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

Authors: Thienpont, Linda M., Katleen Van Uytfanghe, Linde A. C. De Grande, Dries Reynders, Barnali Das, James D. Faix, Finlay MacKenzie, et al.

In: *Clinical Chemistry* 63 (7): 1248–1260, 2017

Harmonization of serum thyroid-stimulating hormone measurements paves the way for the adoption of a more uniform reference interval

Running head: Harmonization of serum TSH measurements

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], Paul Williams[15] on behalf for the IFCC Committee for Standardization of Thyroid Function Tests (C-STFT).

[1]Department of Pharmaceutical Analysis, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium.

[2]Linda M. Thienpont is now at 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

[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, USA.

[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.

*Corresponding author: (current affiliation) Thienpont & Stöckl Wissenschaftliches Consulting GbR, Erlbacher Strasse 11, Rennertshofen (OT Bertoldsheim), Germany. Tel. +49 8434 94 365 22, email: linda.thienpont@ugent.be

Key words

Thyrotropin; TSH; traceability; all-procedure trimmed mean; method comparison study; harmonization; standardization.

Words: 3094 (max 3500)

Figures: 5

Tables: 1

[16]Abbreviations

Thyroid-stimulating hormone (TSH); reference interval (RI); in vitro diagnostic (IVD); Committee for Standardization of Thyroid Function Tests (C-STFT); all-procedure trimmed mean (APTM); confidence interval (CI); total error (TE); lower limit (LL); upper limit (UL).

Abstract (249)

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 multi-assay method comparison study with clinical 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 supply 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 post-recalibration RIs with a non-parametric 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 2.5, 50 and 97.5 percentile estimates were reduced from 0.27, 0.89 and 2.13 mIU/L to 0.12 , 0.29 and 0.77 mIU/L.

Conclusion

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.

Introduction

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) 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 the thyroid function in patient care (3-7). Reference intervals (RI) reported along with the laboratory data are an integral part of the interpretation process (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 due 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 health care 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 IFCC's 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 multi-assay 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 reference interval 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 the online Supplemental, Section 3). 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 USA-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 Bioethic Committee and written informed consent from patients were received. The de-identified samples were accompanied by a short description of the patients' clinical background (type of thyroid dysfunction, comorbidities, surgery/treatment, ethnicity, gender, 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 anti-thyroperoxidase antibody screening and a serum TSH

concentration <10 mIU/L (cut-off 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, two sets of 101 sample-specific value assignments – for the harmonization panel, referred to below (for historical reasons) as 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 3rd 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 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 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 re-assignment 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 post-recalibration 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 two 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 (i) the pre- and post-recalibration median deviation (%) to the target in distinct concentration intervals; (ii) the mean deviation or bias (%) (and one-sided 95% confidence interval (CI)) to the target after recalibration; (iii) the pre- and post-recalibration CVs (%) of the assay means, and (iv) the total error (TE, %) for the first replicate after recalibration.

For treatment of the pre- and post-recalibration 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 2.5 and 97.5 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 datasets was normally distributed, nor after log transformation ($p<0.01$), we opted for the non-parametric bootstrap (500 replicates) procedure (25). We also estimated the pre- and post-recalibration 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 post-recalibration 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 Supplemental, Sections 1 and 2.

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 more elevated measurement result. The concentrations in the follow-up panel were between 0.002 and 169 mIU/L (based on APTM-4). In the online Supplemental, Figure 1S and Figure 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 the online Supplemental, Figure 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 [mean $results_{recal\ to\ the\ APTM\text{-}11}$] = 0.987 [mean $results_{recal\ to\ the\ APTM\text{-}4}$] + 0.055 ($R^2$ = 0.9999); the mean difference was -2.2% (see the online Supplemental, Figure 4S). Based on this outcome and appreciating the asset 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 Supplemental, 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 (Figure 1A and Figure 1B) show the assays' deviations to the APTM-4 before (Figure 1A) and after recalibration (Figure 1B). Note, the latter was constructed using the measurement data mathematically recalculated with the re-assigned master calibrators. Figure 2A and Figure 2B demonstrate the assay-specific median deviations (%) to the APTM-4 before and after recalibration in 3 concentration intervals. Figure 2A shows the combined picture of the deviations with indication of the 15^{th}, 50^{th} and 85^{th} centiles, while Figure 2B represents for each assay the magnitude and sign of the deviations. From the details listed in the 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 & C, L & C, and B/L & 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.

Figure 3 shows that the bias (%) (and one-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 the 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 interval is shown in the online Supplemental, Figure 5SA. In terms of TE, 8 of the recalibrated TSH assays (A, B, D, F, I, J, L, N) met the

specification (less than 5% of the differences >23.8%), while for the other 6 assays, 7% to 15% were outside the limits (Figure 4).

Reference interval study

Figure 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 the online Supplemental, Table 5S). Figure 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). The 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 the online Supplemental, Figure 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 Supplemental, Section 1 for details).

Discussion

Our attempt to harmonize commercially available TSH immunoassays began with a method comparison using specimens 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 specimens (Phase II) established a proof-of-concept that the approach to harmonization was feasible (30). Recalibration to the APTM was similarly successful using specimens 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 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 out of the 14 assays had pre-harmonization 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 15^{th} and 85^{th} 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 specimens, 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 make it possible for 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 does not mean “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 the Clinical and Laboratory Standards institute (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 set already 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.

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); ZD. Chen, W.Li, H.Xu and JY.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 co-authors are also grateful to the Center for Statistics of Ghent University for the support given under the FIRE Statistical Consulting program.

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

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. American Thyroid Association; American Association of Clinical Endocrinologists. 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. Endocrine Practice 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 [Review].
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 GRD, Barker A, Tate J, LimC-F, 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;44:203–08.
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. 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 February 2017).
16. 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.

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. CLSI document EP17-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2012.
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. EUR 12282 EN. The certification of progesterone in two lyophilized serum materials, CRM 347 and CRM 348. Reference materials Report. Brussels - Luxembourg, 1989.
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.; for the IFCC Working Group on Standardization of Thyroid Function Tests. 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. 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.

31. 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.
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). 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 Edition. CLSI document EP28-A3c. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.
36. De Grande LAC, Goossens K, Van Uytfanghe K, Das B, MacKenzie F, Patru M-M, Thienpont LM; for the 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 2016 April 27. doi: 10.1016/j.cca.2016.04.032. [Epub ahead of print].

Tables

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***	Code	Reference Interval (mIU/L)	Measurement Interval (mIU/L)[e]	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]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)].

