Valentin Braun, Uta Ceglarek, Alexander Gaudl, Joanna Gawinecka, Daniel Müller, Manfred Rauh, Matthias Weber and Christoph Seger*

Evaluation of five multisteroid LC–MS/MS methods used for routine clinical analysis: comparable performance was obtained for nine analytes

https://doi.org/10.1515/cclm-2023-0847
Received August 4, 2023; accepted November 3, 2023;
published online December 4, 2023

Abstract

Objectives: A mass spectrometry (LC–MS/MS)-based interlaboratory comparison study was performed for nine steroid analytes with five participating laboratories. The sample set contained 40 pooled samples of human serum generated from preanalyzed leftovers. To obtain a well-balanced distribution across reference intervals of each steroid, the leftovers first underwent a targeted mixing step.
Methods: All participants measured a sample set once using their own multianalyte protocols and calibrators. Four participants used in-house developed measurement platforms, including IVD-CE certified calibrators, which were used by three participants; the 5th lab used the whole LC–MS kit from an IVD manufacturer. All labs reported results for 17-hydroxyprogesterone, androstenedione, cortisol, and testosterone, and four labs reported results for 11-deoxycortisol, corticosterone, cortisone, dehydroepiandrosterone sulfate (DHEAS), and progesterone.
Results: Good or acceptable overall comparability was found in Bland–Altmann and Passing–Bablok analyses. Mean bias against the overall mean remained less than ±10 % except for DHEAS, androstenedione, and progesterone at one site and for cortisol and corticosterone at two sites (max. –18.9 % for androstenedione). The main analytical problems unraveled by this study included a bias not previously identified in proficiency testing, operator errors, non-supported matrix types and higher inaccuracy and imprecision at lower ends of measuring intervals.
Conclusions: This study shows that intermethod comparison is essential for monitoring the validity of an assay and should serve as an example of how external quality assessment could work in addition to organized proficiency testing schemes.

Keywords: LC–MS/MS; endocrinology; steroid measurement; laboratory medicine; method comparison

***Corresponding author: Dr. Christoph Seger**, PD, Institute of Pharmacy/Pharmacognosy, CCB – Centrum of Chemistry and Biomedicine, University of Innsbruck, Innrain 80-82, 6020 Innsbruck, Austria; and Dr. Risch Ostschweiz AG, Lagerstrasse 30, 9470 Buchs, Switzerland, E-mail: christoph.seger@uibk.ac.at. https://orcid.org/0000-0002-3984-461X
Valentin Braun, Institute of Pharmacy/Pharmacognosy, CCB – Centrum of Chemistry and Biomedicine, University of Innsbruck, Innsbruck, Austria; and Dr. Risch Ostschweiz AG, Buchs, Switzerland. https://orcid.org/0000-0002-8991-0646
Uta Ceglarek and Alexander Gaudl, Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, Leipzig University, Leipzig, Germany. https://orcid.org/0000-0002-4034-5535 (U. Ceglarek). https://orcid.org/0000-0002-7197-3868 (A. Gaudl)
Joanna Gawinecka, Institute of Clinical Chemistry, University Hospital Zurich, Zürich, Switzerland. https://orcid.org/0000-0003-3859-0934
Daniel Müller, Institute of Clinical Chemistry, University Hospital Zurich, Zürich, Switzerland; and Department of Clinical Chemistry and Laboratory Medicine, University Hospital Basel, Basel, Switzerland
Manfred Rauh, Department of Pediatrics and Adolescent Medicine, University Hospital Erlangen, Erlangen, Germany
Matthias Weber, Bioscientia MVZ Labor Karlsruhe, Karlsruhe, Germany

Introduction

Modern laboratory diagnostics increasingly rely on liquid chromatography–mass spectrometry (LC–MS/MS) methods for measuring steroidal hormones. Over the past two decades, the possibilities and limitations of this technology have become well understood [1–3]. Metrological traceability has been demonstrated for several key analytes, and the IVD industry has established certified assay formats [1, 4, 5]. In addition, the in-house development of measurement methods (lab-developed tests, LDTs) is widespread in the field of endocrinology. For example, to target prepubertal estradiol and testosterone levels, the measurement range of commercially available kits must be significantly extended at its lower end; as a result, clinical needs remain unmet in certain populations (e.g., in pediatrics) [6, 7]. In addition, the sensitivity of IVD kits is often

limited due to poorly optimized chromatography and mass spectrometry settings [8]. A mixed form is achieved by LDTs, which use certified and traceable commercial IVD calibration systems to minimize the bias input into the method, which is possibly associated with the production of calibrator samples.

Overall, the situation is complex, but the monitoring interlaboratory testing schemes showed a satisfactory outlook; these schemes first and foremost include the United Kingdom National External Quality Assessment Service (UK NEQAS) proficiency testing (PT) scheme for steroids, which works with real patient materials [9]. Furthermore, mass spectrometric laboratories obtain measurements with comparable precision or better than the analytical realization on the immunological high-throughput automats. However, mass spectrometric examination procedures are decisively advantageous because with these methods, an analytical design that is principally free of interference can be obtained and patient-specific systematic errors due to nonspecific measurements of the measurand in the ligand-binding assay ("cross-reactivity") can be suppressed.

In the present study, five laboratories from Switzerland and Germany, designated Lab A–Lab E, attempted to test a central hypotheses that explains the advantage of using mass spectrometry in routine clinical diagnostics. All participating laboratories possess multianalyte LC–MS/MS methods for diagnostic use; all laboratories are accredited according to either ISO 15189 or ISO 17025. Most of these laboratories use IVD-certified calibrator materials from one vendor (Supplementary Table S1). The aims of this study were threefold. First, the study was performed to investigate the level by which random and systematic errors contribute towards the results. Second, evaluations were performed to determine whether these figures of merit were comparable to the interlaboratory scatter found in the UK NEQAS PT scheme. As a third goal, the desirable total allowable error (TAE) derived from biological variation (BV) data was compared to the experimentally found intra- and interlaboratory errors to unveil whether these numbers exceed the TAE goals [10, 11]. The role of using IVD-certified calibrator materials is of special interest since the use of these materials may lead to reduced interlaboratory deviations [12–15]. To test these hypotheses as realistically as possible, one participant purposefully prepared 40 multianalyte pooled serum samples that covered the reference interval of the individual analytes as completely as possible. In this respect, the approach differs from the recently published HarmoSter study, which focused more on targeted individual sample analyses and material comparisons [12, 13].

