Clinical Chemistry 62:9
1255–1263 (2016)

Analytical Bias Exceeding Desirable Quality Goal in 4 out of 5 Common Immunoassays: Results of a Native Single Serum Sample External Quality Assessment Program for Cobalamin, Folate, Ferritin, Thyroid-Stimulating Hormone, and Free T_4 Analyses

Gunn B.B. Kristensen,[1*] Pål Rustad,[1] Jens P. Berg,[2] and Kristin M. Aakre[3]

BACKGROUND: We undertook this study to evaluate method differences for 5 components analyzed by immunoassays, to explore whether the use of method-dependent reference intervals may compensate for method differences, and to investigate commutability of external quality assessment (EQA) materials.

METHODS: Twenty fresh native single serum samples, a fresh native serum pool, Nordic Federation of Clinical Chemistry Reference Serum X (serum X) (serum pool), and 2 EQA materials were sent to 38 laboratories for measurement of cobalamin, folate, ferritin, free T_4, and thyroid-stimulating hormone (TSH) by 5 different measurement procedures [Roche Cobas (n = 15), Roche Modular (n = 4), Abbott Architect (n = 8), Beckman Coulter Unicel (n = 2), and Siemens ADVIA Centaur (n = 9)]. The target value for each component was calculated based on the mean of method means or measured by a reference measurement procedure (free T_4). Quality specifications were based on biological variation. Local reference intervals were reported from all laboratories.

RESULTS: Method differences that exceeded acceptable bias were found for all components except folate. Free T_4 differences from the uncommonly used reference measurement procedure were large. Reference intervals differed between measurement procedures but also within 1 measurement procedure. The serum X material was commutable for all components and measurement procedures, whereas the EQA materials were noncommutable in 13 of 50 occasions (5 components, 5 methods, 2 EQA materials).

CONCLUSIONS: The bias between the measurement procedures was unacceptably large in 4/5 tested components. Traceability to reference materials as claimed by the manufacturers did not lead to acceptable harmonization. Adjustment of reference intervals in accordance with method differences and use of commutable EQA samples are not implemented commonly.

Despite the effort that has been put into standardization and harmonization of laboratory test results, large method-dependent differences persist between measurement procedures *(1)*. Method-dependent differences may have clinical consequences, leading to considerable practical problems when patients move between healthcare institutions, thus challenging the use and interpretation of research data. The scope of external quality assessment (EQA)[4] has evolved considerably during recent years *(2)*. With the increasing worldwide concern about the use of common reference intervals and medical decision limits *(1, 3–5)*, modern EQA schemes should aim at assessing the standardization and harmonization status of commercial in vitro diagnostic tests. A prerequisite is the use of commutable EQA samples, i.e., samples that demonstrate the same numeric relationship between measurement procedures as is observed in real patient samples. Results from commutable EQA samples can be used to evaluate trueness and the calibration standardization to a reference measurement pro-

[1] The Norwegian Clinical Chemistry EQA program, Bergen, Norway; [2] Department of Medical Biochemistry, Oslo University Hospital and Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [3] Laboratory of Clinical Biochemistry, Haukeland University Hospital, Bergen, Norway.
* Address correspondence to this author at: NKK, Box 6165, 5892 Bergen, Norway. Fax +47–55979510; e-mail gunn.kristensen@noklus.no.
Received April 11, 2016; accepted May 26, 2016.
Previously published online at DOI: 10.1373/clinchem.2016.258962

[4] Nonstandard abbreviations: EQA, external quality assessment; NKK, Norwegian Clinical Chemistry EQA Program; NFKK, Nordic Federation of Clinical Chemistry; serum X, NFKK Reference Serum X; TSH, thyroid-stimulating hormone; USP, US Pharmacopeial Convention; DEKS, Danish Institute for External Quality Assurance for Laboratories in Health Care; Mt, total mean; CV_T, total biological variation; CV_I, within-participant biological variation; CV_G, between-participant biological variation; CV_A, analytical coefficient of variation; PI, prediction interval.

cedure or calibration harmonization when no reference method is available *(2)*.

We present the results of an EQA survey performed by the Norwegian Clinical Chemistry EQA Program (NKK) using commutable native single serum samples. The main aim of this survey was to investigate method-dependent differences for a few components analyzed with commercially available immunoassays. Both components that were traceable to a reference material and components being traceable to industry standards were included. Further, we evaluated if the use of local reference intervals may compensate for method differences so that the clinical interpretation of the test results will be unaffected. The last aim was to test the commutability of 2 commercial EQA sample materials and Nordic Federation of Clinical Chemistry (NFKK) Reference Serum X (serum X) that are used in ongoing EQA programs.

Methods

SURVEY OUTLINE

The survey was offered to clinical chemistry laboratories in Norway as part of the EQA program from NKK. Five components [cobalamin, folate, ferritin, thyroid-stimulating hormone (TSH), and free T_4] were measured using 5 different measurement procedures [Roche Cobas (n = 15) (Roche Diagnostics), Roche Modular (n = 4) (Roche Diagnostics), Abbott Architect (n = 8) (Abbott Diagnostics), Beckman Coulter Unicel (n = 2) (Beckman Coulter), and Siemens ADVIA Centaur (n = 9) (Siemens Healthcare Diagnostics)] at 38 different clinical-chemistry laboratories. Supplemental Table S1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol62/issue9 shows the traceability for the different immunoassays as specified by the manufacturers. The first or third WHO reference materials were used for ferritin, whereas the second or third WHO reference materials were used for TSH. The folate assay from Abbott Diagnostics reported traceability to a WHO reference material. The other assays claimed traceability to different US Pharmacopeial Convention (USP) or in-house industry standards. No assay reported traceability to the published reference method for free T_4 *(6)*.