Figure Captions

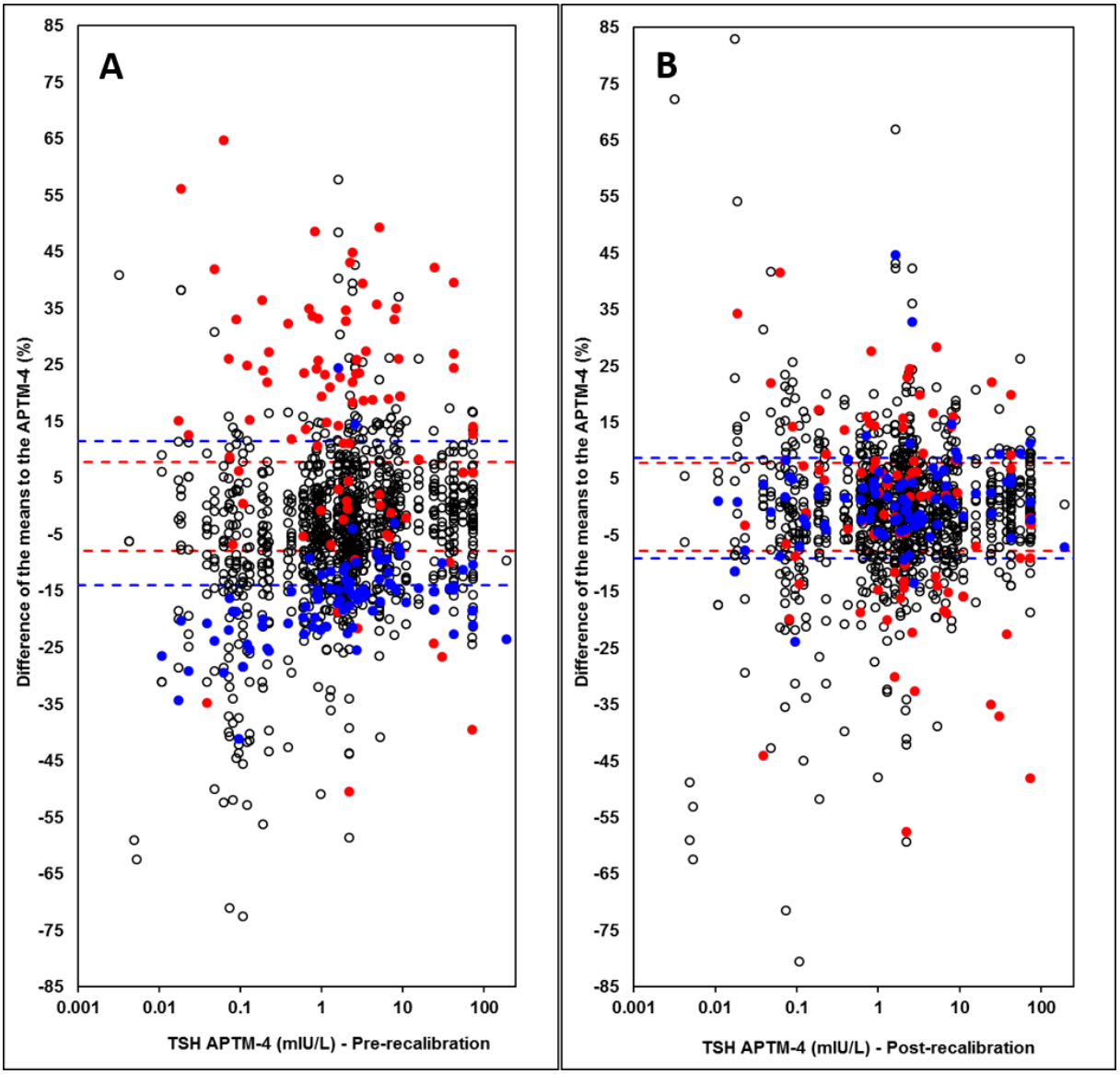


Figure 1: Combined difference (%) plots to the APTM-4 before (A) and after recalibration (B).

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 ~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.

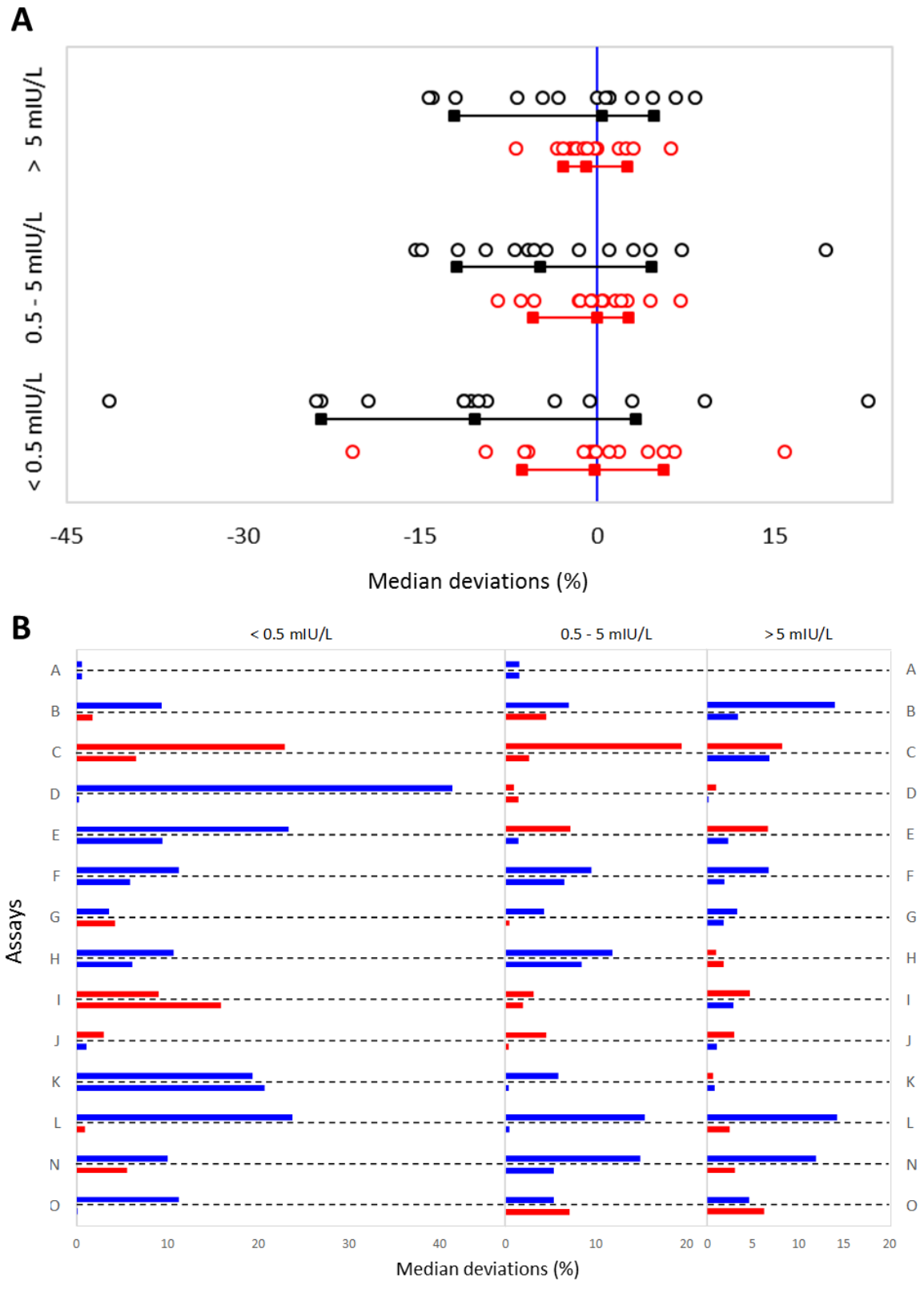
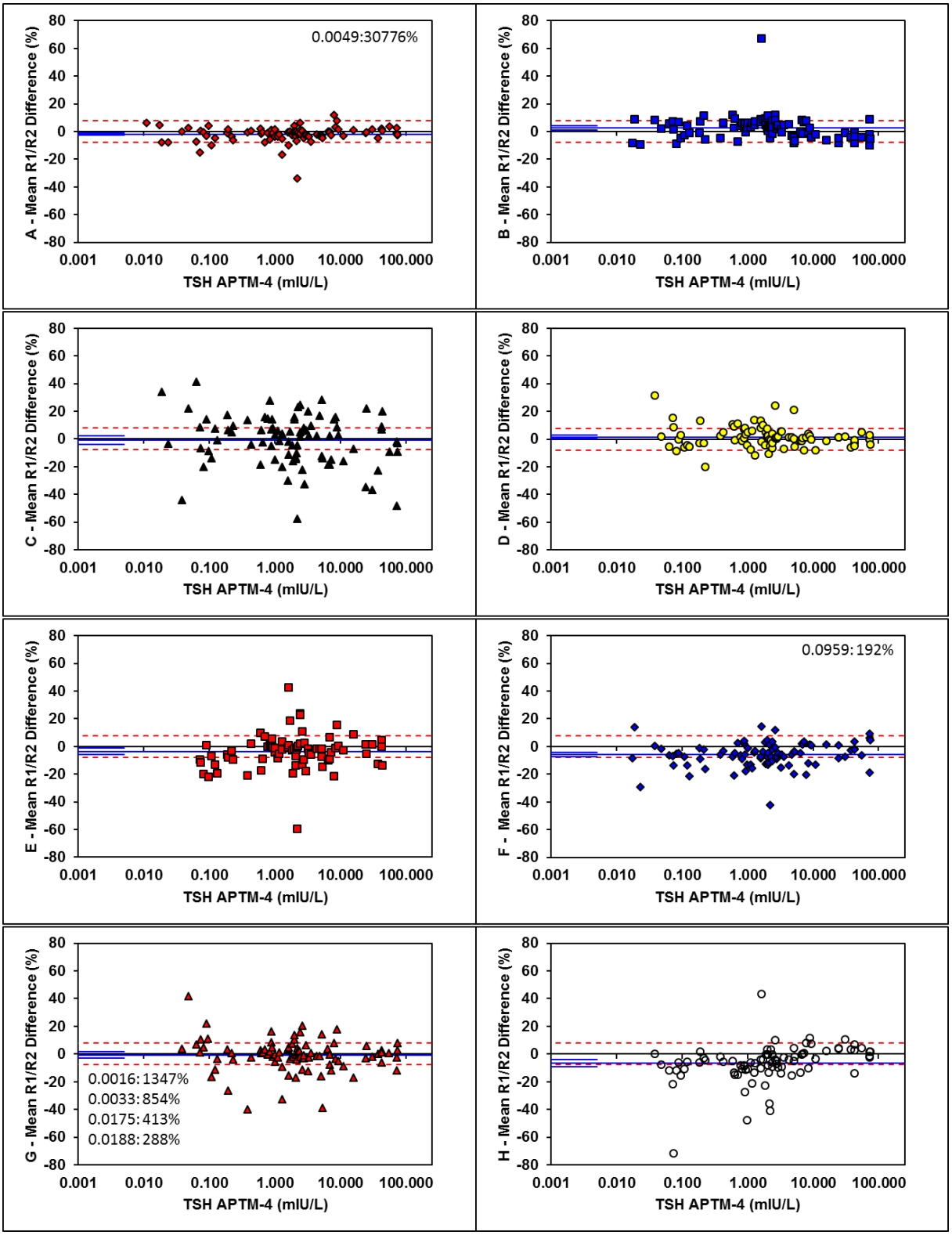


Figure 2: Median deviations (%) of the assays to the APTM-4 before and after recalibration in 3 concentration intervals: low: <0.5 mIU/L, mid: ≥0.5 <5.0 mIU/L, high: ≥5.0 mIU/L.

