Clinical Chemistry 63:10
1642–1652 (2017)

Standardization of Free Thyroxine Measurements Allows the Adoption of a More Uniform Reference Interval

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

BACKGROUND: The IFCC Committee for Standardization of Thyroid Function Tests intended to standardize free thyroxine (FT_4) immunoassays. We developed a Système International d'Unités traceable conventional reference measurement procedure (RMP) based on equilibrium dialysis and mass spectrometry. We describe here the latest studies intended to recalibrate against the RMP and supply a proof of concept, which should allow continued standardization efforts.

METHODS: We used the RMP to target the standardization and reference interval (RI) panels, which were also measured by 13 manufacturers. We validated the suitability of the recalibrated results to meet specifications for bias (3.3%) and total error (8.0%) determined from biological variation. However, because these specifications were stringent, we expanded them to 10% and 13%, respectively. The results for the RI panel were reported as if the assays were recalibrated. We estimated all but 1 RI using parametric statistical procedures and hypothesized that the RI determined by the RMP was suitable for use by the recalibrated assays.

RESULTS: Twelve of 13 recalibrated assays had a bias, meeting the 10% specification with 95% confidence; for 7 assays, this applied even for the 3.3% specification. Only 1 assay met the 13% total error specification. Recalibration reduced the CV of the assay means for the standardization panel from 13% to 5%. The proof-of-concept study confirmed our hypothesis regarding the RI but within constraints.

CONCLUSIONS: Recalibration to the RMP significantly reduced the FT_4 immunoassays' bias, so that the RI determined by the RMP was suitable for common use within a margin of 12.5%.

The diagnosis of metabolic thyroid disorders and/or monitoring of treatment is based on laboratory testing of serum thyroid-stimulating hormone (TSH)[16] and free thyroxine (FT_4). Provided the hypothalamic-pituitary-thyroid axis is intact, a first-line TSH result may suggest a number of thyroid disorders that could be clarified by follow-up measurement of FT_4; however, immediate combined measurement is indicated for the differential diagnosis between mild (subclinical) primary hyperthyroidism and secondary (central) hypothyroidism. Furthermore, combined measurement is warranted during the first days/weeks of the follow-up of patients with severe thyroid dysfunction, when TSH has not yet returned to a euthyroid baseline concentration and thus is not representative of the actual thyroid functional status (e.g., in patients with autoimmune Graves' disease and

[1] Department of Pharmaceutical Analysis, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium; [2] Ref4U, Laboratory of Toxicology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium; [3] Department of Applied Mathematics, Computer Science and Statistics, Faculty of Sciences, Ghent University, Ghent, Belgium; [4] Biochemistry and Immunology Laboratory, Kokilaben Dhirubhai Ambani Hospital and Medical Research Institute, Mumbai, India; [5] Clinical Chemistry and Immunology, Montefiore Medical Center, and Department of Pathology, Albert Einstein School of Medicine, New York, NY; [6] Birmingham Quality/UK NEQAS, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK; [7] Department of Endocrinology, University Hospitals Leuven, Leuven, Belgium; [8] Department of Infection Control and Clinical Laboratory Medicine, Dokkyo Medical University, Tochigi, Japan; [9] Department of Endocrinology, Ghent University Hospital, Ghent, Belgium; [10] Laboratory of Endocrinology, Department of Laboratory Medicine, AZ Maria-Middelares Sint-Jozef, Campus Maria-Middelares, Ghent, Belgium; [11] Department of Endocrinology, OLV Ziekenhuis Aalst-Asse-Ninove, Aalst, Belgium; [12] Department of Endocrinology, General Hospital Sint Jan, Bruges, Belgium; [13] Department of Endocrinology, Universitair Ziekenhuis Brussel, Brussels, Belgium; [14] Department of Endocrinology, Royal Prince Alfred Hospital, Camperdown, Australia; [15] Thienpont & Stöckl Wissenschaftliches Consulting GbR, Rennertshofen (OT Bertoldsheim), Germany.
* Address correspondence to this author at: Thienpont & Stöckl Wissenschaftliches Consulting GbR, Erlbacher Strasse 11, D-86643 Rennertshofen (OT Bertoldsheim), Germany. E-mail linda.thienpont@ugent.be.

Received March 30, 2017; accepted June 13, 2017.
Previously published online at DOI: 10.1373/clinchem.2017.274407

[16] Nonstandard abbreviations: TSH, thyroid stimulating hormone; FT4, free thyroxine; RI, reference interval; IVD, in vitro diagnostic; RMP, reference measurement procedure; MC, method comparison; ED-ID, equilibrium dialysis-isotope dilution; CI, confidence interval; TE, total error; A-D, Anderson–Darling test.

high titers of TSH receptor antibodies or with increased human chorionic gonadotropin concentrations). On the other hand, FT_4 is the primary test for the titration of levothyroxine replacement in patients with central hypothyroidism and/or with high-risk differentiated thyroid cancer with need for a suppressed TSH *(1–5)*. For maximum effectiveness, current FT_4 immunoassays would benefit from improved clinical and analytical consistency *(6, 7)*. Additionally, the issue of substantial intermethod variability needs to be resolved for improved everyday patient care because it requires interpretation of laboratory results against assay-specific reference intervals (RIs) and prevents incorporation of common decision levels in evidence-based practice guidelines *(7, 8)*. Therefore, the IFCC Committee for Standardization of Thyroid Function Tests was commissioned to standardize FT_4 measurements globally *(9)*. The committee's efforts have been endorsed by the clinical community, which also called for general standardization of hormonal assays in the 21st century *(10)*.

The committee conducted the standardization activities of FT_4 measurements in partnership with the same in vitro diagnostic (IVD) manufacturers (with 1 exception) that had been involved in the TSH harmonization *(11)*. The committee pursued a process similar to that used for the TSH assays, except for FT_4 they developed and used a reference measurement system with traceability to the Système International d'Unités (SI) *(12, 13)*. The committee defined the measurand and developed/validated a conventional reference measurement procedure (RMP) based on equilibrium dialysis combined with isotope dilution LC-MS/MS (ED-ID-LC-MS/MS) *(14–16)*, and undertook several method comparisons (MCs) with single-donation and commutable serum samples (Phase I–III studies) according to the "step-up" approach *(8, 17–19)*. Each of the studies had a different focus, including documentation of the assays' intrinsic quality and demonstration of the feasibility of standardization of assay results by recalibrating the immunoassays to the RMP.

