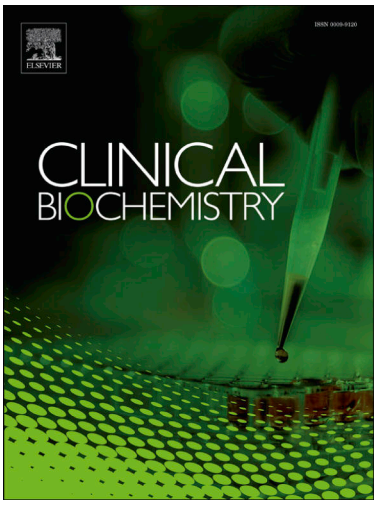
Journal Pre-proofs



Accuracy-based proficiency testing for estradiol measurements

Zhimin Tim Cao, Robert Rej, Hubert Vesper, J. Rex Astles

PII: S0009-9120(23)00228-X
DOI: https://doi.org/10.1016/j.clinbiochem.2023.110700
Reference: CLB 110700

To appear in: *Clinical Biochemistry*

Received Date: 29 August 2023
Revised Date: 26 November 2023
Accepted Date: 30 November 2023

Please cite this article as: Z. Tim Cao, R. Rej, H. Vesper, J. Rex Astles, Accuracy-based proficiency testing for estradiol measurements, *Clinical Biochemistry* (2023), doi: https://doi.org/10.1016/j.clinbiochem.2023.110700

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier Inc. on behalf of The Canadian Society of Clinical Chemists.

Accuracy-based proficiency testing for estradiol measurements

Commented [A1]: Please include 3-5 highlights of up to 85 characters each, per journal style.

Zhimin Tim Cao[1,2], Robert Rej[1,3], Hubert Vesper[4], J. Rex Astles[4]*

1. Wadsworth Center, New York State Department of Health, Albany, NY.

2. College of Arts and Sciences, University at Albany, State University of New York, Albany, NY.

3. Department of Biomedical Sciences, School of Public Health, University at Albany, State University of New York, Albany, NY.

4. Centers for Disease Control and Prevention, Atlanta, GA.

Cao's Current affiliation: Department of Pathology, Upstate Medical University, Syracuse, NY.

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 the New York State Department of Health.

*Address correspondence to this author at: Division of Laboratory Systems, Centers for Disease Control and Prevention, 1600 Clifton Road, NE, MS V24-3, Atlanta, GA 30329-4018. Phone: (404) 498-2296, E-mail: jda4@cdc.gov

Keywords: Estradiol, proficiency testing, standardization, accuracy

Nonstandard abbreviations: BV, bBiological Vvariation; CDC, US Centers for Disease Control and Prevention; CDC HoSt E2, CDC Hormones Standardization Program for estradiol; CLIA, Clinical Laboratory Improvement Amendments; LC-MS/MS, lLiquid Cchromatography-tandem mMass Sspectrometry; NYSDOH, New York State Department of Health; PT, Pproficiency tTesting.

HIGHLIGHTS

- Traditional proficiency testing (PT) cannot identify inaccurate estradiol methods
- Estradiol results from commutable PT samples were compared to CDC target values

Journal Pre-proofs

- Biases were 34% (-17% to 175%) and 40% (-33% to 386%) for the two lowest samples
- Scores were affected by the acceptance limit criteria and use of peer-grouping
- Improvement in estradiol measurement is needed, particularly at low concentrations

Journal Pre-proofs

ABSTRACT

Objectives: Accuracy of estradiol measurements is important but conventional proficiency testing (PT) cannot assess accuracy when possibly non-commutable samples are used and method peer-group means are the targets. Accuracy-based assessment of estradiol measurements is needed.

Design and Methods: Five serum samples were prepared from single- donors, frozen, and distributed overnight to 76 New York State Department of Health (NYSDOH)-certified laboratories. Participants analyzed samples for estradiol. The biases of group means were assessed against the Centers for Disease Control and Prevention (CDC)-defined targets, evaluated using the Hormones Standardization Program (HoSt) E2 performance criteria~~a~~on of ±12.5%. Each laboratory's performance was evaluated using total allowable error (acceptance limits) of target ±25% or ±15 pg/mL (55.1 pmol/L) (whichever was greater, NYSDOH), target ±30% (Clinical Laboratory Improvement Amendments [CLIA]), and target ±26% (minimal limit based on biological variation [BV]).

Results: The biases (range) were 34% (-17% to 175%), 40% (-33% to 386%), 16% (-45% to 193%), 5% (-27% to 117%), and -4% (-31% to 21%), for samples at estradiol of 24.1, 28.4, 61.7, 94.1 and 127 pg/mL, or 88.5, 104.3, 226.5, 245.4 and 466.2 pmol/L, respectively. Large positive method/analytical systematic biases were revealed for 9 commonly used method/analytical systems in the United States at low estradiol concentrations. Of the 9 analytical systems, 0, 0, 3, 7 and 6 met the HoSt criterion for the samples with estradiol at the five respective concentrations. PT evaluation showed that 59%, 69% and 87% of laboratories would receive a PT event pass score when the CDC-defined target and a criterion of NYSDOH, CLIA or BV was used, respectively. However, >95% laboratories would obtain PT pass score if method peer-group means were used as targets regardless of the criterion used.

Conclusions: Improvement in accuracy of estradiol measurements is needed, particularly at low estradiol concentrations. Accuracy-based PT provides unambiguous information about the accuracy of ~~analytical/method~~ methods/analytical systems.

