THYROID
Volume 35, Number 5, 2025
© Mary Ann Liebert, Inc.
DOI: 10.1089/thy.2024.0728

Evaluation of the Current State of Thyroid Hormone Testing in Human Serum—Results of the Free Thyroxine and Thyrotropin Interlaboratory Comparison Study

Ashley Ribera, Otoe Sugahara, Tatiana Buchannan, Norma Vazquez, Alicia N. Lyle, Li Zhang, Uliana I. Danilenko, and Hubert W. Vesper

Background: Performance of thyroid function assays can vary significantly. To address this issue, the Centers for Disease Control and Prevention (CDC) Clinical Standardization Programs conducted an interlaboratory comparison of free thyroxine (fT4) immunoassays (IAs) and laboratory-developed tests (LDTs). This assessment aimed to determine the current performance characteristics of these assays as a first step toward measurement standardization. Thyrotropin (TSH) IAs were also evaluated.

Methods: Assays measured 41 blinded individual-donor sera, including a sample from a pregnant woman (for fT4 analysis only) and three serum pools, with 11.3–32.1 pmol/L (0.881–2.49 ng/dL) fT4 and 0.337–21.6 mIU/L TSH in duplicate over 2 days. Passing–Bablok regression analysis performed pre-recalibration compared assays performance to the CDC fT4 reference measurement procedure (RMP) or TSH all-lab mean (ALM). Additionally, the impact of linear regression-based recalibration of assays to the CDC fT4 RMP or TSH ALM was estimated. Inter-assay agreement of sample classification according to the assay-specific reference interval (RI) was assessed pre- and post-recalibration.

Results: A total of 21 fT4 and 17 TSH assays participated. Pre-recalibration, median biases of TSH measurements to the ALM were –1.2% [confidence interval or CI –1.8% to –0.4%], and good classification agreement among TSH assays was observed. fT4 assays all showed a negative median bias to the RMP, with higher bias among IAs (median: –20.3%, CI [–21.5% to –19.4%]) than LDTs (median: –4.5%, [CI –6.1% to –3.2%]). Of the individual-donor sera, only 21 out of 40 samples were classified uniformly by all fT4 assays, indicating poor inter-assay agreement. Post-recalibration, agreement improved to 33 out of 40 individual-donor sera correctly classified by all tested IAs and LDTs. Similar improvement in post-recalibration median percent bias was observed for fT4 IAs (median: –0.2, [CI –1.2% to 0.6%]) and LDTs (median: –0.3%, [CI –2.5% to 1.4%]).

Conclusions: The comparison among fT4 assays emphasizes the need for measurement standardization to improve accuracy and comparability. This and previous studies demonstrate the possibility to develop common fT4 RIs via standardization, enabling the use of evidence-based clinical guidelines universally in patient care. Recalibration can effectively address high variability in fT4 assays, ensuring consistent diagnostic classification.

Keywords: free thyroxine, thyroxine standardization, thyrotropin (TSH), immuno assays, laboratory-developed tests, interlaboratory comparison study

Introduction

Thyrotropin (TSH) and free thyroxine (fT4) are the two hormones initially measured in blood to assess hypo- and hyperthyroidism and to guide treatment decisions.[1–3] Some guidelines and recommendations suggest specific concentrations for these hormones to guide decision making.[4] Therefore, accurate and reliable TSH and fT4 tests are needed to ensure correct patient care, making the standardization of thyroid function tests to assess both

Division of Laboratory Sciences, Centers for Disease Control and Prevention (CDC), Atlanta, Georgia, USA.

thyroid gland function and therapeutic drug monitoring a priority.[5,6]

Poor accuracy and comparability of TSH and fT4 tests have been described,[7] and reference systems enabling consistent assay calibration and improving other analytical factors affecting inaccurate results have been established by the International Federation for Clinical Chemistry and Laboratory Medicine (IFCC) Committee on Standardization of Thyroid Function Tests in collaboration with the Centers for Disease Control and Prevention (CDC) Clinical Standardization Programs (CSP).[7–9] Studies assessing the analytical performance of TSH and fT4 assays after these reference systems became available are very limited. As part of the CDC's CSP for thyroid hormones, a study was conducted to obtain information about the current analytical performance of TSH and fT4 assays to assess potential improvements and to guide further standardization activities.

Materials and Methods

Materials

Forty individual-donor sera (PS), three serum pools (Pool1–3), and one sample from a pregnant woman (PN) from the third trimester used in this study were collected using protocols described in the Supplementary Data.[10,11] Use of blood by CDC is consistent with the institutional review board approval and donor consent. No personal identifiers were provided to the CDC. The blood samples used in this project were from commercial sources. This activity was reviewed by the CDC, deemed research not involving human subjects, and was conducted consistent with applicable federal law and CDC policy. Information about donors' medical history and medications was based on self-reporting. Four donors reported taking levothyroxine; however, the dosage information is unknown. The donors were not asked about thyroidectomy, and no donors reported taking biotin.

Concentrations for the PS were 11.3–32.1 pmol/L (0.881–2.49 ng/dL) for fT4 and 0.337–21.5 mIU/L for TSH. Pool1–3 concentrations were 15.0, 16.5, and 16.9 pmol/L for f(1.17, 1.28, and 1.31 ng/dL) and 1.09, 2.26, and 1.65 mIU/L for TSH. The fT4 concentration of PN was 11.4 pmol/L (0.885 ng/dL), but due to volume limitations, TSH was not measured.

Procedures for material shipment, storage, and analysis by assays are described in the Supplementary Data. The fT4 reference values were assigned to all samples using the CDC fT4 reference measurement procedure (RMP) described previously, and details can be found in the Supplementary Data.[12]

Assays included in the study

A total of 21 fT4 assays, 4 laboratory-developed tests (LDTs) based on equilibrium dialysis (ED) liquid chromatography–tandem mass spectrometry (LC-MS/MS), and 17 immunoassays (IAs) were included in the study. Among the fT4 IAs, 8 (comprising 15 different platforms) were operated by the independent assay manufacturer and 2 by clinical laboratories. Eight IA manufacturers (comprising 16 different platforms) and 1 clinical laboratory measured TSH. Several independent manufacturers were represented by more than one platform. Further information about the assays used in the study is summarized in Table 1.

Data analysis

The study design was based on the Clinical and Laboratory Standards Institute (CLSI) document EP09-A2.[13] Agreement among LDT and IA manufacturers was assessed using the 40 PS. The concentration of one PS was outside the reportable range for most assays and therefore excluded from TSH analysis. Percent bias of sample replicate means to either the fT4 reference value or the all-lab mean (ALM)

Table 1. List of Assays Included in the Interlaboratory Comparison Study

ID	*Participant type*	*Platform/method principle*	*fT4*	*TSH*
A	Manufacturer	Immunoassay	x	x
B, V	Manufacturer	Immunoassay	x[a]	x
C, D	Manufacturer	Immunoassay	x	x
E, F	Manufacturer	Immunoassay	x	x
G	Clinical laboratory	Immunoassay	x	x
H	Manufacturer	ED-based LC-MS/MS	x	
I	Manufacturer	ED-based LC-MS/MS	x	
J	Manufacturer	ED-based LC-MS/MS	x	
K	Manufacturer	ED-based LC-MS/MS	x	
L, M	Manufacturer	Immunoassay	x	x
N	Manufacturer	Immunoassay	x	x
O	Manufacturer	Immunoassay	x	x
P, Q, R, S, T	Manufacturer	Immunoassay	x	x
U	Clinical laboratory	Immunoassay	x	

A total of 22 assays, 4 LDTs and 18 IAs, performed by 14 independent entities (12 manufacturers and 2 clinical laboratories) were included in the interlaboratory comparison. Each assay was analyzed independently. Assay manufacturers that submitted data for more than one platform are indicated as multiple letters in the ID column. Twenty-one of the laboratories measured serum fT4 and 17 measured serum TSH with commercially available IAs. Four of the laboratories performed serum fT4 measurements with LDTs based on ED LC-MS/MS methods.

[a] Assay manufacturer "V" did not submit fT4 data.

