Report of the IFCC Working Group for Standardization of Thyroid Function Tests, Part 2: Free Thyroxine and Free Triiodothyronine

Linda M. Thienpont,1* Kathleen 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 Toussaint10 for the IFCC Working Group on Standardization of Thyroid Function Tests

BACKGROUND: Free thyroxine (FT4) and free triiodothyronine (FT3) measurements are useful in the diagnosis and treatment of a variety of thyroid disorders. The IFCC Scientific Division established a Working Group to resolve issues of method performance to meet clinical requirements.

METHODS: We compared results for measurement of a panel of single donor sera using clinical laboratory procedures based on equilibrium dialysis–isotope dilution–mass spectrometry (ED-ID-MS) (2 for FT4, 1 for FT3) and immunoassays from 9 manufacturers (15 for FT4, 13 for FT3) to a candidate international conventional reference measurement procedure (cRMP) also based on ED-ID-MS.

RESULTS: For FT4 (FT3), the mean bias of 2 (4) assays was within 10% of the cRMP, whereas for 15 (9) assays, negative biases up to −42% (−30%) were seen; 1 FT3 assay was positively biased by +22%. Recalibration to the cRMP eliminated assay-specific biases; however, sample-related effects remained, as judged from difference plots with biologic total error limits. Correlation coefficients to the cRMPs ranged for FT4 (FT3) from 0.92 to 0.78 (0.88 to 0.30). Within-run and total imprecision ranged for FT4 (FT3) from 1.0% to 11.1% (1.8% to 9.4%) and 1.5% to 14.1% (2.4% to 10.0%), respectively. Approximately half of the manufacturers matched the internal QC targets within ±5%; however, within-run instability was observed.

CONCLUSIONS: The study showed that most assays had bias largely correctable by establishing calibration traceability to a cRMP and that the majority performed well. Some assays, however, would benefit from improved precision, within-run stability, and between-run consistency.

Regardless of the first-line strategy used in thyroid function testing, it is generally accepted that definitive diagnosis and/or initiation of treatment require ascertainment of the relationship between thyroid-stimulating hormone and free thyroid hormones (FTHs) (1–5). The analytical validity and standardization of free thyroxine (FT4) and free triiodothyronine (FT3) testing, however, has been a matter of ongoing debate (6–16). Many attempts have been made to resolve the debate by applying methods based on ultrafiltration and equilibrium dialysis in combination with direct quantification of the separated FTHs (2, 17). When compared, however, even these methods gave discrepant results (17, 18). Therefore, the establishment of FTH reference measurement systems was advocated (2, 17, 18). The IFCC Scientific Division established the Working Group for Standardization of Thyroid Function Tests (WG-STFT) with the task to develop reference measurement systems for THs in general and FTHs in particular. The group started with the definition of the measurand FT4 and proposed an international conventional reference measurement procedure (for simplicity further referred to as RMP) based on a specified procedure for equilibrium dialysis combined with isotope dilution–mass spectrometry (ED-ID-MS).

1 Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium; 2 Department of Clinical Biochemistry, Royal Infirmary, Glasgow, UK; 3 Stanford Medical School, Palo Alto, CA; 4 Department of Clinical Laboratory Medicine, Dokkyo University School of Medicine, Mibu, Tochigi, Japan; 5 Department of Pathology, Virginia Commonwealth University, Richmond, VA; 6 Department of Medicine, Loma Linda University School of Medicine, Loma Linda, CA; 7 Laboratoire de Neurobiologie, Centre National de la Recherche Scientifique, Marseille, France; 8 Department of Chemical Endocrinology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands; 9 Laboratory of Endocrinology, University Medical Center Utrecht, Utrecht, The Netherlands; 10 Institute for Reference Materials and Measurements (IRMM), Joint Research Centre, European Commission, Geel, Belgium.

* Address correspondence to this author at: Harelbekestraat 72, B-9000 Gent, Belgium. E-mail linda.thienpont@ugent.be.

Received November 6, 2009; accepted March 17, 2010.

Previously published online at DOI: 10.1373/clinchem.2009.140194

1 Nonstandard abbreviations: FTH, free thyroid hormone; FT4, free thyroxine; FT3, free triiodothyronine; WG-STFT, Working Group for Standardization of Thyroid Function Tests; ED-ID-MS/MS, equilibrium dialysis–isotope dilution–mass spectrometry; RMP, reference measurement procedure; cRMP, candidate RMP; IQC, internal quality control; CVw, within-run CV; CVt, total CV.
liquid chromatography/tandem mass spectrometry (ED-ID-LC-MS/MS) (19, 20). An operationally defined RMP was selected owing to the lack of sufficient experimental evidence that the hormone concentration measured after physical separation truly represents the original FT4 concentration.