Commented [A2]: This should be reversed "method/analytical" to be consistent

INTRODUCTION and OBJECTIVES

Estrogens are responsible for the development of the secondary female sex characteristics and play an important role in female reproductive processes. Estradiol measurements have a wide range of clinical utilities, e.g., diagnosis of fertility disorders, gynecomastia in males, estrogen-producing ovarian and testicular tumors, disorders of sex steroid metabolism, monitoring low-dose female hormone replacement therapy in postmenopausal women, and antiestrogen therapy (1, 2). To meet those clinical needs, accurate measurement of estradiol in patient care at all clinically relevant concentrations is needed; however, information on measurement accuracy is limited. Furthermore, ~~it~~ accurate measurement provides information that can be used to improve quality of manufacturers' products, to assess the effectiveness of manufacturer standardization, and to advance current evaluations performed as part of activities related to meeting regulatory requirements.

Although proficiency testing (PT) is an effective tool in monitoring quality performance of clinical laboratories and analytical systems, it has limitations (3). Conventional ~~proficiency~~

Journal Pre-proofs

~~testing~~PT often uses non-commutable samples or modified samples whose commutability is unknown, and therefore only evaluates participants' results using method peer-group mean values as targets (4). Therefore, due to presumed existence of matrix effects, conventional PT can only assess whether a laboratory's analysis can meet acceptance limits relative to its peers using the same method. Miller *et al.* (4) demonstrated that by using commutable materials and a *bona fide* reference method, it is possible to differentiate calibration bias from artifactual "matrix bias". However, as it is commonly performed, conventional PT typically cannot differentiate between calibration bias and matrix bias; therefore, it cannot assess whether the results obtained are sufficiently accurate to meet clinical needs (5). In contrast, accuracy-based PT~~,~~ uses authentic, unaltered samples and target values determined by a reference method measurement procedure. Thus, it can assess the proficiency of a laboratory analysis using an analytical system as intended: the accuracy, and reliability of measurement results obtained with the instrument in the context of clinical needs. Because, for many reasons, accuracy-based PT is relatively expensive to perform, it has seen somewhat limited use by external quality assessment~~EQA~~ programs. New economical approaches using commutable samples are needed.

In 2022, the US Centers for Medicare & Medicaid Services (CMS) finalized changes to the Clinical Laboratory Improvement Amendments of 1988 (CLIA) regulations for ~~proficiency testing~~PT, including an acceptance limit of ± 30% for scoring estradiol PT results (6). A more stringent evaluation criterion, of target ± 25% or 15 pg/mL (55.1 pmol/L) whichever is greater, was being used by the New York State Department of Health (NYSDOH) PT program at the time this study was done. This is consistent with the estimated "minimal" analytical performance (total error) (26%) based upon measurements of biological variation (BV) for estradiol (7, 8) and using an approach similar to that used by Miller *et al.* (4). For assessing accuracy of a method or analytical system, the Centers for Disease Control and Prevention (CDC) Hormone Standardization Program for estradiol (CDC HoSt E2) uses performance criteria derived from epidemiological studies (9) of ±12.5% bias for samples with estradiol of >20 pg/mL (73.4 pmol/L), and ±2.5 pg/mL (9.2 pmol/L) absolute bias for estradiol <20 pg/mL (73.4 pmol/L); in this study all specimens exceeded the 20 pg/mL (73.4 pmol/L) threshold concentration.

Objectives of this study were to assess accuracy of measurement procedures for total estradiol using an accuracy-based PT and to explore the effect of different possible acceptance limits using either the CDC-defined target or method peer-group mean.

DESIGN AND METHODS:

Five serum samples, prepared from apparently healthy single donors (2 male and 3 female) according to the procedure described in the Clinical Laboratory Standards Institute ~~CLSI~~ document C37A (10), were obtained from Solomon Park Research Laboratories and the sample collection process was approved by their institutional review board. These human serum specimens were screened and found negative for hepatitis B, hepatitis C and human immunodeficiency virus~~HIV~~. They were aliquoted to 1.0 ml fractions in 2.0 mL cryogenic vials (Corning Inc.) and stored at −80 °C until use within one year after collection. This study (not including sample collection process) was approved by the institutional review board of NYSDOH. The portion of the study conducted by the CDC laboratory was determined not to constitute engagement in human subject research.

Journal Pre-proofs

The serum specimens were distributed overnight, frozen on ice, to 76 NYSDOH-certified clinical laboratories. The laboratories were instructed 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. Estradiol has been shown to be stable in serum at these conditions (11). Participant laboratories were asked to (a) handle the serum samples in the same manner as patient samples for clinical testing, (b) analyze samples for total estradiol with their respective test methods (as shown in Fig. 2 and Tables 1 and 2), and (c) report results within 2 weeks of receipt. The CDC HoSt E2 Program established target values using its isotope dilution liquid chromatography tandem mass spectrometry (LC-MS/MS) reference measurement procedure (12).