Here we report, on behalf of the Committee for Standardization of Thyroid Function Tests, our latest activities in the standardization process. We performed a Phase IV MC study between 13 immunoassays and the RMP. There were 2 objectives: first, to establish calibration traceability of the participating assays to the SI-traceable RMP; second, to validate the efficiency of the process to eliminate the assay-specific biases. Subsequently, we conducted a RI study with a new panel of samples to test the proof of concept that, after standardization, immunoassays might accord sufficiently with the RMP to enable adoption of a common RI for diagnosis and follow-up of patients with thyroid dysfunction.

Material and Methods

PANELS OF CLINICAL SAMPLES AND VALUE ASSIGNMENT

We collected standardization and RI panels. The standardization panel comprised 91 clinically relevant samples and was intended to facilitate the calibration adjustment/readjustment by the manufacturers to the IFCC RMP. The aim of the RI panel was to let manufacturers evaluate their recalibration, for which we used 120 samples donated by apparently healthy American volunteers. The sources, eligibility and exclusion criteria, conditions for sampling, processing, and storage were those described before for the TSH harmonization effort *(11)*. Approval from a bioethics committee and informed consent from the patients were obtained along with a short description of the clinical background of the donating patients. The target values (mean of minimum 3 independent measurements) were assigned with the IFCC conventional RMP performed at the reference laboratory of Ghent University. Both are listed in the Database of the Joint Committee for Traceability in Laboratory Medicine *(20)*.

STUDY PARTICIPANTS AND MEASUREMENT PROTOCOL

Thirteen IVD manufacturers participated in the current studies, each with 1 assay (coding and further details on the platforms/assays in Table 1). We requested that the IVD manufacturers perform all measurements according to a proposed randomized sequence, in singleton on each of 2 days, and include their master calibrators for measurement in parallel with the panel samples. The individual results were reported. The samples for the RI study were measured in order of their ascending identification number, in singleton and within a single run. Of note, the organization and interpretation of internal quality control was left to the discretion of each manufacturer.

RECALIBRATION OF IMMUNOASSAYS

After submitting the results for measurement of the standardization panel with the assays' current calibrators, the IVD manufacturers received from us a preliminary validation report, comprising the target concentrations determined by the RMP. These were intended for use in value reassignment of the master calibrators. The manufacturers were entitled to use their in-house mathematical procedure to determine the relationship of their assay results to those from the RMP *(11)*. After the readjustment of the master calibrators, the manufacturers recalculated and reported back the results for the standardization panel as if they were obtained with the recalibrated assays. The results for measurement of the RI panel were similarly reported after transformation to the revised calibration.

Table 1. Study participants (ordered by code), inclusive of the platforms/FT_4 assays examined for standardization.[a]

IVD manufacturer; *platform/immunoassay*	Code	Reference interval (pmol/L)	Measurement interval (pmol/L)[d–h]
Siemens Healthineers (Tarrytown, NY); *Advia Centaur XP*	A	11.5–22.7 (n = 388)	1.3–155[d]
Abbott Diagnostics (Abbott Park, IL); *Architect i2000*	B	9.0–19.1 (99%, n = 411)	5.2–77[e]
Ortho-Clinical Diagnostics (Buckinghamshire, UK); *Vitros ECi*	D	10.0–28.2 (98%, n = 535)	0.9–90[d]
bioMérieux SA (Marcy-l'Etoile, France); *Vidas*	E	10.6–19.4 (95%, n = 623)	1.1–100[f]
Beckman Coulter Inc. (Brea, CA); *Access 2*	F	7.9–14.4 (95%, n = 316)	3.2–77[f]
DiaSorin S.p.A (Saluggia, Italy); *Liaison® Analyser*	G	10.3–21.9 (95%, n = 517)	1.3–129[d]
Sichuan Maccura Biotechnology Co., Ltd. (Chengdu, China)[b]; *IS1200*	H	12.2–21.2 (95%, n = 175)	2.0–100[f]
Roche Diagnostics GmbH (Mannheim, Germany); *Elecsys (Cobas e 601)*	I	12.0–22.0 (95%, n = 801)	3.0–100[g]
Tosoh Corporation (Tokyo, Japan); *AIA-2000*	J	10.6–21.0 (95%, n = 618)	1.3–103[f]
Snibe Co., Ltd., (Shenzhen, China)[b]; *Maglumi 2000*	K	11.5–22.1 (95%)	1.3–154[h]
Fujirebio Inc. (Tokyo, Japan)[b]; *Lumipulse G1200*	L	9.7–19.8 (95%, n = 141)	1.0–129[d]
LSI Medience Corporation (Tokyo, Japan)[c]; *STACIA*	N	12.5–26.5	1.3–103[f]
Sysmex Corporation (Kobe, Japan)[c]; *HISCL-5000*	O	9.9–20.5	3.2–77[f]

[a] The listed reference and measurement intervals are those stated in the kit inserts.
[b,c] Manufacturers who only joined in 2015[b] and/or 2016[c] for participation in the Phase IV method comparison study.
[d–h] The lower limit of the measurement intervals is limit of detection (according to the CLSI's EP-17 protocol)[d]; functional sensitivity (CV 10%)[e]; functional sensitivity (CV 20%)[f]; limit of quantification at a TE of ±30% (CLSI EP-17)[g]; limit of quantification (CLSI EP-17) (21)[h].

DATA TREATMENT

For consolidation of the MC study data, we used Microsoft EXCEL® 2010. We concentrated on demonstrating and validating the efficiency of the recalibration process. We calculated for each assay (*a*) the pre- and post-recalibration median deviation (%) to the RMP in several FT_4 concentration intervals, (*b*) the mean deviation (%) or bias [and 1-sided 95% confidence interval (CI)] after recalibration, (*c*) the total error (TE, %) from the first replicate after recalibration, and (*d*) the differences between the replicates (in % of the mean). We also compared the pre- and post-recalibration CVs (%) of the assay means.

