HHS Public Access

Author manuscript

Clin Chim Acta. Author manuscript; available in PMC 2017 November 20.

Published in final edited form as:

Clin Chim Acta. 2017 June ; 469: 31–36. doi:10.1016/j.cca.2017.03.010.

Accuracy-based proficiency testing for testosterone measurements with immunoassays and liquid chromatography-mass spectrometry☆

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

[a]Wadsworth Center, New York State Department of Health, Albany, NY, United States

[b]College of Arts and Sciences, University at Albany, State University of New York, Albany, NY, United States

[c]Centers for Disease Control and Prevention, Atlanta, GA, United States

[d]School of Public Health, University at Albany, State University of New York, Albany, NY, United States

Abstract

Background—Accurate testosterone measurements are needed to correctly diagnose and treat patients. Proficiency Testing (PT) programs using modified specimens for testing can be limited because of matrix effects and usage of non-reference measurement procedure (RMP)-defined targets for evaluation. Accuracy-based PT can overcome such limitations; however, there is a lack of information on accuracy-based PT and feasibility of its implementation in evaluation for testosterone measurements.

Methods—Unaltered, single-donor human serum from 2 male and 2 female adult donors were analyzed for testosterone by 142 NYSDH-certified clinical laboratories using 16 immunoassays and LC-MS/MS methods. Testosterone target values were determined using an RMP.

Results—The testosterone target concentrations for the 4 specimens were 15.5, 30.0, 402 and 498 ng/dl. The biases ranged from −17.8% to 73.1%, 3.1% to 21.3%, −24.8% to 8.6%, and −22.1% to 6.8% for the 4 specimens, respectively. Using a total error target of ±25.1%, which was calculated using the minimum allowable bias and imprecision, 73% of participating laboratories had ≥3 of the 4 results within these limits.

Conclusions—The variability in total testosterone measurements can affect clinical decisions. Accuracy-based PT can significantly contribute to improving testosterone testing by providing reliable data on accuracy in patient care to laboratories, assay manufacturers, and standardization programs.

☆Disclaimer: The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official views or positions of the Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry and Wadsworth Center of New York State Department of Health. 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 New York State Department of Health.

*Corresponding author at: Division of Laboratory Sciences, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, GA, United States. hvesper@cdc.gov (H. Vesper).

Keywords

Testosterone; Proficiency; Testing; Standardization; Accuracy

1. Introduction

Abnormalities in serum testosterone concentrations are associated with, or can cause, many clinical manifestations, such as hypogonadism, delayed or precocious puberty, polycystic ovary syndrome, and certain cancers [1,2]. Accurate measurements of testosterone concentrations are critical for providing biochemical evidence to support clinical decisions in the diagnosis, treatment, and prevention of androgen disorders.

Measurements of testosterone are commonly carried out using immunoassays and mass spectrometry-based methods. In the clinical laboratory, the majority of testosterone measurements are performed on so-called “direct immunoassays” using automated analyzers. These procedures omit analyte extraction and/or chromatographic separation prior to immune-reaction. While these assays allow for fast and convenient measurements, some of them have been found to be inaccurate, especially at low testosterone concentrations [3,4]. Mass spectrometry-based assays for testosterone are increasingly used [5], which is evidenced by a nearly 5-fold increase of these methods in the College of American Pathologists PT survey from 2012 to 2015 [6]. Measurements performed with this technology allow- for quantification of the analyte and fragment by the molecular weight allowing for increased specificity. In addition, mass spectrometry-based methods typically include isolation of the analyte prior to analysis through a sample cleanup procedure, as well as chromatographic separation. However, variability in measurement bias of mass spectrometry-based methods to a reference method among these assays has also been described [7,8]. The variability and inaccuracy among all testosterone assays, as well as their clinical significance, were emphasized in editorials, commentaries, an Endocrine Society Position Statement, and other publications by professional organizations and experts [9–17]. The CDC established the Hormone Standardization (HoSt) Program to improve the accuracy and reliability of testosterone assays [18].

The availability of reference measurement procedures (RMPs) [19–22], reference materials, and the CDC standardization program enabled immunoassay manufacturers and laboratories to improve the accuracy and reliability of testosterone assays and helped to generate results that are comparable and accurate. Despite these improvements, analytical performance of testosterone measurements remains a concern, especially for women and pre-pubertal children [23,24]. Furthermore, the analytical performance of assays not participating in standardization programs is unknown. Current inter-laboratory comparison studies investigated measurement accuracy, as expressed in the bias between a routine method and a reference method, using special study designs. In these studies, samples may not be analyzed together with regular patient samples and may be tested differently, for example, using replicate measurements. In addition, these studies often are conducted by research laboratories or assay manufacturers. Therefore, these inter-laboratory studies may not reflect

the accuracy of measurements performed in routine health care settings. Little information is available about the accuracy of testosterone measurements performed in patient care.

Clinical laboratories in the U.S. are required to participate in proficiency testing (PT) programs. These programs require PT samples to be measured in the same manner as patient samples, and thus could provide information about the variability of testosterone measurements performed in patient care. Current data from PT programs often show high inter-laboratory variations and inter-method discrepancies of up to 130% [6]. However, these findings may not correctly describe the actual variability among assays because of the unknown quality of the materials with regard to commutability and the lack of target values defined by reference measurement procedures (RMP) [25,26]. As a result, a laboratory's performance is often evaluated against the target derived from the mean value of results obtained from participants using the same method or measurement system. These so-called 'peer group' assessments limit the effectiveness of PT programs in assessing the accuracy of testosterone measurements in patient care.

