Asian Journal of Andrology (2014) 16, 178–184

www.asiaandro.com; www.ajandrology.com

Open Access

INVITED REVIEW

Male Endocrinology

Challenges and improvements in testosterone and estradiol testing

Hubert W Vesper, Julianne C Botelho, Yuesong Wang

Assays that measure steroid hormones in patient care, public health, and research need to be both accurate and precise, as these criteria help to ensure comparability across all clinical and research applications. This review addresses major issues relevant to assay variability and describes recent activities by the US Centers for Disease Control and Prevention (CDC) to improve assay performance. Currently, high degrees of accuracy and precision are not always met for testosterone and estradiol measurements; although technologies for steroid hormone measurement have advanced significantly, measurement variability within and across laboratories has not improved accordingly. Differences in calibration and specificity are discussed as sources of variability in measurement accuracy. Ultimately, a combination of factors appears to cause inaccuracy of steroid hormone measurements, with nonuniform assay calibration and lack of specificity being two major contributors to assay variability. Within-assay variability for current assays is generally high, especially at low analyte concentrations. The CDC Hormone Standardization (HoSt) Program is improving clinical assays, as evidenced by a 50% decline in mean absolute bias between mass spectrometry assays and the CDC reference method from 2007 to 2011. This program provides the measurement traceability to CDC reference methods and helps to minimize factors affecting measurement variability.

Asian Journal of Andrology (2014) **16**, 178–184; doi: 10.4103/1008-682X.122338; published online: 20 January 2014

Keywords: CDC hormone standardization; CDC HoSt; estradiol; steroid hormone testing; testosterone

INTRODUCTION

Our understanding regarding the relevance of steroid hormones in health and disease has undergone substantial advancement, most notably with pertinence to testosterone and estradiol: where steroid hormones were once thought to be responsible for only the development of secondary sexual characteristics, current comprehension establishes that their significance is much broader, as exemplified by their influence over a wide range of biological functions within organs, such as bones and within the cardiovascular system as well. Steroid hormone levels that are excessive or deficient are associated with numerous diseases and chronic conditions such as hypogonadism, polycysticovary syndrome (PCOS), osteoporosis,[1] metabolic syndrome,[2,3] diabetes,[4,5] cancer,[6–8] and cardiovascular diseases.[9] The potential impact of testosterone deficiency on public health was estimated in one study, wherein it was suggested that, over a 20-year period, testosterone deficiency would be involved in the development of approximately 1.3 million new cases of cardiovascular diseases, 1.1 million new cases of diabetes mellitus, and over 600 000 osteoporosis-related fractures.[10] The total cost of evaluating and providing care to women with PCOS in the United States was estimated in 2005 to be US$4.36 billion.[11]

The increased awareness about the importance of steroid hormones relative to human health has stimulated further research efforts, the results of which have already led to new guidelines for patient care. Many of these new research findings suggest that even relatively small changes in steroid hormone levels can cause changes in health status and augment disease risk. It is therefore crucial that hormone levels be detected and measured accurately, even at low concentrations. The reliable detection of very low levels of hormones in blood is also required by some treatments, such as those using aromatase inhibitors. Assays for measuring steroid hormones in routine patient care, public health, and research must meet the aforementioned needs and requirements.

Detection of testosterone and estradiol in serum necessitates assays that accurately measure these compounds over a wide concentration range, which spans from less than 0.1 nmol l^{-1} to over 30 nmol l^{-1} for testosterone, and from less than 4 pmol l^{-1} to over to 10 000 pmol l^{-1} for estradiol.[12–14] Moreover, such assays should be sufficiently precise to enable the detection of biologically relevant differences in hormone levels, such as the detection of androgen excess in premenopausal women, elevation in postmenopausal women, and deficiency in men. They must also maintain a high level of sensitivity to assure reliable detection of hormone levels in children, to monitor patients on aromatase inhibitor therapy, and at the other extreme, to quantitate the hyperstimulation of steroidogenesis due to gonadotropin-releasing hormone agonist therapy. Finally, these assays must be specific enough to measure only the analyte of interest so as to allow the correct assessment of patient status and treatment response. This can be very challenging, especially considering that more than 100 conjugated and unconjugated

Clinical Chemistry Branch, Division of Laboratory Sciences, Centers for Disease Control and Prevention, Atlanta, Georgia, USA.
Correspondence: Dr. HW Vesper (hvesper@cdc.gov)
Received: 07 August 2013; Accepted: 08 August 2013

estradiol metabolites and exogenous hormones[15–18] can occur in blood and potentially interfere with the measurement.

The previously outlined requirements for steroid hormone assays are not met for all clinical and research applications, which has been noted by professional organizations[19–21] and experts alike.[22–27] Even a clinical guideline that acknowledges the benefits of measuring steroid hormone levels for comprehensive patient care nevertheless recommends an omission of testing, a suggestion made due to the lack of reliable analytical methods.[28] Currently, the lack of accurate and comparable estradiol measurements prevents the implementation of research findings in patient care. For example, certain clinical and epidemiological studies concluded that postmenopausal women with elevated levels of estradiol are at increased risk of breast cancer; however, a generally accepted, specific estradiol concentration with which to categorize women of increased risk cannot be defined. Because the measurements are not accurate, the concentrations that may be considered elevated are not comparable across studies,[29] and thus, women at increased risk of breast cancer cannot be identified using current estradiol tests.

