ORIGINAL PAPER

Demonstrating the comparability of certified reference materials

David L. Duewer · Katrice A. Lippa · Stephen E. Long · Karen E. Murphy · Katherine E. Sharpless · Lorna T. Sniegoski · Michael J. Welch · Wataru Tani · Masao Umemoto

Received: 30 March 2009 / Revised: 15 June 2009 / Accepted: 1 July 2009 / Published online: 22 July 2009

Abstract Certified reference materials (CRMs) enable the meaningful comparison of measurement results over time and place. When CRMs are used to calibrate or verify the performance of a measurement system, results produced by that system can be related through the CRM to well-defined, stable, and globally accessible reference(s). Properly done, this directly establishes the metrological traceability of the results. However, achieving the meaningful comparison of results from measurement systems calibrated and/or verified with different CRMs requires that the different materials truly deliver the same measurand, that is, are "the same" within stated uncertainty except for differences in the level of the analyte of interest. We here detail experimental and data analysis techniques for establishing and demonstrating the comparability of materials. We focus on (1) establishing a uniform interpretation of the common forms of CRM uncertainty statements, (2) estimating consistent measurement system response uncertainties from sometimes inconsistent experimental designs, (3) using "errors-in-variables" analysis methods to evaluate comparability studies and novel graphical tools for communicating results of the evaluation to reviewing authorities and potential CRM customers, and (4) augmenting established comparability studies with new materials using measurements provided by the certifying institution. These experimental and data analytic tools were developed in support of the Joint Committee for Traceability in Laboratory Medicine's efforts to enhance the reliability of clinical laboratory measurements and are illustrated with potassium and cholesterol measurands of clinical relevance; however, these tools can be applied to any group of materials that deliver the same nominal measurand with stated value and uncertainty.

Keywords Certified reference material (CRM) · Cholesterol in human serum · Errors-in-variables regression · Graphical communication · Joint Committee on Traceability in Laboratory Medicine (JCTLM) · Metrological traceability · Potassium in human serum

Acronyms and symbols

ANOVA	Analysis of variance
CRM	Certified reference material
FREML	Linear functional relationship estimation by maximum likelihood
GLS	Generalized least-squares regression
ID-GC/MS	Isotope dilution gas chromatography mass spectrometry

Electronic supplementary material The online version of this article (doi:10.1007/s00216-009-2949-8) contains supplementary material, which is available to authorized users.

D. L. Duewer (✉) · K. E. Sharpless
Analytical Chemistry Division,
National Institute of Standards and Technology (NIST),
100 Bureau Drive, Stop 8390,
Gaithersburg, MD 20899-8390, USA
e-mail: david.duewer@nist.gov

K. A. Lippa · L. T. Sniegoski · M. J. Welch
Analytical Chemistry Division,
National Institute of Standards and Technology (NIST),
100 Bureau Drive, Stop 8392,
Gaithersburg, MD 20899-8390, USA

S. E. Long · K. E. Murphy
Analytical Chemistry Division,
National Institute of Standards and Technology (NIST),
100 Bureau Drive, Stop 8391,
Gaithersburg, MD 20899-8390, USA

W. Tani · M. Umemoto
Reference Material Institute for Clinical Chemistry Standards (ReCCS), KSP R&D A205,
3-2-1 Sakato Takatsu,
Kawasaki, Kanagawa 213-0012, Japan

ID-ICPMS	Isotope dilution inductively coupled plasma mass spectrometry
JCCRM	Japanese Clinical Certified Reference Material; a CRM produced by ReCCS
JCTLM	Joint Committee for Traceability in Laboratory Medicine
MC	Monte Carlo
NIST	National Institute of Standards and Technology
ReCCS	Reference Material Institute for Clinical Chemistry Standards
RegViz	Regression visualization
SRM	Standard reference material, a CRM produced by NIST
[Chol]	Concentration of cholesterol in mg/dL
[K]	Concentration of potassium in mmol/L
a	Intercept
b	Slope
c	Certified value
$\hat{c}$	Predicted certified value from a function relating certified values and responses
$F(.)$	Generic function of a specified set of quantity (ies) and parameter(s)
k	Generic expansion factor
$k_{p,\nu}$	Expansion factor for specified percent coverage and number of degrees of freedom (ν)
n_b	Number of independent measurements per material
n_m	Number of different certified materials compared
n_w	Number of replicate measurements per independent measurement
r	Response of a measurement process
$\bar{r}$	Mean response
$\hat{r}$	Predicted response from a function relating certified values and responses
$s(.)$	Standard deviation of a given set of measurements
$s_b(.)$	Between-set imprecision
$s_t(.)$	Total imprecision of a measurement process
$s_w(.)$	Within-set imprecision
t	Generic two-sided Student's t
$t_{p,\nu}$	Student's t for specified two-sided percent probability and number of degrees of freedom
$u(.)$	Combined uncertainty associated with a specified quantity
$u_\infty(.)$	Large-sample combined uncertainty associated with a specified quantity
$U(.)$	Expanded uncertainty associated with a specified quantity
$U_p(c)$	Expanded uncertainty; the interval $c \pm U_p(c)$ is expected to include the true value of the measurand in all CRM units with a level of confidence of about percent
$U_{95/95}(c)$	Expanded uncertainty; the interval $c \pm U_{95/95}(c)$ is expected to include the true value of the measurand in about 95% of the CRM units with a confidence of about 95%
ν	Effective number of degrees of freedom for a specified estimate
x	Quantity of a specified measurand
$\%s(.)$	Percent relative standard deviation

Introduction

Certified reference materials (CRMs) help enable the meaningful comparison of measurement results over time and place. By definition, a certified value (a quantity and its associated expanded uncertainty) is traceable to some well-defined, stable, and accessible reference(s) such as those provided by the International System of Units [1, 2]. When calibration CRMs (usually either materials of well-characterized composition or simple solutions thereof) are used to calibrate a measurement system, results produced by that system are related through the CRM to the references. When properly done, this directly establishes the *metrological traceability* of the results—formally defined as the "property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty" [3]. Natural-matrix CRMs are sometimes used as calibrants but are more commonly used to verify the metrological traceability of results from an otherwise calibrated measurement system. While the particular CRM(s) used to calibrate and/or verify the calibration of a particular system may not be globally available or may degrade, go out of stock, or otherwise become unavailable over time, new CRMs providing equivalent traceability can in principle be produced. However, achieving the meaningful comparison of results from measurement systems calibrated and/or verified with different CRMs requires that suitable materials are available, accessible, and truly provide equivalent traceability.

While there are international standards for the certification of CRMs [1] and guidance for comparing measurement methods [4], there is little established guidance for evaluating the comparability of CRMs beyond that developed for primary gas standards [5]. To address this need, a protocol for the direct one-time comparison of multiple materials was recently developed [6]. While intended to support the Joint Committee for Traceability in Laboratory Medicine's (JCTLM's) efforts to enhance the reliability of clinical laboratory measurements [7], the protocol can be applied to any group of certified or other materials that deliver a quantity value and its associated uncertainty for the same measurand. However, the current protocol does not fully address data analysis best practice or methods for augmenting the comparison as CRM producers introduce replacement and/or new materials.

We here elaborate on the technical aspects of demonstrating comparability among certified materials that deliver the same nominal measurand. We give particular attention to (1) interpreting CRM uncertainty statements, (2) estimating consistent measurement system response uncertainties, (3) analyzing results obtained from a designed comparability study and communicating those results, and (4) augmenting established comparability studies with new materials using measurements provided by the certifying institution. We propose the use of a well-established but underused "errors-in-variables" data analytic technology for evaluating relationships between two sets of data where both sets have known and significant associated uncertainties. We also introduce novel graphical tools for reporting the comparison results. We illustrate the data analysis methodology and tools with CRMs for potassium and for cholesterol in human sera.

We believe that the proposed comparability analysis methods are well suited for demonstrating the absence of significant differences among a group of suitably similar materials as assessed by a given measurement procedure. However, it must be noted that these methods are neither complete nor completely objective. Should a group of materials be found insufficiently comparable, it may be necessary to consider several different hypotheses and exercise considerable chemical judgment to just identify *which* materials are least comparable. Identifying *why* those materials appear less comparable (e.g., inadequately specified measurand, assigned certified value, or certified uncertainty; systematic biases among the methods used to certify different materials; material degradation; measurement method-specific interferences; analytical blunders; etc.) may require further experimentation. Further, comparability assessed with a single measurement procedure does not and cannot address material commutability—an attribute of the behavior of a material assessed in two or more measurement procedures [3].

