MethodsX 7 (2020) 100919

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Method Article

Experimental and statistical protocol for the effective validation of chromatographic analytical methods

Eugenio Alladio[a,b,1], Eleonora Amante[a,b,1,*], Cristina Bozzolino[a], Fabrizio Seganti[b], Alberto Salomone[a,b], Marco Vincenti[a,b], Brigitte 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, Montreal, Quebec, Canada

A B S T R A C T

The validation of analytical methods is of crucial importance in several fields of application. A new protocol for the validation of chromatographic methods has been proposed. The overall protocol is described in a parallel paper, where the case of a multi-targeted gas chromatography – mass spectrometry (GC–MS) method for the determination of androgens in human urine is in-depth discussed. The purpose of this paper is to report the details about the GC–MS separation and detection of the target analytes, and to provide the mathematical formulas needed to perform the validation of the principal parameters. Briefly, the validation protocol foresees the repetition of three calibration curves in three different days, providing a total amount of nine replicates. Such a structured design allows to use the same experiments to

- perform a rigorous calibration study, by the evaluation of heteroscedasticity, comparison of several weights and linear/quadratic calibration curves.
- determine several parameters which are traditionally computed from dedicated experiments, namely intra- and inter-day accuracy and precision, limit of detection, specificity, selectivity, ion abundance repeatability, and carry over.
- Finally, few further experiments are necessary to evaluate the retention time repeatability, matrix effect and extraction recovery.

DOI of original article: 10.1016/j.talanta.2020.120867
* Corresponding author at: Dipartimento di Chimica, Università degli Studi di Torino, Turin, Italy.
E-mail address: eleonora.amante@unito.it (E. Amante).
[1] The authors equally contributed to this work.

https://doi.org/10.1016/j.mex.2020.100919

ARTICLE INFO
Method name: Effective validation protocol for chromatography mass spectrometry analytical methods
Keywords: Chromatographic method, Validation protocol, Multiresidual analysis, GC-MS
Article history: Received 9 April 2020; Accepted 6 May 2020; Available online 16 May 2020

Specifications Table

Subject area:	Chemistry
More specific subject area:	Analytical Chemistry
Method name:	Effective validation protocol for chromatography – mass spectrometry analytical methods
Name and reference of original method:	Not applicable
Resource availability:	Not applicable

Method details

This paper accompanies the paper entitled "*Effective validation of chromatographic analytical methods: the illustrative case of androgenic steroids*" [1], which presents a new, systematic validation protocol for chromatographic analytical methods. As case study, the fully validation of a multiresidual GC–MS method for the detection of androgens is human urine is discussed. The details related to the separation and acquisition methods are reported in this paper; specifically, the oven temperature program of the gas chromatograph is reported in Fig. 1, together with the typical total ion current (TIC) profile of a real urine sample. Moreover, details about the mass spectrometer (MS) detection of the 18 target compounds (i.e. retention time, quantifier and monitored ions) plus the molecular weight after trimethylsilyl derivation are in Table 1.

