Harmonization of Serum Thyroid-Stimulating Hormone Measurements Paves the Way for the Adoption of a More Uniform Reference Interval

Linda M. Thienpont,[1,2*] Katleen Van Uytfanghe,[3] Linde A.C. De Grande,[1] Dries Reynders,[4] Barnali Das,[5]
James D. Faix,[6] Finlay MacKenzie,[7] Brigitte Decallonne,[8] Akira Hishinuma,[9] Bruno Lapauw,[10]
Paul Taelman,[11] Paul Van Crombrugge,[12] Annick Van den Bruel,[13] Brigitte Velkeniers,[14] and Paul Williams[15]
on behalf of the IFCC Committee for Standardization of Thyroid Function Tests (C-STFT)

BACKGROUND: The IFCC Committee for Standardization of Thyroid Function Tests developed a global harmonization approach for thyroid-stimulating hormone measurements. It is based on a multiassay method comparison study with clinical serum samples and target setting with a robust factor analysis method. Here we describe the Phase IV method comparison and reference interval (RI) studies conducted with the objective to recalibrate the participating assays and demonstrate the proof-of-concept.

METHODS: Fourteen manufacturers measured the harmonization and RI panel; 4 of them quantified the harmonization and first follow-up panel in parallel. All recalibrated their assays to the statistically inferred targets. For validation, we used desirable specifications from the biological variation for the bias and total error (TE). The RI measurements were done with the assays' current calibrators, but data were also reported after transformation to the new calibration status. We estimated the pre- and postrecalibration RIs with a nonparametric bootstrap procedure.

RESULTS: After recalibration, 14 of 15 assays met the bias specification with 95% confidence; 8 assays complied with the TE specification. The CV of the assay means for the harmonization panel was reduced from 9.5% to 4.2%. The RI study showed improved uniformity after recalibration: the ranges (i.e., maximum differences) exhibited by the assay-specific 2.5th, 50th, and 97.5th percentile estimates were reduced from 0.27, 0.89, and 2.13 mIU/L to 0.12 , 0.29, and 0.77 mIU/L.

CONCLUSIONS: We showed that harmonization increased the agreement of results from the participating immunoassays, and may allow them to adopt a more uniform RI in the future.

Given the prevalence and gravity of thyroid disorders, timely diagnosis, initiation, and monitoring of therapy are important to restrict the impact of the disease on public health. Measurement of serum thyroid hormone concentrations is an indispensable tool to confirm the disease, particularly because the clinical symptoms often resemble other disorders or are subtle in case of subclinical thyroid dysfunction *(1, 2)*. The main clinical scenarios for measurement of serum thyroid-stimulating hormone (TSH)[16] are screening for thyroid dysfunction, evaluation of thyroid hormone replacement for primary hypothyroidism, and assessment of suppressive therapy in patients with follicular cell-derived thyroid cancer. Professional practice guidelines incorporate laboratory testing of thyroid function in patient care *(3–7)*. Reference intervals (RI) reported along with the laboratory data are an integral part of the interpretation process

[1] Department of Pharmaceutical Analysis, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium; [2] Current affiliation: Thienpont & Stöckl Wissenschaftliches Consulting GbR, Rennertshofen (OT Bertoldsheim), Germany; [3] Ref4U, Laboratory of Toxicology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium; [4] Department of Applied Mathematics, Computer Science, and Statistics, Faculty of Sciences, Ghent University, Ghent, Belgium; [5] Biochemistry and Immunology Laboratory, Kokilaben Dhirubhai Ambani Hospital and Medical Research Institute, Mumbai, India; [6] Clinical Chemistry and Immunology, Montefiore Medical Center, and Department of Pathology, Albert Einstein School of Medicine, New York, NY; [7] Birmingham Quality/UK NEQAS, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK; [8] Department of Endocrinology, University Hospitals Leuven, Leuven, Belgium; [9] Department of Infection Control and Clinical Laboratory Medicine, Dokkyo Medical University, Tochigi, Japan; [10] Department of Endocrinology, Ghent University Hospital, Ghent, Belgium; [11] Laboratory of Endocrinology, Department of Laboratory Medicine, AZ Maria-Middelares Sint-Jozef, Campus Maria-Middelares, Ghent, Belgium; [12] Department of Endocrinology, OLV Ziekenhuis Aalst-Asse-Ninove, Aalst, Belgium; [13] Department of Endocrinology, General Hospital Sint Jan, Bruges, Belgium; [14] Department of Endocrinology, Universitair Ziekenhuis Brussel, Brussels, Belgium; [15] Department of Endocrinology, Royal Prince Alfred Hospital, Camperdown, Australia.

* Address correspondence to this author at: Thienpont & Stöckl Wissenschaftliches Consulting GbR, Erlbacher Strasse 11, Rennertshofen (OT Bertoldsheim), Germany. E-mail linda.thienpont@ugent.be.

Received November 17, 2016; accepted April 4, 2017.
Previously published online at DOI: 10.1373/clinchem.2016.269456

[16] Nonstandard abbreviations: TSH, thyroid-stimulating hormone; RI, reference interval; IVD, in vitro diagnostic; C-STFT, Committee for Standardization of Thyroid Function Tests; APTM, all-procedure trimmed mean; TE, total error; LL, lower limit; UL, upper limit.

(8, 9). Since many laboratory measurements are not yet comparable, RIs are typically established for each assay and are considered assay-specific. For physicians who only use one laboratory and are aware of these technical issues, this practice is fine. However, those who request test results from different laboratories, are often faced with challenges owing to different RIs. Assay-specific RIs are also problematic for patients who regularly move between geographic locations and/or are seen by different doctors *(10)*. More generally, assay-specific measurement results prevent the development of modern public health standards, such as clinical guidelines quoting fixed decision limits and integration of electronic patient records in the healthcare system *(11)*. Paramount to the goal of using common RIs is the establishment of metrological traceability of in vitro diagnostic (IVD) medical devices—also called standardization *(12–14)*. As the IFCC Committee for Standardization of Thyroid Function Tests (C-STFT) members, we decided to focus our efforts on immunoassays for TSH and free thyroxine in partnership with the IVD industry *(15)*. Our premise was that, if possible, we should adhere to the concept for traceability recommended by the International Organization for Standardization *(16)*. Although a reference measurement procedure existed for free thyroxine, we considered this option for TSH unlikely and developed a pragmatic approach to harmonization rather than standardization *(17, 18)*. To circumvent the often encountered commutability issues in establishing calibration traceability of IVD assays, it was a premise for C-STFT that harmonization should be done from a multiassay method comparison study with a panel of native and clinically relevant samples *(19–21)*. We developed a robust factor analysis method for estimation of the harmonization targets and demonstrated the equivalence of the approach to standardization to a reference measurement procedure *(22, 23)*.

Here we report on behalf of the C-STFT the most recent Phase IV studies in our TSH harmonization efforts in which we demonstrate that establishing calibration traceability of commercially available immunoassays enables the adoption of a more uniform RI for TSH.

Materials and Methods

PANELS OF CLINICAL SAMPLES

To allow manufacturers to adjust their calibration to the harmonization basis we developed, we performed a new method comparison for Phase IV. We sourced samples from 2 commercial companies (in.vent Diagnostica GmbH; Solomon Park Research Laboratories) but also with the aid of 8 different outpatient thyroid clinics in Belgium, Japan, and Australia. The goal was to obtain a harmonization and first follow-up panel each comprising samples with concentrations that reasonably cover the measurement intervals of the participating TSH immunoassays. C-STFT provided the eligibility and exclusion criteria (see Section 3 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol63/issue7). Blood (ca. 50 mL per donor) was collected in serum separator tubes to mimic routine conditions and locally processed into off-the-clot serum. Samples were stored at −70 °C and transported under dry ice to either the Europe- or US-based company for aliquoting. The aliquots of the 1st follow-up panel are stored in the facilities of the National Institute for Biological Standards and Control (UK). For all collections the approval of a Bioethics Committee and written informed consent from patients were received. The deidentified samples were accompanied by a short description of the patients' clinical background (type of thyroid dysfunction, comorbidities, surgery/treatment, ethnicity, sex, etc.). The TSH harmonization and first follow-up panels comprised 101 and 95 samples, respectively.