2. Materials and methods

Serum samples from 4 healthy adult donors (2 female: Samples A and B, 2 male: Samples C and D) prepared according to the procedure described in the CLSI document C37A [27] were obtained from Solomon Park Research Laboratories under an approved process by the institutional review board, IRB Services. These authentic human serum specimens were screened and found negative for Hepatitis B, Hepatitis C and HIV. They were aliquoted to 1.0-ml fractions in 2.0-mL cryogenic vials (Corning Inc.) and stored at −80 °C until use. This study was approved by the institutional review board of New York State Department of Health. The portion of the study conducted by the Centers for Disease Control and Prevention (CDC) laboratory was determined not to constitute engagement in human subjects research.

The serum specimens were shipped frozen overnight on ice to 142 clinical laboratories in September 2012 (Samples A and D) and January 2013 (Samples B and C). The laboratories were asked to either store the specimens at 0–8 °C upon receipt or freeze the samples if the analysis could not be carried out within 24 h of receipt. Testosterone has been shown to be stable in serum at these conditions [28]. Participants were asked to handle the serum specimens in the same manner as patient samples for clinical testing. Thus, it was requested that single measurements of each sample be made and for results to be reported within 2 weeks of receipt.

The testosterone target values for the serum specimens were determined using the RMP operated by the CDC reference laboratory [22]. Briefly, serum specimens were treated with ammonium acetate to release testosterone from binding proteins, followed by a liquid/liquid extraction using a mixture of ethyl acetate and hexane solvents to separate proteins and lipids. The organic extracts were dried and reconstituted in ammonium carbonate solution, and a second extraction was performed with hexanes to remove polar lipids. The organic phase was dried and reconstituted, followed by analysis on LC-MS/MS. The target values for the serum specimens were 15.5 ng/dl (Sample A), 30.0 ng/dl (Sample B), 402 ng/dl

(Sample C), and 498 ng/dl (Sample D), respectively, with relative expanded uncertainty of 2.9%.

Data were grouped into 5 groups of assay manufacturers with >4 participants (Ortho Clinical Diagnostics, Beckman Coulter, Siemens, Roche Diagnostics, and LC-MS/MS). Further, data were grouped by analytical platforms with >3 participants (referred in the text as “Analytical Systems”).

Statistical analyses were carried out using JMP (ver 11.1.1) and Microsoft Excel. Individual measurement results were grouped by assay manufacturer and by measurement system. Results reported by individual laboratories were within the method group mean ± 2.5 SD, with the exception of 2 results, which were considered outliers, and were excluded from assessments. Sample bias was calculated as the percent difference from the assigned target value for each sample. Laboratory bias was determined as the mean bias of all samples reported by a participant (referred to as “calibration bias”). The bias of an assay manufacturer group was the mean of each participant’s calibration bias in that group. Percentages of coefficients of variation (%CV) were determined for an assay manufacturer or analytical system group for each sample, because only single measurements were made by individual participants.

3. Results

The 142 clinical laboratories, using a total of 17 analytical platforms, participated in one or both PT events and reported 133 sets of testosterone results (Table 1). Data for all 4 samples were reported by 115 participants.

Individual results reported by all participants (n = 133) ranged from 7.1–39.8 ng/dl (Sample A), 20.0–53.7 ng/dl (Sample B), 239.5–471.6 ng/dl (Sample C), and 303.0–589.0 ng/dl (Sample D). The among-laboratory coefficients of variation was 35% (Sample A), 19% (Sample B), 15% (Sample C), and 13% (Sample D). Laboratories operating two assays with a limit of quantitation of 20 ng/dl (Siemens Immulite 1000 and Immulite 2000 XPi) correctly reported values for Sample A as <20 ng/dl (Table 1). The mean percent bias (range) between individual reported (n = 133 participants) results and the RMP target values was 28% (−55% to 156%), 14% (−34% to 78%), −12% (−40% to 17%), and −11% (−39% to 18%), for Samples A, B, C, and D, respectively (Fig. 1). The laboratory bias, defined as the mean bias of the 4 samples for all 133 laboratories, was calculated from the bias of each reported sample and ranged from −35% to 57%.

The Assay Manufacturer’s mean calibration bias (range) was −24.1% (−35.0% to −16.5%), 13.1% (−10.0% to 24.9%), 8.0% (−1.8 to 23.4), 3.5% (−13.9% to 17.2%), and −1.2% (−26.6% to 34.4%) for Ortho Clinical Diagnostics, Beckman Coulter, LC-MS/MS, Roche Diagnostics, and Siemens, respectively (Fig. 2). The measurement bias for individual samples was inconsistent among Assay Manufacturers. Laboratories using Beckman and Siemens systems mostly reported a positive bias for samples with low concentrations (Samples A and B) and a negative bias for samples with high concentrations (Samples C and

D). Laboratories operating Ortho Clinical Diagnostics systems showed mostly a negative bias, while LC-MS/MS systems mostly showed a positive bias for all samples.

