Procedure for the selection and validation of a calibration model
II —Theoretical basis

Brigitte Desharnais[a,b,*], Félix Camirand-Lemyre[c], Pascal Mireault[a,b], Cameron D. Skinner[b]

[a]*Department of Toxicology, Laboratoire de sciences judiciaires et de médecine légale*
1701 Parthenais Street, Montréal, Québec, Canada H2K 3S7
[b]*Department of Chemistry & Biochemistry, Concordia University*
7141 Sherbrooke Street West, Montréal, Québec, Canada H4B 1R6
[c]*Department of Mathematics, Université de Sherbrooke*
2500 boulevard de l'Université, Sherbrooke, Québec, Canda J1K 2R1

Abstract

In the first part of this paper (I — Description and application), an automated, stepwise and analyst independent process for the selection and validation of calibration models was put forward and applied to two model analytes. This second part presents the mathematical reasoning and experimental work underlying the selection of the different components of this procedure. Different replicate analysis designs (intra/inter-day and intra/inter-extraction) were tested and their impact on test results was evaluated. For most methods, the use of intra-day/intra-extraction measurement replicates is recommended due to its decreased variability. This process should be repeated three times during the validation process in order to assess the time stability of the underlying model. Strategies for identification of heteroscedasticity and their potential weaknesses were examined and a unilateral F-test using the lower limit of quantification and upper limit of quantification replicates was chosen. Three different options for model selection were examined and tested: ANOVA lack-of-fit (LOF), partial F-test and significance of the second-order term. Examination of mathematical assumptions for each test and LC-MS/MS experimental results lead to selection of the partial F-test as being the most suitable. The advantages and drawbacks of ANOVA-LOF, examination of the standardized residuals graph and residuals normality testing (Kolmogorov-Smirnov or Cramer-Von Mises) for validation of the calibration model were examined with the last option proving the best in light of its robustness and accuracy. Choosing the correct calibration model improves QC accuracy, and simulations have shown that this automated scheme has a much better performance than a more traditional method of fitting with increasingly complex models until QC accuracies pass below a threshold.

[*]Author to whom correspondence should be addressed. Email: brigitte.desharnais@msp.gouv.qc.ca

1. Introduction

Choosing an appropriate calibration model is an important part of quantitative method validation. The analyst has several choices to make: forcing the calibration through the origin, choosing a weighting factor and selecting model order (e.g., quadratic or linear). Recommendations as to how to make these choices are frequently vague and do not address all necessary decisions [1, 2, 3, 4], although some recent papers have begun to address these issues [5, 6, 7, 8, 9].

The first decision concerns which experimental design to employ to obtain the replicates necessary for data analysis. The goal is to mimic the process used in the production setting once validation is complete. It was shown in the first part of this paper that seven or more replicates are beneficial to improve the accuracy of model selection. One must also consider whether replicate measurements should be performed on different days or not, and with samples extracted separately or several injections of the same extract. Some sources suggest the use of an inter-day/inter-extracts setup [2, 10], but this recommendation does not appear to be supported by a reasoned decision based on mathematical concepts or experimental results.

An option available to the analyst is to force the calibration through the origin. Although no paper actually advocates this procedure and some explicitly discourage it [2, 5, 10], this is a software option available and has been used in papers [11, 12, 13]. This practice therefore needs to be clearly addressed.

The absolute error of replicate measurements can be independent of concentration (homoscedastic data) or scale with concentration (heteroscedastic data) [5, 14]. Regression minimizes the sum of squared error (difference between the measured values and predicted values) by selecting optimal calibration coefficients (slope(s) and intercept). When regression is performed on heteroscedastic data, greater importance should be given to the data with the smaller absolute error [15]. Choosing the proper weighting level from the common options (uniform weighting [“no weight”], $^1/_x$ or $^1/_{x^2}$) is an important part of obtaining a calibration robust to normal changes in individual measurement values [7, 15, 16]. Whereas SWGTOX guidelines points toward the use of residuals graph to select proper weighting [2], others use different tests to confirm the presence of heteroscedasticity [5, 6, 17] and select the proper weighting factor [7].

Model order selection is frequently necessary since non-linear behavior is expected with some methodologies (e.g., LC-MS/MS). Appropriate calibration models capture the systematic behavior of the instrument’s response but do not model the random error. Excessive model order (“overfitting”) results in inclusion of the random error in the model and actually reduces accuracy when the model is used [18]. This topic has been covered more thoroughly in the literature than weight selection, with suggestions of using the analysis of variance lack-of-fit (ANOVA-LOF) [2, 6, 18], residuals graph [2, 5, 10, 15], partial F-test [18] and significance of the second-order term [2, 6, 18] for appropriate model order selection.

Once the calibration model is selected, a good but often overlooked step is to validate that the model describes only the systematic behavior of the data. ANOVA-LOF [2, 5, 6, 10, 14, 18, 19] and examination of the residuals graph [2, 10, 15] have been suggested as methods for validating final calibration models.

In the first paper, we outlined a generalized method for selecting and validating a calibration model. The procedure first tested for heteroscedasticity using an F-test on the lower limit of quantification (LLOQ) and the upper limit of quantification (ULOQ) measurements. Weight selection was performed using variance evaluation to examine which weighting (no weight, $^1/_x$ or $^1/_{x^2}$) produced the smallest spread in weighted normalized variances. A partial F-test was then used to select the model order (quadratic or linear). Finally, the model was validated through testing the residuals for normality (Kolmogorov-Smirnov (KS) or Cramer-Von Mises (CVM) test). This procedure is an automated, analyst-independent approach to selection and validation of calibration models.

