Progress in standardizing and harmonizing thyroid function tests[1,2]

*James D Faix[3] and W Greg Miller[4]**

[3]Montefiore Medical Center, Bronx, NY; and [4]Virginia Commonwealth University, Richmond, VA

ABSTRACT

Iodine is an essential component of thyroid hormone. Because thyroid hormone synthesis is affected by iodine deficiency on the one hand and by excess iodine intake on the other, thyroid function biomarkers may be useful for assessing iodine status and studying the effects of iodine supplementation. However, reference intervals for some of the most useful thyroid function biomarkers, including serum concentrations of thyroid-stimulating hormone (TSH), free thyroxine (FT4), and thyroglobulin, vary widely due to variability in the commercially available immunoassays for these tests. Recognizing the need for standardization of thyroid function testing, the International Federation of Clinical Chemistry and Laboratory Medicine established a working group, later restructured as the Committee for Standardization of Thyroid Function Tests, to examine its feasibility. The committee has established a conventional reference measurement procedure for FT4 and an approach to harmonization of results for TSH. Panels of single-donation human blood specimens that span the measuring interval of the immunoassays were used to assess the performance of commercially available immunoassays and form the basis for their recalibration. Recalibration of the manufacturers' methods for both FT4 and TSH has shown that the variability among immunoassays can be successfully eliminated for euthyroid individuals as well as for patients with thyroid disease. The committee is not investigating the standardization of thyroglobulin at the present time. *Am J Clin Nutr* 2016;104(Suppl):913S–7S.

Keywords: clinical laboratory tests, harmonization, international, iodine, standardization

INTRODUCTION

Although urinary iodine concentration is a sensitive indicator of recent iodine intake, laboratory tests that detect abnormal thyroid function may be more useful for identifying individuals with chronic iodine deficiency or excessive iodine intake and for monitoring the effects of iodine supplementation. Among the most helpful of these measures are serum concentrations of thyroid-stimulating hormone (TSH)[5], serum concentrations of the thyroid hormone thyroxine (T4) as total T4 and free (i.e., unbound) T4 (FT4), and serum concentrations of the thyroid protein thyroglobulin, which plays an important role in the synthesis of thyroid hormone, including the iodination steps. Other helpful measures include serum concentrations of the thyroid hormone triiodothyronine (T3) as total T3 and free T3 (FT3). Most of the thyroid hormone in the body is stored and circulated as T4; T3 is the active form.

TSH is regulated by circulating FT4. Moderate to severe iodine deficiency may produce lower serum FT4 and consequent elevation of serum TSH. If the iodine deficiency is prolonged, goiter may develop and there may be an increase in circulating concentrations of thyroglobulin (1). Excess iodine may induce hyperthyroidism in some patients, resulting in elevated FT4 and depressed TSH (2).

Evidence-based practice guidelines for the use of the above thyroid function tests have been hampered by the lack of uniformity in the methodologies available. Standardization (or an alternative approach called harmonization) will facilitate clinical studies that may be used to establish accurate thresholds for dietary iodine in terms of their effects on thyroid function.

REFERENCE MEASUREMENT PROCEDURES

In vitro diagnostic (IVD) assays are laboratory tests that are intended for use in the diagnosis of disease or the determination of the state of health and are performed on specimens taken from the human body. Standardization of IVD assays used in clinical laboratory medicine means that the calibration of all assays is traceable to a reference measurement system and, consequently, all assays give equivalent results for any patient's specimen. Ideally, such a system must include a definition of the substance to be measured (measurand), a primary (i.e., "pure") reference material, and a reference measurement procedure (RMP). This reference system allows the assignment of values to secondary reference materials, which may be used to establish and validate the calibration of individual assays and ensure the uniformity of results. [See Vesper et al. (3) in this supplement issue for more information on standardization and traceability.]

In 2005, the International Federation for Clinical Chemistry and Laboratory Medicine (IFCC) recognized the need for standardization

[1] Presented at the workshop "Assessment of Iodine Intake: Analytical Methods and Quality Control" held by the NIH Office of Dietary Supplements in Rockville, MD, 22–23 July 2014.