Materials and methods

Study design

Five laboratories participated in this comparison study. Lab A coordinated the study, provided the sample sets, and gathered and analyzed the results. Pooled serum samples were used that contained deidentified leftover materials from authentic patient samples, which were collected from the laboratory archive. Therefore, no ethical approval was necessary for this study. All participating laboratories analyzed the samples using their own in-house protocols and materials for their multisteroid LC–MS/MS methods, some of which were also published previously [8, 16, 17]. Up to 15 analytes were measured by a single method, including androstenedione, corticosterone, cortisol, cortisone, dehydroepiandrosterone sulfate (DHEAS), 11-deoxycortisol, 21-deoxycortisol, 17-hydroxyprogesterone, estradiol, progesterone, and testosterone. Method details and covered analytes for each laboratory based on a standardized reporting format [18] are summarized in the Supplementary Table S1.

Sample collection, storage, and distribution

Forty samples from pooled human serum were specially created for this experiment from leftovers patient samples with the target to cover the maximum the biological reference intervals as well as pathological values of every analyte. For this purpose, reference intervals published by Eisenhofer et al. were selected for comparison because the intervals were determined in a population from the same geographic region as the participating laboratories [19]. The procedure used to create the serum pools is described in detail in the Supplementary Text S1. All sets of the 40 pooled samples were stored at −20 °C at Lab A between production and measurements or shipments. Sample sets were shipped to Labs B–E on dry ice with an express service providing next-day delivery.

LC–MS/MS measurements and sample stability evaluation

In principle, each lab conducted one measurement per sample by their own in-house protocol, instrumentation, and procedure, as summarized in the Supplementary Table S1. Three laboratories performed a second round of measurements six months after the first measurement. Lab C analyzed a complete set of samples a second time due to analytical problems in the first round, and Lab E split the analysis of the sample set into two batches with 20 samples each. Lab A reanalyzed the whole sample set after six months to evaluate sample stability and to support longitudinal comparability of the overall measurement results.

Data analysis

Only results above the limits of quantification (LOQs) of the specific methods were considered for data analysis. Analytes with results from less than four participating laboratories and individual samples with less than three reported results were excluded from the analysis. To calculate the bias of a measurement result, the mean of all laboratories' measurement results for the same sample (overall mean) was used as a reference. Agreement of the different LC–MS/MS methods was assessed

by comparing their individual results against the overall mean using Bland–Altman plots [20] and Passing–Bablok regression analysis [21]. An average interlaboratory CV (iCV) was calculated for every analyte from CVs that were calculated for each pooled sample from the measurements of all laboratories. For analytes androstenedione, cortisone, DHEAS, estradiol, progesterone and testosterone, iCV data of this study were compared to iCV data from the UK NEQAS PT scheme by analyzing results from 24 distributions (PT rounds 470–494).

In addition, iCV and overall bias variability data were compared to the TAE, a performance criterion commonly used in clinical chemistry. The TAE of an analyte is based on the concept of deriving intra- and interindividual variation values (CVi and CVg, respectively) from the estimated BV of the investigated analysis. When available, CVi and CVg values were sourced from a BV database hosted by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) [11, 22, 23]. For androstenedione and cortisone, CVi and CVg values were obtained from recent publications [24, 25] and for 11-deoxycortisol from the Westgard Database [26]. TAE values were calculated and categorized according to the EFLM database recommendations. Only TAE figures of merit rated as "desirable" and "optimal" were used for performance evaluations in this study. Analogous to the HarmoSter study, 11-deoxycortisol TEA was also used for corticosterone, for which no biological variation data were available [13]. For further details, see Supplementary Table S2.

Results

The following analytes were included in this comparison study: 17-hydroxyprogesterone, androstenedione, cortisol, and testosterone, with all five participants reporting results, and 11-deoxycortisol, corticosterone, DHEAS, cortisone and progesterone, with results available from four sites. Depending on method performance, some results were excluded from data analysis because the concentrations were below the respective locally established LOQ; one sample result was affected for testosterone and 11-deoxycortisol (Labs A and C, respectively), and eight sample results had to be excluded for DHEAS (Lab B). For progesterone, a total of 16 sample results from multiple sites were below the LOQ (Labs B–D); thus, three samples were completely excluded because the limit of three participant results per sample was not achieved. For further details, see Table 1. After the first samples were sent out to the participating laboratories, six months were needed for all samples to be processed at all sites. Sample stability over this period was assessed in Lab A and found to be sufficient (for details, see Supplementary Text S2).

The concentrations realized in the pooled samples ranged from 0.1 nmol/L for testosterone to more than 8,000 nmol/L for DHEAS. These values matched very well with biological reference intervals for healthy adults and included samples clearly below (e.g., cortisol) or above (e.g., progesterone) these cutoffs. These concentrations can be found in certain healthy subpopulations, such as children or pregnant women, or under individuals taking medications, e.g., glucocorticoid treatment. The data analysis results for these samples as the mean and range, bias and iCV data are summarized in Table 1. Bias values against the overall mean for individual samples are illustrated for each laboratory in Bland–Altman styled plots in the Supplementary Figures S1 to S9. Generally, mean bias values were greater than ±10 % for all nine analytes at each laboratory. For androstenedione, DHEAS and progesterone, there were mean deviations from the overall mean of >10 % at one site and for cortisol and corticosterone at two sites. The maximum mean bias observed was −18.9 % at Lab C for androstenedione.

The SD of bias was constantly <10 % at all laboratories apart from Lab D, with an elevated SD of bias ranging from 8.5 to 22.3 % for all measured analytes. The SD of bias was elevated with all methods for the analytes progesterone and testosterone compared to other analytes, mainly due to higher variations at lower concentrations. The SD of bias ranged between 6.1 and 15.3 % and between 18.0 and 46.8 % for testosterone and progesterone, respectively. If the analysis is limited to samples with an overall mean of below 2 nmol/L or below 3 nmol/L for progesterone or testosterone, respectively, the SD of bias was 7.7–18.9 % for testosterone (depending on the laboratory) and 23.7–67.5 % for progesterone; for samples above those cutoffs, the SD of bias was 2.6–10.0 % and 3.5–9.7 % for testosterone and progesterone, respectively (see Supplementary Figures S10–S13).

Plots of Passing–Bablok regression analysis conducted with the results of all laboratories are displayed in Figure 1. Detailed results from Passing–Bablok regression analysis for intercept, slope, and the coefficient of correlation (r^2) together with their respective 95 % CIs are summarized in Table 2.

Good agreement between laboratories was observed for the analytes 17-hydroxyprogesterone, cortisone and progesterone, with the 95 % CIs of the individual (Lab A–Lab E) slopes overlapping with the overall 95 % CI. For all other analytes, overlapping 95 % CI intervals were attained for all but one method per analyte (see bolded values in Table 2). The results obtained for the LC–MS/MS methods used in Lab C and Lab B tended to be lower than those of the other participants for androstenedione, cortisol, and corticosterone and for DHEAS, respectively, while the results of Lab B and Lab D were higher on average for 11-deoxycortisol and testosterone, respectively.

