Qualitative method validation and uncertainty evaluation via the binary output
I – Validation guidelines and theoretical foundations

Félix Camirand Lemyre[a,b,c,*], Brigitte Desharnais[d,e,**], Julie Laquerre[d], Marc-André Morel[d], Cynthia Côté[d], Pascal Mireault[d], Cameron D. Skinner[e]

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

Abstract

Qualitative methods have an important place in forensic toxicology, filling central needs in, amongst others, screening and analyses linked to *per se* legislation. Nevertheless, bioanalytical method validation guidelines either 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". Since the goal of any method validation is to demonstrate fitness for use under production conditions, guidelines should evaluate performance by relying on the discrete results, instead of the continuous measurements obtained (e.g. peak height, area ratio).

We have developed a tentative validation guideline for decision point qualitative methods by modeling measurements and derived binary results behaviour, based on the literature and experimental results. This preliminary guideline was applied to an LC-MS/MS method for 40 analytes, each with a defined cut-off concentration. The standard deviation of measurements at cut-off (s) was estimated based on 10 spiked samples. Analytes were binned according to their %RSD (8.00%, 16.5%, 25.0%). Validation parameters calculated from the analysis of 30 samples spiked at $-3s$ and $+3s$ (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. Subsequent examination found that this was attributable to an appreciable shift in the standard deviation of the area ratio between different batches of samples analyzed. Keeping this behaviour in mind when setting the validation concentrations, the developed guideline can be used to validate qualitative decision point methods, relying on binary results for performance evaluation and taking into account measurement uncertainty.

An application of this method validation scheme is presented in the accompanying paper (II – Application to a multi-analyte LC-MS/MS method for oral fluid).

---

1. Introduction

Qualitative methods are best described by contrasting them with quantitative methods. The output of qualitative methods is categorical (or discrete) in nature, typically binary: presence or absence (qualitative identification methods), above or below threshold (qualitative decision point methods). On the other hand, quantitative methods produce concentration estimates on a continuous scale.

There is an abundant literature dealing specifically with quantitative methods and their validation procedures [1, 2, 3, 4, 5, 6, 7, 8, 9]. However, the literature and guidelines dealing with qualitative methods is much sparser.

SWGTOX [1] and AAFS Standards Board [10] both contain recommendations for decision point qualitative methods. According to these guidelines, LC-MS/MS qualitative decision point method validation should include interference and carryover studies, dilution integrity and stability if necessary, as well as precision evaluation. Precision of the measured signal should be evaluated at $\geq 50\%$ of the decision point (DP) or cut-off concentration, at the DP concentration and at $\leq 150\%$ of the DP concentration. The method is considered to be validated if %RSD $\leq 20\%$ and $(\bar{x}_{50\%} + 2s_{50\%}) < \bar{x}_{DP} < (\bar{x}_{150\%} - 2s_{150\%})$, i.e. the mean $\pm$ two standard deviations at 50% and 150% of the decision point do not overlap with the mean measurement at cut-off.

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 fact that continuous measurements (area, height, area ratio, luminescence, etc.) are transformed 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 evaluation of the method's uncertainty of measurement (UM). The requirement for UM evaluation in qualitative methods has been recently introduced in the ISO 17025:2017 [11] standard, therefore its absence from published guidelines is not surprising. Nonetheless, given this new requirement, adequate UM evaluation procedures for qualitative methods are required.

---

*F. Camirand Lemyre and B. Desharnais contributed equally to this work and are named in alphabetical order.

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

In any method validation, the goal is to demonstrate the quality of the analytical method by producing objective proof that predefined performance criteria are met [1, 10]. Importantly, this verification of the fitness for use has to occur under the same preparation, analysis and data processing procedures which will be used for 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 the adequacy of its performance, since this is the result ultimately produced in a production setting.

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

In order to answer such questions, the behaviour of the response variable (area ratio, luminescence, etc.) in relation to the encoded, binary outcome must be understood. This subject is touched upon sparingly in the literature [12, 13, 14] where diverse validation procedures are suggested.

In this paper, we draw upon these various sources, computer simulations and experimental data to study the behaviour of the binary above/below threshold output of qualitative decision point methods, in order to put forward a tentative validation guideline. This guideline is heavily based on the performance evaluation of another kind of categorical test: medical tests for the presence of a diseased state [15, 16]. This validation process is then evaluated by using an LC-MS/MS method for 40 analytes in blood. Results guided modifications to the guidelines. The final version of the qualitative decision point method validation guidelines is applied to an LC-MS/MS method for 97 analytes in oral fluid in Part II of this paper.