(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.


A - Mean R1/R2 Difference (%)
0.0049:30776%
TSH APTM-4 (mIU/L)
B - Mean R1/R2 Difference (%)
TSH APTM-4 (mIU/L)
C - Mean R1/R2 Difference (%)
TSH APTM-4 (mIU/L)
D - Mean R1/R2 Difference (%)
TSH APTM-4 (mIU/L)
E - Mean R1/R2 Difference (%)
TSH APTM-4 (mIU/L)
F - Mean R1/R2 Difference (%)
0.0959:192%
TSH APTM-4 (mIU/L)
G - Mean R1/R2 Difference (%)
0.0016:1347%
0.0033:854%
0.0175:413%
0.0188:288%
TSH APTM-4 (mIU/L)
H - Mean R1/R2 Difference (%)
TSH APTM-4 (mIU/L)

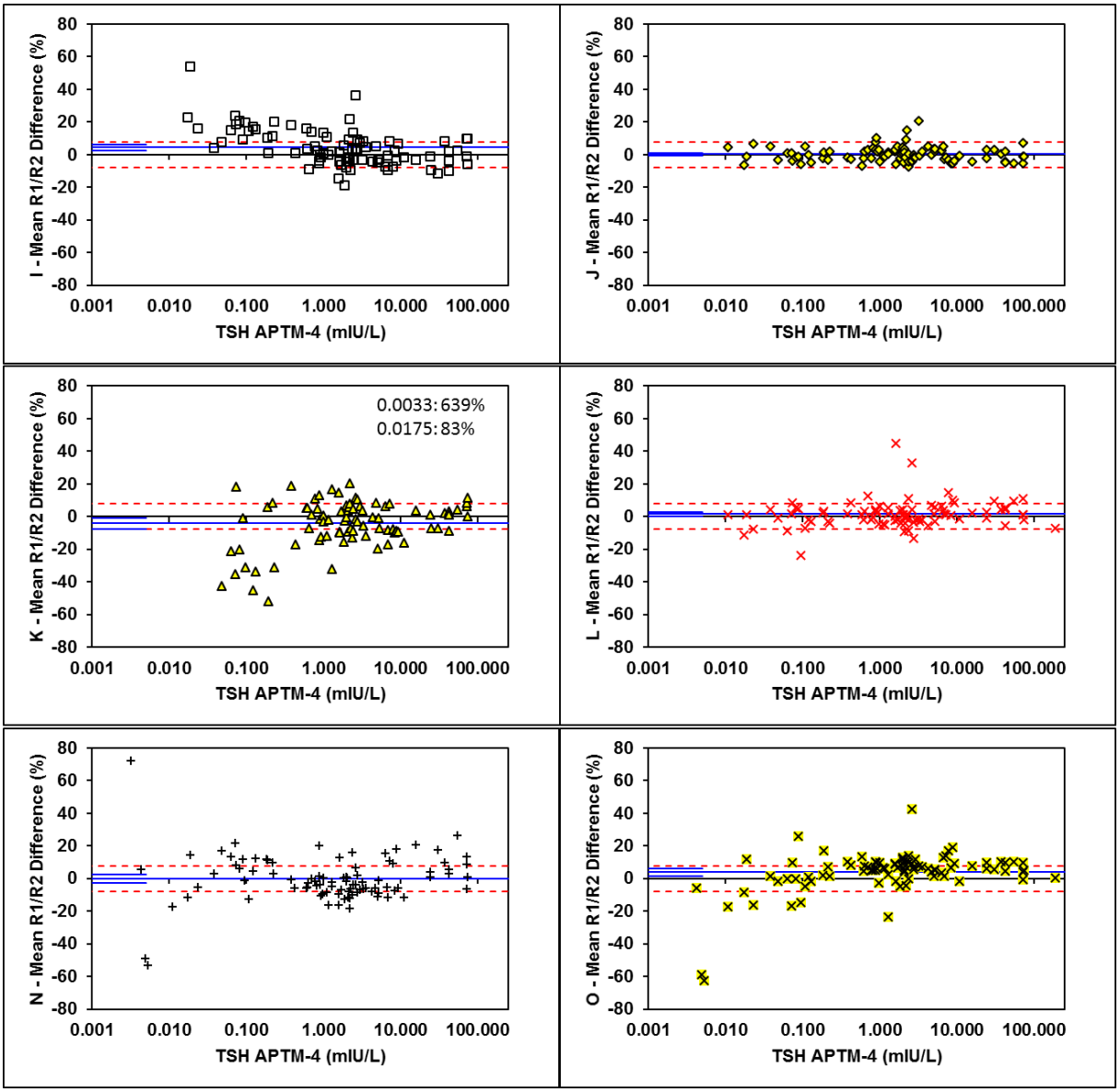
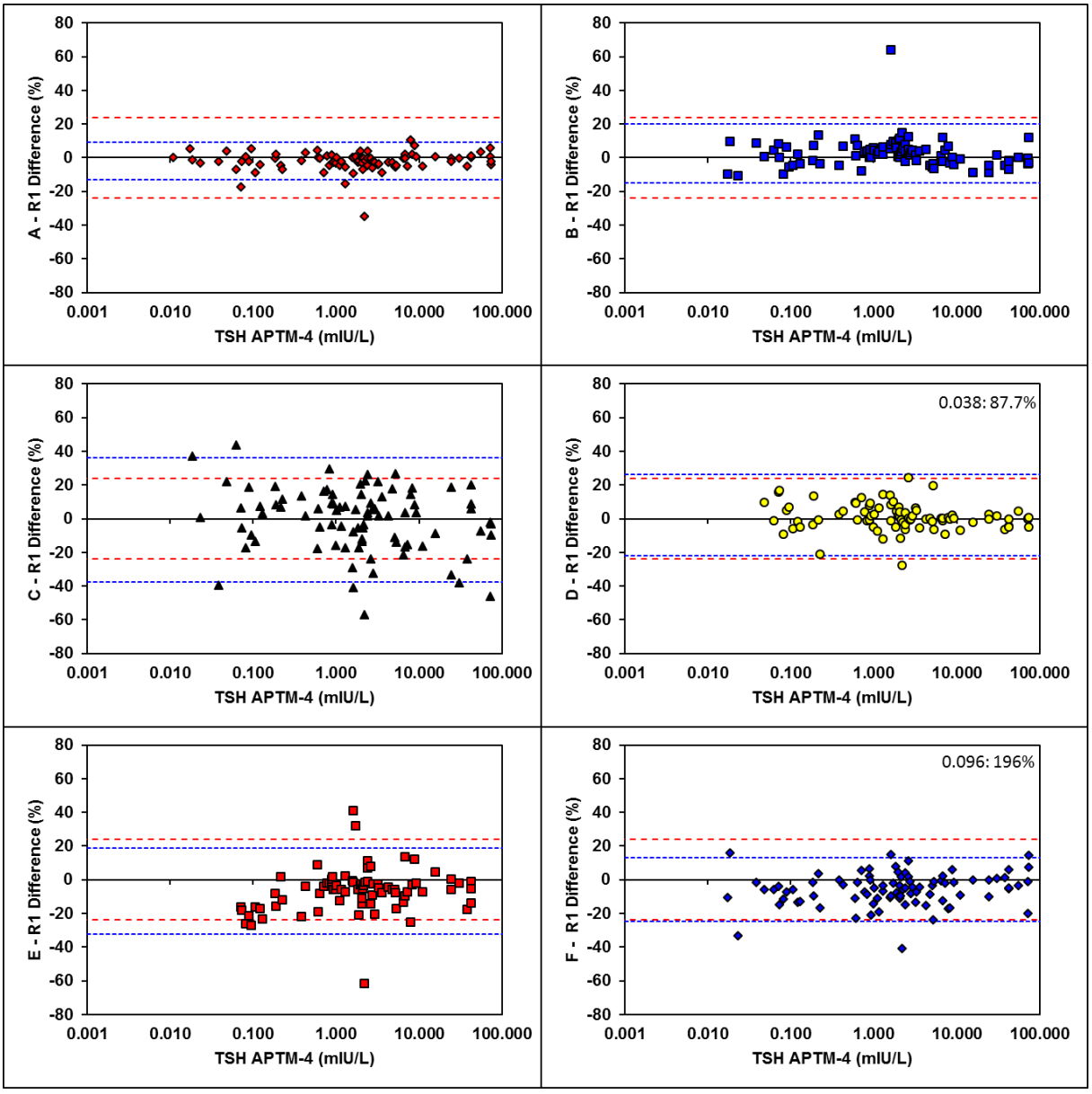