Overcoming these challenges requires ongoing programs to continuously assess the analytical performance of assays, to identify potential problems, and to properly address their causes. Furthermore, these programs are needed to maintain the quality of these tests. This review discusses major factors affecting estradiol and testosterone measurements and describes activities by the US Centers for Disease Control and Prevention (CDC) to improve measurement reliability.

ASSAY TECHNOLOGIES FOR MEASURING STEROID HORMONES IN PATIENT CARE

Initially, steroid hormone measurements were performed using radioimmunoassays (RIAs). RIAs include steps to isolate hormones from serum before the actual RIA measurement, and these steps typically consist of some form of chromatography. Because the actual hormone measurement is performed after isolation of the analyte, these assays are called 'indirect' assays. To reduce cost, throughput, and the overall complexity of operating these indirect assays, so-called 'direct' assays were developed that measure steroid hormones directly in serum. These direct RIAs were further simplified by replacing radioisotope measurements with other techniques. Currently, almost all immunological testosterone and estradiol methods performed in patient care are direct assays.[20,30]

Analytical methods for measuring steroid hormones by using mass spectrometry have been in use since the 1960s. Similar to the indirect immunoassays, these methods include a step to isolate the steroid hormones from serum. After isolation, the steroid hormones are derivatized to enable further chromatographic separation of isomers and other structurally similar compounds by gas chromatography. After chromatographic separation, the steroid hormones are detected based on their specific mass using mass spectrometry. These methods are highly specific and can provide highly accurate measurements of multiple steroid hormones simultaneously. Because of their complexity, low throughput, and high specimen volume requirements, these methods were used only as reference methods and to address specific research questions.[31] Advancements in mass spectrometry enabled the use of liquid chromatography instead of gas chromatography, and the introduction of tandem mass spectrometry increased the detection specificity and sensitivity. Other technical developments enabled automation of sample preparation and analysis. As a consequence, liquid chromatography coupled with tandem mass spectrometry is increasingly used in patient care, public health, and research.[30–35]

Testosterone circulates in the blood either tightly bound to sex hormone binding globulin (SHBG), loosely bound to albumin, or unbound (free). Testosterone bound to SHBG is considered unavailable to target tissues, and only free and albumin-bound testosterone are considered bioavailable. For many clinical applications, measurement of total testosterone can be considered adequate for the evaluation of a patient. However, certain conditions, such as those where SHBG levels might be altered, may require measurement of free or bioavailable testosterone.

Free testosterone is traditionally estimated using equilibrium dialysis. With this method, serum is dialyzed across a semipermeable membrane that retains all protein-bound hormones, and the testosterone that crosses the membrane is measured and considered to represent free, unbound testosterone.[36–38] Other methods employ ultrafiltration to separate unbound and SHBG-bound hormones.[39] These methods can be very precise but are time-consuming, and they are difficult to operate and automate. To overcome these challenges, so-called 'analog-based assays' were developed, which intend to measure free hormone directly in serum. However, concerns have been raised about the reliability of analog-based assays,[40–42] leading to the recommendation not to use analog-based assays for determining free hormones.[19,27]

The profound technical improvements in steroid hormone measurement technologies resulted in a wide variety of methods available to physicians, researchers, and public health professionals. However, these technical improvements did not result in profound improvements in patient care. A large body of patient data and public health information is created with these different methods, as reflected in the increasing number of scientific publications, clinical studies, and increased demand for tests in patient care; however, generally accepted reference ranges, clinical decision levels, and guidelines for patient care do not exist or cannot be applied in most cases. One reason for this situation is the variability and lack of comparability of measurement results among different methods, laboratories, and over time. The variability in measurement results of testosterone and estradiol assays has been known for many years,[43–47] and it can be attributed mainly to differences in assay accuracy, specificity, and imprecision.

ASSESSMENT OF FACTORS CONTRIBUTING TO VARIABILITY IN MEASUREMENT ACCURACY

Variability in measurement accuracy, defined as closeness of agreement between a measured value and its true value, can be caused by inconsistent or inaccurate calibration or by interfering compounds.

Studies comparing measurement results for testosterone and estradiol obtained with immunoassays against those measurement results obtained with mass spectrometric assays found high disagreement between the assays.[46,48–51] In most studies, the absolute measurement bias between an assay and the comparison method was highest at low analyte concentrations and can be more than 100% with individual samples. The patterns of individual sample bias over the investigated concentration ranges differed among assays (**Figure 1**). One comparison study of commercial immunoassays for testosterone against a mass spectrometry method reported a reasonable overall agreement ($R^2 > 0.92$) at testosterone concentrations of >4 nmol l^{-1} and a mean bias ranging between 35% and 50%. The correlation was less strong (R^2, 0.59–0.97) at concentrations of <4 nmol l^{-1}, where the mean bias ranged from 5% to 220%.[52] Another study evaluated an immunoassay standardized for testosterone by the CDC Hormone Standardization (HoSt) Program against a mass spectrometric assay using samples from 3174 adult males (testosterone levels: >8 nmol l^{-1}).

