Clinical Biochemistry 68 (2019) 37–43

Contents lists available at ScienceDirect

Clinical Biochemistry

journal homepage: www.elsevier.com/locate/clinbiochem

Impact of testosterone assay standardization efforts assessed via accuracy-based proficiency testing

Zhimin (Tim) Cao[a,b], Julianne Cook Botelho[c], Robert Rej[a,d], Hubert Vesper[c], J. Rex Astles[c,*]

[a] Wadsworth Center, New York State Department of Health, Albany, NY, United States of America
[b] College of Arts and Sciences, University at Albany, State University of New York, Albany, NY, United States of America
[c] Centers for Disease Control and Prevention, Atlanta, GA, United States of America
[d] Department of Biomedical Sciences, School of Public Health, University at Albany, State University of New York, Albany, NY, United States of America

ARTICLE INFO

Keywords:
Testosterone
Proficiency testing
Standardization
Accuracy
Harmonization
External quality assessment

ABSTRACT

Background: We reported observations on analytical performance in testosterone measurements of various methods/assays from the study carried out using accuracy-based proficiency testing (PT) during 2012–2013. In 2016, we re-evaluated analytical performance of testosterone assays using accuracy-based PT to assess effectiveness of CDC efforts toward standardization.

Methods: Five single-donor human serum samples from female and male adult donors were analyzed for testosterone by New York State Department of Health-certified clinical laboratories using 16 immunoassays and LC-MS/MS methods. Target values were determined using the CDC reference measurement procedure.

Results: Testosterone targets for the 5 samples were 43.5, 160, 294, 457, and 534 ng/dL. The biases of individual result of the 65 participant laboratories against the target for each sample were calculated. Of participants, 87.7% had ≥4 of the 5 results within the minimum allowable total error limits (± 25.1%), a 14.7% increase from the previous study. The improved PT scores were attributed to better analytical accuracy and precision, and laboratories' selection of more accurate assays/methods.

Conclusions: Improved analytical accuracy and precision for testosterone assays were demonstrated over a 3.5-year period after the first CDC-directed accuracy-based proficiency testing. Additional effort is needed to improve accuracy/precision of measurements, especially at low concentrations.

1. Introduction

Accuracy-based proficiency testing (PT), or external quality assessment is needed for reliably assessing accuracy of analytical measurement systems and participant laboratories, and through comparisons over time for determining the impact of standardization/harmonization efforts. In our previous study carried out during 2012–2013 [1], we reported observations from accuracy-based PT study on analytical biases and levels of imprecision for testosterone measurements using 17 analytical systems including immunoassays and liquid chromatography-tandem mass spectrometry (LC-MS/MS) methods. We conducted this study in May 2016 to monitor changes in accuracy and precision of testosterone testing during the interim.

Accurate testosterone measurement is important in the diagnosis, treatment, and prevention of a series of clinical manifestations, such as hypogonadism, delayed or precocious puberty, polycystic ovary syndrome, and certain cancers [2,3]. Inaccurate testosterone measurements have been reported, especially at the low concentrations found in women and pre-pubertal children [4,5]. The importance of analytical variability and inaccuracy of testosterone assays on quality of patient diagnosis and treatment is well-documented [6–8]. Commonly used methods for testosterone analysis include immunoassays and mass spectrometry. The CDC Hormone Standardization (CDC HoSt) Program aims to improve the accuracy of testosterone assays [9].

Although PT can monitor and aid in improving quality performance of clinical laboratories and commercial products, PT providers often use altered rather than authentic human specimens as a matrix to accommodate large volumes. Such artificially altered matrices may affect

Abbreviations: CDC, US Centers for Disease Control and Prevention; CAP, College of American Pathologists; HoSt, hormone standardization; LDT, laboratory developed test; LC-MS/MS, liquid chromatography-tandem mass spectrometry; NYSDH, New York State Department of Health; PT, proficiency testing; RMP, reference measurement procedure.

* Corresponding author at: Division of Laboratory Systems, Center for Epidemiology, Laboratory, and Surveillance Services, Centers for Disease Control and Prevention, 4770 Buford Hwy, MS G-23, Atlanta, GA 30341, United States of America.
E-mail address: jda4@cdc.gov (J.R. Astles).

https://doi.org/10.1016/j.clinbiochem.2019.03.014
Received 21 January 2019; Received in revised form 7 March 2019; Accepted 27 March 2019
Available online 28 March 2019
0009-9120/ Published by Elsevier Inc. on behalf of The Canadian Society of Clinical Chemists.

Table 1
Statistical summary of testosterone results (ng/dL) reported by all participant laboratories.