For the RI study, 120 samples from American individuals were sourced under identical conditions from Solomon Park Research Laboratories. Selection criteria were negativity in antithyroperoxidase antibody screening and a serum TSH concentration <10 mIU/L (cutoff recommended for starting with replacement therapy; testing performed with the Tosoh AIA-2000 platform) *(4, 5)*.

STUDY PARTICIPANTS

Fourteen IVD manufacturers participated, each with one immunoassay (coding and further details in Table 1).

ASSIGNMENT OF TARGET VALUES

Two "targets"—actually, 2 sets of 101 sample-specific value assignments—for the harmonization panel, referred to below (for historical reasons) as all-procedure trimmed mean (APTM)-11 and APTM-4, were assigned using a robust factor analysis model *(22)*. The first target, the APTM-11, was derived from the results reported by all manufacturers but 3, i.e., manufacturer E whose assay design was in contrast to that of all others not real third generation, and N and O who joined the project 1 year after the validation of the target setting described in this report had been completed. The second target, APTM-4, was based on the results of 4 manufacturers only (identified in Table 1), i.e., those who measured both the harmonization and first follow-up panel in the same run. The data from these 2 panels (n = 196) were pooled to statistically estimate the APTM-4 targets.

STUDY MEASUREMENT PROTOCOL

In the method comparison study, all IVD manufacturers quantified the harmonization panel. The samples

Table 1. Study participants (ordered by code given in this report), inclusive the platforms/TSH assays and number of samples considered for validation of the recalibration process. The listed reference and measurement intervals are those stated in the kit inserts.

IVD manufacturer *Platform/Immunoassay*[a,b]	Code	Reference Interval (mIU/L)	Measurement Interval (mIU/L)[e,f]	N[g]
Siemens Healthineers (Tarrytown, NY) *Advia Centaur XP*	A[c,d]	0.55–4.78 (n = 229)	0.008–150	89
Abbott Diagnostics (Abbott Park, IL) *Architect i2000*	B[c,d]	0.35–4.94 (99%, n = 549)	0.010–100	88
[a]Shenzhen Mindray Bio-Medical Electronics Co., Ltd. (Shenzhen, China) *CL-2000i*	C[d]	0.35–5.10	0.020–100	87
Ortho-Clinical Diagnostics (Buckinghamshire, UK) *Vitros ECi*	D[d]	0.47–4.68 (95%, n = 525)	0.015–100	85
bioMérieux SA (Marcy-l'Etoile, France) *Vidas*	E	0.25–5.00 (n = 60)	0.050[f]–60.0	77
Beckman Coulter Inc. (Brea, CA*) Access 2*	F[d]	0.34–5.60 (95%, n = 217)	0.015–100	86
DiaSorin S.p.A (Saluggia, Italy) *Liaison® Analyser*	G[d]	0.30–3.60 (95%, n = 519)	0.020–100	90
[a]Sichuan Maccura Biotechnology Co., Ltd (Chengdu, China) *IS1200*	H[d]	0.30–4.04 (95%, n = 146, Chinese) 0.37–3.76 (95%, n = 299, Europeans)	0.020–100	86
Roche Diagnostics GmbH (Mannheim, Germany) *Elecsys (Cobas e 601)*	I[c,d]	0.27–4.20 (95%, n = 516)	0.014–100	88
Tosoh Corporation (Tokyo, Japan) *AIA-2000*	J[c,d]	0.38–4.31 (95%, n = 497)	0.010–100	89
[a]Snibe Co., Ltd. (Shenzhen, China) *Maglumi 2000*	K[d]	0.30–4.50 (95%)	0.020–100	87
[a]Fujirebio Inc. (Tokyo, Japan) *Lumipulse G1200*	L[d]	0.31–3.07 (95%, n = 140)	0.0042[f]–200	90
[b]LSI Medience Corporation (Tokyo, Japan) *STACIA*	N	0.48–4.15	0.002[f]–100	88
[b]Sysmex Corporation (Kobe, Japan) *HISCL-5000*	O	0.34–4.22 (n = 134)	0.002–100	91

[a,b] Manufacturers who only joined in 2015[a] and/or 2016[b] for participation in the Phase IV method comparison study.
[c] Data from these manufacturers were used to calculate the APTM-4.
[d] Data from these manufacturers were used to calculate the APTM-11.
[e,f] The lower limit of the measurement intervals is the functional sensitivity unless differently stated as [f]limit of quantitation defined by CLSI's EP17 *(24)*.
[g] Actual number of samples taken into consideration in the validation of the recalibration [this number was related to each assay's measurement interval and was maximum 101 (total number of samples in the harmonization panel)].

were measured in a randomized sequence specified by us, in singleton on each of 2 days; the individual results were reported. The manufacturers also included their master calibrators (note, these are the calibrators used for in-house value assignment to the product calibrators) for measurement in parallel with the panel samples and according to the same protocol. In the RI study, which was performed a minimum 6 months after the method comparison, the samples were measured in order of ascending ID number, in singleton and within run. Organization and interpretation of internal QC was left to the discretion of each manufacturer.

RECALIBRATION OF IMMUNOASSAYS

We calculated both the APTM-11 and APTM-4 targets for the harmonization panel and sent the IVD manufacturers a preliminary report with the intention that both targets would be used in recalibration. Manufacturers recalibrated by value reassignment of their master calibrators to the APTM-11 and APTM-4 targets following their in-house mathematical procedure without disclosing it to us. In essence the process consisted of fitting the respective APTM values and instrumental response data for the patient samples into an equation, and solving it for concentrations as a function of the responses registered for the master calibrators; the process continued with recalculating the results for the patient samples as if the revised master calibrators were used for calibration. The manufacturers reported back 2 sets of results, i.e., recalibrated to either the APTM-11 or APTM-4. For the measurements of the RI panel, manufacturers also reported the pre- and postrecalibration results; the latter were based on mathematical transformation of the former using the master calibrators revised in the harmonization study.

DATA TREATMENT

For data treatment in the method comparison study, we used Microsoft EXCEL®. We focused on 2 objectives: decide which APTM (APTM-11 or APTM-4) to use as a basis for harmonization, and demonstrate/validate the suitability of the recalibrated results to meet the analytical specifications stated below. For the first objective, we calculated/plotted the differences (%) between the 2 APTMs relative to their mean; in addition, we compared the outcome of the recalibration of the assays to each of the APTMs by ordinary linear regression analysis. To do so, we calculated for each sample the overall mean concentration from the results reported by the manufacturers after recalibration to the APTM-11 (*y* axis) and APTM-4 (*x* axis). For the second objective, we considered for each assay (*a*) the pre- and postrecalibration median deviation (%) to the target in distinct concentration intervals; (*b*) the mean deviation or bias (%; and 1-sided 95% CI) to the target after recalibration; (*c*) the pre- and postrecalibration CVs (%) of the assay means, and (*d*) the total error (TE; %) for the first replicate after recalibration. For treatment of the pre- and postrecalibration data for the RI study we used the CBstat software (version 5.1, K. Linnet, www.cbstat.com). It comprises the Anderson–Darling test to assess the data for normality, before selecting the appropriate procedure to estimate the RI characteristics [among others, the 2.5th and 97.5th percentiles, further referred to as lower limit (LL) and upper limit (UL), respectively]. In addition, the software supplies the 90% CIs of the estimates. Since none of the data sets was normally distributed, also not after log transformation ($P < 0.01$), we opted for the nonparametric bootstrap (500 replicates) procedure *(25)*. We also estimated the pre- and postrecalibration overall RI, after applying the robust factor analysis model on the results of the 14 participating assays. To investigate the effect of recalibration on the uniformity of the RI characteristics, we calculated the reduction of the CV (%) of the assay means, and compared the pre- and postrecalibration medians and percentiles of the individual RIs to those of the overall RI.