2. Materials and methods

2.1. Analytical method

The experimental data used for prospective and confirmatory studies of qualitative decision point 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 [17] and will not be discussed in this paper. For every qualitative analyte, a cut-off concentration was selected based on analytical (sensitivity across multiple LC-MS/MS systems) and toxicological (relevant concentrations for effects) considerations. The full list of substances and their designed cut-off is available in Supplementary Data 1.

2.1.1. Sample preparation

Samples were brought to room temperature over 1 hour. Following vortex mixing for 10 seconds, 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).

For the purpose of this study, blood samples were spiked at the cut-off concentration or its multiples (e.g. 50%, 150%, 200%, etc.). Postmortem cardiac and femoral blood with negative screening results, as well as antemortem blood purchased from UTAK (Valencia, California, USA) were used. All compounds used for spiking purposes were purchased from Cerilliant (Round Rock, Texas, USA), except for 3-hydroxy bromazepam and N-desmethyl diphenydramine which were purchased from Toronto Research Chemicals (North York, Ontario, Canada). 10 μL of stable isotope labelled internal standards solution (IS, Cerilliant, Round Rock, Texas, USA), at concentrations indicated in Supplementary Data 1, were added to the blood sample and mixed using vortexing.

In order to obtain a more finely granular precipitate, 100 μL of methanol:0.2% formic acid in water (50:50 v:v) solution was mixed into the blood sample. Then 400 μL of acetone:acetonitrile (30:70 v:v) mixture was used to precipitate the proteins. Following mixing, the plate was centrifuged at 3200 $\times g$ for 5 minutes. 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 0.2% formic acid in water and vortexed.

2.1.2. LC-MS/MS analysis

A 5 μL aliquot of diluted extract was separated on an Agilent Zorbax Eclipse Plus C18 column (2.1 x 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 with regards to the liquid chromatography and mass spectrometry acquisition are available in Supplementary Data 1.

2.2. Preliminary validation guidelines

Based on the picture of the behaviour of binary results in qualitative decision point methods described in the literature and our exploratory experimental data analysis (described in Section 3.1), a preliminary set of validation guidelines was determined and applied.

The standard deviation was estimated by analyzing a minimum of 10 different samples all spiked at the cut-off and calculating the standard deviation of the response variable used, in this case the ratio of analyte peak area to IS peak area.

Probability curves (akin to the one shown in Figure 1b) plotting the positivity rate (or above cut-off 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 and a blank response of zero.

The core of the validation procedure consisted of analyzing 30 samples spiked at $-3s$ and $+3s$ (upper (UURL) and lower (LURL) unreliability limits, see explanation in Section 3), which were used to calculate the method's validation parameters (see Section 2.2.1). Care was taken to ensure that these samples were prepared, injected and analyzed 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 cut-off.

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).

Carryover and interference studies should be carried out, as well as stability evaluation if deemed necessary. Although several procedures and guidelines exist, SWGTOX's practices are recommended in forensic toxicology [1]. If applicable, dilution integrity can be verified by repeating the main validation procedures (standard deviation estimation, evaluation of performance parameters on diluted samples). These studies were carried out for the method presented here, but will not be discussed since they are not the main focus of this paper.

2.2.1. 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 [12, 13, 15]:

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

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

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

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

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

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.

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

A qualitative decision point method validated under these production conditions (2 measured cut-off samples, rates estimated over 30 samples) can be considered fit for purpose if the $FNR \leq 7\%$, $FPR = 0\%$, $RLR \geq 93\%$, $SLR = 100\%$, $SNR \geq 93\%$ (Figure 1d) and ion ratio, carry-over and interference studies are successful. These expected performance levels were calculated using the RStudio script presented in Supplementary Data 2, Section 3. Expected performance levels under a different number of measured cut-offs and number of samples for rate estimation can be computed from the same R script: readers are encouraged to use it to define criteria under their own production and validation conditions.