We calculated the biases of participant laboratories' results against the CDC-defined target values, expressed as percent difference of each laboratory's results from the target values (Fig. 2). We grouped the results according to instrument/method. We then calculated the method/instrument peer-group means after excluding outliers, and the biases between the peer-group means against the CDC-defined target value (Table 2). We performed PT evaluation on all individual participant laboratories using the former NYSDOH PT program's acceptance limit of target ± 25% or target ± 15 pg/mL (55.1 pmol/L) (whichever was greater), the CLIA criterion of target ± 30% (6), and the criterion of target ± 26% based on the "minimal" requirement for allowable total error derived from the estimated ~~biological variation~~BV for estradiol (7, 8), respectively. The ± 26% "minimal" total error (TE) specification was obtained using the equation: $TE < 1.65 \times 0.75\ CV_W + 0.375\ (CV_W^2 + CV_G^2)^{1/2}$, with the most current median estimates of within-subject ~~biological variation (~~BV~~)~~ (CV_W = 15.0) and between-subject BV (CV_G = 13.0) provided by the European Federation of Clinical Chemistry and Laboratory Medicine ~~(EFLM)~~ website for estradiol (8). Statistical analyses were carried out using Microsoft 365 Excel programs. We used the method of Dixon (13) as modified by Reed *et al*. (14) to identify outliers.

RESULTS

The CDC-defined target values were 24.1 pg/mL (88.5 pmol/L) (Sample I), 28.4 pg/mL (104.3 pmol/L) (Sample II), 61.7 pg/mL (226.5 pmol/L) (Sample III), 94.1 pg/mL (345.4 pmol/L) (Sample IV), and 127 pg/mL (466.2 pmol/L) (Sample V). The 76 participant laboratories analyzed the five samples using 14 analytical systems. The mean, range of reported values, and coefficient of variation calculated with all reported results were 32 pg/mL (117.5 pmol/L) (20-66 pg/mL or 73.4-242.3 pmol/L, 28%) for Sample I, 40 pg/mL (146.8 pmol/L) (19-138 pg/mL or 69.7-506.6 pmol/L, 36%) for Sample II, 72 pg/mL (264.3 pmol/L) (34-181 pg/mL or 124.8-664.5 pmol/L, 53%) for Sample III, 99 pg/mL (363.4 pmol/L) (69-204 pg/mL or 253.3-748.9 pmol/L, 17%) for Sample IV, and 122 pg/mL (447.9 pmol/L) (88-154 pg/mL or 323-565.3 pmol/L, 14%) for Sample V. The mean bias (range) from all reported results against the CDC-defined target values for Samples I, II, II, IV and V was 34% (-17% to 175%), 40% (-33% to 386%), 16% (-45% to 193%), 5% (-27% to 117%), and -4% (-31% to 21%), respectively (Fig. 1).

Of the 14 ~~analytical~~ methods/analytical systems, 9 had ≥ 4 users accounting for 64 participant laboratories. Their method peer-group mean, median and range are shown in Table 1. Individual results for method groups of fewer than 4 participants are also listed. Of the 320 results reported from the 64 participant laboratories, 10 results from 4 laboratories of four different method

Commented [A3]: This word order was switched for consistency throughout

Journal Pre-proofs

groups were identified as outliers, therefore, they were excluded from the analysis for the method group mean (~~SD~~standard deviation), median, result range in the Table 1 and the method mean biases in the Table 2; however, these results were otherwise included in the rest of the analysis. Of 6 laboratories using Siemens Dimension Vista method, 5 reported < 20 pg/mL (74.3 pmol/L) and 1 reported 21 pg/mL (77.1 pmol/L) for Sample I, therefore only one data point is shown in Fig. 2 for that sample.

The peer-group means of Beckman Coulter and Roche systems had high positive biases at the low estradiol concentrations, observed in Samples I - III; a mixture of both slightly positive and obviously negative biases was seen in Sample IV; and slightly negative biases were seen for sample V at the highest concentration (Table 2). The peer-group means of Siemens systems had positive biases at all estradiol concentrations, while the Dimension showed biases within a range of -19.2% and 14.1% for samples II – V. (Fig. 2, Table 2). Assessment of biases for the ~~nine~~ 9 method/analytical systems~~,~~ that had more than 3 participants, using the CDC HoSt criterion showed that all 9 analytical systems exceeded the criterion for Samples I and II, while 3, 7 and 6 of the 9 analytical systems met the criterion for Sample III, IV and V, respectively (Table 2).

The evaluations of individual laboratory performance were first done using the respective CDC-defined target, but with the 3 different acceptance limits. The evaluation results are summarized in Table 3 (middle section for each sample). Laboratories' results falling within the acceptance limits according to each of the three evaluation criteria for ≥4 out of 5 samples received a satisfactory PT event score (right side of Table 3) according to the CLIA'88 criterion for scoring PT events (15). The percentage of laboratories with satisfactory performance for the PT event are summarized in Table 3 along with an average of actual PT event scores for each instrument/method peer group, as evaluated with the three different criteria. The evaluations were recalculated using each instrument/method peer-group mean as the target, and the 3 different acceptance limits as shown in Table 4.

DISCUSSION

The results showed that the magnitudes of biases were both concentration- and method-dependent. Upon examining all results, regardless of method peer groups, we observed a general trend toward positively-biased results at low estradiol concentrations, which declined as concentration increased up to about 100 pg/mL (367.1 pmol/L) (Fig. 1). To compare performances by method ~~-~~groups, we analyzed biases of each analytical method or peer group with users ≥ 4. We observed that all method groups produced highly biased results at low estradiol concentrations, i.e., Sample I and II, exceeding the CDC HoSt criterion, while Siemens Dimension Vista was not applicable for Sample I because ~~three~~ five users reported results of < 20 pg/mL (73.4 pmol/L) (Fig. 2, Table 2). Three method-group users (Siemens Immulite series, ADVIA Centaur, and Abbott Architect) had nearly half of their results beyond the NYSDOH acceptance limit at low estradiol concentrations for Sample II. This was consistent with an earlier report that 14 of 17 estradiol methods exceeded the suggested maximum allowable bias of ±12.5% (8). All the immunoassays showed this bias with decreasing estradiol concentrations, suggesting that compounds other than estradiol might contribute to the measurement result, such as estradiol analogs and metabolites as has been previously reported (16, 17).