Methods and materials

Potassium in human serum

Materials As of 2009, the only JCTLM-listed higher-order CRMs for potassium in human serum have been produced by the National Institute of Standards and Technology (NIST) or Reference Material Institute for Clinical Chemistry Standards (ReCCS) [8]. The four CRMs available from or produced by these organizations during the period 2003–2009 are ReCCS Japanese Clinical Certified Reference Material (JCCRM) 111-5 Certified Reference Material for Ion Selective Electrode, NIST Standard Reference Material (SRM) 909b Human Serum, and NIST SRMs 956a and 956b Electrolytes in Frozen Human Serum. Each unit of these CRMs delivers two or three different materials having different levels of potassium concentration, [K]; Table 1 lists the characteristics for all 11 of these constituent materials.

While multiple measurands were certified in all four CRMs, all but SRM 909b were primarily intended as serum electrolyte standards. SRM 909b has more general utility, was prepared, and is supplied in larger quantity and, as a lyophilized material, is expected to have a longer shelf-life than the frozen materials. JCCRM 111-5, SRM 909b, and SRM 956a were certified using very similar isotope dilution-thermal ionization mass spectrometry methods [9]. SRM 956b was certified using an isotope dilution method employing cold-plasma inductively coupled plasma mass spectrometry (ID-ICPMS) [10].

Comparability studies In early 2003, [K] in the eight materials from the three CRMs then available were evaluated in a study designed for demonstrating comparability. All measurements were made at NIST using ID-ICPMS. Single vials of each material were prepared as directed in their respective certificates. Independently prepared aliquots were evaluated in singlicate under repeatability conditions in two measurement campaigns, each campaign of 1-day duration with a break of 1 day between campaigns. The value and a 95% expanded uncertainty on the value estimated from the measurement model and previously established uncertainty components were reported for each material in each campaign.

In 2004, SRM 956b was prepared as a replacement for SRM 956a, all of which had been sold. All of the certification measurements for SRM 956b were made at NIST using ID-ICPMS. The certified values were assigned from singlicate measurements on six vials of each level. As part of the certification process, measurements were also made on two independent aliquots from the last available vial of each of the three levels of SRM 956a. All measurements for all materials were reported.

Cholesterol in human serum

Materials As of 2009, the only JCTLM-listed higher-order CRMs for cholesterol in human serum also have been produced by NIST or ReCCS [8]. The nine CRMs available from or produced by these organizations during the period between 2003 and 2009 are ReCCS JCCRMs 211-1 and 211-2 Certified Reference Material for Measurement of Total Cholesterol in Human Serum; NIST SRM 909b Human Serum; NIST SRMs 968c and 969d Fat-Soluble Vitamins, Carotenoids, and Cholesterol in Human Serum; NIST SRM 1589a Polychlorinated Biphenyls, Pesticides,

Table 1 Certified reference materials for potassium [K] in human serum, mmol/L

	Ancillary information					Certified value					2003 NIST		2005 NIST	
CRM	Status[a]	Year[b]	Form[c]	mL[d]	°C[e]	c	$U(c)$	Type[f]	ν[g]	$u_\infty(c)$	r	$s_t(r)$	r	$s_t(r)$
JCCRM 111-5H	Out	2001	Fr	1.2	25	5.69	0.02	95	9	0.01	5.702	0.012		
JCCRM 111-5L	Out	2001	Fr	1.2	25	3.25	0.02	95	9	0.01	3.241	0.007		
JCCRM 111-5M	Out	2001	Fr	1.2	25	4.40	0.02	95	9	0.01	4.410	0.009		
SRM 909b-1	Avail	1996	Ly	10	20-25	3.424	0.025	95/95	?	0.011	3.431	0.007		
SRM 909b-2	Avail	1996	Ly	10	20-25	6.278	0.052	95/95	?	0.021	6.267	0.019		
SRM 956a-1	Out	1996	Fr	2	RT	6.008	0.020	95	12	0.010	5.996	0.013	6.014	0.014
SRM 956a-2	Out	1996	Fr	2	RT	3.985	0.020	95	11	0.010	3.993	0.008	4.029	0.008
SRM 956a-3	Out	1996	Fr	2	RT	2.025	0.008	95	11	0.004	2.019	0.004	2.021	0.003
SRM 956b-1	Avail	2004	Fr	2	RT	5.973	0.045	95	>30	0.023			5.973	0.003
SRM 956b-2	Avail	2004	Fr	2	RT	3.983	0.029	95	>30	0.015			3.983	0.002
SRM 956b-3	Avail	2004	Fr	2	RT	1.987	0.014	95	>30	0.007			1.987	0.001

[a] Current status of CRM: “Avail” = available for purchase; “Out” = sold out

[b] Year of original certification

[c] Form of sample matrix: “Fz” = frozen, “Ly” = lyophilized

[d] mL of serum (if frozen serum matrix) or reconstitution volume (if lyophilized) per unit

[e] Use temperature of sample in °C: “RT” = room temperature, unspecified range

[f] Type of expanded uncertainty interval: 95% coverage or 95%/95% tolerance

[g] Number of degrees of freedom: “?” = certificate does not specify

Polybrominated Diphenyl Ethers, and Dioxins/Furans in Human Serum; NIST SRMs 1951a and 1951b Lipids in Frozen Human Serum; and NIST SRM 1952a Cholesterol in Human Serum (freeze-dried). Each unit of these CRMs delivers one, two, or three different materials having different levels of cholesterol concentration, [Chol]; Table 2 lists the characteristics for all 17 of these constituent materials.

Only the JCCRM 211, SRM 1951, and SRM 1952 CRM families are explicitly intended for use as cholesterol standards. While including certified [Chol] values, SRM 909b was designed to provide stable and homogenous materials for a wide variety of clinical measurands; the SRM 968 family is primarily intended to serve clinical and nutritional communities concerned with fat-soluble vitamins; and SRM 1589a is primarily intended for use in evaluating analytical methods for the determination of selected environmental pollutants in tissue matrices. Regardless of the intended usage, [Chol] in all of these CRMs was certified using similar isotope dilution gas chromatography electron impact mass spectrometry (ID-GC/MS) assays [11].

Comparability studies In early 2003, [Chol] in the 12 materials of the six then-available CRMs (JCCRM 211-1, SRM 909b, SRM 968c, SRM 1589a, SRM 1951a, and SRM 1952a) were evaluated in a study designed for demonstrating comparability. All measurements were made at NIST using ID-GC/MS. Measurements were made in two independent measurement campaigns separated by about a week in time and a round of equipment maintenance. Each campaign evaluated all 12 materials; each sample prepared from one vial that was thawed, reconstituted, and/or mixed as specified in the CRM certificate. Each sample was analyzed in duplicate over 2 days under repeatability conditions; the run order on the second day was the reverse of that on the first day.

In 2004, SRM 1951b was prepared to replace SRM 1951a. All of the certification measurements for the SRM 1951b materials were made at NIST using ID-GC/MS. The certified [Chol] values were assigned from duplicate measurements on nine vials of each material. As part of the process, three vials of each of the SRM 1952a levels were analyzed in singlicate.

In 2005, JCCRM 211-2 was prepared to replace JCCRM 211-1. All of the certification measurements for JCCRM 211-2 were made at ReCCS using ID-GC/MS. The certified [Chol] values were assigned from triplicate analyses of at least 38 vials of each material. As part of the certification process, 3 to 13 aliquots of the JCCRM 211-1 and SRM 1951b levels were also analyzed in triplicate.

In 2008, SRM 968d was prepared to replace SRM 968c. The certification measurements for the SRM 968d material were made at NIST using ID-GC/MS. The certified [Chol] values were assigned from duplicate measurements on 12 vials of each material. As part of the process, two or three vials of each of the SRM 1951b levels were analyzed in duplicate.