Samples (target values)*		Sample-I (43.5 ng/dL)			Sample-II (160.0 ng/dL)			Sample-III (294.0 ng/dL)			Sample-IV (457.0 ng/dL)			Sample-V (534.0 ng/dL)		
Method#	n**	Median	Mean (%CV)	Min-Max	Median	Mean (%CV)	Min-Max	Median	Mean (%CV)	Min-Max	Median	Mean (%CV)	Min-Max	Median	Mean (%CV)	Min-Max
Abbott																
Architect c & i System	5	56.0	56.5(9.1)	49.0–61.6	172.4	171.0(3.6)	160.0–175.1	306.5	296.7(7.6)	260.0–315.0	502.6	504.5(5.1)	472.0–544.0	593.0	596.6(4.0)	573.0–636.0
Beckman Coulter																
UniCel DxI 800	6	83.7	82.2(3.2)	78.0–84.0	150.4	151.5(2.6)	148.0–157.0	289.4	285.3(3.7)	267.0–294.0	403.0	401.6(3.9)	374.0–420.0	433.7	427.2(3.5)	406.0–442.9
UniCel DxI 600	9	81.0	81.5(4.9)	76.0–87.9	150.0	152.9(4.7)	143.0–165.0	290.0	288.8(3.0)	276.0–302.0	415.0	410.5(4.1)	376.0–430.0	437.0	429.0(5.9)	382.0–453.0
ACCESS	7	81.1	79.8(5.9)	72.0–85.0	156.8	155.4(3.2)	148.0–161.0	311.1	307.2(4.3)	278.0–317.6	442.7	436.1(5.0)	400.0–465.9	481.1	474.4(5.0)	438.0–500.3
Siemens																
ADVIA Centaur	15	52.9	53.4(19.6)	34.4–73.5	143.0	144.7(9.7)	119.8–180.8	243.1	249.9(15.5)	203.8–339.3	405.2	423.8(16.8)	316.6–570.0	422.0	422.1(16.4)	311.0–566.0
Immulite 2000	8	39.8	39.8(6.7)	36.2–44.0	148.0	148.3(7.6)	129.0–163.0	339.5	329.6(9.6)	287.0–375.0	420.5	438.8(12.7)	366.0–534.0	417.0	414.3(7.9)	372.0–481.0
Roche Cobas																
e601–2 & modular	5	64.0	64.5(3.0)	62.9–67.8	164.7	165.0(4.8)	154.5–176.8	307.6	306.0(4.2)	289.6–320.4	492.1	482.8(4.4)	446.8–498.2	542.2	531.7(3.5)	510.8–550.0
Tosoh Bioscience	4	57.4	59.7(16.0)	52.0–71.9	150.0	148.9(6.4)	136.2–159.3	349.2	357.3(7.8)	334.5–396.3	541.7	533.8(3.6)	505.0–547.0	583.1	617.3(12.4)	571.1–732.0
DiaSorin Liaison	1	50.0			130.0			235.0			393.0			497.0		
Siemens Immulite 1000	1	46.5			160.0			422.0			538.0			461.0		
Ortho Vitros 5600	1	58.8			122.0			282.0			407.0			518.0		
LDT-LC-MS/MS	1	49.0			149.0			263.0			381.0			415.0		
Immuno-Biological Laboratories	1	198.5***			151.9			341.9			493.9			536.4		
LDT-immunoassay	1	168.9***			244.5***			> 420.0			> 420.0			> 420.0		

* Target values were determined using the CDC reference measurement procedure.
** n denotes number of participant laboratories.
*** Outlier results.
# All immunoassays are competitive and heterogenous, but vary in antibodies, labeled testosterone, solid phases and detection methods [4].

