Report of the IFCC Working Group for Standardization of Thyroid Function Tests; Part 1: Thyroid-Stimulating Hormone

Linda M. Thienpont,1* Katleen Van Uytfanghe,1 Graham Beastall,2 James D. Faix,3 Tamio Ieiri,4 W. Greg Miller,5 Jerald C. Nelson,6 Catherine Ronin,7 H. Alec Ross,8 Jos H. Thijssen,9 and Brigitte Toussaint,10 for the IFCC Working Group on Standardization of Thyroid Function Tests

BACKGROUND: Laboratory testing of serum thyroid-stimulating hormone (TSH) is an essential tool for the diagnosis and management of various thyroid disorders whose collective prevalence lies between 4% and 8%. However, between-assay discrepancies in TSH results limit the application of clinical practice guidelines.

METHODS: We performed a method comparison study with 40 sera to assess the result comparability and performance attributes of 16 immunoassays.

RESULTS: Thirteen of 16 assays gave mean results within 10% of the overall mean. The difference between the most extreme means was 39%. Assay-specific biases could be eliminated by recalibration to the overall mean. After recalibration of singlicate results, all assays showed results within the biological total error goal (22.8%), except for 1 result in each of 4 assays. For a sample with a TSH concentration of 0.016 mIU/L, 6 assays either did not report results or demonstrated CVs >20%. Within-run and total imprecision ranged from 1.5% to 5.5% and 2.5% to 7.7%, respectively. Most assays were able to match the internal QC targets within 5%. Within-run drifts and shifts were observed.

CONCLUSIONS: Harmonization of TSH measurements would be particularly beneficial for 3 of the 16 examined assays. These data demonstrate that harmonization may be accomplished by establishing calibration traceability to the overall mean values for a panel of patient samples. Although a majority of assays showed excellent quality of performance, some would benefit from improved within-run stability.

© 2010 American Association for Clinical Chemistry

Thyroid disease is an important health problem, with a prevalence of 4%–8% (1–3). Although laboratory testing is an essential tool for diagnosis and management of thyroid disease (4, 5), the standardization status and quality of performance of thyroid hormone assays remains an issue of concern (e.g., (6, 7)). For that reason, the IFCC Scientific Division established the Working Group for Standardization of Thyroid Function Tests (WG-STFT).11 The tests covered by this group are total thyroxine (TT4), total triiodothyronine (TT3), free thyroxine (FT4), free triiodothyronine (FT3), and thyroid-stimulating hormone (TSH). The first phase of the project intended to investigate the status of result comparability and assess the key performance attributes of the individual assays for the above tests. Immunoassay manufacturers (for FT4 and FT3, also mass spectrometry laboratories) were invited to participate in a method comparison study performed with a panel of 40 sera selected from 200 apparently healthy donors, based on their FT4 values. The target values for TT4, TT3, FT4, and FT3 were determined at Ghent University with isotope dilution mass spectrometry. For TSH, the overall method mean was used as target. All measurements were made in accordance with a study-defined measurement protocol. In this first part, the WG-STFT reports on the method comparison for...
TSH (for FT4/FT3 and TT4/TT3, we refer to parts 2 and 3, respectively).

Materials and Methods

SERUM PANEL
We selected donors from screening of 200 apparently healthy subjects (Solomon Park Research Laboratories, www.solomonpark.com) to obtain a panel of 40 single-donor sera with a maximum range of FT4 values. We did not screen the sera for antithyroid peroxidase and antithyroglobulin. The panel was produced according to the C37-A protocol from the Clinical and Laboratory Standards Institute (except that the serum was not filtered or pooled) (8). Details including an overview of donor characteristics can be found in the common Data Supplement that is referenced and shared by all 3 parts of this WG-STFT report. The Data Supplement accompanies the online version of this article at www.clinchem.org/content/vol56/issue6.