In choosing each part of this procedure, the main considerations were accuracy of the result, robustness, ease of use, mathematical soundness and how adequately it fit with real situations faced in toxicological analyses. In this paper, the different procedures tested are detailed. The mathematical reasoning justifying the selected procedures, buttressed by experimental work with 50 analytes quantified by LC-MS/MS are presented.

2. Materials and methods

2.1. LC-MS/MS quantification

Fifty analytes were spiked in bovine blood at concentrations of 5, 10, 15, 50, 75, 100, 400, 500 and 1000 ng/mL to produce a set of calibration standards. The analytes were obtained from Cerilliant (Round Rock, TX, USA) and belonged to the benzodiazepine, opiate, cocaine and amphetamine families and are listed in Supplemental Data 1. Amphetamine-D_8, benzoylecgonine-D_3, clonazepam-D_4, cocaetylene-D_8, cocaine-D_3, ephedrine-D_3, diazepam-D_5, MDEA-D_5, oxycodone-D_3, methamphetamine-D_5 and codeine-D_3 (Cerilliant) were used as internal standards for the analytes (concentrations and internal standard assignation are available in Supplemental Data 1). Sample preparation and analysis details can be found in the “Materials and methods” section of the accompanying paper. This method has been validated according to ISO 17025 and CAN-P-1578 guidelines and is currently used as a routine quantification method. Although in the first paper, seven measurement replicates were found to improve the success rate, we present here data collected in accordance with current SWGTOX guidelines which dictate five measurement replicates. To generate inter-day, inter-extraction data, a set of standards was extracted and analyzed in the same day for 5 different days. To generate intra-day, inter-extraction data, five aliquots of the set of standards were extracted separately and injected on the same day. To generate intra-day, intra-extraction data, a set of standards was extracted and five injections of the extract were performed on the same day. Finally, inter-day, intra-extraction data were generated by extracting one set of standards and injecting an aliquot of the extract each day for 5 different days. Data analysis was performed with Multiquant^{TM} (AB Sciex, Framingham, MA, USA).

2.2. Simulated data sets

The procedure used to generate simulated data sets for the six different calibration models was described in the "Materials and methods" section of the accompanying paper.

2.3. Heteroscedasticity testing

The procedure and calculations used to perform an F-test for heteroscedasticity testing were described in the "Materials and methods" section of the accompanying paper.

2.4. Tests for weight selection

2.4.1. Examination of the variance graph

A plot of the variance as a function of the concentration was generated for each analyte. The variance for the five measurements at each concentration level was obtained in Excel 2010 (Microsoft, Redmond, WA, USA) by the formula $= VAR.S(Measurements)$. Constant variance across the calibration range indicated unweighted regression, while a linear increase in variance indicated $^1/_x$ and a parabolic increase indicated that a $^1/_{x^2}$ weighting factor should be selected [5, 7, 10]. Weight selection was thus based on visual inspection of the graph by the analyst.

2.4.2. Variance evaluation

The procedure and calculations used to perform the variance evaluation for weight selection were described in the "Materials and methods" section of the accompanying paper.

2.5. Tests for model order

2.5.1. Analysis of variance lack-of-fit

Coefficients for linear ($y_L = b_1 \cdot x + b_0$) and quadratic ($y_Q = b_2 \cdot x^2 + b_1 \cdot x + b_0$) equations were obtained for each analyte using the data for all replicates ($n = 45$). The mean square for the pure (or experimental) error (MS_{PE}) was first calculated by [18]

$$MS_{PE} = \frac{SS_{PE}}{dof_{PE}} = \frac{\sum_i \sum_j W_i \times (y_{ij} - \overline{y_i})^2}{n_{ij} - k} \tag{1}$$

where SS_{PE} was the sum of squares of the pure error, dof_{PE} was the number of degrees of freedom of the SS_{PE} value, W_i was the weighting applied at the i^{th} concentration level (e.g., a $^1/_x$ weighting at the 5 ng/mL concentration level will be $^1/_5 = 0.2$), y_{ij} was the j^{th} measurement at the i^{th}i concentration level, $\overline{y_i}$ was the average of all measurements at the i^{th} concentration level, n_{ij} was the total number of measurements (9 concentration levels × 5 replicates = 45 measurements) and k was the number of concentration levels (maximum value of i, here 9).

The mean square for the lack-of-fit (MS_{LOF}) was then calculated by

$$MS_{LOF} = \frac{SS_{LOF}}{dof_{LOF}} = \frac{\sum_i \sum_j W_i \times (\overline{y_i} - \hat{y_i})^2}{k - z} \tag{2}$$

where SS_{LOF} was the sum of squares of the lack-of-fit, dof_{LOF} was the degrees of freedom of the SS_{LOF} value, $\hat{y_i}$ was the predicted measurement at the i^{th} concentration level (obtained by inserting the concentration in the calibration equation) and z was the number of regression parameters ($z = 3$ for quadratic models and 2 for linear models).

The F value was obtained through

$$F_{model} = \frac{MS_{LOF}}{MS_{PE}} \tag{3}$$

and the probability (P) associated with the F statistic was calculated using the Excel 2010 function $= F.DIST.RT(Fmodel; (k - z); (n_{ij} - k))$.

For each analyte, these calculations were performed for the quadratic and linear calibration models. The P_{quad} (P-value for the quadratic model) was compared with the P_{linear}, and the model order with the largest P-value was retained for calibration.