Overall bias variability was compared to desirable TAE values by utilizing Bland–Altman styled plots (Figure 2). For 17-hydroxyprogesterone, cortisol, 11-deoxycortisol, corticosterone, and DHEAS, the 1.96 SD interval of the bias values was found within TAE limits. For androstenedione and

Table 1: Overview of measurement results. Sample numbers (n) with concentrations above the LOQ of the individual laboratories (Labs A–E), laboratory-specific and overall mean of found sample concentrations and concentration ranges with comparison to biological reference intervals, mean bias against overall mean with SD of bias and imprecision statistics including number of samples exceeding desired or optimal TAE goals based on biological variation data.

Analyte	Lab	n	Sample concentrations, nmol/L		Biological reference interval, nmol/L [19]	Bias to overall mean, %		Individual sample CV, %[a]		Samples>TAE (des/opt), %
			Mean	Range		Mean	SD	Mean	Range	
17OHP	All	200	4.53	(0.54–21.8)	0.28–6.26	–	–	8.1	(1.5–20.0)	0.0/10.0
	A	40	4.54	(0.54–20.8)		0.8	3.0			
	B	40	4.55	(0.56–21.5)		1.5	4.2			
	C	40	4.25	(0.49–20.9)		−6.6	5.8			
	D	40	4.80	(0.69–25.8)		5.8	11.6			
	E	40	4.53	(0.41–20.3)		−1.5	9.6			
AND	All	200	3.32	(0.30–8.75)	1.16–8.01	–	–	12.8	(7.2–32.1)	5.0/47.5
	A	40	3.38	(0.36–9.13)		3.1	5.2			
	B	40	3.57	(0.27–8.89)		7.7	6.8			
	C	40	2.75	(0.2–7.73)		−18.9	6.0			
	D	40	3.39	(0.39–9.37)		4.3	11.7			
	E	40	3.49	(0.25–8.64)		3.8	7.0			
CL	All	200	361	(29.0–902)	126–665	–	–	9.1	(4.2–24.4)	0.0/15.0
	A	40	358	(29.4–831)		−0.4	3.4			
	B	40	368	(28.7–953)		1.6	3.9			
	C	40	321	(27.0–813)		−11.0	3.2			
	D	40	405	(34.1–1,050)		12.0	9.4			
	E	40	353	(28.1–861)		−2.2	2.7			
TES	All	199	7.68	(0.10–52.9)	0.26–32.7	–	–	10.7	(0.9–47.4)	17.5/47.5
	A	39	7.71	(0.11–52.1)		−3.8	6.1			
	B	40	7.12	(0.11–48.5)		−7.1	6.1			
	C	40	7.51	(0.16–53.8)		−1.1	11.2			
	D	40	8.53	(0.05–55.7)		9.8	15.3			
	E	40	7.73	(0.08–54.3)		2.1	12.5			
DHEAS	All	152	3,148	(423–8,325)	914–9,390	–	–	11.2	(5.4–18.8)	0.0/25.0
	A	40	3,446	(459–9,130)		8.5	4.7			
	B	32	3,176	(1,512–7,709)		−13.4	5.2			
	D	40	3,334	(400–8,670)		5.5	8.8			
	E	40	3,076	(410–8,711)		−3.2	4.3			
PROG	All	144	23.94	(0.12–199)	0.05–43.09	–	–	19.6	(1.2–94.6)	21.1/39.5
	A	40	24.51	(0.10–220)		−5.2	19.0			
	B	31	23.06	(0.34–190)		−4.7	13.1			
	D	38	27.16	(0.17–196)		7.9	17.0			
	E	35	24.14	(0.01–191)		−1.5	38.8			
CC	All	160	11.7	(0.51–44.8)	1.69–41.23	–	–	13.0	(7.5–22.6)	0.0/37.5
	A	40	12.0	(0.50–45.2)		1.5	3.8			
	B	40	11.7	(0.62–47.5)		0.3	6.0			
	C	40	9.87	(0.44–37.7)		−15.6	2.7			
	E	40	13.3	(0.50–48.9)		13.8	6.0			
11-DF	All	159	3.05	(0.20–15.8)	0.12–2.18	–	–	7.7	(0.6–19.7)	0.0/7.5
	A	40	2.93	(0.22–15.2)		−1.9	4.5			
	B	40	3.28	(0.17–17.4)		5.0	8.3			
	C	39	2.98	(0.43–15.7)		−5.4	6.5			
	E	40	3.08	(0.22–14.8)		2.1	5.4			

Table 1: (continued)

Analyte	Lab	n	Sample concentrations, nmol/L		Biological reference interval, nmol/L [19]	Bias to overall mean, %		Individual sample CV, %[a]		Samples>TAE (des/opt), %
			Mean	Range		Mean	SD	Mean	Range	
CN	All	160	51.5	(7.62–75.7)	28.1–90.4	–	–	9.7	(4.8–20.0)	5.0/72.5
	A	40	48.5	(7.24–73.4)		−5.8	7.2			
	B	40	49.0	(7.40–75.4)		−4.6	4.0			
	C	40	56.4	(9.10–84.6)		9.8	5.1			
	D	40	52.1	(6.75–84.5)		0.5	8.5			

11-DF, 11-deoxycortisol; 17OHP, 17-hydroxyprogesterone; AND, androstenedione; CC, corticosterone; DHEAS, dehydroepiandrosterone sulfate; CL, cortisol; CN, cortisone; PROG, progesterone; TES, testosterone; TAE, total allowable error; des, desirable; opt, optimal; CV, coefficient of variation; SD, standard deviation; n, sample number; LOQ, limit of quantification. [a]Mean CV (%) of UK NEQAS PT results: 17OHP: 12.7 %, AND: 9.5 %, CL: 7.3 %, TES: 8.8 %, DHEAS: 7.7 %, PROG: 13.4 %.