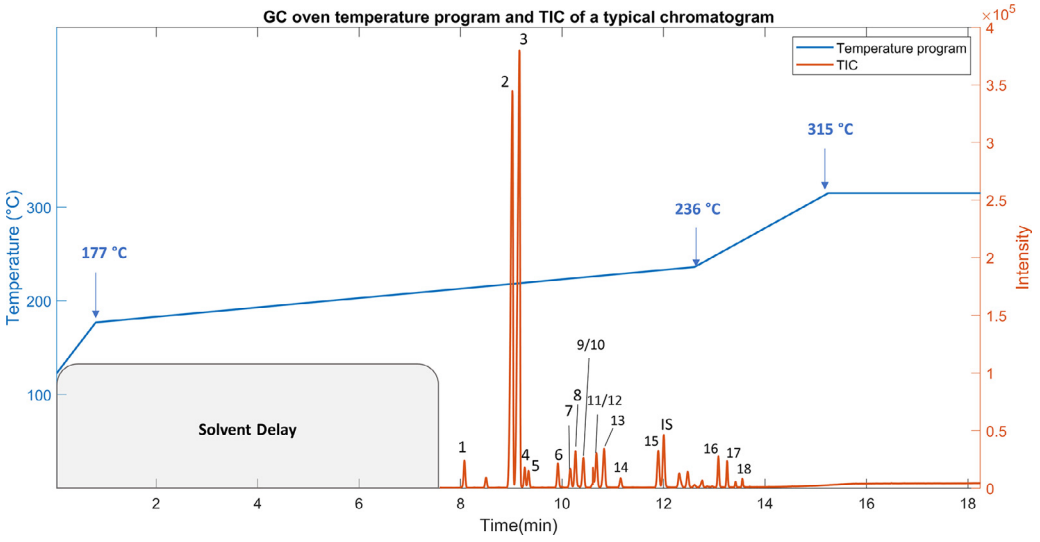

Fig. 1. Temperature program of the GC oven (blue line) and typical chromatographic profile (orange line). Coded target analytes are: (1) 5β-androstan-3,17–dione, (2) A, (3) Etio, (4) 5α-adiol, (5) 5β-adiol, (6)DHEA, (7) 5-androsten-3,17-diol, (8) E, (9) 4,6-androstadien-3,17–dione, (10) DHT, (11) 4-androsten-3,17–dione, (12) Δ6-testosterone, (13) testosterone + testosterone-D3, (14) 7α-hydroxytestosterone, (IS) 17-methyl-testosterone, (15) 7β-OH-DHEA, (16) Formestane, (17) 4-hydroxytestosterone, (18) 16α-hydroxyandrosten-3,17–dione (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Table 1
List of the analytes included in Mix I and Mix II, with the relative CAS number and the internal standard used for their quantitation. The concentrations at the different calibration levels are also reported.

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

Furthermore, the validation protocol is described in the Experimental Design Section, and all the parameters (homoscedasticity evaluation, linearity tests such as ANOVA, Mandel's test and Lack of Fit, limit of detection, intra- and inter-day accuracy and precision, matrix effect, extraction recovery) are defined, together with the equations for their computations.

Experimental design, materials, and methods

Analytical method

Samples pre-treatment

The sample preparation involved the fortification of 6 mL of urine 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 h 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 min, centrifuged at 6.24 g for 5 min 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/NH4I/dithioerythritol – 1000:2:4 v/w/w), the reaction was allowed to proceed at 70 °C for 30 min. 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. Mix I and II had distinct calibration ranges (Table 1), selected on the basis of the expected physiological concentrations, as reported in literature [2,3].

GC–MS separation and detection

The GC–MS method optimization was the subject of another study [3]. The GC separation was performed using an Agilent 6890 N instrument (Agilent Technologies, Milan, Italy) equipped with a J&W Scientific HP-1, 17 m x 0.2 mm (i.d.) x 0.11 mm (f.t.) capillary column. Helium was employed as the carrier gas at a constant pressure of 18.5 psi. The temperature program of the GC oven was

set as follows: initial temperature equal to 120 °C, then a 70 °C/min heating rate was applied until the temperature of 177 °C was reached. Subsequently, the temperature was raised to 236 °C with a 5 °C/min gradient. A final heating rate of 30 °C/min allowed to rise the temperature of 315 °C, which was hold for 3 min. The GC injector and transfer line were maintained at 280 °C. The temperature program is reported in Fig. 1 (blue line).

The trimethylsilyl derivatives of the analytes were ionized and fragmented in EI at 70 eV using an Agilent 5975 inert mass-selective detector (Agilent Technologies, Milan, Italy). The MS was operated in the selected ion monitoring mode and three diagnostic ions for each analyte were monitored with dwell times of 20–50 ms. The details about the retention times and the monitored ions are reported in Tables 1 and S2 of the parallel paper [1] In Fig. 1 (orange line and Arabic numbers) is reported the typical Total ion current (TIC) profile of a spiked urine sample.