2.5.2. Partial F-test

The procedure and calculations used to perform a partial F-test for selection of model order are described in the "Materials and methods" section of the accompanying paper.

2.6. Tests for validation of the chosen model

2.6.1. Standardized residuals graph

Standardized residuals were calculated for each measurement by [20]

$$s_{e_{ij}} = \frac{(y_{ij} - \hat{y_i})}{s_e \times \sqrt{1 - h_{ij}}} \tag{4}$$

where $s_{e_{ij}}$ was the standardized residual for the j^{th} measurement at the i^{th} concentration level, $\hat{y_i}$ was the predicted measurement at the i^{th} concentration level, s_e was the estimate of the standard deviation of the residuals and h_{ij} was the leverage value for observation ij.

The estimate of the standard deviation of the residuals se was calculated by

$$s_e = \sqrt{\frac{\sum_1^{i \times j} (y_{ij} - \hat{y_i})^2}{n - z}} \tag{5}$$

where z was the number of regression parameters, i.e., $z = 2$ for linear regressions or 3 for quadratic regressions.

The leverage value for each observation h_{ij} was calculated by

$$h_{ij} = \frac{1}{n} - \frac{(x_{ij} - \overline{x})^2}{\sum_{1}^{i \times j} (x_{ij} - \overline{x})^2} \tag{6}$$

where x_{ij} was the concentration associated to the measurement y_{ij} , n was the total number of measurements and $\overline{x}$ was the average concentration over all standards measured (9 concentration levels x 5 replicates = 45 concentration values).

All standardized residuals were then plotted against their concentration level (i) [2]. If the calibration model is appropriate, the residuals should be distributed randomly around the $y = 0$ line [15, 21].

2.6.2. Analysis of variance - lack-of-fit

The ANOVA-LOF procedure can be used as a test for both model order and model validation. Once a model order had been chosen using the procedure presented above in Section 2.5, the P-value could be (re-)used to validate the calibration model. A P-value above the 0.05 threshold indicated that the error attributable to the LOF was not significant compared with the experimental or pure error [15]; therefore, the model was validated. A P-value below 0.05 marked an LOF error significantly larger than the experimental error; the model used was therefore not an adequate fit to the experimental data.

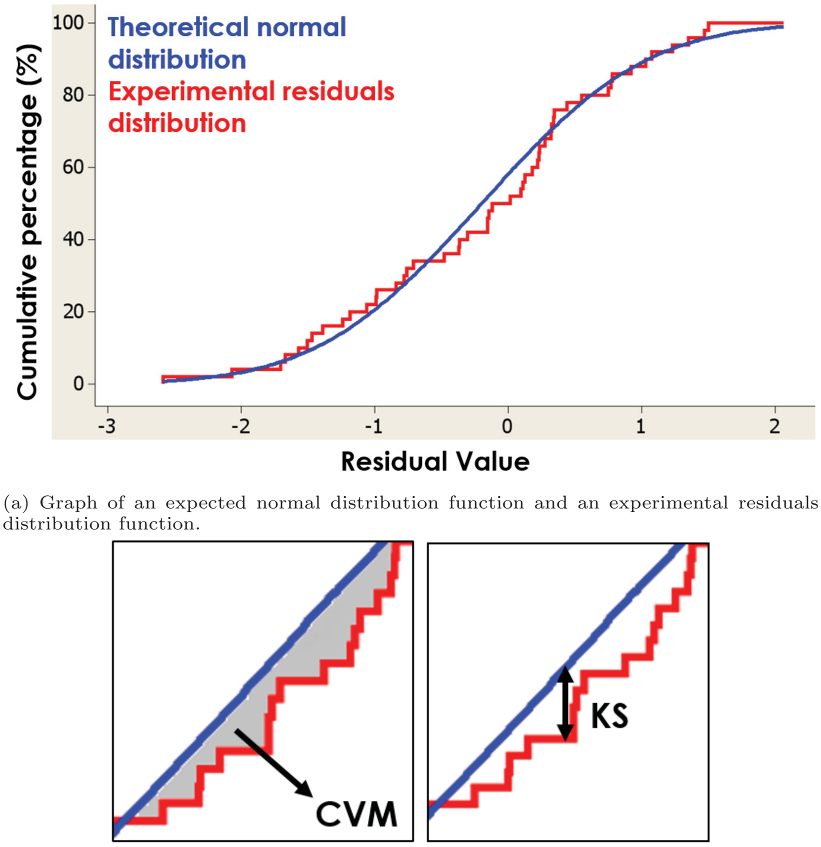
2.6.3. Normality of the residuals

Calculations for testing the normality of the standardized residuals will not be detailed here since they are beyond the scope of this paper and require a specialized knowledge of statistics [22]. However, a general description of the operations performed is provided here. Readers interested in a fuller understanding can consult the R scripts available in Supplemental Data 3 of the first paper which details all calculations.

When the chosen calibration model accurately describes all systematic trends in the data, the residuals correspond to pure random error. This means the residuals should be randomly distributed around zero (i.e., follow a normal distribution). Thus, normality testing of the residuals is an appropriate means of validating a calibration model [21]. This approach is conceptually similar to the standardized residuals plot, but produces a definitive, analyst-independent result.

To evaluate whether the residuals are normally distributed, the distribution function of the residuals is compared with the distribution function obtained from the expected normal distribution (see Figure 1).

A distribution function plots the proportion of data points that are smaller or equal to all values present in the distribution. Normally distributed residuals produce a sigmoidal distribution curve. In Figure 1a for example, at residual = 0 (when the experimental data value = predicted value by the model), the normal distribution function has $y = 50\%$; therefore, 50% of all the residuals are smaller than or equal to 0.