ED LC-MS/MS, equilibrium dialysis liquid chromatography–tandem mass spectrometry; fT4, free thyroxine; IA, immunoassay; LDT, laboratory-developed test; TSH, thyrotropin.

for TSH, median bias, confidence interval (CI) of the median bias, and the percent coefficient of variation (CV) for replicate measurements were calculated for each assay. Outliers detected among replicate fT4 or TSH data were removed using the guidelines described in CLSI document EP09-A2.[13] Suspected transcription errors (with biases to the ALM of 585–609%) for replicate TSH measurements of the same study sample by assay "O" were removed prior to analysis. Analysis of Pool1–3 or PN sera assessing the impact of serum pooling and pregnancy on fT4 measurement was performed independently of the analysis of the 40 PS.

Included assay manufacturers (LDT and IA) and clinical laboratories were compared to the CDC fT4 RMP (or the TSH ALM) using Passing–Bablok regression, chosen to account for the measurement error in the assay and CDC RMP data without making assumptions about the distribution of residuals. The mean biases of replicate measurements of the 40 PS samples among assays were compared to the criteria for acceptable performance, the greater of 0.3 ng/dL or ±15% for fT4 and 0.2 mIU/L or ±20% for TSH, set by the Centers for Medicare & Medicaid Services (CMS) to assess method performance pre- and post-recalibration.[14] Methods for estimating the impact of pregnancy, sample pooling, and T4 supplementation on measurement accuracy are described in the Supplementary Data. Data were analyzed using Microsoft Excel® with the Analyse-it® (version 5.90) add-in and the R statistical environment in R Studio® (version 4.1.2).

fT4 and TSH in-silico recalibration

Linear regression-based assay recalibration was performed in-silico using assay PS results and the reference value (or TSH ALM) for that sample. Pooled and PN samples were not included in linear regression models. In-silico recalibration was conducted by subtracting the intercept from the original submitted data and by dividing the slope. Classification of serum samples was performed post-recalibration by comparing the recalibrated results to a unified reference interval (RI), which was calculated according to the principles of transference discussed in the CLSI document EP28-A3c.[15] Prospective transferred RIs were chosen from the RIs of assays (Fig. 1) with linear regression slopes closest to 1 and intercepts closest to 0.

Results

Comparison of fT4 measurements to the CDC RMP

Based on the data as received from assays included, the all-sample median percent bias with CI to the CDC RMP observed for the 40 PS among all assays was −17.1% [CI −18.1% to −15.7%). LDTs had better accuracy, with median bias [CI] of −4.5% [CI −6.1 to −3.2], compared to −20.3% [CI −21.5 to −19.4] for IAs (Fig. 2 and Table 2). Pre-recalibration, 3 out of the 21 fT4 assays met the CMS criteria for all PS, with 5–40 of the 40 PS meeting the bias requirement among fT4 assays (Table 3). The mean percent of samples meeting the CMS requirements [±CI] for individual fT4

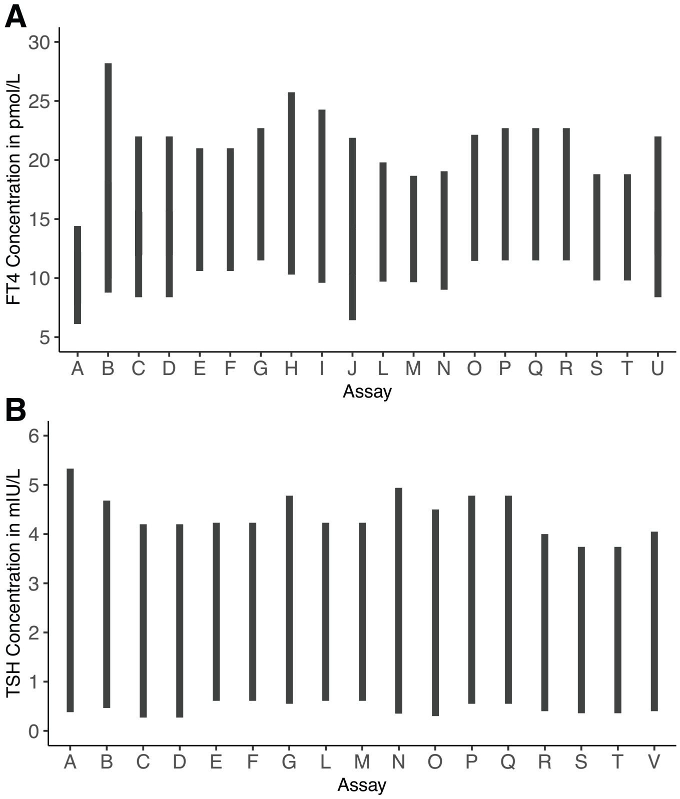

FIG. 1. Assay-specific reference intervals among included fT4 assays. Assay-specific reference intervals for assays are shown for fT4 (**A**) and TSH (**B**). Assay-specific reference intervals in picomoles per liter for fT4 and milli-international units per liter for TSH for adults are shown as dark grey bands for each of the included assays. RIs that could not be confirmed were excluded. fT4, free thyroxine; RI, reference interval; TSH, thyrotropin.

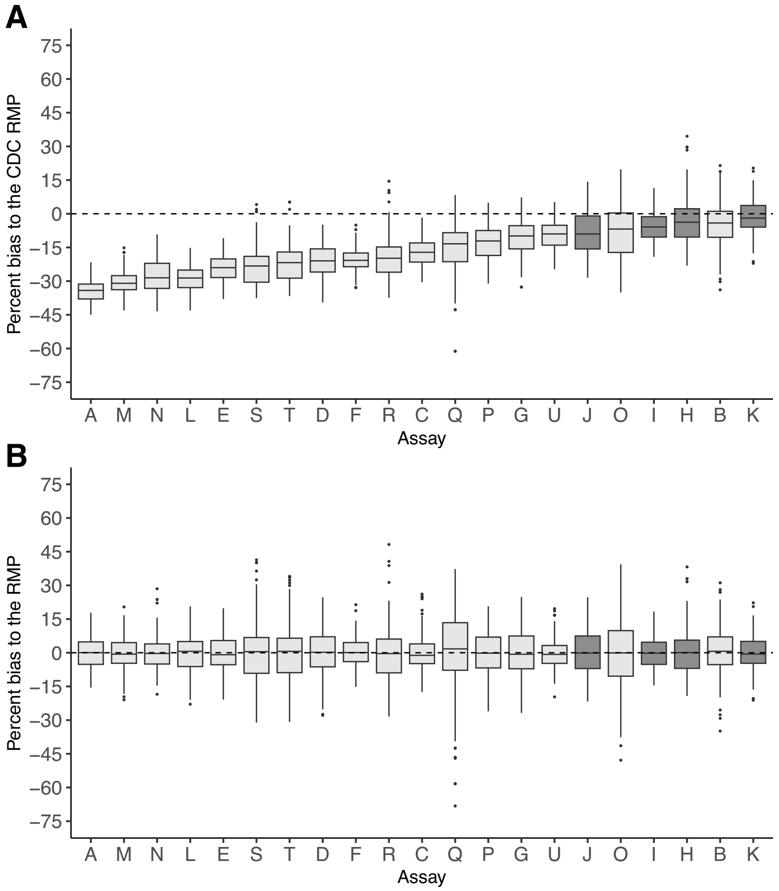

FIG. 2. fT4 distribution of biases observed among fT4 assays from the 40 single-donor study samples (PS) to the fT4 RMP. Percent biases of each assay's individual replicate fT4 measurements to the CDC RMP were calculated for all 40 PS pre-recalibration **(A)** and post-recalibration **(B)**, excluding pooled samples and samples from pregnant women. Boxplots are arranged in order of increasing median bias, indicated as the horizontal black bars, and zero bias is represented by a horizontal dashed line. Maximum and minimum biases are indicated by the upper and lower whiskers, with results beyond ±1.5 times the interquartile range shown as black dots. Darker grey boxplots indicate LDT results. Post-recalibration, a surrogate CDC fT4 RMP RI was created following the procedures recommended in the CLSI document EP28-A3c.[15] In brief, linear pairwise regression analysis of each included assay to the CDC RMP was determined (Supplementary Fig. S1). Participant "H" was selected for transference of the RI because the slope and intercept of the linear regression were closest to 1 and 0, respectively. This surrogate RI is only intended to estimate the impact of recalibration on patient classification. CDC, Centers for Disease Control and Prevention; CLSI, Clinical and Laboratory Standards Institute; LDT, laboratory-developed test; RMP, reference measurement procedure.

measurements was higher [99.0% ± 1.8%] among fT4 LDTs than IAs [58.9% ± 12.9%], which improved to 99.5% ± 1.6% for fT4 LDTs and 97.0% ± 1.9% for fT4 IAs post-recalibration. The bias distributions were narrow with an average interquartile range (IQR) across all assays of 9.0% (Fig. 2) for fT4. The fT4 bias becomes more negative with increasing fT4 concentrations for IA, while LDT bias remains constant (Supplementary Fig. S2). Most assays were well correlated with the CDC RMP with correlation coefficients >0.8 (Supplementary Fig. S1). The Passing–Bablok regression parameters among LDTs indicate well-calibrated assays with slope and intercept CIs encompassing 1 and 0, respectively. Only one of the 17 Passing–Bablok regression parameters among the IAs met the same criteria for slope and intercept (Table 2 and Supplementary Fig. S3). The mean precision (range of mean CV) was 3.8% (1.7–8.0) for all fT4 assays pre-recalibration and 4.3% (1.9–10.3) post-recalibration (Table 2). Comparing the fT4 replicate mean biases among the four PS from donors taking levothyroxine to the fT4 mean biases of the remaining 36 PS by two sample *t*-test suggests a significant difference in biases between these two groups for IAs (*p*-value <0.05) but not for LDTs (*p*-value >0.05) at the 5% significance level. No significant difference between the two groups was observed among TSH assays (*p*-values >0.05).