2.3. Computer simulations

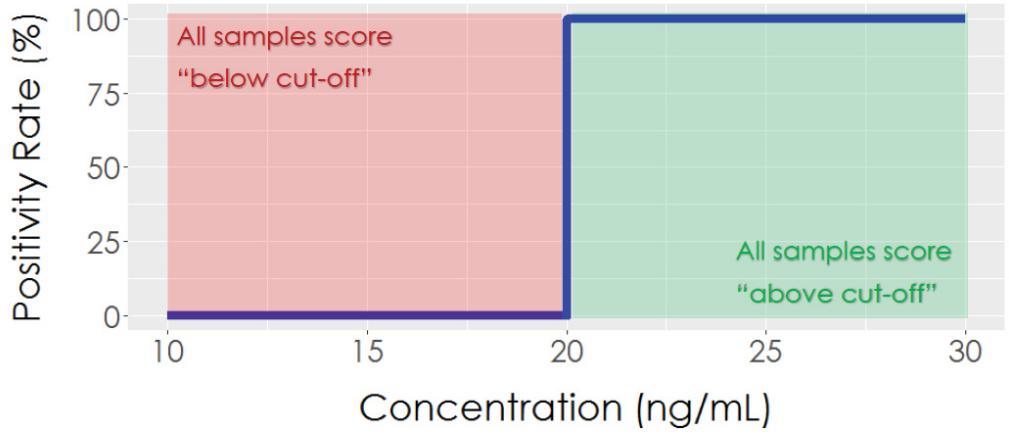
In preparation for the simulations, a set of 30 different samples spiked at cut-off were extracted and analyzed to determine their area ratios (analyte peak area/internal standard peak area). Application of the Cramer-von Mises normality test did not show significant departures from normality in all but two 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 Supplementary Data 3. This is in accordance with the implicit statement consensus in the literature, namely that measurements (e.g. area, area ratios), including those made on an LC-MS/MS instrument, can be approximated by a normal distribution [12, 13, 14, 18].

Based on these results, response values for simulations were modeled 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 cut-off was set to vary between 0.008 and 1.050, based on the experimentally observed area ratios at the cut-off 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. R scripts used to carry out these simulations are available in Supplementary Data 2.

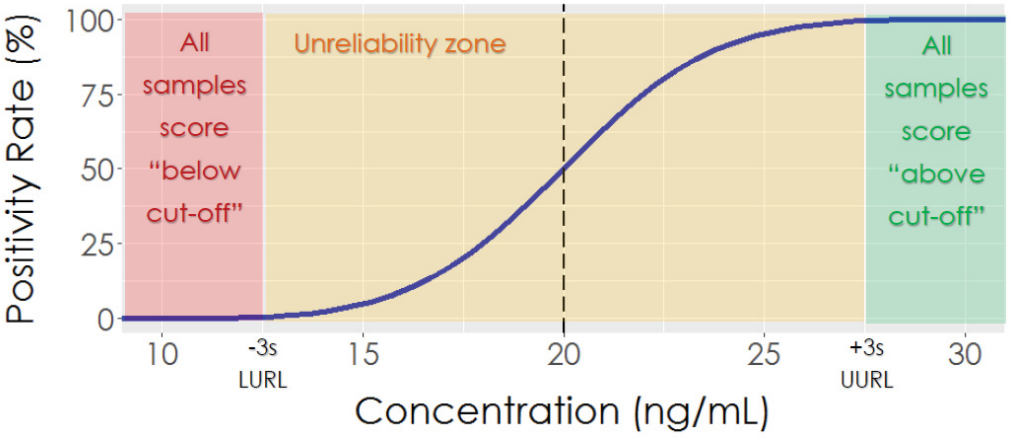
3. Results and Discussion

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 1a. 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 analyzed, and samples with a concentration higher than cut-off will similarly score positive (or above cut-off) systematically.



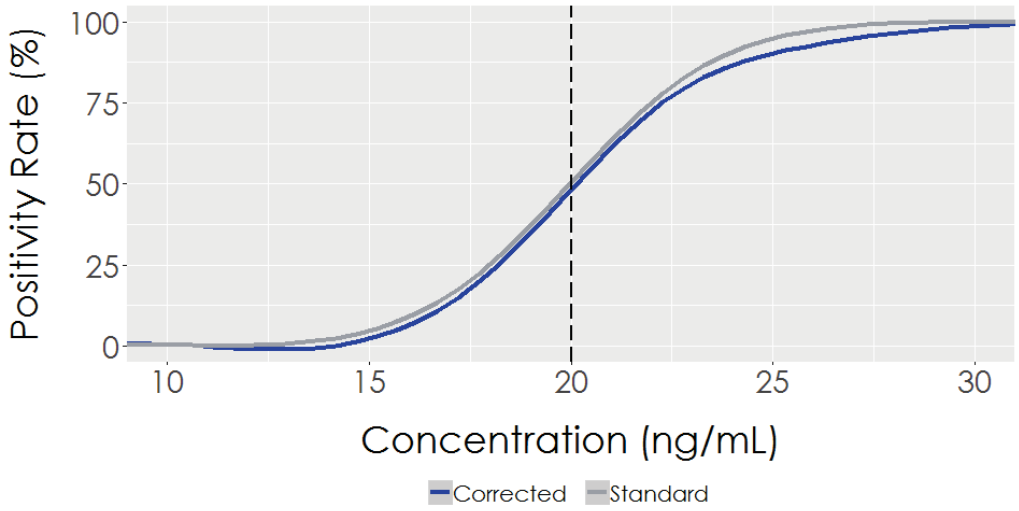
(a) Idealized behaviour of decision point qualitative methods