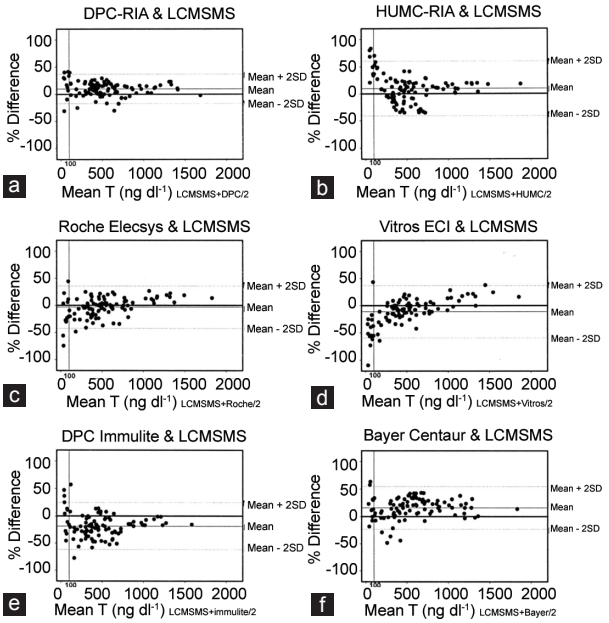

Figure 1: Differences in serum testosterone levels between commercial immunoassays and a liquid chromatography tandem mass spectrometry method of individual patient samples. Adapted with permission from Wang, *et al*.[49]

The authors found good agreement ($R^2 = 0.93$) between the two assays with the 95% confidence interval (CI) of the slope and intercept of the regression plot ranging from 0.96 to 0.99 and 0.18 to 0.65 nmol l^{-1}, respectively. The mean bias was 0.77%. No notable differences in the population distribution were observed when comparing data from both assays. However, the comparison of estradiol measurements performed in this population differed notably, with the 95% CI of the slope and intercept ranging from 1.15 to 1.25 and 0.75 to 7.81 pmol l^{-1}, respectively. The mean bias was 30.1%. The authors concluded that this immunoassay provides reliable data for testosterone in normal men, while it is unreliable for estradiol in men.[53]

Mass spectrometric methods are frequently used as a comparison method, because they are considered more specific than direct immunoassays. However, studies investigating the variability among mass spectrometry methods used in research and patient care found these methods to be inaccurate as well; relative to immunoassays, the mass spectrometric methods presented similar bias patterns, albeit much less pronounced.[54,55] Therefore, assessment and improvement of measurement accuracy need to include all assay technologies.

The different bias patterns among assays and the changes in bias with the analyte concentration within assays suggest that different factors affect the measurement accuracy of an assay and that each assay is affected differently by these factors.

VARIABILITY CAUSED BY DIFFERENCES IN MEASUREMENT SPECIFICITY

Typically, increasing positive bias with decreasing analyte concentration (**Figure 1b**) suggests that the assay is measuring other compounds in addition to the analyte and lacks specificity. Certain medications and other steroids such as estriol or dehydroepiandrosterone sulfate are known to potentially interfere with steroid hormone measurements.[44,56,57] Improvement in measurement accuracy has been shown upon introduction of liquid-liquid extraction or chromatography that removes compounds with similar characteristics from the sample prior to the actual measurement, especially at low analyte concentrations typically observed for testosterone in women and estradiol in men and postmenopausal women.[14,29,52,58]

Increasing negative bias with decreasing analyte concentration (**Figure 1d**) could be explained by incomplete dissociation of the hormone from the SHBG prior to measurement. Also, interfering compounds with higher affinity for the antibody relative to the analyte's affinity may cause incomplete recovery of the analyte, resulting in a negative bias; interfering compounds with low cross-reactivity but present at high concentration levels may also result in incomplete recovery.[53]

Measurement specificity can be assessed by determining the measurement accuracy after adding potentially interfering compounds to patient samples. However, the compounds interfering with the measurement are frequently not known. Therefore, measuring panels of individual patient samples with target values assigned by a reference measurement procedure can help to identify inaccuracies caused by interfering compounds as described later in this review.

VARIABILITY CAUSED BY DIFFERENCES IN MEASUREMENT CALIBRATION

Measurement bias that is constant and independent of the analyte concentration is typically caused by differences in calibration and is expressed as mean bias calculated from a panel of samples spanning the reportable range of the assay.

Calibration bias can be minimized by use of a common calibrator or reference material and by following ISO standard 17511 for calibration and target value assignment.[59] Pure compound, certified reference materials for testosterone and estradiol are available from the National Measurement Institute in Australia and the National Institute of Advanced Industrial Science and Technology in Japan, respectively. Many clinical immunoassays are optimized for measurements in serum and cannot use pure compound materials as calibrators. Serum-based reference materials created from pooled sera are available from the National Institute for Standards and Technology (NIST) for testosterone and the Institute for Reference Materials and Measurements (IRMM) for estradiol. Furthermore, panels of sera from individual donors with values for testosterone and estradiol assigned by a reference method are available from the CDC. The target values assigned to these serum-based reference materials are traceable to the certified, pure compound reference materials in line with requirements described in ISO standard 17511.

The use of a common calibrator and establishing metrological traceability as described in ISO 17511 in itself does not assure accurate measurements in patients. The process of calibration and assigning target values can introduce measurement bias if not conducted carefully. Non-commutable reference materials applied as calibrators or for calibration verification can result in or suggest measurement bias, especially with direct immunoassays optimized for use in patient care. The importance of commutability of materials used for calibration was emphasized in several reviews and commentaries.[60–62] Commutability is a material characteristic that describes how well measurement results obtained with a reference material correlate with measurement results obtained with patient samples. Protocols for assessing commutability of materials were established by the Clinical and Laboratory Standards Institute (CLSI).[63] An alternate approach that does not require commutability assessments employs the use

single donor patient samples for calibration or calibration verification. By using panels of patient samples for bias assessment rather than few pooled materials, the likelihood for an interfering compound to mask a calibration bias can be minimized.