Comparison of TSH measurements to the ALM

The 17 included TSH IAs were in better agreement compared to fT4 assays with a median percent bias (Fig. 3 and Table 4) to the ALM of –1.2% [CI –1.8% to –0.4%]. The sample replicate mean bias distribution among all TSH assays showed a narrow 5.4% mean IQR. Out of the 17 TSH assays, 10 met the CMS measurement criteria for

Table 2. A Comparison of Assay Free Thyroxine Values to Centers for Disease Control and Prevention Free Thyroxine Reference Measurement Procedure

Assay	*Measurement principle*	*Median bias, % [CI]*	*Regression equation,* y = a *[CI]*x + b *[CI]*	*Mean CV (%)*
A				
A	IA	−34.1 [−35.3 to −33.2]	*y* = 0.569 [0.502–0.638] *x* + 1.48 [0.406–2.53]	
B	IA	−4.1 [−5.8 to −3.2]	*y* = 0.908 [0.817–1.05] *x* + 0.793 [−1.58 to 2.14]	
C	IA	−17.2 [−18.1 to −15.8]	*y* = 0.686 [0.630–0.788] *x* + 2.27 [0.762–3.20]	
D	IA	−21.0 [−22.5 to −19.3]	*y* = 0.690 [0.622–0.783] *x* + 1.70 [0.222–2.84]	
E	IA	−24.0 [−25.7 to −23.0]	*y* = 0.709 [0.636–0.793] *x* + 0.675 [−0.584 to 1.88]	
F	IA	−20.8 [−21.5 to −19.7]	*y* = 0.764 [0.699–0.827] *x* + 0.496 [−0.601 to 1.63]	
G	IA	−9.9 [−11.5 to −8.2]	*y* = 0.677 [0.533–0.771] *x* + 3.67 [2.23–5.67]	
H	LDT	−3.8 [−6.8 to −2.4]	*y* = 0.971 [0.839–1.11] *x* −0.372 [−2.50 to 1.87]	
I	LDT	−5.9 [−7.2 to −3.9]	*y* = 0.921 [0.825–1.01] *x* + 0.389 [−1.08 to 1.89]	
J	LDT	−9.0 [−10.6 to −6.4]	*y* = 0.965 [0.766–1.09] *x* − 1.33 [−3.16 to 2.05]	
K	LDT	−2.0 [−3.1 to 0.2]	*y* = 1.01 [0.912–1.10] *x* − 0.350 [−1.77 to 1.19]	
L	IA	−28.6 [−30.3 to −27.2]	*y* = 0.624 [0.550–0.723] *x* + 1.45 [−0.0400 to 2.76]	
M	IA	−31.0 [−31.7 to −29.2]	*y* = 0.567 [0.512–0.625] *x* + 2.07 [1.20–2.91]	
N	IA	−28.5 [−30.0 to −27.5]	*y* = 0.432 [0.391–0.483] *x* + 4.75 [3.77–5.38]	
O	IA	−6.8 [−10.5 to −5.1]	*y* = 0.492 [0.435–0.597] *x* + 6.91 [5.34–7.87]	
P	IA	−12.1 [−14.6 to −10.8]	*y* = 0.652 [0.545–0.774] *x* + 3.52 [1.57–5.24]	
Q	IA	−13.4 [−16.3 to −11.8]	*y* = 0.619 [0.474–0.767] *x* + 3.54 [1.39–6.04]	
R	IA	−19.8 [−22.5 to −18.2]	*y* = 0.569 [0.502–0.656] *x* + 3.70 [2.29–4.96]	
S	IA	−23.3 [−24.2 to −22.4]	*y* = 0.590 [0.485–0.703] *x* + 2.91 [1.13–4.57]	
T	IA	−21.8 [−22.9 to −20.2]	*y* = 0.620 [0.536–0.720] *x* + 2.73 [0.977–4.11]	
U	IA	−9.0 [−12.1 to −7.5]	*y* = 0.762 [0.680–0.846] *x* + 2.33 [1.17–3.42]	
All assays		−17.1 [−18.1 to −15.7]	n/a	3.8
IA		−20.3 [−21.5 to −19.4]	n/a	3.3
LDT		−4.5 [−6.1 to −3.2]	n/a	6.0
B				
A	IA	−0.4 [−2.8 to 1.1]	*y* = 0.968 [0.847–1.09] *x* + 0.580 [−1.38 to 2.46]	3.7
B	IA	−1.6 [−3.2 to 0.8]	*y* = 1.05 [0.944–1.19] *x* − 0.725 [−3.22 to 0.852]	5.5
C	IA	−1.8 [−2.1 to −0.5]	*y* = 1.05 [0.952–1.19] *x* −0.876 [−3.07 to 0.614]	2.0
D	IA	0.2 [−1.3 to 1.7]	*y* = 1.07 [0.960–1.21] *x* − 1.15 [−3.31 to 0.613]	5.2
E	IA	−2.0 [−2.7 to −0.5]	*y* = 1.01 [0.905–1.13] *x* − 0.403 [−2.12 to 1.42]	3.8
F	IA	0.4 [−1.1 to 1.1]	*y* = 1.02 [0.932–1.11] *x* − 0.394 [−1.83 to 1.22]	2.9
G	IA	−0.7 [−2.7 to 2.2]	*y* = 1.18 [0.929–1.36] *x* − 2.77 [−5.58 to 0.808]	2.9
H	LDT	0.0 [−3.0 to 1.8]	*y* = 0.976 [0.843–1.13] *x* + 0.120 [−2.32 to 2.26]	5.0
I	LDT	−0.3 [−1.9 to 1.6]	*y* = 0.989 [0.882–1.08] *x* + 0.233 [−1.37 to 1.83]	4.2
J	LDT	−0.6 [−3.2 to 1.1]	*y* = 0.984 [0.788–1.11] *x* − 0.212 [−2.10 to 3.19]	7.0
K	LDT	0.1 [−1.7 to 1.8]	*y* = 1.03 [0.942–1.13] *x* − 0.442 [−1.94 to 1.03]	5.3
L	IA	0.6 [−2.6 to 2.2]	*y* = 1.07 [0.939–1.24] *x* − 1.21 [−3.58 to 1.06]	2.2
M	IA	−0.8 [−1.4 to 0.7]	*y* = 1.05 [0.937–1.16] *x* − 0.701 [−2.44 to 0.950]	2.1
N	IA	0.3 [−1.6 to 1.6]	*y* = 1.03 [0.933–1.14] *x* − 0.496 [−2.60–0.908]	2.6
O	IA	0.5 [−1.8 to 2.5]	*y* = 1.12 [0.970–1.38] *x* − 2.42 [−5.99 to 0.254]	9.2
P	IA	−0.7 [−2.6 to 1.5]	*y* = 1.05 [0.883–1.30] *x* − 0.813 [−4.54 to 1.75]	2.7
Q	IA	2.7 [−0.5 to 5.4]	*y* = 1.33 [1.02–1.64] *x* − 5.78 [−10.2 to −0.483]	10.3
R	IA	−1.8 [−4.2 to 0.4]	*y* = 1.08 [0.945–1.23] *x* − 1.63 [−4.04 to 0.911]	6.8
S	IA	0.2 [−2.7 to 2.2]	*y* = 1.10 [0.937–1.29] *x* − 1.61 [−4.61 to 1.09]	1.9
T	IA	−0.4 [−3.0 to 1.1]	*y* = 1.11 [0.966–1.30] *x* − 1.65 [−4.78 to 0.630]	2.4
U	IA	−0.7 [−2.8 to 0.7]	*y* = 1.07 [0.951–1.19] *x* −1.13 [−3.05 to 0.400]	2.7
All assays		−0.2 [−1.1 to 0.5]	n/a	4.3
IA		−0.2 [−1.2 to 0.6]	n/a	4.1
LDT		−0.3 [−2.5 to 1.4]	n/a	5.4