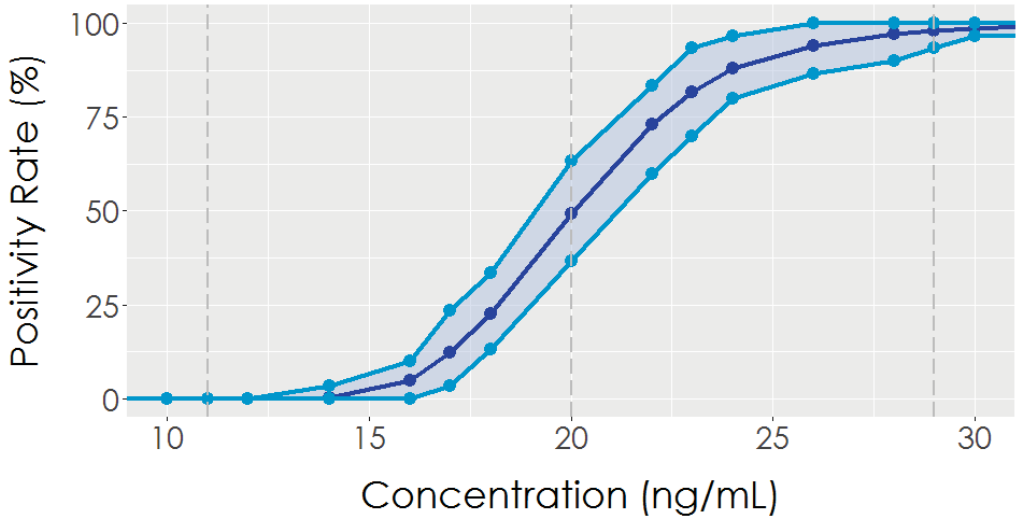
(b) Positivity curve for normally distributed measurements compared to a 20 ng/mL threshold

Figure 1: Positivity curves under different models

While this would be incredibly helpful, it is unfortunately impossible. A sample at a given concentration subjected to experimental manipulations and measured by a device



(c) Corrected positivity curve, accounting for heteroscedasticity and sampled threshold



(d) Average positivity rate when 30 spiked samples are measured and compared to a sampled cut-off (two measurements to establish threshold). 90% of positivity rate results fall within the shaded area (5% to 95% quantiles).

Figure 1: Positivity curves under different models

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 [19]. Thus, as can be seen in Figure 2a, repeated measurements on a sample with a concentration exactly equal to the cut-off will yield a normal distribution with an average response equal to the cut-off. 50% of these measurements will be reported as "below cut-off" and 50% as "above cut-off" (50% positivity rate).

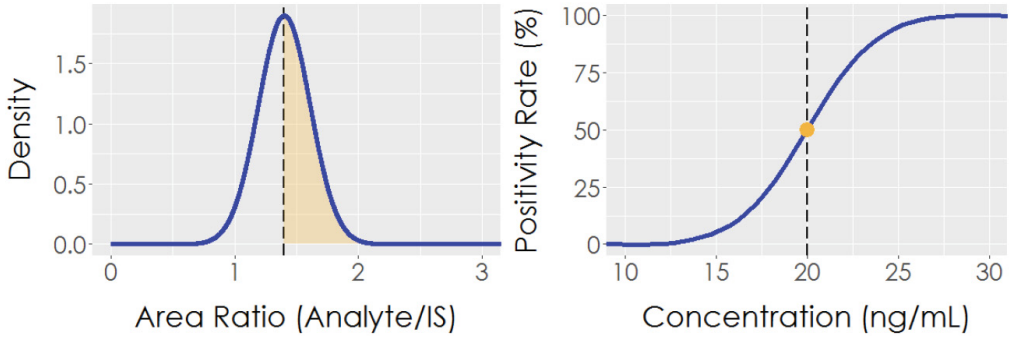
If the sample analyzed has a concentration far enough away from the cut-off, e.g. if the mean measurement for that concentration is 3σ above or below from the cut-off value (Figure 2b), then almost all responses ($> 99.7\%$) will be reported as "above cut-off", or "below cut-off" respectively.