We used CBstat (version 5.1) for statistical evaluation of the data from the RI study. This software evaluated normality of data distributions by the Anderson–Darling (A-D) test ($P \geq 0.05$), did outlier testing on the basis of power-transformed values (limit 4 SD), and supplied parametric (direct on the original data and/or after transformation) and nonparametric procedures to estimate the RI characteristics. For the normally distributed data sets, we used the direct parametric procedure [RI estimated as mean $\pm$ $1.96(1/\{1 - 1/[4(n - 1)]\}) \times SD$]. For those data sets for which normality did not apply, we selected the procedure after a sequence of investigations, i.e., in addition to the detection of statistical outliers, we did a visual screening for aberrant differences (%) to the RMP targets. If after omission of the detected values the A-D test allowed acceptance of the hypothesis of normally distributed data, we again selected the direct parametric procedure; if not, we verified the data for normality after log-transformation. If the A-D *P* value was then $\geq$0.05, we applied the parametric procedure. Finally, 1 data set remained, which was submitted to the nonparametric bootstrap (500 replicates) procedure to generate bootstrap estimates of the (2.5/100)n + 0.5 and (97.5/100)n + 0.5 ordered values *(22)*. To test the hypothesis that after recalibration a common RI could be used by all manufacturers, we first investigated whether the probabilities that the 2.5 and 97.5 percentiles (further also referred to as lower and upper limits, respectively), estimated from the data sets of the immunoassays, were located within the 90% CI from the RMP data percentiles (further referred to as reference percentiles) and were reasonably large (>90%). We repeated the probability testing while using limits of 12.5% around the reference percentiles. Probability estimations were done in R 3.2.3 for all assays but assay K (Table 1), for which the CIs were determined by CBstat; for the latter, we used the R statistical software to perform a bootstrap procedure on the original RI data set to simulate the distribution of the percentiles.

ANALYTICAL SPECIFICATIONS

We demonstrated/validated the suitability of the recalibrated results to meet desirable specifications for bias and TE based on the biological variation, i.e., 3.3% and 8.0%, respectively *(23)*. However, because of the ex-

treme stringency of these values, we also used the empirical bias limit of 10% that was considered state of the art in previous MC studies, and expanded the TE specification to 13% to account for any imprecision of the RMP *(8, 16, 18, 19)*. The 12.5% limit used for testing the RI hypothesis was based on the state-of-the-art bias specification used above but would additionally account for the uncertainty of the location of the reference percentiles.

HOMOGENEITY AND STABILITY STUDY

We assessed the homogeneity and stability of the FT_4 standardization panel in the same way as described for TSH *(11)*.

Results

CONCENTRATION RANGE COVERED BY THE PANELS OF CLINICAL SAMPLES

The FT_4 standardization panel covered a concentration range from 4.5 pmol/L to 164 pmol/L (determined by the RMP). The expanded uncertainty of the targets (coverage factor k = 2) was estimated to be on the order of 7.0% *(16)*. The central 95% of the RI panel covered the range from 13.5 pmol/L (±0.7 pmol/L; 90% CI) to 24.3 pmol/L (±0.7 pmol/L) with the mean at 18.9 pmol/L.

VALIDATION OF THE EFFICIENCY OF RECALIBRATION

The combined difference plots (Fig. 1) reflect the assays' calibration biases to the RMP before (Fig. 1A) and after (Fig. 1B) recalibration. The effect of recalibration on the assay-specific median deviations (%) to the RMP targets in 4 concentration intervals is shown in Fig. 2A by a combined picture with indication of the 15th, 50th, and 85th centiles, and in Fig. 2B by the individual deviations (see Table 1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol63/issue10 for more details). Before recalibration, deviations were negative across the FT_4 measurement range for all but assay N (<10 pmol/L). Moreover, the deviations increased with increasing concentration. The highest median manufacturer deviations were −40.8% (assay J) (<10 pmol/L), −37.9% (assay F) (≥10 and <25 pmol/L), −57.7% (assay B) (≥25 and <100 pmol/L), and −72.7% (assay B) (≥100 pmol/L). The lowest median manufacturer deviations were 7.4% (assay N), −13.7% (assay N), −25.6% (assay O), and −30.2% (assay G). Hence, the most discrepant assay pairs (assays J/N, F/N, B/O, and B/G) deviated by 48.2%, 24.2%, 32.1%, and 42.5%, respectively. After recalibration, the ranges of the median deviations became −12.0% (assay O) to +8.2% (assay A) (<10 pmol/L), −8.9% (assay O) to +1.7% (assay H) (≥10 and <25 pmol/L), −8.4% (assay H) to +9.5% (assay F) (≥25 and <100 pmol/L), and −12.5% (assay O) to +11.9% (assay G) (≥100 pmol/L). Fig. 3 shows the post-recalibration differences (%) and the assay biases (%) reflected against the used specifications. From the numbers in Table 2 in the online Data Supplement, we can confidently assert that after recalibration the bias (and 1-sided 95% CI) of all assays but assay O complied with the empirical specification of 10% at a 95% probability; the bias of 7 assays (A, B, D, E, I, J, and N) complied when assessed against the 3.3% specification (see Table 2 in the online Data Supplement) *(24)*. With regard to the assays' TE after recalibration, only assay I met the expanded specification, i.e., had 95% of its differences within 13%, whereas for the other assays, 8% to 35% of the differences violated it (see Fig. 1 in the online Data Supplement). The median differences between the replicates from 2 runs ranged from −1.5% (assay K) to 4.1% (assay F), and the SD_{diff} ranged from 2.5% (assay H) to 5.9% (assay A) (see Table 3 in the online Data Supplement). Fig. 2 in the online Data Supplement shows that for several assays the differences (%) between replicates were concentration-dependent. After recalibration, the CV of the assay means (the latter calculated for each assay from all results) decreased from 13% to 5%.

RI STUDY

The RI characteristics from the ED-ID-LC-MS/MS measurements were obtained with the direct parametric procedure. This procedure was also used for the other normally distributed data sets, which excluded the assays A, G, H, and K. Despite a negative outlier test in CBstat for these 4 data sets, visual inspection of the plots of assays G and H (see Fig. 3 in the online Data Supplement) revealed aberrant differences (%) to the RMP targets (4 for assay G and 3 for assay H, respectively). After omission of these aberrant data, the A-D *P* values became >0.26 and >0.25, respectively, which justified application of the direct parametric procedure to these assays. For the assay A, the hypothesis of normality was accepted after log-transformation of the data, again justifying the use of the parametric procedure; only for assay K did we have to use a nonparametric bootstrap procedure. Table 4 in the online Data Supplement lists the main characteristics of the respective RIs. The widths of the RIs by the immunoassays ranged from 9.4 pmol/L to 12.0 pmol/L vs 10.7 pmol/L for the RMP. The CIs for the respective percentiles ranged from 1.1 pmol/L to 2.4 pmol/L (at the 2.5 percentile) and 1.2 pmol/L to 2.4 pmol/L (at the 97.5 percentile) vs 1.4 pmol/L (for both percentiles of the RMP). The range of the means/medians of the RIs was from 17.2/17.0 pmol/L to 20.8/20.5 pmol/L vs 18.9/18.8 pmol/L for the RMP. Tables 5 and 6, plus Figs. 4 and 5 (all in the online Data Supplement), demonstrate that none of the calculated probabilities for the assays met the minimum requirement of >90%. However, after expanding the reference percentile intervals to 12.5%, they did for assays E, F, G, H, I, J, L, and N at the 2.5 percentile. For the 97.5 percentile, the >90% requirement was achieved by all

