Clinical Chemistry 00:0
1–10 (2023)

Special Report

Recommendations for Setting a Criterion for Assessing Commutability of Secondary Calibrator Certified Reference Materials

W. Greg Miller,[a,*] Thomas Keller,[b] Jeffrey Budd,[c] Jesper V. Johansen,[d] Mauro Panteghini,[e] Neil Greenberg,[f] Vincent Delatour,[g] Ferruccio Ceriotti,[h] Liesbet Deprez,[i] Robert Rej,[j] Johanna E. Camara,[k] Finlay MacKenzie,[l] Alicia N. Lyle,[m] Eline van der Hagen,[n] Chris Burns,[o] Pernille Fauskanger,[p] and Sverre Sandberg,[p,q,r] for the IFCC Working Group on Commutability in Metrological Traceability

A secondary higher-order calibrator is required to be commutable with clinical samples to be suitable for use in the calibration hierarchy of an end-user clinical laboratory in vitro diagnostic medical device (IVD-MD). Commutability is a property of a reference material that means results for a reference material and for clinical samples have the same numeric relationship, within specified limits, across the measurement procedures for which the reference material is intended to be used. Procedures for assessing commutability have been described in the literature. This report provides recommendations for establishing a quantitative criterion to assess the commutability of a certified reference material (CRM).

The criterion is the maximum allowable noncommutability bias (MANCB) that allows a CRM to be used as a calibrator in a calibration hierarchy for an IVD-MD without exceeding the maximum allowable combined standard uncertainty for a clinical sample result ($u\text{max}_{CS}$). Consequently, the MANCB is derived as a fraction of the $u\text{max}_{CS}$ for the measurand. The suitability of an MANCB for practical use in a commutability assessment is determined by estimating the number of measurements of clinical samples and CRMs required based on the precision performance and nonselectivity for the measurand of the measurement procedures in the assessment. Guidance is also provided for evaluating indeterminate commutability conclusions and how to report results of a commutability assessment.

Introduction

Clinical laboratory test results are used for making decisions regarding the medical condition of patients. Equivalent results for clinical samples (CSs) among different end-user in vitro diagnostic medical devices (IVD-MDs) used in clinical laboratories are important for making medical decisions using clinical practice guidelines and decision thresholds. Establishing metrological traceability of results for CSs to a certified reference material (CRM), used as a secondary commutable calibrator in position m.3 in the calibration hierarchies described in the International Organization for Standardization (ISO) standard 17511:2020 (1), is an accepted approach to achieve equivalent results for the CSs irrespective of the IVD-MD used for making measurements. A higher-order commutable secondary calibrator used in the calibration hierarchies of IVD-MDs

[a]Department of Pathology, Virginia Commonwealth University, Richmond, VA, United States; [b]ACOMED Statistic, Leipzig, Germany; [c]Jeff Budd Consulting, St. Paul, MN, United States; [d]Radiometer Medical ApS, Copenhagen, Denmark; [e]Research Centre for Metrological Traceability in Laboratory Medicine, University of Milan, Milan, Italy; [f]Neil Greenberg Consulting, LLC, Rochester, NY, United States; [g]Laboratoire national de métrologie et d'essais, Paris, France; [h]Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy; [i]European Commission, Joint Research Centre, Directorate F, Geel, Belgium; [j]Department of Biomedical Sciences, School of Public Health, University at Albany, State University of New York, Albany, NY, United States; [k]National Institute of Standards and Technology, Gaithersburg, MD, United States; [l]Birmingham Quality/UK NEQAS, University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom; [m]Centers for Disease Control and Prevention, Atlanta, GA, United States; [n]Queen Beatrix Hospital, Winterswijk, the Netherlands; [o]National Institute for Biological Standards and Control, A Centre of the MHRA, Hertfordshire, United Kingdom; [p]Norwegian Organization for Quality Improvement of Laboratory Examinations (Noklus), Haraldsplass Deaconess Hospital, Bergen, Norway; [q]Norwegian Porphyria Centre, Department of Medical Biochemistry and Pharmacology, Haukeland University Hospital, Bergen, Norway; [r]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway.

*Address correspondence to this author at: PO Box 980286, Richmond, VA 23298-0286, United States. E-mail greg.miller@vcuhealth.org.

Disclaimer: The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention or the Agency for Toxic Substances and Disease Registry. Use of trade names is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention, the Public Health Service, or the US Department of Health and Human Services.

Received April 3, 2023; accepted June 12, 2023.
https://doi.org/10.1093/clinchem/hvad104

may be a CRM consisting of the measurand in a matrix that closely simulates that of CSs or may be prepared by adding a pure substance CRM into a matrix suitable for use as a calibrator. In either case, a secondary calibrator is required to be commutable with CSs for all of the end-user IVD-MDs for which it is used as part of their calibration hierarchy (2, 3). In addition, the calibrator needs to be commutable with CSs when used with a manufacturer's selected, or standing, measurement procedure (MP) in a calibration hierarchy when that MP is different than the end-user IVD-MD (3). If there are differences between the selected or standing MP and the end-user IVD-MD, then the manufacturer of the IVD-MD is responsible for determining the suitability for use of the CRM in its calibration hierarchy.