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 2c). Samples with a concentration generating a mean 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 responses below the cut-off with positivity rates decreasing as the concentration decreases.

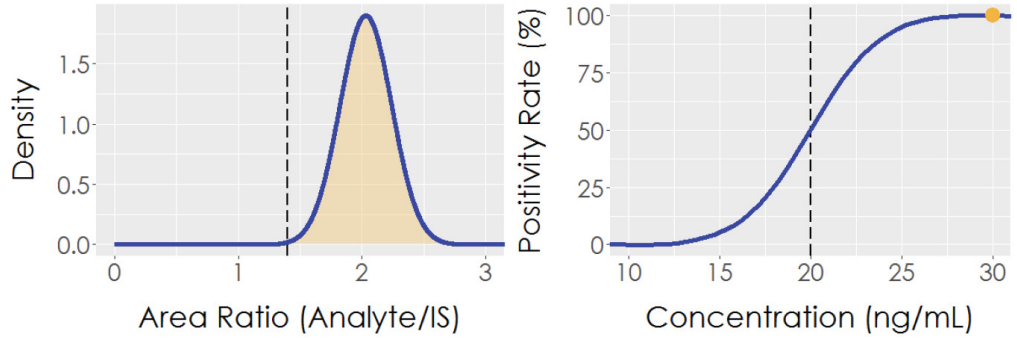
The positivity curve for normal measurements compared to a threshold thus takes a sigmoidal form (Figure 1b). The uncertainty of measurement associated with qualitative methods is evident in this figure. Surrounding the cut-off is a range of concentrations where repeated measurement of the same sample will not always yield the same classification result (and the positivity rate takes an intermediate value). This unreliability (UR) zone, stemming from the uncertainty of measurement, is an ontological characteristic of qualitative decision point methods, and there is no possible way to avoid it. While 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 since the unreliability zone would just follow right along with it. In much the same way that one cannot avoid measurement uncertainty in quantitative methods, 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.

The positivity curve shown in Figure 1b is the one typically reported in the literature. But in order to adequately represent realistic (LC-MS/MS) data, it must be modified to take into account two important factors. First, real measurements are typically heteroscedastic, even over small concentration ranges. This is clearly demonstrated in Figure 3, which displays normal distribution curves based on 30 spiked samples replicates at different concentrations for buprenorphine. The variance increases with increasing concentrations, as made evident by the decreasing distribution maxima. This must be accounted for in validation guidelines.

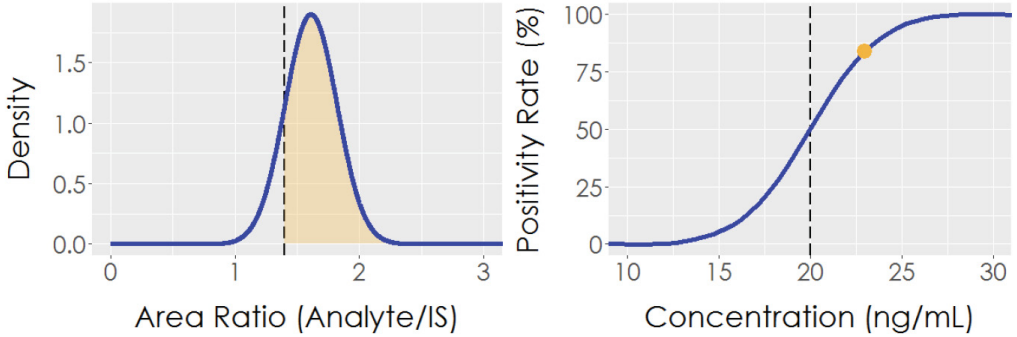
Second, the standard positivity curve presumes that the response (measurement) at cut-off is a known and fixed value. But of course, this is not the case; in a production setting, this value is estimated based on a few measurements (typically 1 to 3) made



(a) Sample spiked at the threshold concentration (20 ng/mL): distribution of measurements (density plot, left) and positivity curve (right). Exactly 50% of measurements are above the threshold measurement, resulting in a 50% positivity rate.



(b) Sample spiked at $> 3\sigma$ above the cut-off concentration (30 ng/mL): distribution of measurements (density plot, left) and positivity curve (right). The whole distribution is far from the measurement at cut-off, yielding a 100% positivity rate.



(c) Sample spiked at an intermediate concentration (between the cut-off concentration and 3σ above the cut-off 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%).