The following EQA materials were included in the survey: 20 fresh native single serum samples; 1 native serum pool consisting of fresh sera from 18 donors; 1 fresh-frozen vial of serum X, which is a native 120-donor serum pool that has been stored at −80 °C for 12 years *(7)* [M. Pedersen, personal communication, February, 2016, Danish Institute for External Quality Assurance for Laboratories in Health Care (DEKS)]; and, finally, 2 commercial EQA materials from Labquality's EQA scheme Hormones A (2300) survey 6–2013 (EQA1, a liquid pooled serum; EQA2, a human-based lyophilized serum).

EQA1 was an unmodified fresh-frozen human serum pool from 8 men having hemochromatosis. The serum was collected "on-the clot" *(7)*, stored at −80 °C and filtered using filters with pore sizes of 80 and 0.45 μm. The filling procedure used was aseptic (M. Pedersen, personal communication, May, 2016, DEKS).

The fresh native single serum samples and the fresh serum pool were produced by the Norwegian Quality Improvement of Primary Care Laboratories (Noklus). Samples were collected from 20 healthy volunteers after obtaining written informed consent. The venipunctures were performed using a BD Vacutainer® Safety-Lok™ butterfly device [Becton, Dickinson and Company (BD), ref. no. 367282]. Blood was drawn into 25 gel tubes of 8.5-mL size [BD Vacutainer® SST™ II Plus Advance (BD, ref. no. 367953)]. The gel tubes were allowed to clot for 30 min at room temperature before they were centrifuged twice at 2000*g* for 10 min. Serum was separated after each centrifugation, and 1 mL serum was aliquoted in 2-mL tubes and stored at 4 °C until shipping (within 1 day after production). Significant hemolysis, lipemia, icterus, or bacterial growth was not present in any of the samples, and stability testing was done for 5 days showing acceptable data for all components (data not shown). Temperature control was included in 4 deliveries. The minimum temperature registered was −1.0 °C, and the maximum was +8.5 °C, which was regarded as acceptable for stability. Detailed instructions for analyzing and registering the results accompanied the samples.

All laboratories except 2 received and analyzed the sample material within 4 days and all within 8 days. The laboratories reported the locally used reference intervals for a woman of 40 years of age for all the included components.

DETERMINATION OF THE TARGET VALUE

A reference method was available only for the free T_4 assay *(6)*. Reference-method values were established for 5 selected native serum samples and for the serum X material at the Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Gent University (Belgium), using this method *(6)*. For the other components, the target value was defined as the total mean (Mt) value determined from 3 single mean values for the method groups Roche, Abbott, and Siemens. We decided to include all Roche methods (Cobas and Modular, n = 19) in a single Roche group for the determination of Mt because the difference between these 2 groups was small, and the Roche methods would have become very dominating for determining Mt if they were regarded as 2 separate methods. However, for the method comparisons, Roche Cobas and Modular were treated separately because between-instrument variation may occasionally be seen. The mean value from the Beckman Coulter Uni-

cel group was not included owing to the small number of responders (n = 2). For each method, the mean deviations of the results are presented as difference plots between the method mean and Mt.

Grubbs test *(8)* was used for outlier testing, and fewer than 1% of the single test results were removed as outliers.

Quality specifications for evaluating method differences were based on biological variation data and were calculated according to the formula *(9–11)*:

$$\text{Bias} < 0.25 \times CV_T, \quad (1)$$

where CV_T is defined as: $CV_T = \sqrt{CV_I^2 + CV_G^2}$

CV_T is total biological variation, CV_I is the within-subject biological variation and CV_G is the between-subject biological variation. The following quality goals for desirable bias were defined: cobalamin 17.7%, folate 19.2%, ferritin 5.2%, free T_4 3.3%, and TSH 7.8% *(11)*.

Method comparison data between the different assays were analyzed using weighted Deming regression.

CALCULATION OF METHOD REPRODUCIBILITY CV

To reveal typical method repeatability and robustness of calibration/measurement procedures, we calculated the pooled within-method analytical coefficient of variation (CV_A) for the different components as:

$$CV_A = \sqrt{(\Sigma CV_{a^2}/N)}, \quad (2)$$

where CV_a is the analytical CV for each serum sample analyzed in different laboratories that use the same measurement procedure, and N is the number of sera (N = 20). Quality specifications for CV_A were defined according to biological variation *(9–11)*:

$$CV_A < 0.5 \times CV_I. \quad (3)$$

The following desirable CV_A values were defined: cobalamin 7.5%, folate 12.0%, ferritin 7.1%, free T_4 3.6%, and TSH 14.7% *(11)*.

COMPARISON OF REFERENCE INTERVALS

To compare the different reference intervals between laboratories, the consensus value for serum X was calculated as the mean value of the single means of Roche, Abbott, and Siemens results, as described above. This value was divided by the individual laboratory's serum X result. The resulting serum X–based factor was used to multiply the upper and lower reference limits so that differences in reference intervals because of method differences were eliminated.

COMMUTABILITY OF EQA SAMPLES

The EQA materials and serum X were evaluated as commutable if the concentration was within the 95% prediction interval of the native serum samples.