Commutability is a property of a reference material that means results for a reference material and for CSs have the same numeric relationship, within specified limits, across the measurement procedures for which the reference material is intended to be used. Experimental designs for assessing commutability have been published (2, 4–6). For the difference in bias approach (4), IVD-MDs are evaluated in pairs, and a CRM is considered commutable for use with those IVD-MDs that consistently meet the criterion when compared across most pairs in the assessment. For the calibration effectiveness approach (5), all IVD-MDs are evaluated simultaneously, and a CRM is considered commutable for use with those IVD-MDs that fulfill the criterion. A CRM is considered noncommutable for use with an IVD-MD for which the magnitude of noncommutability bias plus its uncertainty is outside the criterion. In such cases, the CRM may still be suitable for use if a correction for noncommutability bias with small enough uncertainty is included in the calibration hierarchy for such an IVD-MD (7).

In this report, we provide a process to determine a quantitative criterion for acceptable commutability of a CRM, or of a calibrator prepared from a pure substance CRM by dilution in a suitable matrix. The criterion is based on an allowable contribution from noncommutability bias in the calibration hierarchy of an end-user IVD-MD such that the total combined uncertainty for the CS result allows the maximum allowable combined standard uncertainty for the CS ($u\text{max}_{CS}$) to be fulfilled. The same approach is applicable for trueness controls used to verify the calibration traceability of an end-user IVD-MD.

Propagation of Measurement Uncertainty to a Clinical Sample Result in the Calibration Hierarchy of an IVD-MD

There is a measurement uncertainty (MU) associated with each step in a calibration hierarchy. The MU of each calibration step is propagated through the calibration hierarchy and contributes to the combuined MU at each step. There is a combined MU for the CS result that reflects the cumulative uncertainty contributions from all steps in the calibration hierarchy. A $u\text{max}_{CS}$ is intended to represent an acceptable risk of harm when the CS result is used for medical decisions. Note that usually the MU is expressed as standard uncertainty (u), corresponding to the 1 SD (absolute) or the CV (relative) of a material's assigned value. The expanded uncertainty defines an interval where the true value is expected to lie with a stated level of probability. A coverage factor (k) indicates the expansion, where $U = k \cdot u$. A k of 2 is typically used to define a coverage probability of 95.45%. How to establish an $u\text{max}_{CS}$ is out of scope for this report, and the reader is referred to guidance in the literature (8, 9).

When a commutability assessment is performed, the difference in bias between a CRM and CS results measured for a pair of IVD-MDs is determined and is called the noncommutability bias. The uncertainty associated with an acceptable noncommutability bias will be propagated through the calibration hierarchy to the CS result but is small enough that $u\text{max}_{CS}$ is fulfilled. A decision that a CRM is commutable with CSs is made when the noncommutability bias is within a specified criterion that is consistent with the noncommutability bias being small enough that the $u\text{max}_{CS}$ is fulfilled without correcting for that amount of noncommutability bias. The magnitude of a noncommutability bias is determined at a point in time when a commutability study is performed. In current practice, we assume that the noncommutability bias observed in a commutability assessment is representative of what would be observed if the assessment were repeated and does not change throughout the life of a CRM (7). This assumption implies that the IVD-MDs used in the commutability assessment have no changes regarding reagent formulation or operating parameters.

Determining the Criterion for Assessing Commutability of a CRM

A commutability assessment typically includes a number of IVD-MDs with varying performance characteristics for selectivity and repeatability. Consequently, the magnitude of noncommutability biases will vary from one IVD-MD to another. Each IVD-MD can have a different noncommutability bias because the influence of the CRM matrix on the measurement response of each IVD-MD can be different. An acceptable noncommutability bias can have any value within an interval where the criterion is met. Consequently, for commutability assessment, the noncommutability biases from a group of IVD-MDs are assumed to be randomly distributed. Therefore, the acceptable limit of the randomly

distributed individual IVD-MD noncommutability biases is expressed as the maximum allowable u from noncommutability ($u\text{max}_{NC}$). The maximum allowable noncommutability bias (MANCB) is derived from the $u\text{max}_{NC}$ which is derived from the $u\text{max}_{CS}$. Thus, the MANCB is derived from the $u\text{max}_{CS}$.

There is no preset fraction of $u\text{max}_{CS}$ that is assigned to each step in a calibration hierarchy. The MU from each step can have any reasonable value as long as the combined u for the CS fulfills the $u\text{max}_{CS}$. A model has been proposed to estimate allowable MU in the calibration hierarchy by partitioning the $u\text{max}_{CS}$ among 3 major steps: the MU of the value assigned to a commutable CRM, the MU of the value assigned by the IVD-MD manufacturer to the end-user calibrator, and the MU of the CS that includes all uncertainty sources associated with the IVD-MD measurement process and the individual medical laboratory's operation of the IVD-MD to produce a result for a CS (10, 11). As shown in Fig. 1, we modified this model to include a MU contribution from the allowable interval of noncommutability biases. Historically, an allowable noncommutability bias of a commutable CRM has been assumed to be negligible and thus not corrected for in the calibration hierarchy. However, to set a MANCB requires that an estimate of an acceptable noncommutability bias be made. In Fig. 1, we propose assigning 3/8 of the $u\text{max}_{CS}$ to the allowable individual standard uncertainty for the $u\text{max}_{NC}$, with the assumption that the allowable noncommutability bias can be adequately represented as an uncertainty contribution to $u\text{max}_{CS}$. This model is intended to allow sufficient u for each step in the calibration hierarchy. A different fraction for any of the sources of u, including the $u\text{max}_{NC}$, is allowed as long as the $u\text{max}_{CS}$ is fulfilled.