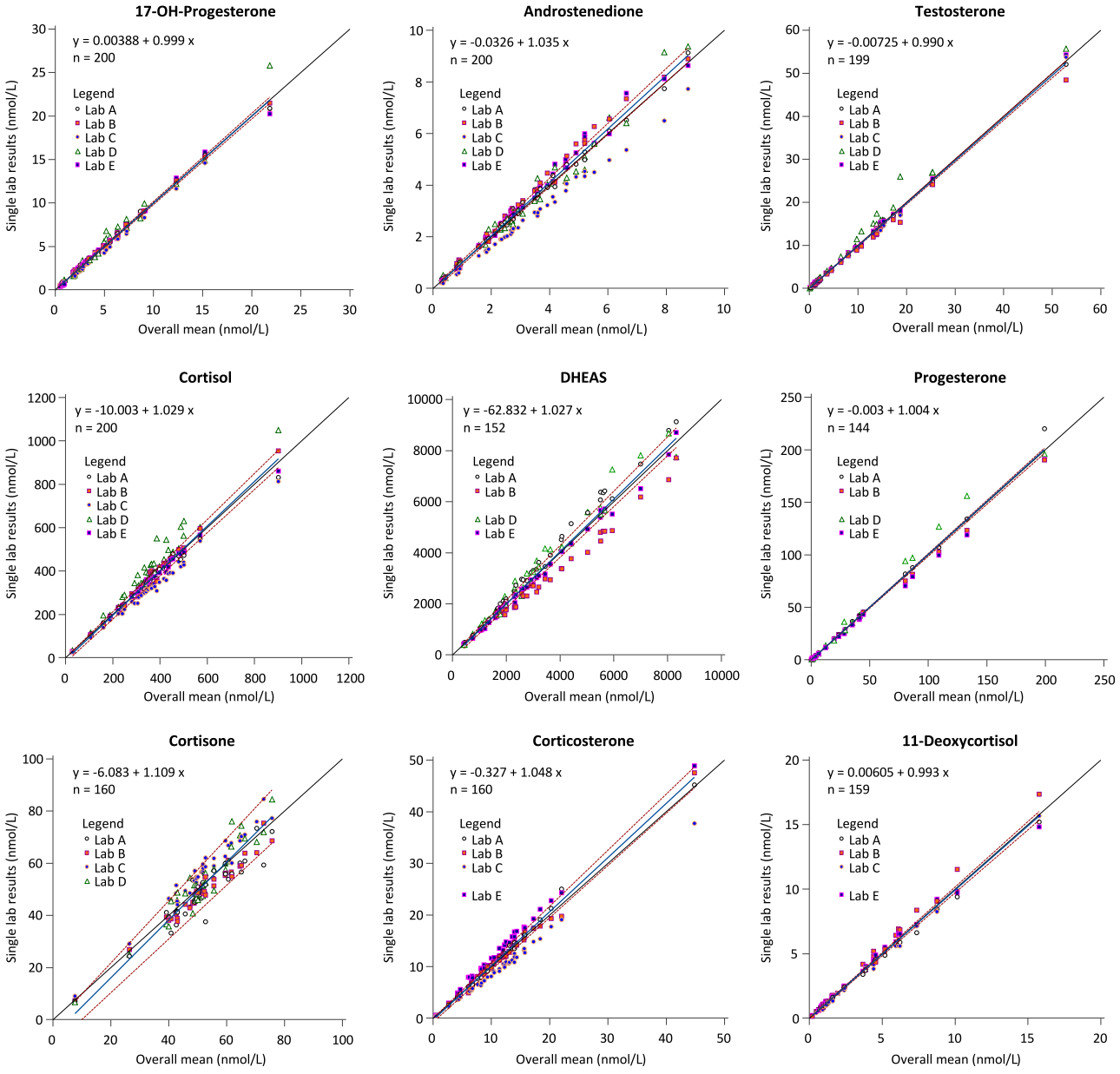

Figure 1: Passing–Bablok regression analysis comparing single lab results with the overall mean.

Table 2: Results from Passing–Bablok regression analysis comparing the overall mean and single measurements of each laboratory (Labs A–E). Laboratory results for the slope parameter are bolded if they are significantly different from the other laboratory results (no overlapping 95 % CI interval).

Analyte	Lab	Intercept	(95 % CI)	Slope	(95 % CI)	R^2	(95 % CI)	n
17OHP	All	0.004	(−0.037 to 0.041)	0.999	(0.985–1.013)	0.990	(0.987–0.993)	200
	Lab A	−0.003	(−0.019 to 0.045)	1.010	(0.991–1.019)	0.996	(0.992–0.998)	40
	Lab B	0.077	(0.020–0.134)	0.984	(0.967–1.000)	0.995	(0.990–0.997)	40
	Lab C	−0.050	(−0.141 to 0.045)	0.949	(0.918–0.970)	0.993	(0.986–0.996)	40
	Lab D	−0.098	(−0.358 to 0.143)	1.081	(0.987–1.155)	0.992	(0.986–0.996)	40
	Lab E	−0.048	(−0.152 to 0.066)	1.018	(0.975–1.046)	0.992	(0.985–0.996)	40
AND	All	−0.033	(−0.095 to 0.029)	1.035	(1.011–1.061)	0.979	(0.972–0.984)	200
	Lab A	0.060	(−0.008 to 0.109)	1.000	(0.973–1.030)	0.995	(0.991–0.998)	40
	Lab B	−0.010	(−0.093 to 0.105)	1.086	(1.045–1.121)	0.996	(0.993–0.998)	40
	Lab C	−0.098	(−0.166 to −0.038)	**0.853**	**(0.830–0.882)**	0.998	(0.996–0.999)	40
	Lab D	0.065	(−0.089 to 0.203)	0.989	(0.929–1.055)	0.984	(0.970–0.992)	40
	Lab E	−0.030	(−0.147 to 0.047)	1.063	(1.027–1.105)	0.995	(0.990–0.997)	40
CL	All	−10.0	(−21.1 to −0.6)	1.029	(0.996–1.063)	0.932	(0.911–0.948)	200
	Lab A	8.5	(−4.6 to 27.1)	0.971	(0.920–1.021)	0.985	(0.971–0.992)	40
	Lab B	−8.4	(−19.4 to 0.4)	1.048	(1.011–1.079)	0.979	(0.960–0.989)	40
	Lab C	5.4	(−4.8 to 18.9)	**0.878**	**(0.839–0.909)**	0.972	(0.948–0.985)	40
	Lab D	−19.5	(−56.1 to 5.0)	1.160	(1.064–1.268)	0.939	(0.888–0.968)	40
	Lab E	−0.3	(−8.1 to 9.8)	0.986	(0.957–1.008)	0.984	(0.970–0.992)	40
TES	All	−0.007	(−0.037 to 0.008)	0.990	(0.979–1.001)	0.994	(0.992–0.996)	199
	Lab A	−0.036	(−0.064 to −0.019)	0.983	(0.973–0.996)	0.998	(0.997–0.999)	39
	Lab B	−0.020	(−0.052 to 0.025)	0.938	(0.919–0.953)	0.998	(0.997–0.999)	40
	Lab C	0.015	(−0.060 to 0.034)	0.971	(0.955–0.998)	0.998	(0.996–0.999)	40
	Lab D	0.029	(−0.063 to 0.122)	**1.081**	**(1.058–1.135)**	0.996	(0.992–0.998)	40
	Lab E	−0.005	(−0.023 to 0.046)	1.000	(0.988–1.011)	0.998	(0.996–0.999)	40
11-DF	All	0.006	(−0.021 to 0.028)	0.993	(0.974–1.014)	0.989	(0.985–0.992)	159
	Lab A	0.039	(0.017–0.075)	0.947	(0.930–0.964)	0.994	(0.989–0.997)	40
	Lab B	−0.068	(−0.123 to −0.010)	**1.102**	**(1.067–1.139)**	0.989	(0.980–0.994)	40
	Lab C	−0.031	(−0.080 to 0.013)	0.965	(0.942–0.986)	0.986	(0.974–0.993)	39
	Lab E	0.016	(−0.044 to 0.046)	1.020	(0.994–1.055)	0.992	(0.986–0.996)	40
CC	All	−0.327	(−0.738 to −0.006)	1.049	(1.011–1.085)	0.964	(0.951–0.973)	160
	Lab A	−0.229	(−0.528 to −0.031)	1.045	(1.020–1.070)	0.995	(0.990–0.997)	40
	Lab B	0.064	(−0.337 to 0.366)	0.998	(0.962–1.034)	0.995	(0.991–0.998)	40
	Lab C	−0.002	(−0.240 to 0.149)	**0.845**	**(0.829–0.868)**	0.995	(0.991–0.997)	40
	Lab E	0.325	(−0.108 to 0.601)	1.107	(1.076–1.142)	0.993	(0.987–0.996)	40
CN	All	−4.571	(−9.066 to −1.107)	1.096	(1.024–1.180)	0.885	(0.846–0.915)	160
	Lab A	−1.854	(−9.709 to 2.688)	0.978	(0.886–1.133)	0.912	(0.839–0.953)	40
	Lab B	1.224	(−2.599 to 4.307)	0.926	(0.858–1.014)	0.955	(0.916–0.976)	40
	Lab C	0.542	(−4.342 to 4.535)	1.089	(1.010–1.182)	0.970	(0.943–0.984)	40
	Lab D	−9.309	(−18.944 to −1.642)	1.174	(1.026–1.365)	0.899	(0.817–0.946)	40
DHEAS	All	−62.8	(−153.6 to 20.1)	1.027	(0.994–1.065)	0.983	(0.976–0.987)	152
	Lab A	−60.9	(−120.3 to −7.9)	1.116	(1.092–1.146)	0.997	(0.995–0.999)	40
	Lab B	106.5	(35.1–177.7)	**0.833**	**(0.802–0.862)**	0.994	(0.988–0.997)	32
	Lab D	−29.5	(−111.5 to 98.6)	1.061	(1.007–1.118)	0.990	(0.981–0.995)	40
	Lab E	−53.0	(−92.8 to −6.4)	0.998	(0.977–1.019)	0.998	(0.996–0.999)	40
PROG	All	−0.003	(−0.056 to 0.023)	1.004	(0.987–1.012)	0.987	(0.983–0.991)	144
	Lab A	−0.058	(−0.081 to −0.006)	1.015	(1.006–1.026)	0.987	(0.975–0.993)	40
	Lab B	−0.035	(−0.080 to 0.060)	1.004	(0.954–1.015)	0.989	(0.978–0.994)	31
	Lab D	0.041	(−0.044 to 0.064)	1.031	(1.009–1.114)	0.993	(0.986–0.996)	38
	Lab E	−0.041	(−0.116 to 0.067)	0.948	(0.921–0.959)	0.981	(0.963–0.990)	35