Figure 2: Normally distributed measurements in relation to a fixed threshold

on a sample spiked at the cut-off concentration. In other words, the threshold value is sampled, not fixed, and this means an unknown error of variable size is attached to the estimated value. This implies that the estimated threshold will move from experiment to

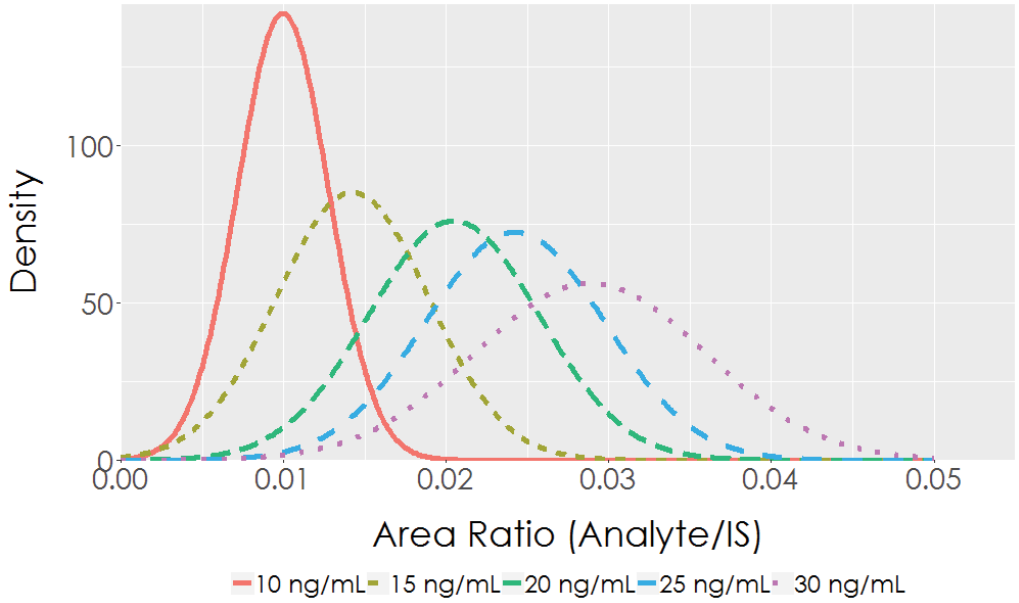


Figure 3: Fitted normal distribution curves for buprenorphine samples ($n = 30$) spiked at different concentrations

experiment, which has a domino effect on which samples get called "above" or "below" cut-off, and thus on the positivity curve.

Fortunately, these measurement characteristics can be modeled and taken into account in establishing validation criteria, i.e. their impact on the positivity curve can be calculated. Details of the modeling performed in RStudio can be found in Supplementary Data 2 (Section 2). The resulting positivity curve is shown in Figure 1c. Notable differences result, particularly at the high concentration end, which affects the expected false negative rate and other parameters relying on it (reliability and sensitivity rates).

This software tool can also be used in establishing appropriate validation criteria by modelling realistic behaviour (heteroscedastic data, cut-off with sampling error) of the measurements and the derived binary results.

3.2. Derived method validation guidelines

The derived method validation guidelines presented in Section 2.2 utilize the performance of the actual method's output, the binary "above cut-off"/"below cut-off" results. With these guidelines we recognize the presence of measurement unreliability due to measurement error. Consequently, evaluation of the method's performance and validation needs to be performed outside of the unreliability zone but provides the most pertinent figures of merit when performed near these boundaries (lower and upper unreliability limits). 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 of it through an overestimation of the UR size (e.g. $\pm 50\%$ of the cut-off 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 cut-off should be reported as "likely below cut-off", and those between the measurement at cut-off and $+3s$ (UURL) should be reported as "likely above cut-off". This will reflect the fact that repeated measurements on these samples might yield different results, and will adequately convey measurement uncertainty in the final analysis report.

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 cut-off. 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. For the probability curves alone, 40 analytes x 10 samples x 9 concentration levels = 3 600 spiked samples would need to be analyzed, an unmanageable workload for the laboratory. Instead, analytes were classified as belonging to one of three standard deviation (%RSD) bins: 8%, 16.5% or 25% (Table 1). This binning process reduced the requirements to 3 bins x 10 samples x 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 analyzed, and a smoothed conditional mean was fitted to the calculated positivity rate. Generally, all analytes produced the expected sigmoidal curve outcome, with some expected deformations attributed to the binning process.