In this part of the study, we report on method comparisons for FT4 and FT3 between a candidate RMP (cRMP), 2 clinical laboratory ED-ID-LC-MS/MS procedures, and 15 FT4 or 13 FT3 routine immunoassays for a panel of single-donor sera.

**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 for FT4 with a maximum range of FT4 values (37 sera were used for FT3). 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) (21). 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

**ASSIGNMENT OF TARGET VALUES**

Laboratory personnel at Ghent University (Ghent, Belgium) measured FT4 and FT3 concentrations of the serum samples with an ED-ID-LC-MS/MS cRMP (22), procedure extended to FT3. The cRMP measured the sera in 4 replicates in 4 independent runs. The expanded uncertainty \((k = 2)\) of measurement for this protocol was estimated to be 5%. For details and modifications, see the online common Data Supplement and the online Data Supplement for this article.

**METHOD COMPARISON STUDY**

The study protocol required duplicate measurements in 3 runs with preferably 3 different combinations of instruments/calibrator lots/reagent lots, controlled with manufacturers’ and MS laboratories’ internal quality control (IQC) materials (see Table 1 for manufacturer conformance to intended design).

**STUDY PARTICIPANTS**

In total, 9 diagnostic manufacturers participated, with 15 automated immunoanalyzers for FT4 and 13 for FT3. For details on the platforms and assays used, see the online Data Supplement.

In addition, 2 laboratories that offer FTH routine clinical services with ED-ID-LC-MS/MS were included (1 performed FT4 and FT3, the other only FT4) (23, 24).

**DATA ANALYSIS**

Visual inspection of the data in a scatter diagram led to exclusion of the sample with the lowest concentration (remaining range referred to as the reduced range). In this article, we used a combination of 3 plots to summarize the status of standardization of FT4 and FT3 assays. One plot displays the mean concentration by each assay in comparison to the mean by the cRMP within an empirical standardization limit of 10% (ordinate). For all assays, the 1-sided 95% CIs of the mean are shown. They 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 other 2 graphs are a scatter and percentage difference plot of the individual sample results by the different assays. The \(x\)-axis of each plot represents the mean by the cRMP, and the \(y\)-axis, each assay’s mean of 6 singlicate results and the percentage difference of this mean compared to the mean of the cRMP. Although the scatter and difference plots show the complete concentration range, only the reduced range was used for Deming regression analysis. The regression equations/lines are shown for the assays whose relationships with the cRMP were the most extreme.

We investigated the improvement of standardization by mathematical recalibration of the results for each assay (mean of 6 singlicate results; reduced concentration range). We achieved recalibration by use of the reverse relationship (Deming regression equation) between the means by the individual assay (\(x\)) and the means by the cRMP (\(y\)). Regression analysis was preceded by examination of each assay’s data for outliers by the Grubbs test (95% probability level) on the percentage differences of the means from the means of the cRMP. The outliers were removed from regression analysis, but not from recalibration.

We applied various methods to evaluate the quality of performance of the assays. We used correlation analysis and difference analysis after recalibration as measures of the performance compared to the cRMP, whereas imprecision, calibration consistency, within-run stability, and between-run differences were used as more general performance measures. The \(r^2\) values, obtained from Pearson correlation analysis for each set of singlicate measurements (note that 6 sets of singlicate results were produced per assay), were averaged, removing, where necessary, any series giving (a) grossly discordant \(r^2\) value(s). The mean \(r^2\) was then used to
rank the assays according to descending correlation. The most representative set of singlicate data for each assay was selected based on its $r^2$ value being closest to the mean $r^2$. We used the difference in concentration units of the selected, recalibrated (same procedure used as described above) set of singlicate results ($y$) plotted against the mean of the cRMP ($x$) to estimate the influence of the variation due to assay imprecision and
sample-related components (i.e., the combined random error effects). The observed differences were interpreted against total error limits of 11% for both FT4 and FT3. These limits were based on the allowable total error from the biologic variation concept [total error goal 9.9%, from desirable imprecision 3.8% (FT4) and 4% (FT3); desirable bias 3.6% (FT4); for FT3, the same 3.6% bias goal was taken, because of the nonavailability of between-subject biological variation data in the cited reference], but expanded to 11% for the imprecision of the cRMPs (25).

We derived imprecision from the results of patient samples as 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 assessed the calibration consistency from the percentage difference of the IQC results from the target. Stability of runs was investigated by visual inspection of a plot of the percentage differences of the replicates vs the sample number. Systematic differences between runs were investigated by comparing the percentage differences of the means of the different runs.