Median bias and confidence interval (CI) of bias, Passing–Bablok regression equation, and average coefficient of variation (CV) were determined with 40 individual donor sera for fT4, pre-recalibration (**A**) and post-recalibration (**B**) among immunoassay (IA) and laboratory-developed tests (LDT).

n/a, not available.

all PS, with 31–39 of the 39 PS meeting the bias requirement among TSH assays (Table 3). Results of Passing–Bablok regression analysis indicate 5 out of the 17 TSH assays are well-harmonized to the ALM, with CIs of slope and intercept including 1 and 0, respectively (Table 4 and Supplementary Fig. S6). The mean precision (range of mean CV) was 2.6% (1.0–5.6) for all TSH assays pre-recalibration and 2.2% (0.9–5.4) post-recalibration (Table 4).

Table 3. Percent of Sample Replicate Mean Biases Meeting Center for Medicare Services Acceptable Bias Criteria for Free Thyroxine and Thyrotropin for Each Assay Tested

		fT4 (%)		*TSH (%)*	
Assay	*Measurement principle*	*Pre-recalibration*	*Post-recalibration*	*Pre-recalibration*	*Post-recalibration*
A	IA	12	100	100	100
B	IA	100	98	97	97
C	IA	75	98	100	97
D	IA	55	100	79	100
E	IA	45	100	97	97
F	IA	60	100	97	97
G	IA	82	95	100	100
H	LDT	98	98	n/a	n/a
I	LDT	100	100	n/a	n/a
J	LDT	98	100	n/a	n/a
K	LDT	100	100	n/a	n/a
L	IA	32	100	100	100
M	IA	25	100	100	100
N	IA	30	98	95	100
O	IA	85	95	100	100
P	IA	78	92	100	100
Q	IA	70	88	100	100
R	IA	58	98	92	100
S	IA	48	92	100	100
T	IA	55	95	100	100
U	IA	92	100	n/a	n/a
V	IA	n/a	n/a	97	97
Mean of all IAs ± CI		58.9 ± 12.9	97.0 ± 1.9	97.3 ± 2.7	99.1 ± 0.7
Mean of all LDTs ± CI		99.0 ± 1.8	99.5 ± 1.6	n/a	n/a
Mean of all assays ± CI		66.6 ± 12.5	97.5 ± 1.6	97.3 ± 2.7	99.1 ± 0.7

Absolute biases of the assays' replicate measurements to the CDC RMP (or all-lab mean for TSH) we calculated among all non-pregnant, single donor samples and compared to the CMS criteria for acceptable assay performance among tested immunoassays (IAs) and laboratory-developed tests (LDTs).[14] The percentages of the 40 (39 for TSH) samples meeting these criteria for each lab are shown pre- and post-recalibration. A value of "n/a" indicates a lack of reported data for that assay and analyte.

Comparison of end user and manufacturer data for fT4 and TSH

Agreement between the manufacturer and the end user results varied. For IA manufacturer "P," Passing–Bablok regression analysis of fT4 data indicated no significant difference between manufacturer "P" and end user ("G") with CIs of slope and intercept including 1 and 0, respectively (Fig. 4). This contrasts with the comparison of manufacturer "P" to the clinical lab ("G") for TSH measurement (Fig. 4) and the comparison of fT4 manufacturer "C" to the clinical lab "U" (Fig. 4), where the CI of the slope parameter is >1. Use of the same lots of calibrators and reagents was confirmed for fT4 measurement between assays "P" and "G"; however, different calibrator and reagent lots were used in TSH assays "P" and "G." Reagent and calibrator lot numbers could not be determined for fT4 assays "U" and "C."

Improvements in sample classification after in-silico recalibration

After recalibration of the fT4 assays, classification was repeated using the recalibrated PS sera results and the RI transferred from assay "H" (9.75–25.1 pmol/L or 0.758–1.95 ng/dL). Classification of TSH results post-recalibration was performed in a similar way, by first performing transference testing using the ALM for comparison, and then reclassifying assay results using the transferred RI from TSH assay "M" (0.606–4.32 mIU/L).

Agreement among fT4 assays on sample classification was poor before recalibration; 21 out of the 40 PS were classified uniformly by all assays (Table 5). fT4 classification agreement improved upon recalibration to 33 out of the 40 PS uniformly classified by all assays. TSH classification agreement was consistent pre-recalibration (33 out of the 39 PS) and post-recalibration (32 out of the 39 PS) (Table 6). Post-recalibration fT4 assay median percent biases (CI) improved to −0.2% [CI −1.1 to 0.5] overall, −0.2% [CI −1.2 to 0.6] for IA, and −0.3% [CI −2.5 to 1.4] for LDT (Fig. 2). Median percent bias among TSH assays improved to 0.5% [CI 0.0–0.8] post-recalibration (Table 4).

Agreement between high-quality individual donor and pooled materials for fT4 and TSH assays

Pool1–3 sera were found to be comparable to individual donor units among all IAs. The median percent biases (−21.0% for fT4 and 0.5% for TSH assays) of these samples were similar to the −20.3% median percent bias for fT4 and −1.2% median percent bias for TSH IAs measuring the 40 PS (Table 7).

IAs generally reported higher values (−1.9% median percent bias to the RMP, [CI −27.8 to 19.8]) for the PN compared to the −20.3% median bias [CI −21.5 to −19.4] of PS materials (Table 7). This increase occurred despite assay-

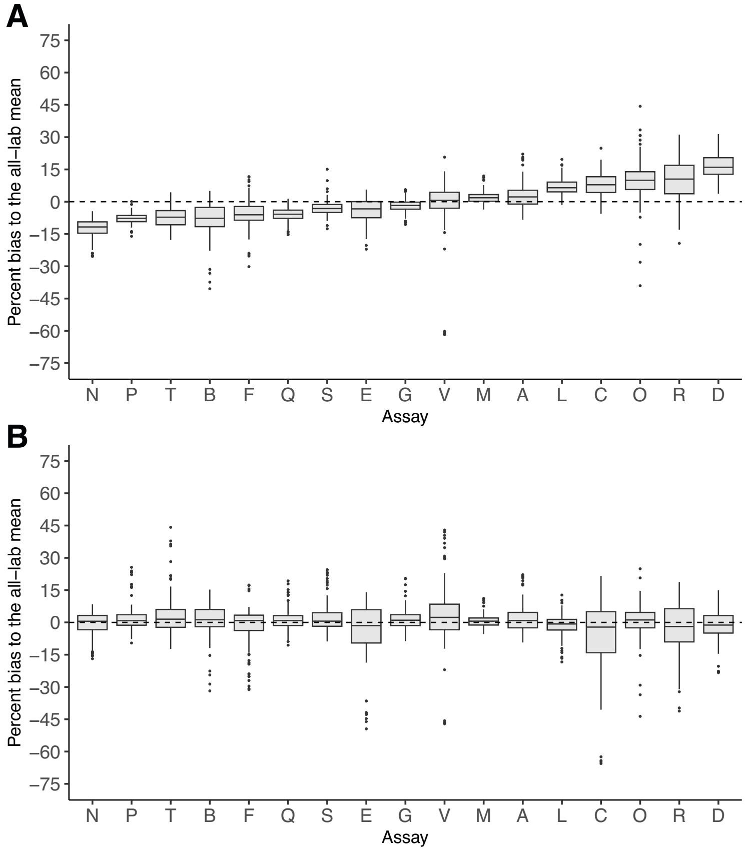

FIG. 3. Distribution of TSH biases to the all-lab mean. Mean percent biases of each IA's replicate TSH measurements to the mean of biases reported for all labs were calculated for all 39 study samples pre-recalibration (**A**) and post-recalibration (**B**). Boxplots are arranged in order of increasing median bias, indicated as the horizontal black bars, and zero bias is represented by a horizontal dashed line. Maximum and minimum bias are indicated by the upper and lower whiskers, with results beyond ±1.5 times the interquartile range shown as black dots. Post-recalibration, a unified TSH RI was estimated following the procedures recommended in the CLSI document EP28-A3c.[15] In brief, linear pairwise regression analysis of each included assay to the all-lab mean was determined (Supplementary Fig. S4). Assay "M" was selected for transference of the RI because the slope and intercept of the linear regression were closest to 1 and 0, respectively. This unified RI is only intended to estimate the impact of recalibration on patient classification. IA, immunoassay.