Figure 3: Difference (%) plots after recalibration of the assays to the APTM-4.

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 one-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.


A - R1 Difference (%)
B - R1 Difference (%)
C - R1 Difference (%)
D - R1 Difference (%)
E - R1 Difference (%)
F - R1 Difference (%)
TSH APTM-4 (mIU/L)
0.038: 87.7%
0.096: 196%

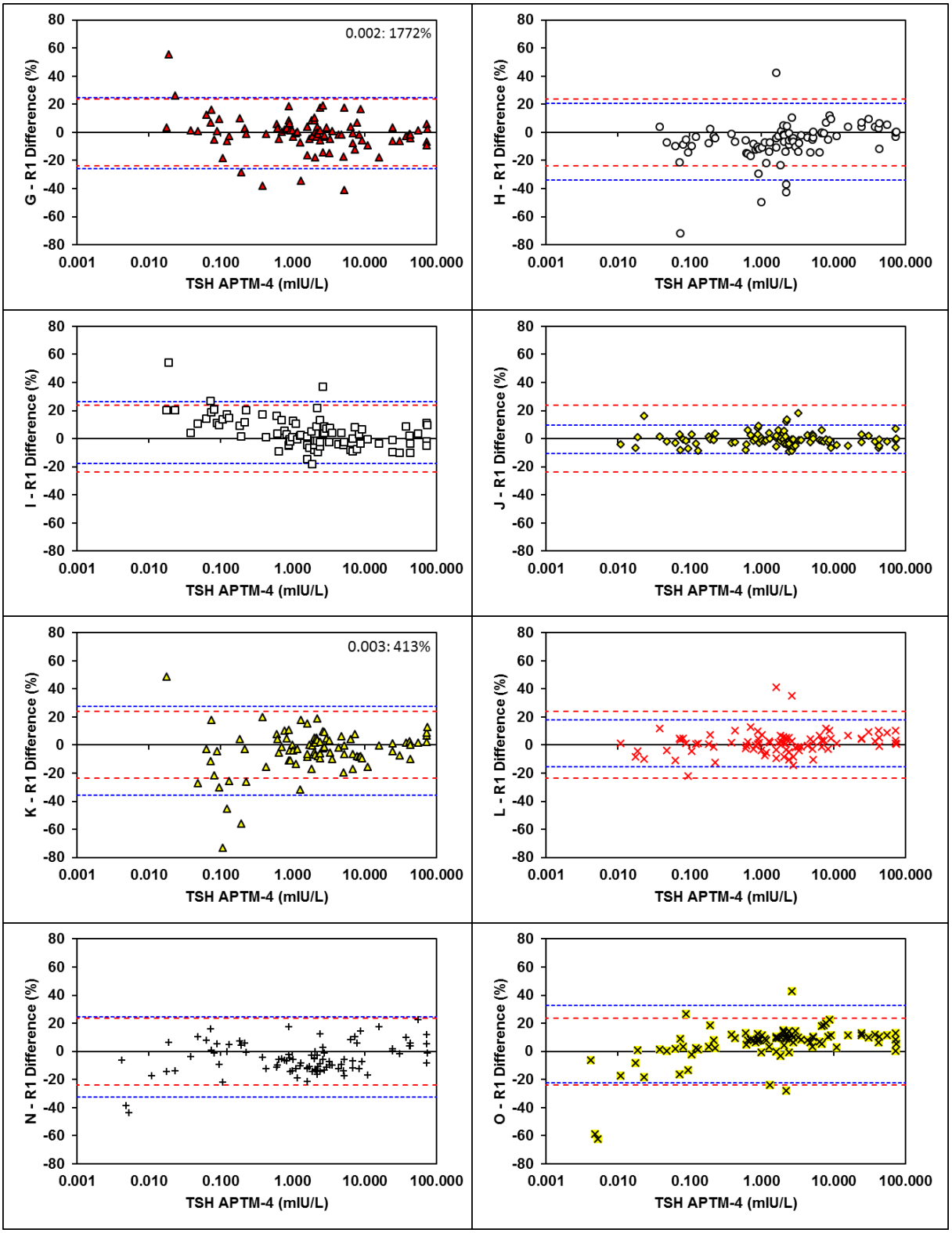


Figure 4: Total error (%) plots of the first replicate after recalibration to the APTM-4. The TE was estimated from the % difference to the APTM-4 of the first replicate after recalibration. It was validated against the 23.8% specification derived

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.

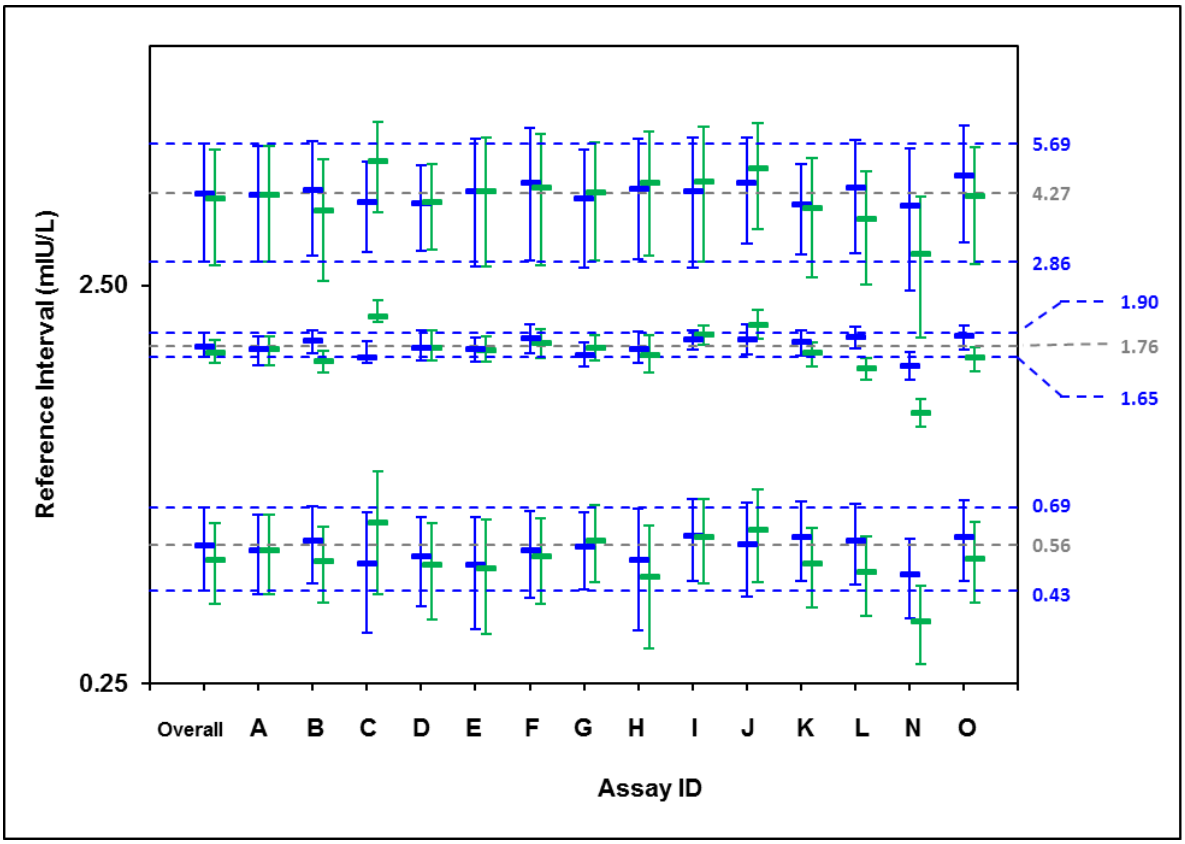


Figure 5: Comparison of the pre- and post-recalibration reference intervals of the individual immunoassays to the overall RI (n = 120).