Validation protocol

The validation protocol is in-depth described in the parallel paper [1]. Briefly, nine replicates of the calibration curve are analyzed in three different days (three replicates/die). This peculiar experimental design allowed the simultaneous evaluation of several parameters, which are typically evaluated performing dedicated experiments, resulting in expensive and timewasting protocols. Among these, a particular focus was put on the study of the calibration curve, with tests of homoscedasticity, quadraticity, ANOVA, Lack of Fit and goodness of the back calculation. The calibration curves were also used for the evaluation of the limit of detection (LOD, by Hubaux and Vos' approach) and intra- and inter-day accuracy and precision. Furthermore, ion abundance repeatability, selectivity, specificity and carry over were studied employing the same experiments. Lastly, few further experiments were performed to determine matrix effect and extraction recovery.

The principal equations employed are reported below.

Nomenclature

In this article, the calibration levels are indicated as *1,2,…i,….k* and the replicates as *1,2,…j,….l* and the total number of samples analyzed is $k \times l = n$.

Computation of the calibration model

Test for heteroscedasticity. The homoscedasticity was tested twice, e.g. using a partial F-Test integrated in the R routine (Eq. (1)) developed by Desharnais et al. [4] and the Levene equation (Eq. (2)) [5]. In the first case, the presence of heteroscedasticity was investigated using a unilateral F-test for the calculation of the probability that the variance of measurements at the upper limit of quantification (ULOQ) was equal to or smaller than the variance of measurements at the lower limits of quantification (LLOQ). The Rstudio function used for the computation is

$$var.test\left(Measurements_{LLOQ}, Measurements_{ULOQ},\ alternative = \text{“less”}\right) \tag{1}$$

Unlike the unilateral F-test described above, the Levene test was applied on all the calibration levels. It is a robust alternative to the F-test and was used to confirm the results obtained with the calibration routine. The equation of Levene test is the following:

$$W = \frac{(n-k)}{(k-1)} \times \frac{\sum_{i=1}^{k} l\left(\bar{Z}_i - \bar{Z}\right)^2}{\sum_{i=1}^{k}\sum_{j=1}^{l}\left(Z_{ij} - \bar{Z}_i\right)^2} \tag{2}$$

With

$$Z_{ij} = \left|y_{ij} - \bar{y}_i\right|$$

k is the number of calibration levels tested, $\bar{Z}_i$ is the average of all the Z_{ij} of a calibration level and $\bar{Z}$ is the average of all the Z_{ij}, in the original version of the test, or their median, from the Brown-Forsythe modification, which is more robust towards heavy-tailed distributions [6]. The RStudio function *levene.test* was used to perform the calculations (in the Brown-Forsythe version).

The W statistics can be compared to an F distribution with {(k-1),(n-k)} degrees of freedom. If the p-value is smaller than the α level of significance chosen (in our case, 0.05), then the variances are considered as significantly different, i.e. the data are heteroscedastic. If $p > \alpha$, the data is consistent with an equality of variances.

Partial F-test for the quadratic term

The Partial F-test is a hypothesis test which relies on comparing the sum of squares of the regression to the mean square of residuals (Eq. (3)):

$$F_{exp} = \frac{SS_{reg,Q} - SS_{reg,L}}{\left(\frac{SS_{res,Q}}{n-3}\right)} \tag{3}$$

Where $SS_{reg,Q}$ and $SS_{reg,L}$ are the sum of squares of the regression in the quadratic and linear models, respectively (Eq. (4)). SS_{res} is the sum of mean squares in residuals (Eq. (5)).

$$SS_{reg} = \sum_{i=1}^{k} w_i \times l \times \left(\hat{y}_i - \bar{y}_{ij}\right)^2 \tag{4}$$

$$SS_{res} = \sum_{i=1}^{k}\sum_{j=1}^{l} w_i \times \left(y_{ij} - y_i\right)^2 \tag{5}$$

The p-value associated with F_{exp} can be found using the RStudio command *1-pf(F_{calc},1,(n-3))*. A $p < 0.05$ denotes a significant improvement in the model fit brought by the use of a quadratic model.

