Effective validation of chromatographic analytical methods: The illustrative case of androgenic steroids

Elsevier Editorial System(tm) for Analytica Chimica Acta or its open access mirror
Manuscript Draft

Manuscript Number:

Title: Effective validation of chromatographic analytical methods: the illustrative case of androgenic steroids

Article Type: Full Length Article

Section/Category: MASS SPECTROMETRY

Keywords: Validation; calibration; chromatography - mass spectrometry; androgens

Corresponding Author: Dr. Eleonora Amante,

Corresponding Author's Institution: Università degli Studi di Torino

First Author: Eugenio Alladio, PhD

Order of Authors: Eugenio Alladio, PhD; Eleonora Amante; Cristina Bozzolino; Fabrizio Seganti; Alberto Salomone; Marco Vincenti; Brigitte Desharnais, PhD

Manuscript Region of Origin: ITALY

Suggested Reviewers: Riccardo Leardi
Università di Genova - DIFAR
leardi@difar.unige.it

Mario Thevis
Centre for Preventive Doping Research (ZePräDo)
thevis@dshs-koeln.de

Francisco Raposo
Spanish National Research Council | CSIC · Industrial Processes & Environment Unit-Instituto de la Grasa
fraposo@cica.es

Pedro Araujo
National Institute of Nutrition and Seafood Research, PO
pedro.araujo@nifes.no

Cover Letter(including Novelty Statement)

Università degli Studi di Torino
Dipartimento di Chimica
Via P. Giuria, 7 10125 Torino Italy

chimica

Eleonora Amante, MSc
phone: +39 3467891901
e-mail: eleonora.amante@unito.it

Torino, October 21th, 2019

Editor, **Analytica Chimica Acta**

Dear Editor,

This letter accompanies submission to Analytica Chimica Acta of a paper entitled “**Effective validation of chromatographic analytical methods: the illustrative case of androgenic steroids**” and of a parallel MethodsX paper entitled “**Experimental and statistical protocol for the effective validation of chromatographic analytical methods**”.

The authors are:

Eugenio Alladio, Eleonora Amante, Cristina Bozzolino, Fabrizio Seganti, Alberto Salomone, Marco Vincenti and Brigitte Desharnais.

Please address all correspondence to:
Eleonora Amante
Dipartimento di Chimica, Università degli Studi di Torino, Via Pietro Giuria 7, 10125, Torino (Italy)
Phone +39-3467891901
e-mail: eleonora.amante@unito.it

The validation of analytical methods is of crucial importance in several fields of application. The expounding case of a gas chromatographic-mass spectrometric method for the urinary endogenous steroid profiling is presented to illustrate a validation strategy that combines rigorous estimation of validation parameters with highly efficient use of the collected data. This work was inspired by two papers from Desharnais et al. published on the Journal of Analytical Toxicology (doi 10.1093/jat/bkx001 and 10.1093/jat/bkx002), which proposed a routine for the evaluation of calibration models. In practise, the validation protocol we describe requires three replicates of the calibration curve performed in three different days, for a total of nine replicates and 54 experiments. Such an operating scheme allows to evaluate several validation parameters using the same set of experiments. Among them, the calibration model is meticulously defined for each analyte, using several statistical tests for heteroscedasticity and linearity. With the same procedure, intra- and inter–day accuracy and precision are calculated.
This work provides an in-depth discussion of the results obtained with different statistical tools, using as representative example the case of a multi-targeted GC-MS method for the detection of androgen steroids in urine. All the equations and reported and described in the MethodsX parallel paper.

Novelty statement: This work is new and original and is not under consideration elsewhere. In the scientific literature, analytical method validations are frequently reported without clear relationship between the objectives of validation and the strategies of data collection and interpretation. Moreover, wrong statistical tools and unjustified assumptions are repeatedly used. We believe that our study represents an important tool of reflection for analytical chemists that can significantly contribute to standardize and improve the reliability of the validation process in the field of chromatography hyphenated with mass spectrometry.

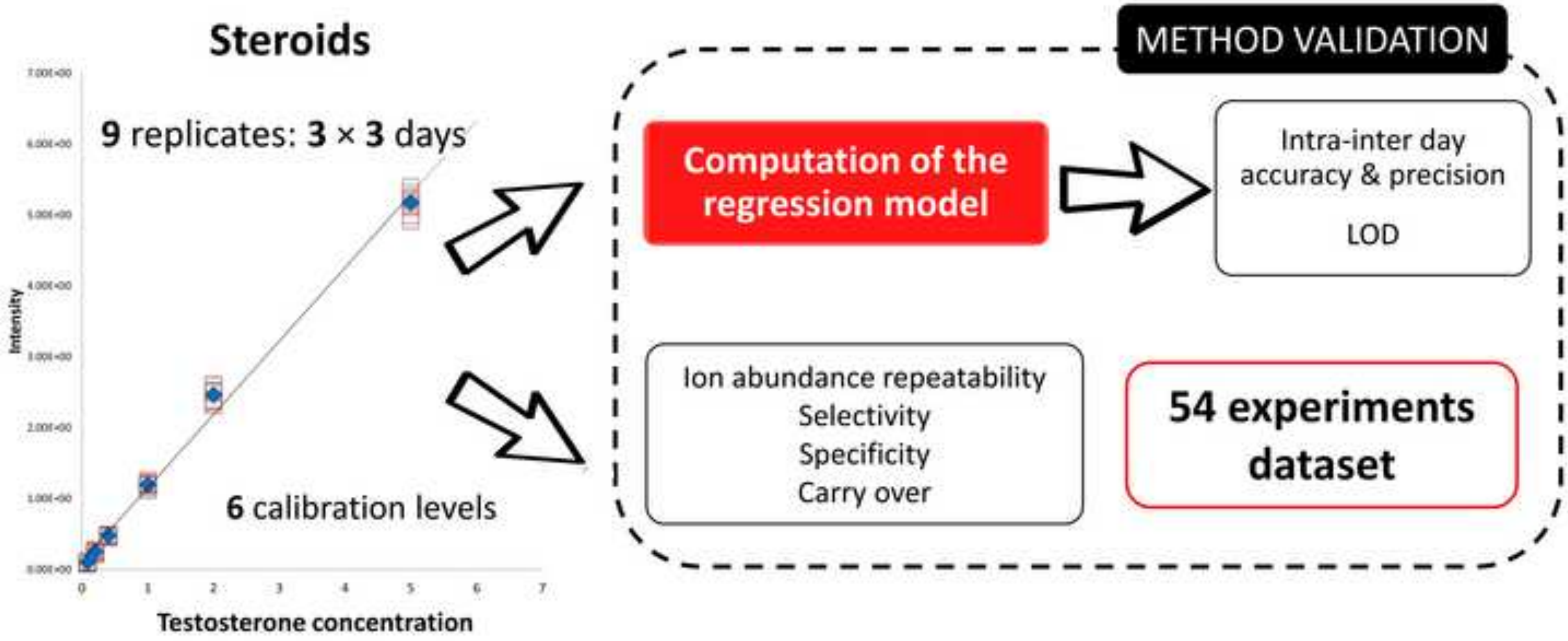
Thank you for considering the paper for Analytica Chimica Acta.

Best regards.

Yours faithfully,

Eleonora Amante

*Graphical Abstract



9 replicates: **3** × **3** days

6 calibration levels

METHOD VALIDATION

Computation of the regression model

Intra-inter day
accuracy & precision

LOD

Ion abundance repeatability
Selectivity
Specificity
Carry over

54 experiments dataset

Highlights

- The case study of a multitargeted method for urinary steroids is reported;
- An efficient and comprehensive validation strategy is proposed;
- From nine replicates of calibration data-points most validation parameters are calculated;
- Appropriate statistical tests are used and discussed.

Effective validation of chromatographic analytical methods: the illustrative case of androgenic steroids.

E. Alladio [a,b*], E. Amante [a,b*#], C. Bozzolino [a,b], F. Seganti [b], A. Salomone [a,b], M. Vincenti [a,b], B. Desharnais [c].

[a] Dipartimento di Chimica, Università degli Studi di Torino, Turin, Italy
[b] Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", Orbassano (Turin), Italy
[c] Laboratoire de sciences judiciaires et de médecine légale, Montréal, Québec, Canada

* The authors equally contributed to this work.

# Corresponding Author:

Eleonora Amante,

Address**:** Dipartimento di Chimica, Università degli Studi di Torino, Via Pietro Giuria 7, 10125, Torino (Italy)
Telephone: +390116705264
Fax number: +390116705242
e-mail: eleonora.amante@unito.it

Keywords: Validation, calibration, chromatography – mass spectrometry, androgens.

Abstract

The increasing need to develop quantitative chromatographic methods with upgradable multi-targeted approach, allowing flexible and reliable application on large daily workload, makes the implementation of an efficient strategy of method's validation and maintenance crucial for the quality assurance policy. The expounding case of a gas chromatographic-mass spectrometric method for the urinary endogenous steroid profiling is presented to illustrate a validation strategy that combines rigorous estimation of validation parameters with highly efficient use of the collected data. The analysis of blank urine samples fortified at six concentration levels with 18 targeted steroids was replicated nine times in three working sessions along twelve days. This dataset of 54 analysis formed the groundwork on which the statistical evaluation of several validation parameters was founded, including calibration, intra- and inter-day accuracy and precision, limit of detection (LOD), limit of quantification, ion abundance repeatability, selectivity, specificity, and carry-over. The preliminary comparison of the response variances at different concentration levels provided the evaluation for heteroscedasticity. Then, the most appropriate calibration model was determined for each steroid, in terms of order (linear vs. quadratic) and weighting, allowing to complete their quantitation in each solution. Intra- and inter-day accuracy and precision were calculated therefrom. LOD values were computed with the Hubaux-Vos method from the weighted linear segment of the calibration curves. Only the assessment of recovery and matrix effect required the execution of further independent experiments. The case study demonstrated that the application of adequate statistical testing typically produced non-homogeneous models of calibration curves, mostly arising from heteroscedastic and quadratic distribution of datasets, unlike what is reported in overly simplified approaches. The misleading information obtained from the regression coefficient R^2 to evaluate linearity was evidenced. The strong dependence of calculated LOD and accuracy from the selected calibration parameters was highlighted, making the implementation of an adequate calibration maintenance policy highly advisable.