Sera depleted of testosterone or estradiol using charcoal ('charcoal stripped' sera) are frequently used as matrices for preparing calibrator solutions, especially with mass spectrometric assays. Remaining charcoal fines in such modified sera can adsorb testosterone or estradiol and thus lead to inaccurate calibrator solutions. Therefore, the accuracy of calibrator solutions needs to be assured before they can be used as calibrators.

Serial dilution of calibrators can result in propagation of dilution errors and in inaccurate calibration solutions.[64] Minimizing serial dilutions and appropriately correcting for any dilution error can minimize this source of variability.

VARIABILITY IN PRECISION

Measurement precision, defined as the closeness of agreement between independent test results obtained on the same sample and with the same assay, mainly depends for the robustness of the measurement procedure and instrumentation used with the procedure. The assay precision is typically described by the imprecision and expressed as percent coefficient of variation (CV).

The within-assay imprecision for estradiol and testosterone assay is very inconsistent at low analyte concentrations. The imprecision of commercial immunoassays for testosterone was found to be <8% at concentrations >24 nmol l^{-1} and between 6.1% and 22% at concentrations <3 nmol l^{-1}.[48] The imprecision of commercial immunoassays for estradiol was reported to range from 1.2% to 42.6% CV at concentrations from 18 to 846 pg ml^{-1} (66–3105 pmol l^{-1}). At an estradiol concentration of 18 pg ml^{-1} (66 pmol l^{-1}), the imprecision was very inconsistent among these assays, and it became more uniform at levels of 93 pg ml^{-1}, with CVs ranging from 2.5% to 9.4% CV.[46] The reasons for these differences in within-assay imprecision are not fully understood. One explanation could be the analytical system having problems distinguishing the signal obtained from the instrument as a true analyte signal rather than background noise as the concentration approaches zero.

One study investigated the among-assay precision by comparing measurement results from the College of American Pathologists Survey (Y Ligand Special, September 2004), where the same sample was measured by different laboratories using the same testosterone assay. This study found that three out of nine assays were unable to produce peer group CVs lower than 10% at 12.8 nmol l^{-1}, a concentration level typically observed in men. Not a single assay was able to produce a peer group CV below 10% at 2.7 nmol l^{-1}, a concentration level typically observed in women. These findings were found to be consistent with data from other external quality assesment/proficiency testing (EQA/PT) programs. The authors pointed out that data from PT/EQA programs may also reflect other sources of variability, such as potential data entry errors, reagent lot-to-lot variability, and specimen stability. Despite these limitations, the high imprecision among laboratories using the same assay warrants further investigation.[65]

Mass spectrometric methods employ different types of sample preparation to isolate hormones from the serum matrix. Typical procedures are solid-phase extraction and liquid-liquid extraction, either with or without prior protein precipitation. One study investigated the impact of different sample preparation procedures on imprecision and data distribution. The researchers found that the different procedures can have different precisions. The authors assumed these observations to be related to differences in analyte recovery and removal of the hormone from SHBG.[66]

THE CDC HOST PROGRAM TO IMPROVE TESTOSTERONE AND ESTRADIOL MEASUREMENTS

The CDC HoSt Program helps in establishing traceability of measurements performed in patient care to a common reference material by assigning target values to individual donor sera using its reference methods.[67] These reference methods are calibrated with pure compound, certified reference materials. The use of single donor sera for calibration or calibration verification overcomes potential commutability issues frequently observed with pooled or otherwise modified sera. Aligning assays to a common reference material is accomplished in phase 1 of the CDC Hormone Standardization Program, where 40 samples along with target values are provided to participants for calibration or calibration verification (**Figure 2**).

To assure the established calibration remains accurate and consistent over time, the participant is challenged with 10 samples per calendar quarter with analyte concentrations unknown to the participant; evaluation of measurement bias is accomplished using data from four consecutive challenges. Testosterone assays with a mean bias of ±6.4% are considered sufficiently accurate and standardized. This bias criterion was reviewed and suggested by an expert group of the Partnership for the Accurate Testing of Hormones (PATH).[68] All participants were asked to perform replicate measurements on the same samples, which allowed for the assessment of assay imprecision as well as bias.

Measurement performance—specifically measurement accuracy—can change over time as new batches of calibrators and reagents are prepared and as instrumentation and other measurement conditions change. Interlaboratory comparison studies provide information about assay performance at one point in time, but they cannot detect changes in variability over time. Furthermore, results from different interlaboratory comparison studies performed with the same assays may not be comparable because of differences in study designs such as type and concentration range of samples used.