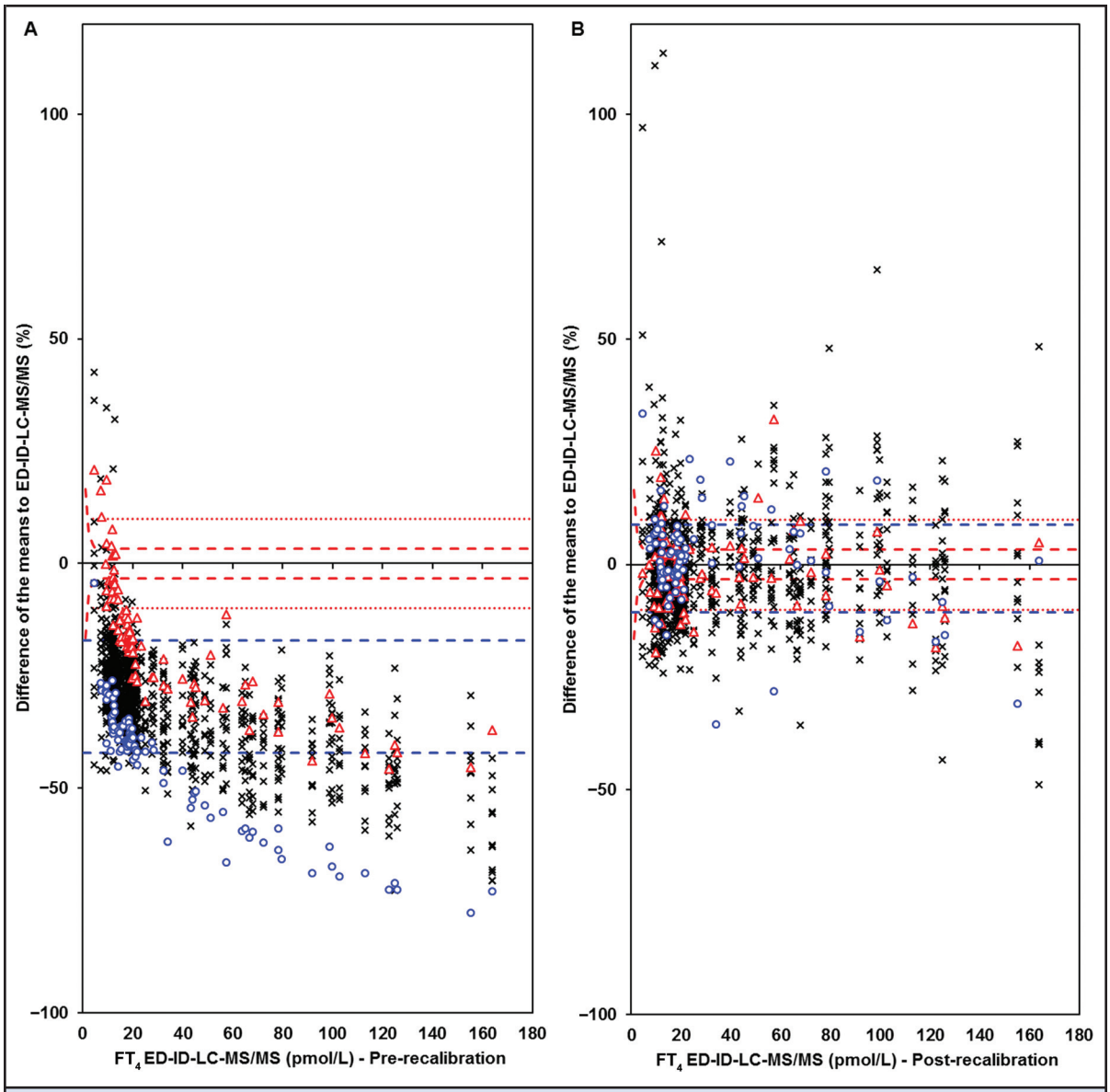

Fig. 1. Combined difference (%) plots of the immunoassay results to those by ED-ID-LC-MS/MS, before (A) and after (B) recalibration.

The most discrepant assays before recalibration are highlighted by colored symbols [blue circles for assay B (<25 pmol/L) and assay F (>25 pmol/L); red triangles for assay N)], whereas all other assays are indicated with the symbol X. The red broken lines are the bias limits based on the biological variation concept: ±3.3% (note that we converted the percentage limit to 0.165 pmol/L for concentrations ≤5 pmol/L), whereas the red dotted lines are the empirical bias limits of 10% *(8, 18, 19)*. The blue broken lines represent the 15th and 85th percentiles.

but assay A. The graphical overview of the respective RIs (Fig. 4) shows that assays A and B had the most discrepant 2.5 percentiles (calculated to the mean of both percentile values, they were 28% apart), while this was the case for assays A and F for the 97.5 percentiles (21% apart).

HOMOGENEITY STUDY

Statistical testing confirmed that the hypothesis of homogeneity of the aliquots in the standardization panel ($P >$ 0.05; see Table 7 in the online Data Supplement) could be accepted. The stability study is still ongoing.