Commented [A4]: The results section states 5 results <20, clarify.

Commented [A5R5]: Edited to correct. Should be "five"

Journal Pre-proofs

We observed a wide range of results reported by the participant laboratories on the same sample. In some cases, the highest values were about 7 times higher than the lowest value (Sample II). Overall, high variability was observed at low estradiol concentrations which may be due to differences in: (1) calibration of the assays, (2) differences in assay selectivity, which is more pronounced at low concentrations, and (3) differences in instrument operation or reliability. Such variability could lead to different clinical interpretations in patient care. Our study revealed that analytical performance of some existing methods cannot meet the clinical needs for testing low estradiol concentrations that may occur with various clinical conditions associated with post-menopausal women, breast cancer patients treated with aromatase inhibitors, men, and pre-pubertal/pubertal children (2, 17). A European Menopause and Andropause Society ~~(EMAS)~~ position statement uses an estradiol threshold of 50 pmol/L (13.6 pg/mL) to diagnose premature ovarian failure (18). In our study, for Sample I with a target value of 24.1 pg/mL (88.5 pmol/L), the reported results ranged from 20 to 66 pg/mL (73.4 to 242.3 pmol/L), and these could result in clinical misinterpretation. A similar observation was also reported in a study by Vesper *et al*. (9), which revealed not only the immunoassays' inaccuracy, but also a high inter-laboratory result variability at low estradiol concentrations. These immunoassays therefore could not guarantee consistent and reliable measurement results towards diagnosis of premature ovarian failure and other medical conditions seen in various subgroups of patient populations as mentioned above and elsewhere.

Of the 76 participant laboratories, two used laboratory-developed LC-MS/MS methods and results for Sample II showed about a two-fold difference, (19 vs 39 pg/mL or 69.7 vs 143.2 pmol/L), and both high positive and negative biases (-33.1% and 37%) relative to the CDC reference method, indicating that this technology can be subject to inaccuracy. It is common practice in a PT evaluation that if the number of participants in a method peer-group is low, then their scores are ungradable because the target value cannot be reliably defined. However, it is exactly these laboratories, in this case those using "laboratory-developed tests," that can most benefit from participating in proficiency testing if results *could* be evaluated. They could be evaluated if commutable samples were used because peer grouping would not be required.

The results of our PT evaluation performed for the participant laboratories certified by NYSDOH and CLIA showed that the majority of laboratories were able to meet the requirements set by NYSDOH, CLIA, and according to BV, but only at high estradiol concentrations. A PT evaluation criterion comprised of a combination of percentage and an absolute value was workable in this case, especially for PT samples containing low estradiol concentrations. Like the NYSDOH PT program, for many analytes PT program providers include more tolerant limits at lower concentrations by switching the criterion from a percentage to an absolute value or whichever be greater. The use of combination limits is sometimes necessary based upon the analytical performance that modern analyzers can achieve at low concentrations, but this was found to be unnecessary during pilot testing for the revised CLIA PT regulations. For some analytes this is reasonable because accuracy at very low concentrations may be clinically less important and when clinical interpretation is not affected. For estradiol it is important to be accurate at lower concentrations, as otherwise clinically-relevant decisions may be misguided.

Although peer grouping to score results is a common practice and the necessity of this practice has been demonstrated (19, 20), for practicality and cost, peer grouping to set targets is commonly performed without first establishing noncommutability of PT materials as CLIA regulations intended. Comparison of the average event scores and the percentage of laboratories

that would pass the PT event, i.e., achieve a satisfactory score, using a single target as shown in Table 3, versus using the peer-group target (Table 4) illustrates the extent to which scores are affected by this practice. Comparing data at the right-hand side of the Tables illustrate the substantial additional tolerance that peer grouping introduces. Without peer grouping to set the target, the overall PT scores using NYSDOH, CLIA, and BV criteria were 86%, 79% and 75%, respectively; whereas with peer grouping to set the target, the respective values were 96%, 94% and 92%. The pattern for laboratories with passing PT event scores was similar; nearly all laboratories achieved passing scores with peer grouping.

Ideally, ~~proficiency testing~~PT assesses the accuracy of both the analytical system and the ability of the laboratory operation for accurate analyses, both of which can affect analytical accuracy, and ultimately patient care. P~~roficiency testing~~T should not be unnecessarily punitive, but rather, serve as an effective tool for assuring that participant laboratories achieve minimally acceptable accuracy to support clinical needs. Ideally, it should also provide evidence to document improvements in accuracy over time. Unfortunately, in the modern practice of ~~proficiency testing~~PT there is lack of a reference method-defined target value and PT materials are presumed, rather than proven, to be noncommutable, thus the *de facto* practice is to use peer-group means as targets. Effective detection of inadequate analytical performance requires comparison to a single definitive target defined using a reference procedure as done in this study. Traditional PT that relies on peer grouping to set the targets cannot contribute to method standardization, nor to providing information to users on a method's quality performance. Furthermore, to be successful in detecting problems in a particular portion of the analytical measurable range, a PT challenge set should cover the dynamic range of most method systems (4).