The 95% prediction interval (PI) for the value x_0 was calculated as:

$$PI = 2 \cdot t_{0.025,\,n-2} \cdot S_{y \cdot x} \sqrt{(1 + 1/n + (x_0 - M_X)^2/SS_X)}. \quad (4)$$

where $t_{0.025,n-2}$ is the Student t variable for 95% probability and n − 2 degrees of freedom, where n is number of native sera (n = 21, i.e., including the native single sera and the fresh 18-donor serum pool), $S_{y/x}$ is the SE of y on x, M_X and SS_X are the mean and variance, respectively, of all x.

The residuals for the EQA samples and serum X were further divided by 0.5 × PI. Values >1 or < −1 indicated a result outside the prediction interval and were interpreted as noncommutability.

Results

Fig. 1 shows substantial differences between the various measurement procedures for 4 of 5 components compared with the target values. Only the folate assays showed measurement differences within the limit for desirable bias. Differences were registered also for methods in which all manufacturers claimed traceability to reference materials that are traceable to each other (see online Supplemental Table 1). The Roche methods showed very similar results for all components except for folate, in which a slightly constant positive bias was seen for the measurements done on the Roche Cobas instrument compared with Roche Modular. Furthermore, the Roche assays deviated from the other measurement procedures for TSH and ferritin, and the Beckman Coulter Unicel assay deviated from the other measurement procedures for cobalamin and free T_4. Fig. 1F shows that all of the commercial immunoassays measured too low free T_4 concentrations compared with the reference method. Online Supplemental Table S2 shows the correlation between the different measurement procedures.

Table 1 shows the consensus target values for serum X. The bias for the different components is compared to the quality specifications for bias based on biological variation (see Methods). The free T_4 data included both the deviation from the consensus target value and from the reference method value.

The pooled within-method analytical variation for the sera is shown in Fig. 2. The Roche methods showed low analytical variation for the cobalamin and free T_4 assays, and acceptable CV_A for the other components. The CV_A for the Abbott instruments was low for ferritin and acceptable for folate, free T_4, and TSH, whereas the CV_A was larger than desirable for cobalamin. Beckman Coulter Unicel had fairly low CV_A for all components, except cobalamin and free T_4, but the

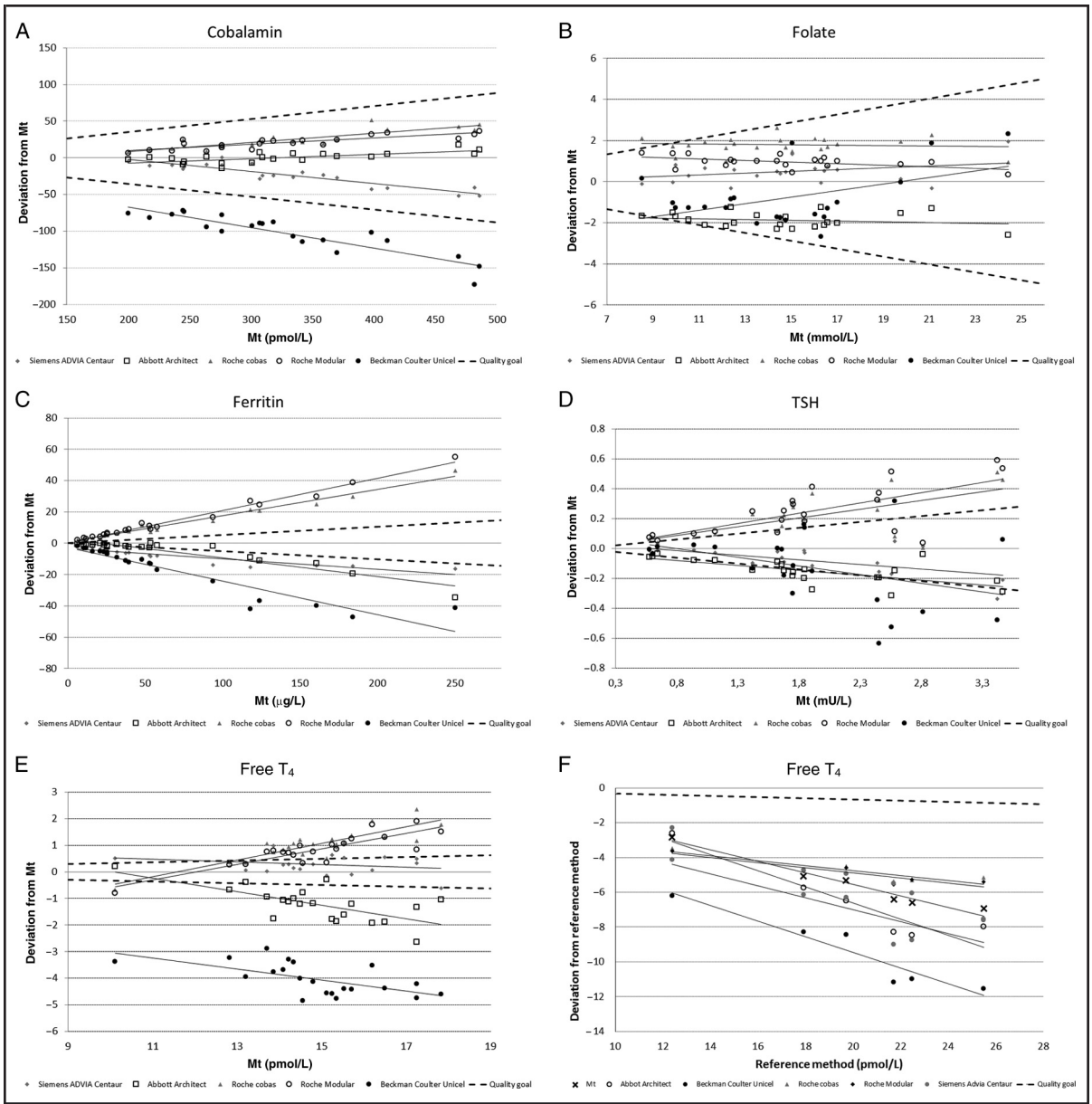

Fig. 1. Difference plots showing the bias between the measurement procedures for the 5 components.