Table 2 Certified reference materials for cholesterol [Chol] in human serum, mg/dl

CRM	Ancillary information					Certified value					2003 NIST		2004 NIST		2005 ReCCS		2008 NIST	
	Status[a]	Year[b]	Form[c]	mL[d]	°C[e]	c	$U(c)$	Type[f]	ν[g]	$u_\infty(c)$	$\bar{r}$	$s_t(\bar{r})$	$\bar{r}$	$s_t(\bar{r})$	$\bar{r}$	$s_t(\bar{r})$	$\bar{r}$	$s_t(\bar{r})$
JCCRM 211-1H	Out	2001	Fz	0.5	RT	262.4	1.0	95	14	0.5	259.68	0.68			259.68	0.50		
JCCRM 211-1M	Out	2001	Fz	0.5	RT	205.2	0.8	95	14	0.4	204.00	0.24			205.34	0.36		
JCCRM 211-2H	Avail	2005	Fz	0.5	25	230.8	0.4	95	>30	0.2					230.75	0.17		
JCCRM 211-2M	Avail	2005	Fz	0.5	25	191.4	0.3	95	>30	0.2					191.41	0.12		
SRM 909b-1	Avail	1996	Ly	10	20-25	146.4	1.8	95/95	?	0.8	146.31	0.18						
SRM 909b-2	Avail	1996	Ly	10	20-25	235.3	3.0	95/95	?	1.3	233.63	0.21						
SRM 968c-1	Out	1999	Ly	1	RT	133.5	1.3	95	?	0.7	132.24	0.20						
SRM 968c-2	Out	1999	Ly	1	RT	166.9	1.7	95	?	0.9	166.15	0.18						
SRM 968d	Soon	2008	Fz	1	RT	133.5	0.4	95	?	0.2							133.46	0.27
SRM 1589a	Decert	2000	Ly	10	20-25	157.8	0.4	95/95	5	0.2	155.24	0.29						
SRM 1951a-1	Out	1997	Fz	1	20-25	182.15	0.45	95	8	0.23	181.86	0.27	181.92	0.34				
SRM 1951a-2	Out	1997	Fz	1	20-25	276.67	0.55	95	8	0.28	277.05	0.18	277.23	0.44				
SRM 1951b-1	Avail	2004	Fz	1	RT	185.76	0.55	95	?	0.28			185.76	0.07	186.15	0.20	182.87	0.88
SRM 1951b-2	Avail	2004	Fz	1	RT	266.58	0.84	95	?	0.42			266.58	0.11	267.44	0.42	266.55	1.22
SRM 1952a-1	Avail	1990	Ly	3	20-25	147.5	1.4	95/95	12	0.5	147.12	0.18						
SRM 1952a-2	Avail	1990	Ly	3	20-25	233.4	1.4	95/95	12	0.5	230.47	1.22						
SRM 1952a-3	Avail	1990	Ly	3	20-25	333.0	2.4	95/95	12	0.9	328.87	1.59						

[a] Current status of CRM: “Avail” = available for purchase; “Decert” = decertified for this measurand, value taken from original certificate; “Soon” = in process; “Out” = sold out

[b] Year of original certification

[c] Form of sample matrix: “Fz” = frozen, “Ly” = lyophilized

[d] mL of serum (if frozen serum matrix) or reconstitution volume (if lyophilized) per unit

[e] Use temperature of sample in °C: “RT” = room temperature, unspecified range

[f] Type of expanded uncertainty interval: 95% coverage or 95%/95% tolerance

[g] Number of degrees of freedom: “?” = certificate does not specify

CRM certificates

Certificates for all ReCCS CRMs are available at: http://www.reccs.or.jp/e_materials.html; certificates for all NIST SRMs are available at http://www.nist.gov/srm.

Computation

All calculations and data displays used in this study are from freeware or purpose-built spreadsheet systems implemented in Excel 2003 (Microsoft Corp, Redman, WA, USA), running on ordinary PC hardware. The purpose-built software is available on request from the corresponding author.

Results and discussion

In principle, demonstrating whether a series of certified materials successfully deliver the same measurand (i.e., are the "the same" except for differences in the level of the analyte of interest) is *relatively* straight forward: (1) determine a response to the quantity of interest in one or more units of each material with a suitable measurement system over a short period of time under well-controlled measurement conditions, (2) establish a consensus functional relationship between the certified values and measurement responses, and (3) evaluate the significance of any deviations from the relationship given the CRM uncertainties and measurement method imprecision.

Certified reference values

By definition, certified values are stated as a value, c, associated with an expanded uncertainty, $U(c)$, that specifies the quantity of a specifically defined measurand, x. Typically, the interval $c \pm U(c)$ is asserted to include the "true" value of x in a specified proportion of the CRM units with a defined level of confidence. However, the $U(c)$ in different CRMs may specify different proportions or different levels of confidence. Before such materials can be validly compared, all of the reported $U(c)$ must be transformed to have at least approximately the same meaning.

Standardizing expanded uncertainties Expanded uncertainties are generally estimated by combining estimates for multiple uncertainty sources (which may include but are not limited to bias among analytical methods, method imprecision, material heterogeneity, and material instability) into a combined uncertainty, $u(c)$, that has the statistical properties of a standard deviation estimated with some effective number of degrees of freedom, v [12]. The $u(c)$ are expanded to provide the desired coverage by multiplying by an appropriate factor, k: $U(c) = k \cdot u(c)$. The particular value of k is dictated by the v, the desired level of confidence in the coverage, and the nature of the desired coverage interval.

Most commonly, $U(c)$ is stated as a symmetric half-interval such that $c \pm U(c)$ is expected to include the true value of the *average over all units* quantity with a given percent level of confidence; the coverage factor and resulting expanded uncertainty can be denoted $k_{\mathrm{p,v}}$ and $U_{\mathrm{p}}(c)$. Note that the expansion process transforms the sample-based $u(c)$ into a "large-sample" estimate no longer associated with a specific v. The most commonly specified coverage is 95%, but occasionally 99% is specified. When the certificate does not specify the value of k but does explicitly or implicitly specify v, $k_{\mathrm{p,v}}$ can be assumed to be the two-sided normal-distribution Student's t value for percent confidence and v degrees of freedom, $t_{\mathrm{p,v}}$. When neither k nor v is specified, it is likely that a default large-sample expansion factor has been used: k_{95}=2 (which accords well with $t_{95,\infty} \approx 1.96$) or k_{99}=3 (which overestimates the $t_{99,\infty} = 2.58$, perhaps in partial compensation for fatter-than-normal extreme tails typical of many "real life" distributions) [12].

At NIST, many older lyophilized CRMs (where variability in the mass of the sample and in the volume of the liquid used in its reconstitution are concerns) were certified with 95%/95% tolerance intervals where $c \pm U(c)$ is expected to include the true value in about 95% of the *individual fully reconstituted* units with about a 95% level of confidence [13]; here, the coverage factor and resulting expanded uncertainty can be denoted $k_{95/95,v}$ and $U_{95/95}$. For a given v, $U_{95/95}(c)$ will be somewhat larger than $U_{95}(c)$.

In general, when the coverage interval is specified as $U(c)$ and the expansion factor k and v are known, then $U_{95}(c) \approx U(c) \cdot k_{95,v}/k$. When v is not explicitly or implicitly specified, it is reasonable to assume that its value is between about 10 (the minimum number of CRM units currently recommended for homogeneity assessment) and about 30 [1]. Electronic supplementary material Figure S1 displays $k_{95,v}/k_{95/95,v}$ and $k_{95,v}/k_{99,v}$ as functions of v. Given that the $k_{95,v}/k_{95/95,v}$ and $k_{95,v}/k_{99,v}$ ratios are monotonic increasing with increasing v, when the true value for v is unknown, we suggest assuming v=30 to provide the largest defensible ratio value.

Estimating large-sample combined uncertainties When it is desired to combine a CRM uncertainty with estimates from other sources, it is typically necessary to transform $U_{95}(c)$ back to a form that has the properties of a standard deviation, $u(c)$. Unless the degrees of freedom for all uncertainty sources are to be explicitly used in the calculations, it is conventional to use the large-sample estimate: $u_\infty(c) = U_{95}(c)/2$. This follows from the observation that $c \pm U_{95}(c)$ and a Gaussian distribution with mean c and standard deviation $U_{95}(c)/2$ covers the same 95% interval. The 11th columns (under "Certified values") of Tables 1 and 2 list $u_\infty(c)$ for the [K] and [Chol] CRMs, respectively.

Measurement responses

The measurement system used in a comparison of certified materials must provide a response, r, that is a well-characterized, adequately sensitive, and stable function of measurand quantity: $r = F(x)$. While not strictly necessary, evaluating the data is simpler when $F(x)$ is linear: $r = a + b \cdot x + \text{error}$. It is not, however, necessary that the measurement system be externally calibrated. Since the comparison requires establishing a functional relationship between the observed responses and the measurand quantity values as specified by certified values, external calibration is largely superfluous.

Given that the certified materials being compared (1) nominally deliver (by definition) the same measurand and (2) are evaluated (by design) under strictly controlled conditions, many of the uncertainty components relevant to certification can be expected to be about the same for all of the materials and thus irrelevant to the comparison. The dominant relevant uncertainty component will generally be related to measurement imprecision. However, when results from multiple vials and/or multiple measurement campaigns are combined, it may also be necessary to include a between-vial or between-campaign component. While not difficult, computing these estimates does require use of analysis of variance (ANOVA) methods and attention to detail. The following sections detail the calculations used for each data set. The right-hand columns of Tables 1 and 2 summarize the results for all the [K] and [Chol] studies; electronic supplementary material Tables S1 thru S6 list the data and detail the intermediate calculations.