1. Introduction

The whole process of an analytical method validation has found various definitions. For example, the American Food and Drug Administration (FDA) defines it as "the process of demonstrating that an analytical procedure is suitable for its intended purpose" [1]. In practice, several organizations and scientists have tried to standardize the validation procedure, according to the purpose of the analysis (*e.g.*, qualitative, quantitative) and the application field, recommending specific parameters to be evaluated and tests to be performed [1–3]. Features of importance will differ depending on the particular application field (e.g. bioanalysis [4–6]) or instrumental technique used (e.g. chromatography [1,4]).

For the validation of quantitative methods, a feature of utmost importance is the calibration, which is the process that transforms the raw data obtained from the analytical instruments into useful concentration units by means of the statistical technique of regression [7]. Building an appropriate regression model requires the analysis of a series of standard samples within a defined range of concentrations, and the subsequent study of the mathematical relationships occurring between these concentrations and the corresponding analytical responses [7,8]. Consequently, the quality of the quantitative data that a novel analytical method will provide is highly dependent on the quality of the calibration model used [9]. Although most instrumental systems should theoretically exhibit analytical signals directly proportional to concentrations, hence generating linear calibration curves, in reality some interfering physical and chemical phenomena may result in a deviation from the expected linear trend [9,10] and/or heterogeneous distribution of data-point at different concentration levels [11]. To account for the latter problems, several publications and official documents have proposed the use of various statistical tests to compare linear and quadratic calibration curves and weighted least squares (WLS) regression strategies to correct heteroscedastic distributions [9,10,12–19]. The International Union of Pure and Applied Chemistry (IUPAC) guideline recommends to choose the model (*e.g.* ordinary linear regression, second order calibration function or weighted regression) which provides the lowest measurement uncertainty [15]. Baumann and Watzig [14] developed a stepwise approach aimed to find the best weighted calibration curve. In 2005, Singtoroj et al. [12] developed a systematic method to calculate and compare regression models during pre-validation and validation of bioanalytical methods. Their approach considered both quadratic and linear fitting, forcing through origin, transformation of data strategies (*e.g.* log-log, Box-Cox) and applied weighting (1, 1/x or 1/x^2). More recently, Desharnais et al. [10,20] published an R routine devoted to the automatic testing and selection of the best calibration model, including order (linear or quadratic) and weighting (1, 1/x or 1/x^2). In 2016, Raposo [21] reviewed the validation guidelines of several organizations and compared the tests recommended to evaluate fitting and linearity of the calibration curves, concluding that several different ways were valid to assess the linearity of calibration curves. The Scientific Working Group for Forensic Toxicology (SWGTOX), Irish National Accreditation Board (INAB), Joint Research Centre (JRC) and IUPAC all recommended executing several statistical tests within their validation protocols.

Another fundamental determination included in all validation procedures is the limit of detection (LOD), which is calculated either by examining the variance of the residual signal at zero concentration, using a large number of independent blank samples, or – more practically with chromatographic methods – by estimating its value from the lowest levels of the calibration curve [22]. The regression-based Hubaux-Vos' algorithm is a widespread technique for this estimation [22,23], but one of the prerequisites to use this approach is that the residuals from the linear model are homoscedastic, i.e. have a uniform distribution along the whole calibration range [13,22,23]. In contrast, the occurrence of heteroscedasticity is commonly observed in routine analytical models that cover several orders-of-magnitude concentrations. To overcome this problem, the use of an unweighted model can be replaced by a weighted least squares (WLS) calibration model in the Hubaux-Vos' LOD calculation [13].

The calculation of several other validation parameters is recommended in all guidelines, including trueness, precision, specificity, recovery, matrix effect, carry over, and others [1–4,8,15,18,22]. The evaluation of each of these parameters generally requires repeated independent experiments, making the whole validation process quite demanding.

In our daily work with biological matrices, LC-MS/MS and GC-MS-based methods are continuously developed and/or updated to support the ongoing evolution of clinical and toxicological requirements. To reduce the number of experiments needed to achieve a comprehensive validation of our analytical methods, we studied an integrated approach, resulting in the development of an efficient and rigorous validation protocol. This integrates the Desharnais' routine R procedure of calibration [10] into an inclusive strategy to estimate further validation parameters: intra- and inter-day accuracy and precision, LOD, limit of quantification (LOQ), ion abundance repeatability, selectivity, specificity and carry-over. At present, this protocol is routinely used in our laboratory for the validation of both GC-MS and UHPLC-MS/MS bioanalytical methods devoted to the determination of endogenous metabolites and xenobiotics [24–26]. In the present study, the whole validation strategy is presented, using as a case study the analytical method developed for the endogenous androgenic steroids.

2. Material and Methods

2.1 Chemicals and reagents

All steroid standards were purchased as pure powders from Steraloids (Newport, RI, USA). Methanol, methyl tert-butyl ether (TBME), ethyl acetate, 17α-methyltestosterone, dithioerythritol, N-Methyl-N-(trimethylsilyl)trifluoroacetamide (MSTFA) and synthetic urine were provided by Sigma-Aldrich (Milan, Italy). β-glucuronidase from *Escherichia Coli* was purchased from Roche Life Science (Indianapolis, IN, USA). Ultra-pure water was obtained using a Milli-Q® UF-Plus apparatus (Millipore, Bedford, MA, USA). Solid-phase extraction (SPE) C-18 endcapped cartridges were from UCT Technologies (Bristol, PA, USA).

Standards solutions were prepared in methanol at the concentration of 1 mg/mL. Then, two working solution mixtures were prepared by dilution (MIX I = 3 µg/mL, MIX II = 100 µg/mL, internal standard solutions = 10 µg/mL). Two internal standards were used: testosterone-D_3 for the quantification of mix I; 17α-methyltestosterone for mix II (Table 1).

2.2 Samples pre-treatment

The investigated samples were either synthetic or negativized real urine specimen, depending on the specific experiment. For the evaluation of repeatability of retention times and ion abundance ratios authentic urine samples were negativized by extracting its steroid content by Solid-Phase Extraction (SPE) using C-18 endcapped cartridges. The absence of analytes at concentrations above the limits of detection was verified by the analysis of a non-spiked sample. Both authentic and synthetic urine samples were treated identically. 6 mL urine aliquots were fortified with testosterone-D_3 and 17α-methyltestosterone at the final concentration of 25 ng/mL and 125 ng/mL, respectively. The pH was then adjusted to a value between 6.8 and 7.4 by adding 2 mL phosphate buffer 0.1 M and drop(s) of NaOH 1 M, if necessary. A volume of β-glucuronidase solution corresponding to 83 units was added and then the mixture was incubated at 58 °C for 1 hour. After cooling at room temperature, 2 mL carbonate buffer 0.1 M was added to the aqueous solution, together with drop(s) of NaOH 1 M, until the final pH = 9 was reached. Then, liquid-liquid extraction (LLE) was performed with 10 mL of TBME; the samples were shaken in a multi-mixer for 10 minutes, centrifuged at 6.24 x g for 5 minutes and the organic supernatant was transferred into a glass tube. The extracts were subsequently dried under a nitrogen flow at 70 °C. After addition of 50 µL derivatizing solution (MSTFA/NH_4I/dithioerythritol – 1,000:2:4 v/w/w), the reaction was allowed to proceed at 70 °C for 30 minutes. The resulting solutions were transferred into conical vials and a 1 µL aliquot was injected by autosampler into the GC-MS working in the splitless mode. Further instrumental details are available in a previously published version of this method [25] and in the MethodsX article accompanying this work [27]. Mix I and II had distinct calibration ranges (Table 1), selected on the basis of the expected physiological concentrations, as reported in literature [25,28].

2.3 Validation protocol

A comprehensive scheme of the validation protocol is depicted in Figure 1. A total of nine independent replicated analyses at each concentration level (6 levels) were executed in three different days, resulting in three calibration data points collected on day 1, three on day 5, and three on day 12. The 12-days timeframe during which the experiments were scheduled allowed evaluating the stability of the method with time. Additionally, the 3×3×6 distribution of the 54 analysis made the evaluation of other validation parameters possible without carrying out further independent experiments. These included precision, accuracy, limit of detection (LOD), limit of quantification (LOQ), ion abundance repeatability, selectivity, specificity, and carry-over (adding blank samples after the analysis of the spiked samples with highest concentration). Independent experiments were conducted to determine the extent of matrix effect and recovery. An *ad hoc* Excel® sheet was built in-house to adapt the routine developed by Desharnais *et al.* [10] to our needs. All the equations employed to compute the validation parameters have been omitted from this text and can be found in a companion paper [27].

2.3.1 Calibration

A stepwise standard approach [10] was applied to calculate the best calibration models, as schematized in Figure 1. Initially, the heteroscedasticity of data points was tested by comparing the variance of the area ratios at the lowest and highest calibration level, using an F-Test integrated in the R routine [20,29]. Also the Levene test [30] was executed (in the version modified by Brown and Forsythe), in order to confirm the F-Test results with a procedure robust to non-normal distributions, operating on all the calibration levels. If the variance increased with concentration, i.e. the system was heteroscedastic, a weighted model was adopted, using a 1/x weighting factor when the variance increased proportionally to the concentration, or a $1/x^2$ weighting when a quadratic increase of the variance was observed. Case by case, the weighting generating the smallest variance of weighted normalized variances was selected.

In the second step, the order of the calibration model was established by comparing the captured variances of quadratic and linear (weighted) models by a partial F-test, which differs from Mandel's test because it compares the sum of squares of the regression (not of the residuals) to the mean squares of residuals [27]. If the quadratic calibration model significantly improved the captured variance of the data in comparison with the linear model, the former was accepted and the algorithm computed it. Lastly, the analysis of variance lack of fit (ANOVA-LoF) test was performed, to verify the goodness of the calculated calibration model.