Discussion

The approach to the standardization of commercial FT_4 immunoassays was similar to that previously described for TSH *(11)*. The Phase I MC demonstrated that mathematical recalibration of measurement results for samples

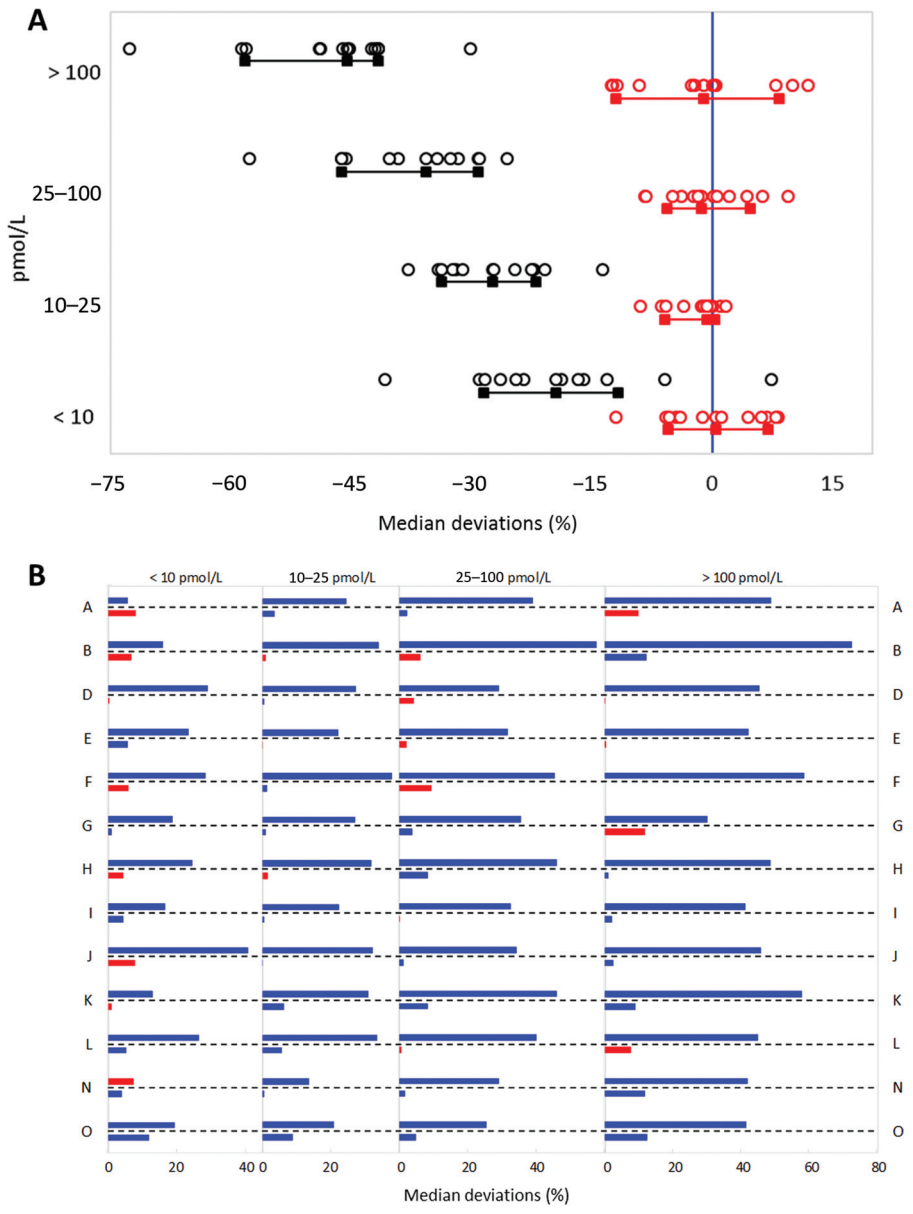

Fig. 2. Median deviations (%) of the immunoassays to ED-ID-LC-MS/MS before and after recalibration in 4 concentration intervals: <10 pmol/L, 10–25 pmol/L, 25–100 pmol/L, and >100 pmol/L.

The overall improvement in terms of the median deviations (%) by recalibration (A). 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 the 15th, 50th, and 85th centiles. The median deviations (%) of each assay by a pair of bars; the upper and lower bar shows the median deviation before and after recalibration, respectively (B). Note that the bars represent the unsigned magnitudes, whereas the colors refer to the signs (blue: negative, red: positive).

from presumably healthy volunteers could align the different immunoassays to the RMP. The Phase II and III MCs extended the findings for euthyroid individuals to patients with hypothyroidism and hyperthyroidism, and provided proof of concept that manufacturers could also do the recalibration by adjusting their calibrators *(8, 17–19)*. The current Phase IV MC was the natural next step in our standardization project, and the RI study was in-