11-DF, 11-deoxycortisol; 17OHP, 17-hydroxyprogesterone; AND, androstenedione; CC, corticosterone; DHEAS, dehydroepiandrosterone sulfate; CL, cortisol; CN, cortisone; PROG, progesterone; TES, testosterone; CI, confidence interval; n, sample number; R^2, coefficient of determination.

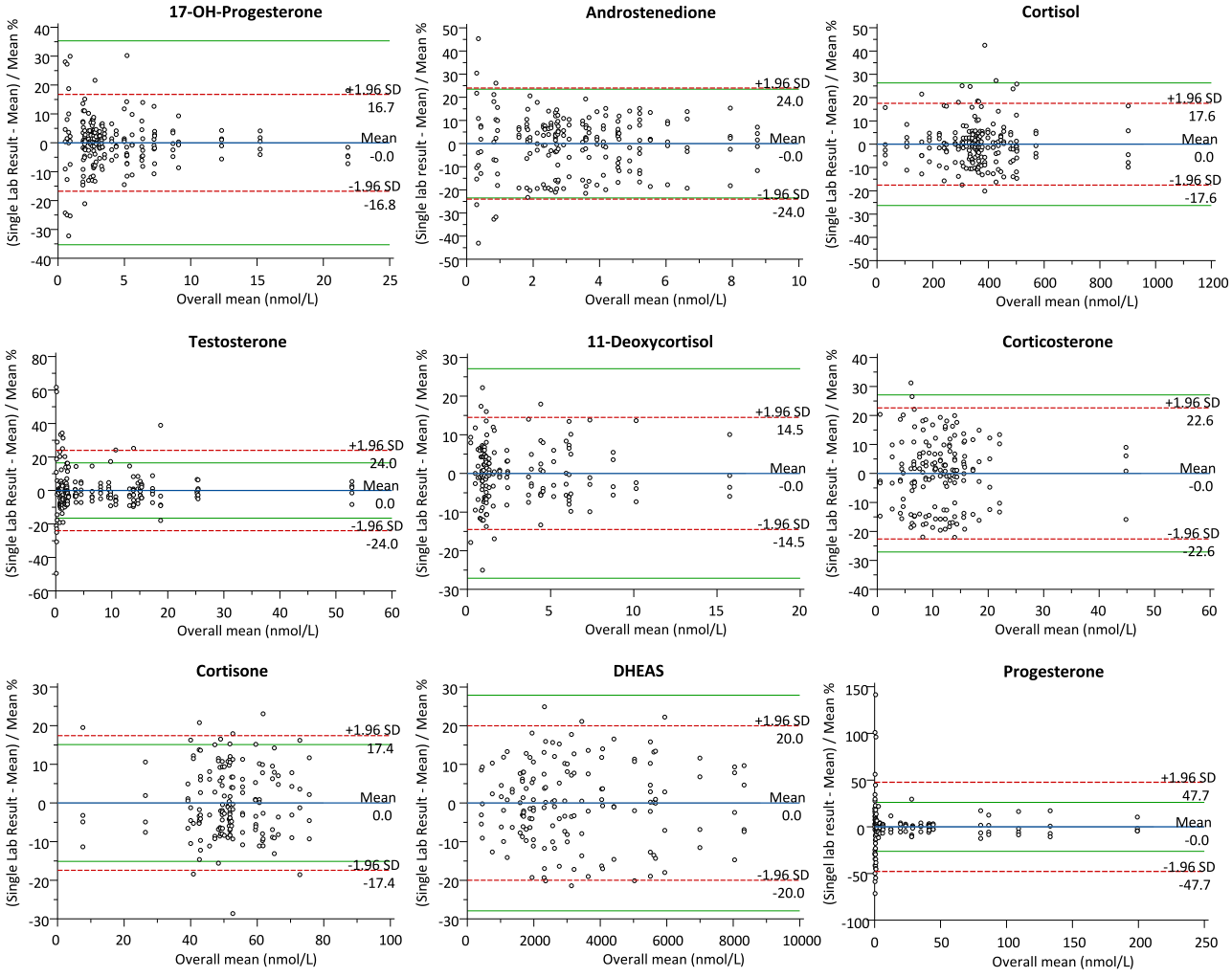

Figure 2: Bland–Altman styled scatter plots showing the percent difference of single results of each laboratory against the overall mean (y-axis) drawn against the overall mean (x-axis). Consequently, the mean difference against the overall mean is always 0.0 %, as illustrated with a blue line. The overall 1.96 SD interval (dashed red line) is compared with the total allowable error (green line) derived from biological variation data (see Supplementary Table S2).

cortisone, the TAE limits closely matched the 1.96 SD interval. For cortisol, progesterone, and testosterone, a higher bias variability in samples with a mean concentration below 2 nmol/L was causal for missing the TAE targets.