Once $u\text{max}_{NC}$ is determined, the MANCB can be derived. Using the difference in bias approach (4), the acceptable noncommutability bias is allowed to have any random value in the interval [–MANCB; +MANCB]. A rectangular distribution gives the most conservative (largest) estimate of such variability with width (W) which has the SD indicated in Eq. 1 (12):

$$SD = \frac{W}{2\sqrt{3}} \quad (1)$$

Since $W = 2 \cdot \text{MANCB}$ (i.e., −MANCB to +MANCB), the SD is equal to:

$$SD = \frac{2 \cdot MANCB}{2 \cdot \sqrt{3}} \quad (2)$$

Since $u\text{max}_{NC}$ is the SD for the distribution of acceptable noncommutability values, the MANCB is:

$$MANCB = \sqrt{3} \cdot u\text{max}_{NC} \quad (3)$$

Using the difference in bias approach for commutability assessment (4), any noncommutability bias plus its CI falling entirely within the interval [−MANCB; +MANCB] is accepted, and the CRM is considered commutable with CSs and suitable for use with all combinations of IVD-MD pairs in the commutability assessment that satisfied this criterion. The CRM commutability is considered "inconclusive" for use with an IVD-MD if the CI for the difference in bias overlaps the criterion. A CRM determined to be "inconclusive" needs further evaluation as discussed in a later section. A CRM is considered noncommutable with CSs for IVD-MDs in a pairwise comparison for which the noncommutability bias plus its CI exceeds the MANCB criterion. A CRM determined to be noncommutable is not suitable for use in the calibration hierarchies of IVD-MDs unless a correction for noncommutability bias is included (7).

In the calibration effectiveness approach (5), each IVD-MD is calibrated using the CRM in its calibration hierarchy. The average difference (bias) from an expected value is determined for each CS for each IVD-MD. If the CS bias values are normally distributed, the mean bias, and its CI, is determined for each IVD-MD. If, due to nonselectivity of an IVD-MD for the measurand in some CS, the distribution is skewed or has high kurtosis, then the median bias, and its CI, is determined. The IVD-MDs are sorted in order of their average biases and the highest bias minus the lowest bias is computed as the inter-MP bias range. The CRM is considered commutable for use with those IVD-MDs for which the inter-MP bias range does not exceed 2 · MANCB. After excluding IVD-MDs where the CRM is not commutable with CSs, the CRM commutability is considered "inconclusive" for those IVD-MDs whose bias estimates and their associated CIs overlap the criterion.

Evaluating the MANCB Criterion for Commutability Assessment as Practical for Use

Before conducting a commutability experiment, the MANCB should be assessed for its suitability for use with a practical experimental design to demonstrate that the CRM is commutable with CSs within predefined limits. As described in the Supplemental Material Parts 2 and 3, the statistical approach is a test for equivalence with the null hypothesis that the CRM is not commutable with CSs. Sample sizes are estimated to determine the number of CSs and the number of replications of both CS and CRM measurements. Because commutability assessment is a complex experimental design, closed sample size formulas using effect size, SD(s),

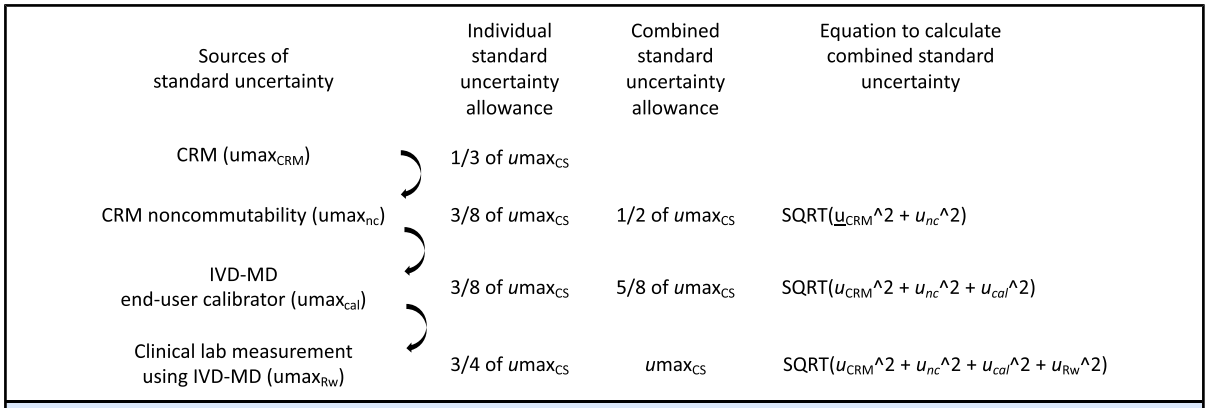

Fig. 1. Allocation of standard measurement uncertainty (*u*) to the major steps of a calibration hierarchy. *u*max$_{CS}$ represents the maximum allowable combined *u* for a CS result. The other *u* terms are what are allowable for each contributor, assuming the model allocation is applicable, and are derived from the *u*max$_{CS}$. The *u*max$_{CRM}$ is the uncertainty of the value assigned to the commutable CRM, *u*max$_{nc}$ is the uncertainty from noncommutability of the CRM, *u*max$_{cal}$ is the uncertainty of the value assigned to the end-user calibrator, and *u*max$_{Rw}$ is the uncertainty estimate based on data obtained under intermediate precision conditions of measurement as defined in ISO/TS 20914 (Medical laboratories—practical guidance for the estimation of measurement uncertainty. ISO, Geneva, Switzerland, 2019). Contributions to *u* at positions below the *u*max$_{CRM}$ are shown as individual contributions from that step and for the combined *u* including that step. Note that the diagram is not itself a calibration hierarchy. This figure is expanded from the concept described in (10).