Analysis of Variance - Lack of Fit (ANOVA-LoF) to verify the goodness of the calibration model

The ANOVA-LoF hypothesis test is used to evaluate the fit of data-points with the final calibration model. The null-hypothesis is that there is no lack of fit and the F is computed as follows (Eq. (6)):

$$F_{LoF} = \frac{\frac{SS_{LoF}}{DoF_{LoF}}}{\frac{SS_{Pure}}{DoF_{Pure}}} = \frac{\frac{\sum_{i=1}^{k}\sum_{j=1}^{l} w_i \times \left(\overline{y_i} - \hat{y_i}\right)^2}{k-q}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{l} w_i \times \left(y_{ij} - y_i\right)^2}{n-k}} \tag{6}$$

It is important to underline that this test is very sensitive to experimental design, in particular the number of replicates and/or the number of calibration levels. Hence, if accuracy and precision are within the limits of acceptability, it is possible to ignore the outcome of this test.

Limit of detection (LOD)

The limit of detection is the lower concentration detectable with the specified analytical method. It can be evaluate using several different approaches; here, we propose the Hubaux and Vos' computation [7].

The approach relies on five hypotheses:

1. The standards are independent
2. The contents of the standards are accurately known
3. The observed signals have a gaussian distribution
4. A linear regression model is adequate for the data at hand
5. The variance of the error is constant (i.e. homoscedastic data).

Assuming that the first three prerequisites are met, it is necessary to focus on numbers 4 and 5, which are not necessarily respected. When linearity is not respected, it is possible to reduce the calibration range excluding the upper calibration levels, in order to exclude the quadraticity.

If the homoscedasticity is not respected, the weights need to be introduced into the Hubaux and Vos equation (Eqs. (7)–(11):

$$X_{LOD} = t_{(0.05,n-2)} \times s_{\hat{y}0} \tag{7}$$

Where t is the Student's test, value at 0.05 confidence limit and n-2 degrees of freedom, and s_{y0} is equal to

$$s_{\hat{y}0} = S_{y/x}\sqrt{1 + \frac{1}{\sum_{i=1}^{k} l \times w_i} + \frac{\left(-\bar{x}_w\right)^2}{\sum_{i=1}^{k} l \times w_i \left(x_i - \bar{x}_w\right)^2}} \tag{8}$$

$S_{y/x}$ and x_w are, respectively:

$$S_{y/x} = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{l} w_i \times \left(y_i - y_{ij}\right)^2}{n-2}} \tag{9}$$

$$\bar{x}_w = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} w_i x_i}{\sum_{i=1}^{k}\sum_{j=1}^{l} w_i} \tag{10}$$

And, finally,

$$Y_c = bX_c + a \tag{11}$$

a is the intercept of the calibration curve and b the slope of the calibration curve.

Once the concentration of analyte constituting the LOD is mathematically obtained, an experimental verification is needed. It consists in the fortification of blank matrix at the computed X_{LOD} and the measurement of the Signal-to-Noise, which has to be higher than 3.

Accuracy

The accuracy is a measure of the closeness of a measured value to the actual value. The three replicates measured in each validation day allow the computation of the intra-day accuracy, and the 12-days timeframe of the overall validation procedure allows the evaluation of the inter-day accuracy. The two computations are performed employing the R routine developed by Desharnais et al. [4,8] following the operating scheme presented here [1]. The method's accuracy is expressed in terms of bias%, which it is measured as follows:

$$bias_{\%} = \left(1 - \frac{x_{real}}{\bar{x}_{exp}}\right) \times 100 \tag{12}$$

Where x_{real} is the spiked concentration and $\bar{x}_{exp}$ is the experimental result.