A comparison of iCV data from this study to iCV data from the UK NEQAS PT scheme and to TAE targets is presented in Figure 3. The iCV of most samples, including those from this study's and the UK NEQAS sample sets, remained within desirable TAE targets (Table 1). Generally, the results for iCV were comparable between both sample sets in terms of absolute values and trends (e.g., higher iCV values at lower concentrations). The highest differences were observed in DHEAS and 17-hydroxyprogesterone. For DHEAS, iCV values were higher on average in this study than in the UK NEQAS sample set and lower for 17-hydroxyprogesterone (mean iCVs 8.1 vs. 12.7 % and 11.2 vs. 7.7 %, respectively).

Discussion

It was demonstrated that comparable results were obtained in most participating laboratories for the addressed analytes in terms of the mean bias against the overall mean value and the variance of biases. Most results remained clearly within the desirable TAE performance targets derived from the EFLM database biological variability data. This also includes progesterone and testosterone, even though their scatter expressed as a 1.96 SD interval in the Bland–Altman plot clearly exceeded the TAE targets. This effect was observed only if the scatter was calculated from all samples and was assignable to poor interlaboratory agreement for very low-concentration samples. This problem may not involve general interassay agreement but rather a lack of sensitivity or specificity; in addition, low

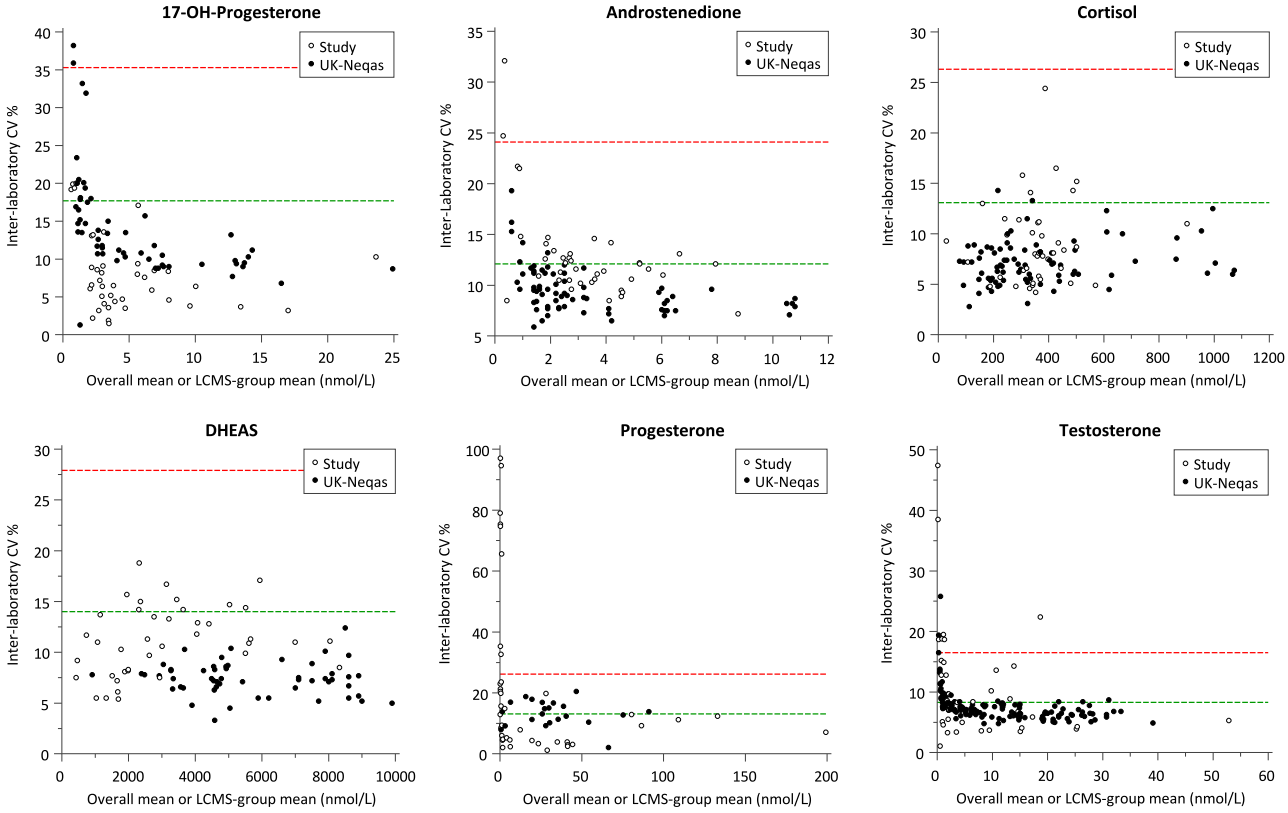

Figure 3: Comparison of interlaboratory CV values obtained from this study (white dots) and from the UK-NEQAS proficiency testing scheme (black dots) as a function of the mean sample concentration. The green line represents the optimal total allowable error (TAE), and the red line represents the desirable TAE, which was derived from the EFLM database or from the literature [8, 22–24]. LCMS group CV data from UK NEQAS PT distributions 470–494 were used, and only samples in the concentration range of the study samples were included. Study samples: all analytes, n=40; UK NEQAS samples: 17-OH-progesterone, n=56; androstenedione, n=63; cortisol, n=82; testosterone (female and male), n=150; DHEAS, n=54; progesterone, n=22.

concentrations could be missed when the calibration range is too narrow for individual assays.

The study provided new insights into the quality of assay performances at all laboratories. For example, the measurement bias found at Lab C for androstenedione was not identified in the PT scheme the laboratory participated. In that particular scheme, reconstituted serum spiked with steroids was used instead of an authentic serum or pooled serum, hence an oversimplified matrix was applied. Similar problems with the commutability of external quality materials using samples with a manipulated or artificial matrix were also found in the HarmoSter study [9, 10]. However, the major reason for the deviations observed at Lab C was partly due to assay calibration issues, which was demonstrated by exchanging calibration materials with Lab C and Lab B (data not shown).