(a) Graph of an expected normal distribution function and an experimental residuals distribution function.

(b) Illustration of the calculation of the CVM statistic.

(c) Illustration of the calculation of the KS statistic.

Figure 1: Graphical representation of normality testing process.

Two commonly used statistics exist to estimate whether there is a significant difference between the expected (normal) distribution and the experimental distribution of the residuals: CVM and KS. The CVM statistic is the integral of the squared difference between both distribution functions, meaning the area between both curves as seen in Figure 1b [22]. The KS statistic is the maximal vertical distance observed between both distribution functions, as shown in Figure 1c [23].

Once the test statistic is obtained for the experimental residuals data set, a P-value needs to be calculated. It represents the probability of obtaining a statistic of that value when the residuals are indeed normally distributed. The distribution of the CVM and KS statistics is governed by complex laws. The best approach is therefore to estimate those distributions using a bootstrap approach [24]. Briefly, this numerical method synthesizes large numbers of residuals data sets by employing resampling of the experimental residuals [25]. Then, y_{ij} values associated with these synthetic residuals sets are utilized to determine parameters b_0, b_1 and b_2, which in turn permit the calculation of residuals.

Since these residuals are derived from bootstrapped data, they are pseudo-residuals. The CVM or KS statistic from each set of pseudo-residuals is then calculated. Inclusion of a correction factor is necessary when calculating the results of the CVM or KS laws from pseudo-residuals [26]. Using a large ensemble (here, 1000) of the derived CVM or KS statistic allows the probability of observing the experimental statistic to be determined.

3. Results and discussion

3.1. Raw data necessary and replicate analysis

Validation guidelines from the Scientific Working Group in Toxicology (SWGTOX) state that "at least six different non-zero concentrations" should be used to choose the calibration model, but that additional calibration levels may be required to characterize properly higher order models [2]. This is a generally agreed upon standard [1, 4, 27].

Regarding the number of replicate measurements necessary for an adequate calibration model selection and validation process, the SWGTOX suggests that "a minimum of five replicates per concentration is required" [2]. Other authors suggest increasing the number of replicates up to nine [6]. In the opinion of the authors, in the context of method validation in toxicology, seven replicates is a good compromise between adequate statistical test performance and the amount of work that must be done in the laboratory (see results for the simulated data in the first paper). It is important to understand that with larger %RSD, more replicate measurements are needed before the estimation from the data converges to the real variance value.

The SWGTOX guidelines suggest that an experimental setup using inter-day and inter-extraction replicate analysis should be used [2]. However, using such a setup results in an increased variability in the data compared with an intra-day and/or intraextraction setup. The less robust the method is, the more this is true. Good analytical methods aim to curb this inter-day and interextraction variability by instrument maintenance and suitability tests as well as the use of stable isotope-labeled internal standards. Despite all of these measures, it is expected that there will always be some inter-day and inter-extraction variability left. The problem with this increased variability is that, when it is included in the calibration, it can mask patterns in the data, such as quadraticity. This will result in a higher mis-selection rate for the order of the model. It is important to point out that this is not a failure of this particular algorithm for the selection of the calibration model. Masking of the patterns occurs in the data set itself, and any calibration model selection scheme will suffer from this.

The purpose of any calibration model selection scheme is to find out the model underlying the data generated in production. Therefore, the authors suggest that calibration model selection should be performed on data obtained with an experimental setup that represents what will occur during that production process. If the method states that two series of standards are to be extracted and analyzed with each batch to create the calibration curve (intra-day, inter-extraction), then this is the data that should be used for selection of the calibration model during the validation process. More often than not, the method calls for one series of standards to be extracted and analyzed (sometimes in duplicate) with each batch to create the calibration curves. This was the case with the LC-MS/MS method presented here. In this situation, intra-day, intra-extraction data should be used to select the calibration model.

In our algorithm, pattern masking through increased variability creates problems on two fronts: the test for model order selection and the test for validation of the chosen model. Both showed a lower performance due to increased variability introduced by the inter-day/ inter-extraction setup.

Using simulated quadratic data sets, we observed that increased variability can result in a masking of the quadratic nature of the data. A data set that is, in fact, quadratic can therefore be identified as linear when performing the partial F-test for selecting model order. The experimental design combining inter-day/inter-extraction data can therefore incorporate systematic variability, which is then treated mathematically as random error during regression. Therefore, any improvement afforded by a quadratic fit can be obscured by the artificially large variability of the measurements and mask the underlying nature of the data. An increased variability will have no impact when the instrument's pattern of response is linear, since the null hypothesis of the test is that the model is linear. These predictions are demonstrated by the experimental results obtained. When intra-day, intra-extraction data were used, the partial F-test concluded that a quadratic model should be used for 47 out of 50 analytes (Supplemental Data 1). On the other hand, this number dropped to 40, 43 and 41 out of 50 when intra-day/inter-extraction, inter-day/intra-extraction and inter-day/inter-extraction data are used, respectively. This phenomenon was confirmed using simulated data based on experimentally obtained calibration curves (see Supplemental Data 1 of the first paper). These results show that as the variability (%RSD) increases in quadratic models with greater variability at the high end of the curve ($^1/_x$ or $^1/_{x^2}$ weighting), the success rate of the partial F-test for order selection drops when the curvature is weak, as expected.