specific RI being lower in general during pregnancy than the adult RI.[16]

Discussion

To assess the need for standardization and monitor improvements in fT4 measurement, it was imperative to collect updated information on the status of fT4 measurements by IAs and LDTs to support the IFCC efforts. The conducted interlaboratory comparison study demonstrated high variability among commercial fT4 IAs. The large negative median assay biases of up to −34.1% to the fT4 RMP observed in this study can lead to patient misclassification as indicated by only 21 of 40 samples used in the study being classified uniformly. The observation of negative bias for IAs is consistent with the results of the College of American Pathologists (CAP) Harmonized Thyroid (ABTH) survey that uses high-quality pooled serum materials with fT4 reference values assigned by the CDC fT4 RMP. Recent ABTH 2022-B and 2023-A surveys reported sample-specific and calibration measurement inconsistencies among IAs measuring six pooled samples, with IAs underestimating fT4 concentrations and LDTs being in better agreement with the RMP.[17,18] The results of ABTH surveys are consistent with our findings despite differences in the number of samples measured and the evaluation of multiple assay's measurement results in peer groups in the CAP surveys.[17,18] The results of a previous fT4 interlaboratory comparison study conducted by IFCC and performed at Ghent University in 2017 were compared to the present study to determine what potential improvements in accuracy have been made, although some of the assays in both studies differed.[7] Comparing the median percent bias for fT4 measurements of samples in the 10–25 pmol/L (0.8–1.9 ng/dL) range among all assays that participated in the 2017 study and all assays included in the present study showed an improvement in median percent bias (which is calibration-related) from −37.7% in 2017 to −16.8% in 2022.

In many situations, while assay-specific RIs are assumed to overcome issues with data interpretation, their use did not prevent misclassification, especially for samples near cutoff values. In the example provided (Table 5), two samples with

Table 4. Comparison of Assay Thyrotropin Values to All-Lab Mean

Assay	*Median bias, % [CI]*	*Regression equation,* y = a *[CI]* x + b *[CI]*	*Mean CV (%)*
A			
A	2.2 [1.1–3.2]	*y* = 1.03 [0.997–1.07] *x* – 0.00992 [–0.0571 to 0.0324]	2.7
B	–7.7 [–8.8 to –6.0]	*y* = 0.972 [0.922–1.01] *x* – 0.0662 [–0.119 to –0.00720]	1.9
C	7.9 [6.7–9.3]	*y* = 1.13 [1.08–1.15] *x* – 0.0528 [–0.109 to –0.00728]	1.3
D	16.0 [14.5–17.1]	*y* = 1.16 [1.12–1.19] *x* + 0.0131 [–0.0486 to 0.0584]	2.4
E	–3.3 [–4.5 to –1.9]	*y* = 0.966 [0.923–0.995] *x* + 0.0114 [–0.0571 to 0.0500]	2.6
F	–6.1 [–6.9 to –5.4]	*y* = 0.887 [0.856–0.907] *x* + 0.0873 [0.0506–0.144]	1.4
G	–1.8 [–2.2 to –1.1]	*y* = 1.00 [0.991–1.02] *x* – 0.0292 [–0.0519 to –0.0135]	2.1
L	6.5 [5.6–7.4]	*y* = 1.06 [1.03–1.09] *x* + 0.00489 [–0.0243 to 0.0529]	1.7
M	1.8 [1.2–2.2]	*y* = 1.02 [0.995–1.03] *x* + 0.00190 [–0.0162 to 0.0299]	1.0
N	–11.7 [–12.3 to –10.9]	*y* = 0.879 [0.859–0.898] *x* + 0.00286 [–0.0299 to 0.0411]	2.0
O	9.9 [8.8–11.3]	*y* = 1.06 [1.02–1.10] *x* + 0.0632 [–0.0101 to 0.106]	5.2
P	–7.7 [–8.3 to –7.2]	*y* = 0.936 [0.923–0.949] *x* – 0.0182 [–0.0428 to –0.000331]	1.7
Q	–5.8 [–6.2 to –5.2]	*y* = 0.958 [0.945–0.969] *x* – 0.0262 [–0.0421 to –0.00294]	2.2
R	10.5 [9.3–12.5]	*y* = 1.09 [1.02–1.16] *x* + 0.0164 [–0.0768 to 0.100]	5.6
S	–3.2 [–3.8 to –2.5]	*y* = 0.982 [0.967–1.00] *x* – 0.0148 [–0.0443 to 0.00765]	2.7
T	–7.2 [–8.6 to –6.4]	*y* = 0.946 [0.919–0.974] *x* – 0.0275 [–0.0680 to 0.00429]	3.6
V	0.6 [–0.3 to 2.1]	*y* = 1.03 [0.986–1.08] *x* – 0.0239 [–0.101 to 0.0282]	3.8
All assays	–1.2 [–1.8 to –0.4]	n/a	2.6
IA only	–1.2 [–1.8 to –0.4]	n/a	2.6
B			
A	0.8 [–0.3 to 1.8]	*y* = 0.932 [0.900–0.968] *x* + 0.118 [0.0746–0.160]	2.1
B	1.1 [0.0–2.8]	*y* = 0.995 [0.943–1.04] *x* + 0.0220 [–0.0342 to 0.0855]	1.6
C	–2.7 [–6.9 to –0.7]	*y* = 1.17 [1.12–1.20] *x* – 0.294 [–0.355 to –0.254]	1.5
D	–1.3 [–2.2 to 0.0]	*y* = 1.05 [1.01–1.07] *x* – 0.0771 [–0.133 to –0.0381]	2.5
E	–1.5 [–3.9 to 1.5]	*y* = 1.13 [1.08–1.16] *x* – 0.187 [–0.260 to –0.147]	2.7
F	0.9 [–0.1 to 1.3]	*y* = 1.06 [1.01–1.08] *x* – 0.0855 [–0.127 to –0.0171]	1.5
G	0.8 [0.1 to 1.9]	*y* = 0.974 [0.961–0.989] *x* + 0.0579 [0.0360–0.0735]	1.6
L	–0.8 [–1.9 to –0.5]	*y* = 1.03 [1.00–1.05] *x* – 0.0610 [–0.0888 to –0.0146]	1.5
M	0.6 [0.1–0.9]	*y* = 0.988 [0.965–1.00] *x* + 0.0224 [0.00512–0.0539]	0.9
N	0.6 [–0.5 to 1.6]	*y* = 1.03 [1.00–1.05] *x* – 0.0406 [–0.0779 to 0.00358]	1.7
O	1.6 [1.1 to 2.5]	*y* = 0.998 [0.954–1.04] *x* + 0.0183 [–0.0446 to 0.0596]	3.7
P	0.9 [0.2–1.9]	*y* = 0.959 [0.944–0.972] *x* + 0.0751 (0.0496–0.0950]	1.4
Q	0.7 [0.3–1.5]	*y* = 0.980 [0.966–0.990] *x* + 0.0419 [0.0257–0.0643]	1.6
R	–1.2 [–3.3 to 1.5]	*y* = 1.07 [1.01–1.14] *x* – 0.127 [–0.217 to –0.0523]	5.4
S	–0.2 [–0.7 to 0.6]	*y* = 0.951 [0.937–0.972] *x* + 0.0916 (0.0620–0.113]	2.2
T	0.6 [–0.8 to 1.8]	*y* = 0.928 [0.900–0.956] *x* + 0.137 [0.0987–0.169]	2.6
V	1.9 [–0.7 to 3.3]	*y* = 0.927 [0.890–0.975] *x* + 0.156 [0.0880–0.203]	2.6
All assays	0.5 [0.0–0.8]	n/a	2.2
IA only	0.5 [0.0–0.8]	n/a	2.2

Mean bias and confidence interval (CI) of bias, Passing–Bablok regression equation, and average coefficient of variation (CV) were determined with 38 individual donor sera for TSH, pre-recalibration (**A**) and post-recalibration (**B**).

fT4 concentrations of 11.3 and 26.5 pmol/L (0.881 and 2.06 ng/dL) were classified as euthyroid and hyperthyroid by the fT4 RMP, respectively. Most of the 21 assays misclassified these samples with challenging fT4 concentrations (15 misclassified the 11.3 pmol/L sample and 14 misclassified the 26.5 pmol/L sample).