and alpha and beta error are not available. Consequently, simulations were used to investigate the number of CSs and replications needed to achieve a predefined probability of a successful experiment. This probability is known as the power of the experiment and is typically chosen as 80% or 90%. The power can be estimated by performing simulations for example data sets using the analysis program provided in the difference in bias approach (4). The simulations were based on estimates of precision of CS and CRM measurements, magnitude of acceptable noncommutability bias, and the influence of nonselectivity for the measurand, which was assumed to be an additional normally distributed random error contributing to the measured values of the CSs. The simulations were performed by applying scenarios of varying numbers of CSs and replicate measurements and evaluated as the proportion of simulations when a conclusion of "commutable" was reached for a given MANCB. This proportion is the statistical power of the experiment. The number of CSs and replicates needed to reach the desired power can then be evaluated for their suitability to be implemented. Details and a table showing possible experimental designs are described in the Supplemental Material Part 3.

Because simulation requires advanced statistical support, we developed a table of sample sizes based on an alternative estimation using the two one-sided *t*-test (TOST) equivalence test for a series of commonly encountered commutability assessment designs. The sample size estimation via the TOST approach provides the overall total numbers of measurements of CS and CRM but ignores the clustered variance structure associated with a specified number of replicate measurements of a CS or CRM. Consequently, the user must allocate the total number of measurements into a reasonable number of CSs and replicates of each CS. For example, TOST might indicate that 40 measurements are needed but does not distinguish between 1 single measurement of 40 CSs, 4 replicate measurements of 10 CSs, or 40 replicate measurements of 1 CS. However, TOST provides an easy-to-use table of possible numbers of measurements to estimate the size and practicality of an experimental design. If needed, nonselectivity for the measurand can be taken into account by adding an additional normally distributed random error component to the SD for the CS measurement results. Details for using the TOST approach and a table showing possible experimental designs are described in the Supplemental Material Part 2.

From simulations shown in Supplemental Table 3.2 in the supplemental material, the smallest numbers of samples and measurements are needed when the same number of measurements are made for the CRM and for the CSs. However, in most cases, it is not logistically

practical to have the same number of measurements for the CRM and for the CSs. Consequently, the TOST approximation of sample sizes in Supplemental Table 2.2 includes lower ratios of 1:12, 1:8, and 1:4, i.e., 1 CRM measurement for every 12, 8, and 4 CS measurements, respectively. Ratios lower than 1:4 have little effect on the number of CRM measurements needed and a large influence on increasing the number of CS measurements needed. A ratio of 1:4 is a practical choice for an effective commutability assessment.

Examples Deriving the Commutability Criterion MANCB and Evaluating the Practicality of a Commutability Assessment Experiment Using a Given MANCB

The examples that follow evaluate the practicality of an MANCB using the difference in bias approach for commutability assessment (4). The Supplemental Material Part 1 shows the evaluation using the calibration effectiveness approach (5). Note that this report describes how to derive a commutability assessment criterion (MANCB) from a $u\text{max}_{CS}$ but does not address how to determine the $u\text{max}_{CS}$, which is described elsewhere (8, 9).

SERUM CREATININE

For serum creatinine we used a $u\text{max}_{CS}$ of 7.6% based on recommendations from the National Kidney Disease Education Program (13). Using the model in Fig. 1, the $u\text{max}_{NC}$ is 2.85%. Using Eq. 3, the MANCB is 4.94%. For the difference in bias approach, this criterion is ±0.05 mg/dL at 1.0 mg/dL (4.6 μmol/L at 88.5 μmol/L) and ±0.20 mg/dL at 4.0 mg/dL (17.5 μmol/L at 354 μmol/L). Based on internal quality control data for serum creatinine (14), the intermediate within-laboratory SD was approximately 0.02 mg/dL at 0.8 mg/dL, 2.5% CV (1.8 μmol/L at 71 μmol/L) and 0.06 mg/dL at 7.0 mg/dL, 0.9% CV (5.3 μmol/L at 619 μmol/L).

The sample sizes necessary for the commutability experiment for serum creatinine were investigated using TOST and simulation approaches. The measurement SD (or CV) as a fraction of MANCB for both IVD-MDs in a comparison for commutability assessment was 0.05 (2.5%/4.94%), thus the sample sizes in row 4 of Supplemental Table 2.2 based on TOST apply (see blue highlighted cells). For 90% power, 1:4 ratio of CRM:CS and noncommutability bias 0.33 of MANCB (which leaves the remaining 0.67 of the MANCB to accommodate the CI for the noncommutability bias), 14 measurements of CRM and 56 measurements of CSs are needed.

For simulation under the same assumptions, Supplemental Table 3.2, scenario 7 (see blue highlighted cells) shows that the CRM measured in triplicate in 5 positions distributed among the CSs, and 20 CSs measured in triplicate were needed, which is a total 15 measurements of CRM and 60 measurements of CSs. The 2 approaches gave similar estimates of the total number of CRM and CS measurements needed.

The estimates suggest a reasonable experimental design and we conclude the MANCB of 4.94% (rounded to 5%) is reasonable for assessing commutability of a CRM for creatinine without considering nonselectivity of IVD-MDs for creatinine.