Compared to the other laboratories, Lab D exhibited problems with higher variance across the concentration range for some analytes. The possible reasons for this problem remain speculative. The pooled samples used in this study included serum from different kinds of collection tubes and from gel separator tubes. The assay in Lab D was not optimized for these materials because they are not used in routine testing at that site. The assay in Lab D also has the shortest runtime, which could limit its chromatographic resolution toward interferences from the matrix, suggesting a selectivity issue.

Furthermore, valuable information was provided for all laboratories on the individual performance of their methods in the quantification of analytes, such as 11-deoxycortisol and corticosterone; to our best knowledge, there is no existing PT scheme available using authentic materials.

The iCV data in this study were comparable to iCV values derived from the UK NEQAS PT data. This is a good indicator for the quality of the data, since UK NEQAS uses mainly pooled serum samples, although the samples can be spiked with analyte or known interferences. Differences observed between both datasets may be influenced by the low number of participants in this study or in the UK NEQAS sample set, as observed for progesterone. Since most iCV values of both data sources were within optimal TAE goals, good general agreement between LC–MS/MS methods and sufficient performance for clinical needs can be assumed.

The influence of calibration materials could be examined in this study since two types of calibrators were used by the participating laboratories. Lab C was the only laboratory that used in-house prepared calibration materials, while the other laboratories used commercial IVD certified calibrators from one provider. The largest deviation from the overall mean was found at androstenedione for Lab C. Additionally, significant bias contributions were found at DHEAS for Lab B or at testosterone for Lab D. Therefore, the type of calibration may influence the comparability of the methods but is not the exclusive cause of bias. Differences in results between methods may also be caused by different sample preparations, chromatography settings, and different specificities against matrix interferences. Additionally, compared to the calibration type, lot-to-lot variability could exert a more significant impact since commercial and in-house calibrator materials are based on reference materials spiked into surrogate matrices. These findings also correspond with other comparison studies mentioned above.

To the best of our knowledge, this is the first published study to examine the interlaboratory performance of LC–MS/MS measurements for DHEAS. The overall comparability of results was sufficient for clinical needs, since the overall bias variability and iCV values remained within TAE targets. However, additional variation was introduced by Lab B and Lab D, leading to lower-than-expected overall performance for DHEAS from an analytical point of view. While Lab B showed a nonlinear relative bias against the overall mean, hinting at a linearity problem, Lab D had a higher bias across the whole concentration range than that of Lab A or E. Considering that all laboratories used calibrators from the same manufacturer and that DHEAS is the most abundant steroid in the bloodstream, this result scatter is surprising.

Compared to other steroids, DHEAS exhibits an advantage in MS analysis due to its inherent properties; however, these properties could be a possible explanation for the poor analytical performance. Multisteroid methods are usually optimized for analytes that are the most challenging to measure. DHEAS, as a charged analyte, has very different physicochemical properties from those of other steroids and has a concentration approximately 10 times higher. This could lead to DHEAS measurement issues, such as mass spectrometry linearity problems due to saturation effects or unstable chromatography affecting specificity, e.g., if unbuffered mobile phases are used (e.g., ammonium fluoride in methanol/water).

There are only a few other published studies to date that compare interlaboratory LC–MS/MS methods for steroid analytes. The published studies treat single or only a few steroidal analytes (testosterone, androstenedione, DHEA, estradiol) [27–33]. A single study, which was partly published recently (HarmoSter), dealt with a similar number of analytes as in this study [12, 13]. The HarmoSter study design is significantly different, especially the type of samples. In the study presented here, pooled serum samples were used exclusively, whereas in HarmoSter, samples were taken from individual donors with three different collection tubes each and from PT schemes. While single donor samples are the best solution in terms of commutability, the procedure we have chosen for this study is by far easier to perform (leftover-samples, sample number, sample type). Using our targeted approach in sample collection, the concentration ranges of the samples could be better controlled. Thus, we could achieve good coverage of biological reference intervals and pathological levels in a compact format of 40 samples without the need to artificially spike the samples to reach target concentrations. Hence, the commutability of these pooled serum samples should be sufficient for the purpose of this study – to evaluate the performance and comparability of routine clinical LC–MS/MS methods.

Limitations of this study are the low number of participating laboratories and the restricted sample number, which could affect the validity of the results. However, the results of this study were substantiated by comparison to the UK NEQAS sample set and to other studies mentioned above. Furthermore, since the sample set was not value assigned by reference methods, individual measurement deviation (bias) could not be calculated against a true value but only against the overall mean.

Conclusions

This study provides a good overview that compares and describes the current state of multisteroid LC–MS/MS-LDTs. The results provide valuable feedback to the study participants on their respective methods. In addition, generalized conclusions on the performance of LDT solutions in modern routine laboratories could be obtained. The results of this study show that, despite remarkable instrumental heterogeneity, the LC–MS/MS assays provided comparable measurement results. The result scatter was within the generally accepted performance limit (TAE) based on the BV of the endogenously present analytes. Thus, it can be concluded that the application of LC–MS/MS for steroid analysis in clinical laboratories has developed so that an increased risk for patients from LDTs cannot be identified [34]. This observation is especially true when considering risks associated with the limitations of currently widely used fully automated immunoassays, such as cross-reactivities or limited assay sensitivities [35].

However, as for other laboratory instrumentation, constant performance monitoring of LC–MS/MS installations in routine is necessary. This study should serve as an example of how interlaboratory quality assessment could work in addition to organized PT schemes, especially for analytes not covered by PT providers. No general decisive advantage for the comparability of measurement results could be found when commercial calibrators were utilized, but circumventing the laborious and error-prone in-house production of calibrators may increase the robustness of an assay and may facilitate traceability and standardization. Therefore, we think that the best solution for LDT-based LC–MS/MS steroid analysis involves developing the an in-house measurement method tailored to the analyte panel needed locally and the use of standardized, traceable calibrators. This approach will enable modern personalized laboratory medicine that provides the best possible results in a flexible and timely manner as the basis of individualized decision-making by care-taking clinicians.

Research ethics: All procedures were in accordance with the Declaration of Helsinki.

Informed consent: Not applicable as only anonymized leftover samples were used.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: The authors state no conflict of interest. C.S. and V.B. have been employees of Dr. Risch Ostschweiz AG at the time the study measurements have been carried out.

Research funding: None declared.

Data availability: Not applicable.