The in-silico recalibration increased the number of samples classified uniformly by all assays from 21 to 33 of the 40 PS, suggesting that while some sample-specific bias exists, significant improvement in overall consistency of sample classification can be achieved upon recalibration. While some sample-specific scatter existed for most of the assays (Supplementary Table S2), recalibration improved both classification agreement among assays and bias to the CDC RMP. After recalibration is complete, manufacturers can focus on improving sample-specific bias. The study demonstrated that recalibration significantly improved the accuracy of assays and the consistency of classifications. Stakeholders should be aware that assay values may change by as much as 34% or more following standardization. However, the IFCC study suggests that this shift in assay values due to recalibration is viewed as a positive change and is not expected to pose a problem to stakeholders.[9,19]

While the study results revealed problems with the accuracy of many IAs, some promising data were observed. The improved median percent bias from the 2017 study to the present study suggests that some of the IA manufacturers have already initiated standardization activities, successfully decreasing calibration bias. The significant fT4 median bias of the present study indicates further improvement is needed,

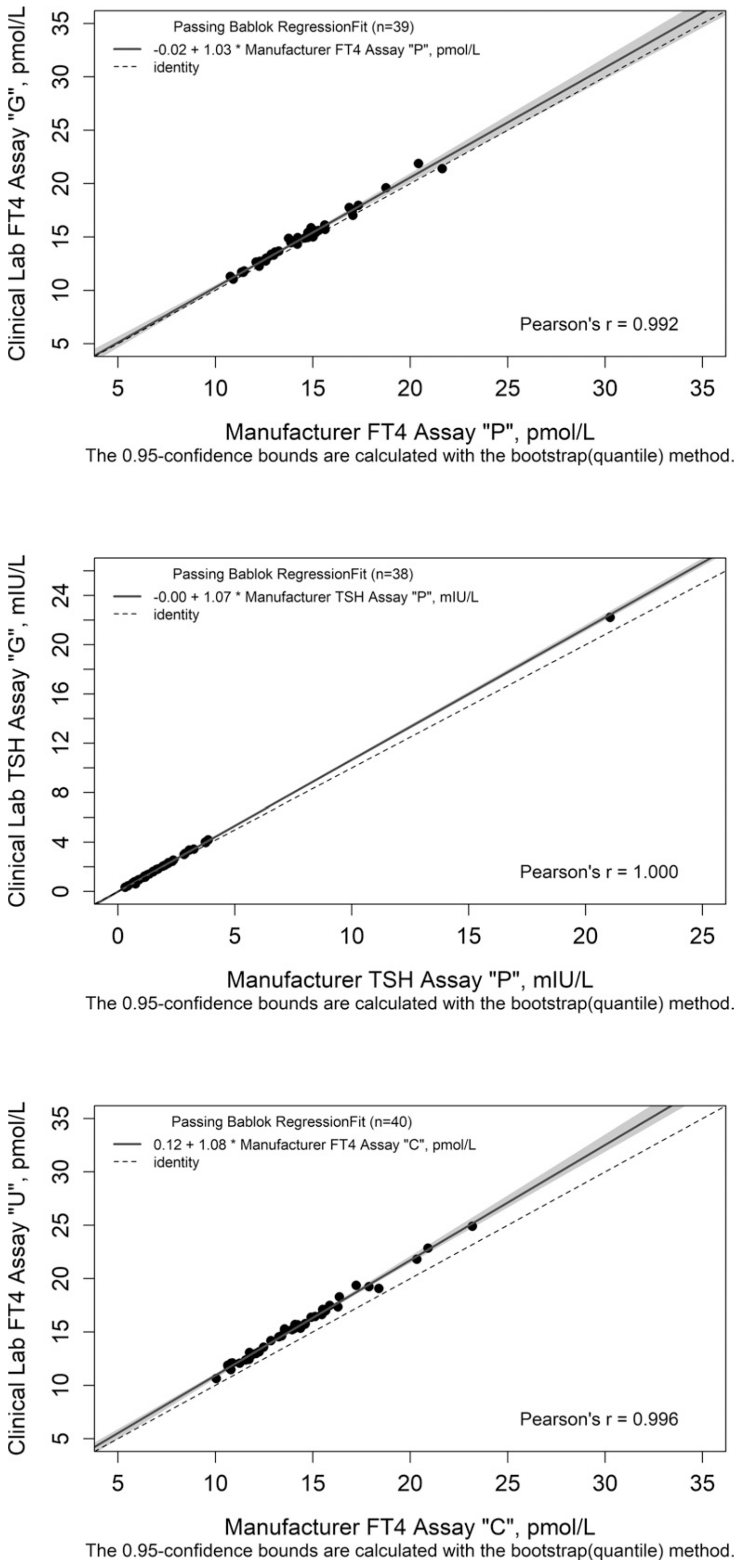

FIG. 4. Comparison of bias to the RMP for manufacturers and their corresponding clinical labs. Mean percent biases were compared among clinical and manufacturer lab pairs measuring fT4 or TSH with the same method. Passing–Bablok regression analysis of manufacturer "P" and clinical lab "G" paired labs are shown in plot A (fT4) and plot B (TSH). Passing–Bablok regression analysis of fT4 manufacturer "C" and clinical lab "U" paired labs are shown in plot C. The $y = x$ identity line is indicated by the dashed horizontal line, and Passing–Bablok regression with confidence interval (CI) is indicated by the solid lines and grey bands. The CI of the slope and intercept parameters are [0.938–1.12] (slope) and [–1.30 to 1.24] (intercept) for plot A, [1.03–1.08] (slope) and [–0.031 to 0.070] (intercept) for plot B, and [1.04–1.11] (slope) and [–0.282 to 0.720] (intercept) for plot C.

Table 5. Free Thyroxine Classification Agreement Among Included Immunoassay and Liquid Chromatography–Tandem Mass Spectrometry–Based (Laboratory-Developed Test) Assays

Sample number	*fT4, pmol/ L (ng/dL)*	*Participants reporting below RI (%)*	*Participants reporting within RI (%)*	*Participants reporting above RI (%)*	*Not reported (%)*
PS 1	11.3 (0.881)	71 (5)	**29 (95)**	0 (0)	0 (0)
PS 2	12.0 (0.932)	52 (5)	**48 (95)**	0 (0)	0 (0)
PS 3	13.2 (1.03)	48 (0)	**52 (100)**	0 (0)	0 (0)
PS 4	12.3 (0.952)	67 (0)	**33 (100)**	0 (0)	0 (0)
PS 5	12.4 (0.967)	57 (0)	**43 (100)**	0 (0)	0 (0)
PS 6	12.7 (0.988)	38 (0)	**62 (100)**	0 (0)	0 (0)
PS 7	13.0 (1.01)	19 (0)	**81 (100)**	0 (0)	0 (0)
PS 8	13.2 (1.02)	33 (0)	**67 (100)**	0 (0)	0 (0)
PS 9	13.9 (1.08)	5 (0)	**95 (100)**	0 (0)	0 (0)
PS 10	12.6 (0.98)	14 (0)	**86 (100)**	0 (0)	0 (0)
PS 11	14.2 (1.11)	43 (5)	**57 (95)**	0 (0)	0 (0)
PS 12	14.4 (1.12)	14 (0)	**86 (100)**	0 (0)	0 (0)
PS 13	15.3 (1.19)	10 (0)	**90 (100)**	0 (0)	0 (0)
PS 14	15.7 (1.22)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 15	16.4 (1.27)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 16	15.0 (1.16)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 17	17.6 (1.37)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 18	16.8 (1.3)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 19	16.0 (1.24)	5 (0)	**95 (100)**	0 (0)	0 (0)
PS 20	16.4 (1.27)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 21	16.3 (1.27)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 22	16.3 (1.27)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 23	17.4 (1.35)	5 (0)	**95 (100)**	0 (0)	0 (0)
PS 24	17.6 (1.37)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 25	18.4 (1.43)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 26	18.9 (1.46)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 27	18.7 (1.45)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 28	18.9 (1.47)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 29	18.3 (1.42)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 30	19.2 (1.49)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 31	18.2 (1.41)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 32	18.8 (1.46)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 33	19.8 (1.54)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 34	21.2 (1.65)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 35	22.4 (1.74)	0 (0)	**95 (95)**	5 (5)	0 (0)
PS 36	19.9 (1.54)	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 37	23.4 (1.82)	0 (0)	**100 (95)**	0 (5)	0 (0)
PS 38	24.0 (1.86)	0 (0)	**90 (52)**	10 (48)	0 (0)
PS 39	26.5 (2.06)	0 (0)	67 (0)	**33 (100)**	0 (0)
PS 40	32.1 (2.49)	0 (0)	24 (5)	**71 (90)**	5 (5)