METHOD COMPARISON STUDY
The study protocol required duplicate measurements in 3 runs with preferably 3 different combinations of instruments/calibrator and reagent lots (see Table 2 for manufacturer conformance to intended design), controlled with manufacturers’ internal quality control (IQC) materials (for details, see the online common Data Supplement).

STUDY PARTICIPANTS
Nine manufacturers participated in the study, with a total of 16 immunoassays. For details on the platforms and assays used, see the online Data Supplement on TSH.

DATA ANALYSIS
We investigated the status of harmonization of TSH measurements by use of weighted Deming regression analysis of each assay’s mean of 6 singlicate results per sample to the overall mean (for the reduced concentration range, after exclusion of the 2 lowest and 2 highest sample to the overall mean (for the reduced concentration range). The plot contains empirical harmonization limits of within 10% and shows the 1-sided 95% CIs of the mean, which were calculated from the SDs obtained by ANOVA from the means of the different sets of singlicate results (2 sets per run, 3 runs in total) and the respective degrees of freedom from the Satterthwaite approximation. The second graph represents, for each assay, the percentage difference of the mean of 6 singlicate results (ordinate) from the overall mean per sample (abscissa).

We investigated improvement of harmonization by mathematical recalibration of the results using the reverse regression equation of each individual assay method comparison data (x = mean of 6 singlicate results per sample, y = overall mean). Note that we restricted recalibration to the reduced concentration range. Before performing regression analysis, we examined the data for the presence of outliers by use of the Grubbs test (95% probability level) on the percentage differences of the individual assay mean results from the overall means. After removal of the outliers, we performed linear and polynomial (second- and third-order) regression analyses, followed by selection of the best-fitting equation on the basis of the highest \( r^2 \) value. Although the outliers were omitted from the regression analysis, they were included in the recalibration process.

To evaluate the quality of performance delivered by the manufacturers, we used a set of performance attributes. Some were measures of performance compared to the overall mean (correlation and difference after recalibration, both restricted to the reduced data range). Others captured more general performance characteristics such as limit of quantification, imprecision, calibration consistency, within-run stability, and between-run differences.

Each assay produced 6 sets of singlicate results (duplicate measurements in each of 3 runs). We performed third-order correlation analysis separately for each set of singlicate measurements. The individual \( r^2 \) values were averaged and used to rank the assays according to descending correlation as a measure of scatter in results. For each assay, we identified 1 set of the 6 singlicate results as being the most representative, when its \( r^2 \) value was closest to the mean \( r^2 \) for that particular assay. We estimated the influence of the variation due to assay imprecision and sample-related effects (combined random error components) by use of a difference plot in concentration units of the most representative, recalibrated (same procedure used as described above) set of singlicate results. The differences were interpreted against total error limits, as derived from the biological variation concept (total error goal within 22.8%, derived from a desirable bias within 6.9% and a desirable imprecision ≤9.7%) (9).

We investigated the limit of quantification by applying 2 criteria to the results for the 2 samples with lowest TSH concentration. We defined sample concentrations as below the limit of quantification when ei-
ther 2 of the 6 replicates were not reported or reported as 0 or when the CV of the 2 samples was >20%.

We assessed imprecision from the results for the human samples as the within-run CV (CVw) (calculated from the duplicates and pooled for the 3 runs) and total CV (CVt) (calculated from the first and second replicate of each run and pooled).

We investigated calibration consistency by means of the percentage difference of the IQC results from the target provided by the manufacturer. Some gave ranges only, in which case the mean of the range was taken as target.

We investigated stability of runs by inspecting a plot of the percentage differences of the replicates vs the sample number. We investigated systematic differences between runs by comparing the percentage differences of the means of the different runs.

For weighted Deming regression analysis, we used CBstat software (version 5.1, K. Linnet); all other statistical analysis and plotting used Microsoft Excel® 2002.