ANALYTICAL SPECIFICATIONS

For validation of the recalibration data we used the desirable specifications for bias and TE based on the biological variation, i.e., 7.8% (bias) and 23.8% (TE) *(26)*.

HOMOGENEITY AND STABILITY STUDY

We assessed the homogeneity from a subset of 12 samples (12 aliquots per sample) collected in parallel with the samples for the method comparison study (but not included in the harmonization panel). The TSH concentrations in this sample set were in the low, mid, and high range (4 test samples per interval). Because 2 companies had been involved in aliquoting, we did this study for both. A protocol described for certified reference materials was adopted *(27)*. Note that the stability study is ongoing. For details on both studies, see the online Supplement, Sections 1 and 2. All Supplemental Tables and Figs also are available in the online Supplement.

Results

CONCENTRATION INTERVAL COVERED BY THE CLINICAL SAMPLES IN THE METHOD COMPARISON STUDY

The full TSH concentration interval of the harmonization panel was from 0.001 mIU/L to 172 mIU/L (based on APTM-11) and 0.002 mIU/L to 193 mIU/L (based on APTM-4). Note, the reason for the discrepancy between the highest TSH concentration according to the APTM-4 and APTM-11 was that, coincidentally, the 4 selected assays in APTM-4 all reported a higher measurement result. The concentrations in the follow-up panel were between 0.002 and 169 mIU/L (based on APTM-4). In online Supplemental Fig. 1S and online Supplemental Fig. 2S the uncertainties of the APTM-4 estimates are shown. The overall relative uncertainties amounted to 0.7% (for the upper part of the CI of the estimate) and 1.0% (the lower part CI). The mean difference between the APTM-4 and APTM-11 targets relative to their mean was −0.6% (see online Supplemental Fig. 3S). Regression analysis of the overall mean results calculated from the results reported by the manufacturers after recalibration to either the APTM-11 or APTM-4 gave $[\text{mean results}_{\text{recal to the APTM-11}}] = 0.987\ [\text{mean results}_{\text{recal to the APTM-4}}] + 0.055\ (R^2 = 0.9999)$; the mean difference was −2.2% (see online Supplemental Fig. 4S). Based on this outcome and recognizing the value of using targets inferred from the results by the 4 assays that measured both the harmonization and first follow-up panel in the same run (details in the online Supplement, Section 13), we decided to use the APTM-4 for recalibration.

VALIDATION OF THE EFFECTIVENESS OF RECALIBRATION

Only the results within the assays' claimed measurement intervals were used (see Table 1). The combined difference (%) plots (Fig. 1, A and B) show the assays' deviations to the APTM-4 before (Fig. 1A) and after recalibration (Fig. 1B). Note, the latter was constructed using the measurement data mathematically recalculated with the reassigned master calibrators. Fig. 2, A and B, demonstrate the assay-specific median deviations (%) to the APTM-4 before and after recalibration in 3 concentration intervals. Fig. 2A shows the combined picture of the deviations with indication of the 15th, 50th, and 85th centiles, while Fig. 2B represents for each assay the magnitude and sign of the deviations. From the details listed

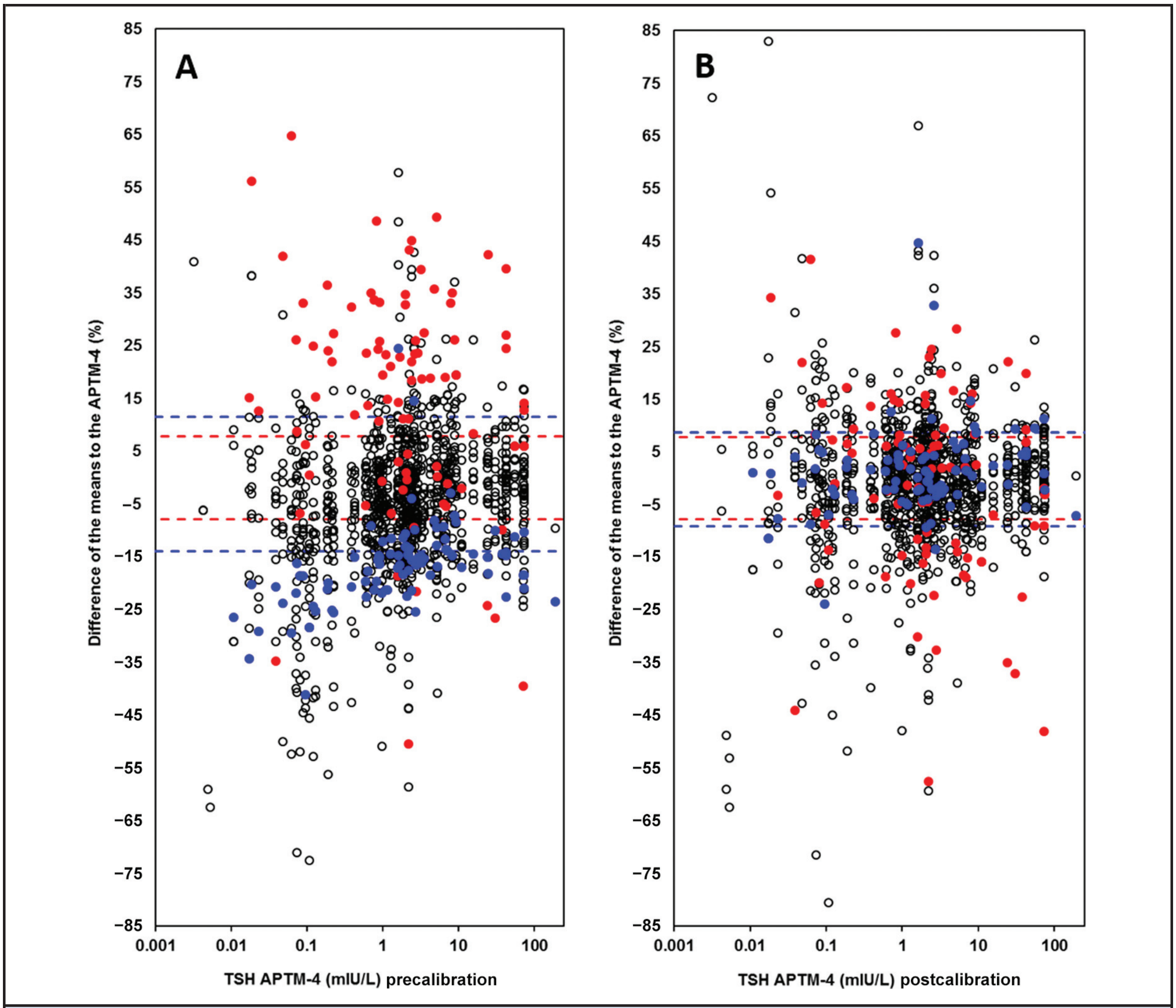

Fig. 1. Combined difference (%) plots to the APTM-4 before (A) and after recalibration (B).

For each assay and sample, the difference of the mean from duplicate measurements is plotted. The differences of the most discrepant assays before recalibration are highlighted by filled and colored circles: assay C, red (highest positive mean difference at approximately 15%), assay L, blue (highest negative mean difference at −16.5%); those of all other assays are shown by open black circles. For the sake of resolution, the plots do not include samples with a % difference beyond ±85% (13 and 10 samples before and after recalibration, respectively). The red broken lines are the 7.8% bias limits based on biological variation; the blue broken lines represent the 15th and 85th centiles. Note that as a result of recalibration, the symbols of the most discrepant assays are centered around zero % difference, and that the % differences of the centiles are reduced by one-third.

in online Supplemental Table 3S, one can see that before recalibration, the median deviations ranged from −41% (D) to +23% (C; <0.5 mIU/L), −15% (L) to +19% (C; ≥0.5 mIU/L to 5 mIU/L), and −14% (B, L) to 8% (C; ≥5 mIU/L), hence, the deviations of the most discrepant assay pairs (D and C, L and C, and B/L and C) were respectively 64%, 34%, and 22% apart from each other. After recalibration, the ranges of the median deviations were reduced, from −20.7% (K) to +16% (I), −8.0% (H) to +7% (B), −7% (C) to 6% (O), respectively.