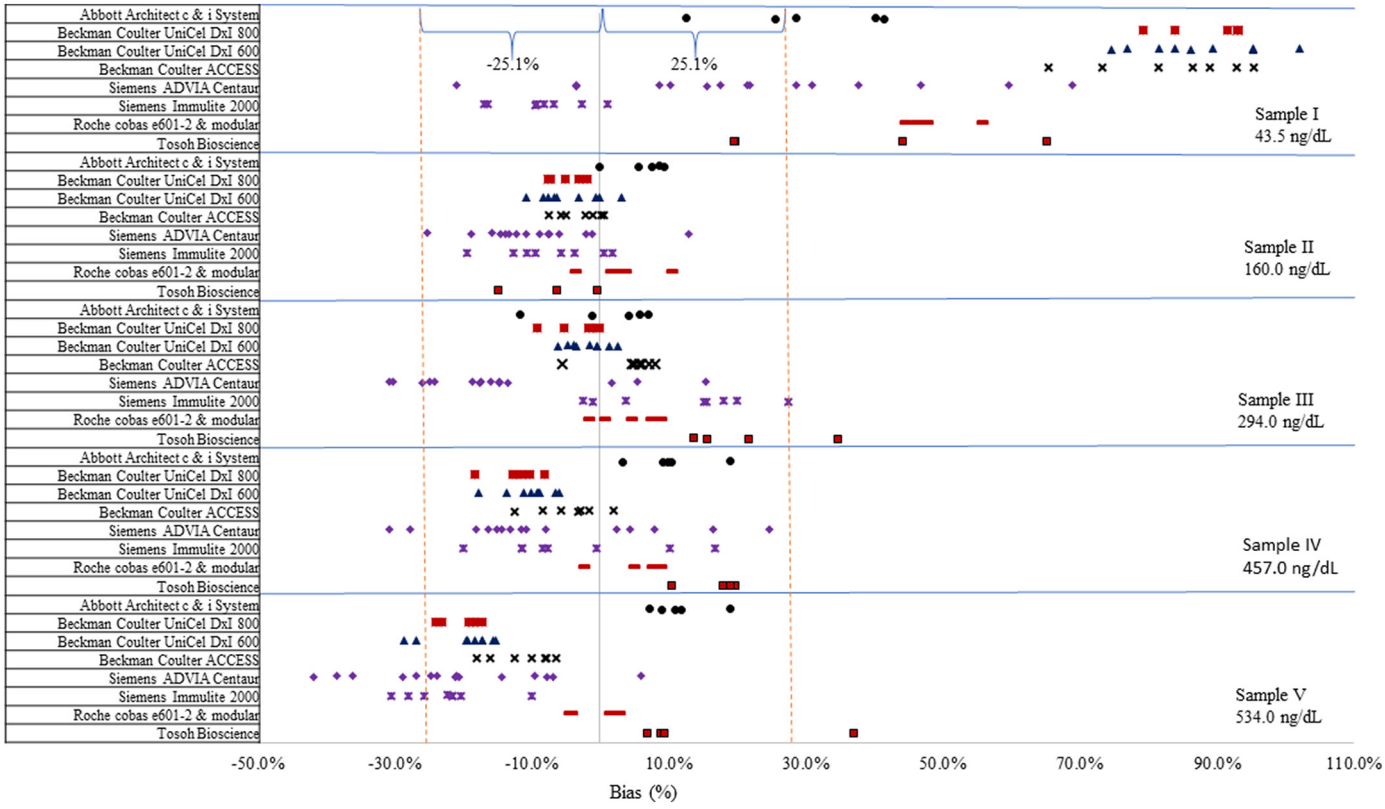

Fig. 1. Analysis of bias (%) of each individual total testosterone result sorted by analytical systems from the target value defined using the reference measurement procedure (RMP) by CDC for the five authentic human serum samples. Sample identification and target values are shown on the right portion of the figure. The dotted lines denote ± 25.1% as minimum total analytical error limits, derived from the calculation based on biological variability (16).

commutability of PT specimens, impacting the accuracy of some analytical methods [10]. "Peer grouping" results by analytical system is still commonly done for most analytes [11], without first demonstrating that PT samples are not commutable [12]. Using non-commutable samples for PT, or assuming non-commutability exists without its demonstration, limits the effectiveness of PT for assessing the accuracy of testosterone measurements. Therefore, non-commutable samples for PT cannot effectively contribute to assay improvements and standardization. These shortcomings can be eliminated by accuracy-based PT, which uses commutable samples from patients and target values defined by reference measurement procedures.

2. Materials and methods

Serum samples from 5 healthy adult donors (1 female: Sample I, 4 males: Samples II to V) were obtained from Solomon Park Research Laboratories (Seattle, WA), under a process approved by their institutional review board. The serum samples were prepared according to the procedure described in Clinical and Laboratory Standards Institute (CLSI) guideline C37A [13]. Donor samples were screened for hepatitis B and C, and HIV, and found negative. Each of these samples was aliquoted to 1.0-ml fractions in 2.0-mL cryogenic vials (Corning Inc., Corning, NY) and stored at −80 °C until use. Based on previous studies conducted on sera prepared the same manner as the samples used in this study, it can be assumed that the samples used in this study were commutable. This project was reviewed by the institutional review board of New York State Department of Health (NYSDH) and was determined to be exempt. The portion of the study conducted by the CDC laboratory did not involve human subjects research.

In May 2016, samples were shipped frozen overnight on ice to 65 U.S. clinical laboratories certified by NYSDH, 14 of which were outside New York State. Participant laboratories were instructed that the samples be processed for analysis, stored at 0–8 °C upon receipt, or kept frozen if the analysis could not be performed within 24 h of receipt. Serum testosterone is stable when stored in these conditions [14]. Laboratories were required to handle the samples in the same manner as those from patients for clinical testing, perform single analysis for each sample, and report results within two weeks of receipt. Target values for testosterone concentration of the samples were determined by the CDC reference laboratory from duplicate analysis using the reference measurement procedure (RMP), which is traceable according to International System of Units (SI) units as outlined in the International Organization for Standardization (ISO) document 17511 to SI through use of previously described primary reference material NMI M914 [15].

Statistical analyses were carried out using Microsoft Excel, calculating the mean, median, range, standard deviation (Excel STDEV·S) and coefficients of variations (%CVs). Data were grouped into five assay manufacturers (Abbott, Beckman Coulter, Siemens, Roche Diagnostics, and Tosoh Bioscience), and were further sub-grouped based on analytical platforms or analytical systems used by ≥4 participants. Bias of laboratory results was calculated as the percent difference from the target value as assigned by the RMP. As in our previous report, variation of measurement results was determined as %CV of each assay manufacturer or analytical system sub-group for each sample. The biases of the analytical systems were compared with the suggested allowable analytical bias goals derived from biological variability data: "minimal," ± 9.5%; "desirable," ± 6.4%, and "optimal," ± 3.2% [16]. To determine the difference between the results of this study and those from our previous study for analytical system bias and variation, we compared results for Samples I and IV to results for Samples B, C and D [1], respectively, based on similarity of their target values (Table 3). Results that exceeded 2.5 SD from their method group mean were removed from further analysis, as was done in our previous study [1].

Table 2
Biases of laboratory results - sorted by analytical system.

Sample (target value, ng/dL)		Sample I (43.5)	Sample II (160.0)	Sample III (294.0)	Sample IV (457.0)	Sample V (534.0)		Calibration bias[#]
Assay Manufacturer Analytical System	n[*]	Mean bias in % (95% CI)					n	Mean bias in % (95% CI)
Abbott							5	11.8 (7.5 to 11.6)
Architect c & i System	5	29.8 (19.4 to 40.2)	6.3[d] (3.0 to 9.7)	0.9[o] (−5.8 to 7.6)	10.4 (5.5 to 15.3)	11.7 (7.8 to 15.7)		
Beckman Coulter							22	11.3 (9.6 to 13.0)
UniCel Dxl 600	9	87.3 (81.4 to 93.3)	−4.4[d] (−7.4 to −1.5)	−1.8[o] (−3.7 to 0.1)	−10.2 (−12.6 to −7.8)	−19.7 (−22.8 to −16.6)		
UniCel Dxl 800	6	89.0 (83.8 to 94.3)	−5.3[d] (−7.5 to −3.2)	−2.9[o] (−6.1 to 0.2)	−12.1 (−15.1 to −9.1)	−20.0 (−22.5 to −17.5)		
ACCESS	7	83.4 (75.4 to 91.4)	−2.9[o] (−5.2 to −0.6)	4.5[d] (1.1 to 7.9)	−4.6[d] (−8.1 to −1.1)	−11.2 (−14.4 to −7.9)		
Siemens							23	−6.0 (−10.5 to −1.6)
ADVIA Centaur	15	22.7 (10.5 to 34.9)	−9.6 (−14.0 to −5.1)	−15.0 (−21.7 to −8.3)	−7.3[m] (−15.1 to 0.6)	−21.0 (−27.5 to −14.4)		
Immulite 2000	8	−8.5[m] (−12.7 to −4.2)	−7.3[m] (−12.2 to −2.4)	12.1 (4.7 to 19.6)	−4.0[d] (−12.5 to 4.5)	−22.4 (−26.7 to −18.2)		
Roche							5	12.1 (9.2 to 15.1)
Cobas e601–2 & Modular	5	48.3 (44.5 to 52.2)	3.0[o] (−1.3 to 7.5)	4.1[d] (0.2 to 7.9)	5.7[d] (1.6 to 9.8)	−0.4[o] (−3.5 to 2.6)		
Tosoh Bioscience	4	37.2 (15.6 to 58.7)	−7.0[o] (−12.8 to −1.1)	21.5 (12.2 to 30.8)	16.8 (12.6 to 21.0)	15.6 (1.5 to 29.7)	4	16.8 (11.4 to 22.3)