Among the Assay Manufacturers, the variability (%CV) was highest for Siemens in Samples A, B and C, and Ortho Clinical Diagnostics in Sample D (Table 1). Calibration bias of the manufacturer can be assessed using the mean bias calculated from all samples analyzed by each laboratory operating assays from the same manufacturer. Among the 5 manufacturers, Siemens and Roche Diagnostics had a mean bias within the recommended limits of ±6.4%, with −1.2% and 3.5%, respectively, while LC-MS/MS methods had a mean bias (8.0%) within the minimal bias goal of ±9.5% [29] (Table 2).

The Analytical Systems group means ranged from 12.7 and 26.8 ng/dl (Sample A), 27.3–36.4 ng/dl (Sample B), 302.2–436.4 ng/dl (Sample C), and 361.0–532.0 ng/dl (Sample D) (Table 1).

The biases between RMP target values and Analytical Systems ranged from −17.8% to 73.1%, 3.1% to 21.3%, −24.8% to 8.6%, −22.1% to 6.8%, for Samples A, B, C, and D, respectively (Table 2). The intra-Analytical Systems group coefficients of variation ranged from 5.2%–26.8%, 6.1%–22.7%, 2.9%–11.4%, and 1.0%–10.5%, for Samples A, B, C, and D, respectively (Table 1). The variability among the Analytical Systems, as expressed in the coefficients of variation, was higher for samples with low concentrations (Samples A and B) compared to those with high concentrations (Samples C and D) with median coefficients of variation of 20.1%, 12.8%, 6.4%, and 7.1% for Samples A to D, respectively.

The biases of the Analytical Systems were compared with the suggested bias goals for total testosterone of ±9.5% (minimal bias goal), ±6.4% (desirable bias goal), and ±3.2% (optimal bias goal) [29]. The bias of the analytical systems varied based on concentration. For samples with high concentrations (Samples C and D), of the 10 Analytical Systems groups, 5 groups (Beckman Coulter Access 2, Roche Cobas (e411, e601, e602), Roche Modular, Roche Elecsys and LC-MS/MS) met the minimal bias goal with one additional group (Siemens- Coat-A-Count) meeting the minimal bias goal for Sample D only. For samples with low concentrations, only one group (Roche Elecsys) met this goal for Sample A and one group (LC-MS/MS) for Sample B. For samples with high concentrations, desirable bias goals were met by 3 groups (Roche Cobas (e411, e601, e602), Roche Modular, and LC-MS/MS) for Sample C and by 5 groups (Beckman Coulter Access 2, Siemens Coat-A-Count, Roche Cobas (e411, e601, e602), Roche Modular, and RocheElecsys) for Sample D. For samples with low concentrations, only one group met the desirable performance crieria for Samples A (Roche Elecsys) and B (LC-MS/MS). Beckman Coulter Unicel Dxl 800, Siemens Advia Centaur (XP, XPT), Siemens Immulite 1000, and Simens Immulite 2000 XPi did not meet any of the suggested bias goals (Table 2).

The proportion of individual laboratories reporting data for all 4 samples ($n = 115$) and passing the suggested minimum total error goal of ±25.1% [29], was 55%, 71%, 86%, and 89% for Samples A, B, C, and D, respectively (Table 3). The percentage of these laboratories having ≥3 results within this limit was 73%. Participants using Siemens Coat-A-Count, Roche Cobas (e411, e601, e602), Roche Modular, Roche Elecsys and LC-MS/MS

had 100% of results within the limit. Laboratories using Beckman Coulter Unicel Dxl 600, Siemens ADVIA Centaur (XP, XPT, and CP), Siemens Immulite 1000, Ortho Clinical Diagnostics Vitros (ECiQ and 5600), Abbott AxSYM, and Tosoh Bioscience A1A had <60% of results with ≥3 samples within the total error goal.

4. Discussion

In this accuracy-based survey, authentic human serum specimens from single donors were analyzed in the same manner as regular patient samples. Measurement results were compared to a target value determined with an established RMP. Two specimens from adult female donors (A and B) and adult male donors (C and D), covering a concentration range from 15.5 ng/dl to 498 ng/dl, were selected. These testosterone concentrations are within the respective reference intervals [28, 30].

The 2010 Endocrine Society Guideline on androgen deficiency in men suggests 300 ng/dl testosterone to be the lower end of normal [17]. For Sample C (target value 402 ng/dl), laboratories (n = 133) reported results ranging from 240 to 472 ng/dl, with 15% of all reported results being values below 300 ng/dl, suggesting that patients can be misclassified as being androgen deficient depending on the laboratory where the testing was performed. Typical cut-off values for testosterone suppression therapy in patients with prostate cancer are 20 ng/dl for chemical castration and 50 ng/dl for therapy with luteinizing hormone-releasing hormone analogue [31]. For Samples A (target value 15.5 ng/dl) and B (target value 30.0 ng/dl), laboratories reported values ranging from 7.1 to 39.8 ng/dl (Sample A) and 20.0 to 53.7 ng/dl (Sample B), with 44% of all reported results for Sample A being above 20 ng/dl and 2.5% of all reported results for Sample B being above 50 ng/dl. These observations suggest that incorrect conclusions about the efficacy of testosterone suppression therapy might be made depending on the laboratory where testing was performed.

Ortho Clinical Diagnostics showed a consistent negative bias across all samples, suggesting that measurement bias is mainly caused by inaccurate calibration. Beckman Coulter and Siemens showed a consistently negative mean bias for samples at high concentrations (Samples C and D) and consistently positive mean bias for samples at low concentrations (Samples A and B). This suggests that inaccuracy in these systems is caused by a combination of inaccurate calibration and insufficient analytical specificity.