2.3.2 LOD

The limit of detection (LOD) was estimated by means of the Hubaux-Vos' algorithm [23], which is an implementation of Currie's method to calculate the LOD [31]. This technique estimates the concentration associated with the confidence interval for the signal of a blank sample (i.e. the intercept). This is a robust means to estimate the LOD, since it relies on the entire data set to do so. Although the original Hubaux-Vos technique applies to homoscedastic data, Sanchez [13] has demonstrated that it can be used for heteroscedastic data if properly weighted parameters (standard error of the regression, slope and intercept, etc.) are used. Moreover, to respect Hubaux-Vos linearity assumption, when a quadratic calibration trend was recorded, the highest concentration(s) (typically the last level) were discarded from the calculation until a linear dependence from concentration was observed, as confirmed by the same partial F-test previously employed for the definition of the calibration model. The calculated LOD values were experimentally tested by spiking the blank matrix (synthetic urine) with the targeted analytes at the approximate LOD concentration and verifying that the signal-to-noise ratio (S/N) was higher than 3.

2.3.3 Accuracy

In most validation protocols reported in the literature, accuracy and trueness [32] are calculated by executing at least five determinations per concentration at a minimum of three concentration levels, and a deviation below 15% from the expected value [33] is recommended as an acceptance criterion. In contrast, our routine relies on an independent back-calculation of each data-point using only the calibration curves that did not include the sequence containing the specific data-point considered. The operating scheme is summarized in Figure 2A. For the calculation of intra-day accuracy, each sequence of six-level analysis was imagined to provide a separate calibration line, theoretically yielding three calibrations per each day. Then, (i) two sequences were used to compute the calibration model using the R routine, as described above; (ii) this model (which might be weighted or unweighted, linear or quadratic, depending on the former conclusions) was used to back-calculate the data-points of the third sequence; (iii) all concentration results from the back-calculation were averaged, and the overall bias was calculated.

The inter-day accuracy was computed following a similar operating scheme (Figure 2B), where the back-calculations of the data-points from a specific day were performed on the calibration model built using the six sequences of the other two days.

Accuracy results were obtained for all six-concentration levels of the calibration curve along which quantitative determinations were made. Accuracy was considered optimal if the bias was below 15% and acceptable if it fell between 15% and 20%.

2.3.4 Precision

Precision was expressed by the coefficient of variation (%CV or %RSD) of calculated concentrations from repeated analysis of homogeneous urine aliquots [33–35] spiked to provide the calibration curves. The intra-assay precision was calculated independently for the three days of analysis using the three replicates obtained in each day. A calibration model was computed for each day of validation and it was used to back calculate the three experimental replicates performed the same day. Then, the %CV was calculated [27].

The inter-assay (or intermediate) precision, which applies to within laboratory variations [36], was computed by back calculating all the nine replicates using the comprehensive calibration curve (i.e. the one built using all the performed experiments). Our protocol allowed to compare the performance of the same analysis in different days and by different operators, involving two operators working in alternate sequence.

Satisfactory results were expected to lie within ± 15% for both intra and inter-assay precision.

2.3.5 Selectivity and Specificity

The presence of potentially interfering substances in urine, including endogenous matrix components, metabolites, and decomposition products, was checked by examining the selected ion chromatograms around the expected retention times for all the analytes of interest. The presence of interfering peaks with S/N > 3 around the retention time of the analytes was examined for all samples in each experiment. Identification criteria for the analytes were established by checking the presence of all qualifying ions and their relative abundance at the expected retention time.

2.3.6 Repeatability of retention times and ion abundance ratios

Retention time repeatability was evaluated using the calibration standards and 30 blank (negativized) real urine samples fortified with the target analytes at different concentration levels within the dynamic range of the analytical methodology. The latter were prepared from 6 mL real urine aliquots by extracting the steroid fraction by C-18 SPE and subsequently spiking the eluates with the target analytes. Deviations below 1% from calibrators and controls are usually considered acceptable. Ion abundance ratio (quantifying to qualifying ion) repeatability was evaluated for each target analyte at all calibration levels, with acceptance limit of ±20% with reference to the control.

2.3.7 Matrix effect

The matrix effect was estimated at three concentration levels (*e.g.* low, medium and high concentration, within the linear range of the method) by comparing the experimental results obtained from synthetic blank urine samples and blank deionized water samples, both spiked after the extraction step. The matrix effect for each target analyte was expressed as the percentage ratio between the two measured concentrations.

2.3.8 Extraction recovery

The extraction recovery was determined by comparing the experimental results obtained from synthetic urine samples respectively spiked before and after the extraction step. It was expressed as the percentage ratio between the two quantified concentrations and estimated at the first, third and last calibration levels, in triplicate, for a total of 18 samples. In practice, only the experiments involving the spiking after the extraction step were added to the sequences performed to build the calibration curves.

2.3.9 Carry-over effect

The carry-over was evaluated by injecting distilled water extracts after the highest point of each calibration curve. If the signal-to-noise ratio was lower than 3 for each selected ion, the carry-over effect was considered negligible.

3. Results and Discussion

3.1 Validation protocol

To exemplify the step-by-step validation protocol implemented within our quality assurance policy, the analytical method based on gas chromatography-mass spectrometry (GC-MS) for the detection of 18 endogenous anabolic androgenic steroids (EAAS) in the urine of male individuals is used in this paper. This analytical procedure was developed for diagnostic purposes, to screen individuals with suspected prostate cancer [25]. All monitored steroids are reported in Table 1, with their CAS numbers, retention times and internal standard. Validation results obtained for all analytes are reported in Tables 2 to 6 and in the Supplementary Material. An in-depth discussion of the procedure is conducted in the following paragraphs for three steroids representing different and emblematic casework conditions:

- A target analyte (testosterone (T)) quantified using the corresponding isotopically-labelled homologue (testosterone-d3 (T-d3));
- A target analyte (4,6-androstadien-3,17-dione (6-D)) quantified using an isotopically-labelled compound with different structure (testosterone-d3), but belonging to the same chemical class (*viz.* C-19 steroids);
- A target analyte (androsterone (Andro)) quantified in a higher and wider calibration range using an exogenous compound (17α-methyltestosterone (17α-methyl-T)), not isotopically-labelled.

These three cases were selected as representative of experimental conditions commonly found in most analytical laboratories. The first condition is ideal, but it is not always feasible, for example due to the unavailability of isotopically-labelled standards for all members of a large series of target analytes. Conversely, the second and third conditions are easier to fulfil and are largely adopted by laboratories.

3.2 Linear Dynamic Range and Calibration

The linear instrumental response was evaluated within the concentration range C_S = 2-125 ng/mL for mix I, and C_S = 100-2250 ng/mL for mix II. The calibration curves of T, 6-D and Andro are reported in Figure 3, the values on the X-and Y-axis being C_S/C_{IS} and A_S/A_{IS}, respectively (where C_{IS} is the fixed concentration of the internal standard -considered adimensional-, A_S and A_{IS} are the area of the steroid and internal standard chromatographic peaks, respectively). The slope of the curves represents the response factor for each analyte and it is close to unity for T, as expected.

Although T and 6-D were spiked using the same working solution, the data-point variability exhibited by 6-D is considerably higher than what is observed for T. The same phenomenon occurs for other analytes in MIX I, with the exception of E (T-epimer). This reflects the parallel signals fluctuation of T and E with T-d3, which does not occur for the other analytes. It appears that the overall variability of experimental results increases as the difference between the analyte and internal standard structure becomes more imposing. Following this trend, data collected for the calibration of Andro is more scattered than for 6-D, since Andro and 17α-methyl-T display a larger structural difference than 6-D and T-d3.

All data-point distributions were found to be heteroscedastic, as indicated by congruous results from F-Test and Levene test, with the only exceptions of 4-hydroxytestosterone and 4-androsten-3,17-dione, both being characterized by low regression coefficients and limited dependence on the choice of alternative regression models. The best performing weighting factor was applied for all analytes (Table 2), namely $1/x^2$ weighting for most of them, with the exception of 6-D, 4-androsten-3,17-dione, and 5-androstendiol. Among the representative analytes studied (Figure 3), heteroscedasticity is apparently more pronounced for Andro, as the response variance becomes quite pronounced at the high concentration levels. This result can be attributed to the combined effect of structural difference with the internal standard, as discussed above, and the higher concentration range explored with respect to it. On the other hand, the significance of the heteroscedasticity F-test is extremely high for T, despite the limited spread of data-points observed at any concentration.

Indeed, for T and 6-D, the distributions at the lower concentration levels (from 2 to 25 ng/mL) appear to be almost homoscedastic, but a wider spread of the replicates is observed at the higher concentration levels (50 and 125 ng/mL). The described trends have been highlighted in previous studies [13,18].

The order of the calibration model was subsequently chosen. For most analytes (including T and 6-D) a quadratic model proved superior, while the introduction of the second order term turned out negligible for Andro and few of other target analytes, making the linear fitting more advisable for them. All the equations for the final calibration models are reported in Table 2. These models showed the best performance under the partial F-test for the quadraticity.

The ANOVA-LoF was computed for all the 18 analytes using the complete set of validation. This test was intended to verify the fit of data-points with the final calibration model, irrespective of its order and weighting. For 10 of them, the calibration model was rejected, despite the good accuracy and precision performances (see Table 6 and Supplementary Table 1). This is in line with the findings from Desharnais *et al.* [20], who concluded that the excessive sensitivity of this test to the chosen experimental design (in particular the number of replicates and of calibration levels) limits its practical applicability. Hence, we can consider not to use the outcome of this testing as a strict acceptance criterion within the validation protocol.

Even if this is widely done in method validation procedures and in the literature, Table 2 demonstrates that the adoption of identical calibration models (weighting factor, order) for large sets of target analytes does not provide the best fit for the data. Quite often, the software for data analysis provided with most instrumentation encourages the use of homogeneous models by asking their definition prior to testing. Additionally, insufficient statistical testing frequently characterizes the method validations reported in the literature.