The pre- and post-recalibration 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.

Supplement to "Harmonization of serum thyroid-stimulating hormone measurements paves the way for the adoption of a more uniform reference interval" by Linda M. Thienpont, Katleen Van Uytfanghe, Linde A.C. De Grande, Dries Reynders, Barnali Das, James D. Faix, Finlay MacKenzie, Brigitte Decallonne, Akira Hishinuma, Bruno Lapauw, Paul Taelman, Paul Van Crombrugge, Annick Van den Bruel, Brigitte Velkeniers, Paul Williams for the IFCC Committee for Standardization of Thyroid Function Tests (C-STFT).

Contents

1 Details of the homogeneity study

As mentioned in the main text, we used for the homogeneity study the protocol described for certified reference materials of the European Commission (Ref. 27 in the main text). Note that one of the IVD manufacturers volunteered to perform measurements for the homogeneity study, i.e., Roche with the Cobas Elecsys TSH. Of the 12 aliquots per sample, 4 were randomly selected and pooled; 8 measurements were done for this pool. The remaining 8 aliquots were measured in singleton. All measurements (of the individual aliquots and the pool) were done in an alternating sequence, within run. The combined data were tested for outliers with a Grubbs test. Then the variances of the measurement data for the pool and the individual aliquots were assessed for significant difference by means of an F-test (95% confidence level) (see table below).

Table 1S: Summary of the results of the homogeneity study.

Sample ID	Mean (mIU/L) (aliquots)	CV (%) (aliquots)	Mean (mIU/L) (pool)	CV (%) (pool)	p (F-test, 95% CL[a])
1	1.681	0.6	1.713	0.6	0.9
2	2.604	0.7	2.647	0.4	0.3
3	25.23	0.5	25.71	0.3	0.4
4	67.39	0.4	68.58	0.5	0.7
5	0.008	10.7	0.009	9.6	1.0
6	0	N/A[b]	0	N/A	N/A
7	0.981	0.8	0.989	0.7	0.4
8	0.031	2.3	0.031	2.5	0.9
9	0.703	0.7	0.681	0.5	0.3

10	5.374	0.6	5.318	0.9	0.4
11	0.241[c]	1.1	0.241	1.5	0.6
12	10.53	0.9	10.47	0.5	0.2

[a]CL: confidence level

[b]: N/A: not applicable

[c]: 1 Outlier identified with the Grubbs test

2 Details of the stability study

The stability study has already started, but will last in total 2 and 5 years. For each panel, 9 samples in total (3 times 3 with a concentration in the hypo-, eu- and hyperthyroid range, respectively) will be stored for different time periods at the effective storage temperature (-70°C) or at the reference temperature (liquid N2). Storage is done in the facilities of the National Institute for Biological Standards and Control. Note that the samples were collected as described for the homogeneity study in the main text. The time storage points are 0, 8, 16 and 24 months (2 year study) or 0, 36, 48 and 60 months (5 year study). To avoid any complications due to measurement errors/variation, all samples will be measured within-run at the end of each study. A schematic overview of the design of the stability study is given in Table 1S: at time points 8 and 16 or 36 and 48, one box of samples will be moved from the reference to the storage temperature.

Table 2S: Stability study: experimental set-up for storage of the samples.

Sample	0-8 months	8-16 months	16-24 months
A1-ref 1	Store at L N2	Store at L N2	Store at L N2
A2-ref 1	Store at L N2	Store at L N2	Store at L N2
A1-t1 1	Store at L N2	Store at L N2	Store at -70°C
A2-t1 1	Store at L N2	Store at L N2	Store at -70°C
A1-t2 1	Store at L N2	Store at -70°C	Store at -70°C
A2-t2 1	Store at L N2	Store at -70°C	Store at -70°C
A1-t3 1	Store at -70°C	Store at -70°C	Store at -70°C
A2-t3 1	Store at -70°C	Store at -70°C	Store at -70°C
Sample	**0-36 months**	**36-48 months**	**48-60 months**
A1-ref 2	Store at L N2	Store at L N2	Store at L N2
A2-ref 2	Store at L N2	Store at L N2	Store at L N2
A1-t1 2	Store at L N2	Store at L N2	Store at -70°C
A2-t1 2	Store at L N2	Store at L N2	Store at -70°C
A1-t2 2	Store at L N2	Store at -70°C	Store at -70°C
A2-t2 2	Store at L N2	Store at -70°C	Store at -70°C
A1-t3 2	Store at -70°C	Store at -70°C	Store at -70°C
A2-t3 2	Store at -70°C	Store at -70°C	Store at -70°C

Note: A stands for aliquot; L N2 for liquid nitrogen.

3 Sample sourcing – Eligibility and exclusion criteria

Inclusion criteria:	-Individuals are at least 18 years old and competent to give informed consent, as considered by the physician, study nurse or other health care professional interviewing the patient. -Individuals being evaluated for a thyroid disorder and classified into one of the following groups (if possible evenly distributed): A: Hyperthyroid (n = 30) A1: 10 patients with suppressed TSH, around 0.01 mIU/L A2: 10 patients with TSH values between 0.01 – 0.1 mIU/L A3: 10 patients with TSH values between 0.1 – 0.3* mIU/L B: Euthyroid (n = 30) Patients with TSH values between 0.3 – 3.0 mIU/L* C: Hypothyroid (n = 40) C1: 20 patients with TSH values between 3.0 – 50 mIU/L* C2: 20 patients with TSH values > 50 mIU/L up to 100 mIU/L. Donors treated for thyroid dysfunction can be included, provided information on the type of treatment and start of the treatment is available.
Exclusion Criteria	-Those individuals previously enrolled into this clinical study. -Individuals diagnosed with a severe non-thyroidal illness (NTI), defined as a state of dysregulation where levels of T3, T4, FT3 and/or FT4 are abnormal although the thyroid gland does not appear to be dysfunctional. In practice, NTI is reported to be usually associated with critical illness or starvation. Examples: chronic renal failure, liver cirrhosis, advanced (active) malignancy, sepsis, trauma, prolonged fasting or starvation, heart failure, MI, and any psychiatric disorder. -Individuals with known pregnancy.

-Those patients not meeting the established inclusion criteria.

*Note: these values are only indicative because they depend on the measurement range and the reference interval of the assay used to evaluate the TSH status.

4 Uncertainty of the APTM targets

Figure 1S: Plots showing the uncertainties of the statistically estimated APTM-4 target values for the TSH harmonization panel. The uncertainty for each target value was estimated by use of a bootstrapping procedure (n = 1500) with sampling from the distribution of data used in the robust factor analysis model for deriving the APTM (Ref. 22 in the main text). This procedure provided for each APTM estimate the 95% confidence interval, which was used as a measure of the uncertainty. The reason why the bootstrap simulations leads to asymmetric uncertainties is that the results by the different assays per individual sample were not normally distributed around that APTM, e.g., when more assays gave for a certain sample a higher concentration than the estimated APTM, or when 1 assay gave a much higher (or lower) value than the other assays for a certain sample, the uncertainty was broader at the positive side (and vice versa). To better visualize the relative magnitudes of the uncertainties, we plotted them for the samples sorted by increasing but normalized concentration (concentration/concentration); the horizontal bars represent for each sample the upper and lower limit of 95% confidence interval around the APTM estimate expressed as ratio to the estimate to which both relate. We finally estimated the overall uncertainty (%) from the mean of the relative uncertainties at both sides of the APTM targets, which amounted to 0.9% (lower limit) and 0.7% (upper limit). Note: 1) the plotted data are for samples with concentrations above the mean of the assays' lower limits of the stated measurement intervals (0.0175 mIU/L; for the individually claimed measurement intervals, see Table 1 in the main text); 2) the uncertainty of the APTM-11 was similar (not shown, because we decided to only use the APTM-4; see "Results" in the main text).