References

1. French D. Clinical utility of laboratory developed mass spectrometry assays for steroid hormone testing. J Mass Spectrom Adv Clin Lab 2023; 28:13–9.
2. Keevil BG. LC–MS/MS analysis of steroids in the clinical laboratory. Clin Biochem 2016;49:989–97.
3. Taylor AE, Keevil B, Huhtaniemi IT. Mass spectrometry and immunoassay: how to measure steroid hormones today and tomorrow. Eur J Endocrinol 2015;173:D1–12.
4. Travison TG, Vesper HW, Orwoll E, Wu F, Kaufman JM, Wang Y, et al. Harmonized reference ranges for circulating testosterone levels in men of four cohort studies in the United States and Europe. J Clin Endocrinol Metab 2017;102:1161–73.
5. Vesper HW, Thienpont LM. Traceability in laboratory medicine. Clin Chem 2009;55:1067–75.
6. 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.
7. 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.
8. Braun V, Stuppner H, Risch L, Seger C. Design and validation of a sensitive multisteroid LC-MS/MS assay for the routine clinical use: one-step sample preparation with phospholipid removal and comparison to immunoassays. Int J Mol Sci 2022;23:14691.
9. MacKenzie F. Letter in response to: John W Honour. Standardization of steroid tests and implications for the endocrine community. Annals of Clinical Biochemistry, Vol. 54(6): 618–630. Ann Clin Biochem Int J Lab Med 2018;55:409.
10. Oosterhuis WP, Bayat H, Armbruster D, Coskun A, Freeman KP, Kallner A, et al. The use of error and uncertainty methods in the medical laboratory. Clin Chem Lab Med 2018;56:209–19.
11. Aarsand AK, Fernandez-Calle P, Webster C, Coskun A, Gonzales-Lao E, Diaz-Garzon J, et al. The EFLM biological variation database [online]. https://biologicalvariation.eu [Accessed 26 Sep 2023].
12. Fanelli F, Cantù M, Temchenko A, Mezzullo M, Lindner JM, Peitzsch M, et al. Report from the HarmoSter study: impact of calibration on comparability of LC-MS/MS measurement of circulating cortisol, 17OH-progesterone and aldosterone. Clin Chem Lab Med 2022;60: 726–39.
13. Fanelli F, Bruce S, Cantù M, Temchenko A, Mezzullo M, Lindner JM, et al. Report from the HarmoSter study: inter-laboratory comparison of LC-MS/MS measurements of corticosterone, 11-deoxycortisol and cortisone. Clin Chem Lab Med 2022;61:67–77.
14. Yates AM, Bowron A, Calton L, Heynes J, Field H, Rainbow S, et al. Interlaboratory variation in 25-hydroxyvitamin D2 and 25-hydroxyvitamin D3 is significantly improved if common calibration material is used. Clin Chem 2008;54:2082–4.
15. Annesley TM, Mckeown DA, Holt DW, Mussell C, Champarnaud E, Harter L, et al. Standardization of LC-MS for therapeutic drug monitoring of tacrolimus. Clin Chem 2013;59:1630–7.
16. Gaudl A, Kratzsch J, Bae YJ, Kiess W, Thiery J, Ceglarek U. Liquid chromatography quadrupole linear ion trap mass spectrometry for quantitative steroid hormone analysis in plasma, urine, saliva and hair. J Chromatogr A 2016;1464:64–71.
17. Gaudl A, Kratzsch J, Ceglarek U. Advancement in steroid hormone analysis by LC–MS/MS in clinical routine diagnostics – a three year recap from serum cortisol to dried blood 17α-hydroxyprogesterone. J Steroid Biochem Mol Biol 2019;192:105389.
18. Vogeser M, Schuster C, Rockwood AL. A proposal to standardize the description of LC–MS-based measurement methods in laboratory medicine. Clin Mass Spectrom 2019;13:36–8.
19. Eisenhofer G, Peitzsch M, Kaden D, Langton K, Pamporaki C, Masjkur J, et al. Reference intervals for plasma concentrations of adrenal steroids measured by LC-MS/MS: impact of gender, age, oral contraceptives, body mass index and blood pressure status. Clin Chim Acta 2017;470: 115–24.
20. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;327: 307–10.
21. Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part I. Clin Chem Lab Med 1983;21:709–20.
22. Garde AH, Hansen AM, Skovgaard LT, Christensen JM. Seasonal and biological variation of blood concentrations of total cholesterol, dehydroepiandrosterone sulfate, hemoglobin A1c, IgA, prolactin, and free testosterone in healthy women. Clin Chem 2000;46:551–9.

23. Garde AH, Hansen ÅM, Skovgaard LT, Christensen JM. Erratum: seasonal and biological variation of blood concentrations of total cholesterol, dehydroepiandrosterone sulfate, hemoglobin A_{1c}, IgA, prolactin, and free testosterone in healthy women (Clin Chem 2000;46: 551–9). Clin Chem 2001;47:1877.
24. van der Veen A, van Faassen M, de Jong WHA, van Beek AP, Dijck-Brouwer DAJ, Kema IP. Development and validation of a LC-MS/MS method for the establishment of reference intervals and biological variation for five plasma steroid hormones. Clin Biochem 2019;68: 15–23.
25. Røys EÅ, Guldhaug NA, Viste K, Jones GD, Alaour B, Sylte MS, et al. Sex hormones and adrenal steroids: biological variation estimated using direct and indirect methods. Clin Chem 2023;69:100–9.
26. Westgard QC. Desirable biological variation database specification [Online]. https://www.westgard.com [Accessed 3 May 2023].
27. Büttler RM, Martens F, Fanelli F, Pham HT, Kushnir MM, Janssen MJW, et al. Comparison of 7 published LC-MS/MS Methods for the simultaneous measurement of testosterone, androstenedione, and dehydroepiandrosterone in serum. Clin Chem 2015;61:1475–83.
28. 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.
29. Thienpont LM, Van Uytfanghe K, Blincko S, Ramsay CS, Xie H, Doss RC, et al. State-of-the-art of serum testosterone measurement by isotope dilution-liquid chromatography-tandem mass spectrometry. Clin Chem 2008;54:1290–7.
30. Owen LJ, MacDonald PR, Keevil BG. Is calibration the cause of variation in liquid chromatography tandem mass spectrometry testosterone measurement? Ann Clin Biochem 2013;50:368–70.
31. 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.
32. French D, Drees J, Stone JA, Holmes DT, van der Gugten JG. Comparison of four clinically validated testosterone LC-MS/MS assays: harmonization is an attainable goal. Clin Mass Spectrom 2019;11:12–20.
33. Büttler RM, Martens F, Ackermans MT, Davison AS, van Herwaarden AE, Kortz L, et al. Comparison of eight routine unpublished LC–MS/MS methods for the simultaneous measurement of testosterone and androstenedione in serum. Clin Chim Acta 2016;454:112–8.
34. Seger C, Salzmann L. After another decade: LC–MS/MS became routine in clinical diagnostics. Clin Biochem 2020;82:2–11.
35. Ghazal K, Brabant S, Prie D, Piketty ML. Hormone immunoassay interference: a 2021 update. Ann Lab Med 2021;42:3–23.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cclm-2023-0847).