The potential influence of nonselectivity of IVD-MDs can be estimated by arbitrarily increasing the SD for the measurement of CSs to reflect an increase in random variability of the biases for CSs caused by nonselectivity for the measurand. The magnitude of increase should consider the analytical capability of the IVD-MDs in the commutability assessment but is an assumption only used to determine if the MANCB is realistically suitable for use. If the nonselectivity is known to be too large, a MP will be excluded from commutability assessment (2).

For the creatinine example, the SD for CS measurements was arbitrarily doubled to reflect the potential influence of nonselectivity. Using the same measurement procedure parameters as previously described, the TOST estimation in Supplemental Table 2.2, row 6 (see green highlighted cells) shows that a total of 20 measurements of CRM and 80 measurements of CS are needed. The simulation results when the same SD contribution for nonselectivity in CS measurements is included is shown in Supplemental Table 3.3, scenario 7 (see green highlighted cells). In this simulation, the CRM measured in triplicate in 8 positions and 30 CS measured in triplicate were needed, which is a total 24 measurements of CRM and 90 measurements of CSs. The 2 approaches gave similar estimates of the total number of CS and CRM measurements needed. Assuming a larger SD or increased power would increase the total number of measurements needed. Note that when nonselectivity is present, more CSs are needed, not more replicates of the same number of CSs because replication does not reduce the influence of nonselectivity. The size of the experiment remains reasonable, suggesting that the MANCB for creatinine is suitable for use.

SERUM SODIUM

For serum sodium, we used 0.4% for $u\text{max}_{CS}$, which is the minimum value based on biological variability of this measurand (0.75×median within-individual CV) (15). Using the model in Fig. 1, the $u\text{max}_{NC}$ is 0.15%. Using formula 3, the MANCB is 0.26%. For the difference in bias approach, this criterion is ±0.36 mmol/L at 140 mmol/L. Based on internal

quality control data for serum sodium (14), the intermediate within-laboratory SD was approximately 1 mmol/L between 114 and 160 mmol/L, which is CV 0.7% at 140 mmol/L. The sample sizes necessary for the difference in bias commutability experiment for sodium were estimated using TOST. The SD for all measurements was 2.7 times MANCB (0.7%/0.26%). The sample sizes in row 46 of Supplemental Table 2.2 for a 2–2 ratio of SDs were used (see yellow highlighted cells). We used the 2–2 ratio because 2.7 exceeds 2 and the number of replicates at or above a 2–2 ratio are very large, making the 2–2 row suitable for determining the feasibility of a commutability assessment experiment. Table 1 shows that 140 measurements of CRM and 560 measurements of CS would be needed for a 1:4 allocation design, 80% power, and noncommutability bias that is 0.33 of the MANCB from Supplemental Table 2.2. We conclude that the experimental design using this MANCB is not realistic. When an initially proposed MANCB is unrealistic, a larger value needs to be used and the influence of the noncommutability bias on fulfilling $u\text{max}_{CS}$ needs to be investigated and stated. Table 1 shows several possible MANCB values and the corresponding numbers of measurements needed using the TOST approach from Supplemental Table 2.2. An MANCB of 0.56% is likely not achievable because of the cost of the total number of measurements and the typical fill volume of a CRM may require an excessively large number of vials to support the assessment. If the MANCB is increased to 0.70%, the experimental design may be feasible with 36 CRM and 144 CS measurements, although this is a fairly large number of measurements with associated cost and logistical challenges.

If an 0.70% MANCB was used for sodium, then the influence of the noncommutability bias on fulfilling $u\text{max}_{CS}$ needs to be investigated. The noncommutability bias of a single IVD-MD could be close to 0 or just within the MANCB criterion for the CRM to be considered commutable for use with that IVD-MD. When the acceptable noncommutability biases of all IVD-MDs in an assessment are considered, the MANCB can be treated as a random contribution to the combined u of the CS results. Following the approach in Fig. 1, Table 2 shows the combined u for the final CS result using the larger 0.7% MANCB. The combined u of 0.84% for a CS result is approximately double the 0.4% $u\text{max}_{CS}$ criterion based on the biological variability model. Serum sodium is an example of a measurand with stringent $u\text{max}_{CS}$ for which it is challenging to demonstrate commutability of a CRM with CSs. Suggesting alternatives to the $u\text{max}_{CS}$ for sodium is beyond the scope of this report on determining the MANCB criterion, but such an investigation may be indicated (16).

Discussion and Recommendations

SETTING A CRITERION FOR COMMUTABILITY ASSESSMENT

The MANCB is established by first setting a $u\text{max}_{CS}$ for the CS result. Following Fig. 1, the $u\text{max}_{NC}$ is set as 3/8 of the $u\text{max}_{CS}$ assuming the magnitude of noncommutability bias is randomly distributed among the IVD-MDs in a commutability assessment. Calculate the MANCB as $\sqrt{3}\cdot\ u\text{max}_{NC,}$ which incorporates the random distribution of noncommutability biases. Finally, determine that the MANCB is consistent with a reasonable number of CS and CRM measurements in a commutability assessment experiment using the TOST table, a TOST statistical application or a simulation approach based on the performance characteristics of the IVD-MDs in the study. Consequently, the MANCB is related to the $u\text{max}_{CS}$ which is intended to represent clinically relevant measurement results.

CONSIDERATIONS WHEN THE CRITERION FOR COMMUTABILITY ASSESSMENT IS NOT PRACTICAL FOR USE

If the experimental design required to use an MANCB is not reasonable from a logistical or cost perspective, i.e., a large number of CSs or replicates is needed, then a larger MANCB needs to be used. The influence of a larger MANCB on the combined u for CS results needs to be considered from the perspective of potential increased risk of harm to a patient from larger acceptable noncommutability biases when the CRM is used in the calibration hierarchies of IVD-MDs.