The *x* axis shows the Mt or the concentrations measured by the reference method (1F). The *y* axis shows the deviation from Mt (or reference method, i.e., Fig. 1F) for the different measurement procedures. The dashed lines indicate the allowable bias. (A), Cobalamin (pmol/L); (B), folate (nmol/L); (C), ferritin (μg/L); (D), TSH (mU/L); (E), free T_4 (pmol/L); and (F), reference values for free T_4 (pmol/L).

results were based on data from only 2 laboratories. Siemens ADVIA Centaur showed the highest CV_A, with values above the desirable analytical quality for 4 of the components (cobalamin, folate, ferritin, and free T_4).

Fig. 3 shows the adjusted reference intervals (see Methods) for the components. Large variations were seen, especially regarding the upper limits. The upper reference limit for ferritin varied from 109 to 355 μg/L, and the lower limit varied from 7.7 to 18.7 μg/L. The highest within-method variation for reference intervals was observed for Siemens ADVIA Centaur.

The serum X and the 2 EQA materials were tested for commutability for all 5 components on all 5 measure-

Table 1. Bias from the target value (Mt or the concentrations measured by the reference method[a]) for the concentrations of serum X.

	Serum X	Quality goal[b]	Abbott Architect	Beckman Coulter Unicel	Roche Cobas	Roche Modular	Siemens ADVIA Centaur
Cobalamin, pmol/L	329	±58	6	**−108**	15	15	−21
Folate, nmol/L	14.0	±2.7	−1.4	−1.5	2.1	0.6	0.1
Ferritin, μg/L	62.4	±3.2	−1.5	**−13.8**	**9.4**	**11.1**	**−8.7**
TSH, mU/L	1.69	±0.13	−0.15	0.01	**0.16**	**0.20**	−0.03
Free T_4, pmol/L	14.3	±0.5	**−1.2**	**−3.1**	**0.8**	**0.7**	0.4
Free T_4, pmol/L[a]	19.7	±0.7	**−6.5**	**−8.4**	**−4.5**	**−4.6**	**−4.9**

[a] Reference method values [IFCC *(6)*].
[b] The quality goal (±absolute values) for desirable bias is based on biological variation data: cobalamin 17.7%, folate 19.2%, ferritin 5.2%, free T_4 3.3%, and TSH 7.8% *(11)*. Biases exceeding the quality goal are shown in bold.

ment procedures. Serum X results were commutable for all components and methods. Noncommutability for the EQA materials was observed on 13 out of 50 occasions (5 components, 5 methods, and 2 EQA samples) (Table 2). As expected, the poorest commutability (n = 10) was seen for the human-based lyophilized serum (EQA2). The liquid pooled serum (EQA1) was commutable for 4/5 components. None of the commercial EQA samples (1 and 2) were commutable for folate measurements performed by the Abbott Architect and Siemens ADVIA Centaur assays. Both samples were commutable for free T_4 analysis for all 5 measurement procedures. In 5 instances, the bias demonstrated by the EQA samples was found to be in an opposite direction compared