Comparison between Figure 3 and Table 2 highlights the misleading role played by the squared correlation coefficient (R^2, also called global goodness of fit) [37,38] in the common practice of using it as a test for linearity. Although R^2 values higher than 0.99 were calculated for the linear calibration of T and 6-D (Figure 3) and could have been interpreted as a confirmation of the validity of the linear model, both visual inspection and Partial F-test test outcomes (Table 2) clearly indicate that the quadratic term should not be neglected in the studied setting.

After the choice of the calibration model was completed, back-calculation of the concentration for all levels was performed using the averaged signal from nine replications (Table 3). The mean deviation from the real value was below the threshold of 20% for most analytes at any level, with few exceptions at the lowest concentration levels (5-androstendiol, formestane, 5β-androstan-3,17-dione). The overall satisfactory results from this back-calculation process indicate that a proper calibration model was chosen for each analyte, which will be confirmed further by accuracy and precision results (Tables 5, 6, and S1).

3.3 LOD and LOQ

The adoption of the Hubaux-Vos method to calculate the LOD relies on the assumption that the calibration curve is linear. The method's performance is also greatly improved by the use of a concentration range restricted to the lower interval of the dynamic range. These two aspects are easily combined together, since deviation from linearity typically manifests itself in the upper part of the calibration range. Thus, linear calibrations can generally be obtained when the high concentration level(s) are excluded from the regression. For example, in Figure 3, the data-points alignment for 6-D and T is lost in correspondence to the last calibration point. If the latter point is excluded, the quadratic term loses statistical significance and a linear regression becomes the calibration model of choice.

When quadraticity is noted in the data, several calibration options are available: (i) use the quadratic calibration model to quantify the real samples; (ii) split the calibration into two ranges both fitted with a linear model; (iii) reduce the calibration range to a narrower linear interval. Whereas the second approach might be advisable whenever the calibration range covers two or more orders of magnitude, the latter strategy is a relevant choice to calculate the LOD. Figure 4 shows linear regressions for T and Andro with and without the sixth and highest calibration level. Obviously, the intercept of the green curve is closer to the origin of the y-axis, which is closer to the expected outcome and results in lower LOD values and reduced uncertainty in their estimation.

To illustrate this concept, LOD values were calculated using linear models and 6, 5, and 4 calibration levels, respectively (Table 4). For T and 6-D, the LOD values are 2- to 4-times lower when computed using the five-points calibration models instead of the six-points, while the LOD for Andro remained virtually unaffected, as expected. This is in agreement with previous studies, which demonstrated that LODs are overestimated whenever deviation from homoscedasticity and linearity are ignored [9,12]. Similar trends were observed for the remaining steroids: LOD values generally tend to decrease when the concentration levels used within the Hubaux-Vos method are reduced to the lowest calibration levels. On the other hand, this increases the risk of LOD underestimation and makes the experimental verification with blank samples spiked at the LOD level highly advisable. The results of this verification are reported in Table 4, where the most reliable calculated values are displayed in bold. This dependable LOD value was chosen after careful inspection of the data point distribution and experimental confirmation.

In the present method, targeted analytes were endogenous steroid, whose actual concentration in real samples may largely exceed the LOD values. While LOQ values could theoretically be defined as two or three times the LOD, we preferred to assess them as the minimal concentrations allowing to guarantee quantitative determinations with acceptable accuracy, within the range of physiological values. This makes LOQ and accuracy strictly conjoined concepts, as discussed below.

3.4 Accuracy and precision

An interesting feature of the proposed method validation procedure is that the large data set collected throughout several days for calibration purposes can be exploited to calculate intra and inter-day accuracy and precision at six concentration levels (instead of the common low, intermediate, and high levels) without requiring any further experimental work. These results for the studied method are reported in Tables 5 and 6.

The first highlight obtained from Table 5 is that none of the 216 accuracy results exceeds ±25, whereas 18 (8.3 %) exceed ±20, 21 (9.7 %) are comprised between ±15 and ±20, and the remaining 177 (82%) are within ±15. Only two substances (4-androsten-3,17-dione and formestane) display intra-day and inter-day accuracy which are not satisfactory at several concentrations. Beside these analytes, the overall accuracy of the method is acceptable and even fully satisfactory for most of the analytes, yet not perfect, as is expected for a wide set of authentic results. In particular, 7 out of the 18 accuracy results exceeding ±20 are found in the highest concentration level, while the others are randomly scattered throughout the other concentration levels. Again, this is to be expected since the calibration uncertainty is maximal at the extremes of the concentration range. Notably, the six steroids included in the steroidal module of the WADA Athlete Biological Passport (T, Andro, E, Etio, 5α-adiol, 5β-adiol), consistently provided accuracy data largely below 10, except 5α-adiol and 5β-adiol at the highest concentration level.

Another interesting observation is that inter-day accuracy frequently proved higher than intra-day accuracy, except for 7α-hydroxytestosterone. Together with the absence of any day-specific clustering of data-points in the nine-replicated analysis, this leads to the conclusion that under stable instrumental (GC-MS) conditions, it is preferable to use averaged calibration curves built from data collected on different days rather than changing the calibration curve daily with a new set of standards. Considering that the first and third sets of experiments were separated by twelve days, it is safe to assume that analytical data from standard solutions can be collected over a two weeks interval to build robust calibration curves from a large number of replicates. Possibly, these conclusions may not hold for LC-MS/MS methods, where large inter-day variation of the signal is commonly observed.

The tests used to verify intra-day and inter-day precision overall mirrors the conclusions drawn from the accuracy testing. Table 6 displays the data for the steroids included in the WADA Athlete Biological Passport, plus 6-D, while the rest of the variation coefficients is provided in the Supplementary Material Table S1.

Similarly to accuracy data, the precision performance was highest for T (likely due to the use of the isotopically labelled analogue T-d3 as the internal standard for signal correction), followed by the group of steroids included in the WADA list for the Athlete Biological Passport. All these steroids are either metabolites or direct precursors of T and have closely related chemical structures, hence yields in the critical steps of sample processing (extraction, derivatization, *etc.*) are likely more closely correlated with T-d3, thus achieving better signal correction.

All together, relatively good performances were obtained in terms of inter-day accuracy and precision, suggesting that the analytical method is robust and the response of the GC-MS instrument is stable within a two weeks' time frame. On a daily basis, a restricted number of standards can be analysed to evaluate the stability of the instrument response in the routine activity before starting the analytical session with real samples. The obtained results may serve the purpose of quality control for the current session, while new calibration data obtained afterwards could possibly be added to the ongoing calibration, to replace the oldest data. For example, updated calibration curves could be obtained in triplicate each week, and the corresponding validation parameters recalculated automatically using a worksheet or a routine such as the ones described elsewhere [10,27]. The quality control results obtained using the existing calibration confirms that correct operating conditions are maintained or, if the confidence limits are exceeded, that careful revision of the analytical steps and recalibration are needed.

In general, several method maintenance strategies are compatible with the validation scheme presented herein, depending on the specific method features and validation results. For example, the ongoing control of intra-day accuracy and precision may require a higher number of repetitions than the three sets of three repetitions proposed above. On the other hand, meticulous verification of intra-day variability may prove excessive if inter-day parameters turn out to be wholly under control.

3.5 Other validation parameters

Retention time repeatability proved satisfactory, as ascertained by both the calibration analyses and those made on fortified negativized samples (see Materials and Methods). No significant deviation from the expected retention time was observed. Repeatability of ion abundance ratios, tested for all the calibration levels, provided results within the limits of acceptance (±20%).

Likewise, the analytical method proved selective and specific for all the targeted compounds, since no interfering peaks appeared in any real urine sample around the retention time of the analytes with S/N above threshold of 3, during both the validation procedure and laboratory routine analyses.

Matrix effect and extraction recovery at three calibration levels (first, third, and sixth levels of the concentration range) yielded satisfactory outcomes, as expressed in term of percentage ratio between the concentrations measured in synthetic urine and water (Table S2), and synthetic urine spiked before and after the extraction step (Table S3), respectively. All percentages fell within the range of tolerance (85-115).

Lastly, no evidence of carry-over was detected, since the blank samples injected after the higher level of calibration showed no signal (*viz.* S/N < 3) for all the analytes.

4. Conclusions

The case study presented hereby investigated the complexity of developing a rigorous validation procedure for chromatographic methods devoted to multi-analyte targeted quantitative determinations of (endogenous) analytes. This complexity arises from the lack of homogeneity of the data amongst the different analytes, the need to verify several performance properties with serious statistical testing along long periods of potential application, and finally the need to ease the daily routine work when high throughput is required.

To meet these requirements, we proposed an operating protocol whose core consists in the systematic replication of three calibration curves in three different days (for a total of nine replicates) within a time lapse of 12 days. This protocol allowed robust evaluation of the calibration curves and simultaneous assessment of other validation parameters, including LOD, LOQ, intra- and inter-day precision and accuracy, ion abundance repeatability, selectivity, specificity and carry over.

Along this case study, we demonstrated that the application of different statistical tests for linearity and homoscedasticity typically produced non-homogeneous results across analytes and tests, which have to be evaluated independently with care before drawing any conclusion. In general, heteroscedastic and quadratic distribution of calibration data-sets were more frequent that linear trends, except when very limited concentration ranges were considered. Even analytes with similar chemical structures (*i.e.*, androgenic steroid) and similar concentration ranges may require different calibration criteria and should be selected case by case, unlike what is reported in numerous literature reports which appear overly simplified and optimistic. Moreover, the inadequacy of the regression coefficient R^2 to evaluate the linearity was demonstrated once more: quadratic and heteroscedastic data distribution proved compatible with high and misleading R^2 values.

The interdependence of LOD and LOQ values with calibration, accuracy, and precision parameters has been clearly shown. If LOD values are calculated partly or completely from the calibration regression equation, for example because real blank samples are lacking, then the number of concentration levels used for the regression and the distribution of data-points may strongly influence the outcome. An experimental confirmation of LOD and LOQ values and their verification with accuracy and precision testing is always advisable.

Lastly, our data showed that the collection of multiple calibration results, their averaging, and their continuous refreshing within a quality control process produced more accurate quantitation than the use of a single calibration collected on a daily basis.