Fig. 3 shows that the bias (%; and 1-sided 95% CI) of 13 of the 14 recalibrated TSH assays met the specification of 7.8%. For assay H (bias: −6.6%) the specification was not met with 95% confidence *(28)* (for details on the interpretation, see online Supplemental Table 4S).

Recalibration reduced the CV of the assay means for the harmonization panel from 9.5% to 4.2% (concentration interval from 0.5 mIU/L to 5.0 mIU/L) and from 7.5% to 4.4% (concentration interval between 0.0175 mIU/L and 74 mIU/L). The CV profile for the larger

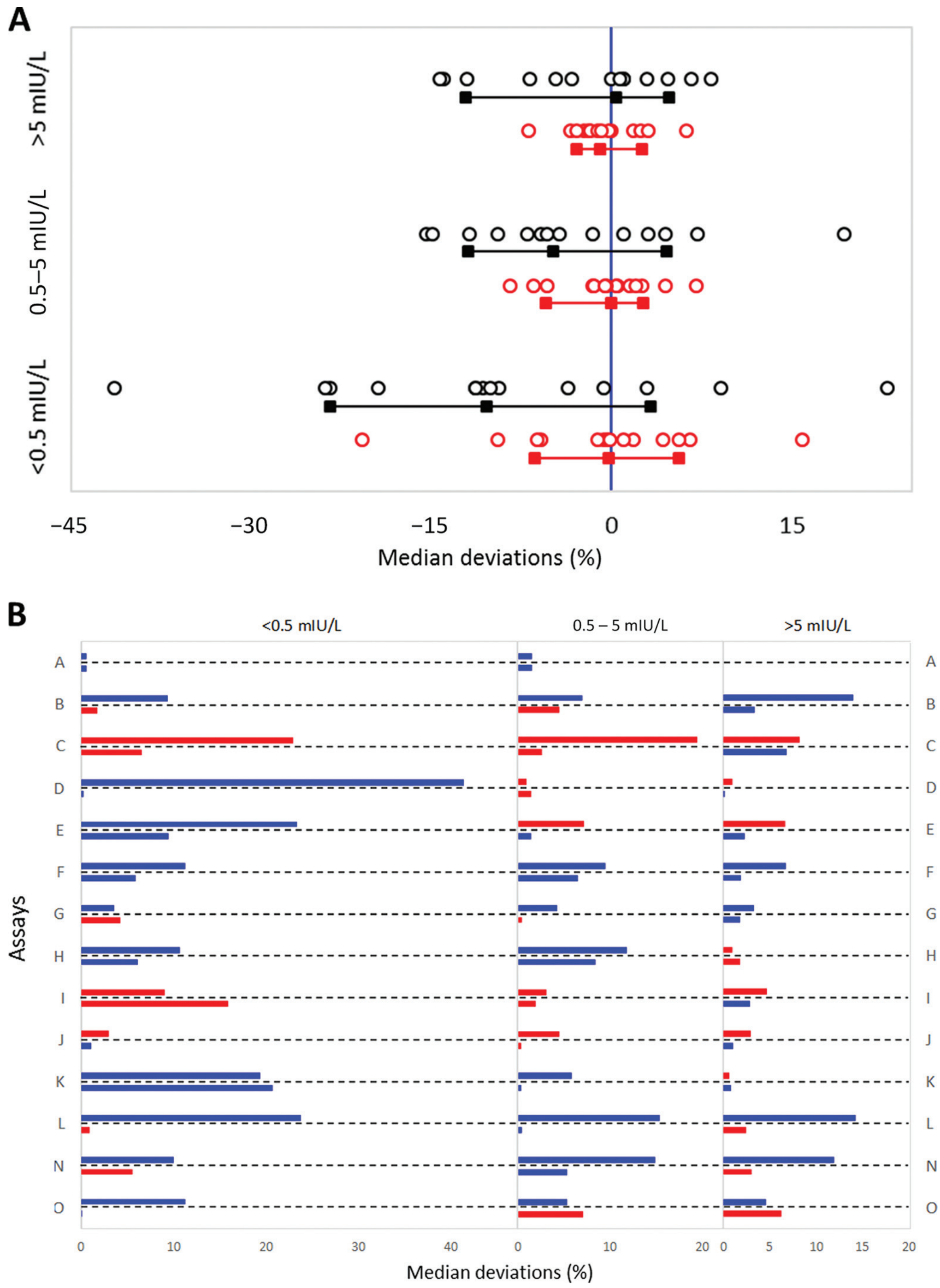

Fig. 2. Median deviations (%) of the assays to the APTM-4 before and after recalibration in 3 concentration intervals: low: <0.5 mIU/L, mid: ≥0.5 to <5.0 mIU/L, high: ≥5.0 mIU/L.

(A), Summarizes the overall improvement in terms of the median deviations (%) by recalibration. For each concentration interval, 2 pairs of data are shown; the black and red dots show the combined assay-specific median deviations before and after recalibration, respectively; the lines represent their 15th, 50th, and 85th centiles. (B), Represents the median deviations (%) of each individual assay by a pair of bars; the upper and lower bar shows the median deviation before and after recalibration. Note that the bars show the unsigned magnitudes, but the colors represent the signs (blue: negative, red: positive). Note, for assay A (>5 mIU/L) the deviations were zero.