Similarly, normality testing of the residuals can have a hard time pinpointing the inadequacy of a model when the apparent variability of the measurements is increased by using inter-day/inter-extraction data. Departures from normality would have to be much greater to be considered significant. Therefore, a model that would be identified as incorrect using intra-day, intra-extraction data could justifiably be validated if an inter-day and/or interextraction data set is used, which would mean an inadequate model would be missed by the test.

That being said, we do need to consider sources of variability that might change the instrument's pattern of response over time if we want to select a robust model that can be used over a long time period. The way to do this is not to use inter-day/inter-extraction data, as one might first think. Rather, the procedure for selection and validation of a calibration model should be performed multiple times throughout the whole validation process (which will most likely span two or more months). A minimum of three times is recommended by the authors, taking care to change (as appropriate) analysts, instruments, standard lots, spiking matrix lots, etc. If all procedures choose the same calibration model, it is a robust model and should be kept for future analyses. In our experience, when results alternate between linear and quadratic model and there is no instrument stability issue, then it is because results of the partial F-test are near the threshold (P-value is near 0.05), and often the curvature of the quadratic equation is small (quadratic term is close to zero). In this situation, both models would give similar results in terms of accuracy and precision, and we recommend that the simplest model (linear) should be used.

To summarize, we recommend the use of intra-day/intra-batch data (if this is the calibration scheme that will be used in production) in order to obtain the most accurate selection of the calibration model as possible, and repetition of this selection process over three different days to take into account the inter-batch/inter-day variability. Using inter-batch and/or inter-day data to feed to our calibration model selection and validation scheme is possible, however, the user has to be aware that some tests will have a lower performance as a result.

3.2. Forcing a calibration equation through the origin

Several data analysis packages offer the option of forcing the calibration function through the origin. In principle, the reasoning behind this option is that the signal is expected to be zero when the concentration is zero. However, the SWGTOX states in their validation guidelines that "the origin shall not be included as a calibration point" [2] which we agree is the correct approach. Forcing the calibration function through the origin means creating data that have not been measured. Although theoretically, the calibration function is expected to go through the origin, there can be several valid reasons why it may not do so experimentally. These reasons may vary from technique to technique but would include blank contamination, undetected non-linear behavior near the blank, insufficient background/baseline removal etc. Additionally, it is in conflict with the other statistical approaches used to choose and validate a calibration model, since it artificially modifies the error at the low end of the calibration curve and alters the calibration coefficients. Due to these mathematical concerns, the approach of forcing a calibration function through the origin was not used and is not recommended.

3.3. Heteroscedasticity testing

With heteroscedastic data, the observed variance and the interrelated precision of the measurement change across the concentration range. In standard least-squares regression, the "best fit" occurs when the sum of the squared errors between the data and the calibration function is minimized. In a heteroscedastic situation, the data with the largest variance will make the largest contribution to the sum of the squared errors and dominate selection of the coefficients - which is opposite to what is appropriate. Reliability of the parameter estimates will be reduced [15]. This problem can be largely "overcome by introducing weighting factors inversely proportional to the variance" [15]. This scaling increases the contribution to the squared error of the high precision data and reduces the contribution of the low precision data when establishing the calibration coefficients.

The presence of heteroscedasticity can be tested in different ways via the F-test, Cochran, Hartley and Bartlett tests [6]. The F-test was chosen for its simplicity of application. It compares the variance of two groups of data, in this application the LLOQ and ULOQ [15].

In many analytical chemistry experimental systems, there is a sound theoretical basis for heteroscedastic data [19, 28, 29], which can mostly be simplified to two primary types of noise (variance): additive (constant) and multiplicative (scales with concentration). Methods where multiplicative noise is the dominant source of error will produce heteroscedastic data [29]. The expectation that the variance increases with concentration when using MS detection dictates that a unilateral F-test should be applied (alternative hypothesis being that the variance at the ULOQ level is greater than the variance at the LLOQ level).

Two situations can create false negatives when the F-test is applied this way. In the first case, the heteroscedasticity pattern is more exotic than what is expected in the simple multiplicative noise case and the variance is not significantly different at the ULOQ and the LLOQ but higher or lower in the center of the calibration range. This is extremely atypical of toxicology methodologies, and standard data analysis software should not be used in this situation since there is no appropriate weighting scheme. In any event, a quick examination of the variance plot, produced as a PDF document when the R script is run, should ensure that this situation does not go unnoticed.

The second situation that might produce a false negative would be the presence of an outlier at the LLOQ. This can artificially increase the observed variance at the LLOQ and mask the heteroscedasticity. Analysts should always be attentive to data points vastly different (4σ or more away from the rest of the group), and understand that inclusion will affect their results. Of course, if the selection and validation of calibration model analysis is repeated more than once during the validation process, then the probability that this issue would occur each time is extremely small.

All of the 50 analytes studied in the LC-MS/MS method tested positive for heteroscedasticity, with P-values from 2.2×10^{-10} to 2.2×10^{-4} (Supplemental Data 1). When large dynamic ranges are analyzed on LC-MS/MS, this type of heteroscedasticity is quite common [1, 30, 31, 32].

3.4. Selection of weight

Once a data set had been shown to be heteroscedastic, the next step was to select an appropriate weighting factor. To correct completely for the heteroscedasticity issue, the inverse of the variance at each concentration level should be used as the weighting factor [15]. However, the variance typically follows set patterns across the calibration range: it increases in proportion to the concentration (x) or with the square of the concentration (x^2) [5, 10]. As these are the most common variance patterns, they are included as weighting options in almost all data analysis software. The use of the experimentally measured variance pattern (that might even vary over time) does not appear a good choice.