In parallel with conventional PT, we suggest that accuracy-based assessment, similar to our approach, might become a routine practice- for clinically important analytes for which reference methodology is available. This could be similar to the method used by Miller *et al.* (4) but would only need a small, representative fraction of all laboratories in the PT program to participate voluntarily to estimate accuracy of the peer group. Such an approach need not be burdensome and would not be intended to identify poorly performing laboratories.- Using a small, but statistically valid number of participants to represent each method peer -group, it should be possible to estimate the peer-group mean and thereby identify methods with systematic problems. This approach conducted yearly or biennially, for example, and intentionally designed to eventually cover the dynamic range of most methods, could detect potentially inaccurate methods. This would be similar to the HoSt program's emphasis on using volunteers for assessing and improving the accuracy of method peer -groups, but it would be based on the analytical performance of individual end-user laboratories for detecting systematic post-marketing problems. With more reference methods being developed and available, this could eventually be applied to most analytes.

LIMITATIONS

Our findings were limited to one event of five samples, so we cannot make conclusions about accuracy over time. A relatively small number of laboratories participated, but these are representative of the available estradiol assays in use. We were not able to test the five samples for possible endogenous of exogenous steroids or drugs that could have interfered with some assays.

Journal Pre-proofs

CONCLUSIONS

In summary, we showed that at high (> 60 pg/mL or 220.3 pmol/L) estradiol concentrations in a majority of laboratories were able to meet minimal requirements for accuracy based upon ~~biological variability~~BV. However, we concluded that the immunoassay measurements for estradiol are inaccurate at low concentrations. Applying desirable or optimal criteria would result in more misses and a different overall assessment. Our results are consistent with reports of inaccuracy and variability in estradiol measurements, which could impact patient care, particularly for samples with low concentrations (16, 20 - 21). Efforts to standardize estradiol measurements were made over two decades ago by multiple institutions (2, 20, 22 – 24). This snapshot assessment of the accuracy of estradiol measurements would benefit by following up with similar accuracy-based assessments to determine whether these efforts are continuing to increase accuracy. Only accuracy-based PT, like the approach we describe here, can assess absolute accuracy. Future efforts would be best focused on lower concentrations, where improvements are most needed.

REFERENCES:

1. Cole TJ. Hormones. In: Rifai N, Horvath AR, Wittwer CT, editors. Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, 6th Ed. St. Louis (MO): Elsevier; 2018. p. 626-38. ISBN: 978-0-323-35921-4

2. 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. PMID: 23463657

3. Shahangian S. Proficiency testing in laboratory medicine: uses and limitations. Arch Pathol Lab Med. 1998;122:15-30

4. 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;132:838–46. doi.org/10.5858/2008-132-838-SOTAIT

5. Rej R, Norton-Wenzel CS, Cao TZ. Target values and method evaluation in proficiency testing programs. Clin Chem 2001;47:2185-6. doi.org/10.1093/clinchem/47.12.2185

6. Clinical Laboratory Improvement Amendments of 1988 (CLIA) Proficiency Testing Regulations Related to Analytes and Acceptable Performance. Fed Reg 2022;87:41194-242. https://www.govinfo.gov/content/pkg/FR-2022-07-11/pdf/2022-14513.pdf

(accessed August 2023)

Journal Pre-proofs

7. Fraser CG: Biological variation: from principles to practice. AACC Press, 2001, Washington DC. ISBN 10: 1890883492

8. European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Biological Variation Database. https://biologicalvariation.eu/meta_calculations (accessed October 2022).

9. Vesper HW, Botelho JC, Vidal ML, Rahmani Y, Thienpont LM, Caudill SP. High variability in serum estradiol measurements in men and women. Steroids. 2014;82:7-13. https://doi.org/10.1016/j.steroids.2013.12.005

10. 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.

11. Kushnir MM, Rockwood AL, Bergquist J, Varshavsky M, Roberts WL, Yue B, et al. High-sensitivity tandem mass spectrometry assay for serum estrone and estradiol. Am J Clin Pathol 2008;129:530–9. https://doi.org/10.1309/LC03BHQ5XJPJYEKG

12. Botelho JC, Ribera A, Cooper HC, Vesper HW. Evaluation of an Isotope Dilution HPLC Tandem Mass Spectrometry Candidate Reference Measurement Procedure for Total 17-β Estradiol in Human Serum. Anal Chem 2016;88:11123-9. DOI: 10.1021/acs.analchem.6b03220

13. Dixon WJ. Processing data for outliers. Biometrics 1953;9:74-89. https://doi.org/10.2307/3001634

14. Reed AH, Henry RJ, Mason WB. Influence of statistical method used on the resulting estimate of normal range. Clin Chem 1971;17:275-84.

15. Clinical Laboratory Improvement Amendments of 1988 (CLIA) Subpart H- Participation in Proficiency Testing for Laboratories Performing Nonwaived Testing. eCFR :: 42 CFR 493.843 -- Standard; Endocrinology. (accessed August 2023)

16. Jaque J, Macdonald H, Brueggmann D, Patel SK, Azen C, Clarke N, et al. Deficiencies in immunoassay methods used to monitor serum Estradiol levels during aromatase inhibitor treatment in postmenopausal breast cancer patients. SpringerPlus 2013;2:5 (2013). https://doi.org/10.1186/2193-1801-2-5

Journal Pre-proofs

17. 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.

https://doi.org/10.1158/1055-9965.EPI-10-0081

18. Vujovic S, Brincat M, Erel T, Gambacciani M, Lambrinoudaki I, Moen MH, et al. EMAS position statement: Managing women with premature ovarian failure. Maturitas. 2010;67:91-3. PMID: 20605383