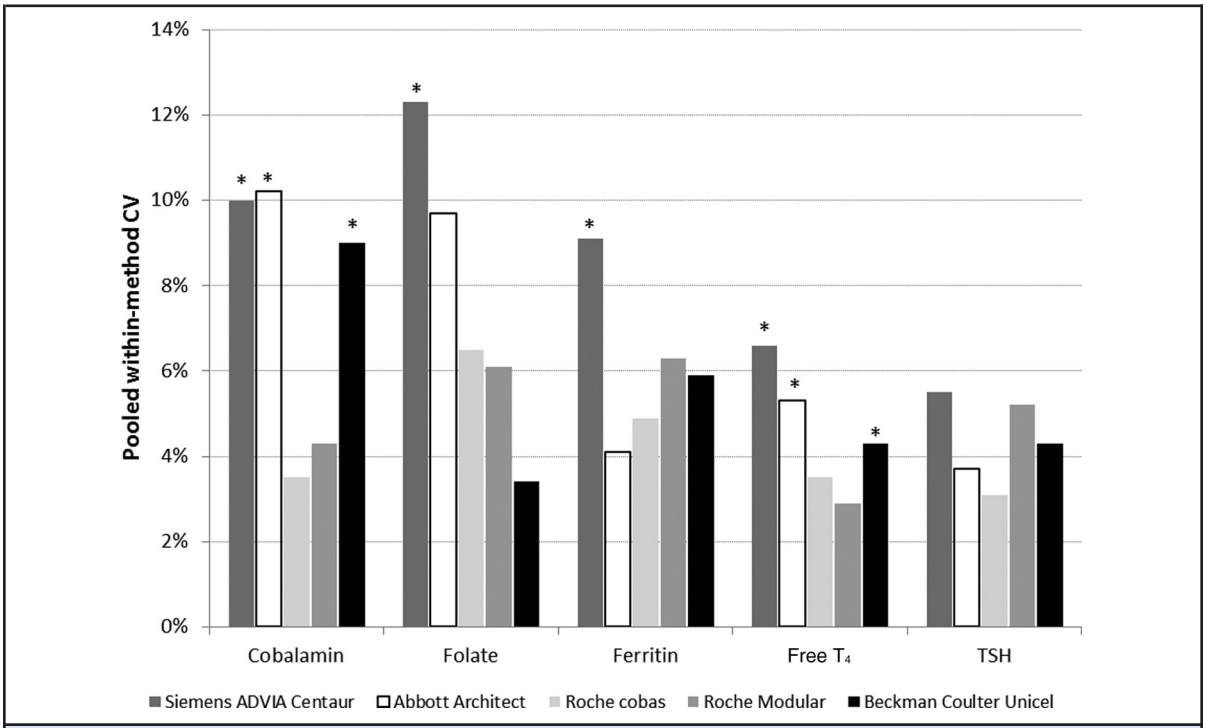

Fig. 2. Pooled within-method analytical variation for the components measured by 5 measurement procedures at 38 different laboratories.

Desirable CV_A were 7.5% (cobalamin), 12.0% (folate), 7.1% (ferritin), 14.7% (TSH), and 3.6% (free T_4) *(11)*. Assays exceeding the desirable CV_A are marked with an asterisk.

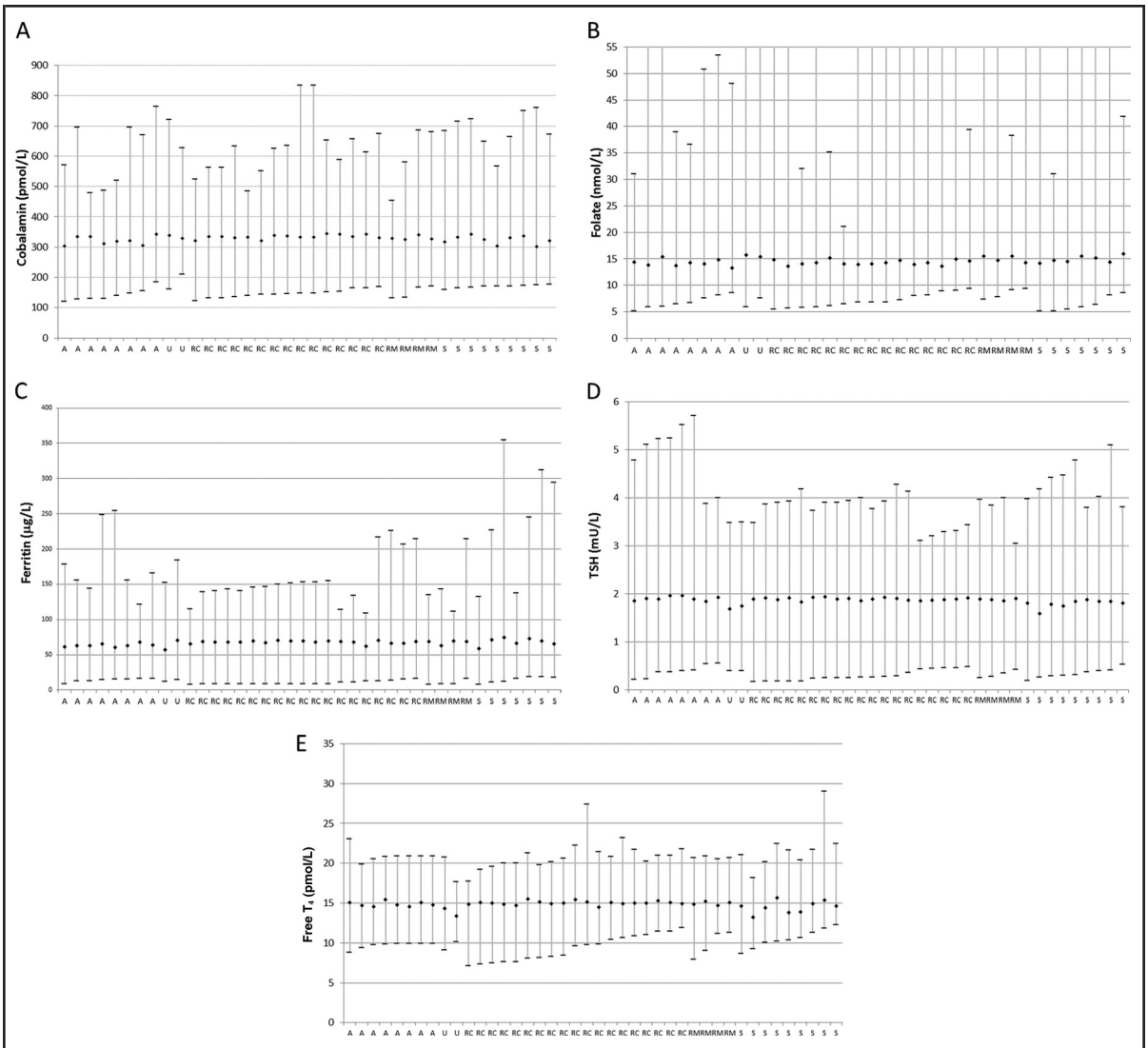

Fig. 3. Reference intervals reported after the lower and upper limits have been corrected by use of a factor derived by the single method mean of serum X (see Methods).