Specification of $u\text{max}_{CS}$ is challenging where different models exist and within each model different "submodels" can be presented (8). In principle, the preferred model depends on the measurand and its biological and clinical characteristics (8, 9). The commonly accepted approaches to determine $u\text{max}_{CS}$ may have important limitations (17–23). In situations when the MANCB will not allow the $u\text{max}_{CS}$ to be fulfilled, the $u\text{max}_{CS}$ may need to be reconsidered for its suitability and a different approach considered to establish the $u\text{max}_{CS}$. A MANCB that will allow a CRM to be useful in a calibration hierarchy to improve harmonization among CS results may be sufficient justification to establish an MANCB criterion for commutability assessment even if an $u\text{max}_{CS}$ is not fulfilled. It is important that scientists and the medical profession cooperate to set analytical performance specifications for different measurands and for different clinical situations and thereby establish a $u\text{max}_{CS}$ that is suitable for medical decisions.

ASSESSING INCONCLUSIVE COMMUTABILITY DECISIONS

The conclusion about commutability is "inconclusive" when the CI overlaps the MANCB criterion. Inconclusive means the experiment did not provide an

Table 1. Number of measurements of CSs and CRM needed for different MANCB values for sodium.

			80% Power; 1:4 (CRM:CS); NCB is 0.33 of MANCB		Possible experimental design			
MANCB (%)	CV (0.7%)/ MANCB	Table S2.2 row	Total number of CRM measurements	Total number of CS measurements	CRM positions in run	CRM Rep	CS	CS Rep
0.26	2.0	46	140	560	28	5	112	5
0.56	1.25	27	55	220	11	5	44	5
0.70	1.0	20	36	144	9	4	36	4
0.93	0.75	12	21	84	7	3	28	3

Abbreviations: NCB, noncommutability bias; Rep, repetitions.

Table 2. Calculation of combined uncertainty of clinical sample results for sodium based on an MANCB of 0.7%.

Source of *u* (see Fig. 1)	Individual *u* (%)	Combined *u* (%)	Comment
u_{CRM}	0.2		From reference (10)
u_{nc}	0.4	0.45	Eq. 3 and 0.7% MANCB
u_{cal}	0.15	0.47	3/8 of $umax_{CS}$ per Fig. 1
u_{Rw}	0.7	0.84	From reference (14) at 140 mmol/L

Abbreviations: nc, noncommutability; cal, calibrator; u_{Rw}, uncertainty for end-user clinical laboratory result based on data obtained under intermediate precision conditions of measurement (see Fig. 1).

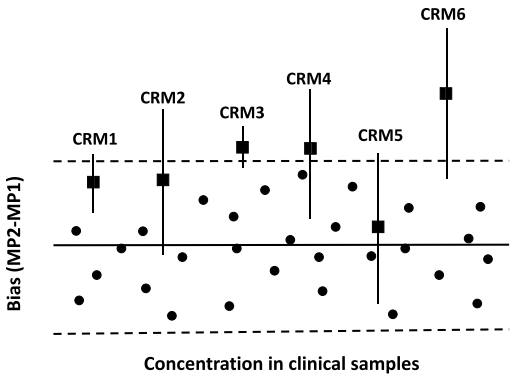

Fig. 2. Examples of possible inconclusive commutability assessment situations using the difference in bias approach. The solid line is the mean bias between the 2 measurement procedures for CSs shown as solid circles and the dashed lines are the MANCB. The solid squares are the mean bias for each CRM and the error bars are the 95% CI for the difference in bias between each CRM and the CSs. Note that the example CRMs are shown vs the same underlying concentration interval of clinical sample results for convenience in the figure, but each CRM is not from the same commutability assessment, which would have the same CI for the difference in bias between each CRM and the clinical samples.

unequivocal decision regarding a CRM being commutable or noncommutable with CSs. The relationship between measurement imprecision, including specimen specific influences, and MANCB determines the frequency that inconclusive conclusions are encountered. Larger imprecisions and smaller MANCB lead to more frequent inconclusive commutability decisions.

The user of a CRM needs to determine if a CRM with an inconclusive commutability status is suitable for use in a calibration hierarchy. Figure 2 shows examples of several inconclusive situations. Example CRM1 has mean noncommutability bias within the MANCB with only a small fraction of the CI exceeding the MANCB. In this example, the relatively small excess uncertainty would likely not cause the $umax_{CS}$ to be exceeded, but its contribution to $umax_{CS}$ should be checked. Example CRM2 has approximately the same noncommutability bias estimate as CRM1, but a much larger fraction of its CI is beyond the MANCB. For CRM2, the larger uncertainty makes it more likely that the $umax_{CS}$ is exceeded unless other contributors to $umax_{CS}$ are small enough to accommodate the

uncertainty in the noncommutability bias. Examples CRM3 and CRM4 are analogous to the preceding except the estimated noncommutability bias exceeds the MANCB increasing the probability that $u\text{max}_{CS}$ may not be fulfilled. Example CRM5 has a very small estimated noncommutability bias with a large CI that is almost contained within the MANCB so would be handled similarly to CRM1. Example CRM6 has a large estimated noncommutability bias and a large CI and would be the most likely to cause the $u\text{max}_{CS}$ to be exceeded. In all the "inconclusive" examples, a CRM with inconclusive commutability may be suitable for use when the combined u for the CS results, including the u of the noncommutability bias, the u from other sources in the calibration hierarchy, and the u for operation of the IVD-MD, is within the $u\text{max}_{CS}$. A CRM with inconclusive commutability can be used if the $u\text{max}_{CS}$ will be fulfilled. Otherwise, a correction for the noncommutability bias can be added to the calibration hierarchy (7).