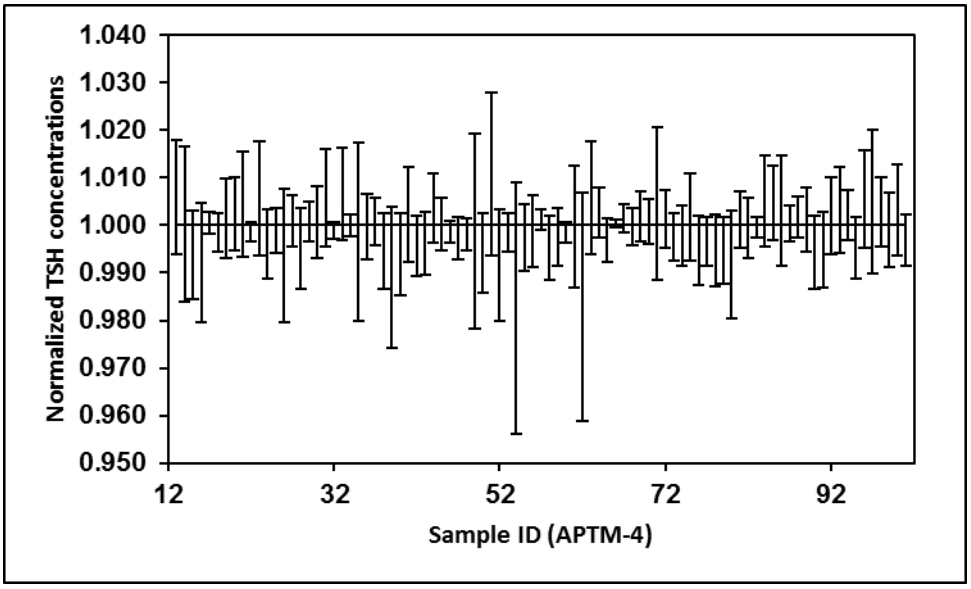
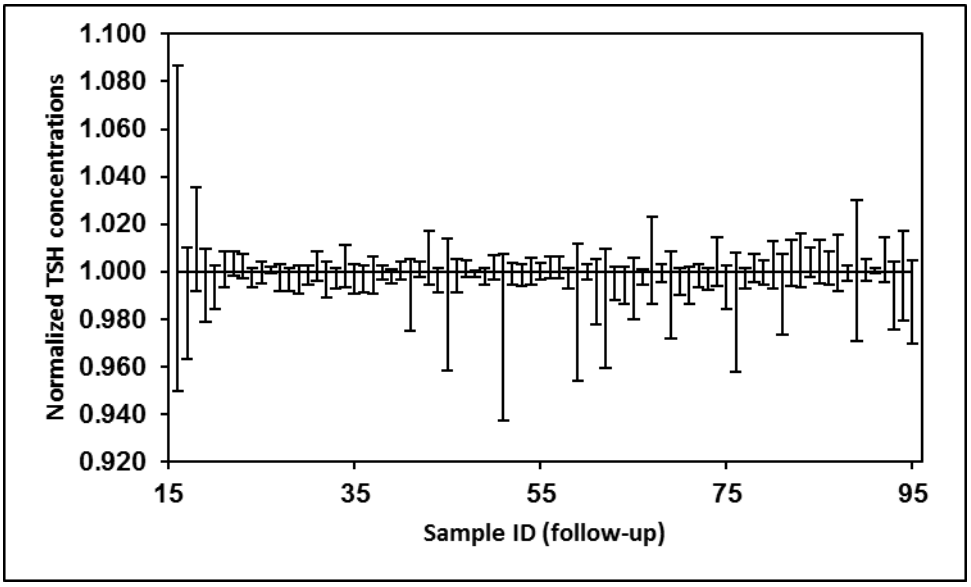
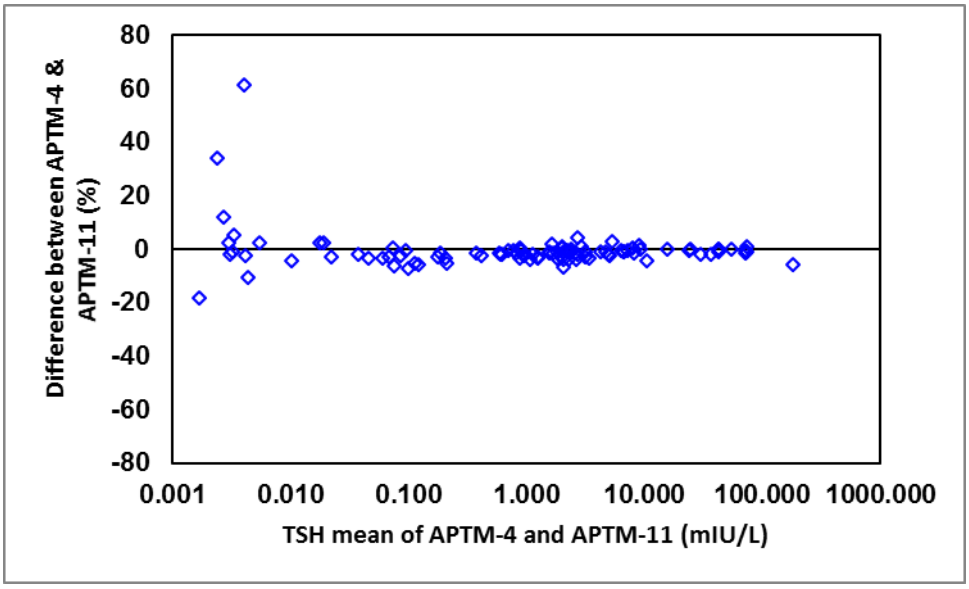


Figure 2S: Similar plot as above but for the first follow-up panel. The mean of the relative uncertainties at both sides of the APTM-4 targets amounted to 1.1% (lower limit) and 0.7% (upper limit).



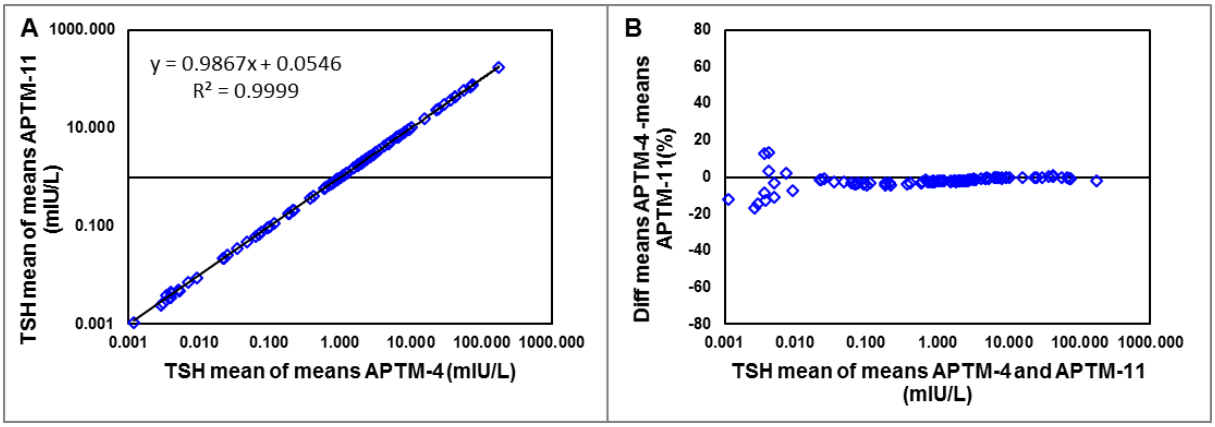
5 Comparison of the APTM targets

Figure 3S: Difference (%) plot of the targets based on the APTM-4 and APTM-11 relative to their mean.



6 Comparison of the recalibration to the APTM-11 and APTM-4

Figure 4S: (A) shows the overall mean concentrations of the samples after recalibration of the assays to the APTM-11 (Y-axis; log) plotted to those after recalibration to the APTM-4 (X-axis; log), as well as the ordinary linear regression equation (with R^2); note that each data point represents the overall mean concentration for a sample calculated from 14 mean concentrations per recalibrated assay ("mean" concentrations because each sample had been measured in duplicate by the respective assays); as described in the main text each manufacturer mathematically recalibrated and reported back 2 sets of measurement results as if his assay was recalibrated either to the APTM-4 (X-axis) or APTM-11 target (Y-axis); (B) plots the data as % difference (mean of means recalibrated to the APTM-4 minus those recalibrated to the APTM-11) relative to their mean. Note that for this comparison, we used all reported results.



7 Assay-specific median deviations (%) (pre- and post-recalibration)

Table 3S: Median deviations (%) of each of the immunoassays to the APTM-4 before and after recalibration in distinct concentration intervals.