[*] n indicates number of participant laboratories.
[m] Indicates the bias of analytical system was within ± 9.5% of the target, "minimal" bias goal.
[d] Indicates the bias of analytical system was within ± 6.4% of the target, "desirable" bias goal.
[o] Indicates the bias of analytical system was within ± 3.2% of the target, "optimal" bias goal.
[#] Calibration bias was defined as the mean of laboratory biases in that analytical system group or sub-group.

3. Results

All 65 participating clinical laboratories reported results for the five samples using a total of 17 analytical systems. Results from some analytical systems that were of the same or similar technology and were used by fewer than 4 participating laboratories were combined. We combined the results for Abbott Architect c and i systems, and Roche cobas e601, e602 and modular systems. Eight of the resulting 14 analytical systems (after combination) had ≥4 participant laboratories. Three results that were reported by two individual laboratories—one using a laboratory-developed test (LDT-immunoassay), and the other using an Immuno-Biological Laboratories assay—fell outside of the method group mean ± 2.5 SD (Table 1), and were removed from the statistical analyses. Results from the 8 systems that had more than one participant were examined for each of the 5 samples for median, mean, %CV and range (Table 1). The differences between each individual result of analysis on the five authentic serum samples and the target value defined by the CDC RMP were calculated and the results shown in Fig. 1. Assay analytical system or manufacturer mean calibration biases and the biases between the RMP-defined target values and analytical system group or sub-group means, are summarized in Table 2. Among the assay manufacturer analytical systems, Siemens ADVIA Centaur had the highest variability (%CV) for all five samples (Table 1).

As shown in Table 2, the bias levels of the analytical systems varied depending on concentration and analytical system. The bias for individual samples was inconsistent between the analytical systems of the same assay manufacturer. For Sample I with low concentration, laboratories using Siemens ADVIA Centaur reported much higher bias (22.7%) than laboratories using Siemens Immulite 2000 (−8.5%), although results from these two analytical systems showed a similar level of bias (−21% vs −22.4%) for Sample V with high concentration (Fig. 1, Table 2). All other analytical systems revealed positive bias for Sample I with low concentration (43.5 ng/dL) and negative bias for Sample V with high concentration (534.0 ng/dL) with exception of Abbott and Tosoh analytical systems.

At low concentration (Sample I) for most methods, the bias in this study was greater than previously seen (Table 3). At higher concentrations (Sample IV, 457.0 ng/dL), Beckman Coulter series had nearly no change; however, ADVIA Centaur and Immulite 2000 substantially reduced their systematic bias, while Roche series showed a slightly increased bias (Table 3).

For intra-analytical system variation (%CV), all analytical systems showed, at low concentration (Sample I), various levels of reduction when compared with the previous event (Table 3) [1]. Whereas, at higher concentrations (Sample IV, 457 ng/dL), UniCel Dxl 800 and ACCESS had nearly no changes, ADVIA Centaur and Immulite 2000 had slight increases and Cobas series showed a slight reduction of the intra-system variation (Table 3).

All 65 laboratories' results for the five samples were evaluated against an acceptance limit (target ± 25.1%), the minimum tolerance for allowable total error based on calculated requirements for accuracy and precision derived from estimates of within-individual biological variability (CVi) and between-person biological variability (CVg) [16]. The percentages of results that met this criterion for Samples I to V ranged from 35% to 99%, respectively (Table 4). Using the CLIA '88 scoring approach [17], we defined "*satisfactory* performance or score" on a PT event as a participant laboratory obtaining results for at least four of five samples within the minimum acceptance limit of target ± 25.1%. Of the 65 participating laboratories, 88% achieved satisfactory performance on this PT event (Table 4). Among the 65 laboratories, 13 laboratories did not participate in our previous study [1]. When results of these 13 laboratories were removed in order to compare the same cohort, the satisfactory score was 86.5% (45/52). In the current study, 10 of the 14 analytical systems had all their user laboratories obtained satisfactory scores. In our previous study, however, only 3 of these 10 analytical systems had all their user laboratories achieve satisfactory

Table 3
Comparison of measuring system's mean biases and analytical variation between this study and the study carried out in 2012–2013 (in italics) (1)