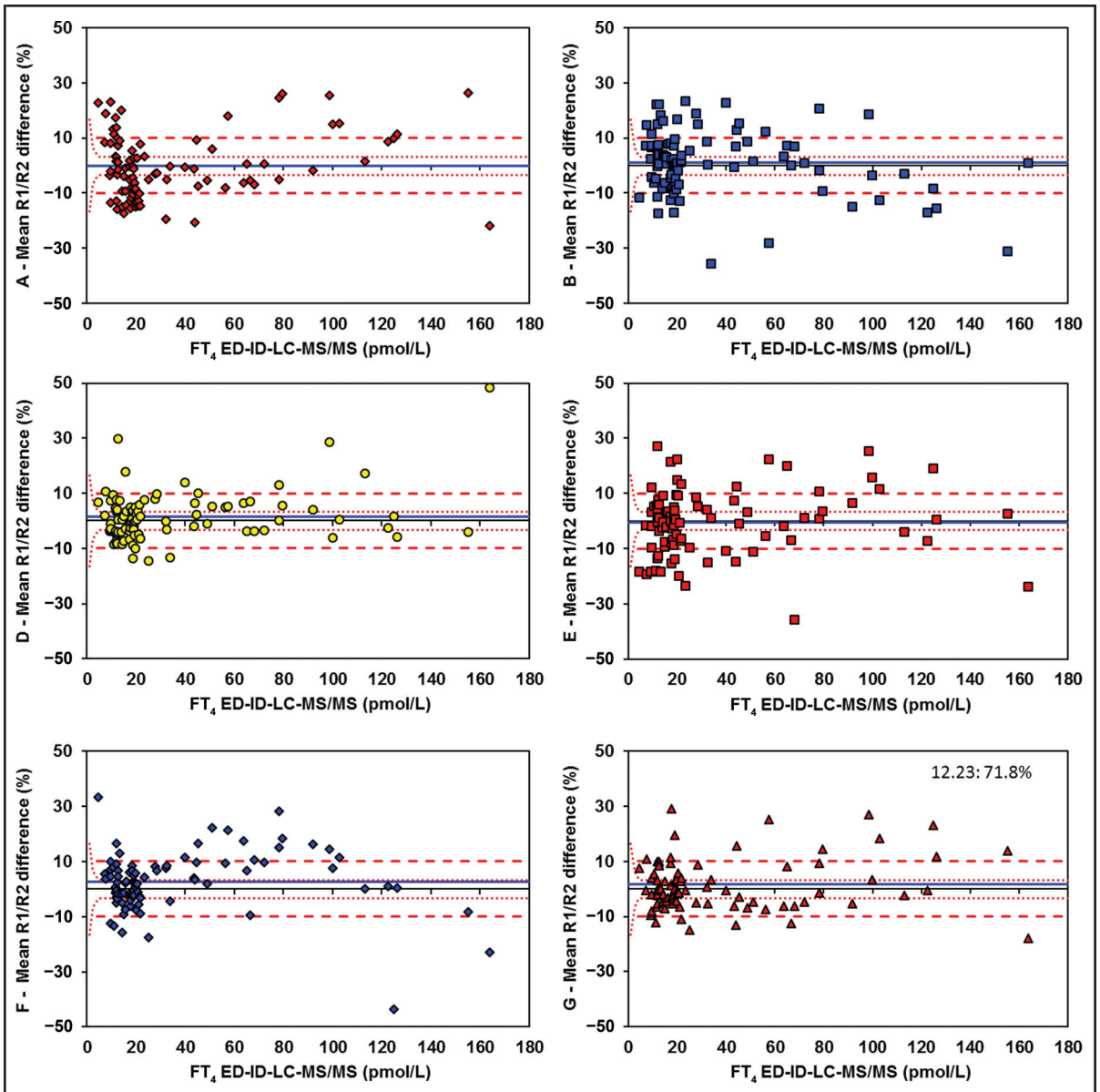

Fig. 3. Difference (%) plots after recalibration of the individual immunoassays.

The red dotted lines are the 3.3% bias limits from the biological variation concept (converted to 0.165 pmol/L for concentrations ≤5 pmol/L), whereas the red broken lines stand for the previously used empirical limits of 10% *(8, 18, 19)*. The blue line represents for each immunoassay the mean deviation or bias (%). The 1-sided 95% CIs given in Table 2 in the online Data Supplement are not shown because of too little graphical resolution. To keep the *y* axes identical in all plots, certain % differences required omission (concentrations and % differences mentioned in the plots).

Continued on page 1649

tended to assess whether recalibration would allow a uniform basis for the use of common RIs. The strengths of the FT_4 standardization approach were the involvement during several years of the globally operating IVD industry and the use of a panel of commutable samples, collected to mimic clinical conditions. The concentrations of the samples spanned the measurement range of current assays because they were sourced from euthyroid individuals and also from patients with overt hypothyroidism and hyperthyroidism.

The current study confirmed that establishing calibration traceability to the RMP significantly reduced the

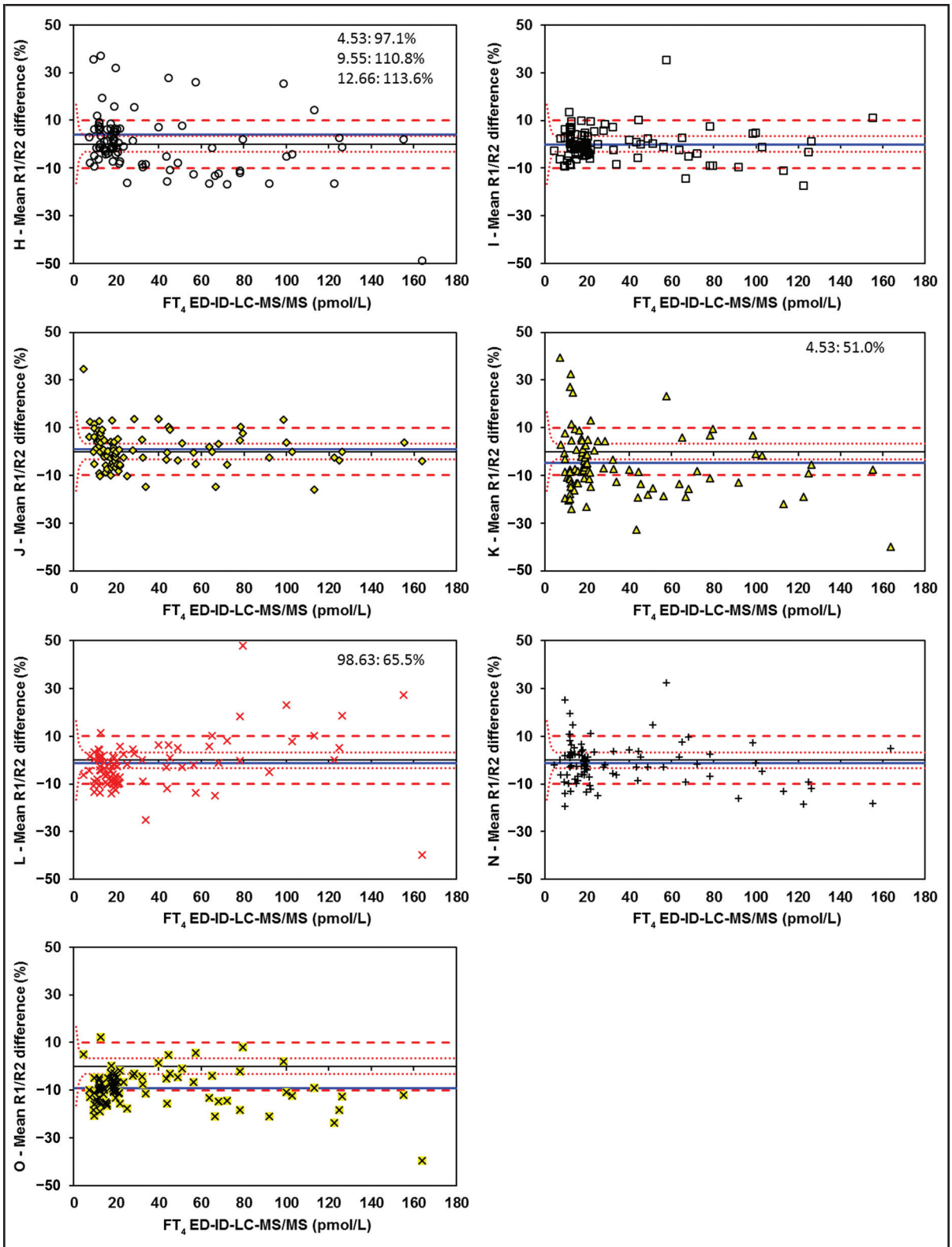

Fig. 3. Continued.