[2] The authors reported no funding received for this study.

*To whom correspondence should be addressed. E-mail: gmiller@vcu.edu.

[5] Abbreviations used: C-STFT, Committee for Standardization of Thyroid Function Tests; FT3, free triiodothyronine; FT4, free thyroxine; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine; IVD, in vitro diagnostic; RMP, reference measurement procedure; TSH, thyroid-stimulating hormone; T3, triiodothyronine; T4, thyroxine; WG-STFT, Working Group for Standardization of Thyroid Function Tests.

First published online August 17, 2016; doi: 10.3945/ajcn.115.110379.

of thyroid hormone testing and established the Working Group for Standardization of Thyroid Function Tests (WG-STFT) to investigate the possibility of achieving this goal. Reference measurement systems had already been established for serum concentrations of total T4 and total T3 (4, 5). The major challenge lay in developing similar reference systems for serum concentrations of TSH and FT4.

The WG-STFT agreed on a definition of the measurand for FT4 as follows: "thyroxine that is not bound to proteins in plasma or serum measured under physiologic conditions in units of pmol/L" (6). The WG-STFT also agreed to separate FT4 from protein-bound T4 by using equilibrium dialysis under defined, near-physiologic conditions of temperature and solution composition, followed by measurement of the FT4 in the dialysate using isotope-dilution liquid chromatography/tandem mass spectrometry. These measurement conditions define a conventional RMP for FT4 (7).

The WG-STFT agreed on a definition of the measurand for TSH as follows: "human TSH—intact, total measured so that different glycoforms are detected in an equimolar fashion in units associated with the existing measurement standard of mIU/mL until transition to mol/L is deemed possible" (8). The measurement standard for TSH is a WHO International Reference Preparation obtained from purified cadaver pituitary. It was deemed unlikely that an RMP for TSH would be feasible in the near future, and it was shown that the WHO International Reference Preparation was noncommutable, meaning that it does not have chemical reaction properties in immunoassays that are the same as those for clinical specimens from patients. As discussed by Vesper et al. (3), commutability is a critically important characteristic of reference materials. When standardization to an RMP cannot be achieved, agreement among measurement procedures can still be obtained through an approach called "harmonization," in which there is a reference system consisting of methods and materials that are not traceable to the International System of Units (Système International; SI) but are agreed upon as references. Consequently, the WG-STFT decided to use harmonization for TSH rather than standardization (9). The proposed harmonization approach uses a panel of single-donation human blood serum specimens that spans the TSH measuring interval of the commercially available methods. TSH values are assigned to each patient sample by using a statistically valid method to calculate the all-procedure trimmed mean. Although the WG-STFT used the term "all-procedure trimmed mean," the terms "all-method trimmed mean" and "all-method mean" are commonly used by others when referring to this summary statistic. Throughout the present article, we use "all-method mean" to describe this statistic when citing published work that uses any of these terms. Assay manufacturers can then use the value-assigned patient serum panel to recalibrate the individual methods. This approach has been well documented as a reliable way to ensure uniform results among different immunoassays (10).

The WG-STFT then embarked on a series of method comparisons for FT4, TSH, and several other tests of thyroid function, with the objective of assessing the current state of variability among commercially available immunoassays. To this end, collaborations were established with most major manufacturers, who agreed to perform the analyses. The study was designed in 3 phases. In 2012, the IFCC replaced the WG-STFT with the Committee for Standardization of Thyroid Function Tests (C-STFT); for that reason, phase 3 was performed under the aegis of the C-STFT. The 3 phases are described in the following sections.