For Deming regression analysis, we used CBstat software (version 5.1, K. Linnet, www.cbstat.com); all other statistical analysis and plotting were done with Microsoft Excel® 2002 (office.microsoft.com).

Results

STATUS OF STANDARDIZATION AND INFLUENCE OF RECALIBRATION

Figure 1 graphically summarizes the results of the method comparison. Note that to maintain confidentiality of results, the codes were randomized resulting in no relationship in the coding schemes between FT4 and FT3. Figure 1A (FT4) (range 12–29 pmol/L by the cRMP) shows that of the 17 assays investigated, 2 (B and E) had means that differed by <10% from the mean of the cRMP, whereas all other assays provided considerably lower measurement results. The Deming regression lines and the equations of the most deviating assays are given in the scatterplot depicted in Fig. 1B. The percentage difference plot (Fig. 1C) highlights the differences in agreement between the assays, varying from good agreement with the cRMP (assay E, triangles) to an average deviation of –42% (assay D, circles).

Figure 1D (FT3) (range 4–7 pmol/L by the cRMP) shows that 4 of the 14 assays (D, F, G, and J) had a bias with respect to the cRMP of <10%, that the mean of 1 assay (K) just exceeded the –10% limit, and that the means of all other assays deviated considerably, mostly in the negative direction, except for assay O, which had a pronounced positive bias. Figure 1E shows the regression plots for the means of individual samples and the Deming equations for the assays with largest differences from the line of identity. The percentage difference plot in Fig. 1F shows that the most biased assays differed on average by –30% (assay A, circles) and +22% (assay O, triangles) from the cRMP.

Figure 2 (A, FT4; B, FT3) illustrates that mathematical recalibration of the results (after removal of 1 outlier for FT4 assays A, C, E, H, I, M, and N and FT3 assays F, H, and J) eliminated the biases. This removal of bias can be appreciated by comparison of the location of the circles and triangles in Fig. 2A and B with Fig. 1C and F. Note also that the average between-assay variation postrecalibration as calculated for the individual samples ranged from −8.5% to +8.9% for FT4 and from −11.5% to +16.5% for FT3 (for details, see the online Data Supplement and online Supplemental Tables 4 and 5). After recalibration, however, large sample-related effects remained. For FT4, they were most pronounced for sample 8 in several assays; for FT3, 4 samples with strongly positive deviations up to 59% were seen in assays D and F (indicated by the squares in Fig. 2B, owing to potential clotting problems discussed below, but these effects should be interpreted with caution). As shown in online Supplemental Fig. 6, it was not possible to use this mathematical approach to recalibrate FT3 assay D, because the Grubbs test did not identify the problematic samples as outliers (for more explanation, see the online Data Supplement). In the same figure, however, the effect of recalibration of the assay after visual selection and removal of outliers is shown.

ASSAY QUALITY

Performance compared with the cRMP. We applied ranking of the assays according to decreasing mean $r^2$ values as a measure of scatter in results compared with the cRMP (for details, see Table 1 and online Supplemental Tables 2 and 3). For FT4, the mean $r^2$ values ranged between 0.92 (Q) and 0.78 (G), and for FT3, between 0.88 (A and M) and 0.30 (N). For assays D and F, $r^2$ was <0.30 owing to sample-related effects, most probably caused by clotting problems (see below).

Difference plots of the most representative, recalibrated sets of singlicate results for assays with the highest (assay Q, 0.92), a moderate (assay J, 0.88), and the lowest (assay G, 0.78) mean $r^2$ are shown for FT4 in the top panel of Fig. 3. The number of differences exceeding the goal of 11% was 2, 4, and 11 for assays Q, J, and G, respectively. The bottom panel of Fig. 3 shows analogous difference plots for the FT3 assays with the highest (assay A, 0.88), a moderate (assay E, 0.76), and the lowest (assay F, 0.06) mean $r^2$ (cautions regarding the interpretation of this difference plot are pointed out below). The
Fig. 1. A (FT4), D (FT3): Assay means for the reduced concentration range (FT4, 12–29 pmol/L; FT3, 4–7 pmol/L; both by the cRMP) (x-axis: different assays).