Sample (target value, ng/dL)	Sample I (43.5)[a]		*Sample B (30.0)*[b]		Sample IV (457.0)[a]	*Sample C (402.0)*[b]	*Sample D (498.0)*[b]
Assay Manufacturer Analytical System	Mean bias in % (95% CI); variation (%CV) of results by each analytical system						
	n		n				
Beckman Coulter							
UniCel DxI 800	6	89 (84 to 94); 3.2%	11	*16 (10 to 22); 9.4%*	−12 (−15 to −9); 3.9%	*−15 (−17 to − 13); 3.5%*	*−13 (−17 to − 10); 7.0%*
ACCESS	7	83 (75 to 91); 5.9%	6	*19 (10 to 27); 9.1%*	−5 (−8 to −1); 5.0%	*−9 (−11 to − 6); 3.4%*	*−4 (−8 to − 1); 4.6%*
Siemens							
ADVIA Centaur	15	23 (11 to 35); 19.6%	39	*21 (14 to 29); 20.4%*	−7 (−15 to 1); 16.8%	*−17 (−20 to − 14); 11.3%*	*−13 (−16 to − 11); 9.2%*
Immulite 2000	8	−9 (−13 to −4); 6.7%	20	*10 (−1 to 21); 22.7%*	−4 (−13 to 5); 12.7%	*−25 (−28 to − 21); 10.4%*	*−22 (−26 to − 19); 10.5%*
Roche							
Cobas e601–2 & Modular*	5	48 (45 to 52); 3.0%	7	*11 (−2 to 25); 16.4%*	6 (2 to 10); 4.4%	*3 (−2 to 8); 6.4%*	*3 (−4 to 10); 7.6%*

[a] Indicates the results of this study.
[b] Indicates the results of the study carried out in 2012–2013 [1]. These results are italicized.
* Results of Cobas e601, e602 and Modular were grouped for Samples I & IV; whereas, results of e411, e601 and e602 were grouped for Samples B, C & D.

scores (Table 4) [1]; note that there were no Coat-A-Count users in the current event.

Investigation to identify the source(s) of increased participant laboratories' satisfactory rates revealed three changes or differences between the studies (Table 4). First, for the current study, three assay systems (Abbott Axsym, Siemens ADVIA Centaur CP and Ortho Vitros ECiQ) either had no results, or no participants used these systems in this study. In the 2012–2013 study, of the eight laboratories that reported results using these three systems, fewer than half achieved a satisfactory event score. Secondly, Abbott Architect c & i assay systems had results from five laboratories, all of which achieved a satisfactory score in this study. No results were reported using these systems in the 2012–2013 study. Finally, eight assay systems (Beckman Coulter UniCel Dxl600/800 and ACCESS, Siemens Immulite 1000/2000 and ADVIA Centaur, Tosoh and Ortho Vitros 5600) had a higher percentage of users that achieved 100% or satisfactory scores in the current study, compared to the 2012–2013 study (Table 4).

4. Discussion

Testosterone concentrations of five authentic human serum samples were analyzed by 65 clinical laboratories using 17 analytical systems.

Table 4
Proficiency evaluation of participant laboratory results.

Methods	n##	Sample I	Sample II	Sample III	Sample IV	Sample V	All samples- 2017	All samples- 2012 & 2013#
		Number (%) of laboratories with results within the allowable limits*					Laboratories with satisfactory event score**	
							Number (%)	(%), all participant number
Abbott Laboratories								
Axsym								(33), *n* = 3
Architect c & i System	5	1 (20)	5 (100)	5 (100)	5 (100)	5 (100)	5 (100)	
Beckman Coulter								
UniCel DxI 600	9	0 (0)	9 (100)	9 (100)	9 (100)	7 (78)	7 (78)	(50), *n* = 2
UniCel DxI 800	6	0 (0)	6 (100)	6 (100)	6 (100)	6 (100)	6 (100)	(90), *n* = 10
ACCESS	7	0 (0)	7 (100)	7 (100)	7 (100)	7 (100)	7 (100)	(83), n = 6
Siemens								
ADVIA Centaur (XP, XPT)	15	9 (60)	15 (100)	12 (80)	13 (86.7)	10 (67)	11 (73)	(56), *n* = 36
ADVIA Centaur (CP)								(33), *n* = 3
Immulite 1000	1	1 (100)	1 (100)	0 (0)	1 (100)	1 (100)	1 (100)	(33), n = 3
Immulite 2000	8	8 (100)	8 (100)	7 (87.5)	8 (100)	5 (63)	8 (100)	(68), *n* = 19
Coat-A-Count								(100), n = 3
Roche cobas e601–2 & e	5	0 (0)	5 (100)	5 (100)	5 (100)	5 (100)	5 (100)	(100), *n* = 21
Tosoh Bioscience	4	2 (50)	4 (100)	3 (75)	4 (100)	3 (75)	3 (75)	(0), n = 1
Ortho Clinical Diagnostics								
Vitros 5600	1	0 (0)	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)	(0), n = 1
Vitros ECiQ								(50), n = 2
DiaSorin Liaison	1	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)	
Immuno-Biological Laboratories	1	0 (0)	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)	
LDT***- LC/MS/MS	1	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)	(100), *n* = 4
LDT-Immunoassay	1	0 (0)	0 (0)	0 (0)	1 (100)	1 (100)	0 (0.0)	(100), n = 1
Total	65	23 (35)	64 (99)	58 (89)	63 (97)	54 (83)	57 (88)	(73), *n* = 115

* Target ± 25.1%.
** Satisfactory score is defined as laboratory's results with the allowable limit for ≥4 of the 5 samples in the event.
*** Laboratory Developed Test.
# Original data were published in reference [1].
n denotes number of participant laboratories.