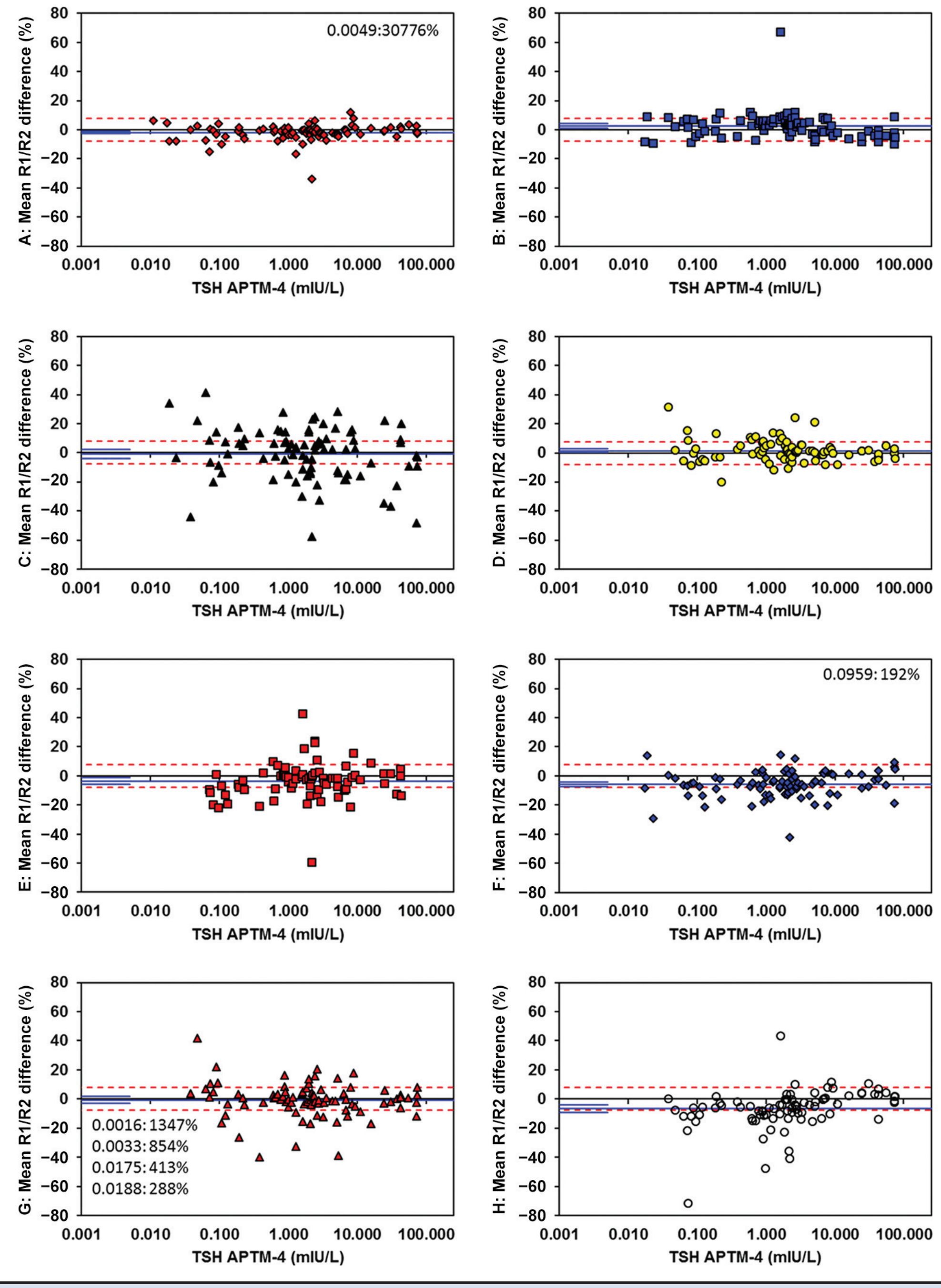

Fig. 3. Difference (%) plots after recalibration of the assays to the APTM-4.

The red broken lines are the bias limits of 7.8%, while the blue full lines represent each assay's mean deviation or bias (%) for the claimed measurement interval (detailed in Table 1). The short and parallel blue lines (left in the plots) represent the limits of the 1-sided 95% CI of the bias. Note that the samples for which the deviation was beyond 80% were not included in the % difference plots; they are identified in the respective graphs by their concentration and % difference. To avoid confusion: the concentration given in the graph is based on the APTM-4, for which the concerned assay reported a result within its measurement interval.

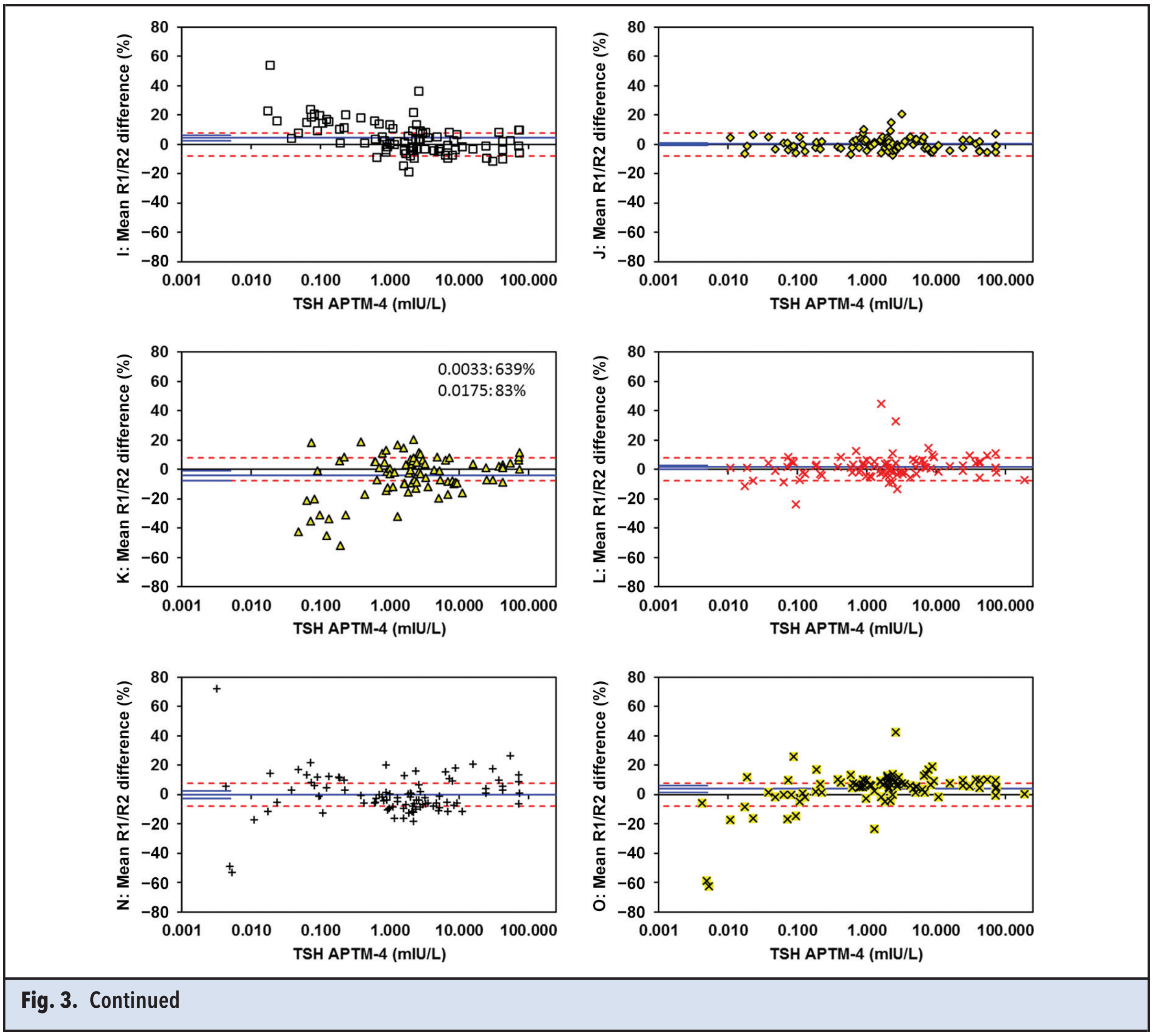

Fig. 3. Continued

interval is shown in online Supplemental Fig. 5SA. In terms of TE, 8 of the recalibrated TSH assays (A, B, D, F, I, J, L, N) met the specification (<5% of the differences >23.8%), while for the other 6 assays, 7% to 15% were outside the limits (Fig. 4).

RI STUDY

Fig. 5 gives an overview of the medians and percentiles (both with the 90% CIs) of the overall and individual RIs before and after recalibration (data available in online Supplemental Table 5S). Fig. 5 shows how the uniformity of the RIs (medians and percentiles) was improved by recalibration, as the latter narrowed the ranges of the medians by approximately one third (expressed relative to the median of the overall RI). The range before recalibration was from 1.20 mIU/L (assay N) to 2.09 mIU/L (assay C), and after recalibration was from 1.58 mIU/L (assay N) to 1.87 mIU/L (assay O). Online Supplemental Table 5S shows a similar effect of recalibration on the percentiles. Before recalibration the maximum deviations for the LL and UL amounted to 53% and 51% (assays C and N), while after recalibration the most deviating assays were 21% apart from each other for the LL (assays I and N) and 18% for the UL (assays O and N). Recalibration also considerably reduced the CV (%) of the assay means for the RI measurements, i.e., from 11.9% to 4.8% (see also online Supplemental Fig. 5SB). This reduction in CV for the RI panel compared well with the CV decrease observed for the same concentration interval of the harmonization panel.

HOMOGENEITY STUDY

Statistical testing confirmed that the hypothesis of homogeneity of the samples in the 3 panels could be accepted ($P >$ 0.05, see the online Supplement, Section 1 for details).