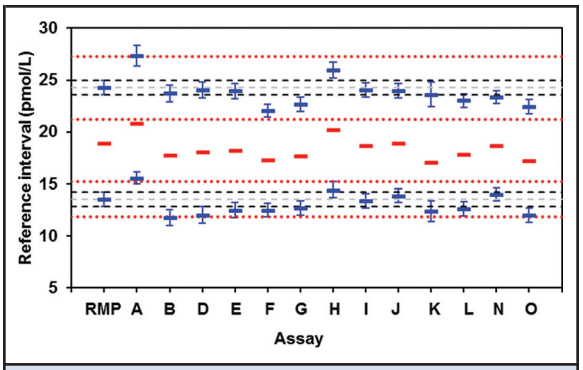

Fig. 4. Comparison of the RI percentiles of the individual immunoassays with those of ED-ID-LC-MS/MS (n = 120).

The blue thick horizontal bars represent the respective 2.5 percentiles and 97.5 percentiles of each RI, whereas the blue vertical lines show the respective 90% CIs. The red thick horizontal bars for each assay stand for the mean (except for assay K, for which it shows the median). The gray and black broken horizontal lines represent the reference percentiles (from the data by the RMP) and the 90% CIs around them, respectively. The red dotted lines are the 12.5% limits of the interval around the reference percentiles.

negative biases of the immunoassays, as well as the CV of the assay means. However, it is also important to appreciate the huge impact that standardization could have on future measurement results and RIs. After recalibration, 12 of 13 immunoassays had their bias (and CI) meet the empirical specification of 10% at a 95% probability, and 7 of them even passed the stringent specification of 3.3% derived from the biological variation. Although this outcome is overall reasonable, it also points to the fact that the recalibration effectiveness was better for some assays than for others.

The fact that the standardization panel comprised sufficient native samples enabled us also to focus on the validation of the post-recalibration TE. This is an important performance attribute because it reflects the accuracy of an assay for measurement of the individual sample. Most assays violated the expanded TE limits despite reasonable recalibration. This might be because the specification was too stringent, even after expansion. However, considering that in the previous MC studies we already highlighted the TE issue of many FT_4 immunoassays because of their susceptibility to sample-related effects, it is more realistic to suggest that our current study confirms this limitation.

Finally, the results on the differences between replicates highlight the occasional high interrun imprecision and lack of robustness of calibration (see Table 3 and Fig. 2 in the online Data Supplement). The importance of continual improvement of these performance attributes across all assays was discussed with the IVD manufacturers.

The aim of the current RI study was primarily to supply a proof of concept that after recalibration the use of a common RI may be feasible. We used the RI estimated from the measurement data by ED-ID-LC-MS/MS as reference and assessed whether the recalibrated assays could share it. We inferred the percentiles and mean of the central 95% of all but 1 RI by a parametric procedure applied to either the original or log-transformed data. Interestingly, the width of the interval by the RMP corresponded reasonably with that calculated from the FT_4 biological variation, i.e., 10.7 pmol/L vs 9.6 pmol/L, as well as that estimated in another study using ED-ID-LC-MS/MS, i.e., 12.1 pmol/L *(23, 25)*. However, it was most important to compare the derived immunoassay percentiles of the RIs with those of the RMP. In the statistical approach used, an immunoassay would be qualified to share the RI of the RMP if the probability that its percentiles were located within the CI around the reference was higher than 90%. None of the assays met this criterion. However, when an interval of $\pm$12.5% was adopted, the probabilities of 8 assays met the $>$90% requirement at the 2.5 percentiles, and of all but 1 assay also at the 97.5 percentiles. We present 3 reasons to justify the hypothesis of testing with the 12.5% margin around the reference percentiles. First is the observation that the magnitudes of the CIs around the reference percentiles were $\leq$5%, thus similar to or narrower than the assays' effective biases in the euthyroid range after recalibration (range 0% up to 9%). Second, we refer to the impact of the lot-to-lot variation on the RI study, which was performed with a time offset of at least 6 months compared with the Phase IV MC. Third, we found it legitimate to account to a certain extent for the uncertainty of the location of the estimated reference percentiles because of the potential impact of an undetectable bias in the measurements with the RMP. Nevertheless, even if the current margin of 12.5% accommodates the current state-of-the-art measurements, we advocate that in the future it should be decreased, particularly because of the low biological variation of serum FT_4. We also recommended the IVD manufacturers of the assays that did not agree with the RMP to share its percentiles, despite adopting the 12.5% margin, to do root cause analysis.

In conclusion, the Phase IV MC study described here showed that, in general, the recalibration process could eliminate the considerable FT_4 calibration biases to the RMP. In addition, the basic RI study provided the proof of concept because the percentiles of the RMP applied for most of the recalibrated assays within a margin of 12.5%. Although this result represents substantial progress in standardization of FT_4 measurements, we recognize that it cannot be extrapolated to all clinical situa-

tions when FT_4 testing is indicated, particularly when binding proteins are abnormal. Therefore, to better understand more-subtle assay differences in other patient cohorts, such as pregnant females and patients with the nonthyroidal illness syndrome, we recommend that our approach serves as model for future studies. We also see surveillance of the sustainability of the recalibration basis as a final key component of our standardization approach. We propose that, after implementation of the recalibrated assays, the surveillance should be done under field conditions to account for the impact of variables like lot-to-lot changes and instrument instability. This could be done by using the Percentiler/Flagger applications described elsewhere as useful tools for continuous monitoring of the stability of performance/flagging frequency in laboratories grouped according to instrument/assay-specific peers *(26)*. Another tool could be the organization of proficiency testing or external quality assessment surveys with commutable samples *(27)*. We also recognize that we should expand the measurement capacity with the conventional RMP. Therefore, we are currently working on establishing a network of competent reference laboratories. Last, but not least, from the perspective that implementing the recalibrated FT_4 assays will have a huge impact on future measurement results and RIs, we are committed to gaining broad consensus on this step *(28)*.