The dotted lines represent the mean of the cRMP within 10%; 1-sided 95% CIs of the means are shown; for the most deviating assays, also the numerical value of the mean is listed. B (FT4), E (FT3): Scatterplot (x = mean of the cRMP, y = mean of 6 singlicate results per assay) with indication of the line of equality (dotted) and the most extreme Deming regression lines/equations [assay D (FT4), assay A (FT3), circles; assay E (FT4), assay O (FT3), triangles]; all other assays are indicated with the same symbol, X. C (FT3), F (FT3): Percentage difference plot with indication of the most negatively (circles) and positively (triangles) biased assays. Note that in order to maintain confidentiality of results, the codes used to designate the assays were completely randomized, without relation between FT4 and FT3.
number of differences exceeding 11% was 2 for assays A and E and 19 for assay F. The complete set of difference plots before and after recalibration can be found in the online Data Supplement. In all cases, we achieved recalibration by use of the reverse Deming regression equation, after removal of 1 outlier for FT4 assays A, C, D F, H, and M, 2 for FT4 assays B and M, 1 for FT3 assays B, E, F, and J, and 2 for FT3 assay G.

General performance characteristics. Imprecision for the assays as estimated from the measurements of patient samples is summarized in Table 1, including the number of instruments, calibrators, and reagent lots used. For FT4 assays C and H, the CV$_w$ was used in the CV$_t$ column, because CV$_t$ < CV$_w$ (a commonly used convention in statistics, since CV$_t$ < CV$_w$ occurred by chance). For FT3, the same applies for assays G, A, and F. The table shows that the manufacturers typically used 2 to 3 reagent lots in the 3 independent runs with differing numbers of calibrators and instruments. For FT3, assay G, the 3 runs were performed with the same reagent lot, calibrator, and instrument. Because of these differences in experimental details between manufacturers, caution is necessary when comparing the values for total imprecision.

Most manufacturers and/or MS laboratories used 3 different IQC samples (results also shown in Table 1), but a few used only 1 or 2 materials. A majority of manufacturers used the same target values for different reagent and calibrator lots (except J for FT3). The IQC data showed that for FT4, approximately half of the assays matched the target within ±5%. Other assays demonstrated considerable biases over all controls (H) or at certain concentrations (P, L, J, R, Q, F, B, and G). For FT3, the performance in IQC was similar for approximately half of the assays, whereas others were unable to match the target within ±5% (G, A, E, H, C, M, I, and O).

Nine FT4 and FT3 assays showed shifts and drifts when results for the first set of patient samples were compared with the second set in a run. The problems were observed, in particular, after the measurement of the midrun IQC samples (measurement sequence: start IQC, samples 1–40 (or 1–37), midrun IQC, samples 40–1 (or 37–1), end IQC). FT4 assay C showed a shift (on the order of 2% to 4%) in all 3 runs; for assays G and H, a shift was observed in 2 runs (G, 15%–20%, H, ±5%, with a small drift), whereas assays R, F, K, and D shifted by 5%–10% in only 1 run; assay N shifted and drifted more than 10% in 1 run (from −8% to +5%). FT4 assays B and E had spuriously high results and/or showed increased variation in some runs ($r^2$ values ≤0.47); in addition, assay B had a shift between replicates in 1 run of approximately 18%.

For the FT3 assays, shifts of 5% to 8% were observed in the 3 runs for assay G, 2 runs for assay D, and 1 run for assays M, A, B, C, E, and J. A drift of approximately 10% was observed in 1 run for assay F. For graphical details, see online Supplemental Figs. 3 and 7.

Systematic differences between the 2 most deviating runs were important for FT4 assays F (11%), Q (12%), and G (21%) and for FT3 assays M (11%) and I (13%). The systematic differences had major influence on the CV$_t$ values of the assays (see Table 1).
Discussion

This method comparison between 15 FT4 and 13 FT3 immunoassay platforms and 2 FT4/1 FT3 ED-ID-LC-MS/MS procedures and the respective cRMPs suggests the need for establishing calibration traceability in most of the examined assays. The clinical impact of the current lack of standardization can be appreciated by calculating with the generated Deming regression equations that FT4 (FT3) concentrations in the range of 12–29 pmol/L (4–7 pmol/L) by the cRMP would translate to FT4 (FT3) values ranging from 7–16 pmol/L for assay D (3–5 pmol/L for assay A) and 12–28 pmol/L for assay E (5–9 pmol/L for assay O). This overlap in ranges shows that results far outside the range of 1 assay are well within that of another. The observation is concordant with other studies that also have reported considerable differences between FT4 assays (the reported means of the reference intervals typically ranged from 13 to 18 pmol/L), and between FT3 assays (range of means from 3.6 to 6.4 pmol/L) (26–28).