2003 potassium In addition to evaluating the comparability of the then-available CRMs, this study was designed to evaluate a minimum-measurement model for conducting such comparisons. Singlicate measurements (i.e., one replicate per independent measurement, n_w=1) were made on n_m= 8 materials using a measurement procedure with known within-campaign precision: the relevant precision function for the NIST ID-ICPMS measurement system in 2003 was $s_w(r)$=0.0029·r (i.e., a relative imprecision of 0.29%).

To evaluate the singlicate approach, complete sets of measurements were made in n_b=2 separate campaigns. While none of the between-campaign differences were significant at the 95% level of confidence as evaluated using the usual two-tailed t test, the difference for the material with the highest [K] was larger than expected. The comparison therefore proceeded using the mean, $\bar{r}$, and standard deviation, $s(r)$, of each pair of responses for each material. Since $s_w(r)$ is based on prior experience, the "pure" between-campaign variability can be estimated using standard one-way ANOVA [14]: $s_b(r) = \sqrt{\mathrm{MAX}\left(0, s^2(r) - s_w^2(r)/n_w\right)}$ where MAX is the function "use the maximum of the arguments". The overall variability combines these two sources: $s_t(r) = \sqrt{s_w^2(r) + s_b^2(r)}$. Note that this sequence of estimates establishes a "floor" for the total variability at $s_w(r)$ while keeping $s(r)$ as the "ceiling". The expected imprecision of the mean is estimated from the overall variability and the number of replicate measurements: $s_t(\bar{r}) = s_t(r)/\sqrt{n_b}$. Electronic supplementary material Figure S2 displays $s(r)$, $s_w(r)$, $s_b(r)$, and $s_t(r)$ for the eight materials.

2005 potassium The 2005 NIST ID-ICPMS measurements were made as part of a material certification project rather than a material comparison. While the number and nature of the independent measurements differed among the materials (triplicates on single vials for SRM 956a and singlicates on multiple vials for SRM 956b; Table S2), if it can be assumed that the between-vial variability for the SRM 956b materials is insignificant, then the relative standard deviations, $\%s(r) = 100 \cdot s(r)/r$, for the n_m=6 sets of measurements can be used to estimate a repeatability function for the study: $s_w(r) \approx \sqrt{\sum_{i=1}^{n_m} \left(\frac{s(r_i)}{r_i}\right)^2 / n_m}$. For these data, $s_w(r)$=0.0014·r or 0.14%. Since the %$s(r)$ for the three SRM 956b levels are as small as or smaller than those for the corresponding SRM 956a levels, the assumption of negligible between-vial variability for the new material appears justified.

The uncertainty estimates are calculated as for the *2003 potassium* data. Because all measurements were "doubly used" to estimate both $s(r)$ and $s_w(r)$, all n_w=1 for the purposes of separating the within- and between-set uncertainty components. Electronic supplementary material Figure S2 also displays $s(r)$, $s_w(r)$, $s_b(r)$, and $s_t(r)$ for this study. The improved measurement precision of these 2005 measurements over those of 2003 reflects reduction of background drift.

2003 cholesterol The 2003 cholesterol comparison assayed n_m=12 certified materials with n_w=2 replication in two measurement campaigns using a separate vial of each material in each campaign (n_b=2). This design therefore probed three potential uncertainty components: within campaign, between campaign, and between vial. A mean response, r_{ij}, and standard deviation, $s(r_{ij})$, were estimated for each set of duplicate measurements, where i indexes materials and j indexes campaigns. A within-campaign measurement imprecision for each material was estimated by pooling: $s_w(r_i) = \sqrt{\sum_{j=1}^{n_b} s\left(r_{ij}\right)/n_b}$. On inspection, the majority of the $s_w(r_i)$ are approximately constant regardless of r_i; therefore, the method imprecision function is estimated by pooling: $s_w(r) = \sqrt{\sum_{i=1}^{n_m} s_w^2(r_i)/n_m} = 0.25$ mg/dL.

The grand mean response for each certified material, $\overline{r}$, is estimated as the mean of the r_i. The remaining calculations were performed as above from the standard deviation of the n_b=2 independent mean responses. Electronic supplementary material Figure S3 displays the $s(r)$, $s_w(r)$, $s_b(r)$, and $s_t(r)$ values.

The $s_t(r)$ are also approximately constant and a little different from $s_w(r)$ for 9 of the 12 certified materials, suggesting that the between-campaign uncertainty component is small. However, the $s_t(r)$ are erratically large for the remaining three materials (JCCRM 211-1H, SRM 1952a 1-2, and SRM 1952a 1-3), suggesting that between-vial heterogeneity may be a dominant source of uncertainty. Note that the two largest $s_b(r)$ are for lyophilized materials (SRM 1952a 1-2 and SRM 1952a 1-3).

2004, 2005, and 2008 cholesterol The three sets of cholesterol ID-GC/MS measurements dating after 2003 were made as part of various material certification projects. The experimental designs for each project are similar to that of the 2003 comparison although n_b and n_w differ among both the studies and the materials evaluated in each study. The within-vial standard deviations, $s_w(r_i)$, for the 2004 and 2005 projects also appear to be approximately constant at $s_w(r)$=0.19 and 0.72, respectively (Tables S4 and S5, respectively). The $s_w(r_i)$ in the 2008 project are erratic (Table S6), but $s_w(r)$=0.0040·r appears to better describe the observed behavior than does a constant value of $s_w(r)$= 0.69. Electronic supplementary material Figure S3 also displays $s(r)$, $s_w(r)$, $s_b(r)$, and $s_t(r)$ values for all materials evaluated in these three studies.

Analysis of comparability studies

Assuming that the mean measurement responses, $\overline{r}$, are functionally related to the quantity of the measurand and that the certified values, c, are good estimates of that quantity, then regression methodologies can be used to evaluate the comparability of a given set of n_m certified materials. If all of the $u_\infty(c)$ are negligibly small relative to the $s_t(\overline{r})$ for all materials, then ordinary least squares regression (using unequal weights if the $s_t(\overline{r})$ are unequal) can efficiently estimate $\hat{r} = \mathrm{F}(\overline{r}, c, \ldots)$. Likewise, if all $s_t(\overline{r})$ are negligibly small relative to the $u_\infty(c)$, then $\hat{c} = \mathrm{F}(\overline{r}, c, \ldots)$ can be efficiently estimated. However, should the $u_\infty(c)$ and $s_t(\overline{r})$ be of similar magnitude, then regression techniques that simultaneously estimate $\{\hat{r}, \hat{c}\} = \mathrm{F}(\overline{r}, s_t(\overline{r}), c, u_\infty(c), \ldots)$ may be required for the truest assessment. The magnitudes of the certified uncertainty and measurement imprecision are indeed about equal in our [K] and [Chol] studies. Since few measurement laboratories have the analytical resources to make more replicate measurements than are needed, we believe that this rough equivalency is likely to be the case for most comparisons among natural-matrix CRMs.

There are numerous approaches to this simultaneous estimation problem, often termed "errors-in-variables" or "total least squares" regression [15, 16]. Deming regression provides a solution in the special case where the ratio $s_t(\overline{r})/u_\infty(c)$ is constant for all materials [17]. When there are no constraints on these uncertainties, more general approaches are needed. Recently, two spreadsheet implementations of such an approach have been made freely available: linear functional relationship estimation by maximum likelihood (FREML) [18] and generalized least-square regression (GLS) [19]. FREML addresses only the linear model, while GLS also supports polynomial models. When used with the linear model, both systems minimize $\sum_i^{n_m}\left(\frac{\hat{r}_i-\overline{r}_i}{s_t(\overline{r}_i)}\right)^2 + \sum_i^{n_m}\left(\frac{\hat{c}_i-c_i}{u_\infty(c_i)}\right)^2$ by iterative estimation of the intercept and slope, α and β, where $\hat{r} = \alpha + \beta \cdot \hat{c}$ or, equivalently, $\hat{c} = (\hat{r} - \alpha)/\beta$.

An alternate, more flexible (if less computationally efficient) exploratory analysis tool that evaluates this model has been developed at NIST. This tool, termed "RegViz", uses the non-linear optimization engine native to the Excel spreadsheet environment. For the above linear model and minimization function, RegViz produces the same parameter estimates (to at least five digits) as do FREML and GLS.