Phase 1

The initial investigations, published in 2010, entailed analyses on a panel of sera from apparently healthy subjects, with the expectation that their thyroid function tests would place them into the euthyroid (i.e., normal) range (11–13). Concentrations of FT4 and FT3 were measured by laboratory personnel at Ghent University in Belgium using conventional RMPs for these analytes. The specimens were also measured by the individual manufacturers using different combinations of instrument and calibrator lots. The results for FT4 identified substantial variability among commercially available immunoassays, with most of the methods showing a negative bias greater than the total error goal based on intraindividual biological variability of ±10% for FT4 (**Figure 1**A) (11). However, mathematical recalibration using regression of the conventional RMP results for the individual patient specimens on the commercial methods' results successfully reduced the systemic deviations and also diminished the negative bias, drastically shifting the magnitudes of the reported results (Figure 1B) (11). There was less variability among commercially available immunoassays for TSH as a percentage of the all-method mean, although a few methods showed variation greater than the total error goal based on intraindividual biological variability of ±23% for TSH (Figure 1C) (12). Mathematical recalibration using regression of the all-method mean values for the individual patient specimens on the commercial methods' results reduced the variability (Figure 1D) (12). The WG-STFT also reported the variability of FT3, total T4, and total T3 in the same panel of sera (11, 13).

Phase 2

For the second method comparison, the manufacturers adjusted their master calibrators by using either the all-method mean values for the individual patients' specimens (for TSH) or the values assigned by the conventional RMP (for FT4) (14). This phase was considered a proof-of-concept demonstration that the ultimate goals of establishing harmonization for TSH and standardization for FT4 were feasible. Phase 2 used a serum panel with characteristics similar to the euthyroid panel used in phase 1, with the exception of a few specimens from donors judged to have probable thyroid disease on the basis of comparison with the results obtained in phase 1. **Figure 2** shows that the dispersions of test results for the phase 1 and phase 2 studies were comparable both before and after recalibration, although the manufacturers' recalibration of master calibrators in phase 2 was slightly less successful than the mathematical recalibration performed in phase 1 (14). Taking into consideration that the phase 1 and phase 2 studies were conducted 1 y apart using different lots of reagents, the WG-STFT expressed confidence in the feasibility of establishing traceability of the calibration of IVD assays for TSH to the all-method mean values and the calibration of IVD assays for FT4 to the conventional RMP values assigned to the individual patients' specimens (14).

Phase 3

The C-STFT performed the third method comparison using specimens from patients with a wide variety of thyroid diseases (15). As opposed to the previous studies, which used specimens primarily from individuals believed to be euthyroid, the goal of this study was to explore the impact of standardization and harmonization across the clinically relevant concentration range.