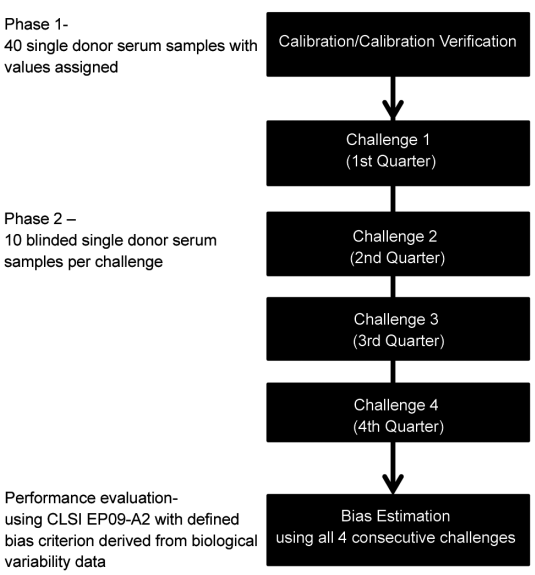

Figure 2: The US Centers for Disease Control and Prevention Hormone Standardization Program scheme for assessing assay calibration (Phase 1) and certifying assay accuracy (Phase 2) using single-donor serum samples.

The CDC HoSt Program is evaluating measurement accuracy over time through its quarterly challenges. The specimens used in this program are collected, processed, and shipped using standardized protocols. Reference values are assigned to these sera using reference methods recognized by Joint Committee for Traceability in Laboratory Medicine, and the measurement bias is assessed following an established protocol from the CLSI.[69] This allows the CDC HoSt Program to consistently and reliably assess measurement performance over time.

Figure 3 show the measurement bias observed with participants in the CDC HoSt Program over 2.5 years (10 quarterly challenges) for testosterone. The measurement bias for participant A is highly variable among the first quarterly challenges, with bias ranging between 10% and 20%. Using the information obtained from quarterly challenges, the participant was able to identify the sources of variability and to minimize measurement bias, which is reflected in the small measurement bias observed in the last three quarters.

With participant B, the individual sample biases obtained from male donor samples appear to be consistent within ±10% over 2.5 years. However, the samples from female donors consistently have high measurement biases of up to ±100% difference from the target value. These data suggest that the assay is calibrated consistently over time, but seems to have interferences that cause high measurement bias in female samples.

The individual sample bias observed with participant C on sera from male and female donors appears mostly consistent within each quarter, suggesting that the measurement accuracy does not change with the testosterone concentration or gender of the serum donor, and thus that the assay has high specificity. However, the mean bias from each set of quarterly samples seems to change over time, ranging between 20 and 10%. Because the assay shows a high degree of specificity, it can be assumed that the observed variability in measurement bias is caused mainly by variability in calibration. Further investigation revealed that changes in in-house prepared calibrator batches coincided with changes in measurement accuracy.

These observations demonstrate that assay accuracy can change over time and is best monitored using individual donor samples that cover the reportable range of the assays. Furthermore, the use of single-donor serum samples can help distinguish between calibration bias as well as bias due to interfering substances. This greatly facilitates identifying and addressing the source of measurement bias and variability among assays.

The impact of the CDC HoSt Program is indicated in improvements in measurement accuracy among mass spectrometric methods for testosterone (**Figure 4**). The among-laboratory variability for the mass spectrometric methods, expressed in the mean absolute bias calculated from the bias observed with different assays on individual samples, declined by approximately 50% across samples from 2007 to 2011. Improvements with immunoassays are anticipated.

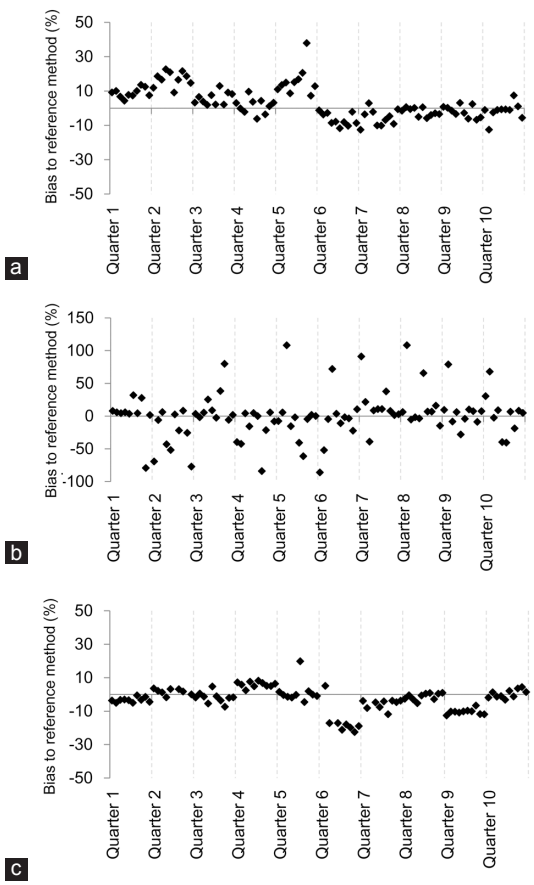

Figure 3: Measurement bias among three assays by HoSt participants (**a, b, c**) and the US Centers for Disease Control and Prevention reference laboratory on individual HoSt samples measured over 10 quarters.

SUMMARY

Clinical measurements of testosterone and estradiol are important for the diagnosis, treatment, and prevention of many diseases. Currently, the analytical performance of individual assays may not be appropriate to meet the needs of all clinical applications. This is especially true with regard to measurements of low analyte concentrations, such as those for testosterone observed in women and estradiol in men and postmenopausal women; accuracy in measuring low analyte concentrations remains challenging for most assays. Therefore, the reliability of a particular assay must be assessed in light of the intended use, and it cannot be generalized to all potential applications. The CDC HoSt Program is improving clinical assays by providing traceability of measurements to the highest available reference and by helping laboratories to understand factors affecting measurement variability. Further research and improvements are needed to minimize measurement variability and assure appropriate patient care.