Results

STATUS OF HARMONIZATION AND INFLUENCE OF RECALIBRATION

Table 1 presents the weighted Deming regression parameters for the method comparisons. Fig. 1A shows that the mean (within the 1-sided 95% CIs) of 3 of 16 assays (C, G, K) deviated >10% from the overall mean (geometric mean of the reduced data range between 0.5 and 6.6 mIU/L). The most extreme assays, C and G, differed by 39% on average. Fig. 1B shows the realignment of the assays’ mean results after mathematical recalibration. The impact of recalibration is additionally shown in the percentage difference plots of Fig. 2, illustrating that the difference between the most-deviating assays C (triangles) and G (circles) (Fig. 2A) is substantially reduced, so that the remaining dispersion of the data was nearly entirely due to within-assay effects (Fig. 2B). Third-order polynomial regression functions gave the best curve fitting in all cases (regression was performed without 1 outlier for B and G and 2 outliers for I).

ASSAY QUALITY

Performance compared to the overall mean (data in the reduced range). We investigated scatter in results compared to the overall mean by use of third-order correlation analysis of the 6 sets of singlicate results for each assay. Ranking of the assays according to descending mean $r^2$ values for the 6 sets of singlicate results showed the highest correlation for assays K, H, and J ($r^2$ between 0.998 and 0.997) and the lowest for assays G, C, and R ($r^2$ between 0.984 and 0.972) (Table 1).

Fig. 3 contrasts the difference plots (before and after recalibration) of the most representative, recalibrated set of singicates of assay K, having the high-
The best-fitting equation was a third-order polynomial for all assays. Assays B, N, E, and R had 1 result outside the 22.8% total error limit. The complete set of difference plots before and after recalibration can be found in the online Data Supplement on TSH.

**General performance attributes.** The results varied considerably for the samples with the lowest concentration (mean 0.009 mIU/L; range 0.002–0.02 mIU/L) and the second lowest concentration (mean 0.016 mIU/L; range 0.002–0.028 mIU/L) used to evaluate performance near the limit of quantification. The CV for the lowest sample was 6% for assay K, ranged from 11% to 18% for assays P, I, J, H, and C, and was >20% for assays N, F, B, and A; for the second lowest sample, the CV ranged from 1.4% to 8% for assays J, P, H, C, N, I, and O, from 10% to 20% for assays A, B, and R, and >20% for assays F, K, and E. The 2 lowest concentrations were not reported or were reported as 0 for 2 of the 6 replicates for assays G, L, and M and for the lowest concentration for assays E, R, and O.

The imprecision data derived from the method comparison data for human samples are shown in Table 2, together with a summary of the measurement design used by the different manufacturers, i.e., the number of instruments, calibrators, and reagent lots. Because for assay G the CVt was smaller than the CVw, the latter is used in the CVt column (a widely used convention in statistics, because it is by chance that CVt < CVw). Manufacturers typically used different reagent lots in the 3 runs; however, the number of calibrators and instruments varied. Two manufacturers (J and O) performed only 2 runs. Because of these differ-
Fig. 3. Difference plots (in concentration units; reduced concentration range) of the most representative set of singlicate results in comparison to the overall mean, before (left) and after (right) mathematical recalibration. The $r^2$ value mentioned in each graph is the mean $r^2$. The dotted lines represent the biological total error goal for TSH (22.8%). Shown are the plots for assays G, C, and R, having the lowest correlation in comparison to assay K, having the highest correlation to the overall mean.
ences in experimental details between manufacturers, caution is necessary when comparing the values for total imprecision. Note that higher CVt values are typically associated with a greater number of instruments, calibrators, and reagents.

The results for the IQC samples (percentage difference from target, listed in Table 2) showed that most manufacturers used 3 different IQC samples; however, some used only 1 or 2 materials (M, C, N, O). Interestingly, mostly the same target values were used for different reagent and calibrator lots (with exceptions for M and O). The majority of manufacturers matched the IQC targets within ±5%. Some showed consistent deviations of ~5%–8% over all controls (B, G). Biases >10% at certain concentrations were observed for assays J, L, and E. Note that the high percentage difference for 1 control of assay E was due to its low concentration (0.04 mIU/L). The low control of other manufacturers had a concentration typically in the range of 0.1–0.5 mIU/L, or even higher.