As shown in Fig. 2, mathematical recalibration of results was able to remove the systematic deviations of the various assays from the cRMP, but note that it drastically changed the magnitude of reported results for the majority of assays. The postrecalibration dispersion of the data was nearly entirely due to combined random error components. Sample-related effects were the dominating factor, with sample 8 in many FT4 assays (see also online Supplemental Fig. 2). No association could be found between the aberrant behavior of this sample and the results for the biochemical testing done (see Table 1, online Common Supplement; note that a more in-depth investigation of potential causes, such as dysglobulinemia, was beyond the scope of the study). Also for FT3, severe problems remained after recalibration for 2 assays (D and F), which exhibited grossly elevated results for 4 sera. To exclude that the elevated results were real in the analyzed sample aliquots (e.g., because of errors in labeling, contamination, bad storage conditions, etc.), manufacturers sent the used/remaining vials to Ghent University. Analysis by the cRMP confirmed the expected (nonelevated) concentrations. The observed effect was most probably due to the clotting process used in the C37-A protocol (for a short description, see online Common Supplement) (21). Therefore, that protocol, originally developed for cholesterol analysis, needs verification for FT3.

Correlation analysis and the number of differences outside the total error limits of 11% after mathematical recalibration (see Fig. 3) showed that the majority of

![Fig. 3. Difference plots in concentration units, after mathematical recalibration of the most representative sets of singlicate results for the assays with the highest [Q (FT4), A (FT3), left], a moderate [J (FT4), E (FT3), middle], and the lowest [G (FT4), F (FT3), right] mean correlation coefficients to the cRMPs (dotted V-shaped lines: biological total error limits of 11%).](image)
the FTH assays were capable of good performance in comparison to the cRMP. However, for 6 of the FT4 assays (P, I, B, R, F, G) and 3 of the FT3 assays (H, O, N) the impact and need for improvement of excessive random error components was obvious, particularly when these assays are used under more challenging conditions, as may be expected for a more diverse range of samples (e.g., from sick patients). Total assay imprecision, when judged against the biological CV limit (3.8% for FT4; 4% for FT3) (25), may need improvement for several of the FT4 (Q, F, E, B, G) and FT3 (I, O) assays, as they had CV, values >7%. Note that in FT4 assays F, Q, and G and FT3 assays M and I, CV, was highly influenced by systematic differences between runs. The rather tight biological bias limit (4%) suggests that improved control of between-run consistency of calibration and stability is needed for some assays. The spuriously high FT4 results observed for assays B and E are of particular concern, because they may be falsely interpreted as hyperthyroidism. Taken together, these data demonstrate that validation of the standardization status for an assay should be combined with assessment of its main analytical performance attributes. Calibration traceability without acceptable performance at the level of the individual sample is not sufficient.

There were several limitations of this pilot study that cannot be overlooked. One was in the application of the CLSI protocol, which is considerably different from a typical blood drawing protocol. Therefore, the sample-related effects observed with 2 FT3 assays emphasize the need to revisit the validity of this protocol. Another limitation was in the use of samples from apparently healthy donors only, without information on medication that potentially influences the distribution of thyroid hormones with their binding proteins. On the other hand, the C37-A protocol is currently the most appropriate one for collection of single donations in high volumes, and at least for this phase of the study, the missing information was not relevant. Therefore, it was the Working Group’s opinion that the samples were adequate to serve the purpose of the pilot study, but in view of these potential limitations the Working Group cautioned against overinterpretation of results. It will be a priority for the Working Group to resolve these limitations for future activities. The next study will be performed using a panel of samples from healthy individuals and patients with thyroid disorders or clinical conditions known to challenge FTH assays (2, 16, 18). However, the performance of assays when they are applied to samples from diseased subjects is difficult to evaluate. For example, an older study reported that, when compared to controls, samples from patients with familial dysalbuminemic hyperthyroxinemia were measured with a negative bias, whereas samples from patients with nonthyroidal illness deviated positively (29). This observation implies that standardization would not resolve such discrepancies. Opposed to this, a more recent study showed absence of systematic differences between an immunoassay and an ED method when ambulant, hospitalized, and pregnant subjects were investigated (9). Other studies of methods applied on typical patient categories are difficult to interpret, because the observed bias may be confounded by the difference in FT4 concentrations (e.g., (26)).

In conclusion, this pilot method comparison demonstrated the need for and feasibility of standardization of FTH assays by establishing calibration traceability to the proposed international conventional RMP. It also showed that a majority of assays had acceptable quality of performance when measuring samples from nondiseased individuals; however, some assays would benefit from improved precision, within-run stability, and between-run consistency. The observed limitations should be used as guide to improve the approach of future studies. In view of the significant change in values that will be introduced by standardization, the implementation should be coordinated between method developers and health care providers/receivers.
Standardization of FT4 and FT3 Measurements

References