Among the 5 manufacturers, Roche Diagnostics and Siemens had a mean bias within the recommended limits of ±6.4%, with −1.2% and 3.5% respectively. However, the mean bias of Analytical Systems produced by the same manufacturer differ highly for systems from Siemens (−11.1% for Immulite 2000 to 5.2% for Advia Centaur) and Beckman Coulter (7.7% for UniCel DxI 800 to 19.8% for ACCESS 2). More consistent bias among analytical systems are observed for those produced by Roche (0.7% for Roche Cobas e601 to 5.8% for Roche Elecsys) and for Ortho (−25.3% for Ortho Vitros ECi/ECiQ and −23.0% for Ortho Vitro 5600). This suggests that different Analytical Systems produced by the same manufacturer are calibrated inconsistently. Further studies using larger numbers of patient samples from male and female donors, as well as larger numbers of participants per

Analytical System, are needed to better distinguish the effects of assay calibration and assay specificity on measurement accuracy.

The Roche Diagnostics immunoassay was the only immunoassay certified by the CDC HoSt Program at the time this study was conducted, [18] and thus demonstrated that the analytical performance of the assay in the hands of the manufacturer meets suggested analytical performance criteria. The agreement achieved at the manufacturer level through the CDC HoSt Program certification is reflected in the agreement of measurements performed at the patient care level. These findings demonstrate that standardization of assays through the CDC HoSt Program can help improve and ensure the accuracy of measurements performed in patient care.

The results of the LC-MS/MS methods showed higher positive bias and inter-laboratory variance for Sample A, as compared to the other samples measured by this group. Sources of the higher positive bias could include inaccurate calibration at the low end of the analytical measurement range for these assays or could suggest that this technology, like immunoassays, can also be affected by interfering compounds and other factors. Considering that the LC-MS/MS methods group represents a very heterogeneous group of assays with different operational procedures, instrumentation employed, operational conditions, and instrument parameters, the observed variability in measurement bias appears small. However, because of this heterogeneity, observations made in this survey cannot be generalized to all LC-MS/MS assays.

The limited information available from the manufacturer's Technical Bulletins and product inserts suggest that all immunoassays are traceable to a higher order reference, as outlined in ISO 17511 [32]. However, the measurement bias observed with most assays suggest that demonstrating metrological traceability does not necessarily lead to accurate measurements in patient care. Therefore, more accuracy-based PT, such as those provided by the College of American Pathologists (CAP), are needed to appropriately assess and monitor the accuracy of testosterone measurements performed in patient care.

Total analytical error (TAE) criteria, calculated using the minimum allowable bias and imprecision, was used to assess the bias and imprecision of a single laboratory measurement [33]. When individual laboratories' results were evaluated against a previous defined [29] target of ±25.1% for TAE, 73% of laboratories would have achieved satisfactory status with at least 3 results being within these limits. This is much lower than that of any given conventional PT carried out by the NYSDH PT Program for testosterone testing, where typically 90% to 95% of laboratories achieve satisfactory results. In this study, all laboratories operating assays shown to be more accurate, such as the Roche Diagnostics, Coat-A-Count, and LC-MS/MS assays, would achieve 100% satisfactory status. Thus, the high fail rate is mainly caused by laboratories operating less accurate assays. These findings suggest that evaluation using an accuracy-based target of ±25.1% is realistic and achievable, and may help minimize incorrect patient assessment and interpretation of treatment efficacy.

Although this study used individual donor serum samples with target values assigned by an RMP, the 4 samples used in this study may not be representative for all specimens measured

in patient care. For example, they do not include samples from the pediatric population. The high sample-to-sample variability observed with the samples from female donors indicates that general conclusions about the assay performance in such samples are very limited, and data from more patient samples, and at lower concentrations, are needed. Because of these limitations, the findings in this study, with regards to bias caused by calibration and bias caused by other factors, may need to be verified in larger studies using more individual patient samples. Also, samples were analyzed in the same manner as patient samples, and therefore no replicate measurements were performed that would have provided information of laboratory imprecision. However, this study provides some insight about the measurement accuracy in patient care and offers suggestions for factors that may contribute to the observed measurement inaccuracy.

In conclusion, the variability in measurement accuracy observed in this study can affect clinical decisions and needs further improvements. This variability appears to be caused by incorrect assay calibration and insufficient analytical specificity. Using assays with a high level of specificity and an analytical performance verified through standardization programs can result in higher measurement accuracy. Accuracy-based PT can majorly contribute to improving the accuracy and reliability of testosterone measurements by providing reliable data on current accuracy in patient care.

Acknowledgments

The authors would like to acknowledge support from the CDC Foundation, CDC Hormone Laboratory members, as well as Krista Poynter and Christopher Ghattas.