We used within-run shifts and drifts in results for human samples as a measure of assay stability. As illustrated in Fig. 4, this was a problem for 7 assays, in particular, after the measurement of the midrun IQC samples (measurement sequence: start IQC, samples 1–40, midrun IQC, samples 40–1, end IQC). Shifts and/or drifts in the order of 3%–5% were seen in all 3 runs in assays C and H, in 2 runs in assays N and B, and in 1 run in assays E and P. Assay R showed a shift of 20% in 1 run.

Systematic differences between the 2 most deviating runs were 10% for assay E and 11% for assay R. These differences were also seen in the IQC data of assay E but not assay R.

Clinical aspects. Clinical interpretation of thyroid status is beyond the scope of this article. In the online Data Supplement on TSH, the correlation of TSH with FT4 and TT4 is given for information only.

Discussion

We performed a method comparison study of 16 immunoassays with 40 samples from apparently healthy

<table>
<thead>
<tr>
<th>Table 2. Measurement design, imprecision, and quality control data.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assay code</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>H</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>J</td>
</tr>
<tr>
<td>K</td>
</tr>
<tr>
<td>L</td>
</tr>
<tr>
<td>M</td>
</tr>
<tr>
<td>O</td>
</tr>
<tr>
<td>P</td>
</tr>
<tr>
<td>R</td>
</tr>
</tbody>
</table>

* Calculated from results for patient samples.
* NA, not applicable.
* As per convention in statistics, CVw was taken because CVt < CVw.
* Mean of range.
* Two runs.
* Three high-normal.
* Different targets used in different runs.
* Two calibrators in run 2.
**Fig. 4.** Shifts and drifts within a run observed in 7 assays (codes in the upper left corner).

The y axis plots for each of the 40 samples the % difference of replicate 2 (R2) in comparison to R1 [note that the samples were measured in duplicate in 3 runs; the duplicates per run were obtained by measuring the samples in ascending (1–40) and descending (40–1) order, respectively]. The diamonds stand for run 1, the squares for run 2, the triangles for run 3.
donors to assess the status of harmonization and the key performance attributes of serum TSH measurements. Despite the samples having been obtained from apparently healthy donors, 2 had suppressed TSH values (<0.4 mIU/L) and 6 had increased TSH values (>4 mIU/L).

The results showed that the most discrepant TSH assays (G and C) differed by an average of 39%, which is comparable to other reports (6, 10) and to results from a proficiency testing survey with a fresh-frozen serum material (11). Using the regression relationships for the most discrepant assays, one can translate a conventional TSH reference range from 0.4 to 4 mIU/L (5) into the ranges of the most discrepant assays, i.e., from 0.33 to 3.32 mIU/L (G) and 0.48 to 4.56 mIU/L (C). This observation indicates that harmonization of TSH assays may be particularly beneficial in view of current clinical practice discussions about lowering the concentration of the TSH decision limit (5). Harmonization would affect only 3 assays to a major extent. The harmonization goal of 10% used in this study was considered current state-of-the-art of immunoassay comparability. It may be substituted in the future by a limit of 6.9%, derived from the biological variation of TSH (9).

The study demonstrated that mathematical recalibration based on regression of each assay results on the overall mean results was able to remove the assay-specific biases, so that the remaining dispersion of the data was nearly entirely due to within-assay effects. Mathematical recalibration was possible because of the excellent (third-order) correlation of most assays to the overall mean (9.7% limit derived from the biological variation of TSH) (9). Under field conditions, however, higher values may be expected and manufacturers of assays O and R with CVt >6% may need to improve the precision. The study also demonstrated from the IQC results, more in particular from the observed deviation from the target, that 5 assays may benefit from improved performance (B, G, H, J, L). Another observation was that 2 assays (E and R) had excessive between-run variability, compared with the 6.9% maximal bias limit derived from the biological variation of TSH (9). Surprisingly, most assays demonstrated within-run instabilities that ranged from minor problems (drifts and shifts in the order of 3%–5%) to severe problems (shifts of 20%). Altogether, the combined assessment of the analytical performance attributes suggests that several assays are in need of improvement by their manufacturers, particularly the assays that scored poorly by the different criteria.