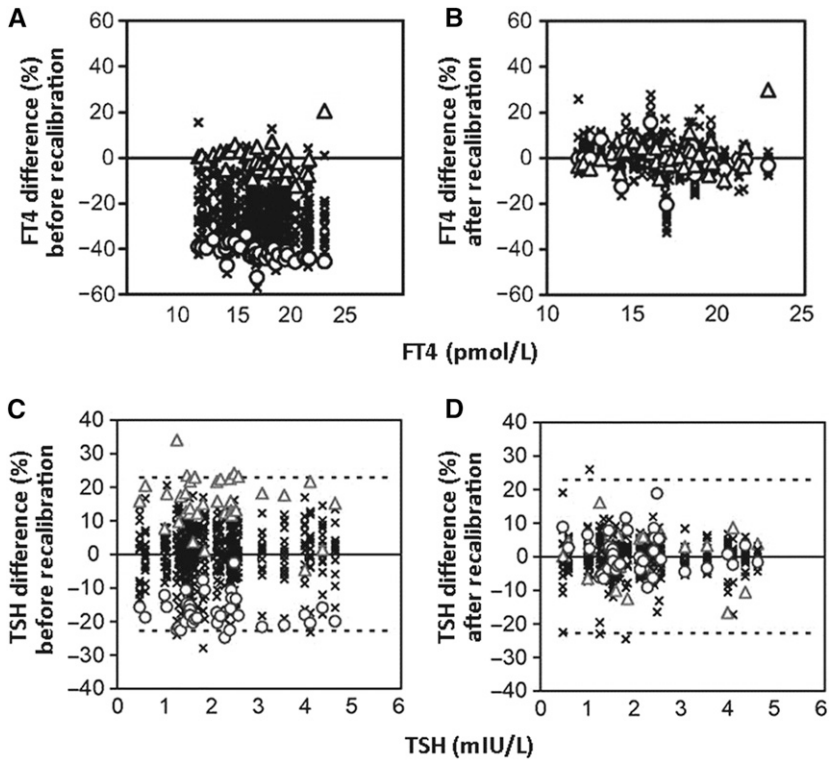

FIGURE 1 FT4 (A and B) and TSH (C and D) results for specimens from euthyroid subjects in the phase 1 study: percentage differences. FT4 measurements were obtained using immunoassay (15 procedures) and ED in combination with ID-LC-MS/MS (2 procedures); these were compared with the mean of FT4 results obtained using the conventional RMP, which was based on ED with ID-LC-MS/MS. TSH measurements were obtained using immunoassay methods (16 procedures); these were compared with values assigned to the individual patients' specimens as the all-method means of TSH results. In panels A and B, the *y* value of each data point represents the mean of 6 results for an individual FT4 assay as a percentage of the RMP mean value, and the *x* value represents the RMP mean value. FT4 data are shown before (A) and after (B) mathematical recalibration using the reverse relation (Deming regression of RMP mean values on individual-assay mean values). In panels C and D, the *y* value of each data point represents the mean of 6 results for an individual TSH assay as a percentage of the all-method mean value, and the *x* value represents the all-method mean value; the dashed lines are drawn at the total error goal based on intraindividual biological variation for TSH (±23%). TSH data are shown before (C) and after (D) mathematical recalibration. ED, equilibrium dialysis; FT4, free thyroxine; ID-LC-MS/MS, isotope dilution–liquid chromatography–tandem mass spectrometry; RMP, reference measurement procedure; TSH, thyroid-stimulating hormone. Adapted from references 11 and 12 with permission.

Serum concentrations of 10–35 pmol/L for FT4 and 0.4–8.9 mIU/L for TSH are typically considered normal (16). The concentrations spanned by the panel of specimens, 3–77 pmol/L for FT4 and 0.04–80 mIU/L for TSH, extended into ranges associated with hyperthyroidism (elevated FT4 and low TSH) and hypothyroidism (low FT4 and elevated TSH) and thus were suitably broad. The study excluded specimens from pregnant women, patients with severe nonthyroidal illness, and patients in other categories known to challenge FT4 immunoassays in ways that may be design-dependent. As in the phase 2 study, the manufacturers were allowed to perform the recalibration themselves, using their master calibrators.

When compared with the conventional RMP, FT4 concentrations were all negatively biased for concentrations in the euthyroid range (similar to the results for the euthyroid serum panel in phase 1 and the primarily euthyroid panel in phase 2) and the hyperthyroid range (>27 pmol/L). Interestingly, in the hypothyroid range (<9 pmol/L), some assays were negatively biased but others were positively biased (**Figure 3**A) (15). Regardless, manufacturers were able to eliminate most of the bias from their individual assays by recalibration to the RMP (Figure 3B) (15).

As can be seen from Figure 3C, 2 of the TSH assays were outliers with regard to the all-method mean, especially in the low TSH (hyperthyroid) range; for TSH values <0.3 mIU/L, one assay (open circles) had a negative bias of ~30−60% and the other (×'s) had a negative bias of ~25−90% (15). Although recalibration to the all-method mean made the dispersion more symmetrical for almost all of the assays and corrected the bias of the former outlier, it failed to correct the negative bias of the latter outlier (Figure 3D) (15). The overall excellent correlation of most of the TSH assays to the all-method mean led the committee to conclude that current immunoassays appear to be "glycosylation blind" (i.e., they measure TSH in an equimolar fashion regardless of differences in glycosylation). The general conclusion of the phase 3 study was that recalibration was successful in achieving uniform FT4 and TSH results, but that there may be a need to improve the approach for the low range of both measurands (15).