If all of the materials in a comparability study truly deliver the same measurand and have been accurately certified and a fit-for-purpose measurement system has been competently used following a fit-for-purpose experimental design, then the uncertainty-scaled Euclidean distance between the observed and estimated values, $d = \mathrm{SIGN}(\hat{r} - \overline{r}) \cdot \sqrt{\left(\frac{\hat{r}-\overline{r}}{s_t(\overline{r})}\right)^2 + \left(\frac{\hat{c}-c}{u_\infty(c)}\right)^2}$, where "SIGN" is the function "take the sign of the value", which can be interpreted directly as a standardized normal distribution having zero mean and unit standard deviation. The probability of observing a deviation as large or larger than $|d|$ for such a "*z*-score" is equal to 2(1−NORMSDIST($|d|$)), where "NORMSDIST" is the standard normal cumulative distribution function. Approximately 95% of the materials with $\{c, \overline{r}\}$ pairs that satisfy the above assumptions will have $|d|$ no larger than 1.96; likewise, approximately 99% will have $|d|$ no larger than 2.58. Values of $|d|$ much larger than 3 are unlikely to arise by chance and therefore suggest that at least one of the prerequisite assumptions is invalid.

2003 potassium Figure 1 (and electronic supplementary material Figure S4) is RegViz graphical output for the 2003 [K] comparison, indicating a very satisfactory [K] comparability for all of the studied materials. The large scatterplot gives an overview of the data for all of the materials and the (quite linear) functional relationship between the certified and response values, but the resolution is insufficient for

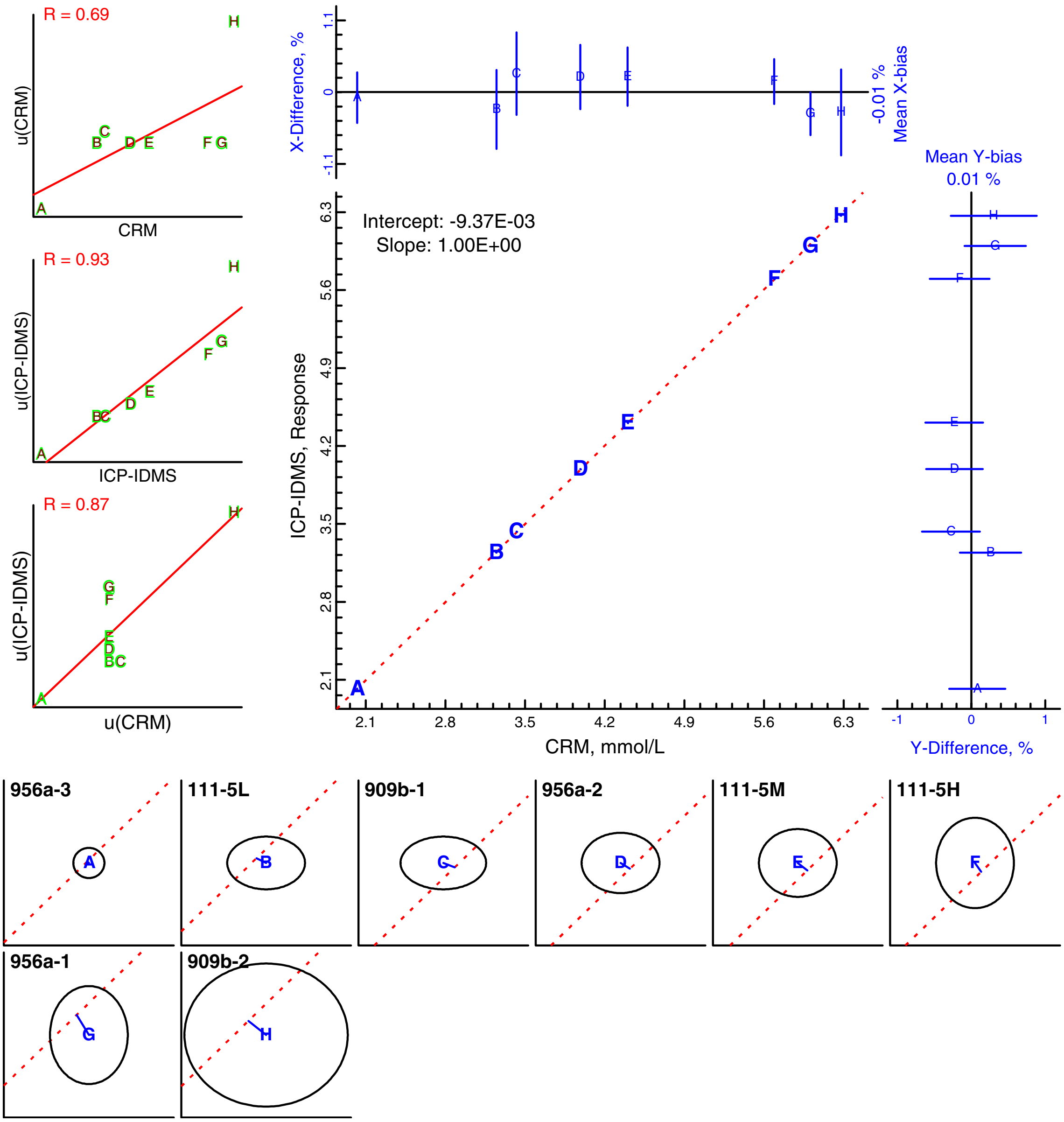


Fig. 1 RegViz graphical results for the 2003 [K] comparison. The large certified value vs. measurement response, $\{c, \bar{r}\}$, scatterplot provides an overview of the paired values, denoted with one-character codes, and the $\hat{r}_i = \alpha + \beta \cdot \hat{c}_i$ regression function, represented as a *dashed line*. Each of the small scatterplots below the overview provide a high-resolution display for the sample named to the *upper left*; the *corresponding code at the center* marks the observed data, $\{c, r\}$. Each *ellipse* bounds an approximate 95% confidence region specified by the values and their uncertainties, $\{c, u_\infty(c); \bar{r}, s_t(\bar{r})\}$. All of the small scatterplots share the same axis scaling, constructed to just contain the largest of the ellipses. The marginal plot above the large scatterplot displays the percent relative bias $100(c \pm u_\infty(c) - \hat{c})/\hat{c}$; the marginal plot to the right displays $100(\bar{r} \pm s_t(\bar{r}) - \hat{r})/\hat{r}$. The three scatterplots to the left display $\{c, u_\infty(c)\}$, $\{\bar{r}, s_t(\bar{r})\}$, and $\{u_\infty(c), s_t(\bar{r})\}$; the *solid line within each of these scatterplots* represents a simple linear fit between the respective values

meaningful evaluation of differences on the scale of $u_\infty(c)$ or $s_t(\bar{r})$ given the wide range spanned by the c and r axes. The small scatterplots, one for each material, along the bottom provide the required details. For all of the materials studied, the d are well within the 95% comparability acceptance bounds. The relative bias plot for the certified values, $\%b = 100(c - \hat{c})/\hat{c}$, just above the large scatterplot reveals that the observed certified values, c, are consistent within about ±0.2%. The three scatterplots to the left of the large scatterplot are intended to help guide selection of an appropriate optimization model and are here of little utility. Figure 2 details the structure of and information embodied in the small scatterplots.

Figure 3 consolidates the critical comparability information contained in the small scatterplots and upper bias plot, displaying the uncertainty-scaled residual (d) as functions of certified value (c) and percent bias ($\%b$). This display is