The target values ranged from 43.5 ng/dL to 534 ng/dL, simulating concentrations typical for children, adult males and females, and approximating cutoff values for clinical diagnosis of hypogonadism for men [14,18,19]. The mean of means for each of the eight analytical systems for Sample III was 302.6 ng/dL (10.5% CV), almost the same as the RMP-defined target value of 294.0 ng/dL and close to the 300 ng/dL decision point defining androgen deficiency or hypogonadism in men according to The Endocrine Society clinical practice guideline [20]. However, results of Sample III had a rather wide range of 203.8–422.0 ng/dL (Table 1). Further analysis revealed that overall performance was best for Sample III among all samples. However, for Sample III, only 3 analytical systems had mean biases within the *optimal* range (± 3.2% of the target) and intra-system CVs ≤7.6%), while 2 analytical systems had mean biases within the *desirable* range (± 6.4% of the target) and intra-system CVs ≤4.3%, and 3 analytical systems had mean biases exceeding the *minimal* bias goal or ± 9.5%, with higher intra-method imprecision than other assay systems (Tables 1 & 2). These results suggest that further improvements in both analytical accuracy and precision are needed for at least 3 commercial analytical systems. For Samples I-III, results from the LDT were more innaccurate than the commercial tests, illustrating the value of for accuracy-based PT for this type of assay.

For Sample IV (target value 457.0 ng/dL) and Sample V (target value 534.0 ng/dL), participant laboratories (n = 65) reported results ranging from 316.6 to 570.0 ng/dL (both seen for Siemens ADVIA Centaur) and from 311.0 to 732.0 ng/dL, respectively, with 100% of reported values above 300 ng/dL, suggesting that no samples for adult males would be misclassified as having an androgen deficiency. For Sample II (target value 160.0 ng/dL), all participant laboratories' results ranged from 119.8 to 180.8 ng/dL (again, both seen for Siemens ADVIA Centaur), with 100% of the values below 300 ng/dL, suggesting that all patients would be correctly classified as hypogonadism for the adult male. In the previous study [1], Sample A had a target testosterone concentration value of 15.5 ng/dL, close to the cut-off value of 20 ng/dL used to indicate achievement of effective treatment of prostate cancer with androgen ablation [21]. In that study for the Sample A, about 60% of the results were reported < 20 ng/dL; however, for the Sample B with a target value of 30.0 ng/dL, none of the results were reported < 20 ng/dL [1]. In the current study, though the testosterone concentration (43.5 ng/dL) of Sample I was the lowest among the five samples, it was still higher than that of the Sample B of the previous study, and the Sample I had the highest biases (Table 3). These results indicate that the problems of false high results with samples at low concentrations observed in the first study still exist.

Our results showed that the Beckman Coulter (UniCel 600, 800 and ACCESS) products, as used by the laboratories participating in this study, had bias means about 3-fold higher than those of Abbott Architect c & i systems for Sample I (Table 2), but with relatively low intra-method variation (3.2% to 5.9% CV) (Table 1). For Sample I with the target value of 43.5 ng/dL, results from almost all other analytical systems fell within a published reference interval for normal female adults (15–70 ng/dL) [18], suggesting no misleading diagnoses would be made for this sample. However, of the 22 results from the Beckman Coulter's 3 analytical systems, only one result fell within the Beckman Coulter's ACCESS recommended reference range of 10–75 ng/dL for female serum sample [22]. In certain cases, analytical bias might be "corrected" by customized reference intervals; however, bias cannot be removed if it is due to sporadic analytical non-specificity. Concerns of a similar issue, but on different analytes, have been expressed by others [23].

Results from Sample I revealed the overall highest biases in terms of percentage, ranging from −8.5% to 89.0% among the five samples, with 63% (41/65) of the reported results falling above the value of target +25.1% or 54.4 ng/dL (Table 2, Fig. 1). It is known that immunoassays frequently produce biased testosterone results with high sample-to-sample variability in bias, especially at low concentrations that are typical of female and early adolescent male samples [4,5]. The Immulite 2000 appears to have a noticeable change in assay performance over recent years. In 2003, bias mean for 53 samples from females (mean 63.4 ± 34.6 ng/dL) was 150% (50% to 156%) compared to the target as defined by a GC/MS method [4]. In our previous study for the Sample B (30.0 ng/dL) [1], the bias mean was 9.8% (−1.2% to 20.7%). In this study for the Sample I (43.5 ng/dL), the bias mean was −8.5% (−12.7% to −4.2%). Such a change may be partially attributable to Immulite 2000's improvement in assay specificity and precision (Table 3). There are several causes of inaccurate results at low testosterone concentration as mentioned by others [5]. Sample I was obtained from a normal adult female and to our knowledge it contained no unusual interfering substances such as heterophilic antibodies or abnormal hormones. However, further studies are needed to ensure that the observed differences in bias are not a result of high sample-to-sample variability in bias.

We observed a noticeable difference in mean biases between the Immulite 2000 and ADVIA Centaur systems for Samples I and III; mean biases were − 8.5% and 12.1% for Immulite 2000 vs. -15.0% and 22.7% for ADVIA Centaur, respectively (Table 2). This is similar to the observations obtained in the 2012/2013 study [1]. The Immulite series and ADVIA Centaur had been products of different manufacturers until 2006 when acquired by Siemens; therefore, they may differ in antibody specificity. Thus, the observed differences can be partially explained with differences in antibody specificity. Further studies with more samples are needed to assess whether these differences also derive from different calibrations. However, the discrepancies were not shown for the Sample II, IV and V, suggesting existence of high sample-to-sample variability in bias.