Abbreviations

CDC	US Centers for Disease Control and Prevention
NYSDH	New York State Department of Health
PT	Proficiency Testing
AHS-PT	Authentic Human Serum Proficiency Test specimen
RMP	Reference Measurement Procedure

References

1. Matsumoto, AM., Bremner, WJ. Testicular disorders. In: Melmed, S.Polonksy, KS.Larson, PR., Kroneberg, HM., editors. Williams Textbook of Endocrinology. 12. Elsevier Saunders; Philadelphia (PA): 2011. p. 688-777.
2. Isbell, TS., Jungheim, E., Gronowski, AM. Reproductive endocrinology and related disorders. In: Burtis, CA.Ashwood, ER., Bruns, DE., editors. Tietz Textbook of Clinical Chemistry and Molecular Diagnostics. 5. Elsevier Saunders; St. Louis (MO): 2012. p. 1945-1990.
3. Tiel Groenestege WM, Bui HN, Kate JT, Menheere PPCA, Oosterhuis WP, Vader HL, et al. Accuracy of first and second generation testosterone assays and improvement through sample extraction. Clin Chem. 2012; 8:1154–1156.
4. Fuqua JS, Sher ES, Migeon CJ, Berkovitz GD. Assay of plasma testosterone during the first six months of life: importance of chromatographic purification of steroids. Clin Chem. 1995; 41:1146–1149. [PubMed: 7628089]

5. Jannetto PJ, Fitzgerald RL. Effective use of mass spectrometry in the clinical laboratory. Clin Chem. 2016; 62:92–98. [PubMed: 26553795]
6. College of American Pathologists, Ligand (Special) Participant Summary, Y-B. Surveys, 2015.
7. Theinpont LM, Van Uytfanhe K, Blincko S, Ramsay CS, Xie H, Doss RC, et al. State-of-the-art of serum testosterone measurements by isotope dilution-liquid chromatography-tandem mass spectrometry. Clin Chem. 2008; 54:1290–1297. [PubMed: 18556330]
8. Vesper HW, Bhasin S, Wang C, Tai SS, Dodge LA, Singh RJ, et al. Interlaboratory comparison study of serum total testosterone measurements performed by mass spectrometry methods. Steroids. 2009; 74:498–503. [PubMed: 19428438]
9. Herold DA, Fitzgerald RL. Immunoassays for testosterone in women: better than a guess? Clin Chem. 2003; 49:1250–1251. [PubMed: 12881438]
10. Rosner W, Auchus RJ, Azziz R, Sluss PM, Raff H. Position statement: utility, limitations, and pitfalls in measuring testosterone: an Endocrine Society position statement. J Clin Endocrinol Metab. 2007; 92:405–413. [PubMed: 17090633]
11. Matsumoto AM, Bremner WJ. Editorial: serum testosterone assays - accuracy matters. J Clin Endocrinol Metab. 2004; 89:520–524. [PubMed: 14764756]
12. Sacks SS. Are routine testosterone assays good enough? Clin Biochem Rev. 2005; 26:43–45. [PubMed: 16278777]
13. Stanczyk FZ, Lee JS, Santen RJ. Standardization of steroid hormone assays: why, how, and when? Cancer Epidemiol Biomark Prev. 2007; 16:1713–1719.
14. Wierman ME, Basson R, Davis SR, Khosla S, Miller KK, Rosner W, et al. Androgen therapy in women: an Endocrine Society clinical practice guideline. J Clin Endocrinol Metab. 2006; 91:3697–3710. [PubMed: 17018650]
15. Legro RS, Arslanian SA, Ehrmann DA, Hoeger KM, Murad MH, Pasquali R, et al. Diagnosis and treatment of polycystic ovary syndrome: an Endocrine Society clinical practice guideline. J Clin Endocrinol Metab. 2013; 98:4565–4592. [PubMed: 24151290]
16. Rosner W, Vesper HW. CDC workshop report improving steroid hormone measurements in patient care and research translation. Steroids. 2008; 73:1285. [PubMed: 18755204]
17. Rosner W, Vesper HW. Endocrine Society and endorsing organizations, toward excellence in testosterone testing: a consensus statement. J Clin Endocrinol Metab. 2010; 95:4542–4548. [PubMed: 20926540]
18. CDC. [accessed 01.10.16] CDC Hormone Standardization Project. Standardization of Serum Total Testosterone Measurements. 2016. http://www.cdc.gov/labstandards/pdf/hs/HoSt_Protocol.pdf
19. Siekmann L. Determination of steroid hormones by the use of isotope dilution-mass spectrometry: a definitive method in clinical chemistry. J Steroid Biochem. 1979; 11:117–123. [PubMed: 491583]
20. Thienpont LM, Van Nieuwenhove B, Stöckl D, Reinauer H, De Leenheer AP. Determination of reference method values by isotope dilution-gas chromatography/mass spectrometry: five years' experience of two European reference laboratories. Eur J Clin Chem Clin Biochem. 1996; 34:853–860. [PubMed: 8933112]
21. Tai SS, Xu B, Welch MJ, Phinney KW. Development and evaluation of a candidate reference measurement procedure for the determination of testosterone in human serum using isotope dilution liquid chromatography/tandem mass spectrometry. Anal Bioanal Chem. 2007; 388:1087–1094. [PubMed: 17530229]
22. Botelho JC, Shacklady C, Cooper HC, Tai SS, Uytfanghe KV, Thienpont LM, et al. Isotope-dilution liquid chromatography–tandem mass spectrometry candidate reference method for total testosterone in human serum. Clin Chem. 2013; 59:372–380. [PubMed: 23213081]
23. Taieb J, Mathian B, Millot F, Patricot MC, Mathieu E, Queyrel N, 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. 2003; 49:1381–1395. [PubMed: 12881456]
24. Wang C, Catlin DH, Demers LM, Starcevic B, Swerdloff RS. Measurement of total serum testosterone in adult men: comparison of current laboratory methods versus liquid chromatography-tandem mass spectrometry. J Clin Endocrinol Metab. 2004; 89:534–543. [PubMed: 14764758]