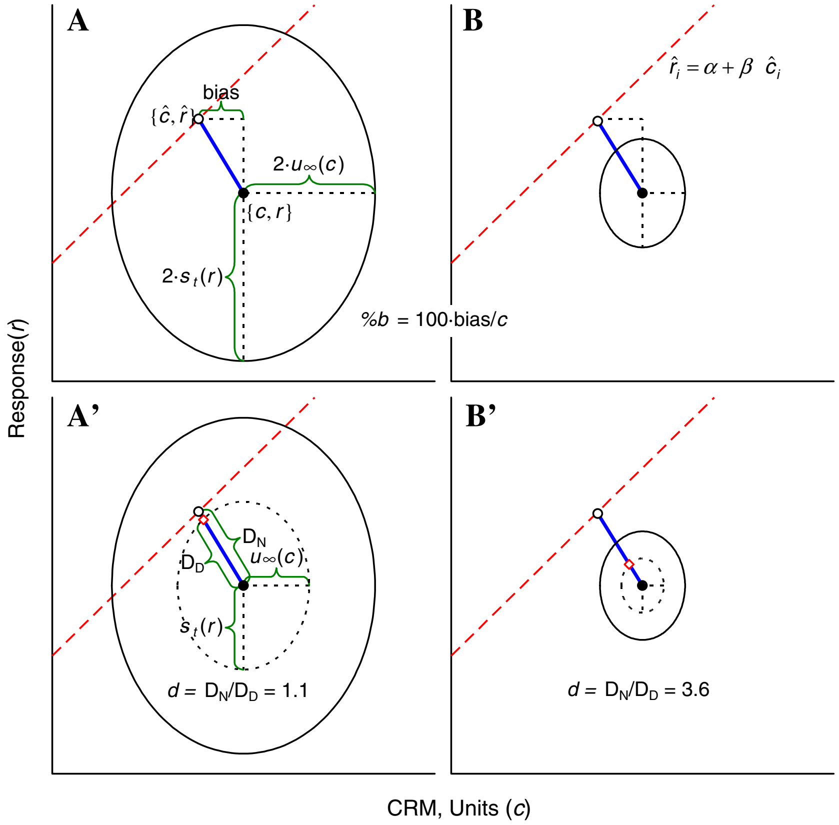


Fig. 2 Properties embodied in the small scattergrams. Each of the small certified value vs. measurement response scattergrams in a RegViz graphic describes the data and comparison results for one material. **A** and **A′** emphasize different aspects of results for an exemplar material with good comparability; **b** and **b′** display aspects for an exemplar of poor comparability. The only difference between the two exemplars (**A** and **B**) is the magnitude of the uncertainties. In all panels, the *solid circles* represent the reported certified value and measured response, $\{c, \bar{r}\}$; the *open circles on the dashed line* represent the predicted values, $\{\hat{c}, \hat{r}\}$; the *dashed line* represents the $\hat{r} = \alpha + \beta \cdot \hat{c}$ regression function; the *thick line* connecting $\{c, \bar{r}\}$ to $\{\hat{c}, \hat{r}\}$ is the Euclidean distance between the observed and predicted values, $\sqrt{(\hat{c} - c)^2 + (\hat{r} - \bar{r})^2}$. In **A** and **B**, the *horizontal dotted line* to the right of the predicted values (*open circles*) represents the bias between the predicted and reported certified values; this bias along the horizontal CRM axis is used to calculate the relative bias, $\%b = 100(\hat{c} - c)/c$. The *solid-line ellipses* bound approximate 95% confidence regions specified by the values and their uncertainties, $\{c, u_\infty(c); \bar{r}, s_t(\bar{r})\}$, where the expansion factor to achieve 95% coverage is asserted to be 2. The *dotted lines* of the horizontal and vertical radii of the *solid-line ellipse* are likewise equal to $2 \cdot u_\infty(c)$ and $2 \cdot s_t(\bar{r})$, respectively. In **A′** and **B′**, an additional *dashed-line ellipse* representing an approximate 68% (1σ) confidence region is displayed; the intersection of this *ellipse* with Euclidean distance line is marked by an *open diamond*. The uncertainty-weighted Euclidean distance, d, is equal to the distance (D_N) from centerpoint (*solid circles*) to the predicted values (*open circles*) divided by the distance (D_D) from centerpoint (*solid circles*) to the ellipse boundary (*open diamond*) along the radius

intended to provide a simpler if less nuanced summary of results: For these materials, the d have very similar magnitude and are all well within the 95% acceptance region; the $\%b$ are also all of similar magnitude and indicate that the materials deliver the same measurand to within about ±0.2%.

2003 cholesterol Electronic supplementary material Figure S5 is the full RegViz output for the 2003 [Chol] comparison, with Fig. 4a summarizing the critical results. While the c vs. $\bar{r}$ functional relationship is again very strong and quite linear, the d for a number of materials are outside the 95% acceptance limits. However, it is necessary to consider all of the available evidence before concluding that either some to all of the certified values and/or their uncertainties are incorrect or the comparison measurements are flawed, or that the materials have degraded over time. Examination of the size and location of the 95% ellipses relative to the “best fit” line reveals several anomalies: (1) The ellipses for SRM 1589a (labeled D), 1951a-1 (F), and 1951a-2 (K) are atypically narrow along the c axis; (2) the ellipses for SRM 1952a-2 (H) and 1952a-3 (L) are atypically long along the r axis; and (3) the ellipse for SRM 1951a-2 (K) is somewhat above the fitted line, while the ellipses for the other high [Chol] materials (G, H, I, J, and L) are trending toward below the line.

The largest d is for SRM 1589a; RegViz results for the analysis with this material excluded from the fit are

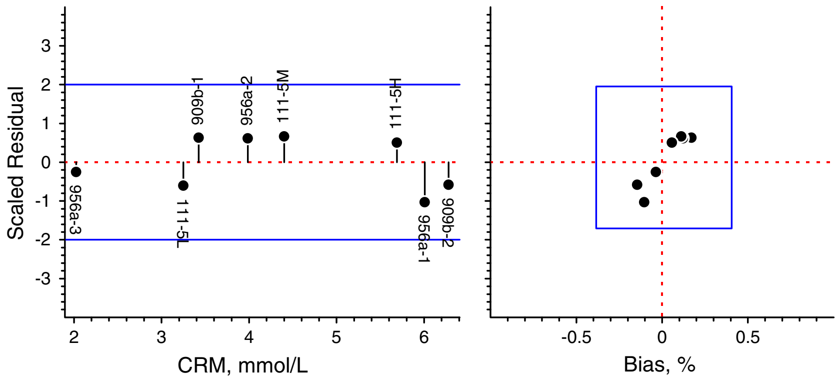


Fig. 3 Comparability summary for the 2003 [K] comparison. The scatterplot to the left displays the uncertainty-scaled Euclidean distance, d, (scaled residual) as a function of certified value, c, (*CRM*). The $\{c, d\}$ pairs are represented as *closed circles*. The *horizontal dashed line* represents d=0; the *solid vertical lines* connecting the zero line to each of the $\{c, d\}$ pairs provides graphical emphasis of the d. The *solid horizontal lines* bound the $-2 \le d \le 2$ comparability acceptance region. The scatterplot to the right displays the d as a function of relative bias, %b. The box bounds the "well behaved" majority of the data, estimated using robust estimates of location and dispersion. The *vertical* and *horizontal dashed lines* represent "zero difference" on the d and %b axis, respectively

displayed in electronic supplementary material Figure S6 and summarized in Fig. 4b. Exclusion of SRM 1589a (D) improves comparability for the other low [Chol] materials but does not much affect the d or %b of the high [Chol] materials. Given the discordance between the SRM 1951a-2 (K) material and the other high [Chol] materials in combination with its atypically small—and therefore relatively influential—$u_\infty(c)$, it is plausible that the SRM 1951a-2 (K) is a major source of discordance. With the additional exclusion of SRM 1951a-2 (K) from the fit (Figure S7 and Fig. 4c), the d for all of the remaining materials are within the acceptance region, although the magnitude of %b for the SRM 968c-1 (A) material becomes atypically large. Excluding SRM 968c-1 (A) from the fit (Figure S8 and Fig. 4d) only marginally reduces d for the remaining materials and appears to increase their %b. Exclusion of the SRM 1952a-2 (H) and SRM 1952a-3 (L) materials (not shown) does not appreciably change any result.

The inadequate [Chol] comparability of SRM 1589a was recognized in the initial analysis of these data [6] and is attributed to a modest degree of cholesterol oxidation. SRM 1589a was immediately decertified for [Chol]; however, there was no evidence for degradation of the pesticide measurands of primary interest in this CRM nor for the "Total cholesterol" measurand (cholesterol plus its simple oxidation products as well as other closely related entities) as determined by spectroscopic assay. SRM 1589a is now sold out and is being replaced with SRM 1957 Organic Contaminants in Human Serum (non-fortified) and SRM 1958 Organic Contaminants in Human Serum (fortified), neither of which is certified for cholesterol.