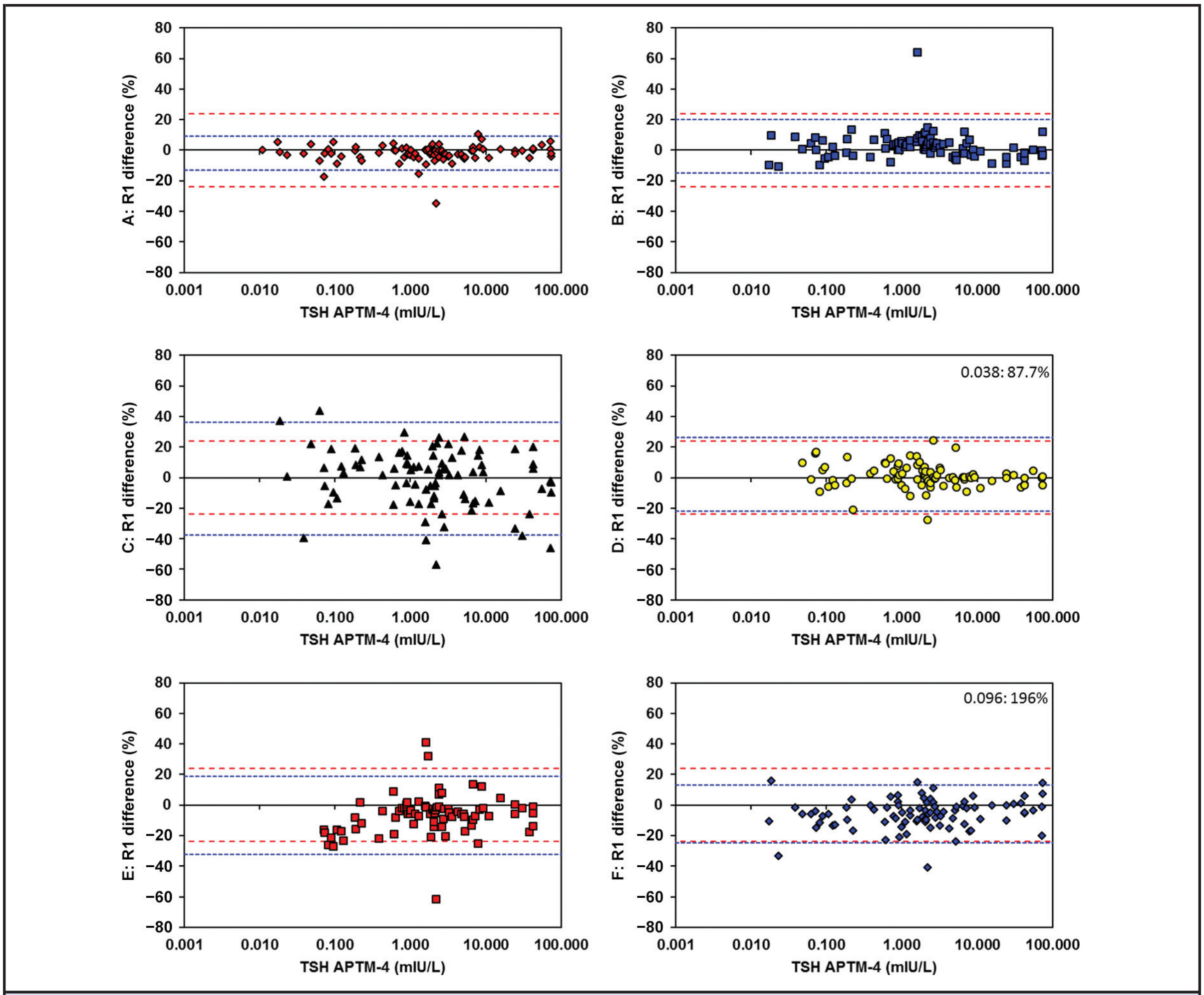

Fig. 4. Total error (%) plots of the first replicate after recalibration to the APTM-4.

The TE was estimated from the % difference to the APTM-4 of the first replicate after recalibration. It was validated against the 23.8% specification derived from the biological variation (red broken lines). The 95% limits of agreement [mean % difference ±1.96 CV_{diff} (%); blue broken lines] emphasize the fact that the magnitude of the scatter in the plots is different from assay to assay. Note that to keep the resolution of the graphs reasonable, the samples for which the deviation was beyond 80% were not included, but are identified in the respective graphs by their concentration and % difference. To avoid confusion: the given concentration is based on the APTM-4, for which the concerned assay reported a result within its measurement interval.

Discussion

Our attempt to harmonize commercially available TSH immunoassays began with a method comparison using samples from presumably healthy individuals (Phase I), in which we showed that recalibration using the APTM significantly increased the agreement of commercially available assays *(29)*. Allowing the manufacturers to individually adjust their own calibrators using the APTM from another method comparison with a similar panel of euthyroid samples (Phase II) established a proof-of-concept that the approach to harmonization was feasible *(30)*. Recalibration to the APTM was similarly successful using samples from patients with thyroid disease (Phase III). In addition, the overall excellent correlation of most of the immunoassays' results to the APTM in patients with both hypo- and hyperthyroidism led the committee to conclude that the assays measured TSH in an equimolar fashion, regardless of differences in glycosylation *(31)*. This report describes our next step (Phase IV), in which we attempt to show that our approach for recalibration may allow manufacturers to have more uniform RIs in the future. Note that the participating manufacturers who only recently joined our effort successfully went

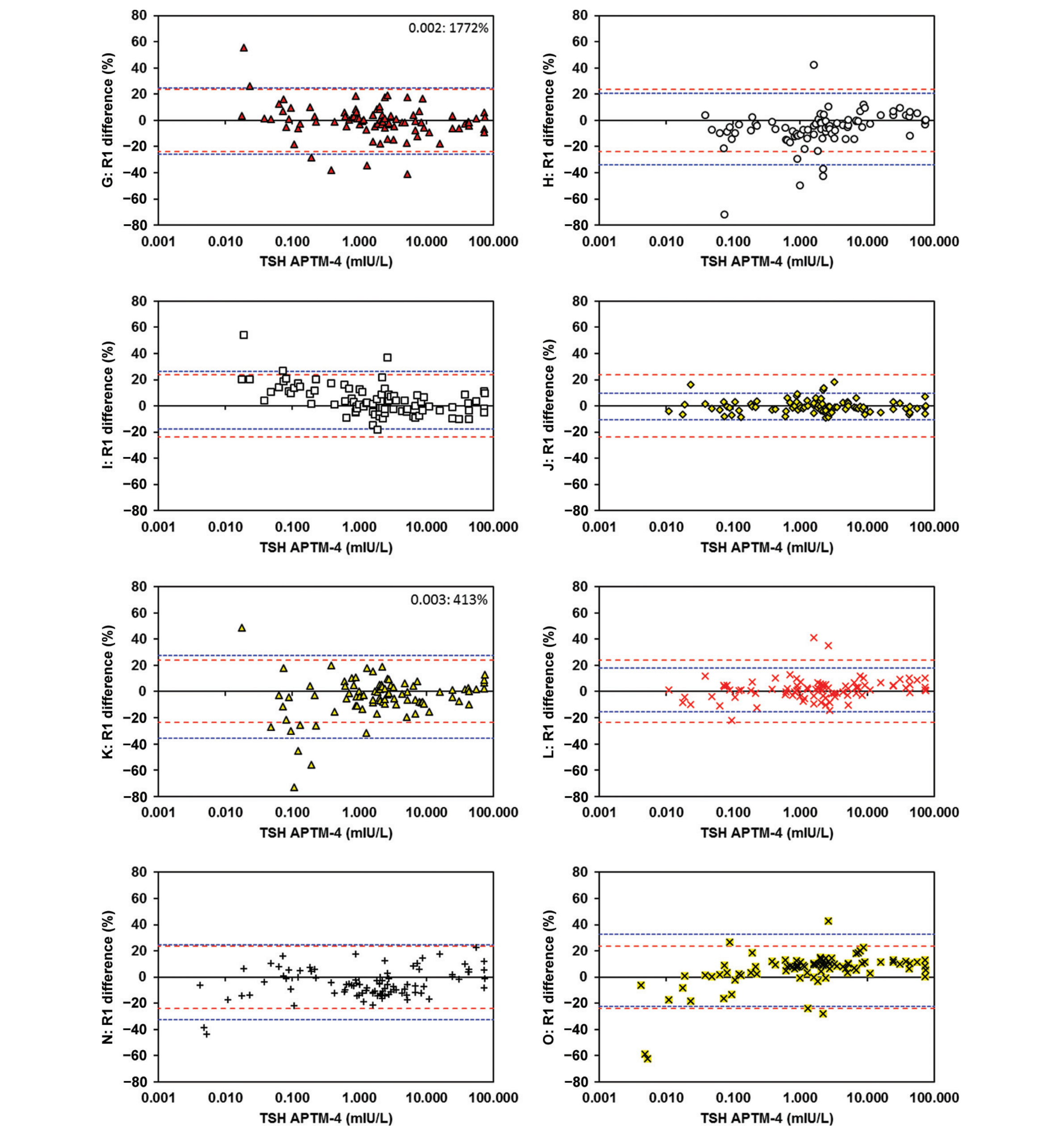

Fig. 4. Continued

through the "step-up" approach previously described *(32)*.