References

[1] U.S. Department of Health and Human Services, Food and Drug Administration (FDA), Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), U.S. Department of Helth and Human Services, Guidance for Industry: Analytical Procedures and Methods Validation for Drugs and Biologics (Draft Guidance), (2014) 1–14. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM386366.pdf.

[2] I. Taverniers, M. De Loose, E. Van Bockstaele, Trends in quality in the analytical laboratory. II. Analytical method validation and quality assurance, TrAC - Trends Anal. Chem. 23 (2004) 535–552. doi:10.1016/j.trac.2004.04.001.

[3] G. Shabir, Step-by-step analytical methods validation and protocol in the quality system

compliance industry, J. Valid. Technol. 10 (2004) 210–218. doi:10.1038/ncomms11248.

[4] O. González, M.E. Blanco, G. Iriarte, L. Bartolomé, M.I. Maguregui, R.M. Alonso, Bioanalytical chromatographic method validation according to current regulations, with a special focus on the non-well defined parameters limit of quantification, robustness and matrix effect, J. Chromatogr. A. 1353 (2014) 10–27. doi:10.1016/j.chroma.2014.03.077.

[5] M. Kaza, M. Karaźniewicz-Łada, K. Kosicka, A. Siemiątkowska, P.J. Rudzki, Bioanalytical method validation: new FDA guidance vs. EMA guideline. Better or worse?, J. Pharm. Biomed. Anal. 165 (2019) 381–385. doi:10.1016/j.jpba.2018.12.030.

[6] D.E. Coleman, L.E. Vanatta, Lack-of-fit testing of ion chromatographic calibration curves with inexact replicates, in: J. Chromatogr. A, Elsevier, 1999: pp. 43–51. doi:10.1016/S0021-9673(99)00369-6.

[7] L.E. Vanatta, D.E. Coleman, Calibration, uncertainty, and recovery in the chromatographic sciences, J. Chromatogr. A. 1158 (2007) 47–60. doi:10.1016/j.chroma.2007.02.040.

[8] E. Rozet, A. Ceccato, C. Hubert, E. Ziemons, R. Oprean, S. Rudaz, B. Boulanger, P. Hubert, Analysis of recent pharmaceutical regulatory documents on analytical method validation, J. Chromatogr. A. 1158 (2007) 111–125. doi:10.1016/j.chroma.2007.03.111.

[9] A.M. Almeida, M.M. Castel-Branco, A.C. Falcão, Linear regression for calibration lines revisited: Weighting schemes for bioanalytical methods, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 774 (2002) 215–222. doi:10.1016/S1570-0232(02)00244-1.

[10] B. Desharnais, F. Camirand-lemyre, P. Mireault, C.D. Skinner, Procedure for the Selection and Validation of a Calibration Model I — Description and Application, J. Anal. Toxicol. 41 (2017) 261–268. doi:10.1093/jat/bkx001.

[11] A. Sayago, A.G. Asuero, Fitting straight lines with replicated observations by linear regression: Part II. Testing for homogeneity of variances, Crit. Rev. Anal. Chem. 34 (2004)

133–146. doi:10.1080/10408340490888599.

[12] T. Singtoroj, J. Tarning, A. Annerberg, M. Ashton, Y. Bergqvist, N.J. White, N. Lindegardh, N.P.J. Day, A new approach to evaluate regression models during validation of bioanalytical assays, J. Pharm. Biomed. Anal. 41 (2006) 219–227. doi:10.1016/j.jpba.2005.11.006.

[13] J. Sanchez, Estimating Detection Limits in Chromatography from Calibration Data: Ordinary Least Squares Regression vs. Weighted Least Squares, Separations. 5 (2018) 49. doi:10.3390/separations5040049.

[14] K. Baumann, H. Wätzig, Appropriate calibration functions for capillary electrophoresis II. Heteroscedasticity and its consequences, J. Chromatogr. A. 700 (1995) 9–20. doi:10.1016/0021-9673(95)00128-A.

[15] S. Bratinova, B. Raffael, C. Simoneau, Guidelines for performance criteria and validation procedures of analytical methods used in controls of food contact materials, 2009. doi:10.2788/49046.

[16] C.P. Da Silva, E.S. Emídio, M.R.R. De Marchi, Method validation using weighted linear regression models for quantification of UV filters in water samples, Talanta. 131 (2015) 221–227. doi:10.1016/j.talanta.2014.07.041.

[17] S. Rudaz, M. Feinberg, From method validation to result assessment: Established facts and pending questions, TrAC - Trends Anal. Chem. 105 (2018) 68–74. doi:10.1016/j.trac.2018.04.013.

[18] F. Schiller, D. Jena, L. a Currie, International Union of Pure and Applied Chemistry. Guideline for calibration in analytical chemistry- Part 1 . Fundamentals and single component, Pure Appl. Chem. 70 (1998) 993–1014. doi:10.3390/rs8030220.

[19] J. Scott Long, L. Ervin, Correcting for Heteroscedasticity with Heteroscedasticity Consistent Standard Errors in the Linear Regression Model: Small Sample Considerations, Indiana

Univ. Bloom. 47405 (1998) 1–33.

[20] B. Desharnais, F. Camirand-Lemyre, P. Mireault, C.D. Skinner, Procedure for the selection and validation of a calibration model II-theoretical basis, J. Anal. Toxicol. 41 (2017) 269–276. doi:10.1093/jat/bkx002.

[21] F. Raposo, Evaluation of analytical calibration based on least-squares linear regression for instrumental techniques: A tutorial review, TrAC - Trends Anal. Chem. 77 (2016) 167–185. doi:10.1016/j.trac.2015.12.006.

[22] E. Bernal, Limit of Detection and Limit of Quantification Determination in Gas Chromatography, Adv. Gas Chromatogr. (2014). doi:10.5772/57341.

[23] A. Hubaux, G. Vos, Decision and Detection limits for linear Calibration Curves, Anal. Chem. 42 (1970) 849–855. doi:10.1021/ac60290a013.

[24] A. Salomone, J.J. Palamar, R. Bigiarini, E. Gerace, D. Di Corcia, M. Vincenti, Detection of Fentanyl Analogs and Synthetic Opioids in Real Hair Samples, J. Anal. Toxicol. (2018). doi:10.1093/jat/bky093.

[25] E. Amante, E. Alladio, A. Salomone, M. Vincenti, F. Marini, G. Alleva, S. De Luca, F. Porpiglia, Correlation between chronological and physiological age of males from their multivariate urinary endogenous steroid profile and prostatic carcinoma-induced deviation, Steroids. 139 (2018) 10–17. doi:10.1016/j.steroids.2018.09.007.

[26] E. Alladio, G. Biosa, F. Seganti, D. Di Corcia, A. Salomone, M. Vincenti, M.R. Baumgartner, Systematic optimization of ethyl glucuronide extraction conditions from scalp hair by design of experiments and its potential effect on cut-off values appraisal, Drug Test. Anal. (2018). doi:10.1002/dta.2405.

[27] E. Alladio, E. Amante, C. Bozzolino, F. Seganti, A. Salomone, M. Vincenti, B. Desharnais, Experimental and statistical protocol for the effective validation of chromatographic

analytical methods, MethodsX. Submitted.

[28] P. Van Renterghem, P. Van Eenoo, H. Geyer, W. Schänzer, F.T. Delbeke, Reference ranges for urinary concentrations and ratios of endogenous steroids, which can be used as markers for steroid misuse, in a Caucasian population of athletes, Steroids. 75 (2010) 154–163. doi:10.1016/j.steroids.2009.11.008.

[29] B. Desharnais, F. Camirand-Lemyre, P. Mireault, C.D. Skinner, Procedure for the Selection and Validation of a Calibration Model I—Description and Application, J. Anal. Toxicol. 41 (2017) 261–268. doi:10.1093/jat/bkx001.

[30] Levene, Robust tests for equality of variances, in: I. Olkin, H. Hotelling (Eds.), Contrib. to Probab. Stat. Essays Honor Harold Hotell., Stanford University Press, 1960: pp. 278–292.

[31] L.A. Currie, Detection and quantification limits: Origins and historical overview, Anal. Chim. Acta. 391 (1999) 127–134. doi:10.1016/S0003-2670(99)00105-1.

[32] V.P. Shah, K.K. Midha, J.W.A. Findlay, H.M. Hill, J.D. Hulse, I.J. Mcgilveray, G. Mckay, K.J. Miller, R.N. Patnaik, M.L. Powell, A. Tonelli, C.T. Viswanathan, Bioanalytical Method Validation—A Revisit with a Decade of Progress, 17 (2000).

[33] M.R. Lakshmi HimaBindu, S. Angala Parameswari, C. Gopinath, A review on GC-MS and method development and validation, Int. J. Pharm. Qual. Assur. 4 (2013) 42–51.

[34] S. V. Karmarkar, Validation of Ion Chromatographic Methods, Appl. Ion Chromatogr. Pharm. Biol. Prod. (2012) 285–308. doi:10.1002/9781118147009.ch15.

[35] G. a Shabir, Validation of high-performance liquid chromatography methods for pharmaceutical analysis, J. Chromatogr. A. 987 (2003) 57–66. doi:10.1016/S0021-9673(02)01536-4.

[36] A.G. González, M.Á. Herrador, A.G. Asuero, Intra-laboratory assessment of method accuracy (trueness and precision) by using validation standards, Talanta. 82 (2010) 1995–

1998. doi:10.1016/j.talanta.2010.07.071.

[37] Scientific Working Group for Forensic Toxicology (SWGTOX), Scientific working group for forensic toxicology (SWGTOX) standard practices for method validation in forensic toxicology, J. Anal. Toxicol. 37 (2013) 452–474. doi:10.1093/jat/bkt054.

[38] P. Araujo, Key aspects of analytical method validation and linearity evaluation, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 877 (2009) 2224–2234. doi:10.1016/j.jchromb.2008.09.030.