The black diamond shows the mean value of the samples from 20 healthy individuals included in the survey. A, Abbott Architect; U, Beckman Coulter Unicel; RC, Roche Cobas; RM, Roche Modular; S, Siemens ADVIA Centaur.

with the native serum samples, e.g., the folate analysis in the noncommutable EQA1 and EQA2 on the Abbott Architect and Siemens ADVIA Centaur platforms (Fig. 4). Folate analyzed in EQA samples on Abbott Architect instruments indicated a positive bias, whereas the native samples showed a negative bias. Siemens ADVIA Centaur results for folate showed EQA results with a negative bias, whereas the native samples showed a small positive bias. Similar findings for folate were also shown for the noncommutable EQA1 on the Beckman Coulter instruments.

Discussion

The main findings of this study are that the method differences for 4 out of 5 tested components were large, and the detected bias exceeded the acceptance criteria as defined by biological variation. In spite of claimed traceability to an international standardized reference material as calibrator (see online Supplemental Table S1), method differences for ferritin and TSH were not diminished compared with the other components that were tested. Reference intervals varied between laboratories and were

Table 2. Commutability of serum X and the EQA samples.[a]

	Cobalamin			Folate			Ferritin			Free T_4			TSH		
	Serum X	EQA1	EQA2	Serum X	EQA1	EQA2	Serum X	EQA1	EQA2	Serum X	EQA1	EQA2	Serum X	EQA1	EQA2
Abbott Architect	0.5	0.8	**1.3**	0.5	**3.5**	**3.7**	0.5	−0.6	**1.1**	−0.1	−0.1	0.3	−0.1	0.1	**−1.1**
Beckman Coulter Unicel	−0.2	−0.7	0.2	−0.2	**1.8**	0.1	0.2	−0.7	−0.1	0.7	0.0	0.7	0.2	−0.1	0.2
Roche Cobas	−0.7	−0.3	−0.3	0.4	0.3	**−1.1**	−0.5	0.6	**−1.8**	0.0	−0.2	0.1	−0.1	−0.1	0.5
Roche Modular	−0.5	−0.5	**−1.2**	−0.8	−0.3	**−1.7**	−0.5	0.4	**−2.2**	0.1	0.0	0.2	−0.1	−0.1	0.4
Siemens ADVIA Centaur	0.1	−0.3	−0.5	−0.3	**−2.6**	**−1.9**	−0.3	0.5	0.1	0.1	0.2	−0.5	0.3	0.0	0.3

[a] Results >1 or < −1 signal a noncommutable material (shown in bold). Serum X was a native serum pool that consisted of sera from 120 donors. It has been stored for 12 years. EQA1 was a commercial liquid pooled serum, and EQA2 was a human-based lyophilized serum.

not in agreement with the demonstrated method differences. A serum pool that had been stored for 12 years showed commutability for all 5 components and was able to reveal true bias between measurement procedures. A commercial human-based serum (EQA1) gave noncommutability on 3 out of 25 occasions (all for folate measurements), and the direction of the bias was opposite to the bias detected by the native material. A human-based lyophilized serum (EQA2) showed noncommutability on 10 out of 25 occasions, and in 2 of these instances, the bias was contradictory when compared to the bias demonstrated by the native material.

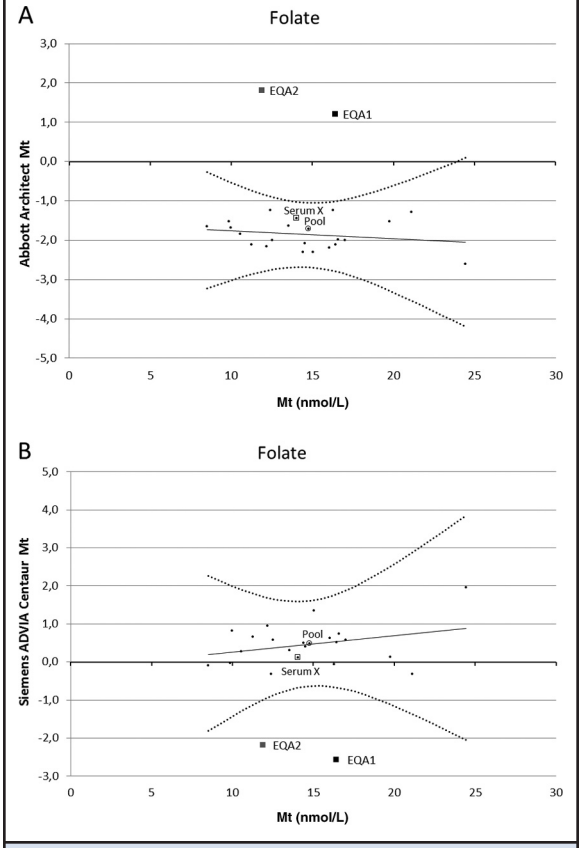

Fig. 4. Bias of the native serum samples, the fresh native serum pool (shown as a dot within a circle), serum X (shown as a dot within a square), and the noncommutable EQA material (EQA1 and EQA2) for folate on the Abbott Architect (A) and Siemens ADVIA Centaur (B) platform.

One of the strengths in this study was that it evaluated method differences for immunoassays that are at various stages along a harmonization process. The study included both components with traceability to an international reference material and components using in-house industry standards. The study also examined the implementation of different processes that might be undertaken to compensate for or reduce the effects of method differences.

Substantial method differences were detected for the 5 components studied, and all components except folate, which is a component with large biological variation and consequently a large acceptance criterion, exceeded the acceptance criteria based on biological variation. For the cobalamin immunoassays, the manufacturers used USP standards or internal standards as calibrators. A clear bias was detected between the measurement procedures; however, this was significant only for the Beckman Coulter Unicel assay and may partly be explained by the fact that this method was not included in the calculation of the Mt.