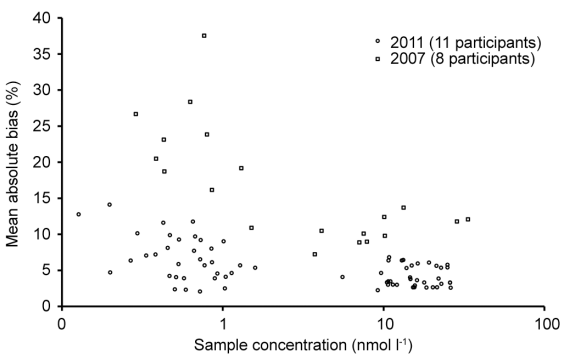

Figure 4: Mean absolute bias between mass spectrometry assays and the US Centers for Disease Control and Prevention reference method calculated from individual samples measured by eight participants in 2007 (squares) and by 11 participants in 2011 (circles).

COMPETING INTERESTS

The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the Centers for Disease Control and Prevention.

REFERENCES

1 Meier C, Nguyen TV, Handelsman DJ, Schindler C, Kushnir MM, *et al*. Endogenous sex hormones and incident fracture risk in older men: the dubbo osteoporosis epidemiology study. *Arch Intern Med* 2008; 168: 47–54.
2 Rodriguez A, Muller DC, Metter EJ, Maggio M, Harman SM, *et al*. Aging, androgens, and the metabolic syndrome in a longitudinal study of aging. *J Clin Endocrinol Metab* 2007; 92: 3568–72.
3 Bhasin S, Jasjua GK, Pencina M, D'Agostino R Sr, Coviello AD, *et al*. Sex hormone-binding globulin, but not testosterone, is associated prospectively and independently with incident metabolic syndrome in men: the framingham heart study. *Diabetes Care* 2011; 34: 2464–70.
4 Feldman HA, Longcope C, Derby CA, Johannes CB, Araujo AB, *et al*. Age trends in the level of serum testosterone and other hormones in middle-aged men: longitudinal results from the Massachusetts male aging study. *J Clin Endocrinol Metab* 2002; 87: 589–98.
5 Selvin E, Feinleib M, Zhang L, Rohrmann S, Rifai N, *et al*. Androgens and diabetes in men: results from the Third National Health and Nutrition Examination Survey (NHANES III). *Diabetes Care* 2007; 30: 234–8.
6 Wurtz AM, Tjonneland A, Christensen J, Dragsted LO, Aarestrup J, *et al*. Serum estrogen and SHBG levels and breast cancer incidence among users and never users of hormone replacement therapy. *Cancer Causes Control* 2012; 23: 1711–20.
7 Lin JH, Zhang SM, Rexrode KM, Manson JE, Chan AT, *et al*. Association between sex hormones and colorectal cancer risk in men and women. *Clin Gastroenterol Hepatol* 2013; 11: 419–24.
8 Zhang X, Tworoger SS, Eliassen AH, Hankinson SE. Postmenopausal plasma sex hormone levels and breast cancer risk over 20 years of follow-up. *Breast Cancer Res Treat* 2013; 137: 883–92.
9 Hougaku H, Fleg JL, Najjar SS, Lakatta EG, Harman SM, *et al*. Relationship between androgenic hormones and arterial stiffness, based on longitudinal hormone measurements. *Am J Physiol Endocrinol Metab* 2006; 290: E234–42.
10 Moskovic DJ, Araujo AB, Lipshultz LI, Khera M. The 20-year public health impact and direct cost of testosterone deficiency in U.S. men. *J Sex Med* 2013; 10: 562–9.
11 Azziz R, Marin C, Hoq L, Badamgarav E, Song P. Health care-related economic burden of the polycystic ovary syndrome during the reproductive life span. *J Clin Endocrinol Metab* 2005; 90: 4650–8.
12 Thienpont L. Meeting Report: First and Second Estradiol International Workshops. *Clin Chem* 1996; 42: 1122–4.
13 Demers LM. Testosterone and estradiol assays: current and future trends. *Steroids* 2008; 73: 1333–8.
14 Stanczyk FZ, Jurow J, Hsing AW. Limitations of direct immunoassays for measuring circulating estradiol levels in postmenopausal women and men in epidemiologic studies. *Cancer Epidemiol Biomarkers Prev* 2010; 19: 903–6.
15 Dawling S, Roodi N, Mernaugh RL, Wang X, Parl FF. Catechol-O-methyltransferase (COMT)-mediated metabolism of catechol estrogens: comparison of wild-type and variant COMT isoforms. *Cancer Res* 2001; 61: 6716–22.
16 Mueck AO, Seeger H, Lippert TH. Estradiol metabolism and malignant disease. *Maturitas* 2002; 43: 1–10.
17 Lee AJ, Cai MX, Thomas PE, Conney AH, Zhu BT. Characterization of the oxidative metabolites of 17beta-estradiol and estrone formed by 15 selectively expressed human cytochrome p450 isoforms. *Endocrinology* 2003; 144: 3382–98.
18 Lepine J, Bernard O, Plante M, Tetu B, Pelletier G, *et al*. Specificity and regioselectivity of the conjugation of estradiol, estrone, and their catecholestrogen and methoxyestrogen metabolites by human uridine diphospho-glucuronosyltransferases expressed in endometrium. *J Clin Endocrinol Metab* 2004; 89: 5222–32.
19 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: 405–13.
20 Rosner W, Hankinson SE, Sluss PM, Vesper HW, Wierman ME. Challenges to the measurement of estradiol: an endocrine society position statement. *J Clin Endocrinol Metab* 2013; 98: 1376–87.
21 Paduch DA, Brannigan RE, Fuchs EF, Kim ED, Marmar JL, *et al*. The Laboratory Diagnosis of Testosterone Deficiency. American Urology Association White Paper. 2013. Available from: http: //www.auanet.org/common/pdf/education/clinical-guidance/Testosterone-Deficiency-WhitePaper.pdf. [Last accessed on 2013 June 30].
22 Herold DA, Fitzgerald RL. Immunoassays for testosterone in women: better than a guess? *Clin Chem* 2003; 49: 1250–1.
23 Matsumoto AM, Bremner WJ. Serum testosterone assays-accuracy matters. *J Clin Endocrinol Metab* 2004; 89: 520–4.
24 Sacks SS. Are routine testosterone assays good enough? *Clin Biochem Rev* 2005; 26: 43–5.
25 Stanczyk FZ, Lee JS, Santen RJ. Standardization of steroid hormone assays: why, how, and when? *Cancer Epidemiol Biomarkers Prev* 2007; 16: 1713–9.