To measure performance parameters and ion ratio reliability, 30 samples spiked at the LURL and UURL for each of the %RSD bins were analyzed. For example, MDEA (cut-off = 20 ng/mL, %RSD = 8%) was spiked at 15 and ng/mL. Results, displayed in Table 1, show that numerous performance parameters fall outside of the expected range (greyed-out cells in Table 1). On the other hand, ion ratio, carryover 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 cut-off and, more importantly, its standard deviation, shifted on a batch to batch basis, with an even more marked difference between days (Supplementary Data 4). The measurement error was therefore not adequately characterized or controlled. This type of analysis thus

Table 1: LC-MS/MS validation results

Analyte	Cut-Off (ng/mL)	s (ng/mL)	Bin	LURL (ng/mL)	UURL (ng/mL)	FNR	FPR	RLR	SLR	SNR	Validated?
α-Hydroxyalprazolam	20	0.13	0.165	10	30	0%	0%	100%	100%	100%	YES
Aripiprazole	10	0.1	0.08	8	12	30%	0%	85%	100%	70%	NO
3-Hydroxy Bromazepam	20	0.17	0.165	10	30	0%	0%	100%	100%	100%	YES
Buprenorphine	5	0.31	0.25	1	9	3%	0%	98%	100%	97%	YES
Hydroxybupropion	20	0.07	0.08	15	25	3%	0%	98%	100%	97%	YES
N-Desmethylcitalopram	20	0.08	0.08	15	25	3%	0%	98%	100%	97%	YES
N-Desmethylclobazam	20	0.11	0.08	15	25	20%	0%	90%	100%	80%	NO
Cocaethylene	20	0.08	0.08	15	25	0%	0%	100%	100%	100%	YES
Norcodeine	20	0.08	0.08	15	25	27%	0%	87%	100%	73%	NO
N-Desmethylcyclobenzaprine	20	0.16	0.165	10	30	0%	0%	100%	100%	100%	YES
Dextrorphan	20	0.11	0.08	15	25	7%	0%	97%	100%	93%	YES
Nordiazepam	20	0.09	0.08	15	25	3%	0%	98%	100%	97%	YES
N-Desmethyl diphenydramine	20	0.11	0.08	15	25	0%	0%	100%	100%	100%	YES
Duloxetine	20	0.15	0.165	10	30	7%	0%	97%	100%	93%	YES
Norfentanyl	0.5	0.13	0.08	0	1	43%	0%	78%	100%	57%	NO
7-Aminoflunitrazepam	20	0.09	0.08	15	25	40%	0%	80%	100%	60%	NO
N-Desmethylflunitrazepam	20	0.12	0.08	15	25	30%	0%	85%	100%	70%	NO
Norfluoxetine	20	0.22	0.25	5	35	0%	0%	100%	100%	100%	YES
2-Hydroxyethylflurazepam	20	0.09	0.08	15	25	80%	0%	60%	100%	20%	NO
Norketamine	20	0.1	0.08	15	25	10%	0%	95%	100%	90%	NO
Lorazepam-glucuronide	40	0.24	0.25	10	70	10%	0%	95%	100%	90%	NO
mCPP	20	0.1	0.08	15	25	0%	0%	100%	100%	100%	YES
MDEA	20	0.08	0.08	15	25	7%	0%	97%	100%	93%	YES
MDPV metabolite	20	0.1	0.08	15	25	17%	0%	92%	100%	83%	NO
Normeperidine	40	0.08	0.08	30	50	3%	0%	98%	100%	97%	YES
α-Hydroxymidazolam	20	0.09	0.08	15	25	23%	0%	88%	100%	77%	NO
N-desmethylmirtazapine	20	0.12	0.08	15	25	0%	0%	100%	100%	100%	YES
6-Acetylmorphine	5	0.1	0.08	4	6	3%	0%	98%	100%	97%	YES
Morphine-6β-D-glucuronide	100	0.31	0.25	25	175	7%	0%	97%	100%	93%	YES
Naloxone	20	0.08	0.08	15	25	3%	0%	98%	100%	97%	YES
Naltrexone	20	0.09	0.08	15	25	7%	0%	97%	100%	93%	YES
Desmethylolanzapine	20	0.3	0.25	5	35	0%	10%	95%	90%	100%	YES
Oxazepam-glucuronide	20	0.2	0.165	10	30	10%	0%	95%	100%	90%	NO
Phenylpropanolamine	30	0.09	0.08	23	37	3%	0%	98%	100%	97%	YES
Norpseudoephedrine	30	0.08	0.08	23	37	7%	0%	97%	100%	93%	YES
Norquetiapine	20	0.23	0.25	5	35	3%	0%	98%	100%	97%	YES
7-Hydroxyquetiapine	20	0.1	0.08	15	25	0%	0%	100%	100%	100%	YES
Temazepam-glucuronide	20	0.19	0.165	10	30	3%	0%	98%	100%	97%	YES
α-Hydroxytriazolam	20	0.13	0.165	10	30	0%	0%	100%	100%	100%	YES
N-Desmethylzopiclone	20	0.09	0.08	15	25	7%	0%	97%	100%	93%	YES