The panel of commutable samples used for recalibration in this round had fairly uniformly distributed concentrations within the typical measurement intervals. Eleven of the 14 assays had preharmonization median deviations within 10% from the APTM. The improved agreement after recalibration is shown by centering of the assays' differences (%) around zero difference from the APTM-4 targets, by the reduced differences (%) of the 15th and 85th centiles and the mean deviations (%) meeting the 7.8% bias specification with 95% confidence for 13 out of 14 assays. Another indicator of successful recalibration was the reduction of the CV (%) of the assay means for the harmonization panel from 9.5% to 4.2%, and for the RI panel from 11.9% to 4.8%.

Because a minimum of 6 months passed between recalibration of the assays and testing of the RI samples,

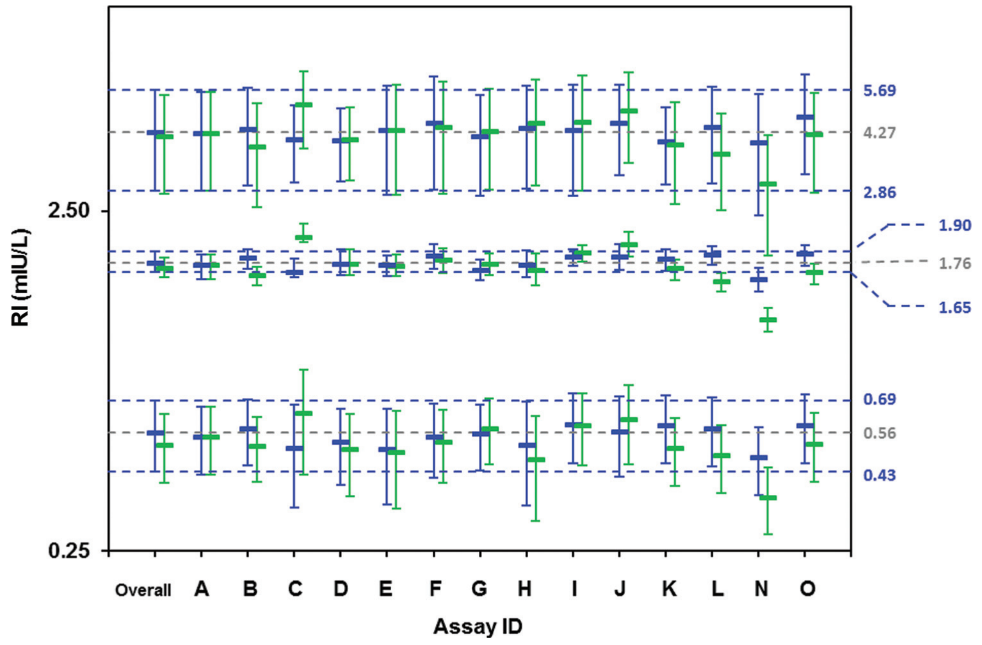

Fig. 5. Comparison of the pre- and postrecalibration RIs of the individual immunoassays to the overall RI (n = 120).
The pre- and postrecalibration RI characteristics are shown in green and blue, respectively; the thick horizontal bars for each assay stand for the 2.5th, 50th, and 97.5th percentiles, while the thin vertical lines represent the 90% CIs of the respective percentiles. The grey and blue broken horizontal lines stand for the post-recalibration 2.5th, 50th, and 97.5th reference percentiles and their respective 90% CIs.

several manufacturers assayed the latter using different reagent lots (12 of 14), different calibrator lots *(10)*, or different instruments *(8)*. This may have contributed to the observed differences of the individual RI percentiles from the reference ones.

We believe that this study provides evidence that harmonization may enable manufacturers to achieve more uniform RIs in the near future. However, we wish to emphasize that the RI presented in this report cannot be seen as the endpoint. It is important that all involved stakeholders understand that uniform RIs do not indicate a "one-size-fits-all RI." Reference intervals may be impacted by factors such as age, ethnicity, iodine intake, etc. IVD manufacturers will need to verify their individual RIs for TSH in accordance with accepted consensus standards, such as those from the IFCC, the National Academy of Clinical Biochemistry and CLSI *(33–35)*.

It will also be important that the traceability anchor achieved through this study is sustained by providing follow-up panels with traceability to the very first harmonization panel. We already have made an important step in this direction by ensuring the perfect link between the first follow-up and harmonization panel (through the target setting of both panels in parallel). For the future, we intend to always develop a new panel before depletion of the previous one, and measure both in overlap. Whether the 4 assays selected here will do the future target setting, will depend on their long-term stability. We will assess this by our Percentiler application described elsewhere *(36)*. Also, collaboration with proficiency testing organizers using commutable samples will be important to provide surveillance of the continuing relationship among different assays.

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

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

Employment or Leadership: None declared.
Consultant or Advisory Role: A. Hishinuma, Corresponding member of IFCC Committee for Standardization of Thyroid Function Tests.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: Abbott Diagnostics, USA; Beckman Coulter Inc., USA; bioMérieux SA, France; DiaSorin S.p.A, Italy; Fujirebio Inc.,

Japan; LSI Medience Corporation, Japan; Ortho-Clinical Diagnostics, UK; Roche Diagnostics GmbH, Germany; Shenzhen Mindray Bio-Medical Electronics Co., Ltd., China; Snibe Co., Ltd., China; Sichuan Maccura Biotechnology Co., Ltd., China; Siemens Healthineers, USA; Sysmex Corporation, Japan; and TOSOH Corp., Japan, sponsored (all contributed equally) the study in terms of sample procurement and funding for the statistical target setting. The Center for Statistics of Ghent University provided support under the FIRE Statistical Consulting program.
Expert Testimony: None declared.
Patents: None declared.

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

Acknowledgments: The Chair of the IFCC C-STFT (L.M. Thienpont) is grateful to (companies in alphabetical order; individual names also): D. Flanagan J. Reid and S. Ruetten (Abbott Diagnostics, USA); A. Adelmann, J. Budd and J. Sackrison (Beckman Coulter Inc., USA); J-M. Barbeaud (bioMérieux SA, France); I. Kutschera and G. Markowitz (DiaSorin S.p.A, Italy); K. Aoyagi, C. Hall and T. Niwa (Fujirebio Inc., Japan); S. Tashiro, and T. Ono (LSI Medience Corporation, Japan); P. Hosimer, M.-M. Patru and C. Thomas (Ortho-Clinical Diagnostics, UK); A. Hoppe and M. Rottmann (Roche Diagnostics GmbH, Germany); shixin, like and Rtliu (Shenzhen Mindray Bio-Medical Electronics Co., Ltd., China); Z.D. Chen, W. Li, H. Xu and J.Y. Yuan (Snibe Co., Ltd., China); Y. De Qian, Y. Tao, and L. Wan Ju (Sichuan Maccura Biotechnology Co., Ltd., China); R. Janzen, P. Sibley, R. Payne and V. Bitcon (Siemens Healthineers, USA); T. Sakata, M. Yamasaki, T. Kagawa, and K. Kishi (Sysmex Corporation, Japan); M. Kasai, S. Marivoet, S. Narayanan, H. Tsukamoto and M. Tsuura (TOSOH Corp., Japan), acting as representative on behalf of their organizations. Their efforts to review and provide comments on the manuscript are highly appreciated. The 14 organizations sponsored (all contributed equally) the study in terms of sample procurement and funding for the statistical target setting.