26 Swerdloff RS, Wang C. Free testosterone measurement by the analog displacement direct assay: old concerns and new evidence. *Clin Chem* 2008; 54: 458–60.
27 Morales A, Collier CP, Clark AF. A critical appraisal of accuracy and cost of laboratory methodologies for the diagnosis of hypogonadism: the role of free testosterone assays. *Can J Urol* 2012; 19: 6314–8.
28 Watts NB, Adler RA, Bilezikian JP, Drake MT, Eastell R, *et al*. Endocrine Society. Osteoporosis in men: an endocrine society clinical practice guideline. *J Clin Endocrinol Metab* 2012; 97: 1802–22.
29 Dowsett M, Folkerd E. Deficits in plasma oestradiol measurement in studies and management of breast cancer. *Breast Cancer Res* 2005; 7: 1–4.
30 Stanczyk FZ, Clarke NJ. Advantages and challenges of mass spectrometry assays for steroid hormones. *J Steroid Biochem Mol Biol* 2010; 121: 491–5.
31 Krone N, Hughes BA, Lavery GG, Stewart PM, Arlt W, *et al*. Gas chromatography/mass spectrometry (GC/MS) remains a pre-eminent discovery tool in clinical steroid investigations even in the era of fast liquid chromatography tandem mass spectrometry (LC/MS/MS). *J Steroid Biochem Mol Biol* 2010; 121: 496–504.
32 Giese RW. Measurement of endogenous estrogens: analytical challenges and recent advances. *J Chromatogr A* 2003; 1000: 401–12.
33 Soldin SJ, Soldin OP. Steroid hormone analysis by tandem mass spectrometry. *Clin Chem* 2009; 55: 1061–6.
34 Blair IA. Analysis of estrogens in serum and plasma from postmenopausal women: past present, and future. *Steroids* 2010; 75: 297–306.
35 Kushnir MM, Rockwood AL, Roberts WL, Yue B, Bergquist J, *et al*. Liquid chromatography tandem mass spectrometry for analysis of steroids in clinical laboratories. *Clin Biochem* 2011; 44: 77–88.
36 Kley HK, Bartmann E, Kruskemper HL. A simple and rapid method to measure non-protein-bound fractions of cortisol, testosterone and oestradiol by equilibrium dialysis: comparison with centrifugal filtration. *Acta Endocrinol* 1977; 85: 209–19.
37 Sinha-Hikim I, Arver S, Beall G, Shen R, Guerrero M, *et al*. The use of a sensitive equilibrium dialysis method for the measurement of free testosterone levels in healthy, cycling women and in human immunodeficiency virus-infected women. *J Clin Endocrinol Metab* 1998; 83: 1312–8.
38 Van Uytfanghe K, Stockl D, Kaufman JM, Fiers T, Ross HA, *et al*. Evaluation of a candidate reference measurement procedure for serum free testosterone based on ultrafiltration and isotope dilution-gas chromatography-mass spectrometry. *Clin Chem* 2004; 50: 2101–10.
39 Vlahos I, MacMahon W, Sgoutas D, Bowers W, Thompson J, *et al*. An improved ultrafiltration method for determining free testosterone in serum. *Clin Chem* 1982; 28: 2286–91.
40 Miller KK, Rosner W, Lee H, Hier J, Sesmilo G, *et al*. Measurement of free testosterone in normal women and women with androgen deficiency: comparison of methods. *J Clin Endocrinol Metab* 2004; 89: 525–33.
41 Van Uytfanghe K, Stockl D, Kaufman JM, Fiers T, De Leenheer A, *et al*. Validation of 5 routine assays for serum free testosterone with a candidate reference measurement procedure based on ultrafiltration and isotope dilution-gas chromatography-mass spectrometry. *Clin Biochem* 2005; 38: 253–61.
42 Fritz KS, McKean AJ, Nelson JC, Wilcox RB. Analog-based free testosterone test results linked to total testosterone concentrations, not free testosterone concentrations. *Clin Chem* 2008; 54: 512–6.
43 Middle JG. External quality assessment of steroid hormone assays in the United Kingdom. *Ann Ist Super Sanita* 1991; 27: 459–65.
44 Thomas CM, van den Berg RJ, Segers MF, Bartelink ML, Thien T. Inaccurate measurement of 17 beta-estradiol in serum of female volunteers after oral administration of milligram amounts of micronized 17 beta-estradiol. *Clin Chem* 1993; 39: 2341–2.
45 Boots LR, Potter S, Potter D, Azziz R. Measurement of total serum testosterone levels using commercially available kits: High degree of between-kit variability. *Fertil Steril* 1998; 69: 286–92.
46 Yang DT, Owen WE, Ramsay CS, Xie H, Roberts WL. Performance characteristics of eight estradiol immunoassays. *Am J Clin Pathol* 2004; 122: 332–7.
47 Kane J, Middle J, Cawood M. Measurement of serum testosterone in women; what should we do? *Ann Clin Biochem* 2007; 44: 5–15.
48 Taieb J, Mathian B, Millot F, Patricot MC, Mathieu E, *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: 1381–95.
49 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: 534–43.
50 Lee JS, Ettinger B, Stanczyk FZ, Vittinghoff E, Hanes V, *et al*. Comparison of methods to measure low serum estradiol levels in postmenopausal women. *J Clin Endocrinol Metab* 2006; 91: 3791–7.
51 Bell A, Meek CL, Viljoen A. Evidence of biochemical hyperandrogenism in women: The limitations of serum testosterone quantitation. *J Obstet Gynaecol* 2012; 32: 367–71.
52 Groenestege WM, Bui HN, ten Kate J, Menheere PP, Oosterhuis WP, *et al*. Accuracy of first and second generation testosterone assays and improvement through sample extraction. *Clin Chem* 2012; 58: 1154–6.