displays a two-part heteroscedasticity: 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 disappointing validation results and should be taken as a precautionary warning. The standard deviation changed between batches and daily, which meant that the size and edges (LLURL, UURL) of the unreliability zone also varied daily. Supplementary Data 4 shows, for example, an average 6-fold increase in the standard deviation between two batches. Therefore, while we thought we were measuring method validation parameters (FNR, FPR, SNR, SLR, RLR) at the binned $-3s$ and $+3s$ for each analyte, we might actually have been making these measurements significantly away from the edges, either inside or outside of 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 to apply adequate criteria on validation parameters, but also to properly take into account measurement uncertainty in production operations and accurately classify samples as below cut-off/likely below cut-off/likely above cut-off/above cut-off. 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 accurate estimation of the standard deviation is more difficult than accurate estimation of an average, since this parameter converges more slowly than the mean. To obtain a standard deviation estimation with lower than 20% average error, one would have to analyze at least 10 samples spiked at cut-off per batch/day. Given the constraints of a production setting, this is impractical.

Therefore, until better mathematical predictive tools are developed, an 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/days and use the largest one, or, based on experience, choose one that insures that the validation points are outside the unreliability zone, e.g. at ±50% of the cut-off concentration. Note that this is essentially what is done with immunoassay method validations [1].

For validation of this LC-MS/MS method, all analytes in the 8%RSD bin were pushed up into the 16.5%RSD bin, yielding LURL and UURL at 50% and 150%. Using these relaxed conditions, all analytes satisfied the validation criteria.

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. Results also 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 of the unreliability zone size which is critical for method validation.

Considering this behaviour, we offer the following validation guidelines:

1. Decide on the validation points to be used above and below cut-off (LURL, UURL). If the 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. 30 samples should be spiked at the validation points (above and below cut-off) and treated as they would in a production setting (i.e. analyze those samples as they would be in a real batch, with the same number of extracted cut-offs as will be used in production, and generate binary results). The validation parameters should satisfy the following criteria (for 2 injected cut-offs and 30 samples used for rate estimation): $FNR \leq 7\%$, $FPR = 0\%$, $RLR \geq 93\%$, $SLR = 100\%$, $SNR \geq 93\%$, ion ratio adequacy $> 95\%$. For other conditions, the R Script in Supplementary Data 2, Section 3 can be used to estimate expected performance.
3. Carryover and interference studies according to SWGTOX's practices should be performed and satisfy pre-established criteria.
4. If appropriate, dilution integrity can be assessed by repeating Step 2) with the desired dilution and verification that validation parameters continue to satisfy the above criteria.
5. In production, samples whose response falls below the low validation point are reported as "below cut-off", samples with a measurement between the low validation point (LURL) and the cut-off as "likely below cut-off", samples with a measurement between cut-off and the high validation point (UURL) as "likely above cut-off" and samples with a measurement above the high validation point as "above cut-off".

Using this validation guideline and method of reporting results not only produces a performance evaluation more in line with the definition of method validation, but also takes into account measurement uncertainty, as required by the new ISO 17025:2017 [11] validation guidelines. In the accompanying paper (II – Application to a multi-analyte LC-MS/MS method for oral fluid), this framework was successfully applied to a method covering 97 analytes in saliva collected using a Quantisal® device.

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 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 37 (2013) 452–474.

[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 1353 (2014) 10–27.

[3] 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.

[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 45 (2007) 82–96.

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

[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 7 (2002) 441–449.

[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 23 (2017) 5442–5454.

[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] 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 830 (2014) 11–22.

[13] 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 891 (2015) 62–72.

[14] 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 24 (2005) 516–524.

[15] 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 56 (2008) 45.

[16] Altman, Douglas G and Bland, J Martin, Diagnostic tests. 1: Sensitivity and specificity, British Medical Journal 308 (1994) 1552.

[17] 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.

[18] Trullols, Esther and Ruisanchez, Itziar and Rius, F Xavier, Validation of qualitative analytical methods, Trends in Analytical Chemistry 23 (2004) 137–145.

[19] G. Y. Yi, Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application, Springer, New York, USA, 2017.