25. Miller WG. Time to pay attention to reagent and calibrator lots for proficiency testing. Clin Chem. 2016; 62:666–667. [PubMed: 26988585]

26. Miller WG, Myers GL, Ashwood ER, Killeen AA, Wang E, Ehlers GW, et al. State of the art in trueness and interlaboratory harmonization for 10 analytes in general clinical chemistry. Arch Pathol Lab Med. 2008; 132:838–846. [PubMed: 18466033]

27. Clinical 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.

28. Kushnir M, Blamires T, Rockwood AL, Roberts WL, Yue BF, Erdogan E, et al. Liquid chromatography – tandem mass spectrometry assay for androstenedione, dehydroepiandrosterone, and testosterone with pediatric and adult reference intervals. Clin Chem. 2010; 56:1138–1147. [PubMed: 20489135]

29. Yun YM, Botelho JC, Chandler DW, Katayev A, Roberts WL, Stanczyk FZ, 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. 2012; 58:1703–1710. [PubMed: 23065474]

30. Soldin OP, Sharma H, Husted L, Soldin SJ. Pediatric reference intervals for aldosterone, 17α-hydroxyprogesterone, dehydroepiandrosterone, testosterone and 25-hydroxy vitamin D3 using tandem mass spectrometry. Clin Biochem. 2009; 42:823–827. [PubMed: 19318024]

31. Gomella LG. Effective testosterone suppression for prostate cancer: is there a best castration therapy? Rev Urol. 2009; 11:52–60. [PubMed: 19680526]

32. International Organization for Standardization (ISO). In Vitro Diagnostic Medical Devices: Metrological Traceability of Values Assigned to Calibrators and Control Materials. ISO; Geneva: 2003. (ISO 17511)

33. Fraser, CG. Biological Variation: From Principles to Practice. AACC Press; 2001.

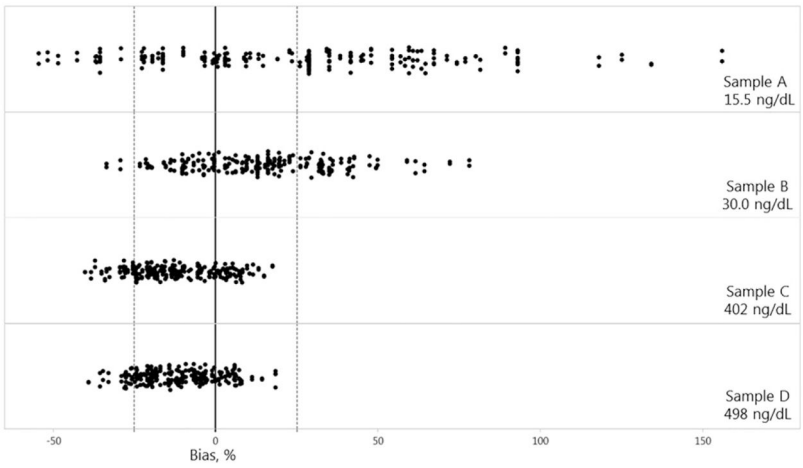

Fig. 1.

Measurement bias (%) between individual total testosterone concentrations reported by 133 clinical laboratories and the target values for 4 individual donor samples with values of 15.5 ng/dl (A), 30.0 ng/dl (B), 402 ng/dl (C), and 498 ng/dl (D). Dashed lines: total error limit (±25.1%).

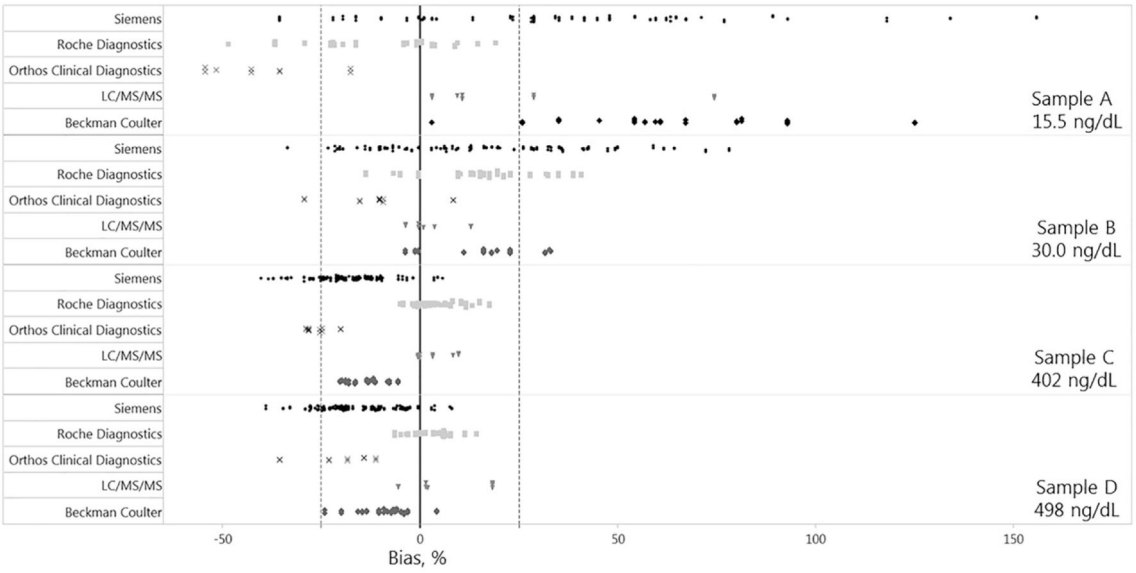

Fig. 2.