Among the eight analytical systems (Table 2), Roche Diagnostics had a mean bias within the recommended "desirable" bias goal of ± 6.4%, ranging from −0.4% to 5.7% for the Samples II though V. However, a mean bias of 48.3% was obtained for Sample I (43.5 ng/dL), greater than four-fold the value of 11% observed for Sample B in the previous study [1]. Given the high sample-to-sample variability in bias, especially in samples with low testosterone concentrations, further studies are needed using more samples to appropriately assess measurement bias and changes in bias over time, and to distinguish between calibration bias and bias caused by non-specificity.

In this study, only Siemens had a calibration bias (−6.0% Table 2) that was within the ± 6.4% mean bias criterion used in CDC's Hormone Standardization Program (CDC HoSt). CDC HoSt program uses samples from 40 individual donors to calculate the mean bias, which can be regarded as calibration bias. In 2016, a Siemens assay obtained certification by CDC's Hormones Standardization Program. In the previous study, a Roche assay was certified by CDC's HoSt program but was no longer listed as certified in 2016 [9]. In this study, the calibration bias of the Roche systems was outside the ± 6.4% mean bias criterion (12.1% Table 2). Furthermore, the CDC HoSt program reports that the Siemens assay had only 20% of individual sample results falling within the ± 6.4% bias criterion [24] which appears consistent with the passing scores listed in Table 4. Thus, information on certification status in CDC HoSt program appears consistent with the analytical performance observed in PT programs. A gradual drift of the analytical performance for all concentrations over time might explain the increase in bias (Table 3), suggesting that periodically monitoring and assessing assays through the CDC HoSt Program may be necessary. More frequent accuracy-based PT challenges are needed to reliably assess and monitor the accuracy of testosterone measurements performed for patient care.

Results from all participant laboratories in this study were evaluated by applying a single target and one criterion, i.e., the RMP-defined target ± 25.1% [16]. Using this criterion 87.7% of participant laboratories would receive a satisfactory PT event score. Since improvements in assay performance are typically made by the manufacturer, with the exception of laboratory developed tests, accuracy-based PT can be used to assess whether improvements of assay performance implemented by

the manufacturer result in improvements at the end-user level. Thus, accuracy-based PT can be used to assess the effectiveness of assay standardization, which ultimately leads to improvement in the quality of patient care and population health. For this study, the satisfactory score of participating laboratories improved in average by 14.7% when compared to the previous study [1]. These improvements could in part be explained by successful participation in CDC's HoSt Program of assay manufacturers and laboratories. Also, the availability of reliable data on measurement accuracy through accuracy-based PT provide valuable information to laboratories for selecting accurate and reliable assays, and for assay manufacturers to further improve assay performance. Information obtained from CDC's HoSt Program and accuracy-based PT is also used by key stakeholders, such as the Partnership for the Accurate Testing of Hormones (PATH), to promote use of accurate and reliable testosterone tests [1,24–26]. Standardization ultimately leads to improved patient care and population health and accuracy based-PT can help assess the effectiveness of standardization efforts.

There are shortcomings in this study, including that these five samples are not representative of all specimens measured in patient care. Ultimately the accuracy of test methods is only measurable (and relevant) in the hands of their users. Thus, like all PT studies, it is impossible to isolate the impact of the test performance by end-users. Therefore, it is possible that improved sample handling or other techniques may have some reflection in the improvements shown here. Follow-up studies are needed to analyze greater numbers of patient samples. It is especially important to study performance at lower concentrations, where biases may impact quality of health care.

In conclusion, improvements in overall analytical accuracy and precision for most analytical systems for testosterone tests were demonstrated over an approximately 3.5-year period after the first CDC-directed accuracy-based PT in 2012/2013. Further effort is needed in improving accuracy of measurement for testosterone, especially at the low concentrations.

Conflicts of interest

All authors declare that they have no conflict of interest in this study.

Disclaimer

The findings and conclusions in this paper have not been formally disseminated by the Centers for Disease Control and Prevention, the Agency for Toxic Substances and Disease Registry, or the Wadsworth Center of the New York State Department of Health, and should not be construed to represent any agency determination or policy. Use of trade names and commercial sources is for identification only and does not constitute endorsement by the U.S. Department of Health and Human Services, the U.S. Centers for Disease Control and Prevention, or the Wadsworth Center of the New York State Department of Health.

References

[1] Z. Cao, J.C. Botelho, R. Rej, H. Vesper, Accuracy-based proficiency testing for testosterone measurements with immunoassays and liquid chromatography-mass spectrometry, Clin. Chim. Acta 469 (2017) 31–36.

[2] A.M. Matsumoto, W.J. Bremner, Testicular disorders, in: S. Melmed, K.S. Polonksy, P.R. Larson, H.M. Kroneberg (Eds.), Williams Textbook of Endocrinology, 12th ed., Elsevier Saunders, Philadelphia (PA), 2011, pp. 688–777.

[3] R.D. Nerenz, E. Jungheim, A.M. Gronowski, Reproductive endocrinology and related disorders, in: N. Rifai, A.N. Norvath, C.T. Wittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, 6th ed., Elsevier, St. Louis (MO), 2018, pp. 1617–1654.

[4] J. Taieb, B. Mathian, F. Millot, M.C. Patricot, E. Mathieu, N. Queyrel, et al., Testosterone measured by 10 immunoassays and by isotope-dilution gas chromatography-mass spectrometry in sera from 116 men, women, and children, Clin. Chem. 49 (2003) 1381–1395.

[5] C. Wang, D.H. Catlin, L.M. Demers, B. Starcevic, R.S. Swerdloff, Measurement of total serum testosterone in adult men: comparison of current laboratory methods versus liquid chromatography-tandem mass spectrometry, J. Clin. Endocrinol. Metab. 89 (2004) 534–543.