	Target analyte	CAS number	T_R (min)	Internal standard
Mix I	**5β-androstan-13,17-dione**	1229-12-5	8.15	Testosterone-D_3
	5α-androstane-3α,17β-diol (5α-adiol)	1852-53-5	9.32	Testosterone-D_3
	5β-androstane-3α,17β-diol (5β-adiol)	1851-23-6	9.40	Testosterone-D_3
	dehydroepiandrosterone (DHEA)	53-43-0	10.00	Testosterone-D_3
	5-androsten-3,17-diol	512-17-5	10.24	Testosterone-D_3
	epitestosterone (E)	481-30-1	10.35	Testosterone-D_3
	4,6-androstadien-3,17-dione (6-D)	633-34-1	10.51	Testosterone-D_3
	dihydrotestosterone (DHT)	521-18-6	10.51	Testosterone-D_3
	4-androsten-3,17-dione	63-05-8	10.69	Testosterone-D_3
	Δ6-testosterone	2484-30-2	10.75	Testosterone-D_3
	testosterone (T)	58-22-0	10.92	Testosterone-D_3
	7α-hydroxytestosterone	62-83-9	11.24	Testosterone-D_3
	7β-hydroxy-dehydroepiandrosterone (7β-OH-DHEA)	2487-48-1	11.98	Testosterone-D_3
	formestane	566-48-3	13.14	Testosterone-D_3
	4-hydroxytestosterone	2141-17-5	13.31	Testosterone-D_3
	16α-hydroxyandrosten-3,17-dione	63-02-5	13.60	Testosterone-D_3
Mix II	**androsterone (Andro)**	53-41-8	9.05	17α-methyl-testosterone
	etiocholanolone (Etio)	53-42-9	9.18	17α-methyl-testosterone

Calibration level	1	2	3	4	5	6
Mix I *(ng/mL)*	2	5	10	25	50	125
Mix II *(ng/mL)*	100	200	500	1000	1500	2250

Table 1. Working mixtures, target analytes, CAS ID numbers, GC-MS retention times, internal standard used for quantification and concentration levels of the target analytes used to build the calibration curves.

Target Analyte	**Calibration Range (ng/mL)**	**Weight**	**Model**	**Equation**	**F-test for heteroscedasticity**	**Levene Test for heteroscedasticity**	**Partial F-test for quadraticity**	**ANOVA - LoF**	**Squared Correlation Coefficient**
T	2-125	$1/x^2$	Quadratic	$-0.038x^2 + 1.25x + 0.005$	**1.3 × 10^{-11}**	**1.4 × 10^{-5}**	**6.8 × 10^{-8}**	**5.1 × 10^{-3}**	0.9942
6-D	2-125	$1/x$	Quadratic	$-0.020x^2 + 0.599x + 0.023$	**2.6 × 10^{-6}**	**1.7 × 10^{-5}**	**1.7 × 10^{-3}**	1.8 × 10^{-1}	0.9949
Andro	100-2250	$1/x^2$	Linear	$0.549x + 0.018$	**2.3 × 10^{-9}**	**4.1 × 10^{-5}**	9.7 × 10^{-1}	1.4 × 10^{-1}	0.9999
16α-hydroxyandrosten-3,17-dione	2-125	$1/x^2$	Quadratic	$-0.013x^2 + 0.164x + 0.006$	**2.0 × 10^{-11}**	**7.2 × 10^{-8}**	**2.9 × 10^{-2}**	6.9 × 10^{-1}	0.9884
4-androsten-3,17-dione	2-125	$1/x$	Quadratic	$-0.001x^2 + 0.017x + 0.010$	**1.2 × 10^{-2}**	5.6 × 10^{-1}	**3.5 × 10^{-3}**	**8.0 × 10^{-4}**	0.9538
4-hydroxytestosterone	2-125	$1/x^2$	Linear	$0.024x + 0.002$	**1.2 × 10^{-10}**	1.1 × 10^{-1}	1.5 × 10^{-1}	7.2 × 10^{-1}	0.9984
5α-adiol	2-125	$1/x^2$	Quadratic	$-0.022x^2 + 0.378x + 0.005$	**1.9 × 10^{-11}**	**5.8 × 10^{-4}**	**6.6 × 10^{-5}**	**4.8 × 10^{-2}**	0.9893
5β-adiol	2-125	$1/x^2$	Quadratic	$-0.020x^2 + 0.319x + 0.005$	**1.5 × 10^{-11}**	**2.6 × 10^{-4}**	**1.1 × 10^{-5}**	**2.4 × 10^{-2}**	0.9865
5β-androstan-3,17-dione	2-125	$1/x^2$	Quadratic	$-0.045x^2 + 0.676x + 0.059$	**4.7 × 10^{-10}**	**8.4 × 10^{-9}**	**2.5 × 10^{-2}**	**1.8 × 10^{-4}**	0.9624
5-androstendiol	2-125	$1/x$	Quadratic	$-0.024x^2 + 0.347x + 0.022$	**1.6 × 10^{-4}**	**3.5 × 10^{-5}**	**2.1 × 10^{-5}**	**2.1 × 10^{-6}**	0.9607
7α-hydroxytestosterone	2-125	$1/x^2$	Linear	$0.171x + 0.011$	**5.8 × 10^{-8}**	**6.8 × 10^{-3}**	8.2 × 10^{-1}	1.2 × 10^{-1}	0.9935
7β-OH-DHEA	2-125	$1/x^2$	Linear	$0.286x + 0.035$	**2.2 × 10^{-9}**	**3.7 × 10^{-4}**	5.6 × 10^{-1}	6.1 × 10^{-1}	0.9964
Δ6-testosterone	2-125	$1/x^2$	Quadratic	$-0.013x^2 + 0.577x + 0.032$	**3.5 × 10^{-5}**	**7.1 × 10^{-3}**	**6.2 × 10^{-3}**	7.8 × 10^{-2}	0.9957
DHEA	2-125	$1/x^2$	Quadratic	$-0.020x^2 + 0.453x + 0.001$	**1.4 × 10^{-9}**	**2.3 × 10^{-3}**	**2.1 × 10^{-7}**	**3.7 × 10^{-3}**	0.9933
DHT	2-125	$1/x^2$	Quadratic	$-0.019x^2 + 0.286x –$	**3.2 × 10^{-11}**	**6.9 × 10^{-6}**	**2.8 × 10^{-4}**	1.2 × 10^{-1}	0.9922

				0.001					
E	2-125	$1/x^2$	Quadratic	$-0.042x^2 + 1.260x + 0.002$	**2.9×10^{-13}**	**4.6×10^{-7}**	**3.5×10^{-12}**	**6.0×10^{-3}**	0.9946
Etio	100-2250	$1/x^2$	Linear	$0.472x + 0.031$	**6.2×10^{-9}**	**1.0×10^{-3}**	5.7×10^{-1}	3.1×10^{-1}	0.9996
formestane	2-125	$1/x^2$	Linear	$0.748x + 0.074$	**6.3×10^{-4}**	**3.0×10^{-4}**	9.2×10^{-2}	**8.4×10^{-3}**	0.9929

Table 2. Calibration model parameters for all targeted analytes. Values of p<5×10-2 (0.05) indicated the occurrence of heteroscedasticity (F-test and Levene test for heteroscedasticity), a relevant contribution of the quadratic term (partial F-test for quadraticity), and a significant deviation from the calibration model (ANOVA-LoF),). The corresponding values are reported in bold type.

	Calibration levels (Deviation %)					
Target Analyte	**1**	**2**	**3**	**4**	**5**	**6**
T	0.4	0.3	3.2	1.2	5.8	2.3
6-D	18.8	11.8	10.4	0.4	4.6	0.8
Andro	0.5	1.0	0.5	1.6	0.1	0.7
16α-hydroxyandrosten-3,17-dione	3.2	11.9	7.7	5.1	10.7	13.0
4-androsten-3,17-dione	9.3	15.5	12.1	11.0	6.9	2.9
4-hydroxytestosterone	7.4	16.2	8.2	10.3	9.8	16.4
5α-adiol	0.1	0.4	0.2	0.9	0.6	0.1
5β-adiol	0.4	1.7	1.3	0.2	0.2	0.1
5β-androstan-3,17-dione	21.7	8.4	12.2	13.5	15.9	3.6
5-androstendiol	37.3	22.2	20.3	5.4	14.3	3.9
7α-hydroxytestosterone	5.1	10.0	8.6	3.2	14.4	3.9
7β-OH-DHEA	16.4	8.4	13.6	4.5	1.8	0.4
Δ6-testosterone	17.8	6.3	13.8	3.5	6.9	1.0
DHEA	0.6	1.1	1.5	1.2	1.9	1.2
DHT	0.1	1.9	0.3	9.9	5.0	12.7
E	0.8	1.7	1.4	0.1	4.0	1.9
Etio	0.5	0.5	1.1	2.4	2.0	1.6
formestane	23.1	21.4	13.3	1.8	14.2	4.5

Table 3. Back-calculation results

Target Analyte	**LOD *(ng/mL)* 6 points**	**LOD *(ng/mL)* 5 points**	**LOD *(ng/mL)* 4 points**	**Experimentally verified LOD (*ng/mL*)**
T	0.7*	**0.2**	0.1	0.5
6-D	1.7*	**0.5**	2.3*	0.5
Andro	0.6	**0.4**	0.3	0.5
16α-hydroxyandrosten-3,17-dione	0.7*	0.8	**0.7**	1.0
4-androsten-3,17-dione	3.1*	3.3	**0.8**	1.0
4-hydroxytestosterone	0.7	0.6*	0.6*	1.0
5α-adiol	0.7*	**0.5**	0.3	0.5
5β-adiol	0.9*	**0.3**	0.2	0.5
5β-androstan-3,17-dione	4.5*	1.1*	**0.5**	0.5
5-androsten-3,17-diol	5.3*	**1.1**	1.8*	1.0
7α-hydroxytestosterone	1.8	1.8*	**0.5**	0.5
7β-OH-DHEA	0.5*	**0.4**	2.2*	0.5
Δ6-testosterone	1.0*	0.4	**0.3**	0.5
DHEA	1.2*	**0.2**	<0.1	0.5
DHT	0.5	**0.4**	<0.1	0.5
E	0.5*	**0.1**	<0.1	0.5
Etio	0.7	**0.5**	0.4	0.5
formestane	0.9	1.8*	**0.4**	0.5

*Although a quadratic trend was detected, it was ignored to compute the LOD values.