Measurement bias (%) between individual reported total testosterone concentrations by assay manufacturer and the target values for 4 individual donor samples with values of 15.5 ng/dl (A), 30.0 ng/dl (B), 402 ng/dl (C), and 498 ng/dl (D). Dashed line: total error limit (±25.1%).

Table 1

A statistical summary of reported testosterone concentrations (ng/dl)[a].

Sample (Target Value) Assay Manufacturer Analytical System	n[b]	Sample A (15.5 ng/dl)			Sample B (30.0 ng/dl)			Sample C (402 ng/dl)			Sample D (498 ng/dl)		
		Mean (%CV)[c]	Median	Min–max	Mean (%CV)[c]	Median	Min–max	Mean (%CV)[c]	Median	Min–max	Mean (%CV)[c]	Median	Min–max
Beckman Coulter (All)	20	25.3 (17.5)	25.2	16.0–35.0	35.0 (6.9)	35.0	29.0–40.1	348.4 (5.1)	349.9	320.0–379.0	447.9 (7.5)	452.0	378.0–519.0
Access 2	6	26.8 (10.6)	26.5	24.0–30.0	35.6 (9.1)	36.0	30.0–39.7	366.7 (3.4)	369.0	349.9–379.0	478.5 (4.6)	476.0	457.0–519.0
Unicel DxI 600	3		30.0	22.6–30.0		35.0	30.0–40.0		333.5	320.0–347.0		431.0	420.0–468.0
Unicel DxI 800	11	23.6 (17.8)	24.8	16.0–35.0	34.7 (9.4)	35.0	29.0–40.1	341.2 (3.5)	347.0	322.0–355.0	432.1 (7.0)	445.9	378.0–466.0
Siemens (All)	75	20.4 (28.0)	20.0	10.0–39.8	34.6 (21.8)	39.8	20.0–53.7	326.6 (12.0)	324.0	239.5–424.0	511.5 (5.4)	407.2	303.0–537.2
ADVIA Centaur (XP, XPT)	42	21.6 (26.8)	22.0	12.1–36.4	36.4 (20.4)	35.6	23.6–53.7	335.1 (11.3)	336.4	239.5–424.0	431.9 (9.2)	427.6	360.0–537.2
ADVIA Centaur CP	3		15.6	10.0–39.8		32.1	20.0–36.0		310.0	302.6–332.7		373.5	359.0–534.8
Immulite 1000	4		<20	n/a	34.1 (19.1)	35.1	25.3–40.8	331.3 (11.4)	328.0	290.0–379.0	361.0 (1.0)	360.0	358.0–365.0
Immulite 2000 XPi	22		<20	n/a	32.9 (22.7)	32.3	23.1–51.9	302.2 (10.4)	305.0	247.0–357.0	388.0 (10.5)	388.0	303.0–476.0
Coat-A-Count	4	13.5 (5.2)	15.0	13.0–<20	27.3 (12.8)	27.0	24.0–31.0	383.3 (8.2)	383.0	352.0–415.0	485.0 (7.2)	490.0	444.0–516.0
Ortho Clinical Diagnostics (All)	8	9.3 (24.8)	8.9	7.1–12.8	26.8 (13.6)	27.0	21.3–27.3	296.8 (4.4)	293.5	285.0–320.0	395.2 (12.0)	406.0	320.2–441.0
VITROS ECiQ	4		8.9	7.5–12.8		25.5	21.3–27.3		288.0	285.0–301.0		427.0	406.0–441.0
VITROS 5600	4		8.5	7.1–10.0		27.0	27.0–32.6		299.0	288.0–320.0		351.0	320.2–382.0
Roche Diagnostics (All)	26	13.8 (20.0)	13.0	8.0–17.8	34.9 (12.2)	35.0	26.0–42.4	419.7 (5.6)	416.0	381.1–471.6	511.5 (5.4)	518.0	465.5–568.9
Cobas e401,e601,e602	7	13.7 (23.9)	13.0	9.8–15.8	33.3 (16.4)	33.0	26.0–40.7	414.8 (6.4)	410.0	381.1–447.7	511.6 (7.6)	504.0	465.5–568.9
MODULAR series	11	12.7 (22.3)	12.8	8.0–17.0	35.5 (9.3)	35.0	30.0–42.4	412.1 (2.9)	415.0	393.0–426.9	505.3 (5.6)	505.5	465.0–554.0
Elecsys	8	15.2 (13.9)	15.8	12.0–17.8	35.6 (12.7)	35.0	28.1–41.8	436.4 (6.4)	433.9	398.5–471.6	519.3 (3.7)	527.4	480.2–536.0
LC-MS/MS	6	19.5 (23.2)	17.2	16.0–27.1	30.9 (6.1)	30.4	29.0–34.0	417.8 (4.5)	414.0	399.0–440.0	532.0 (10.2)	507.0	470.0–589.0
Abbott AxSYM	3		25.2	24.7–26.0		44.7	29.9–51.0		370.9	336.2–436.0		505.8	485.0–517.5
Tosoh Bioscience AIA	2	No result reported					30.9–31.8			355.0–392.4	No result reported		
MP Biomedicals	1		31.0			36.0			423.0			392.0	
All method mean (%CV)		18.4 (29.1)			33.6 (8.2)			374.1 (12.2)			464.5 (12.6)		

[a]Results of participant laboratories ≤3 were not calculated for mean and %CV values.