Precision

The precision is the reproducibility of a measurement, i.e. describes how close are the replicates. It is expressed as coefficient of variance and computed as follows:

$$CV_{\%} = \frac{\frac{\sum_{j=1}^{J}\left(x_{exp} - \bar{x}\right)^2}{J-1}}{\bar{x}} \times 100 \tag{13}$$

Where J is the number of replicates, x_{exp} is the experimental result of the j-replicate and $\bar{x}$ is the mean result.

Matrix effect (ME)

To evaluate the ME, bi-distilled water and synthetic urine are spiked, after the extraction step, at the desired concentration (typically, three concentration levels are tested, i.e. low, middle and high). The ME is provided by the ratio of the means of the replicates (minimum of three):

$$ME_{\%} = \frac{\overline{\left(A_S/A_{IS}\right)_w}}{\overline{\left(A_S/A_{IS}\right)_u}} \times 100 \tag{14}$$

Where A_S and A_{IS} are the area of the standard and the internal standard, respectively; w indicates bi-distilled water and u synthetic urine. Values between 85% and 115% are considered acceptable.

Extraction recovery (ER)

The ER is evaluated comparing the results obtained spiking the standards and internal standards before and after the extraction procedure. The number of replicates and the concentration levels usually tested are the same reported in the ME description. The formula is the following:

$$ER_{\%} = \frac{\overline{\left(A_S/A_{IS}\right)}_{before}}{\left(\overline{A_S/A_{IS}}\right)_{after}} \times 100 \quad (15)$$

Where A_S and A_{IS} are defined as above, *before* and *after* are the samples spiked before and after the extraction, respectively. Again, values between 85% and 115% are considered acceptable.

Method relevance

The traditional validation protocols for analytical methods use dedicated experiments to evaluate each validation parameter, resulting in a large set of experiments to be completed and prolonged execution times. To ease the whole process, it is frequently observed in scientific publications that the validation experiments are cut with detriment to their statistical significance. In particular, most of the published validation procedures show weak and unsubstantial selection of the calibration curve parameters, with regard to heteroscedasticity, weighting, range, and linearity.

The proposed protocol is based on an *ad hoc* design of experiments which allows to use the same set of experiments (e.g., three replicates of the calibration curve in three different days) to produce rigorous evaluation of the most important validation parameters. Core of the present validation procedure is the computation of a reliable calibration model with solid statistical foundation, which is mandatory whenever the achievement of accurate quantitation represents the main objective of the analysis. Limits of detection and quantitation, range, accuracy, and precision parameters are computed thereafter, using the same data with different statistical processing. Thus, the present approach combines the advantage of reducing the number of experiments needed to complete an analytical method validation with the use of a robust statistical apparatus, which compares a wide set of statistical tests and provides the most appropriate mathematical adjustments to the computed parameters.

Funding

This research was funded by the Italian “Ministero dell’Istruzione, dell’Università e della Ricerca” within a PRIN 2017 call for bids (Research Projects of Relevant National Interest—grant 2017Y2PAB8)

Declaration of Competing Interest

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.

References

[1] E. Alladio, E. Amante, C. Bozzolino, F. Seganti, A. Salomone, M. Vincenti, B. Desharnais, Effective validation of chromatographic analytical methods: the illustrative case of androgenic steroids, Talanta 215 (2020) 120867, doi:10.1016/j.talanta.2020.120867.

[2] 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.

[3] 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.

[4] 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.

[5] Levene, Robust tests for equality of variances, in: I. Olkin, H. Hotelling (Eds.), Contributions to Probability and Statistics; Essays in Honor of Harold Hotelling., Stanford University Press, 1960, pp. 278–292.
[6] M.B. Brown, A.B. Forsythe, Robust tests for the equality of variances, J. Am. Stat. Assoc. 69 (1974) 364–367.
[7] A. Hubaux, G. Vos, Decision and Detection limits for linear Calibration Curves, Anal. Chem. 42 (1970) 849–855, doi:10.1021/ac60290a013.
[8] 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.