Assays	Before recalibration			After recalibration		
	<0.5 mIU/L	≥0.5 <5 mIU/L	≥5 mIU/L	<0.5 mIU/L	≥0.5 <5 mIU/L	≥5 mIU/L
A	-0.6	-1.6	0.0	-0.6	-1.6	0.0
B	-9.4	-7.0	-14.0	1.8	4.5	-3.4
C	23.0	19.4	8.3	6.6	2.6	-6.9
D	-41.4	1.0	1.0	-0.3	1.5	-0.2
E	-23.4	7.2	6.7	-9.5	-1.5	-2.3
F	-11.3	-9.5	-6.8	-5.9	-6.5	-1.9
G	-3.6	-4.3	-3.3	4.3	0.5	-1.8
H	-10.7	-11.8	1.0	-6.2	-8.4	1.8
I	9.1	3.1	4.7	15.9	2.0	-2.9
J	3.0	4.5	3.0	-1.1	0.4	-1.1
K	-19.4	-5.9	0.7	-20.7	-0.4	-0.8
L	-23.8	-15.4	-14.3	1.0	-0.5	2.5
N	-10.1	-14.9	-12.0	5.6	-5.4	3.1

O	-11.3	-5.4	-4.6	-0.1	7.1	6.3

8 Post-recalibration biases (%)

Table 4S: Assay biases (%) and one-sided 95% confidence interval (CI) after recalibration to the APTM-4, and their assessment against the specification of 7.8% inferred from the biological variation.

Assay	Bias[1] (%)	One-sided[2] 95% CI (%)	Upper bias limit (%) (Bias + CI)	Lower bias Limit (%) (Bias − CI)
A	-1.8	1.0	-0.8	-2.8
B	2.6	1.6	4.2	1.0
C	-0.8	3.2	2.4	-4.0
D	1.5	1.4	2.9	0.1
E	-3.6	2.4	-1.2	-5.9
F	-5.8	1.6	-4.2	-7.4
G	-0.7	2.2	1.5	-2.9
H	-6.6	2.5	-4.1	-9.1
I	4.3	2.0	6.3	2.3
J	0.3	0.8	1.1	-0.6
K	-4.3	3.5	-0.8	-7.7
L	1.4	1.4	2.9	0.0
N	-0.1	2.5	2.4	-2.6
O	3.7	2.3	6.0	1.5

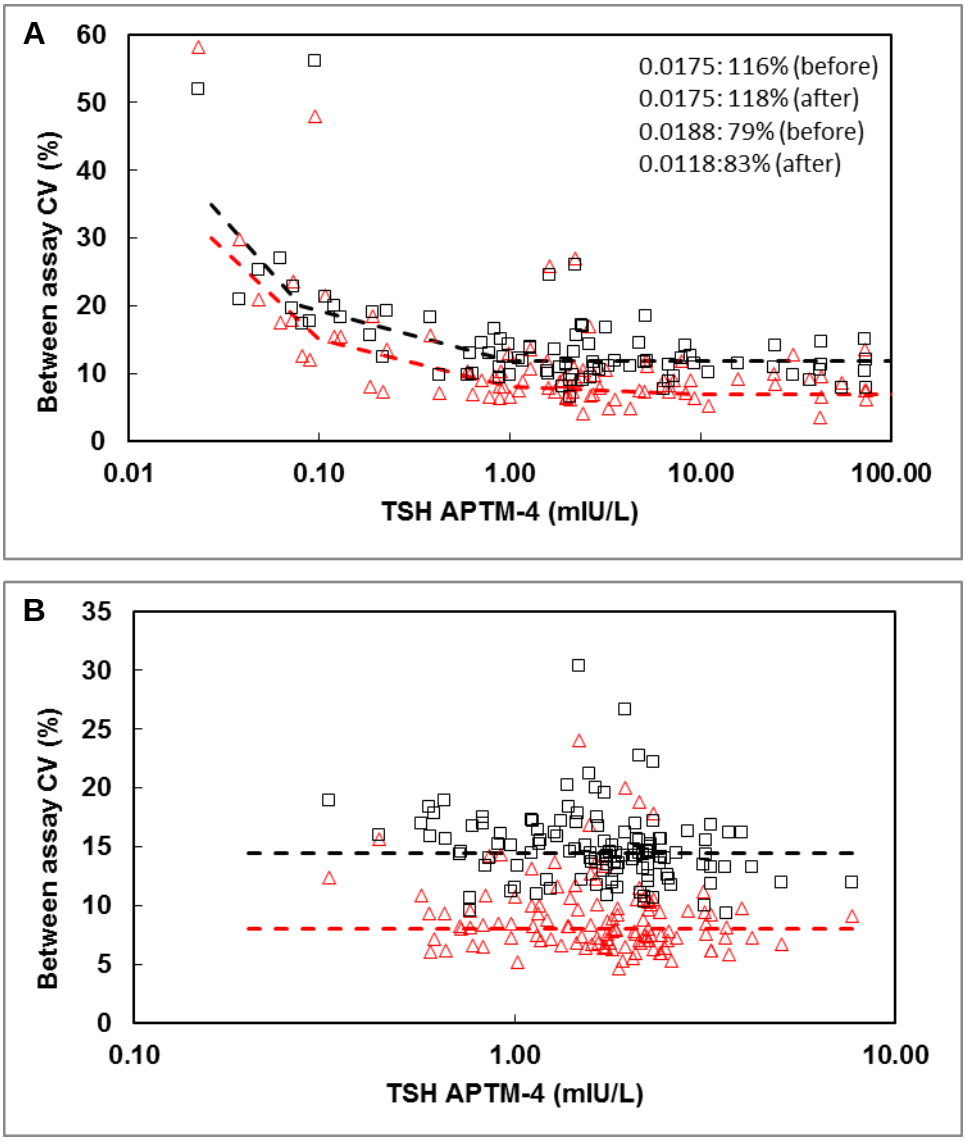
[1]The bias (%) is the mean of the deviations (%) calculated for the claimed measurement ranges.

[2]One-sided *t*-values (obtained from Excel with the function TINV(0.1, df)) were used for calculation of the CI.

Interpretation (Ref. 28 in the main text): for 13 of the 14 assays it can be confidently asserted that their bias meets the 7.8% specification with a 95% probability; for assay H, in spite of a bias of -6.6%, it is not possible to state this with 95% confidence, because the lower limit of the one-sided 95% CI (colored cell) violates the -7.8% limit.

9 CV (%) reduction

Figure 5S: Profile showing the between assay CV (%) (CV calculated from the assay means) for the harmonization panel before and after recalibration, (A) for the concentration interval from 0.0175 mIU/L to 74 mIU/L (note these limits represent the mean of the LL of the claimed measurement ranges and the second highest concentration in the panel; this range shows that the CVs are constant from a concentration of approximately 0.5 mIU/L on), (B) for the concentration range covered by the reference interval panel. The black squares stand for the CV of each individual sample before recalibration, the red triangles after recalibration; the dotted lines in A & B were constructed to fit the data points and match with the median CV per sample over all assays.



10 Reference interval study

Table 5S: Characteristics of the overall and individual TSH reference intervals before and after recalibration of the immunoassays.