For the ferritin and TSH assays, all manufacturers claim traceability to reference materials that are reported to be mutually traceable to each other (see online Supplemental Table S1). The biases detected for these 2 measurement procedures were large but proportional. If proportionality can be demonstrated for the whole measurement range of the assays, it might be possible to harmonize the measurement procedure to the extent that common reference intervals might be used. An easier procedure for harmonizing measurement procedures is the

use of 1 common commutable calibrator. In 2013, Zegers et al. published a study demonstrating that the observed method differences between ceruloplasmin assays were a result of the noncommutability of the certified reference material that was used by the manufacturers (ERM-DA470) *(12)*. Noncommutability for SRM (standard reference material) has also been demonstrated for some troponin I assays *(13)*. The ferritin and TSH assays claim traceability to different WHO reference materials, and until now, these have not been tested for commutability *(14)*. Further efforts should be made to produce commutable reference materials and to ensure that their use has a clear effect on the harmonization of patient results. It is reassuring that several national and international organizations, including the Joint Committee for Traceability in Laboratory Medicine, the WHO, and International Federation of Clinical Chemistry scientific division, have increased their focus on this issue lately *(14, 15)*. The large scatter that was seen around the regression line for several of the TSH assays indicates the possibility of different analytical specificity for these assays.

All manufacturers claim traceability to their own in-house standard for the free T_4 assay, although a reference method has been available for several years *(6)*. Consequently, none of the assays showed agreement with the reference method. A likely reason for the slow implementation of the published reference method may be that all conventional assays show considerable bias to the reference method so implementation will lead to changes in the clinical interpretation of the test results and to redefinition of treatment-decision limits. Another reason might be that substantial changes in standardization will require a new approval of the test from different healthcare authorities (e.g., the US Food and Drug Administration), which is a very large process requiring substantial use of resources by the manufacturers.

As expected, the pooled within-method analytical variation differed between measurement procedures. Quite large variations were seen for some of the tested components or measurement procedures, indicating low method repeatability or low robustness of calibration procedures.

Reference intervals varied substantially between laboratories, and there was no clear correlation between the measurement procedure used and the suggested reference intervals. According to ISO (International Organization for Standardization) standard 15189, all laboratories should validate their reference intervals on a regular basis, and substantial resources are invested in this task. The data reported in our study indicated that this effort might be of little effect. The common assumption that laboratories are able to eliminate method differences by adjusting their reference intervals according to the method they use was not confirmed. Similar results have been shown by others *(16)*. The most likely reasons for this are that laboratories use different sources of information and different study designs with diverse quality when they establish their reference intervals and also when they validate them. The diversity of reference intervals may affect the interpretation of analysis results and have implications for the clinical treatment of patients *(17)*. There is a need for standardization or at least harmonization of reference intervals for different immunoassays *(16)*. A step in the right direction is the United Kingdom Pathology harmony initiative *(18)* and the IFCC launching of an initiative to establish reference intervals in different ethnic populations in several countries in America, Asia, Africa, and Europe *(3–5)*. It is encouraging that the study includes all common clinical-chemistry components and also suggests a procedure for later correlation and transfer of data to different measurement procedures.

Use of commutable EQA material for proficiency testing is essential to detect real method differences and is an important tool to achieve harmonization between measurement procedures *(2, 14, 19, 20)*. Our study showed that noncommutable EQA materials may have opposite direction of bias compared to patient samples, and this is a clear reminder that EQA results always must be interpreted in view of the commutability of the material used. Further, we also showed that liquid commercial serum samples, to some degree, might be noncommutable and consequently must be tested for commutability if evaluation of method differences is intended. The serum X material showed commutability for all components on all measurement procedures, although this pool had been stored for 12 years. Similar findings, i.e., commutability of serum pools and noncommutability for processed EQA materials, have also been demonstrated for immunoassays measuring troponin I *(13)*. Laboratories spend large resources on EQA schemes every year and should, together with health authorities and organizations for laboratory medicine, put pressure on EQA providers to strive for commutability testing of all EQA materials before they are put into use.

A limitation of our study was that the native single sera samples were not prepared in accordance with the CLSI C37A guideline. However, fresh patient samples were used in a manner completely similar to what is done in clinical practice (i.e., when general practitioners mail their samples to the laboratories). Another limitation was the procedure used for establishing the target value (Mt). However, in the absence of a reference method, there are not many other ways of determining a target value *(21)*. The third limitation was that we did not use clinically based quality standards for defining desirable imprecision and bias; only biological variation data were used. Clinically based standards could have been useful for several reasons: (*a*) not all components were gaussian distributed, (*b*) changes that hold clinical meaning for some

components are only 1-sided (cobalamin and folate), and (*c*) it may be argued that diagnostic cutoff limits should be used rather than reference intervals for several components (cobalamin, folate, and ferritin). However, currently there is no consensus regarding clinically based quality standards. It is reassuring that the European Federation of Laboratory Medicine has put this important topic on the agenda, and its members are working to define such standards that will be very useful in the future *(22, 23)*. Lastly, it should be noted that we did not ask the laboratories how their reference intervals were established or validated. No firm conclusions may be drawn regarding the diversity of the reference intervals detected in this study.