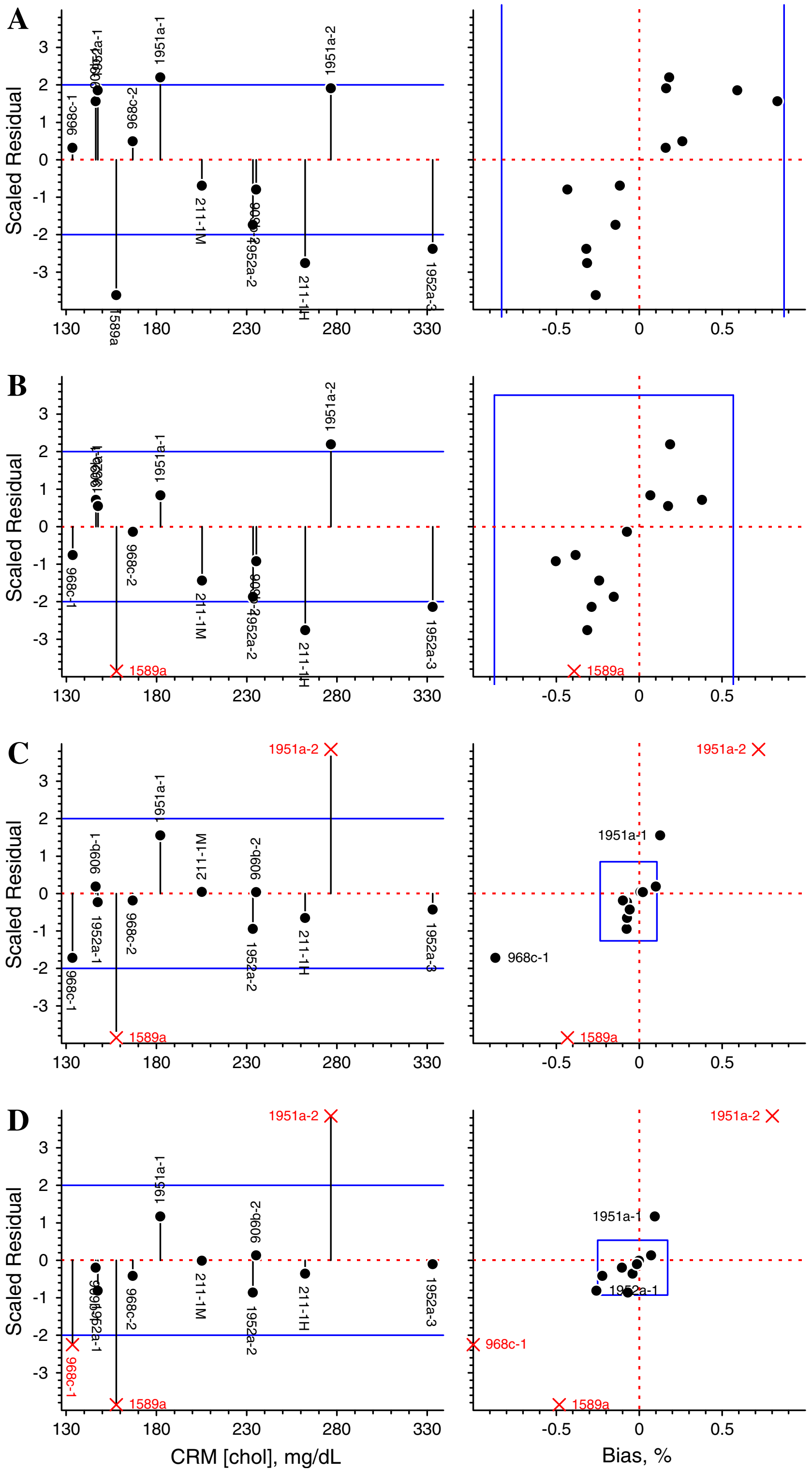
The atypically large—and therefore relatively non-influential—$s_t(\overline{r})$ of the SRM 1952a-2 and SRM 1952a-3 materials are not associated with atypical %b and are therefore unlikely to arise from cholesterol oxidation. While it is likely that the certified uncertainties for the SRM 1951a-1 materials are unrealistically small, the %b are quite in line with those of the other materials. The supply of SRM 1951a was exhausted shortly after completion of the comparability study.

Augmenting comparability demonstrations

Simultaneous analysis under repeatability conditions provides the most convincing evidence for the comparability of certified materials, but it is resource intensive and can only evaluate materials that are available at some given point in time. Mechanisms for augmenting a primary study with new materials as they become available could yield considerable benefit. Fortunately, many if not all CRM producers routinely evaluate old and new materials together when developing new materials. Assuming the functional relationships relating response to measurand quantity for the original and new measurement systems are qualitatively similar, then the responses should be approximately linearly related. The responses obtained with the measurement system used with the new materials (call them r_{new}) can then be transformed to the existing scale: $\hat{r} = \alpha + \beta \cdot \overline{r}_{\text{new}}$ and $s_t(\hat{r}) = \beta \cdot s_t(\overline{r}_{\text{new}})$. The question becomes how best to estimate the linear transformation parameters from the available data.

2005 potassium The SRM 956a materials (956a-1, -2, and -3) included in the 2003 [K] comparability study were also evaluated during the 2005 certification measurements for the replacement CRM, SRM 956b (956b-1, -2, and -3). While the results for SRM 956a-2 were at that time noted as anomalous, no technical cause could be identified, and no additional units of the SRM were available for further investigation. It is therefore necessary to evaluate trans-

Fig. 4 Comparability summary for the 2003 [Chol] comparison. **A** Results when data for all materials are used in the regression, **B** when the SRM 1951a-2 material is excluded from the regression, **C** when both SRM 1951a-2 and 1589a are excluded, and **D** when SRM 1951a-2, 1589a, and 968c-1 are excluded. The format is as described in Fig. 3, with the materials excluded from the regression denoted as *crosses*



formations using (a) all three of the SRM 956a materials and (b) only 956a-1 and -3 materials.

Given three or more pairs of $\{\bar{r}_{\mathrm{new}}, s_{\mathrm{t}}(\bar{r}_{\mathrm{new}}); \hat{r}, s_{\mathrm{t}}(\hat{r})\}$ measurements, the errors-in-variables model implemented in the FREML, GLS, or RegViz systems can be used to estimate best-fit transformation parameters—and the uncertainties for those parameters. While estimation uncertainty is of little concern in comparability assessment where the focus is the relationship between the functional relationship and the $\{c, u_{\infty}(c); \bar{r}, s_{\mathrm{t}}(\bar{r})\}$ intervals, it is critical to establishing appropriate $s_{\mathrm{t}}(\hat{r})$ imprecision estimates for new materials.

The RegViz system uses a Monte Carlo (MC) resampling technique to evaluate parameter uncertainties [20]; with sufficient resamplings, these estimates are congruent with those provided by the FREML and GLS systems. The GLS and RegViz systems both support direct prediction of the response a material "would have had" had it been included in the original study; the RegViz estimates are again congruent with those from GLS. Electronic supplementary material Table S7 provides the numerical results from the RegViz estimation of the transformation parameters; Figure S9 presents the results in graphical form. While the $s_{\mathrm{t}}(\overline{r}_{new})$ for the replacement SRM 956b materials are small due to the large number of measurements made, the transformed $s_{\mathrm{t}}(\hat{r})$ estimates are quite similar to (and a bit larger than) the $s_{\mathrm{t}}(\overline{r})$ of the corresponding 956a materials in the 2003 comparison.

Electronic supplementary material Figure S10 displays the analysis of the SRM 956b-augmented comparability data; panel A of Figure S11 summarizes these results. While all the materials remain acceptably comparable in the augmented set, the SRM 956b-2 (E) and SRM 956b-3 (A) materials are somewhat less comparable and more biased than expected. As can be seen in the small scatterplots of Figure S9, the three-material transformation line "splits the difference" between the 956a-2 (D) and 956a-1 (F) materials—thus partially propagating the potentially biased 956a-2 measurement into the transformation parameters.

Given just two data pairs, the intercept and slope of the connecting line are exactly determined. However, the MC resampling procedure can still propagate the measurement uncertainties through the transformation. Electronic supplementary material Table S8 and Figure S12 present the results of parameter estimation with the SRM 956a-2 material excluded, Figure S13 presents the analysis of the augmented data, and panel B of Figure S10 summarizes the critical results of the analysis. While here the $s_t(\hat{r}_i)$ for the SRM 956b materials are larger than when transformed using all three of the 956a materials, both the comparability and bias for the augmented suite of [K] materials are now changed very little from those of the original set.

2004 cholesterol The SRM 1951a materials (1951a-1 and -2) included in the 2003 [Chol] comparability study were also evaluated during the 2004 certification measurements for the replacement materials, 1951b-1 and -2. With again but two data pairs, the straight line connecting the pairs provides the transformation function. Electronic supplementary material Table S9 and Figure S14 present the results for the two-material transformation, Figure S15 displays the analysis of the augmented data, and panel A of Figure S16 summarizes the results. While the %*b* for both 1951b-1 and -2 are quite acceptably small, the *d* for 1951b-2 approaches the 95% comparability acceptance limit—again perhaps suggesting that the certified uncertainty for this material may be somewhat underestimated.