The percentage of the 21 included IAs and LDTs reporting fT4 concentrations below, within, or above the assay-specific reference interval (RI) for each sample is shown before in-silico recalibration. Post-recalibration, a unified TSH RI was estimated following the procedures recommended in the CLSI document EP28-A3c.[15] In brief, linear pairwise regression analysis of each tested assay to the all-lab mean was determined (Supplementary Fig. S1). Assay "H" was selected for transference of the RI because the slope and intercept of the linear regression were closest to 1 and 0, respectively. This unified RI (9.75–25.1 pmol/L or 0.758–1.95 ng/dL) is only intended to estimate the impact of recalibration on patient classification. Classification agreement after in-silico recalibration is indicated in parentheses. The percentage of assays in which the participant classification matched the classification determined by the CDC fT4 RMP RI is shown in bold.

CLSI, Clinical and Laboratory Standards Institute; PS, individual-donor sera.

with individual assays demonstrating median biases up to –34.1%. The majority of TSH assays were in overall good agreement with the ALM pre-recalibration; 13 of the 17 median bias CIs were within the desirable bias limits based on biological variation (±10.1%),[20] and 33 of the 39 PS samples were classified in the same way by all 17 assays using their method-specific TSH RI, demonstrating success in harmonization of TSH IAs.[8] By contrast, pre-recalibration Passing–Bablok regression parameters among TSH assays indicated most methods were significantly different from the ALM. These differences may be the result of the influence of the single high-concentration TSH sample included in the study, and further studies with additional samples in this concentration range are needed to assess the bias on the full concentration range. The low TSH median bias and narrow bias distributions are likely the result of ongoing TSH harmonization efforts by the IFCC.[8] It is also worth noting that all TSH and fT4 assays had good precision pre- and post-recalibration.

The utility of fT4 measurements of pooled materials was also explored. Using pooled materials can be beneficial when large volumes of the same material are needed, that is,

Table 6. Thyrotropin Classification Agreement Among Included Immunoassays

Sample number	*TSH, mIU/L*	*Participants reporting below RI (%)*	*Participants reporting within RI (%)*	*Participants reporting above RI (%)*	*Not reported (%)*
PS 1	21.5	0 (0)	0 (0)	**100 (100)**	0 (0)
PS 2	3.11	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 3	2.35	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 4	2.40	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 5	2.31	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 6	1.79	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 7	0.696	0 (12)	**100 (88)**	0 (0)	0 (0)
PS 8	2.23	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 9	2.13	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 10	4.01	0 (0)	**71 (82)**	29 (18)	0 (0)
PS 11	1.25	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 12	3.16	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 13	4.13	0 (0)	**59 (82)**	41 (18)	0 (0)
PS 14	1.79	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 15	1.28	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 16	0.644	0 (35)	**100 (65)**	0 (0)	0 (0)
PS 17	0.337	**82 (100)**	18 (0)	0 (0)	0 (0)
PS 18	1.44	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 19	0.855	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 20	1.40	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 21	3.29	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 22	0.801	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 23	1.61	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 24	2.28	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 25	1.56	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 26	1.22	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 27	2.48	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 28	4.07	0 (0)	**65 (94)**	35 (6)	0 (0)
PS 29	1.62	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 30	1.20	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 31	2.57	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 32	3.43	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 33	1.27	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 34	1.86	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 35	0.811	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 36	0.465	**47 (94)**	53 (6)	0 (0)	0 (0)
PS 37	0.969	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 39	1.28	0 (0)	**100 (100)**	0 (0)	0 (0)
PS 40	0.983	6 (6)	**94 (94)**	0 (0)	0 (0)

The percent of the 17 included IAs reporting TSH concentrations below, within, or above the assay-specific reference interval (RI) for each sample is shown before in-silico recalibration. Post-recalibration, a unified TSH RI was estimated following the procedures recommended in the CLSI document EP28-A3c.[15] In brief, linear pairwise regression analysis of each assay to the all-lab mean was determined (Supplementary Fig. S4). Assay "M" was selected for transference of the RI because the slope and intercept of the linear regression were closest to 1 and 0, respectively. This unified RI (0.606–4.32 mIU/L) is only intended to estimate the impact of recalibration on patient classification. Classification agreement after recalibration is indicated in parentheses. The percentage of assays in which the participant classification matched the classification determined by the surrogate TSH RI is shown in bold.

for preparation of trueness or quality controls. However, combining individual units of different binding protein concentrations may result in non-commutable pooled material. When analyzed by each of the included IAs, the high-quality Pool1–3 materials (prepared in accordance with the updated CLSI C-37A procedure) and PS samples provided comparable results when results for each set of samples analyzed were compared to the fT4 RMP.[10] This suggested that high-quality pooled materials may be used in a similar manner as individual donor samples for calibration and performance evaluation. However, additional IA-specific studies of the commutability of pooled materials are needed.

In addition, study participants were asked to measure one sample from a pregnant (third trimester) woman. The results were consistent with the recently published study comparing pregnant women to controls and demonstrated that the majority of the IAs measured lower fT4 in healthy controls compared to PNs.[16,21] The previously published study did not look at differences in fT4 measurements based on trimester, and the reported mean gestational age of the study participants was 24.8 weeks. Based on the limited information from just one sample, IAs overestimated fT4 concentration in the PN sample, while the difference was not as consistent among LDTs. The extent of the overestimation varied depending on the IA used and is most likely due to interferences present in pregnancy samples with the methodologies used.

Table 7. Summary of Median Percent Biases Among Immunoassays and MS-Based Assays (Laboratory-Developed Test) Measuring Free Thyroxine and Thyrotropin to the Centers for Disease Control and Prevention Reference Measurement Procedure (All-Lab Mean for Thyrotropin) for Pooled (Pool1–3) and Pregnancy Samples

A									*B*				
		fT4 IA bias (%)				*fT4 LDT bias (%)*					*TSH IA bias (%)*		
Sample name	N	*Median*	*Minimum*	*Maximum*	N	*Median*	*Minimum*	*Maximum*	*Sample name*	N	*Median*	*Minimum*	*Maximum*
Pool1	16	−26.1	−43.6	−2.1	4	9.7	4.3	21.9	Pool 1	15	0.5	−12.7	16.5
Pool2	16	−17.4	−33.1	−2.2	4	2.5	−1.8	27.8	Pool 2	15	−2.1	−9.5	13.6
Pool3	16	−18.1	−31.8	−3.8	4	0.6	−2.7	16.4	Pool 3	15	0.6	−9.2	12.9
All pools	16	−21.0	−43.6	−2.1	4	3.7	−2.7	27.8	All pools	15	0.5	−12.7	16.5
PN	16	−1.9	−27.8	19.8	4	1.8	−8.7	52.5					

Median percent biases and the minimum and maximum biases among IA and LDT measuring fT4 (**A**) or TSH (**B**) were calculated for all pooled and PN samples. fT4 assay "U" did not report results for these materials. Only TSH IAs are shown because no TSH LDTs participated in this study, and TSH assays did not measure the PN sample. TSH assays "G" and "O" did not provide data for pooled samples.

PN, pregnancy.

Another important aspect of standardization is ensuring that clinical laboratories using the same IAs receive comparable results to manufacturers. One IA manufacturer and clinical laboratory pair using identical assays (Labs "G" and "P") to measure both fT4 and TSH for the study were compared. For fT4 measurement, the pair used identical calibrators and reagents and reported results with no significant differences. For TSH measurement, significant differences in the performance of manufacturer "P" to its paired end user ("G") were observed, which could be explained using different lots of reagents and calibrators reported by the two laboratories. It is important to highlight the importance of monitoring overtime performance of assays performed in different laboratories; changes associated with calibrators and/or reagent lots, among other factors, may cause inconsistencies in results.