STANDARDIZATION OF THYROGLOBULIN IMMUNOASSAYS

Currently, measurement of serum thyroglobulin is used primarily for monitoring the treatment of patients with thyroid carcinoma. However, serum thyroglobulin concentrations are elevated in a variety of thyroid disorders, especially those in which there is active destruction of thyroid cells, such as subacute thyroiditis. Studies on the use of serum thyroglobulin for assessing iodine deficiency are

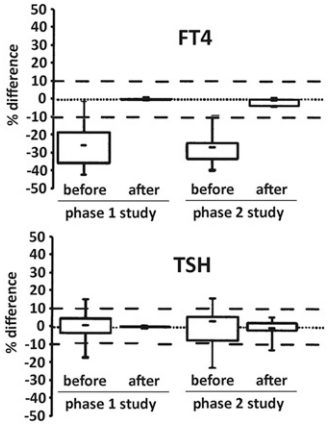

FIGURE 2 FT4 (top panel) and TSH (bottom panel) results for specimens from euthyroid subjects in the phase 1 and phase 2 studies: box plots of percentage differences. In the phase 2 study, manufacturers assayed a new panel of samples similar to the euthyroid specimens from the phase 1 study, using the conventional RMP mean results (for FT4) and the all-method mean values (for TSH) for the individual patients' specimens to readjust their master calibrators, as would be required for implementation of standardization and harmonization. For phase 1 (left-hand plots), results are shown before and after mathematical recalibration as described in Figure 1. For phase 2 (right-hand plots), results are shown before and after manufacturers' adjustment of their master calibrators. The dashed lines are drawn at arbitrary total error values of ±10%. FT4, free thyroxine; RMP, reference measurement procedure; TSH, thyroid-stimulating hormone. Adapted from reference 14 with permission.

limited, but it has been suggested as a useful way to detect and monitor goiter development (17).

The measurement of thyroglobulin in patients with thyroid carcinoma is controversial because of the high incidence of antithyroglobulin autoantibodies in such patients. These endogenous antibodies may interfere with immunoassays in a variety of ways. A novel approach to measuring thyroglobulin using proteolytic digestion followed by liquid chromatography/tandem mass spectrometry detection of a unique target peptide was recently introduced (18). This method, designed to eliminate interference by autoantibodies, could also serve as an RMP for the standardization of thyroglobulin measurement. A similar approach is being investigated by an IFCC committee for the standardization of parathyroid hormone. However, at the current time, this test is not within the scope of the IFCC committee for standardization of thyroid function tests.

THE WAY FORWARD

Although the C-STFT believes that it has shown that both the harmonization of TSH immunoassays and the standardization of FT4 immunoassays are possible, important implementation details remain to be developed. Of particular concern is the impact that recalibration will have on altering the numerical values obtained by individual immunoassays, especially for FT4. It is evident that some of the manufacturers of the assays reported in the first 3 studies (whose identities have been purposely masked) will have to make significant changes in their reference intervals.

The C-STFT is currently engaged in an open collaboration with a broad spectrum of stakeholders, including IVD manufacturers and the US Food and Drug Administration. It will need to reach out to representatives of the clinical laboratory profession, other regulatory agencies and professional societies, other manufacturers of IVD assays in addition to the ones that have supported and participated in its activities so far, and, of course, physicians and their patients. The C-STFT is also looking into