[b]n = number of participating laboratories.

[c]%CV reflects among participant coefficient of variation.

Table 2

Mean bias (%) of laboratories' results against RMP-defined target[a].

Sample (target value) Assay Manufacturer Analytical System	Sample A (15.5 ng/dl)		Sample B (30.0 ng/dl)		Sample C (402 ng/dl)		Sample D (498 ng/dl)		Assay Manufacturer Calibration bias	
	n[b]	Mean bias in % (95% CI)	n	Mean bias in % (95% CI)	n	Mean bias in % (95% CI)	n	Mean bias in % (95% CI)	n	Mean bias in % (95% CI)
Beckman Coulter									20	13.1 (8.8 to 17.4)
Access 2	6	73.1 (58.4 to 87.9)	6	18.7 (10.1 to 27.4)	6	−8.8 (−11.2 to −6.3)	6	−3.9 (−7.5 to −0.4)		
Unicel DxI 800	9	51.7 (32.9 to 70.4)	11	15.8 (9.6 to 21.9)	11	−15.1 (−16.9 to −13.4)	10	−13.2 (−17.0 to −9.5)		
Siemens									75	−1.2 (−4.5 to 2.2)
ADVIA Centaur (XP, XPT)	34	37.9 (26.2 to 49.6)	39	21.3 (13.6 to 29.1)	39	−16.6 (−19.6 to −13.7)	39	−13.3 (−15.8 to −10.8)		
Immulite 1000			4	13.5 (−7.7 to 34.7)	4	−17.6 (−26.8 to −8.4)				
Immulite 2000 XPi			20	9.8 (−1.2 to 20.7)	20	−24.8 (−28.3 to −21.4)	21	−22.1 (−25.6 to −18.6)		
Coat-A-Count							4	−2.6 (−9.5 to 4.3)		
Roche Diagnostics									26	3.5 (0.1 to 6.9)
Cobas e401,e601,e602	5	−11.7 (−30.2 to 6.8)	7	11.1 (−2.4 to 24.5)	7	3.2 (−1.7 to 8.1)	5	2.7 (−4.1 to 9.6)		
MODULAR series	9	−17.8 (−29.7 to −5.8)	11	18.3 (11.8 to 24.7)	11	2.5 (0.7 to 4.3)	10	1.5 (−2.0 to 5.0)		
Elecsys	8	−2.1 (−11.5 to 7.3)	7	18.7 (7.5 to 29.9)	7	8.6 (3.4 to 13.7)	8	4.3 (1.6 to 6.9)		
LC-MS/MS	5	25.5 (0.0 to 51.1)	5	3.1 (−2.5 to 8.6)	5	3.9 (−0.2 to 8.1)	5	6.8 (−2.7 to 16.3)	6	8.0 (0.5 to 15.4)
Ortho Clinical Diagnostics									8	−24.1 (−28.6 to −19.6)

[a]Results from groups with 4 or more participants are reported.

[b]n = number of participating laboratories.

Table 3

Proficiency Evaluation of participant laboratories' results[c].

Assay Manufacturer Analytical System	n[e]	Laboratory (%) results within allowable limits[a] for Sample A–D				Laboratories achieved satisfactory performance[b] (%)
		Sample A (15.5 ng/dl)	Sample B (30.0 ng/dl)	Sample C (402 ng/dl)	Sample D (498 ng/dl)	
Beckman Coulter						
Access 2	6	0	83	100	100	83
Unicel DxI 600	2	0	50	100	100	50
Unicel DxI 800	10	10	90	100	100	90
Siemens						
ADVIA Centaur (XP, XPT)	36	42	50	89	97	56
ADVIA Centaur CP	3	33	66	100	66	33
Immulite 1000	3[d]	66	66	66	0	33
Immulite 2000 XPi	19[d]	100	79	58	68	68
Coat-A-Count	3	100	100	100	100	100
Ortho Clinical Diagnostics						
VITROS ECiQ	2	50	50	0	100	50
VITROS 5600	1	0	100	100	0	0
Roche Diagnostics						
Cobas e401,e601,e602	5	80	60	100	100	100
MODULAR series	9	78	89	100	100	100
Elecsys	7	100	71	100	100	100
LC-MS/MS	4	50	100	100	100	100
Abbott AxSYM	3	0	33	100	100	33
Tosoh Bioscience AIA	1	0	100	100	0	0
MP Biomedicals	1	0	100	100	100	100
Overall performance	115	55	71	86	89	73

[a]Results were within the limits of RMP-defined target ±25.1%.

[b]Satisfactory performance of a laboratory is defined as laboratory's results within the allowable limit for ≥3 of the 4 specimens.

[c]Only the laboratories that submitted results for all 4 samples were evaluated.

[d] Laboratories reported results as <20 and received a pass credit of 100%.

[e] n = number of participating laboratories.