The first test option evaluated was visual examination of the variance plot by the analyst. This procedure worked well, but it was a manual, analyst-dependent process, which was exactly what we were trying to avoid with this new protocol. When analyzing data for the 50 analytes, the analyst often ended-up questioning his or her own weighting choice, especially hesitating between $^1/_x$ and $^1/_{x^2}$ when outliers disturbed the trend in variance.

The second option considered was variance evaluation, where the weighting model with the smallest spread of normalized weighted variances was chosen. This test was automated and the result was analyst independent. Under this test, all 50 analytes were deemed to have $^1/_{x^2}$ (Supplemental Data 1). This result was expected given that samples were analyzed on an LC-MS/MS with calibrations spanning a few orders of magnitude [7, 10]. It is, however, interesting to note that removing a few upper calibration levels would obscure the power relationship and result in an apparently linear variance plot, therefore changing the weighting factor. Using the simulated data for different weighting factors gave a better idea of the performance of this test. The outcome, described in the "Results and discussion" section of the first paper, showed good test performance (2% failure) with 10 measurement replicates.

3.5. Selection of model order

In the SWGTOX guidelines for method validation, selection of model order is merged with validation of the model and tested with the examination of the standardized residuals plot [2]. However, it also stated that "there are other appropriate alternatives to evaluate calibration models (i.e., ANOVA-LOF test for unweighted linear models, checking for the significance of the second-order term in quadratic models, assessment of coefficient of determination for linear models)" [2]. We examined three different options for model order selection: ANOVALOF, significance of the second-order term and the partial F-test. Visual examination of the standardized residuals plot was not considered due to its manual, analyst-dependent nature.

The most common form of ANOVA-LOF test is intended for unweighted linear models, as stated by the SWGTOX [2]. However, in the version presented here, the calculations have been adapted to assess weighted models of n^{th} order [18]. In this test, as with the standardized residuals, selection of model order is somewhat merged with model validation. The result is the selection of the model order based on comparison of two P-values. From a statistical point of view, this comparison is a precarious procedure. Statistical tests should be treated as having a binary outcome (acceptance or rejection of the null hypothesis). Moreover, the P-values are approximations and are therefore not exact values - especially in the extremes. For these reasons, comparing P-values that are nearly identical and/or are both at extreme values is not without risk. Nevertheless, when this procedure was applied to select model order, all but two analytes (naltrexone and phenylpropanolamine) were identified as quadratic models (Supplemental Data 1).

Calculations for the significance of the second-order term were not performed. The first thing to highlight is that calculations for this test are more involved than one might think [18]. An analyst might be tempted to establish confidence intervals of ($\pm^{t \times s}/_{n}$) around the five b_2 values resultant from the five regressions, but this is a faulty strategy. Rather, calculation of the significance of the second-order term uses the variance-covariance matrix of the regression coefficients. Unfortunately, this test relies heavily on an assumption of normality, which is precarious. With the limited data sets obtained for selection and validation of the calibration model, it is probable that normality has not kicked off yet. A bootstrap approach could be used to circumvent this issue, but would require significantly more involved calculations. Taking into account all of these considerations, we have decided not to perform calculations for this test of model order selection and to select another test in the final procedure.

The partial F-test is conceptually and mathematically simple, more so than the ANOVA-LOF and significance of the second-order term tests. Three analytes out of the fifty were classified as linear using the partial F-test: hydroxy-alprazolam, naltrexone and phenylpropanolamine (Supplemental Data 1). This number was higher than with the ANOVA-LOF test because the partial F-test is more stringent against overfitting. Another advantage of this test was that it clearly separated model order selection from model validation, leading to a reduced degree of confusion. Given its several advantages over the other available tests, we decided to retain the partial F-test for model order selection.

3.6. Tests for validation of the chosen model

The main model validation metric test suggested by the SWGTOX is ANOVA-LOF [2]. This test was applied to the 50 analytes of the LC-MS/MS method (Supplemental Data 1). When using intra-day/intra-extraction data, ANOVA-LOF systematically rejected the chosen calibration model. Given the fact that the accuracies and precisions were acceptable for all analytes, this was, on the face of it, a surprising result. It turns out this test is very sensitive to experimental design, in particular the number of replicates and/or the number of calibration levels. As a result, it is difficult to produce a P-value above 0.05 given the typical variances observed in analytical methodologies. The excessive sensitivity of this test limits its practical applicability.

On the other hand, residual normality testing with either the KS or CVM test accepted the chosen calibration model for all 50 analytes. Between 0 and 2% of the simulated data sets saw their chosen calibration model rejected (see Table 2 of first paper). Simulated non-normal data sets (not shown) were also rejected. Therefore, this test can detect most problems with the calibration model, but is more robust than ANOVA-LOF.

Residuals normality testing is based on the same underlying principle as the examination of the standardized residuals graph, which qualitatively verifies the randomness of residuals. Again, the major issue with this visual technique is the lack of automation and the analyst dependency of the result. The problem is that the visual differences between a residuals graph with a distribution deemed normal by the CVM test and a residuals graph with a non-normal distribution can be very slim indeed (see Supplemental Data

2). Knowing the results of the normality tests, the analyst can, a posteriori, find reasons for why the calibration model for Analyte A was rejected and not for Analyte B (Supplemental Data 2), but these examples clearly emphasize the lack of rigor of the visual examination of standardized residuals plot and highlight the potential for analyst bias.

Residuals normality testing by KS or CVM was therefore selected for its bias-free, binary output which is a great advantage over subjective examination of the residuals graph. The drawback to this test is the complexity of the calculations involved, but the tool we have created with RStudio makes implementation of this test fairly easy.

4. Conclusions

The impact of different types of replicate analysis were evaluated. The use of inter-day/inter-extraction validation schemes was found to introduce inappropriate variability into the data set, which masked some underlying trends (quadraticity, departure from normality). It is therefore suggested that validation data should mimic what will be done in production (i.e., use intra-day/intra-extraction data where daily calibration with one set of standards is to be used in production). Ideally, the process of selection and validation of the calibration model will be repeated at multiple points in the global validation process.

Forcing the calibration function through the origin was not applied or tested due to its artificial alteration of the error and calibration coefficients.

Heteroscedasticity was evaluated through the unidirectional F-test applied using the LLOQ and the ULOQ calibration levels. Increasing variance with concentration was the overwhelmingly observed pattern and is encountered with many instrumental analytical methods, which justifies the general use of weighted regression methods.

Selection of the weight was found to be best performed through a variance evaluation. Examination of the variance graph was rejected because of analyst dependency.

The partial F-test was chosen to perform the selection of model order for its conceptual and mathematical simplicity and validity. ANOVA-LOF and significance of the second-order term were examined, but rejected because of their strong dependence on normality of the data, utilization of the dubious P-value comparison procedure and calculation complexity.

Residual normality testing was chosen as the calibration model validation procedure because of its robustness. ANOVA-LOF, the main alternative metric, is too sensitive to experimental design (replicates, calibration levels) to be truly useful. On the other hand, examination of the standardized residuals graph is not a bias-free approach.

Through examination of different testing alternatives, we have created a selection and validation procedure for calibration models that can be used in the overall validation of all quantitative methods in toxicology. This stepwise, bias-free method is a simple automated tool toward better calibration model selection.

5. Funding

Brigitte Desharnais, Félix Camirand-Lemyre and Cameron D. Skinner gratefully acknowledge support of the National Sciences and Engineering Research Council of Canada. Brigitte Desharnais also gratefully acknowledges the support of the Fonds de recherche du Québec - Nature et technologies.

6. Acknowledgements

The authors wish to thank Cynthia Côté for her thorough review and comments as well as Sue-Lan Pearring and Chris H. House for their opinion on key matters. The authors are also grateful to Gabrielle Daigneault, Lucie Vaillancourt, Julie Laquerre and Marc-André Morel for their technical work.

References

[1] Peters, Frank T and Drummer, Olaf H and Musshoff, Frank, Validation of new methods, Forensic Science International 165 (2007) 216–224.

[2] Scientific Working Group for Forensic Toxicology, Scientific Working Group for Forensic Toxicology (SWGTOX) Standard Practices for Method Validation in Forensic Toxicology, Journal of Analytical Toxicology 37 (2013) 452–474.

[3] Shah, Vinod P and Midha, Kamal K and Findlay, John WA and Hill, Howard M and Hulse, James D and McGilveray, Iain J and McKay, Gordon and Miller, Krys J and Patnaik, Rabindra N and Powell, Mark L and others, Bioanalytical method validation - a revisit with a decade of progress, Pharmaceutical Research 17 (2000) 1551–1557.

[4] F. T. Peters, Method validation using LC-MS, in: A. Polettini (Ed.), Applications of LC-MS in Toxicology, Pharmaceutical Press, London, United Kingdom, 2006, pp. 71–96.

[5] Hartmann, C and Smeyers-Verbeke, J and Massart, DL and McDowall, RD, Validation of bioanalytical chromatographic methods, Journal of Pharmaceutical and Biomedical Analysis 17 (1998) 193–218.

[6] Penninckx, W and Hartmann, C and Massart, DL and Smeyers-Verbeke, J, Validation of the calibration procedure in atomic absorption spectrometric methods, Journal of Analytical Atomic Spectrometry 11 (1996) 237–246.

[7] H. Gu, G. Liu, J. Wang, A.-F. Aubry, M. E. Arnold, Selecting the correct weighting factors for linear and quadratic calibration curves with least-squares regression algorithm in bioanalytical LC-MS/MS assays and impacts of using incorrect weighting factors on curve stability, data quality, and assay performance, Analytical Chemistry 86 (2014) 8959–8966.

[8] Burrows, John and Watson, Kenneth, Linearity of chromatographic systems in drug analysis part I: theory of nonlinearity and quantification of curvature, Bioanalysis 7 (2015) 1731–1743.

[9] Pagliano, Enea and Mester, Zoltán and Meija, Juris, Calibration graphs in isotope dilution mass spectrometry, Analytica Chimica Acta 896 (2015) 63–67.

[10] Hubert, Ph and Chiap, Patrice and Crommen, Jacques and Boulanger, Bruno and Chapuzet, E and Mercier, N and Bervoas-Martin, S and Chevalier, P and Grandjean, D and Lagorce, Ph and others, The SFSTP guide on the validation of chromatographic methods for drug bioanalysis: from the Washington Conference to the laboratory, Analytica Chimica Acta 391 (1999) 135–148.

[11] Moore, Christine and Rana, Sumandeep and Coulter, Cynthia, Determination of meperidine, tramadol and oxycodone in human oral fluid using solid phase extraction and gas chromatography–mass spectrometry, Journal of Chromatography B 850 (2007) 370–375.

[12] Cociglio, M and Peyriere, H and Hillaire-Buys, D and Alric, R, Application of a standardized coextractive cleanup procedure to routine high-performance liquid chromatography assays of teicoplanin and ganciclovir in plasma, Journal of Chromatography B: Biomedical Sciences and Applications 705 (1998) 79–85.

[13] Gupta, Anubha and Jansson, Britt and Chatelain, Pierre and Massingham, Roy and Hammarlund-Udenaes, Margareta, Quantitative determination of cetirizine enantiomers in guinea pig plasma, brain tissue and microdialysis samples using liquid chromatography/tandem mass spectrometry, Rapid Communications in Mass Spectrometry 19 (2005) 1749–1757.

[14] Peters, Frank T. and Maurer, Hans H., Bioanalytical method validation and its implications for forensic and clinical toxicology — A review, in: De Bièvre, Paul and Günzler, Helmut (Ed.), Validation in Chemical Measurement, Springer, Berlin, Germany, 2005, pp. 1–9.

[15] D.L. Massart, B.G.M. Vandeginste, L.M.C Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Straight Line Regression and Calibration, in: Handbook of Chemometrics and Qualimetrics: Part A, volume 20A of *Data Handling in Science and Technology*, Elsevier, Amsterdam, Netherlands, 1997, pp. 171–230.

[16] Rozet, Eric and Ceccato, Attilio and Hubert, Cédric and Ziemons, Eric and Oprean, Radu and Rudaz, Serge and Boulanger, Bruno and Hubert, Philippe, Analysis of recent pharmaceutical regulatory documents on analytical method validation, Journal of Chromatography A 1158 (2007) 111–125.

[17] D.L. Massart, B.G.M. Vandeginste, L.M.C Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Some important hypothesis tests, in: Handbook of Chemometrics and Qualimetrics: Part A, volume 20A of *Data Handling in Science and Technology*, Elsevier, Amsterdam, Netherlands, 1997, pp. 93–120.

[18] D.L. Massart, B.G.M. Vandeginste, L.M.C Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Multiple and Polynomial Regression, in: Handbook of Chemometrics and Qualimetrics: Part A, volume 20A of *Data Handling in Science and Technology*, Elsevier, Amsterdam, Netherlands, 1997, pp. 263–303.

[19] Karnes, H Thomas and Shiu, Gerald and Shah, Vinod P, Validation of bioanalytical methods, Pharmaceutical Research 8 (1991) 421–426.

[20] Cook, R. D. and Weisberg, S., Diagnostic methods using residuals, in: Residuals and influence in regression, Chapman and Hall, New York, United States of America, 1982, pp. 10–100.

[21] González, A Gustavo and Herrador, M Ángeles, A practical guide to analytical method validation, including measurement uncertainty and accuracy profiles, Trends in Analytical Chemistry 26 (2007) 227–238.

[22] Darling, Donald A, The kolmogorov-smirnov, cramer-von mises tests, The Annals of Mathematical Statistics 28 (1957) 823–838.

[23] Miller, James and Miller, Jane C, Statistics of repeated measurements, in: Statistics and chemometrics for analytical chemistry, Pearson Education, Harlow, England, 6 edition, 2010, pp. –37–73.

[24] van der Vaart, Aad W. and Wellner, Jon A., The Bootstrap, in: Weak Convergence and Empirical Processes: With Applications to Statistics, Springer, New York, United States of America, 1996, pp. 345–359.

[25] Desharnais, Brigitte and Camirand-Lemyre, Félix and Mireault, Pascal and Skinner, Cameron D, Determination of confidence intervals in non-normal data: application of the bootstrap to cocaine concentration in femoral blood, Journal of Analytical Toxicology 39 (2015) 113–117.

[26] Wellner, Jon A. Wellner and van der Vaart, Aad W., Empirical processes indexed by estimated functions, in: Asymptotics: Particles, Processes and Inverse Problems, volume 55 of *Lecture Notes - Monograph Series*, Institute of Mathematical Statistics, 2007, pp. 234–252.

[27] Zhao, Yue and Liu, Guowen and Shen, Jim X and Aubry, Anne-Francoise, Reasons for calibration standard curve slope variation in LC–MS assays and how to address it, Bioanalysis 6 (2014) 1439–1443.

[28] Hubert, Ph and Nguyen-Huu, J-J and Boulanger, Bruno and Chapuzet, E and Cohen, N and Compagnon, P-A and Dewé, Walthère and Feinberg, M and Laurentie, Michel and Mercier, N and others, Harmonization of strategies for the validation of quantitative analytical procedures: A SFSTP proposal–Part III, Journal of Pharmaceutical and Biomedical Analysis 45 (2007) 82–96.

[29] Ingle, James D. and Crouch, Stanley R., Signal-to-Noise Ratio Considerations, in: Spectrochemical Analysis, Prentice Hall, Englewood Cliffs, United States of America, 1988, pp. 135–163.

[30] Lanckmans, Katrien and Clinckers, Ralph and Van Eeckhaut, Ann and Sarre, Sophie and Smolders, Ilse and Michotte, Yvette, Use of microbore LC–MS/MS for the quantification of oxcarbazepine and its active metabolite in rat brain microdialysis samples, Journal of Chromatography B 831 (2006) 205–212.

[31] Apostolou, Constantinos and Dotsikas, Yannis and Kousoulos, Constantinos and Loukas, Yannis L, Development and validation of an improved high-throughput method for the determination of anastrozole in human plasma by LC–MS/MS and atmospheric pressure chemical ionization, Journal of Pharmaceutical and Biomedical Analysis 48 (2008) 853–859.

[32] Nilsson, Lars B and Eklund, Göran, Direct quantification in bioanalytical LC–MS/MS using internal calibration via analyte/stable isotope ratio, Journal of Pharmaceutical and Biomedical Analysis 43 (2007) 1094–1099.