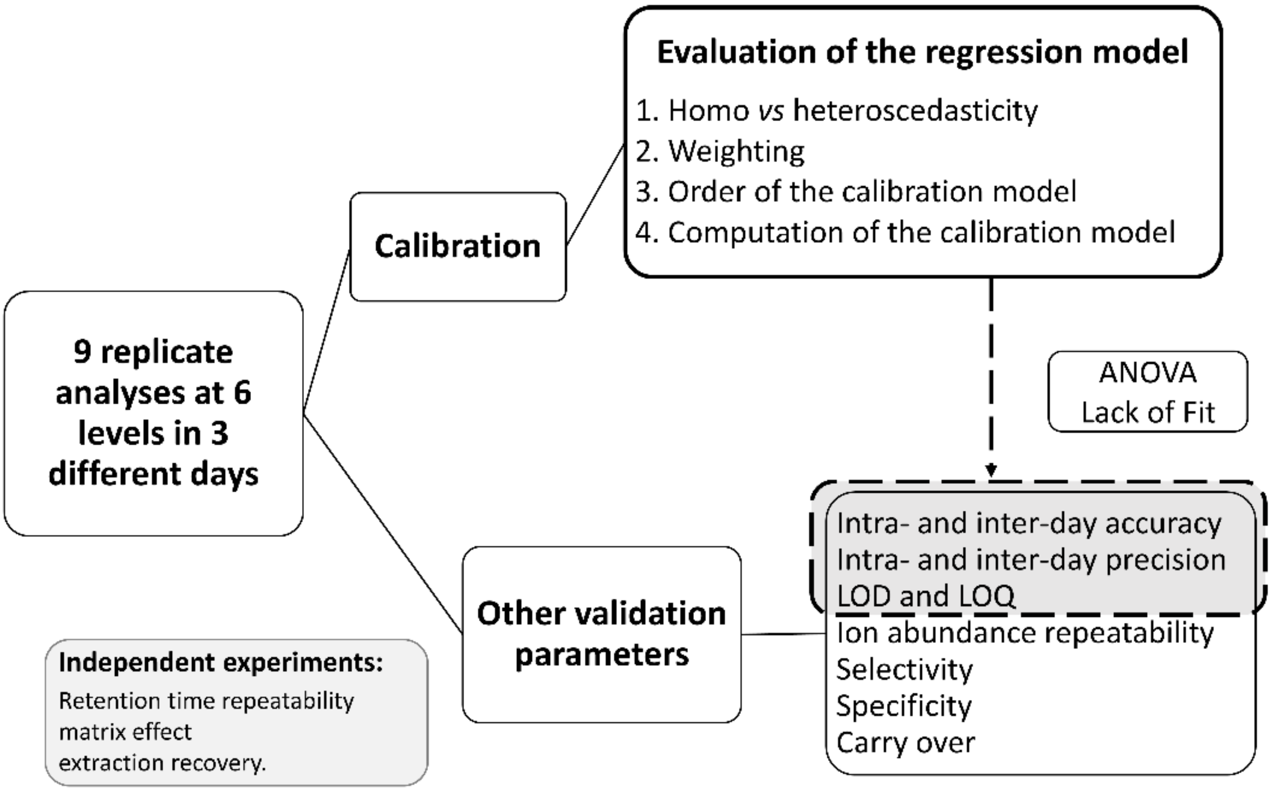
Table 4. The first three columns report the LOD values computed with the Hubaux-Vos algorithm using 6, 5, and 4 calibration levels. Weighting corrections were applied as described in Section 2.3.1. The last column displays the LOD values experimentally verified by spiking blank urine with the concentrations reported.

		Calibration Level (% bias)					
Target Analyte		**1**	**2**	**3**	**4**	**5**	**6**
T	**Intra-day**	0.4	0.5	-3.3	-1.2	5.6	-2.1
	Inter-day	0.4	0.3	-3.3	-1.3	5.4	-2.7
6-D	**Intra-day**	*16.1*	-12.2	-10.3	0.9	4.8	-10.3
	Inter-day	8.1	-12.3	-8.1	3.2	7.6	-1.5
Andro	**Intra-day**	0.2	-1.9	0.9	-0.8	-4.9	-5.1
	Inter-day	1.2	-0.3	0.6	2.3	0.8	-0.1
16α-hydroxyandrosten-3,17-dione	**Intra-day**	**-25.0**	14.5	-6.0	10.9	-11.5	*-19.5*
	Inter-day	-8.9	10.7	-6.5	7.4	-3.7	-0.6
4-androsten-3,17-dione	**Intra-day**	-4.8	*-19.9*	**21.65**	-1.1	*18.0*	**-23.1**
	Inter-day	*-17.3*	-24.9	-14.1	**-21.6**	**-21.9**	**-21.5**
4-hydroxytestosterone	**Intra-day**	0.3	14.6	8.4	-1.1	*17.5*	-1.1
	Inter-day	-8.3	*16.3*	9.9	-9.4	11.4	*-19.9*
5α-adiol	**Intra-day**	-4.9	-3.4	-3.7	-1.4	-4.7	-8.9
	Inter-day	1.3	0.6	1.2	2.0	1.0	**-21.8**
5β-adiol	**Intra-day**	0.5	-1.4	1.3	0.9	-0.1	0.4
	Inter-day	1.2	-0.6	2.9	1.6	2.1	**-20.6**
5β-androstan-3,17-dione	**Intra-day**	12.7	*-16.7*	**-21.5**	-4.0	**20.3**	**-23.1**
	Inter-day	14.9	-6.6	-9.1	-9.3	14.6	-12.2
5-androsten-3,17-diol	**Intra-day**	12.6	**-23.6**	*-16.7*	1.6	*16.5*	-6.4
	Inter-day	-11.5	-14.4	**-23.6**	*-15.1*	7.5	**-20.7**
7α-hydroxytestosterone	**Intra-day**	4.8	-13.6	-14.6	-0.1	10.2	-9.4
	Inter-day	*-19.3*	*-18.3*	*-18.0*	2.7	**-21.5**	**-25.0**
7β-OH-DHEA	**Intra-day**	8.4	-2.7	-9.2	*17.4*	10.6	-0.4
	Inter-day	9.0	-9.0	-12.5	8.3	4.0	5.0
Δ6-testosterone	**Intra-day**	9.0	-11.9	-9.8	-1.7	5.9	-5.6
	Inter-day	11.2	-4.8	-12.5	-0.1	8.8	9.5
DHEA	**Intra-day**	1.4	-0.9	-1.7	1.0	1.6	-0.1
	Inter-day	0.8	-0.9	-1.2	1.4	1.9	-1.2
DHT	**Intra-day**	7.4	-11.0	-11.8	-11.6	-8.2	*-18.5*
	Inter-day	11.0	5.5	5.5	12.4	7.7	-9.2
E	**Intra-day**	5.4	0.1	-0.6	-0.1	3.2	-3.6
	Inter-day	0.8	-1.7	-1.4	0.0	3.9	-1.5
Etio	**Intra-day**	-0.2	0.3	1.3	-1.9	-5.1	-4.1
	Inter-day	0.8	1.8	2.6	3.5	-0.7	-0.6
formestane	**Intra-day**	**25.0**	**-21.9**	*-15.7*	-2.1	9.5	-3.2
	Inter-day	*18.8*	*-17.9*	*-18.6*	-1.6	8.8	-9.0

Table 5. Intra-day and inter-day accuracy results expressed in terms of bias. Good results are expected in the range ± 15%, and acceptable results in the range ± 20% (reported in italics). Results exceeding 20 are indicated in bold.

		Calibration Level (CV %)					
Target Analyte		**1**	**2**	**3**	**4**	**5**	**6**
T	**Intra-day**	4	8	2	2	5	3
	Inter-day	4	8	2	4	5	4
6-D	**Intra-day**	39	21	12	8	10	4
	Inter-day	39	24	15	9	17	8
Andro	**Intra-day**	3	4	6	7	5	2
	Inter-day	10	11	18	15	12	8
E	**Intra-day**	4	3	5	2	3	2
	Inter-day	4	3	6	2	3	5
Etio	**Intra-day**	3	4	6	7	2	2
	Inter-day	13	14	20	16	15	10
5α-adiol	**Intra-day**	0.1	0.1	1	1	0.1	3
	Inter-day	0.2	0.4	2	2	3	26
5β-adiol	**Intra-day**	30	16	8	12	15	5
	Inter-day	16	13	19	16	16	23

Table 6. Intra-day and inter-day precision, expressed in terms of CV. Acceptable results are expected in the range ± 30%


Evaluation of the regression model
1. Homo *vs* heteroscedasticity
2. Weighting
3. Order of the calibration model
4. Computation of the calibration model
Calibration
9 replicate analyses at 6 levels in 3 different days
ANOVA
Lack of Fit
Intra- and inter-day accuracy
Intra- and inter-day precision
LOD and LOQ
Ion abundance repeatability
Selectivity
Specificity
Carry over
Other validation parameters
Independent experiments:
Retention time repeatability
matrix effect
extraction recovery.