19. Miller WG, Myers GL, Ashwood ER, Killeen AA, Wang E, Thienpont LM, et al. Creatinine measurement: state of the art in accuracy and interlaboratory harmonization. Arch Pathol Lab Med. 2005;129:297–304. https://doi.org/10.5858/2005-129-297-CMSOTA

20. Coucke W, Devleeschouwer N, Libeer JC, Schiettecatte J, Martin M, Smitz J. Accuracy andreproducibility of automated estradiol-17β and progesterone assays using native serum samples: results obtained in the Belgian external assessment scheme. Hum Reprod. 2007;22:3204–9. PubMed: 18025029

21. Yang DT, Owen WE, Ramsay CS, Xie H, Roberts WL. Performance characteristics of eight

estradiol immunoassays. Am J Clin Pathol. 2004;122:332–7. PubMed: 15362362

22. Thienpont L. Meeting report: first and second estradiol international workshop. Clin Chem 1996;42:112–4.

23. HoSt/VDSCP: Hormone and Vitamin D Standardization Programs. http://www.cdc.gov/labstandards/hs.html (Accessed August 2023)

24. Reinsberg J, Bätz O, Bertsch T, Bewarder N, Deschner W, Drescher V, et al. Precision and longterm stability of different estradiol immunoassays assessed in a multi-center quality control study. Clin Lab 2009;55:201–6. PubMed: 19728553

Figure legends

Journal Pre-proofs

Fig. 1. Distributions of result biases (% on X-axis) without method peer grouping are shown for the five samples by their corresponding target concentrations as a percentage bias from the CDC target (1 pg/mL = 3.671 pmol/L). The dotted lines indicate bias limits of ± 25%.

Fig. 2. Distributions of results biases (% on X-axis) with method peer grouping of users >4. Along the Y-axis are the seven method systems and their corresponding concentrations as a percentage bias from the CDC target for each sample. The dotted lines for ~~the~~ Samples I and II indicate the allowable limits according to the past NYSDOH PT criterion of target ± 15 pg/mL or 55.1 pmol/L (1 pg/mL = 3.671 pmol/L), whereas, for Samples III through IV the limits are~~of~~ target ± 25%.

Journal Pre-proofs

Table 1. A statistical summary of the reported results

Sample ID (Target value pg/mL)¥	Sample I (24.1)			Sample II (28.4)			Sample III (61.7)			Sample IV (94.1)			Sample V (127)		
Assay manufacturer Analytical system (n)	Mean (SD)	Median	Range	Mean (SD)	Median	Range	Mean (SD)	Median	Range	Mean (SD)	Median	Range	Mean (SD)	Median	Range
Abbott Architect (5)	29 (6.6)	30	20-36	43 (8)	45	31-51	74.8 (5.1)	77	66-79	91.4 (4)	90	88 - 98	123 (4.7)	121	118 - 128
Beckman Coulter (13)	31.1 (4.5)	31.5	23-39	33.5 (7.8)	34	19-47.5	75.4 (8.5)	71.4	54-81	85 (8.2)	83	78-94	111.7 (11.7)	115	88-126.8
UniCel DxI 600 (4)	31.3 (6.6)	31.5	23-39	27.7 (7.8)	30	19-34	66.3 (13.2)	65.5	54-80	95.8 (22.5)	88.5	78-128	104 (13)	104	88-119
UniCel DxI 800 (5)	31.3 (3.8)	33	27-34	37.4 (9.8)	39	24-47.5	75.5 (5.2)	76	69-81	95 (6.4)	93.5	89-104	108 (9.2)	108.5	97-119
Access (4)	30.7 (3.7)	31	26-35	33.9 (3.2)	34.1	30-37.5	69.6 (2.8)	69.9	66-72.7	96 (4.2)	95.9	92-100.2	122 (3.5)	121.1	119-126.8
Roche (12)	30 (4.7)	29.6	20-37	34.4 (3.7)	34.5	26.8-40	68.6 (3.9)	69	63-75	91.5 (11.6)	95	71 - 105	119.1 (12.7)	121.5	97 - 136
Roche e411 (4)	29.6 (5.1)	27.9	25.7-37	32.1 (1.4)	32	30.8 - 33.5	65.6 (3.2)	65	63-69.5	80.9 (14.9)	74.8	71-103	109.2 (18.2)	101.9	97-136
Roche 601 & 602 (8)	30.3 (4.9)	31.1	20-37	35.6 (4)	36.5	26.8-40	70 (3.5)	70.5	63.8-75	96.8 (4.4)	96.5	91-105	124.1 (5.4)	124.2	117.5-133
Siemens (34)	35.2 (8.2)	34.3	20-53	42.8 (8)	43.4	26-59.3	72 (12.4)	72.9	43-109	104.1 (7.7)	103.3	90-119.9	128.1 (17.6)	134.9	93-151.4
ADVIA Centaur (15)	38.4 (7.5)	38	28-53	47 (6)	45.3	38-56	76.6 (6.2)	77.3	66.3-92	106.7 (6.1)	105.6	96.8-117	138.7 (8.4)	137.8	122.7-151.4
Immulite 2000 (13)	32.4 (7.1)	33.2	20-42.2	43.1 (6.9)	43	32.2 - 59.3	74.6 (13.5)	72.9	56-109.3	104.1 (8.7)	102.7	91 - 119.9	128.4 (16.3)	134.9	96 - 148.2