Conclusion

The procedures by which the clinical-chemistry community has recommended keeping method differences within acceptable limits are well known, e.g., traceability to reference materials or methods, adjustment of reference intervals, and use of commutable EQA samples. Even so, our data show that, for commonly used immunoassays, large method differences still persist. Reference intervals used by local laboratories are not always adjusted in accordance with existing method differences, and commercial EQA samples showed noncommutability on several occasions and, therefore, were useless as a tool for harmonization. The elimination of method differences would be a major contribution to future patient care, and the increased focus lately on commutability, as well as the IFCC global campaign on reference values are important steps in the right direction. However, this fundamental task for patient care may only be solved by laboratory professionals, healthcare authorities, manufacturers, and EQA providers jointly demanding commutability at all steps of the traceability chain. The clinical chemistry communities should keep up and increase their focus on this issue until clear improvement is demonstrated.

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

Authors' Disclosures or Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Acknowledgments: Sverre Sandberg at the Norwegian Quality Improvement of Primary Care Laboratories (Noklus) contributed to the planning of the study.

References

1. Koerbin G, Tate JR, Ryan J, Jones GR, Sikaris KA, Kanowski D, et al. Bias assessment of general chemistry analytes using commutable samples. Clin Biochem Rev 2014;35:203–11.
2. Miller WG, Jones GR, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. Clin Chem 2011;57:1670–80.
3. Borai A, Ichihara K, Al Masaud A, Tamimi W, Bahijri S, Armbuster D, et al. Establishment of reference intervals of clinical chemistry analytes for adult population in Saudi Arabia: a study conducted as a part of the IFCC global study on reference values. Clin Chem Lab Med 2016;54:843–55.
4. Ozarda Y, Ichihara K, Aslan D, Aybek H, Ari Z, Taneli F, et al. A multicenter nationwide reference intervals study for common biochemical analytes in Turkey using Abbott analyzers. Clin Chem Lab Med 2014;52:1823–33.
5. Ozarda Y, Ichihara K, Barth JH, Klee G. Committee on Reference Intervals and Decision Limits (C-RIDL), International Federation for Clinical Chemistry and Laboratory Medicine. Protocol and standard operating procedures for common use in a worldwide multicenter study on reference values. Clin Chem Lab Med 2013;51:1027–40.
6. IFCC. Laboratory Medicine Working Group for Standardization of Thyroid Function Tests, Van Houcke SK, Van Uytfanghe K, Shimizu E, Tani W, et al. IFCC international conventional reference procedure for the measurement of free thyroxine in serum: International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Working Group for Standardization of Thyroid Function Tests (WG-STFT)[(1)]. Clin Chem Lab Med 2011;49:1275–81.
7. Henriksen GM, Pedersen MM, Norgaard I, Blom M, Blou L, Blaabjerg O, Uldall A. Minimally processed fresh frozen human reference sera: preparation, testing, and application to international external quality assurance. Scand J Clin Lab Invest 2004;64:293–308.
8. Grubb FE. Sample criteria for testing outlying observations. Ann Math Statist 1950;21:27–58.
9. Fraser CG, Harris EK. Generation and application of data on biological variation in clinical chemistry. Crit Rev Clin Lab Sci 1989;27:409–37.
10. Ricos C, Alvarez V, Cava F, Garcia-Lario JV, Hernandez A, Jimenez CV, et al. Current databases on biological variation: Pros, cons and progress. Scand J Clin Lab Invest 1999;59:491–500.
11. Desirable biological variation database. https://www.westgard.com/biodatabase1.htm (Accessed April 2016).
12. Zegers I, Beetham R, Keller T, Sheldon J, Bullock D, MacKenzie F, et al. The importance of commutability of reference materials used as calibrators: The example of ceruloplasmin. Clin Chem 2013;59:1322–9.
13. Zhang S, Zeng J, Zhang C, Li Y, Zhao H, Cheng F, et al. Commutability of possible external quality assessment materials for cardiac troponin measurement. PLoS One 9:e102046, 2014.
14. Miller WG, Myers GL. Commutability still matters. Clin Chem 2013;59:1291–3.
15. Shinohara K, Hamasaki N, Takagi Y, Yatomi Y, Kikuchi H, Hosogaya S, et al. Multianalyte conventional reference material (MacRM): a useful tool for nationwide standardization of laboratory measurements for medical care–a model study in Japan. Clin Chem 2016;62:392–406.
16. Jones GR, Koetsier SD. RCPAQAP first combined measurement and reference interval survey. Clin Biochem Rev 2014;35:243–50.
17. Tate JR, Yen T, Jones GR. Transference and validation of reference intervals. Clin Chem 2015;61:1012–5.
18. Berg J. The UK pathology harmony initiative; the foundation of a global model. Clin Chim Acta 2014;432:22–6.
19. Tate JR, Johnson R, Barth J, Panteghini M. Harmonization of laboratory testing–current achievements and future strategies. Clin Chim Acta 2014;432:4–7.
20. Miller WG, Myers GL, Ashwood ER, Killeen AA, Wang E, Ehlers GW, et al. State of the art in trueness and interlaboratory harmonization for 10 analytes in general clinical chemistry. Arch Pathol Lab Med 2008;132:838–46.
21. Stepman HC, Tiikkainen U, Stockl D, Vesper HW, Edwards SH, Laitinen H, et al. Measurements for 8 common analytes in native sera identify inadequate standardization among 6 routine laboratory assays. Clin Chem 2014;60:855–63.
22. Horvath AR, Bossuyt PM, Sandberg S, John AS, Monaghan PJ, Verhagen-Kamerbeek WD, et al. Setting analytical performance specifications based on outcome studies–is it possible? Clin Chem Lab Med 2015;53:841–8.
23. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, et al. Defining analytical performance specifications: consensus statement from the 1st strategic conference of the European Federation of Clinical Chemistry and Laboratory Medicine. Clin Chem Lab Med 2015;53:833–5.