The C-STFT and coauthors are also grateful to the Center for Statistics of Ghent University for the support given under the FIRE Statistical Consulting program.

References

1. Thyroid Disease Manager. Guidelines for diagnosis and management of thyroid disease. http://www.thyroidmanager.org (Accessed February 2017).
2. Thienpont LM, Van Uytfanghe K, Poppe K, Velkeniers B. Determination of free thyroid hormones. Best Pract Res Clin Endocrinol Metab 2013;27:689–700.
3. Bahn Chair RS, Burch HB, Cooper DS, Garber JR, Greenlee MC, Klein I, et al. Hyperthyroidism and other causes of thyrotoxicosis: management guidelines of the American Thyroid Association and American Association of Clinical Endocrinologists. Thyroid 2011;21:593–646.
4. Garber JR, Cobin RH, Gharib H, Hennessey JV, Klein I, Mechanick JI, et al. American Association of Clinical Endocrinologists and American Thyroid Association Taskforce on Hypothyroidism in Adults. Clinical practice guidelines for hypothyroidism in adults: co-sponsored by American Association of Clinical Endocrinologists and the American Thyroid Association. Endocr Pract 2012;6:988–1028.
5. Pearce SH, Brabant G, Duntas LH, Monzani F, Peeters RP, Razvi S, Wemeau JL. 2013 ETA Guideline: management of subclinical hypothyroidism. Eur Thyroid J 2013;2:215–28.
6. Biondi B, Bartalena L, Cooper DS, Hegedüs L, Laurberg P, Kahaly GJ. The 2015 European Thyroid Association Guidelines on diagnosis and treatment of endogenous subclinical hyperthyroidism. Eur Thyroid J 2015;4:149–63.
7. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid 2016;26:1–133.
8. Ozarda Y. Reference intervals: current status, recent developments and future considerations. Biochem Med 2016;26:5–16.
9. Miller WG, Horowitz GL, Ceriotti F, Fleming JK, Greenberg N, Katayev A, et al. Reference intervals: strengths, weaknesses, and challenges. Clin Chem 2016;62:916–23.
10. Jones GR, Barker A, Tate J, Lim CF, Robertson K. The case for common reference intervals. Clin Biochem Rev 2004;25:99–104.
11. Beckett G, MacKenzie F. Thyroid guidelines–are thyroid-stimulating hormone assays fit for purpose? Ann Clin Biochem 2007;44:203–8.
12. Vesper HW, Thienpont LM. Traceability in laboratory medicine. Clin Chem 2009;55:1067–75.
13. Ceriotti F. Prerequisites for use of common reference intervals. Clin Biochem Rev 2007;28:115–21.
14. Panteghini M, Ceriotti F. Obtaining reference intervals traceable to reference measurement systems: is it possible, who is responsible, what is the strategy? Clin Chem Lab Med 2011;50:813–7.
15. IFCC. Committee for Standardization of Thyroid Function Tests (C-STFT). http://www.ifcc.org/ifcc-scientific-division/sd-committees/c-stft/ (Accessed February 2017).
16. International Organization for Standardization (ISO). In vitro diagnostic medical devices–measurement of quantities in biological samples–metrological traceability of values assigned to calibrators and control materials. Geneva: ISO; 2003. ISO 17511:2003.
17. Miller GW, Myers GL, Lou Gantzer M, Kahn SE, Schönbrunner ER, Thienpont LM, et al. Roadmap for harmonization of clinical laboratory measurement procedures. Clin Chem 2011;57:1108–17.
18. Thienpont LM, Van Houcke SK. Traceability to a common standard for protein measurements by immunoassay for in-vitro diagnostic purposes. Clin Chim Acta 2010;411:2058–61.
19. Van Houcke SK, Thienpont LM. "Good samples make good assays"–the problem of sourcing clinical samples for a standardization project. Clin Chem Lab Med 2013;51:967–72.
20. Christenson RH, Duh SH, Apple FS, Bodor GS, Bunk DM, Panteghini M, et al. Towards standardization of cardiac troponin I measurements part II: assessing commutability of candidate reference materials and harmonization of cardiac troponin I assays. Clin Chem 2006;52:1685–92.
21. Boulo S, Hanisch K, Bidlingmaier M, Arsene CG, Panteghini M, Auclair G, et al. Gaps in the traceability chain of human growth hormone measurements. Clin Chem 2013;59:1074–82.
22. Stöckl D, Van Uytfanghe K, Van Aelst S, Thienpont LM. A statistical basis for harmonization of thyroid stimulating hormone immunoassays using a robust factor analysis model. Clin Chem Lab Med 2014;52:965–72.
23. Van Houcke SK, Van Aelst S, Van Uytfanghe K, Thienpont LM. Harmonization of immunoassays to the all-procedure trimmed mean–proof of concept by use of data from the insulin standardization project. Clin Chem Lab Med 2013;51:e103–5.
24. CLSI. Evaluation of detection capability for clinical laboratory measurement procedures: approved guideline, second edition. Wayne, (PA): CLSI: 2012. CLSI document EP17–A2.
25. Linnet K. Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. Clin Chem 2000;46:867–9.
26. Westgard QC. Desirable biological variation database specifications. Desirable specifications for total error, imprecision, and bias, derived from intra- and inter-individual biologic variation. https://www.westgard.com/biodatabase1.htm (Accessed February 2017).
27. Commission of the European Communities. The certification of progesterone in two lyophilized serum materials, CRM 347 and CRM 348. Commission of the European Communities: Brussels–Luxembourg; 1989. Report EUR 12282 EN.
28. Stöckl D, Rodríguez Cabaleiro D, Van Uytfanghe K, Thienpont LM. Interpreting method comparison studies by use of the Bland-Altman plot: reflecting the importance of sample size by incorporating confidence limits and predefined error limits in the graphic. Clin Chem 2004;50:2216–8.
29. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri T, Miller WG, et al. Report of the IFCC Working Group for Standardization of Thyroid Function Tests, part 1: Thyroid-Stimulating Hormone. Clin Chem 2010;56:902–11.
30. Thienpont LM, Van Uytfanghe K, Van Houcke S. Standardization activities in the field of thyroid function tests: a status report. Clin Chem Lab Med 2010;48:1577–83.
31. Thienpont LM, Van Uytfanghe K, Van Houcke S, Das B, Faix JD, MacKenzie F, et al. A Progress report of the IFCC Committee for Standardization of Thyroid Function Tests. Eur Thyroid J 2014;3:109–16.
32. Van Uytfanghe K, De Grande LA, Thienpont LM. A "Step-Up" approach for harmonization. Clin Chim Acta 2014;432:62–7.
33. Solberg HE. International Federation of Clinical

Chemistry (IFCC), Scientific Committee, Clinical Section, Expert Panel on Theory of Reference Values, and International Committee for Standardization in Haematology (ICSH), Standing Committee on Reference Values. Approved Recommendation (1986) on the theory of reference values. Part 1. The concept of reference values. J Clin Chem Clin Biochem 1987; 25:337–42.

34. Baloch Z, Carayon P, Conte-Devoix B, Demers LM, Feldt-Rasmussen U, Henry JF, et al. Laboratory support for the diagnosis and monitoring of thyroid disease. Thyroid 2003;13:3–126.

35. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory: approved guideline, third ed. Wayne, PA: CLSI; 2008. CLSI document EP28–A3c.

36. De Grande LAC, Goossens K, Van Uytfanghe K, Das B, MacKenzie F, Patru MM, et al. Monitoring the stability of the standardization status of FT4 and TSH assays by use of daily outpatient medians and flagging frequencies. Clin Chim Acta 2017;467:8–14.