Journal Pre-proofs

Dimension Vista (6)	<21	32.4 (4.8) 31 26-39.8	56.3 (9.1) 58.9 43-67.7	98.2 (6.1) 100.5 90 - 104	102.7 (8.3) 101.5 93 - 117
Tosoh AIA (3)	37.8, 47, 50.8	33.7, 54.4, 57	65.1, 76.4, 80	68.8, 80, 81.3	123.4, 140, 146.2
Siemens ADVIA Centaur CP (3)	26, 30, 37	28, 30, 39	55, 66, 75	77, 87, 93	90, 117, 135
BioMerieux Vidas (1)	35.1	25	70.9	69.8	95.1
Ortho Vitros ECi/ECiQ (3)	<20, 23.2, <20	27.9, 39.8, 41	44, 58.7, 65	91, 96.7, 110	106, 119, 122
LC-MS/MS (2)	20, 21	19, 39	62, 65	88, 113	109, 132

¥ 1 pg/mL = 3.671 pmol/L

Table 2. Mean bias (%) of instruments' and laboratories' results against CDC-defined targets

Sample ID (target value pg/mL) ¥	Sample I (24.1)	Sample II (28.4)	Sample III (61.7)	Sample IV (94.1)	Sample V (127)
Assay Manufacturer Analytical system (n)	Mean bias in % (95% CI)				
Abbott Architect (5)	20.3 (- 3.5 to 44.2)	54 (29.4 to 79.1)	21.2 (14 to 28.5)	-2.9 (-6.6 to 0.8)#	-3.2 (-6.4 to 0.1)
Beckman Coulter (13)					

Journal Pre-proofs

UniCel DxI 600 (4)	29.7 (3 to 56.3)	n/a (n = 3)	7.4 (-13.6 to 28.3)	-9.7 (-19.5 to 0.2)	-18.1 (-28.1 to -8.1)
UniCel DxI 800 (5)	30 (12.2 to 47.8)	31.6 (-2.3 to 65.5)	22.4 (14.1 to 30.6)	1.0 (-5.7 to 7.6)	-7.5 (-22.7 to 7.7)
Access (4)	27.5 (12.5 to 42.5)	19.4 (8.5 to 30.3)	12.8 (8.5 to 17.2)	2 (-2.4 to 6.4)	-3.9 (-6.6 to -1.3)
Roche (12)					
Roche e411 (4)	22.8 (2.1 to 43.5)	12.9 (8.2 to 17.7)	6.4 (1.3 to 11.4)	-14.1 (-29.6 to 1.5)	-14 (-28.1 to 0)
Roche 601 & 602 (8)	25.6 (11.5 to 39.6)	25.2 (15.4 to 34.9)	13.5 (9.5 to 17.4)	2.9 (-0.4 to 6.1)	-2.3 (-5.2 to 0.7)
Siemens (34)					
ADVIA Centaur (15)	59.4 (43.6 to 75.2)	65.4 (54.3 to 76.5)	24.2 (18.9 to 29.5)	13.3 (9.9 to 16.7)	9.2 (5.8 to 12.7)
Immulite 2000 (13)	34.5 (17.7 to 51.3)	51.6 (38.5 to 64.8)	20.9 (8.5 to 33.3)	10.6 (5.6 to 15.7)	1.1 (-5.9 to 8.1)
Dimension Vista (6)	n/a*	14.1 (0.6 to 27.2)	-8.8 (-20.7 to 3)	4.3 (-0.9 to 9.5)	-19.2 (-24.4 to -13.9)
	Bias (%) for each result (n<4)				
BioMerieux Vidas	45.6	-12	14.9	-25.8	-25.1
Ortho Vitros ECi/ECiQ	n/a, -3.7, n/a	-1.8, 40.1, 44.4	-28.7, -4.9, 5.3,	-3.3, 2.8, 16.9	-16.5, -6.3, -3.9
LC-MS/MS	-17, -12.9	-33.1, 37.3	0.5, 5.3	-6.5, 20.1	-14.2, 3.9
Tosoh AIA	56.8, 95, 110.8	18.7, 91.5, 100.7	5.5, 23.8, 29.7	-26.9, -15, -13.6	-2.8, 10.2, 15.1
Siemens ADVIA Centaur CP	7.9, 24.5, 53.5	-1.4, 5.6, 37.3	-10.9, 7, 21.6	-18.2, -7.5, -1.2	-29.1, -7.9, 6.3

Journal Pre-proofs

[¥] 1 pg/mL = 3.671 pmol/L

* Results were reported as <20 pg/mL

[#] High-lighted are the method means within the CDC HoSt criterion (±12.5%)

Table 3. Evaluation of participant laboratories' results using single CDC-defined target and three different criteria

Sample ID (target value pg/mL)[¥]	Laboratories (%) results within the allowable limits															Laboratories in % with pass-		
Assay Manufacturer	Sample I (24.1)			Sample II (28.4)			Sample III (61.7)			Sample IV (94.1)			Sample V (127)			PT** event score (Average event score)		
Analytical system (n)	NY*	CL#	BV[$]	NY	CL	BV	NY	CL	BV	NY	CL	BV	NY	CL	BV	NY	CL	BV
Abbott Architect (5)	100	60	60	400	200	200	800	100	80	100	100	100	100	100	100	80 (84)	60 (76)	60 (76)
Beckman Coulter (13)																		
UniCel DxI 600 (4)	100	50	25	100	100	75	75	100	75	75	75	75	100	100	75	100 (85)	75 (80)	50 (70)
UniCel DxI 800 (5)	60	40	40	80	40	40	40	60	40	80	80	80	100	100	100	80 (72)	40 (64)	40 (60)
Access (4)	100	75	25	100	75	75	100	100	100	100	100	100	100	100	100	100 (100)	75 (90)	75 (80)
Roche (12)																		
Roche e411 (4)	100	75	75	100	100	100	100	100	100	100	100	100	100	100	100	100 (100)	100 (95)	100 (95)