ID	Median [90% CI]	Width RI	LL [90% CI]	UL [90% CI]	Δ^1LL	Δ^1UL
			(mIU/L)		(%)	
Before recalibration						
Overall	1.70 [1.60-1.82]	3.63	0.51 [0.40 - 0.63]	4.14 [2.81 - 5.48]		
A	1.73 [1.57-1.86]	3.69	0.54 [0.42 - 0.66]	4.23 [2.86 - 5.60]	5.4	2.1
B	1.61 [1.51-1.71]	3.37	0.51 [0.40 - 0.62]	3.88 [2.64 - 5.11]	-1.2	-6.5
C	2.09 [2.02-2.29]	4.50	0.63 [0.42 - 0.85]	5.14 [3.82 – 6.45]	23.4	23.9
D	1.75 [1.62-1.93]	3.55	0.50 [0.36 - 0.63]	4.05 [3.07 - 5.03]	-3.4	-2.4
E	1.73 [1.61-1.86]	3.83	0.49 [0.33 - 0.64]	4.32 [2.78 - 5.87]	-4.8	4.3
F	1.79 [1.64-1.94]	3.89	0.52 [0.40 - 0.65]	4.41 [2.81 - 6.01]	1.9	6.4
G	1.74 [1.62-1.87]	3.73	0.57 [0.45 - 0.70]	4.30 [2.89 - 5.71]	11.7	3.7
H	1.67 [1.51-1.87]	4.07	0.47 [0.31 - 0.62]	4.53 [2.97 - 6.09]	-9.4	9.3
I	1.88 [1.77-1.98]	3.97	0.59 [0.45 - 0.73]	4.56 [2.87 - 6.25]	14.1	9.9
J	2.00 [1.83-2.17]	4.31	0.61 [0.45 - 0.77]	4.92 [3.46 - 6.38]	18.3	18.6
K	1.70 [1.56-1.80]	3.41	0.50 [0.39 - 0.61]	3.91 [2.62 - 5.20]	-2.6	-5.7
L	1.55 [1.45-1.64]	3.19	0.48 [0.37 - 0.59]	3.67 [2.51 - 4.84]	-7.0	-11.5
N	1.20 [1.10-1.29]	2.65	0.36 [0.28 - 0.44]	3.01 [1.84 - 4.17]	-29.9	-27.5
O	1.65 [1.52-1.75]	3.67	0.52 [0.4 - 0.64]	4.19 [2.83 - 5.55]	0.7	1.1
After recalibration						
Overall	1.76 [1.65-1.90]	3.72	0.56 [0.43 - 0.69]	4.27 [2.86 - 5.69]		
A	1.73 [1.57-1.86]	3.69	0.54 [0.42 - 0.66]	4.23 [2.86 - 5.61]	-3.2	-1.0
B	1.82 [1.69-1.93]	3.79	0.57 [0.45 - 0.70]	4.36 [2.97 - 5.76]	2.4	2.1
C	1.65 [1.60-1.81]	3.56	0.50 [0.33 - 0.67]	4.07 [3.03 - 5.11]	-10.0	-4.9
D	1.75 [1.62-1.93]	3.50	0.52 [0.39 - 0.66]	4.02 [3.05 - 5.00]	-6.4	-5.9
E	1.73 [1.61-1.85]	3.82	0.50 [0.34 - 0.65]	4.32 [2.79 - 5.85]	-10.8	1.1
F	1.84 [1.69-2.00]	4.01	0.54 [0.41 - 0.67]	4.55 [2.89 – 6.22]	-3.0	6.5
G	1.67 [1.56-1.80]	3.58	0.55 [0.43 - 0.67]	4.13 [2.78 - 5.48]	-1.3	-3.5
H	1.73 [1.59-1.91]	3.86	0.51 [0.34 - 0.68]	4.38 [2.91 - 5.84]	-8.3	2.4

ID	Median [90% CI]	Width RI	LL [90% CI]	UL [90% CI]	Δ[1]LL	Δ[1]UL
			(mIU/L)		(%)	
I	1.83 [1.72-1.92]	3.74	0.59 [0.45 - 0.72]	4.32 [2.77 - 5.88]	5.3	1.2
J	1.84 [1.68-1.99]	3.96	0.56 [0.41 - 0.71]	4.52 [3.18 – 5.87]	0.4	5.8
K	1.81 [1.66-1.93]	3.43	0.58 [0.45 - 0.72]	4.01 [2.98 - 5.04]	4.5	-6.2
L	1.86 [1.74-1.97]	3.83	0.57 [0.44 - 0.70]	4.41 [3.01 - 5.81]	2.7	3.1
N	1.58 [1.45-1.70]	3.49	0.47 [0.36 – 0.58]	3.97 [2.43 – 5.51]	-15.6	-7.2
O	1.87 [1.72-1.98]	4.16	0.59 [0.45 – 0.72]	4.74 [3.21 – 6.27]	4.9	10.9

[1]Difference (%) compared to the overall RI LL and UL, respectively.

Note: The colored cells indicate the ranges of the medians and the percentiles discussed in the main text.

11 Considerations before deciding to use the APTM-11 or APTM-4 for harmonization

Pro's for using the APTM-11

-The harmonization concept developed by C-STFT was based on the APTM inferred from the measurement results by all assays which participated in the method comparison study (Ref. 19).

-Target values based on a greater number of assays are more representative, which is a virtue for harmonization projects.

-All the experience in this project up to now is based on the APTM estimated from the measurement results by all participating manufacturers.

-According to this APTM concept, all manufacturers are and feel equivalent partners in the harmonization process and equally contribute, provided their assay's measurement range covers the concentration range of the method comparison samples and their measurement results correlate sufficiently well with the APTM. However, as described in the main text, in this study with 14 participants, we had already to deviate. Indeed, we calculated the APTM-11, because for one manufacturer the assay design was not real 3rd generation, and another 2 manufacturers joined the project 1 year after completion of the validation of the target setting. Also for the future, new manufacturers who use the follow-up panel will have to accept that they did not contribute to the target setting.

Con's against the APTM-11

The more assays that participate in the target setting of a panel with samples for which a only a restricted volume is available (inherent for a project based on samples from patients), the more sample volume is consumed. For example, in this study with duplicate measurement by 14 assays, the harmonization panel was almost depleted. This would be of particular concern for the follow-up panel, which for obvious reasons should remain with as many sample aliquots as possible after target setting. Therefore, it was clear from the

beginning of the harmonization activity that the target setting of the follow-up panel should be done with fewer assays. This decision required that the following questions be answered:

-How many assays to select for the reduced target setting protocol?

-Which criteria to use for the selection?

-Will the APTM from fewer assays be biased vis-à-vis the APTM from more assays?

-How to ensure that the selected assays are stable in time, so that they can be used for target setting of the future follow-up panels?

The study was designed to be able to answer most of the above questions. In short, the design can be described as follows.

-For the harmonization panel (n = 101 samples) the APTM-11 was calculated from the measurement results by 11 assays.

-Four assays were selected for measuring the first follow-up panel (n = 95) and the harmonization panel, both in the same run.

-The key criteria for the selection of assays were the assays' performance in the Phase III study; and select assays on the basis of results that were symmetrically located about the APTM with, in addition, a low scatter.

-The APTM-4 was calculated from the double sample size (n = 196) using both follow-up and harmonization samples compared to the APTM-11.

-The bias between the APTM-11 and APTM-4 targets was assessed, as well as the outcome of recalibration of all assays against both targets.

The conclusions reached in this study to the above questions were:

-using 4 assays for target setting is sufficient;

-the validity of the selection criteria is confirmed;

- the mean difference between the APTM-4 and APTM-11 targets is only -0.6%;

-regression analysis of the means of means by the immunoassays recalibrated to both APTM targets gives a mean difference of only -2.2%.

Additional pro's for the APTM-4 confirmed by this study

-The APTM-4 calculated from the double sample size ensures a lower level of uncertainty.

-The target setting establishes the link between the harmonization and first follow-up panel, so that the traceability of the latter to the very first harmonization basis is a fact. Note that without this link, traceability would most probably have required a value transfer from the harmonization to the follow-up panel causing additional uncertainty.

-Both panels start with the same low level of uncertainty.

Potential con's against the APTM-4 resulting from this study

Application of the APTM-11 on APTM-4 regression equation, $y = 0.9867\ x + 0.0546$ mIU/L, at TSH reference limits, for example, those ultimately obtained in the study, namely, 0.56 and 4.27 mIU/L, predicts essentially no change at the upper limit, but at the lower limit an increase of 8 or 9%. This change exceeds the mean deviations which can be explained by noise due to the assay variability currently tolerated by IVD manufacturers (approximately 5% in the mean), but also the "desirable specification" for bias (7.8%).

Open question that still needs to be answered

-Will it be possible to use the 4 assays selected here to set the APTM-4 target for future follow-up panels? This decision will depend on the stability of the long-term performance of the assays. It is our plan to assess stability by our Percentiler application described elsewhere (Ref. 36). The decision will also require that the concerned manufacturers transparently communicate on relevant assay changes.

-It might be that the current concept works over several years. However, if one day it is necessary, the harmonization exercise can be repeated.

Final conclusion taken by the C-STFT and its IVD partners

It is obvious that the choice between the two sets of TSH value assignments involved competing considerations (pro's – con's) and finally had to be decided pragmatically. The

considerations were openly discussed in a meeting where nearly all IVD partners were present. They unanimously agreed to the decision to recalibrate to the APTM-4. The following facts were accorded most relative weight in the decision:

-the good comparability between the APTM-11 and APTM-4;

-more aliquots are retained in the follow-up panel targeted by only 4 assays;

-the link established between the first follow-up and harmonization panel (both were assigned with targets (APTM-4) in parallel) assures full traceability of the former to the latter; in addition, both have the same level of uncertainty.

For the future, it will be important that a new follow-up panel is always developed before depletion of the previous one, whereby both are measured in overlap.