REPORTING THE RESULTS OF A COMMUTABILITY ASSESSMENT

We have previously recommended that the CRM certificate of analysis include all IVD-MDs for which a CRM has been assessed for commutability and those for which the CRM was commutable with CSs (2). We recommend that all CRMs have a certification report that compliments the certificate of analysis by providing detailed information on the commutability assessment along with documentation of other aspects of the CRM production and characterization. The CRM certificate or certification report must include the MANCB criterion used in the commutability assessment. Although the magnitudes of noncommutability bias will vary among IVD-MDs for which the CRM is commutable with CSs, the MANCB reflects the maximum potential contribution of noncommutability bias to the combined u of CS results measured using all IVD-MDs.

The CRM certification report should include the magnitude of noncommutability bias and its u for each IVD-MD in a commutability assessment including those for which the CRM was determined to be commutable for use and those with indeterminate or noncommutable conclusions. This detailed information will assist IVD manufacturers to determine how to use the CRM in a calibration hierarchy and if a correction for noncommutability or inconclusive commutability should be considered.

As previously recommended (2, 4, 5), some IVD-MDs with inadequate performance may need to be excluded from assessing commutability of a CRM. In such cases, the CRM would not be suitable for use with those excluded IVD-MDs. We recommend that IVD-MDs excluded from a commutability assessment be identified in the certification report to provide potential users of the CRM with all available information regarding its suitability for use.

LIMITATIONS

Selectivity of IVD-MDs for the measurand is a challenge that affects the estimated noncommutability bias as well as the selection of CS for use in a commutability assessment. Poor selectivity causes sample specific influences that may contribute a bias to some individual CS results. The magnitude of a sample specific bias for an individual CS is a function of the amount of the particular influence quantity in an individual CS. The variable magnitude of sample specific biases across a set of CSs is usually observed as random variability in the set of CSs. Excessive sample specific influences can cause an unacceptable large noncommutability bias and/or an associated large CI leading to an inconclusive commutability decision, or can be a reason to exclude an IVD-MD from a commutability assessment. In either case, the CRM is considered not suitable for use with such an IVD-MD. The calibration effectiveness approach (5) reduces the effect of nonselectivity on results by taking the median value over all CSs, which minimizes the influence of those CSs with larger sample specific biases.

The model described to derive an MANCB applies for all types of measurands including complex and heterogeneous molecules such as protein complexes or situations when nonselectivity changes across the measuring interval due to altered molecular forms in pathological conditions. The influence of nonselectivity of MPs for such measurands will likely increase the magnitude of the MANCB and the experimental design needed to assess commutability. The amount of nonselectivity that can be accepted should be established when designing the commutability assessment experiment. The suitability of an MANCB needs to be evaluated in relation to the influence of noncommutability on the $u\text{max}_{CS}$ specification.

If the required MANCB is small, the u required for the experimental design may not be achievable at an acceptable cost regarding the number of CSs and replicate measurements of each CS and CRM. A larger MANCB may be needed to have a feasible experimental design. In this situation, the potential influence of a larger MANCB on the u of CS results needs to be considered and stated in the certificate or the certification report. Consideration of the MANCB in the context of improvement in harmonization of CS results made possible by a CRM, and lacking suitable alternatives, may be considered even if the $u\text{max}_{CS}$ is not fulfilled.

Sourcing a sufficient number of single-donation CSs to cover the appropriate concentration interval

may be challenging in particular when the value assigned to the CRM is at pathological concentrations. A CRM producer should consider collaborating with other CRM producers, external quality assessment providers, and IVD manufacturers to cost share by including other candidate materials in the commutability assessment. Potential drawbacks of this approach are that inclusion of more materials may require more CSs to suitably cover the concentration interval needed, and the study can become very large as the CRM replicate measurements from Supplemental Table 2.2 need to be made for each individual CRM.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

Nonstandard Abbreviations: CS, clinical sample; IVD-MD, in vitro diagnostic medical device; CRM, certified reference material; ISO, International Organization for Standardization; MP, measurement procedure; u, standard uncertainty; $u\text{max}_{CS}$, combined standard uncertainty for the clinical sample; MU, measurement uncertainty; k, coverage factor (for expanded uncertainty); $u\text{max}_{NC}$, maximum allowable u from noncommutability ; MANCB, maximum allowable noncommutability bias; TOST, two one-sided t-test.

Author Contributions: *The corresponding author takes full responsibility that all authors on this publication have met the following required criteria of eligibility for authorship: (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; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. Nobody who qualifies for authorship has been omitted from the list.*

Greg Miller (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Thomas Keller (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Jeffrey Budd (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Jesper Johansen (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Mauro Panteghini (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Neil Greenberg (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Vincent Delatour (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Ferruccio Ceriotti (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Liesbet Deprez (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Robert Rej (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Johanna Camara (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Finlay MacKenzie (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Alicia Lyle (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Eline van der Hagen (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Chris Burns (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Pernille Fauskanger (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), and Sverre Sandburg (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal).

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

Research Funding: None declared.