Journal Pre-proofs

Roche 601 & 602 (8)	100	50	50	100	50	38	100	100	100	100	100	100	100	100	100	100	100	100	100
	(100)			(100)			(100)			(100)			(100)			(100)			
Siemens (34)																			
ADVIA Centaur (15)	60	13	13	27	00	00	47	73	53	93	93	93	93	93	100	47	13	13	13
	(64)			(77)			(73)			(93)			(93)			(64)			
Immulite 2000 (13)	93	38	38	54	51	11	69	77	69	92	100	100	100	100	100	77	38	38	38
	(82)			(94)			(89)			(92)			(92)			(82)			
Dimension Vista (6)	100	100	100	100	86	67	83	83	83	100	100	100	100	100	100	100	100	100	100
	(87)			(93)			(93)			(93)			(93)			(87)			
BioMerieux Vidas (1)	100	0	0	100	00	00	100	00	00	00	100	100	100	100	100	100	100	100	100
	(80)			(80)			(80)			(80)			(80)			(80)			
Ortho Vitros ECi/ECiQ (3)	100	100	100	100	33	33	66	00	66	00	100	100	100	100	100	100	100	100	66
	(93)			(93)			(87)			(93)			(93)			(93)			
LC-MS/MS (2)	100	100	100	100	55	00	100	00	00	00	100	100	100	100	100	100	100	100	60
	(100)			(100)			(90)			(100)			(100)			(100)			
Tosoh AIA (3)	33	00	00	33	33	33	66	00	66	00	66	66	66	66	66	33	33	33	0
	(60)			(60)			(66)			(60)			(60)			(60)			
Siemens ADVIA Centaur CP (3)	100	66	66	100	66	66	100	00	00	00	100	100	100	100	100	100	66	66	66
	(100)			(100)			(87)			(100)			(100)			(100)			
Average	89	55	49	81	55	50	81	00	92	81	96	96	94	100	100	98	87	69	59
	(86)			(91)			(86)			(86)			(86)			(86)			

[¥] 1 pg/mL = 3.671 pmol/L

*NY indicates a NYSDOH criterion of Target ± 25% or 15 pg/mL whichever is greater.

[#]CL indicates the CLIA criterion of Target ± 30%

[$]BV indicates a criterion of Target ± 26% derived from biological variability.

Journal Pre-proofs

Journal Pre-proofs

**Pass-PT score is obtained if a laboratory has 4 out 5 sample results within the allowable limits.

Journal Pre-proofs

Table 4. Evaluation of participant laboratories' results using peer-group mean as target and three difference criteria

Sample ID (target value pg/mL)[¥]	Laboratories (%) results within the allowable limits															Laboratories in % with pass-		
Assay Manufacturer	Sample I (24.1)			Sample II (28.4)			Sample III (61.7)			Sample IV (94.1)			Sample V (127)			PT event score (Average score)		
Analytical system (n)	NY*	CL#	BV[$]	NY	CL	BV	NY	CL	BV	NY	CL	BV	NY	CL	BV	NY	CL	BV
Abbott Architect (5)	100	80	80	100	100	80	100	100	100	100	100	100	100	100	100	100 (100)	100 (96)	100 (92)
Beckman Coulter																		
UniCel DxI 600 (4)	100	100	75	100	75	50	100	100	100	80	80	80	100	100	100	100 (95)	100 (90)	75 (80)
UniCel DxI 800 (5)	80	80	60	100	80	80	80	80	80	80	80	80	80	80	80	80 (80)	80 (80)	80 (76)
Access (4)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100 (100)	100 (100)	100 (100)
Roche																		
Roche e411 (4)	100	100	100	100	100	100	100	100	100	80	80	80	100	100	100	100 (95)	100 (100)	100(95)
Roche 601 & 602 (8)	100	88	88	100	100	100	100	100	100	100	100	100	100	100	100	100 (100)	100 (98)	100 (98)
Siemens																		
ADVIA Centaur (15)	100	87	67	93	93	93	93	93	93	93	93	93	93	93	93	100 (100)	100 (97)	100 (96)
Immulite 2000 (13)	100	69	69	100	92	92	77	85	85	100	100	100	92	100	100	100 (94)	100 (89)	100 (89)

Journal Pre-proofs

Dimension Vista (6)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100 (100)	100 (100)	100 (100)
BioMerieux Vidas (1)	Non-gradable results																			
Ortho Vitros ECi/ECiQ (3)	Non-gradable results																			
LC-MS/MS (2)	Non-gradable results																			
Tosoh AIA (3)	Non-gradable results																			
Siemens ADVIA Centaur CP (3)	Non-gradable results																			
Average	98	89	82	99	93	88	94	95	95	93	95	93	96	97	97	98 (96)	98 (94)	95 (92)		

[¥] 1 pg/mL = 3.671 pmol/L

*NY indicates a NYSDOH criterion of Target ± 25% or 15 pg/mL whichever is greater.

[#]CL indicates the CLIA criterion of Target ± 30%

[$]BV indicates a criterion of Target ± 26%, derived from biological variability.

Journal Pre-proofs