One of the limitations of this study was the absence of samples with very high and very low fT4 and TSH concentrations. Based on previous studies, IA-based measurements of samples with such concentrations may demonstrate even higher biases. Furthermore, only limited observations were possible for samples from pregnant women and donors taking levothyroxine due to limited sample size.

Conclusions

The CDC CSP interlaboratory comparison among fT4 assays demonstrated that despite previous efforts, no notable improvements have been achieved. The current variability among fT4 assays can affect reproducibility of diagnostic classification. This variability can be easily addressed through recalibration, which is being offered with the CDC CSP fT4 Horome Standardization (HoSt) program. The study also demonstrated that there was better agreement among TSH assays, confirming the success of TSH harmonization efforts conducted by IFCC.

Acknowledgments

The authors gratefully acknowledge support and contributions to the study of Samantha Knoblock, Clark Coffman, Fidelia Pokuah, Danielle Gregory, Alex Doty, and Lynn Collins at the CDC and the Assay Manufacturers and Laboratories that participated in the study.

Authors' Contributions

A.R.: Writing—original draft (co-lead), data analysis (lead), conducting the study (equal), and writing—review and editing (equal). O.S.: Conducting the study (equal) and writing—review and editing (equal). T.B.: Conducting the study (equal) and writing—review and editing (equal). N.V.: Conducting the study (equal) and writing—review and editing (equal). A.N.L.: Study design (supporting), conducting the study (equal), and writing—review and editing (equal). L.Z.: Study design (equal) and writing—review and editing (equal). U.D.: Study design (lead), writing—original draft (co-lead), conducting the study (equal), and writing—review and editing (equal). H.W.V.: Supervision, study conceptualization (lead), study design (supporting), writing—original draft (equal), data analysis (supporting), and writing—review and editing (equal).

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the CDC. Use of trade names and commercial sources is for identification only and does not constitute endorsement by the CDC or the U.S. Department of Health and Human Services.

Author Disclosure Statement

All coauthors are CDC employees involved in a program offering the CDC CSP fT4 HoSt. The authors do not receive any direct reimbursement from the program offered by the CDC. The authors have no conflicts of interest to disclose.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors have no funding information to declare.

Supplementary Material

Supplementary Data
Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Figure S5
Supplementary Figure S6
Supplementary Table S1
Supplementary Table S2

References

1. Jonklaas J, Bianco AC, Bauer AJ, et al.; American Thyroid Association Task Force on Thyroid Hormone Replacement. Guidelines for the treatment of hypothyroidism: Prepared by the American Thyroid Association Task Force on Thyroid Hormone Replacement. Thyroid 2014;24(12):1670–1751; doi: 10.1089/thy.2014.0028
2. Ross DS, Burch HB, Cooper DS, et al. 2016 American Thyroid Association Guidelines for diagnosis and management of hyperthyroidism and other causes of thyrotoxicosis. Thyroid 2016;26(10):1343–1421; doi: 10.1089/thy.2016.0229
3. Alexander EK, Pearce EN, Brent GA, et al. 2017 Guidelines of the American Thyroid Association for the diagnosis and management of thyroid disease during pregnancy and the postpartum. Thyroid 2017;27(3):315–389; doi: 10.1089/thy.2016.0457
4. American Thyroid Association. Hypothyroidism in Pregnancy. Available from: https://www.thyroid.org/wp-content/uploads/patients/brochures/hypothyroidism_pregnancy_brochure.pdf [Last accessed: September 24, 2024].
5. Feldt-Rasmussen U, Bliddal S, Rasmussen AK, et al. Challenges in interpretation of thyroid function tests in pregnant women with autoimmune thyroid disease. J Thyroid Res 2017;2017:4324130; doi: 10.4061/2011/598712
6. Koulouri O, Moran C, Halsall D, et al. Pitfalls in the measurement and interpretation of thyroid function tests. Best Pract Res Clin Endocrinol Metab 2013;27(6):745–762; doi: 10.1016/j.beem.2013.10.003
7. De Grande LAC, Van Uytfanghe K, Reynders D, et al.; IFCC Committee for Standardization of Thyroid Function Tests (C-STFT). Standardization of free thyroxine measurements allows the adoption of a more uniform reference interval. Clin Chem 2017;63(10):1642–1652; doi: 10.1373/clinchem.2017.274407
8. Thienpont LM, Van Uytfanghe K, De Grande LAC, et al.; IFCC Committee for Standardization of Thyroid Function Tests (C-STFT). Harmonization of serum thyroid-stimulating hormone measurements paves the way for the adoption of a more uniform reference interval. Clin Chem 2017;63(7):1248–1260; doi: 10.1373/clinchem.2016.269456
9. Vesper HW, Van Uytfanghe K, Hishinuma A, et al. Implementing reference systems for thyroid function tests—A collaborative effort. Clin Chim Acta 2021;519:183–186; doi: 10.1016/j.cca.2021.04.019
10. Danilenko U, Vesper HW, Myers GL, et al. An updated protocol based on CLSI document C37 for preparation of off-the-clot serum from individual units for use alone or to prepare commutable pooled serum reference materials. Clin Chem Lab Med 2020;58(3):368–374; doi: 10.1515/cclm-2019-0732
11. Ernst DJ, Martel AM, Arbique JC, et al. GP41. Collection of Diagnostic Venous Blood Specimens. 7th Edition. Clinical and Laboratory Standards Institute (CLSI); 2017.
12. Ribera A, Zhang L, Dabbs-Brown A, et al. Development of an equilibrium dialysis ID-UPLC-MS/MS candidate reference measurement procedure for free thyroxine in human serum. Clin Biochem 2023;116:42–51; doi: 10.1016/j.clinbiochem.2023.03.010
13. The National Committee for Clinical Laboratory Standards. EP9-A2. Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition. NCCLS: 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898 USA; 2002.
14. CMS. Center for Clinical Standards and Quality/Quality SOG. Final Rule—Clinical Laboratory Improvement Amendments of 1988 (CLIA) Proficiency Testing - Analytes and Acceptable Performance Final Rule CMS-3355-F). 2024.
15. Horowitz GA, Boyd S, Ceriotti J, et al. CLSI. EP28-A3c. Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline - Third Edition. Clinical and Laboratory Standards Institute; 2010.
16. Jansen HI, van der Steen R, Brandt A, et al. Description and validation of an equilibrium dialysis ID-LC-MS/MS candidate reference measurement procedure for free thyroxine in human serum. Clin Chem Lab Med 2023;61(9):1605–1611; doi: 10.1515/cclm-2022-1134
17. Horowitz GH. A. College of American Pathologists. Educational Discussion: Free T4 Testing 2023-A Harmonized Thyroid (ABTH). 2023. Available from: https://documents.cap.org/documents/2023-A-Harmonized-Thyroid.pdf [Last accessed: February 1, 2025].
18. Horowitz G. College of American Pathologists. Educational Discussion: TSH Testing 2022-B Harmonized Thyroid (ABTH). 2022. Available from: https://documents.cap.org/documents/2022-B-Harmonized-Thyroid.pdf [Last accessed: February 1, 2025].
19. IFCC C-STFT. Summary of activities to collect information about concerns and potential risks associated with changes in reference intervals for free thyroxine and TSH and about communication and interactions among relevant stakeholders. Available from: https://ifcc-cstft.org/research/summary-

of-activities-to-collect-information-about-concerns-and-potential-risks-associated [Last accessed: August 9, 2024].

20. European Federation of Clinical Chemistry and Laboratory Medicine. EFLM Biological Variation Database. Available from: https://biologicalvariation.eu/search?query=Thyroid%20stimulating%20hormone%20(TSH) [Last accessed: February 2, 2025].
21. Jansen HI, van Herwaarden AE, Huijgen HJ, et al. Pregnancy disrupts the accuracy of automated fT4 immunoassays. Eur Thyroid J 2022;11(6); doi: 10.1530/ETJ-22-0145

Address correspondence to:
Uliana I. Danilenko, PhD
Division of Laboratory Sciences
Centers for Disease Control and Prevention (CDC)
4770 Buford Highway NE MS F25
Atlanta
GA 30341
USA

E-mail: udanilenko@cdc.gov