Disclosures: J.E. Camara, National Institute of Standards and Technology Delegate to the Clinical and Laboratory Standards Institute (CLSI). N. Greenberg has received consulting fees from Ortho Clinical Diagnostics. A.N. Lyle, Vice Chair, CLSI EP30 Document Revision Committee. W.G. Miller is Chair for the IFCC Working Group on Metrological Traceability and has received speaking honorarium from Siemens Healthineers and speaking honorarium and travel support from AACC, New York Metro section. S. Sandberg is Chair of the Council of the International Consortium for Harmonization of Clinical Laboratory Results. T. Keller owns a commercial statistics service provider and has business relationships with medical device and pharmaceutical companies.

Role of Sponsor: No sponsor was declared.

Acknowledgment: The authors appreciate the assistance of Stephan Weber at ACOMED statistic, Leipzig, Germany, with the simulations and calculations used in this report.

References

1. International Organization for Standardization. ISO 17511:2020. In vitro diagnostic medical devices—requirements for establishing metrological traceability of values assigned to calibrators, trueness control materials and human samples. 2nd ed. Geneva (Switzerland): International Organization for Standardization; 2020.
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. Miller WG, Greenberg N, Panteghini M, Budd JR, Johansen JV. Guidance on which calibrators in a metrologically traceable calibration hierarchy must be commutable with clinical samples. Clin Chem 2023;69:228–38.
4. 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.
5. Budd JR, Weykamp C, Rej R, MacKenzie F, Ceriotti F, Greenberg N, et al. IFCC working group recommendations for assessing commutability part 3: using the calibration effectiveness of a reference material. Clin Chem 2018;64:465–74.
6. Braga F, Panteghini M. Commutability of reference and control materials: an essential factor for assuring the quality of measurements in laboratory medicine. Clin Chem Lab Med 2019;57:967–73.
7. Miller WG, Budd J, Greenberg N, Weykamp C, Althaus H, Schimmel H, et al. IFCC working group recommendations for correction of bias caused by noncommutability of a certified reference

material used in the calibration hierarchy of an end-user measurement procedure. Clin Chem 2020;66:769–78.
8. Ceriotti F, Fernandez-Calle P, Klee GG, Nordin G, Sandberg S, Streichert T, et al. Criteria for assigning laboratory measurands to models for analytical performance specifications defined in the 1st EFLM strategic conference. Clin Chem Lab Med 2017;55:189–94.
9. Braga F, Panteghini M. Performance specifications for measurement uncertainty of common biochemical measurands according to Milan models. Clin Chem Lab Med 2021;59:1362–8.
10. Braga F, Infusino I, Panteghini M. Performance criteria for combined uncertainty budget in the implementation of metrological traceability. Clin Chem Lab Med 2015;53:905–12.
11. Panteghini M, Braga F, Camara JE, Delatour V, Van Uytfanghe K, Vesper HW, et al. Optimizing available tools for achieving result standardization: value added by Joint Committee on Traceability in Laboratory Medicine (JCTLM). Clin Chem. 2021;67:1590–605.
12. EURACHEM/CITAC. Quantifying uncertainty in analytical measurement. 3rd Ed. EURACHEM/CITAC Guide CG 4. Gembloux (Belgium): EURACHEM/CITAC; 2012.
13. Myers GL, Miller WG, Coresh J, Fleming J, Greenberg N, Greene T, et al. Recommendations for improving serum creatinine measurement: a report from the laboratory working group of the National Kidney Disease Education Program. Clin Chem 2006;52:5–18.
14. Ellis AD, Gross AR, Budd JR, Miller WG. Influence of reagent lots and multiple measuring systems on estimating the coefficient of variation from quality control data; implications for uncertainty estimation and interpretation of QC results. Clin Chem Lab Med 2020;58:1829–35.
15. Aarsand AK, Fernandez-Calle P, Webster C, Coskun A, Gonzales-Lao E, Diaz-Garzon J, et al. The EFLM biological variation database. https://biologicalvariation.eu/ (Accessed December 2022).
16. Oosterhuis WP, Coskun A, Sandberg S, Theodorsson E. Performance specifications for sodium should not be based on biological variation. Clin Chim Acta 2023;540:117221.
17. Oosterhuis WP. Gross overestimation of total allowable error based on biological variation. Clin Chem 2011;57:1334–6.
18. Roraas T, Petersen PH, Sandberg S. Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. Clin Chem 2012;58:1306–13.
19. Aarsand AK, Røraas T, Sandberg S. Biological variation—reliable data is essential. Clin Chem Lab Med 2015;53:153–4.
20. Carobene A, Braga F, Roraas T, Sandberg S, Bartlett WA. A systematic review of data on biological variation for alanine aminotransferase, aspartate aminotransferase and g-glutamyl transferase. Clin Chem Lab Med 2013;51:1997–2007.
21. Marco JD-G, Fernandez-Calle P, Minchinela J, Aarsand AK, Bartlett WA, Aslan B, et al. Biological variation data for lipid cardiovascular risk assessment biomarkers. A systematic review applying the Biological Variation Data Critical Appraisal Checklist (BIVAC). Clin Chim Acta 2019;495:467–75.
22. Coskun F, Braga F, Carobene A, Ganduxe XT, Aarsand AK, Fernández-Calle P, et al. Systematic review and meta-analysis of within-subject and between-subject biological variation estimates of 20 haematological parameters. Clin Chem Lab Med 2019;58:25–32.
23. Braga F, Panteghini M. Biologic variability of C-reactive protein: is the available information reliable? Clin Chim Acta 2012;413:1179–83.