2005 cholesterol The JCCRM 211-1 materials (211-1H and -1M) included in the 2003 [Chol] comparability study and the SRM 1951b materials (1951b-1 and -2) certified in 2004 were evaluated during the 2005 certification measurements for the replacement materials, JCCRM 211-2H and 211-2M. Electronic supplementary material Table S10 and Figure S17 present the results for the four-material best-fit transformation, Figure S18 displays the analysis of the augmented data, and panel B of Figure S16 summarizes the results. Although the transformation analysis used "second-generation" response estimates (for the SRM 1951b materials), both of the JCCRM 211-2 materials have very small %*b* and *d* and thus appear to be nicely comparable to the older materials.

2008 cholesterol The SRM 1951b materials (1951b-1 and -2) certified in 2004 were also evaluated during the 2008 certification measurements for SRM 968d. At that time, it was noted that the measured 1951b-1 [Chol] was significantly lower than expected, while those for 1951b-2 and a gravimetrically prepared primary reference agreed well with expectations. This suggests that the measurement system was in adequate control but that the 1951b-1 material may have become somewhat oxidized. Electronic supplementary material Table S11 and Figures S19 and Figure S20 present transformation and augmentation results for the biased two-material transformation. Since the SRM 968d material has a considerably lower [Chol] than the 1951b materials, the impact of the 1951b-1 bias on 968d is amplified, as indicated by large increase in $\hat{r}$ from 133.46 to 138.03.

Transformation with only the one valid (1951b-2) $\{\overline{r}_{\mathrm{new}}, s_{\mathrm{t}}(\overline{r}_{\mathrm{new}}); \hat{r}, s_{\mathrm{t}}(\hat{r})\}$ pair is possible only if either the slope or intercept of the line can be defensibly established from other information. Since the ID-GC/MS measurement system used for the comparison was fully calibrated, the intercept and slope ideally "should be" 0 and 1, respectively. Electronic supplementary material Table S12 and Figures S21 and S22 present the transformation and augmentation results for a straight line of unit slope that goes through the 1951b-2 values. Table S13 and Figures S23 and S24 present the transformation and augmentation results for a straight line of zero intercept that goes through 1951b-2. Choosing the unit slope transformation since it results in the largest $s(\hat{r}_i)$ and thus minimizes the influence of this material in any future augmentations, the "best guess" critical comparability values are summarized in panel C of Figure S16. Both the *d* and %*b* for SRM 968d are very close to zero, and thus, the material appears to be nicely comparable to the older materials.

Alternate cholesterol comparability analysis

While the SRM 1589a material is unambiguously insufficiently comparable to the rest of the cholesterol CRMs, the requisite exclusion of 1951a-2 is less clear-cut. Electronic supplementary material Figures S25 to S27 display the analyses for the 2004, 2005, and 2008 augmentations with only SRM 1589a excluded; panels A to C of Figure S28 summarize the critical values. Excluding both SRM 1589a and 1951a-2, the 95% interval for the %*b* of all other materials is approximately ±0.4%, and all *d* are within the 95% acceptance interval (see panel C of Figure S16). Excluding only SRM 1589a, the %*b* interval is approximately ±0.5%, and the *d* for an additional three materials (including 1951a-2) are outside the acceptance interval (see panel C of Figure S28). While not dramatically impacting the interpretation, the weight of the evidence is that the SRM 1951a-2 material should be excluded.

Summary and recommendations

When two or more CRMs nominally deliver the same measurand, demonstrating the comparability among the certified materials can inform and reassure CRM users of the materials' fitness for purpose. If done as a routine part of the certification process for new CRMs, comparability comparisons can also help CRM producers assure the quality of their products.

When it is anticipated that multiple CRMs for the same measurand will eventually be produced, it is important that CRM producers properly preserve sufficient supplies of the materials to enable future comparability studies. These reserves are in addition to those intended for use as control materials or in the evaluation of long-term stability [1]. Initial comparability studies can be conducted with as few as three analogous CRMs materials but should include as many materials as possible. We strongly support the JCTLM protocol for administering comparability studies involving materials from multiple producers and encourage all CRM producers to cooperate fully in such studies [6].

Once established, comparability demonstrations are most efficiently augmented by the producer of new CRMs using three or more previously studied materials. Since measurement imprecision for all materials analyzed in an augmentation study, old as well as new, is propagated into the comparability estimates of the new materials, it is critical that sufficient independent measurements are made on all of the materials. To allow for exclusion of data from technically flawed measurements, a minimum design for all "response" measurements is the duplicates of three independently prepared samples of each material.

High-level summaries such as the uncertainty-scaled residual (*d*) and percent bias (%bias) plots used here should be made readily available to CRM customers, perhaps through web portals maintained by interested communities such as the JCTLM [8] as well as by the individual CRM producers. However, particularly when only the producers of the CRMs make measurements and do the data analyses, it is critical that all relevant data are made available to the user communities for independent review and assessment. We have herein attempted to model what we believe this data package should contain, as well as possible ways the measurement data can be analyzed and the results be presented.

Acknowledgments We thank William F. Guthrie, Hung-Kung Liu, and Nien-Fan Zhang of the NIST Statistical Engineering Division and Steven L.R. Ellison of LGC Limited for their advice and insights. We also thank Steve and his colleagues of the Analytical Methods Committee of the Royal Society of Chemistry for providing the world analytical community with the AMC Statistics data evaluation systems.

Disclaimer Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology or the Reference Material Institute for Clinical Chemistry Standards, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

References

1. ISO Guide 35 (2006) Certification of reference materials: general and statistical principles. ISO, Geneva
2. BIPM (2006) The international system of units (SI), 8th edition. Sèvres, France. http://www1.bipm.org/en/si/si_brochure/
3. JCGM 200 (2008) International vocabulary of metrology—basic and general concepts and associated terms (VIM). Joint Committee for Guides in Metrology. Sèvres, France. http://www.bipm.org/en/publications/guides/vim.html
4. Clinical and Laboratory Standards Institute (2002) EP09-A2 Method comparison and bias estimation using patient samples; approved guideline—second edition. CLSI, Wayne
5. Wielgosz RI, Esler M, Viallon J, MoussayP, Oh SH, Kim BM, Botha A, Tshilongo J, Mokgoro IS, Maruyama M, Mace T, Sutour C, Stovcík V, Valková M, S, Castorena AP, Caballero VS, Murillo FR, Konopelko LA, Kustikov YA, Pankratov VV, Gromova EV, Thorn WJ, Guenther FR, Smeulders D, Baptista G, Dias F, Wessel RM, Nieuwenkamp G, van der Veen AMH (2008) Metrologia 45 (Technical Supplement) 08002. http://www.bipm.org/utils/common/pdf/final_reports/QM/P73/CCQM-P73.pdf
6. JCTLM (2006) Joint Committee for Traceability in Laboratory Medicine Quality System Procedure JCTLM WG1-P-04A Process for Comparing Certified Values of the Same Measurand in Multiple References Materials (CRMs). http://www.bipm.org/utils/en/pdf/WG1-P-04A.pdf
7. Armbruster D, Miller RR (2007) Clin Biochem Rev 28(3):105–114

8. JCTLM database: Laboratory medicine and in vitro diagnostics. www.bipm.org/jctlm/
9. Gramlich JW, Machlan LA, Brletic KA, Kelly WR (1982) Clin Chem 28(6):1309–1313
10. Murphy KE, Long SE, Rearick MS, Ertas OS (2002) J Anal At Spectrom 17(5):469–477
11. Ellerbe P, Meiselman S, Sniegoski LT, Welch MJ, White E (1989) Anal Chem 61(15):1718–1723
12. JCGM 100 (2008) Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM). Joint Committee for Guides in Metrology. Sèvres, France. http://www.bipm.org/en/publications/guides/gum.html
13. Natrella MG (2005) Experimental statistics. Dover, Mineola
14. ISO (1994) ISO 5725-2 Accuracy (trueness and precision) of measurement methods and results. Part 2: basic method for the determination of repeatability and reproducibility of a standard measurement method. Geneva, Switzerland
15. Cornbleet PJ, Gochman N (1979) Clin Chem 25(3):432–438
16. ISO (2001) ISO 6143 Gas analysis—comparison methods for determining and checking the composition of calibration gas mixtures. Switzerland, Geneva
17. Deming WE (1943) Statistical adjustment of data. Wiley, New York
18. Ripley BD, Thompson M (1987) Analyst 112(4):337-383. Linear Functional Relationship Estimation by Maximum Likelihood (FREML) freeware available from http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/FREML.asp
19. Milton MJT, Harris PM, Smith IM, Brown AS, Goody BA (2006) Metrologia 43(4):S291–S298. XLGENLINE freeware available through http://www.eurometros.org/component_search.php?component_type=distributions
20. Duewer DL, Kowalski BR, Fasching JL (1976) Anal Chem 48 (13):2002–2010