53 Huhtaniemi IT, Tajar A, Lee DM, O'Neill TW, Finn JD, *et al*. EMAS Group. Comparison of serum testosterone and estradiol measurements in 3174 European men using platform immunoassay and mass spectrometry; relevance for the diagnostics in aging men. *Eur J Endocrinol* 2012; 166: 983–91.
54 Thienpont LM, Van Uytfanghe K, Blincko S, Ramsay CS, Xie H, *et al*. State-of-the-art of serum testosterone measurement by isotope dilution-liquid chromatography-tandem mass spectrometry. *Clin Chem* 2008; 54: 1290–7.
55 Vesper HW, Bhasin S, Wang C, Tai SS, Dodge LA, *et al*. Interlaboratory comparison study of serum total testosterone [corrected] measurements performed by mass spectrometry methods. *Steroids* 2009; 74: 498–503.
56 Cao Z, Swift TA, West CA, Rosano TG, Rej R. Immunoassay of estradiol: Unanticipated suppression by unconjugated estriol. *Clin Chem* 2004; 50: 160–5.
57 Benton SC, Nuttall M, Nardo L, Laing I. Measured dehydroepiandrosterone sulfate positively influences testosterone measurement in unextracted female serum: Comparison of 2 immunoassays with testosterone measured by LC-MS. *Clin Chem* 2011; 57: 1074–5.
58 Cook NJ, Read GF. Oestradiol measurement in women on oral hormone replacement therapy: The validity of commercial test kits. *Br J Biomed Sci* 1995; 52: 97–101.
59 ISO. *In vitro* diagnostic medical devices-Measurement of quantities in biological samples-Metrological traceability of values assigned to calibrators and control materials. ISO document 17511, Geneva, *ISO* 2003.
60 Vesper HW, Miller WG, Myers GL. Reference materials and commutability. *Clin Biochem Rev* 2007; 28: 139–47.
61 Vesper HW, Thienpont LM. Traceability in laboratory medicine. *Clin Chem* 2009; 55: 1067–75.
62 Miller WG, Myers GL. Commutability still matters. *Clin Chem* 2013; 59: 1291-3.
63 CLSI. Characterization and qualification of commutable reference materials for Laboratory Medicine: approved guideline. CLSI document C53-A. Wayne, PA: CLSI 2010.
64 Higgins KM, Davidian M, Chew G, Burge H. The effect of serial dilution error on calibration inference in immunoassay. *Biometrics* 1998; 54: 19–32.
65 Lepage R. Measurement of testosterone and its sub-fractions in Canada. *Clin Biochem* 2006; 39: 97–108.
66 Vesper HW, Botelho JC. Standardization of testosterone measurements in humans. J *Steroid Biochem Mol Biol* 2010; 121: 513–9.
67 Rosner W, Vesper H. Preface. CDC workshop report improving steroid hormone measurements in patient care and research translation. *Steroids* 2008; 73: 1285.
68 Yun YM, Botelho JC, Chandler DW, Katayev A, Roberts WL, *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: 1703–10.
69 CLSI. Method comparison and bias estimation using patient samples: approved guideline. CLSI document EP9-A2. Wayne, PA: CLSI 2002.

How to cite this article: Vesper HW, Botelho JC, Wang Y. Challenges and improvements in testosterone and estradiol testing. *Asian J Androl* 20 January 2014. doi: 10.4103/1008-682X.122338. [Epub ahead of print]