[6] A.M. Matsumoto, W.J. Bremner, Serum testosterone assays - accuracy matters, J. Clin. Endocrinol. Metab. 89 (2004) 520–524.

[7] M.E. Wierman, R. Basson, S.R. Davis, S. Khosla, K.K. Miller, W. Rosner, et al., Androgen therapy in women: an Endocrine Society clinical practice guideline, J. Clin. Endocrinol. Metab. 91 (2006) 3697–3710.

[8] W. Rosner, H.W. Vesper, Endocrine Society and endorsing organizations. Toward excellence in testosterone testing: a consensus statement, J. Clin. Endocrinol. Metab. 95 (2010) 4542–4548.

[9] Centers for Disease Control and, Prevention. CDC Hormone Standardization Project. Standardization of Serum Total Testosterone Measurements, http://www.cdc.gov/labstandards/hs.html (Accessed 21 January 2019).

[10] W.G. Miller, G.R. Jones, G.L. Horowitz, C. Weykamp, Proficiency testing/external quality assessment: current challenges and future directions, Clin. Chem. 57 (2011) 1670–1680.

[11] W.G. Miller, Time to pay attention to reagent and calibrator lots for proficiency testing, Clin. Chem. 62 (2016) 666–667.

[12] Clinical and Laboratory Standards Institute (CLSI), Evaluation of Commutability of Processed Samples; Approved Guideline—Third Edition (CLSI document EP 14-A3), Clinical and Laboratory Standards Institute, Wayne (PA), 2014.

[13] Clinical and Laboratory Standards Institute (CLSI), Preparation and validation of commutable frozen human serum pools as secondary reference materials for cholesterol measurement procedures (CLSI document C37A), Clinical Laboratory Standards Institute, Wayne (PA), 1999.

[14] M. Kushnir, T. Blamires, A.L. Rockwood, W.L. Roberts, B.F. Yue, E. Erdogan, et al., Liquid chromatography – tandem mass spectrometry assay for androstenedione, dehydroepiandrosterone, and testosterone with pediatric and adult reference intervals, Clin. Chem. 56 (2010) 1138–1147.

[15] J.C. Botelho, C. Shacklady, H.C. Cooper, S.S. Tai, K.V. Uytfanghe, L.M. Thienpont, et al., Isotope-dilution liquid chromatography–tandem mass spectrometry candidate reference method for total testosterone in human serum, Clin. Chem. 59 (2013) 372–380.

[16] Y.M. Yun, J.C. Botelho, D.W. Chandler, A. Katayev, W.L. Roberts, F.Z. Stanczyk, et al., Performance criteria for testosterone measurements based on biological variation in adult males: recommendations from the Partnership for the Accurate Testing of hormones, Clin. Chem. 58 (2012) 1703–1710.

[17] US Department of Health and Human Services. Medicare, Medicaid and CLIA programs: regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA), final rule, Fed Regist 57(1992)7002–7186, 42 CFR Ch. IV, 10-1-13 Edition: http://www.gpo.gov/fdsys/pkg/CFR-2013-title42-vol5/pdf/CFR-2013-title42-vol5-part493.pdf. (Accessed 21 January 2019).

[18] K. Adeli, F. Ceriotti, M. Nieuwesteeg, Reference information for the clinical laboratory, in: N. Rifai, A.N. Norvath, C.T. Wittwer (Eds.), Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, 6th ed., Elsevier, St. Louis (MO), 2018, pp. 1745–1818.

[19] O.P. Soldin, H. Sharma, L. Husted, S.J. Soldin, Pediatric reference intervals for aldosterone, 17α-hydroxyprogesterone, dehydroepiandrosterone, testosterone and 25-hydroxy vitamin D3 using tandem mass spectrometry, Clin. Biochem. 42 (2009) 823–827.

[20] S. Bhasin, G.R. Cunningham, F.J. Hayes, A.M. Matsumoto, P.J. Snyder, R.S. Swerdloff, V.M. Montori, Testosterone therapy in adult men with androgen deficiency syndromes: an Endocrine Society clinical practice guideline, J. Clin. Endocrinol. Metab. 91 (2006) 1995–2010.

[21] L.G. Gomella, Effective testosterone suppression for prostate cancer: is there a best castration therapy? Rev Urol. 11 (2009) 52–60.

[22] Beckman Coulter, Inc, ACCESS Immunoassay Systems Test Information, Access Testosterone, Total, https://www.henryschein.com/assets/Medical/3310089.pdf (Accessed 21 January 2019).

[23] K. Adeli, V. Higgins, D. Seccombe, C.P. Collier, C.M. Balion, G. Cembrowski, et al., CSCC reference interval harmonization (hRI) working group, national survey of adult and pediatric reference intervals in clinical laboratories across Canada: a report of the CSCC working group on reference interval harmonization, Clin. Biochem. 50 (2017) 925–935.

[24] Centers for Disease Control and Prevention, CDC Hormone Standardization Program (HoSt) -Testosterone Certified Procedure, https://www.cdc.gov/labstandards/hs_certified_participants.html (Accessed 21 January 2019).

[25] J.C. Botelho, Z. Cao, Y.E. Rahmani, R. Rej, H.W. Vesper, Towards an accuracy-based proficiency testing program - A CDC and NY State Department of Health Partnership, Association of Public Health Laboratories (APHL) Annual Meeting and Environmental Laboratory Conference, Raleigh, NC, June 2 – 5, 2013.

[26] New York State Department of Health, Endocrinology Proficiency Test Program Statistical Summary – May 2016, http://www.wadsworth.org/sites/default/files/WebDoc/1605endofull_0.pdf (Accessed 21 January 2019).