603 Figure 1. Scheme of the validation protocol.

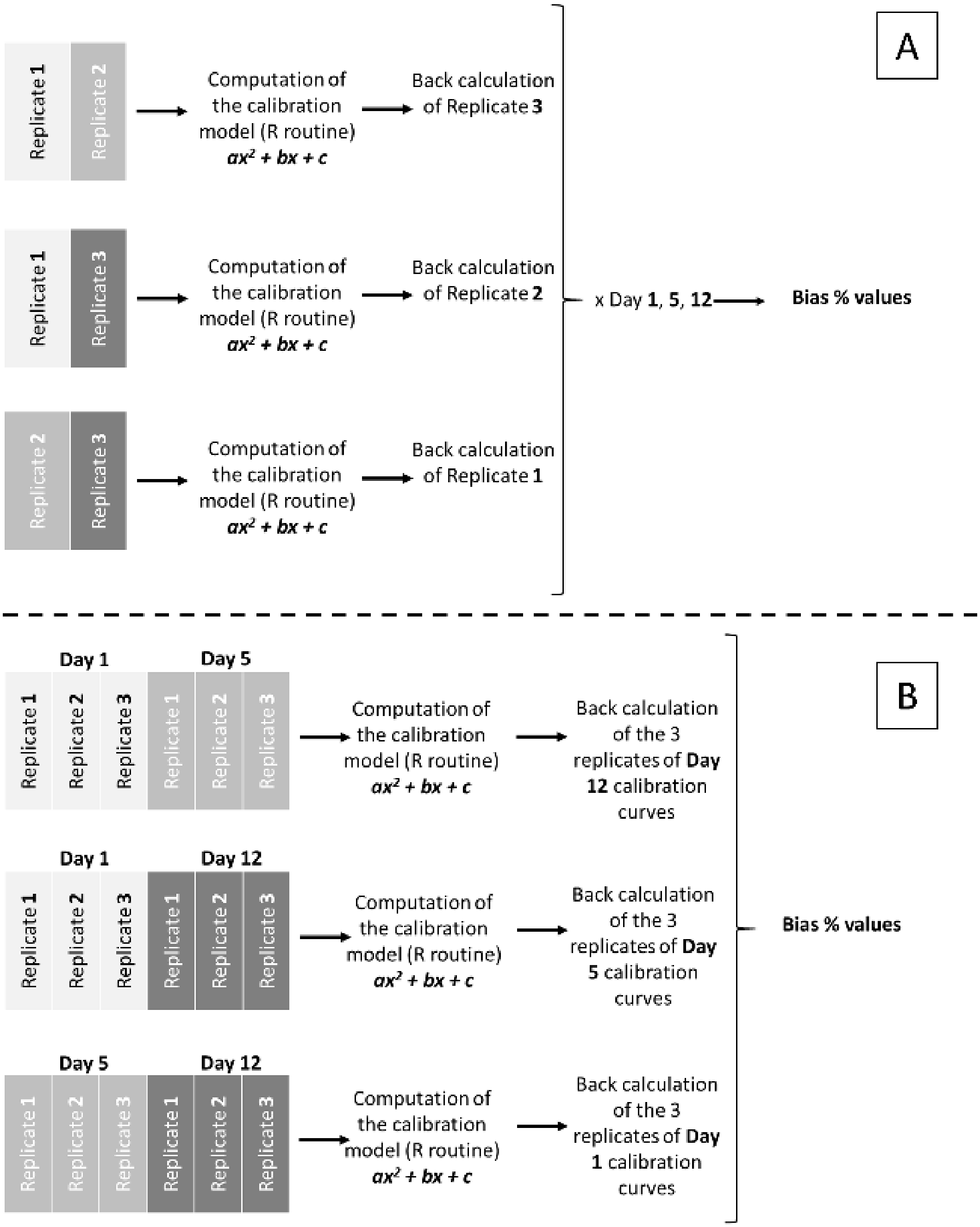

A
Replicate 1
Replicate 2
Computation of the calibration model (R routine) $ax^2 + bx + c$
Back calculation of Replicate 3
Replicate 1
Replicate 3
Computation of the calibration model (R routine) $ax^2 + bx + c$
Back calculation of Replicate 2
Replicate 2
Replicate 3
Computation of the calibration model (R routine) $ax^2 + bx + c$
Back calculation of Replicate 1
x Day 1, 5, 12
Bias % values
B
Day 1
Day 5
Replicate 1
Replicate 2
Replicate 3
Replicate 1
Replicate 2
Replicate 3
Computation of the calibration model (R routine) $ax^2 + bx + c$
Back calculation of the 3 replicates of Day 12 calibration curves
Day 1
Day 12
Replicate 1
Replicate 2
Replicate 3
Replicate 1
Replicate 2
Replicate 3
Computation of the calibration model (R routine) $ax^2 + bx + c$
Back calculation of the 3 replicates of Day 5 calibration curves
Day 5
Day 12
Replicate 1
Replicate 2
Replicate 3
Replicate 1
Replicate 2
Replicate 3
Computation of the calibration model (R routine) $ax^2 + bx + c$
Back calculation of the 3 replicates of Day 1 calibration curves
Bias % values

Figure 2. Operating scheme for the computation of the intra-day (A) and inter-day (B) accuracy.

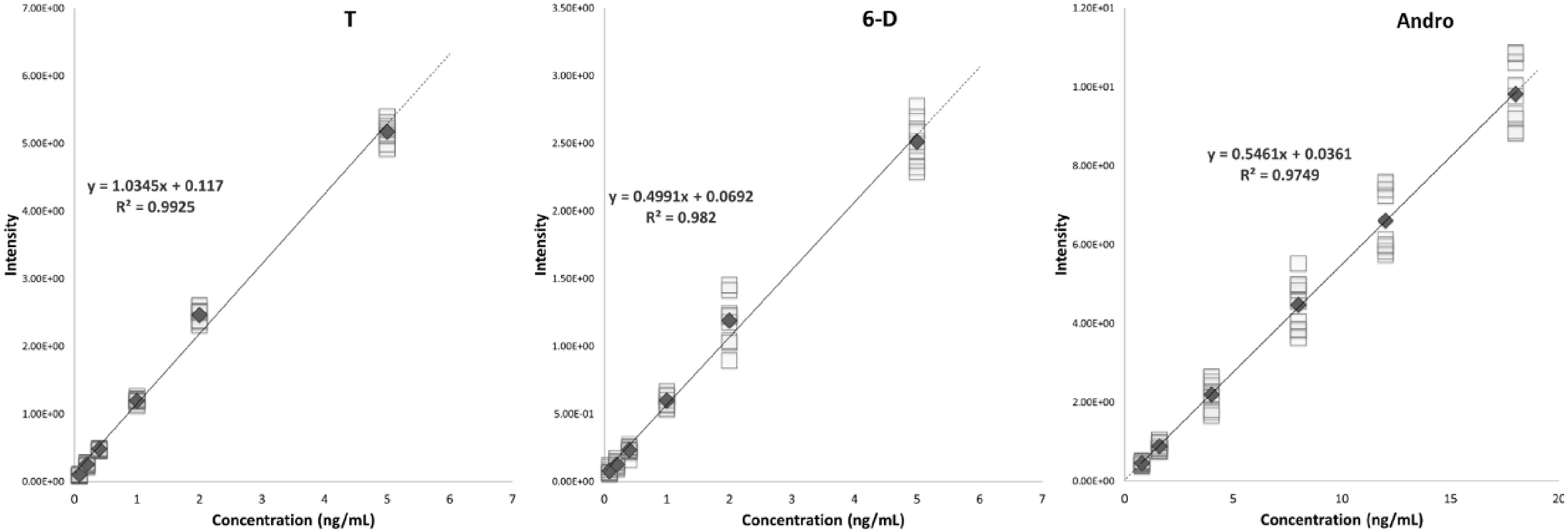
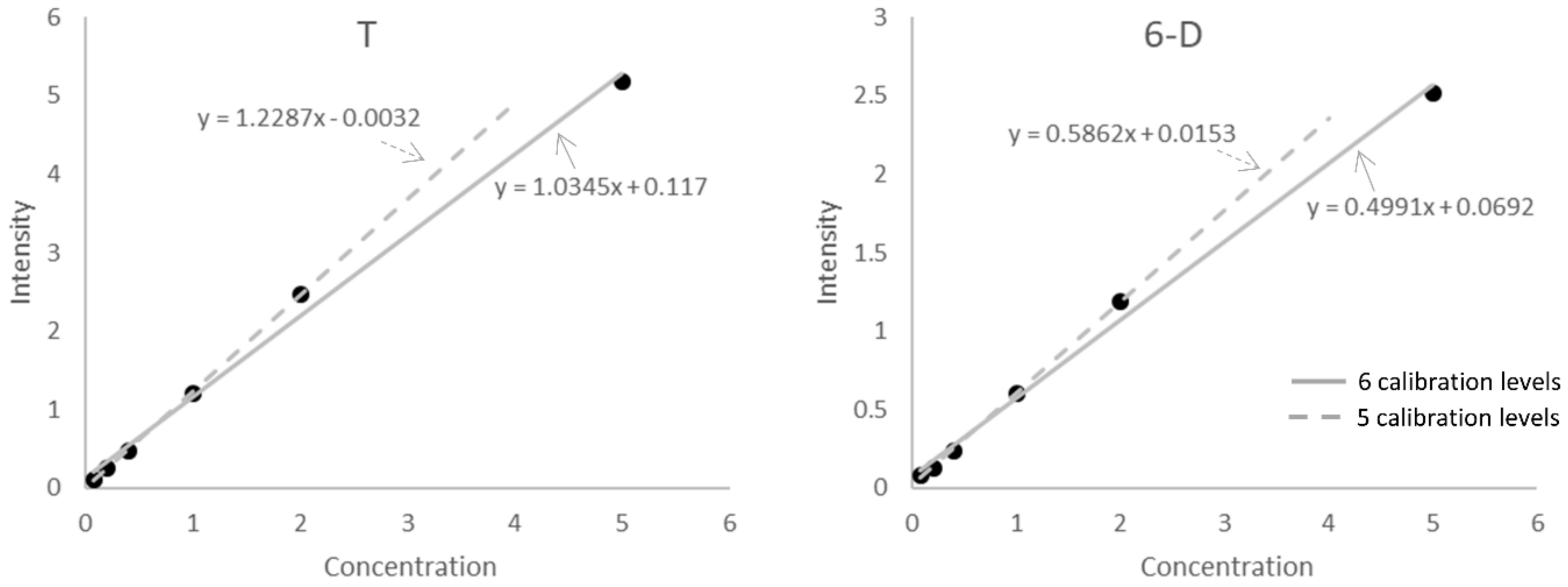


Figure 3. Distribution of the 54 calibration data-points (9 replicates × 6 calibration levels) for testosterone (T), 4,6-androstadien-3,17-dione (6-D), and androsterone (Andro).



610

611 Figure 4. Linear regression for the six-points (solid line) and five-points (dashed line) calibration curves.

***Declaration of Interest Statement**

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: