Received 5 March 2020 | Revised 11 May 2020 | Accepted 27 May 2020 | Published 31 May 2020

DOI: 10.1002/dta.2867

DRUG TESTING AND ANALYSIS POST-PRINT

Qualitative threshold method validation and uncertainty evaluation: A theoretical framework and application to a 40 analytes liquid chromatography–tandem mass spectrometry method

Félix Camirand Lemyre[1,2,3] | Brigitte Desharnais*[4,5] | Julie Laquerre[4] | Marc-André Morel[6] | Cynthia Côté[4] | Pascal Mireault[4] | Cameron D. Skinner[5]

[1]Department of Mathematics, Université de Sherbrooke, 2500 Université Boulevard, Sherbrooke, Québec, J1K 2R1, Canada
[2]School of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria, 3010, Australia
[3]Centre de recherche, Centre hospitalier universitaire de Sherbrooke, 12th Avenue North, Sherbrooke, Québec, J1H 5N4, Canada
[4]Department of Toxicology, Laboratoire de sciences judiciaires et de médecine légale, 1701 Parthenais Street, Montréal, Québec, H2K 3S7, Canada
[5]Department of Chemistry & Biochemistry, Concordia University, 7141 Sherbrooke Street West, Montréal, Québec, H4B 1R6, Canada
[6]Department of Criminalistics, Laboratoire de sciences judiciaires et de médecine légale, 1701 Parthenais Street, Montréal, Québec, H2K 3S7, Canada

Correspondence
*Brigitte Desharnais, 1701 Parthenais Street, Montréal, Québec, Canada, H2K 3S7.
Email: brigitte.desharnais@msp.gouv.qc.ca
F. Camirand Lemyre and B. Desharnais contributed equally to this study and are named in alphabetical order.

Abstract

Qualitative methods hold an important place in drug testing, filling central needs in screening and analyses, among others, linked to per se legislation. Nevertheless, the bioanalytical method validation guidelines do not discuss this type of method or describe method validation procedures ill-adapted to qualitative methods. The output of qualitative methods are typically categorical, binary results, such as presence/absence or above cut-off/below cut-off. As the goal of any method validation is to demonstrate fitness for use under production conditions, qualitative validation guidelines should evaluate performance based on discrete, binary results instead of the continuous measurements obtained from the instrument (e.g. area).

A tentative validation guideline for threshold qualitative methods was developed by in silico modelling of measurements and derived binary results. This preliminary guideline was applied to a liquid chromatography–tandem mass spectrometry method for 40 analytes, each with a defined threshold concentration. Validation parameters calculated from the analysis of 30 samples spiked above and below the threshold concentration (false negative rate, false positive rate, selectivity rate, sensitivity rate and reliability rate) showed a surprisingly high failure rate. Overall, 13 out of the 40 analytes were not considered validated. A subsequent examination found that this was attributable to an appreciable shift in the standard deviation of the area ratio on a day-to-day basis, a previously undescribed and unaccounted-for behaviour in the qualitative threshold method validation literature. Consequently, the developed guideline was modified and used to validate a qualitative threshold method, based on the binary results for performance evaluation and incorporating measurement uncertainty.

KEYWORDS:
method validation, qualitative, threshold, uncertainty of measurement

1 | INTRODUCTION

Qualitative methods are best described by comparing them with quantitative methods. The latter produce concentration estimates on a continuous scale, whereas the output of the former is categorical (or discrete), typically binary. Qualitative identification methods yield a "present" or "absent" result for each analyte. Most screening methods, e.g., time of flight or gas chromatography coupled with mass spectrometry with spectral databases, fall under this category. This paper focuses on qualitative threshold (or decision point, or cut-off) methods, in which a concentration threshold is set, and the analytes are reported as being above or below this threshold.

Qualitative threshold methods, although often brushed aside in the forensic toxicology setting, can be relevant. They can be used when a legislative threshold for an analyte is present, e.g., in the case of a per se limit for driving under the influence (DUID) cases. Some legislations require a quantitative evaluation of said analyte, complete with the uncertainty of measurement evaluation, whereas otherlegislations make no such requirement (e.g. Canada). A qualitative threshold method could then be used, insofar as the uncertainty of this method is properly dealt with. Qualitative threshold methods are also paramount in laboratories running a given qualitative method on multiple instruments, e.g. a laboratory running a method on four distinct liquid chromatography–tandem mass spectrometry (LC–MS/MS) systems of the same brand and model. Each LC–MS/MS system has its specific sensitivity and limit of detection for a given analyte. In such a situation, using a qualitative identification method set-up could result in, e.g., a urine DUID sample being called positive on the first instrument but negative on the second. To achieve fairness between the cases treated, a concentration threshold common to all instruments must be instated in this situation.

Many literature studies deal specifically with quantitative methods and their validation procedures[1,2,3,4,5,6,7,8,9], but only a handful of literature studies and guidelines deal with qualitative methods.

Both SWGTOX[1] and AAFS Standards Board[10] method validation guidelines contain recommendations for qualitative threshold (or decision point) methods. According to these guidelines, LC–MS/MS qualitative threshold method validation should include interference and carry-over studies, dilution integrity and stability if necessary, as well as precision evaluation. Recommendations with regard to the decision point are explicitly addressed only for immunoassays and state that the precision of the measured signal should be evaluated at $\geq$ 50% of the decision point or threshold concentration, at the threshold concentration and at $\leq$ 150% of the threshold concentration. The method is considered validated if the relative standard deviation (%RSD) $\leq$ 20% and $\left(\bar{x}_{50\%} + 2s_{50\%}\right) < \bar{x}_{DP} < \left(\bar{x}_{150\%} - 2s_{150\%}\right)$, that is, the mean $\pm$2 standard deviations at 50% and 150% of the decision point do not overlap with the mean measurement at the threshold concentration.

Two potential weak points can be identified. First, these procedures fail to use the categorical or binary nature of qualitative methods' output, employing instead procedures derived from quantitative method validation which relies on continuous data. One reason likely explaining this state of affairs is the confusion induced by the aspect of the procedure which transforms continuous measurements (area, height, area ratio, luminescence, etc.) into binary results. Moreover, quantitative method validation guidelines are so well developed that they are almost second nature to forensic toxicologists and bioanalysts. It therefore feels natural and safe to fall back on them for the related but distinct problem of qualitative method validation.

A second weak point is the absence of a clear framework for the evaluation of the method's uncertainty of measurement. With the enactment of the most recent "Requirements for the competence of testing and calibration laboratories" (ISO 17025:2017[11] standard), a number of accrediting bodies now require uncertainty of measurement evaluation in all methods[12,13], including qualitative threshold methods, regardless of whether this result appears directly in the final report.

In any method validation, the goal is to demonstrate the quality of the analytical method by producing objective proof that pre-defined performance criteria are met[1,10]. Importantly, this verification of the fitness for use should occur under the same preparation, analysis, and data processing procedures which will be used for the analysis (production)[1,10]. The same holds true for qualitative method validation. Accordingly, binary results (presence/absence, above/below cut-off) yielded by the method should be used to measure its performance because this is the result ultimately achieved in a production setting.

If the binary output of the qualitative methods is to be used, what are the appropriate validation guidelines and the associated minimal performance thresholds, and how should the uncertainty of measurement be evaluated and reported in the final results?

To answer such questions, the behaviour of the response variable (e.g., area ratio, luminescence) in relation to the encoded, binary outcome must be understood. This subject is touched upon sparingly in the literature[14,15,16], where diverse validation procedures are suggested.

In this paper, we use these various sources, computer simulations, and experimental data to study the behaviour of the binary above/below cut-off output of qualitative threshold methods, and put forward a tentative validation guideline. This guideline is similar to the performance evaluation of another kind of categorical test: medical tests for the presence of a diseased state[17,18]. This validation process is then evaluated by using an LC–MS/MS method for 40 analytes relevant to forensic toxicology in blood. The results enabled the discovery of a previously unreported and unaccounted-for behaviour in LC–MS/MS qualitative threshold methods and guided modifications to the guidelines.

2 | MATERIALS AND METHODS

2.1 | Analytical method

The experimental data used for the prospective and confirmatory studies of qualitative threshold methods were derived from a high throughput whole blood LC–MS/MS analysis method for 40 qualitative analytes and 60 quantitative analytes. The quantitative analytes were validated separately[19] and will not be discussed in this paper. Certain active metabolites were analysed in the qualitative portion of the method rather than the quantitative one, either because some issues were encountered during quantitative method validation or because quantification was deemed to bring little contribution in light of the limited literature. Inactive metabolites required a qualitative threshold method due to the presence of multiple LC–MS/MS systems in the laboratory; as discussed in Section 1, in such a laboratory set-up, a threshold is required to ensure fairness for all cases.

For every qualitative analyte, a threshold concentration was selected based on analytical (sensitivity across multiple LC–MS/MS systems) and toxicological (relevant concentrations for effects) considerations. Data S1 includes the full list of substances and their designated cut-off, as well as detailed analytical parameters of the method, including internal standards and concentrations, MS transitions and parameters (Q1, Q3, retention time, DP, EP, CE and CXP). The method's chromatography, MS parameters and instrument models are also specified.

2.2 | Sample preparation

Samples were brought to room temperature over 1 h. Following vortex mixing for 10 s, 100 μL of blood was transferred using a positive displacement pipette into a 96 well-plate with 2 mL square wells (Fisher Scientific, AB-0932, Ottawa, Ontario, Canada).

In this study, blood samples were spiked at the threshold concentration, as well as the upper and lower unreliability limits (UURL and LURL, as described in Section 3). Post-mortem cardiac and femoral blood with negative screening results, as well as ante-mortem blood purchased from UTAK (Valencia, CA, USA), were used. All compounds used for spiking purposes were purchased from Cerilliant (Round Rock, TX, USA), except 3-hydroxy bromazepam and N-desmethyl diphenhydramine, which were purchased from Toronto Research Chemicals (North York, Ontario, Canada). A stable isotope-labelled internal standards (ISs, Cerilliant) solution, at concentrations indicated in Data S1, was added to the blood sample (10 μL) and vortex mixed. Due to pragmatic costs considerations, ISs used are not stable isotope-labelled analogues of the targeted analytes but those of the quantitative compounds analysed in the same method. Each qualitative compound was assigned the IS yielding the smallest %RSD for area ratio across different matrices.

To obtain a more finely granular precipitate, 100 μL of methanol: 0.2% formic acid in water (50:50, v/v) solution was mixed with the blood sample. Then, 400 μL of acetone:acetonitrile (30:70, v/v) mixture was used to precipitate the proteins. After

mixing, the plate was centrifuged at 3200×*g* for 5 min. A 25 μL aliquot of the supernatant was then transferred to a second 96 well-plate with 1 mL round-bottom wells (Canadian Life Science, RT96PPRWU1mL, Peterborough, Ontario, Canada). This extract was diluted with 180 μL of 0.2% formic acid in water and vortexed.

2.3 | LC-MS/MS analysis

A 5 μL aliquot of the diluted extract was separated on an Agilent Zorbax Eclipse Plus C18 column (2.1 × 100 mm, 3.5 μm) using a step/ramp gradient starting from 2:98 methanol:10 mM ammonium formate (pH 3.0) to 50% acetonitrile. The flow from the HPLC (Agilent 1200 or 1260 Infinity) was directed to a Sciex 5500 QTrap triple quadrupole mass spectrometer. Detailed analytical parameters for LC and MS acquisition are available in Data S1.

2.4 | Preliminary validation guidelines

Based on the qualitative threshold methods' literature and our exploratory experimental data analysis (described in Section 3.1), *in silico* simulations of measurement behaviour and derived binary results were carried out using R and RStudio. The R scripts for these simulations are available in Data S2, for both the standard model (described in the literature, Data S2 Section 1) and the model corrected for sampled threshold and heteroscedasticity (Data S2 Section 2). Based on the results of these simulations, the following preliminary validation guidelines were determined and applied.

First, the standard deviation was estimated by analysing a minimum of 10 different samples all spiked at the threshold concentration and by calculating the standard deviation of the response variable used – in this case the ratio of the analyte peak area to the IS peak area. At this stage, some laboratories might want to impose additional criteria, e.g., stating that all measured relative standard deviations (%RSD) should be $< 20\%$. Higher standard deviations can be perfectly acceptable, depending on the purpose of the method, as long as the associated uncertainty of measurement is properly reported. How uncertainty of measurement expresses itself in the final binary results will be discussed in detail later.

Probability curves, plotting the positivity rate (or above threshold rate) as a function of concentration, were built by spiking 10 or more samples at regular concentration intervals from −4 to +4 times the estimated sample standard deviation (s), e.g.$-4s$, $-3s$, $-2s$, $-1s$, cut-off, $+1s$, $+2s$, $+3s$ and $+4s$. This facultative step assumed a linear response of the area ratio as a function of concentration and a blank response of zero.

The core of the validation procedure consisted of analysing 30 different matrix lots spiked at $-3s$ and $+3s$ (upper and lower unreliability limits, UURL and LURL, as described in Section 3), which were used to calculate the method's validation parameters (see Section 2.5). Care was taken to ensure that these samples were prepared, injected and analysed as they would be in a production setting, including for this particular method two replicates of a sample spiked at cut-off, used to establish the threshold measurement and permit classification of samples as being above or below the threshold. Using 30 different matrix lots for this experiment was important, to include matrix effects in the evaluation of the method's performance.

Ion ratio dependability for identification purposes was estimated as the percentage of all samples for which the ion ratio fell within ±30% of the ion ratio measured in reference sample(s), as suggested by the European Commission Directorate General for Health and Food Safety[20].

2.5 | Calculation of validation parameters

The validation parameters [false negative rate (FNR), false positive rate (FPR), reliability rate (RLR), selectivity rate (SLR) and sensitivity rate (SNR)] are calculated as follows[14,15,17]:

$$FNR = \frac{FN}{FN + TP} \times 100 \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \times 100 \tag{2}$$

$$RLR = \frac{TP + TN}{n} \times 100 = 100 - FPR - FNR \tag{3}$$

$$SLR = \frac{TN}{TN + FP} \times 100 \tag{4}$$

$$SNR = \frac{TP}{TP + FN} \times 100 \tag{5}$$

where TN is the number of true negative results, TP is the number of true positive results, FN is the number of false negative results, FP is the number of false positive results, n is the total number of results.

Reliability (RLR) represents the method's overall ability to correctly identify the samples as above or below the threshold; sensitivity (SNR) evaluates the percentage of samples above the threshold that are indeed identified as such, and selectivity (SLR) measures the percentage of samples below the threshold that are indeed identified as such.

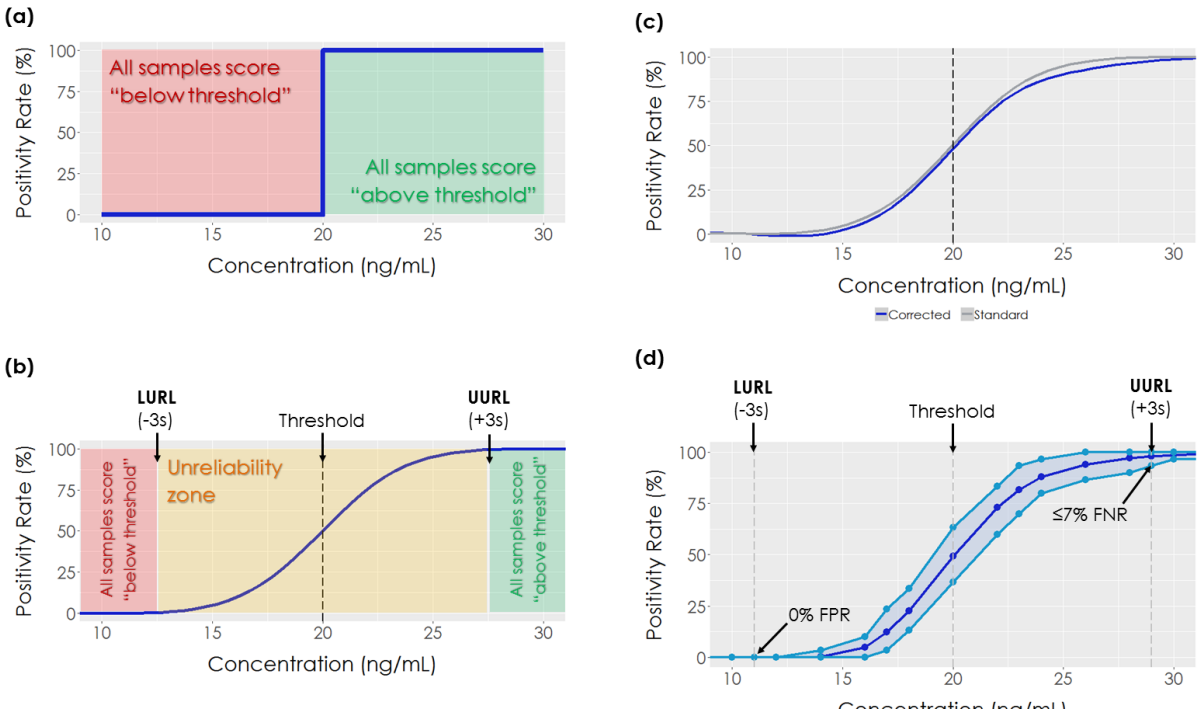
Expected performance levels (validation criteria) are set based on the in silico simulation of the binary results' behaviour (R script available in Data S2, Section 3). A validation experiment is simulated, e.g. 100 times, and the different rates (FNR, FPR, RLR, SLR, SNR) are calculated for each of these simulated experiments, yielding a distribution of probable validation outcomes for the parameters evaluated. The validation criteria are then set as follows: FPR $\leq$ 95% quantile at LURL, FNR $\leq$ 95% quantile at UURL, RLR $\geq$ 5% quantile at UURL, SLR $\geq$ 5% quantile at LURL, SNR $\geq$ 5% quantile at UURL. The expected performance levels under a different set-up (number of threshold samples and number of LURL/UURL samples) will vary, and laboratories should use the R script to define validation criteria under their own production and validation conditions. A qualitative decision point method validated under the production conditions described in this paper (two measured threshold samples, performance estimated over 30 samples) can be considered fit for the purpose if the observed FNR $\leq$ 7%, FPR = 0%, RLR $\geq$ 93%, SLR = 100% and SNR $\geq$ 93% (Figure 1 D), and ion ratio, stability, carry-over and interference studies are successful.

2.6 | Computer simulations

In preparation for the simulations, a set of 30 different samples spiked at the threshold concentration were extracted and analysed as one batch to determine their area ratios (analyte peak area divided by the IS peak area). Cramer–von Mises normality testing did not show significant departures from normality in all but 2 of the 40 analytes ($0.032 < P < 0.922$). The R script used to perform this analysis and the set of complete results are available in Data S3. This is in accordance with the implicit statement consensus in the literature that measurements (e.g. area and area ratios), including those made on an LC–MS/MS instrument, can be approximated by a normal distribution [14,15,16,21].

Based on these results, response values for simulations were modelled using RStudio's normally distributed random number generator `rnorm(n, mean, sd)`, where n is the number of measurements to be generated, mean is the known true value of the measurement and sd is the standard deviation. The number of measurements simulated per concentration level varied as needed between 1 and 100. The known true value of the measurement (area ratio) at threshold was set to vary between 0.008 and 1.050, based on the experimentally observed area ratios at the threshold concentration. A linear function describing the relationship between response and concentration was set as $y = b_1 x$, where b_1 was dictated by the known true concentration of the cut-off selected. Unless otherwise stated, a standard deviation equivalent to 15% of the cut-off response value was applied. The R scripts used to carry out these simulations are available in Data S2.

FIGURE 1 Positivity curves under different models. A, Idealized behaviour of qualitative threshold methods. B, Positivity curve for normally distributed measurements compared to a 20 ng/mL threshold. C, Corrected positivity curve, accounting for heteroscedasticity and sampled threshold. D, Average positivity rate, obtained via in silico simulations, when 30 spiked samples are measured and compared to a sampled threshold (two measurements to establish threshold). Ninety per cent of positivity rate results fall within the shaded area (5%–95% quantiles). The 5% (false positive rate, reliability rate and sensitivity rate) or 95% (false negative rate and selectivity rate) quantiles at the upper and lower unreliability limits ($-3s$ and $+3s$) are used to set the validation criteria



3 | RESULTS AND DISCUSSION

3.1 | Theoretical behaviour of binary results

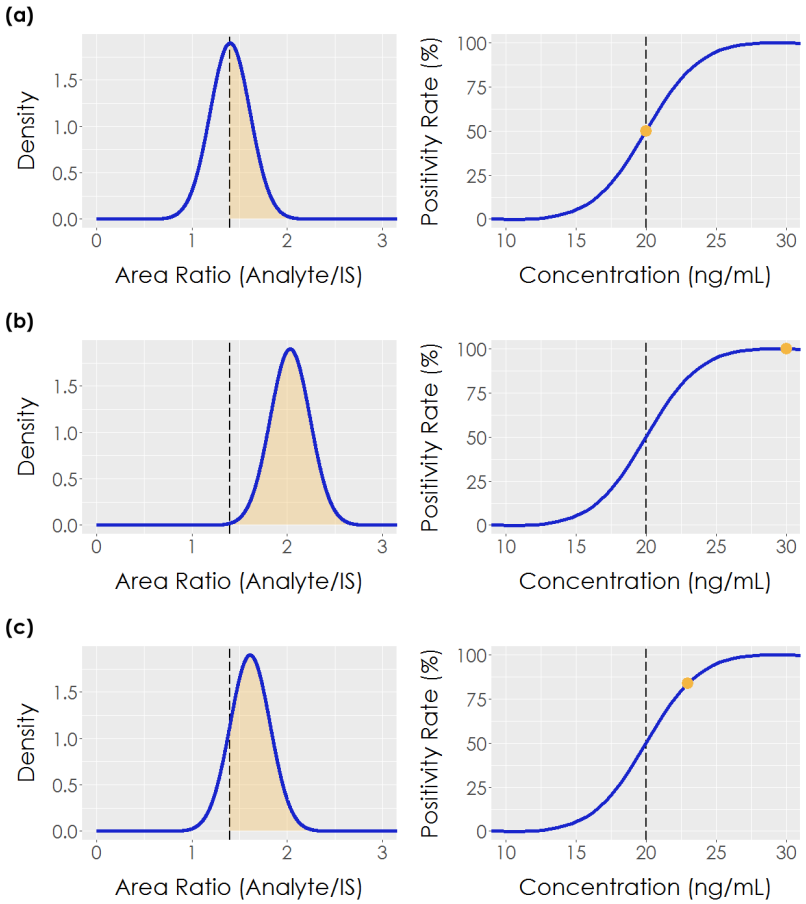
When thinking about decision point or threshold methods, the first reflex is often to assume (or hope) that results behave akin to what is displayed in Figure 1 A. Instinct dictates that all samples with a concentration below the threshold, or cut-off, will produce a low response and therefore score negative every single time they are analysed, and samples with a concentration higher than cut-off will similarly score positive (or above threshold) systematically.

Although this would be incredibly helpful, it is unfortunately impossible. A sample at a given concentration subjected to experimental manipulations and measured by a device that has some degree of imprecision will always produce a range of measured values, typically with a normal distribution. If this measurement error is ignored, important bias will ensue[22]. Thus, as shown in Figure 2 A, repeated measurements on a sample with a concentration exactly equal to the threshold concentration will yield a normal distribution with an average response equal to the threshold. Fifty per cent of these measurements will be reported as "below threshold" and 50% as "above threshold" (50% positivity rate).

If the sample analysed has a concentration far away enough from the threshold, that is, if the mean measurement for that concentration is 3σ above or below the threshold value (Figure 2 B), then almost all responses (> 99.7%) will be reported as "above threshold" or "below threshold" respectively.

Logically, at a point between these two extremes, the normal distribution of responses will overlap to varying degrees with the threshold response, generating an intermediate positivity rate (Figure 2 C). Samples with a concentration generating a mean

FIGURE 2 Normally distributed measurements in relation to a fixed threshold. A, Sample spiked at the threshold concentration (20 ng/mL): distribution of measurements (density plot, left) and positivity curve (right). Exactly 50% of the measurements are above the measurement at threshold, resulting in a 50% positivity rate. B, Sample spiked at $> 3\sigma$ above the threshold concentration (30 ng/mL): distribution of measurements (density plot, left) and positivity curve (right). The whole distribution is far from the threshold measurement, yielding a 100% positivity rate. C, Sample spiked at an intermediate concentration (between the threshold concentration and 3σ above the threshold concentration) (23 ng/mL): distribution of measurements (density plot, left) and positivity curve (right). The distribution of measurements overlaps with the threshold measurement, yielding an intermediate positivity rate (84%)



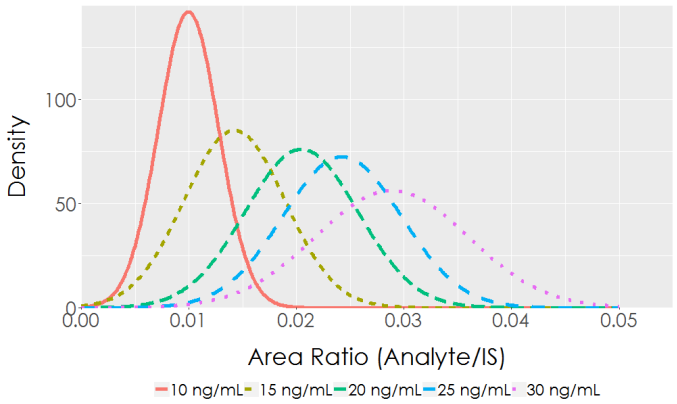
response between the cut-off and $+3\sigma$ above the cut-off will yield positivity rates between 50.0% and 99.7%. The converse also applies to samples with mean responses below the cut-off, with positivity rates decreasing as the concentration decreases.

The positivity curve for normal measurements compared to a threshold thus takes a sigmoidal form (Figure 1 B). How uncertainty of measurement is expressed in qualitative threshold methods is evident in this figure. Surrounding the cut-off is a range of concentrations where repeated measurement of the same sample will not always yield the same classification result: it will sometimes be "above threshold" and sometimes "below threshold". As a consequence, the positivity rate over repeated measurements for any sample with a concentration in this zone will take an intermediate value. This unreliability zone (UR), stemming from the uncertainty of measurement, is an ontological characteristic of qualitative threshold methods, and there is no possible way to avoid it. Although some might reflexively believe that moving the cut-off concentration could avoid this unreliability, it is important to understand that such a strategy is destined to fail, as the unreliability zone would just follow right along with it. In much the same way that one cannot avoid measurement uncertainty in quantitative methods, and estimation is necessary through dedicated calculation[13], the unreliability zone of qualitative decision point methods is here to stay and needs to be acknowledged, identified and estimated, not fought. The only viable strategy to minimize the magnitude of the

unreliability zone is to minimize the standard deviation of the overall analytical process, but although this might reduce the UR size, it will never eliminate it.

The positivity curve shown in Figure 1 B is the one typically reported in the literature for qualitative threshold methods. However, to adequately represent realistic (LC–MS/MS) data, it must be modified to take into account two important factors. First, these measurements are typically heteroscedastic, even over small concentration ranges. This is clearly shown by Figure ?, which displays normal distribution curves based on 30 spiked sample replicates at different concentrations for buprenorphine. The variance increases with increasing concentrations, as made evident by the decreasing distribution maxima. Any performance expectation must encompass this phenomenon.

FIGURE 3 Fitted normal distribution curves for buprenorphine samples ($n = 30$) spiked at different concentrations



Second, the standard positivity curve presumes that the response (measurement) at threshold is a known and fixed value. But, of course, this is not the case; in a production setting, this value is typically estimated based on a few measurements (usually 1–3) made on a sample spiked at the threshold concentration. This is a perfectly valid experimental procedure, which covers instrument performance and batch-to-batch variations. However, it does mean that the threshold value is sampled (estimated), not fixed (absolutely known), and thus an unknown error of variable size is attached to the estimated value. Inevitably, the estimated threshold will be more or less close to the real value from experiment to experiment, which has a domino effect on which samples get called "above" or "below" threshold, and thus on the positivity curve. Although this is not an issue in and of itself, the impact of this behaviour needs to be accounted for in modelling the results.

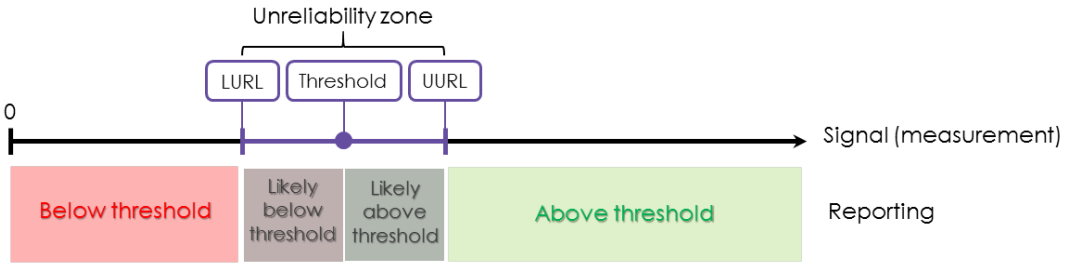
Fortunately, these measurement characteristics can be simulated and incorporated when establishing validation criteria; that is, their impact on the positivity curve and performance parameters can be calculated. Details of the modelling performed in RStudio can be found in Data S2 (Section 2). The resulting positivity curve is shown in Figure 1 C. Notable differences arise, particularly at the high concentration end, which affects the expected false negative rate and other parameters dependent on it (reliability and sensitivity rates).

This software tool, which models realistic behaviour of the measurements and derived binary results (encompassing heteroscedasticity and sampling error) should be used to establish appropriate validation criteria under the validation and production conditions used (number of injected threshold samples and number of UURL/LURL samples for the estimation of validation parameters). Figure 1 D illustrates the application of in silico modelling to establish validation criteria for the different parameters.

3.2 | Derived method validation guidelines

The derived method validation guidelines presented in Section 2.4 use the performance of the actual method's output, the binary "above threshold"/"below threshold" results. With these guidelines, we recognize the presence of result unreliability due to measurement uncertainty (measurement error). Consequently, the evaluation of the method's performance and validation needs to be performed outside the unreliability zone but provides the most pertinent figures of merit when performed near these boundaries (UURL and LURL). However, depending on the purpose of the method, laboratories might find it pertinent to precisely evaluate the size of the unreliability zone, or might be satisfied by performing validation well outside it through an overestimation of the UR size (e.g. ±50% of the threshold concentration). But ultimately, this measurement error must be acknowledged in the production setting as well. Method validation will confirm reliable performance for measurements below $-3s$ (LURL) and above $+3s$ (UURL), but what about measurements between these limits? These fall in the unreliability zone and must be identified and reported as such. Measurements between $-3s$ (LURL) and the measurement at threshold should be reported as "likely below threshold", and those between the measurement at threshold and $+3s$ (UURL) should be reported as "likely above threshold". This will reflect the fact that repeated measurements on these samples will not systematically yield the same outcome ("below/above threshold") and will adequately convey measurement uncertainty in the final analysis report (Figure 4).

FIGURE 4 Reporting of the measurement with respect to limit values (lower unreliability limit, threshold, upper unreliability limit)



Carry-over and interference studies should be carried out, as well as stability evaluation if deemed necessary. Although several procedures and guidelines exist, the OSAC Toxicology Subcommittee's practices are recommended in forensic toxicology[10]. If applicable, dilution integrity can be verified by repeating the main validation procedures (standard deviation estimation and evaluation of performance parameters on diluted samples). These studies were carried out for the method presented here but will not be discussed as they are not the main focus of this paper.

3.3 | Validation of the qualitative decision point LC-MS/MS method

The method validation guideline initially developed was used in an attempt to validate an LC–MS/MS qualitative decision point method for 40 analytes. Standard deviation estimation was performed based on the analysis of 10 samples spiked at the threshold concentration. Each analyte produced a unique standard deviation and therefore an UR zone of unique size. It follows, in principle, that the concentrations used for all subsequent validation steps should also be unique to each analyte, and each analyte would thus need to be spiked in matrices separately. For the probability curves alone, 40 analytes with unique UR size × 10 samples × 9 concentration levels = 3600 spiked samples would need to be analysed, an unmanageable workload for the laboratory. To reduce the amount of experimental work required, three arbitrary UR zone sizes (i.e. %RSD of 8.00%, 16.5%, 25.0%) were selected, and each analyte's %RSD was rounded to the closest of the three predefined, equally spaced, levels (Table 1). This %RSD approximation process reduced the requirements to 3 UR sizes × 10 samples × 9 concentration levels = 270 spiked samples, a reduction by a factor of 13. Samples spiked at $-4s$, $-3s$, $-2s$, $-1s$, cut-off, $+1s$, $+2s$, $+3s$ and $+4s$ were analysed, and a smoothed conditional mean was fitted to the calculated positivity rate. In general, all analytes produced

the expected sigmoidal curve outcome, with some expected deformations attributed to the %RSD rounding process.

TABLE 1 LC–MS/MS validation results

Analyte	Threshold (ng/mL)	%RSD	Rounded %RSD	LURL (ng/mL)	UURL (ng/mL)	FNR (%)	FPR (%)	RLR (%)	SLR (%)	SNR (%)	Validated?
α-Hydroxyalprazolam	20	13	16.5	10	30	0	0	100	100	100	Yes
Aripiprazole	10	10	8.00	8	12	30	0	85	100	70	No
3-Hydroxy Bromazepam	20	17	16.5	10	30	0	0	100	100	100	Yes
Buprenorphine	5	31	25.0	1	9	3	0	98	100	97	Yes
Hydroxybupropion	20	7	8.00	15	25	3	0	98	100	97	Yes
N-Desmethylcitalopram	20	8	8.00	15	25	3	0	98	100	97	Yes
N-Desmethylclobazam	20	11	8.00	15	25	20	0	90	100	80	No
Cocaethylene	20	8	8.00	15	25	0	0	100	100	100	Yes
Norcodeine	20	8	8.00	15	25	27	0	87	100	73	No
N-Desmethylcyclobenzaprine	20	16	16.5	10	30	0	0	100	100	100	Yes
Dextrorphan	20	11	8.00	15	25	7	0	97	100	93	Yes
Nordiazepam	20	9	8.00	15	25	3	0	98	100	97	Yes
N-Desmethyl diphenhydramine	20	11	8.00	15	25	0	0	100	100	100	Yes
Duloxetine	20	15	16.5	10	30	7	0	97	100	93	Yes
Norfentanyl	0.5	13	8.00	0	1	43	0	78	100	57	No
7-Aminoflunitrazepam	20	9	8.00	15	25	40	0	80	100	60	No
N-Desmethylflunitrazepam	20	12	8.00	15	25	30	0	85	100	70	No
Norfluoxetine	20	22	25.0	5	35	0	0	100	100	100	Yes
2-Hydroxyethylflurazepam	20	9	8.00	15	25	80	0	60	100	20	No
Norketamine	20	10	8.00	15	25	10	0	95	100	90	No
Lorazepam-glucuronide	40	24	25.0	10	70	10	0	95	100	90	No
mCPP	20	10	8.00	15	25	0	0	100	100	100	Yes
MDEA	20	8	8.00	15	25	7	0	97	100	93	Yes
MDPV metabolite	20	10	8.00	15	25	17	0	92	100	83	No
Normeperidine	40	8	8.00	30	50	3	0	98	100	97	Yes
α-Hydroxymidazolam	20	9	8.00	15	25	23	0	88	100	77	No
N-Desmethylmirtazapine	20	12	8.00	15	25	0	0	100	100	100	Yes
6-Acetylmorphine	5	10	8.00	4	6	3	0	98	100	97	Yes
Morphine-6β-D-glucuronide	100	31	25.0	25	175	7	0	97	100	93	Yes
Naloxone	20	8	8.00	15	25	3	0	98	100	97	Yes
Naltrexone	20	9	8.00	15	25	7	0	97	100	93	Yes
Desmethylolanzapine	20	30	25.0	5	35	0	10	95	90	100	No
Oxazepam-glucuronide	20	20	16.5	10	30	10	0	95	100	90	No
Phenylpropanolamine	30	9	8.00	23	37	3	0	98	100	97	Yes
Norpseudoephedrine	30	8	8.00	23	37	7	0	97	100	93	Yes
Norquetiapine	20	23	25.0	5	35	3	0	98	100	97	Yes
7-Hydroxyquetiapine	20	10	8.00	15	25	0	0	100	100	100	Yes
Temazepam-glucuronide	20	19	16.5	10	30	3	0	98	100	97	Yes
α-Hydroxytriazolam	20	13	16.5	10	30	0	0	100	100	100	Yes
N-Desmethylzopiclone	20	9	8.00	15	25	7	0	97	100	93	Yes

Abbreviations: FNR, false negative rate; FPR, false positive rate; LC–MS/MS, liquid chromatography–tandem mass spectrometry; LURL, lower unreliability limit; mCPP, *meta*-chlorophenylpiperazine; MDEA, 3,4-methylenedioxy-N-ethylamphetamine; MDPV, methylenedioxypyrovalerone; RLR, reliability rate; RSD, relative standard deviation; SLR, selectivity rate; SNR, sensitivity rate; UURL, upper unreliability limit.

To measure the performance parameters and ion ratio reliability, 30 samples spiked at the LURL and UURL ($+3s$ and $-3s$) for each of the three %RSD levels were analysed. For example, 3,4-methylenedioxy-N-ethylamphetamine (MDEA, cut-off =

20 ng/mL, %RSD = 8%) was spiked at 15 and 25 ng/mL. The results, presented in Table 1 , show that numerous performance parameters fall outside of the expected range (grey-shaded cells in Table 1). On the other hand, ion ratio, carry-over and interference studies were all found to be satisfactory. Overall, 27 out of 40 analytes were considered to be validated.

The validation failure of so many analytes was quite surprising and, in principle, should not have occurred if the theoretical model of measurements and derived binary results was correct. It therefore seemed that something was not taken into account by the model. Further investigation revealed that the average response (area ratio) at threshold and, more importantly, its standard deviation, shifted on a batch-to-batch basis, with an even more marked difference between days. The raw data supporting this conclusion are available as a spreadsheet in Data S4. A set of 30 different blood matrices (antemortem and post-mortem) spiked with all analytes at the threshold concentration was analysed on four different days. For this identical set of matrices, significantly different variances of the area ratio across the different days were observed (using Levene's test for variances). This demonstrates that the changing variance is not linked to matrix effect, but rather is an uncontrolled batch-to-batch variation of the measurement error. This type of analysis thus displays a two-part heteroscedasticity: the first part is the standard deviation changes with the concentration, which is properly accounted for by the model presented here, and the second part is apparently unstructured and unrelated to other factors. This was the key to understanding the validation results and should be taken as a precautionary warning. This previously undescribed behaviour of qualitative threshold methods' results has a major impact on validation and uncertainty of measurement reporting. Indeed, the fact that the standard deviation changed between batches and days means that the size and edges (LURL, UURL) of the unreliability zone also vary daily. For example, Data S4 shows an average sixfold increase in the standard deviation between the two batches. Therefore, although we thought we were measuring method validation parameters (FNR, FPR, SNR, SLR and RLR) at $-3s$ and $+3s$ for each analyte, we might have been making these measurements significantly away from the edges, either inside or outside the unreliability zone on that particular day. This will, naturally, have a major impact on the positivity rate, the method performance, and its measurement uncertainty.

3.4 | Modified method validation guidelines

Having a reliable estimation of the unreliability zone is important not only to apply adequate criteria on validation parameters but also to properly consider measurement uncertainty in production operations and accurately classify samples as below threshold/likely below threshold/likely above threshold/above threshold (Figure 4). Knowing that the size of this unreliability zone varies on a daily basis, the next obvious question is: can the position of its edges (LURL, UURL) be estimated with each batch?

The problem with this approach is that an accurate estimation of the standard deviation is more demanding than an accurate estimation of the average, because this parameter converges more slowly than the mean. To obtain a standard deviation estimation with lower than 20% average error, one would have to analyse at least 10 samples spiked at threshold per batch or day. Given the constraints of a production setting, this is impractical. Moreover, standard uncertainty calculations [13] are designed to be performed on a quantitated concentration, not on the raw instrument measurements, and cannot be applied in this context.

Therefore, until better mathematical modelling and predictive tools are developed, a daily accurate estimate of the size of the unreliability zone seems out of reach. For the moment, the best that can be done is to proceed conservatively. Either perform several estimations of its size on different batches or days and use the largest one or, based on experience, choose one that ensures that the validation points are systematically outside the unreliability zone, e.g. at ±50% of the threshold concentration. In either case, this is not a precise estimation of the unreliability zone on any given day, which again would require more advanced mathematical modelling tools. Note that this is essentially what is done with most immunoassay method validations [1]. Acceptance criteria for this conservatively estimated unreliability zone boil down to fitness for purpose, as discussed by Wille et al. [13] Depending on the application, an unreliability zone of 50% might be considered acceptable (e.g. for post-mortem work) or not (e.g. for driving under the influence cases).

For the validation of this LC–MS/MS method, the first validation results obtained (Table 1), with a surprising number of analytes failing validation, hinted that the size of the UR zone was somehow misestimated initially. Further experiments demonstrated that the standard deviation at threshold (and size of the UR) changed from one batch to the next. To proceed conservatively, we selected a larger UR zone size by increasing the estimated %RSD from 8.0% to 16.5%, thus increasing the

size of the UR zone with LURL and UURL at 50% and 150% of the threshold concentration. Using these more conservative conditions, all analytes satisfied the validation criteria but will have a larger associated UR for reporting purposes.

A threshold qualitative method's report should report samples as "below threshold" if the measured response falls below the LURL, "likely below threshold" if the response falls between the LURL and the threshold, "likely above threshold" if the response is between the threshold and the UURL and "above threshold" if the measurement is above the UURL (Figure 4 . How this translates in the final toxicology report depends on the analysis workflow and institutional perspectives. Indeed, in situations where quantification is performed after a threshold qualitative method, the quantitative result will likely be relied on for the report. For the sake of argument, if we assume that only a threshold qualitative method is carried out, how should the result be expressed to the client in the final report? An institution might hold the perspective that the final report should state the analysis results as transparently as possible and keep the categorization "above/below threshold" and "likely above/likely below threshold". Or an institution might hold the perspective that conservative reporting is better and simpler for the client and report all "below threshold", "likely below threshold" and "likely above threshold" as "below threshold". This intends to avoid false positives and give the benefit of the doubt to the accused. Note that other applications (e.g. medical) might prefer avoiding false negatives; in this case, conservative reporting would mean reporting all "above threshold", "likely above threshold" and "likely below threshold" as "above threshold".

4 | CONCLUSIONS

Qualitative methods yield categorical, binary outputs very different in nature from quantitative methods, and validation guidelines should employ these categorical results to evaluate method performance, not the continuous measurements collected in the process. We have developed a tool to model the measurements and the derived binary results based on the literature and experimental data.

A tentative validation guideline was developed and applied to an LC–MS/MS qualitative decision point method for 40 analytes. The results demonstrated a previously unreported behaviour of this type of measurements: the average area ratio and its variance changes on a daily basis, leading to significant variations in the unreliability zone size which is critical for method validation and uncertainty of measurement reporting.

Considering this behaviour, we offer the following validation guidelines:

1. Decide on the validation points to be used above and below cutoff (LURL, UURL). If the exact size of the uncertainty of measurement is important, the standard deviation can be repeatedly evaluated on different days and the largest one used to conservatively place validation points at $\pm 3s$. If not, a conservatively large size such as cut-off $\pm 50\%$ can be used.
2. Thirty or more samples should be spiked at the validation points (LURL and UURL) and treated as they would in a production setting (i.e. analyse those samples as they would be in a real batch, with the same number of extracted threshold samples as will be used in production, and generate binary results). The validation parameters should satisfy the following criteria, where two injected cut-offs and 30 samples are used for rate estimation: FNR $< 7\%$, FPR $= 0\%$, RLR $> 93\%$, SLR $= 100\%$, SNR $> 93\%$ and ion ratio adequacy $> 95\%$. If other production (number of threshold samples analysed per batch) or validation (number of LURL and UURL samples analysed) conditions are used, the R script in Data S2, Section 3, should be used to establish the validation criteria based on the estimated expected performance.
3. Carry-over, stability and interference studies according to OSAC Toxicology Subcommittee's[10] practices should be performed and satisfy the pre-established criteria.
4. If appropriate, dilution integrity can be assessed by repeating step 2 with the desired dilution and verifiying that validation parameters continue to satisfy the above criteria.
5. The evaluation of matrix effects is folded in method performance evaluation in step 2 using 30 different matrix lots at the different test points. Method performance estimates thus include the impact of a variety of matrices (at any rate, as varied

as the experimenter's matrices choice during validation). However, should a laboratory want to decouple matrix effect evaluation from method performance evaluation (as is often the case in quantitative method validation), they could perform step 2 with 30 replicates of a varied (e.g. 3–10) number of different matrix sources. This would be highly work-intensive.

6. In production (Figure 4), samples whose response falls below the low validation point (LURL) are reported as "below threshold", samples with a measurement between the LURL and the threshold as "likely below threshold", samples with a measurement between the threshold and the high validation point (UURL) as "likely above threshold" and samples with a measurement above the high validation point as "above threshold".

Using this validation guideline and method of reporting results not only produces a performance evaluation more inline with the definition of method validation but also incorporates measurement uncertainty, as required by most accreditation boards under the new ISO 17025:201711 validation guidelines.

However, the unreliability zone associated with each analyte is a conservative estimate. Given the inherent instrumental variability, a precise day-to-day estimation of the unreliability zone will require developing adapted mathematical prediction tools. This is currently a topic of investigation by some of the authors.

Moreover, a validation framework for qualitative identification methods is still lacking. This is a related but distinct problem from the qualitative threshold methods. Indeed, qualitative identification methods not just solely depend on the instrument signal to classify the analytes as "present" or "absent" but also on a series of characteristics such as the retention time, accurate mass and fragments. Given these various inputs to the identification act, in this case, the uncertainty might be better expressed as a probability obtained via Bayesian calculations than as a signal interval as described in this paper for qualitative threshold methods.

5 | ACKNOWLEDGEMENTS

The authors are grateful to Maxime Gosselin for his contribution to the literature review in the early days of the project. Brigitte Desharnais, Félix Camirand Lemyre and Cameron D. Skinner gratefully acknowledge the 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. This research was undertaken thanks in part to the funding from Canada First Research Excellence Fund and the Australian Research Council DP #140100125.

References

1. 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* 2013; 37(7): 452-474. doi: 10.1093/jat/bkt054

2. González, Oskar and Blanco, María Encarnación and Iriarte, Gorka and Bartolomé, Luis and Maguregui, Miren Itxaso and Alonso, Rosa M . 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. *Journal of Chromatography A* 2014; 1353: 10–27. doi: 10.1016/j.chroma.2014.03.077

3. Hartmann, C and Smeyers-Verbeke, J and Massart, DL and McDowall, RD . Validation of bioanalytical chromatographic methods. *Journal of Pharmaceutical and Biomedical Analysis* 1998; 17(2): 193–218. doi: 10.1016/S0731-7085(97)00198-2

4. 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* 2007; 45(1): 82–96. doi: 10.1016/j.jpba.2007.06.032

5. Peters, Frank T and Drummer, Olaf H and Musshoff, Frank . Validation of new methods. *Forensic Science International* 2007; 165(2-3): 216–224. doi: 10.1016/j.forsciint.2006.05.021

6. Peters, Frank T and Maurer, Hans H . Bioanalytical method validation and its implications for forensic and clinical toxicology – A review. *Accreditation and Quality Assurance* 2002; 7(11): 441–449. doi: 10.1007/s00769-002-0516-5

7. Wille, Sarah MR and Coucke, Wim and De Baere, Thierry and Peters, Frank T . Update of standard practices for new method validation in forensic toxicology. *Current Pharmaceutical Design* 2017; 23(36): 5442–5454. doi: 10.2174/1381612823666170714154444

8. Food and Drug Administration . Bioanalytical Method Validation: Guidance for Industry. standard, https://www.fda.gov/downloads/drugs/guidances/ucm070107.pdf; Silver Springs, USA: 2018.

9. European Medicines Agency . Guideline on bioanalytical method validation. standard, https://www.ema.europa.eu/documents/scientific-guideline/guideline-bioanalytical-method-validation_en.pdf; London, United Kingdom: 2011.

10. AAFS Standards Board . Standard Practices for Method Validation in Forensic Toxicology (Draft). standard, https://asb.aafs.org/wp-content/uploads/2018/09/036_Std_Ballot02.pdf; Colorado Springs, USA: 2018.

11. International Organization for Standardization . General requirements for the competence of testing and calibration laboratories. standard, https://www.iso.org/standard/66912.html; Geneva, Switzerland: 2017.

12. Gesellschaft für Toxikologische und Forensische Chemie (GTFCh) . Guideline for quality control in forensic-toxicological analyses. standard, https://www.gtfch.org/cms/index.php/en/guidelines; Germany: 2018.

13. Wille, Sarah MR and Peters, Frank T and Di Fazio, Vincent and Samyn, Nele . Practical aspects concerning validation and quality control for forensic and clinical bioanalytical quantitative methods. *Accreditation and Quality Assurance* 2011; 16(6): 279. doi: 10.1007/s00769-011-0775-0

14. de Souza Gondim, Carina and Coelho, Otávio Augusto Mazzoni and Alvarenga, Ronália Leite and Junqueira, Roberto Gonçalves and de Souza, Scheilla Vitorino Carvalho . An appropriate and systematized procedure for validating qualitative methods: Its application in the detection of sulfonamide residues in raw milk. *Analytica Chimica Acta* 2014; 830: 11–22. doi: 10.1016/j.aca.2014.04.050

15. López, M Isabel and Callao, M Pilar and Ruisánchez, Itziar . A tutorial on the validation of qualitative methods: From the univariate to the multivariate approach. *Analytica Chimica Acta* 2015; 891: 62–72. doi: 10.1016/j.aca.2015.06.032

16. Trullols, E and Ruisánchez, I and Rius, FX and Huguet, J . Validation of qualitative methods of analysis that use control samples. *Trends in Analytical Chemistry* 2005; 24(6): 516–524. doi: 10.1016/j.trac.2005.04.001

17. Parikh, Rajul and Mathai, Annie and Parikh, Shefali and Sekhar, G Chandra and Thomas, Ravi . Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology* 2008; 56(1): 45. doi: 10.4103/0301-4738.37595

18. Altman, Douglas G and Bland, J Martin . Diagnostic tests. 1: Sensitivity and specificity. *British Medical Journal* 1994; 308(6943): 1552. doi: 10.1136/bmj.308.6943.1552

19. Côté, Cynthia and Desharnais, Brigitte and Morel, Marc-André and Laquerre, Julie and Taillon, Marie-Pierre and Daigneault, Gabrielle and Skinner, Cameron D and Mireault, Pascal . High Throughput Protein Precipitation: Screening and Quantification of 106 Drugs and their Metabolites using LC-MS/MS. standard, 2017 Society of Forensic Toxicologists Meeting (SOFT) and 55th Annual Meeting of the International Association of Forensic Toxicologists (TIAFT); Boca Raton, USA: 2018.

20. European Commission Directorate General for Health and Food Safety . Guidance document on analytical quality control and method validation procedures for pesticide residues and analysis in food and feed. standard, https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_mrl_guidelines_wrkdoc_2017-11813.pdf; : 2017.

21. Trullols, Esther and Ruisanchez, Itziar and Rius, F Xavier . Validation of qualitative analytical methods. *Trends in Analytical Chemistry* 2004; 23(2): 137–145. doi: 10.1016/S0165-9936(04)00201-8

22. Yi GY. *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. New York, USA: Springer . 2017.

How to cite this article: Félix Camirand Lemyre, Brigitte Desharnais, Julie Laquerre, Marc-André Morel, Cynthia Côté, Pascal Mireault, and Cameron D. Skinner (2020), Qualitative threshold method validation and uncertainty evaluation: A theoretical framework and application to a 40 analytes liquid chromatography–tandem mass spectrometry method, *Drug Testing and Analysis*, *12:9, 1287–1297*, DOI: 10.1002/dta.2867