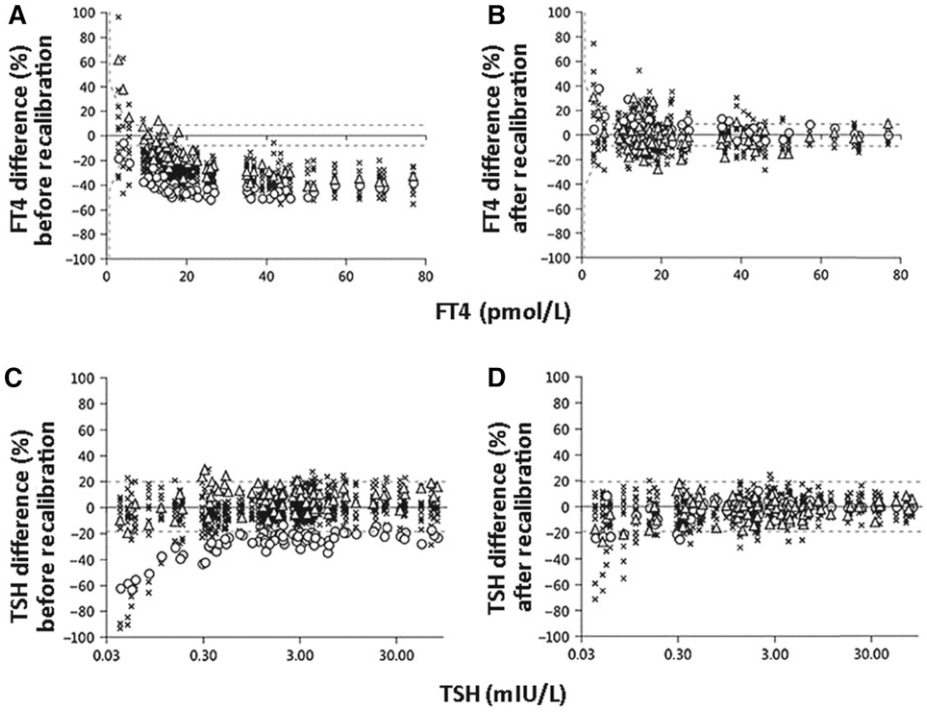

FIGURE 3 FT4 (A and B) and TSH (C and D) results for specimens from patients with thyroid disease in the phase 3 study: percentage differences. In the phase 3 study, manufacturers assayed normal and abnormal specimens both before and after recalibration procedures similar to those used in phase 2. Values are shown before (A and C) and after (B and D) mathematical recalibration as described in Figure 1. The dashed lines are drawn at the total error goal based on intraindividual biological variation of ±10% for FT4 (A and B) and the similarly derived total error goal (rounded to the nearest factor of 10) of ±20% for TSH (C and D). FT4, free thyroxine; TSH, thyroid-stimulating hormone. Adapted from reference 15 with permission.

establishing a network of clinical laboratories capable of offering the conventional RMP for FT4 and of providing an infrastructure for the procurement of serum panels; this would enable standardization and harmonization to be maintained after the initial implementation.

The C-STFT strongly believes that both the harmonization of TSH immunoassays and the standardization of FT4 immunoassays are technically feasible, important for the diagnosis and management of thyroid disease, and important to the use of thyroid biomarkers for iodine status monitoring. However, to be successful, implementation must move forward hand-in-hand with the development of educational materials for manufacturers, clinical laboratories, physicians, clinical researchers, public health scientists, and patients.

We acknowledge the leadership of Linda M Thienpont, Chair of the C-STFT and the former WG-STFT. We also thank Gay Goodman, Iodine Initiative Consultant to the NIH Office of Dietary Supplements, for expert scientific and technical review. JDF is a member of the International Federation for Clinical Chemistry and Laboratory Medicine Committee for Standardization of Thyroid Function Tests. JDF and WGM were members of the International Federation for Clinical Chemistry and Laboratory Medicine Working Group for Standardization of Thyroid Function Tests.

The authors' responsibilities were as follows—WGM: had primary responsibility for the final content; and both authors: wrote the manuscript and read and approved the final manuscript. The authors reported no conflicts of interest related to the study.