The results from this method comparison study suggest that a practical approach to harmonization of TSH assays could be to establish traceability of calibration to the mean values for a panel of serum samples. Each new panel would be measured over-
lacking with the current panel, so that the new panel’s mean values can be matched to the harmonization fixed point of the first panel, similar to the WHO process tracing the international unit. Technical details of this process still have to be developed. A limitation of the current study was the use of samples from apparently healthy donors. Therefore, the full potential impact of this approach must be further explored with a wide range of samples. As already suggested, clinical samples will be needed in view of the evidence of different glycoforms in patients with thyroid disease. If the results of these additional experiments indicate a different behavior of certain immunoassays with clinical samples in comparison to euthyroid samples, standardization would not be possible. Also interpretation of method quality parameters needs assessment over the full clinical range. This shall be addressed in the next phase of the project.

The manufacturers had several concerns regarding the proposed harmonization approach, in particular, because recalibration of an assay can entail major regulatory activities. If recalibration affects the information regarding the lower limit of detection and the upper measuring range, it may require a new assay registration. Other technical considerations include the fact that values must be changed for calibrators, control samples, and reference intervals for various populations (e.g., pregnancy, children, clinical cohorts). These changes must be globally communicated to customers and clinicians. Owing to these concerns raised by the manufacturers, the WG-STFT recognizes that, in addition to the technical aspects of harmonization, it is also important to address issues such as support or acceptance by clinicians, regulatory bodies, and organizers of Proficiency Testing/External Quality Assessment sufficiently early in the process.

In summary, the maximum differences among the 16 TSH immunoassays investigated was 39%, but only 3 of 16 assays differed >10% from the overall mean. The performance quality delivered by the manufacturers, as judged from the correlation of the results to the overall mean, was excellent for nearly all assays. Both facts taken together would allow harmonization of the assays to the overall mean of a panel of serum samples reasonably covering the measuring range. In addition, precision and calibration consistency was good for nearly all assays, but some would benefit from optimization. Surprisingly, within-run drifts/shifts were observed for a considerable number of assays, whereas run-to-run problems were seldom observed.

The next phases of the project are projected to be a proof-of-concept study of establishing calibration traceability to the overall mean, measurement of a panel of clinical samples with suppressed and increased TSH concentrations, investigation of the limit of quantification of the assays, the definition of the measurand TSH, and investigation of the acceptance of the proposed harmonized reporting approach by regulatory authorities.

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 of Potential Conflicts of Interest: Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:

Employment or Leadership: W.G. Miller, AACC and CLSI.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: None declared.
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played a direct role in the preparation and approval of manuscript.

Acknowledgments: The Chair of the IFCC WG-STFT (L.M. Thienpoint) is grateful to (in alphabetical order): M. Banci and G. Markowitz (Diasorin S.p.A., Italy); G. Baudino (bioMérieux, France); R. Janzen, L. Mansbach, and Y. Xue (Siemens Healthcare Diagnostics, USA); M. Kasai (TOSOH Corp., Japan); W. Kivens and C. Walton (Beckman Coulter Inc., USA); M. Minihan (Olympus Life Science Europa GmbH, Ireland); D. Montague (Ortho-Clinical Diagnostics, UK); F. Quinn (Abbott Diagnostics, USA); and M. Rottmann (Roche Diagnostics GmbH, Germany), acting as representatives on behalf of their organizations. The 9 organizations sponsored (all contributed equally) the study in terms of sample procurement. Their efforts to review and provide comments on the manuscript are highly appreciated. L.M. Thienpoint is also indebted to D. Stockl for the graphical and statistical handling of the data and the expert advice in the editing of the report.

References