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: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: Abbott Diagnostics, US; Beckman Coulter Inc., US; bioMérieux SA, France; DiaSorin S.p.A, Italy; Fujirebio Inc., Japan; LSI Medience Corporation, Japan; Ortho-Clinical Diagnostics, UK; Roche Diagnostics GmbH, Germany; Snibe Co., Ltd, China; Sichuan Maccura Biotechnology Co., Ltd., China; Siemens Healthineers, US; Sysmex Corporation, Japan; TOSOH Corp., Japan.
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, or 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, US); A. Adelmann and J. Sackrison (Beckman Coulter Inc., US); 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); Z.D. Chen, H. Xu, J.Y. Yuan, and W. Li (Snibe Co., Ltd, China); Y. Tao, L. Wan Ju, and Y. De Qian (Sichuan Maccura Biotechnology Co., Ltd., China); R. Janzen, P. Sibley, R. Payne, and V. Bitcon (Siemens Healthineers, US); 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 13 organizations sponsored (all contributed equally) the study in terms of sample procurement and funding for the assignment with the reference method values.

References

1. 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.
2. Ross DS, Burch HB, Cooper DS, Greenlee MC, Laurberg P, Maia AL, et al. 2016 American Thyroid Association guidelines for diagnosis and management of hyperthyroidism and other causes of thyrotoxicosis. Thyroid 2016;26:1343–421.
3. 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.
4. Koulouri O, Auldin MA, Agarwal R, Kieffer V, Robertson C, Falconer Smith J, et al. Diagnosis and treatment of hypothyroidism in TSH deficiency compared to primary thyroid disease: pituitary patients are at risk of underreplacement with levothyroxine. Clin Endocrinol (Oxf) 2011;74:744–9.
5. Demers LM, Spencer CA. Laboratory medicine practice guidelines: laboratory support for the diagnosis and monitoring of thyroid disease. Thyroid 2003;13:3–126.
6. Thyroid Disease Manager. Guidelines for diagnosis and management of thyroid disease. http://www.thyroidmanager.org/ (Accessed March 2017).
7. Thienpont LM, Van Uytfanghe K, Poppe K, Velkeniers B. Determination of free thyroid hormones. Best Pract Res Clin Endocrinol Metab 2013;27:689–700.
8. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri T, Miller WG, et al.; for the IFCC Working Group on Standardization of Thyroid Function Tests. Report of the IFCC Working Group for Standardization of Thyroid Function Tests, part 2: free thyroxine and free triiodothyronine. Clin Chem 2010;56:912–20.
9. Committee for Standardization of Thyroid Function Tests (C-STFT). IFCC–Scientific Division (SD). SD Committees. http://www.ifcc.org/ifcc-scientific-division/sd-committees/c-stft/ (Accessed March 2017).
10. Wartofsky L, Handelsman DJ. Standardization of hormonal assays for the 21st century. J Clin Endocrinol Metab 2010;95:5141–3.
11. Thienpont LM, Van Uytfanghe K, De Grande LAC, Reynders D, Das B, Faix JD, et al. IFCC Committee for Standardization of Thyroid Function Tests (C-STFT). Harmonization of serum thyroid-stimulating hormone measurements paves the way for the adoption of a more uniform reference interval. Clin Chem 2017;63:1248–60.
12. Vesper HW, Thienpont LM. Traceability in laboratory medicine. Clin Chem 2009;55:1067–75.
13. ISO 17511 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. ISO 17511:2003. Geneva: ISO; 2003.
14. International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), IFCC Scientific Division Working Group for Standardization of Thyroid Function Tests (WG-STFT), Thienpont LM, Beastall G, Christofides ND, Faix JD, Ieiri T, Miller WG, et al. Measurement of free thyroxine in laboratory medicine–proposal of measurand definition. Clin Chem Lab Med 2007;45:563–4.
15. International Federation of Clinical Chemistry and Laboratory Medicine IFCC, IFCC Scientific Division Working

Group for Standardization of Thyroid Function Tests (WG-STFT), Thienpont LM, Beastall G, Christofides ND, Faix JD, Ieiri T, Jarrige V, et al. Proposal of a candidate international conventional reference measurement procedure for free thyroxine in serum. Clin Chem Lab Med 2007;45:934–6.

16. International Federation of Clinical Chemistry; Laboratory Medicine Working Group for Standardization of Thyroid Function Tests. Van Houcke SK, Van Uytfanghe K, Shimizu E, Tani W, Umemoto M, Thienpont LM. 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). Clin Chem Lab Med 2011;49:1275–81.
17. Van Uytfanghe K, De Grande LA, Thienpont LM. A "step-up" approach for harmonization. Clin Chim Acta 2014; 432:62–7.
18. Thienpont LM, Van Uytfanghe K, Van Houcke S. IFCC Working Group for Standardization of Thyroid Function Tests (WG-STFT). Standardization activities in the field of thyroid function tests: a status report. Clin Chem Lab Med 2010;48:1577–83.
19. Thienpont LM, Van Uytfanghe K, Van Houcke S, Das B, Faix JD, MacKenzie F, et al. IFCC Committee for Standardization of Thyroid Function Tests (C-STFT). A progress report of the IFCC Committee for Standardization of Thyroid Function Tests. Eur Thyroid J 2014;3:109–16.
20. JCTLM database of higher-order reference materials, measurement methods/procedures and services. http://www.bipm.org/jctlm/ (Accessed March 2017).
21. CLSI. Evaluation of detection capability for clinical laboratory measurement procedures; approved guideline. 2nd Ed. CLSI document EP17-A2. Wayne (PA): Clinical and Laboratory Standards Institute; 2012.
22. Linnet K. Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. Clin Chem 2000;46:867–9.
23. 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 March 2017).
24. 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.
25. Yue B, Rockwood AL, Sandrock T, La'ulu SL, Kushnir MM, Meikle AW. Free thyroid hormones in serum by direct equilibrium dialysis and online solid-phase extraction-liquid chromatography/tandem mass spectrometry. Clin Chem 2008;54:642–51.
26. De Grande LAC, Goossens K, Van Uytfanghe K, Das B, MacKenzie F, Patru MM, Thienpont LM; IFCC Committee for Standardization of Thyroid Function Tests (C-STFT). 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.
27. Miller WG, Jones GR, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. Clin Chem 2011;57:1670–80.
28. Thienpont LM, Faix JD, Beastall G. Standardization of free thyroxine and harmonization of thyrotropin measurements: a request for input from endocrinologists and other physicians. Thyroid 2015;25:1379–80.