REFERENCES

1. Zimmermann MB. Iodine deficiency and endemic cretinism. In: Braverman LE, Cooper DS, editors. Werner & Ingbar's the thyroid: a fundamental and clinical text. 10th ed. Philadelphia: Lippincott; 2013. p. 217–41.
2. Roti E, Vagenakis AG. Effect of excess iodide: clinical aspects. In: Braverman LE, Cooper DS, editors. Werner & Ingbar's the thyroid: a fundamental and clinical text. 10th ed. Philadelphia: Lippincott; 2013. p. 241–56.
3. Vesper HW, Myers GL, Miller WG. Current practices and challenges in the standardization and harmonization of clinical laboratory tests. Am J Clin Nutr 2016;104(Suppl):907S–12S.
4. Hopley CJ, Stokes P, Webb KS, Baynham M. The analysis of thyroxine in human serum by an "exact matching" isotope dilution method with liquid chromatography/tandem mass spectrometry. Rapid Commun Mass Spectrom 2004;18:1033–8.
5. Tai SS, Bunk DM, White E, Welch MJ. Development and evaluation of a reference measurement procedure for the determination of total 3,3′,5-triiodothyronine in human serum using isotope-dilution liquid chromatography-tandem mass spectrometry. Anal Chem 2004;76:5092–6.
6. Thienpont LM, Beastall G, Christofides ND, Faix JD, Ieri T, Miller WG, Miller R, Nelson JC, Ross HA, Ronin C, et al; International Federation of Clinical Chemistry and Laboratory Medicine Scientific Division Working Group for Standardization of Thyroid Function Tests. Measurement of free thyroxine in laboratory medicine—proposal of a measurand definition. Clin Chem Lab Med 2007;45:563–4.
7. Van Houcke SK, Van Uytfanghe K, Shimizu E, Tani W, Umemoto M, Thienpont LM. IFCC international conventional reference procedure for the measurement of free thyroxine in serum. Clin Chem Lab Med 2011;49:1275–81.
8. 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.
9. Miller WG, Myers GL, Gantzer LM, Kahn SE, Schonbrunner ER, Thienpont LM, Bunk DM, Christenson RH, Eckfedlt JH, Lo SF, et al. Roadmap for harmonization of clinical laboratory measurement procedures. Clin Chem 2011;57:1108–17.
10. 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.
11. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri JD, Miller WG, Nelson JC, Ronin C, Ross HA, Thijssen JH, et al; IFCC Working Group on Standardization of Thyroid Function Tests. Report of the IFCC Working Group for Standardization of Thyroid Function Tests—part 2: free thyroxine and free triiodothyronine. Clin Chem 2010;56:912–20.
12. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri JD, Miller WG, Nelson JC, Ronin C, Ross HA, Thijssen JH, et al; 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.
13. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri JD, Miller WG, Nelson JC, Ronin C, Ross HA, Thijssen JH, et al; IFCC Working Group on Standardization of Thyroid Function Tests. Report of the IFCC Working Group for Standardization of Thyroid Function Tests—part 3: total thyroxine and total triiodothyronine. Clin Chem 2010;56:921–9.
14. Thienpont LM, Van Uytfanghe K, Van Houcke S; IFCC Working Group on Standardization of Thyroid Function Tests. Standardization activities in the field of thyroid function tests: a status report. Clin Chem Lab Med 2010;48:1577–83.
15. Thienpont LM, Van Uytfanghe K, Van Houcke S, Das B, Faix JD, MacKenzie F, Quinn FA, Rottmann M, Van den Bruel A; IFCC Committee for Standardization of Thyroid Function Tests. A progress report of the IFCC Committee for Standardization of Thyroid Function Tests. Eur Thyroid J 2014;3:109–16.
16. Roberts WL, McMillin GA, Burtis CA, Bruns DE. Reference information for the clinical laboratory. In: Burtis CA, Ashwood ER, Bruns DE, editors. Tietz textbook of clinical chemistry and molecular diagnostics. 5th ed. St. Louis: Elsevier Saunders; 2012. p. 2170–1.
17. Knudsen N, Bulow I, Jorgensen T, Perrild H, Ovesen L, Laurberg P. Serum Tg—a sensitive marker of thyroid abnormalities and iodine deficiency in epidemiological studies. J Clin Endocrinol Metab 2001;86:3599–603.
18. Kushnir MM, Rockwood AL, Roberts WL, Abraham D, Hoofnagle AN, Meikle AW. Measurement of thyroglobulin by liquid chromatography-tandem mass spectrometry in serum and plasma in the presence of antithyroglobulin autoantibodies. Clin Chem 2013;59:982–90.