Eurachem

CITAC

Cooperation on International Traceability in Analytical Chemistry

EURACHEM / CITAC Guide

# Assessment of performance and uncertainty in qualitative chemical analysis

AQA 2021

INTENTIONALLY BLANK

*I personally saw what looked like an animal, but I can't be absolutely positive that it wasn't a mineral. I think that what was involved was really energy rather than matter. Relatively speaking, it would be easiest to describe the whole thing as a phenomenon hovering somewhere on borderland of dimensions and designations, on the abutment of color, shape, odor, mass, length and breadth, contours, shadows, darkness and so on and so forth.*

Slawomir Mrożek, "Streap-Tease"
Translation by Edward Rothert

INTENTIONALLY BLANK

# Assessment of performance and uncertainty in qualitative chemical analysis

First Edition (2021)

## Editors

Ricardo Bettencourt da Silva (Faculdade de Ciências da Univ. de Lisboa),
Stephen L R Ellison (LGC, UK)

## Composition of the Working Group*

| | |
|---|---|
| ***Eurachem members*** | |
| R. Bettencourt da Silva (Chair) | *Univ. Lisboa, Portugal* |
| S. Ellison (Secretary) | *LGC, United Kingdom* |
| A. Togola | *BRGM, France* |
| D. Ivanova | *Eurachem Bulgaria* |
| E. Theodorsson | *LIU, Sweden* |
| E. Totu | *University Politehnica of Bucharest, Romania* |
| H. Emons | *European Commission, European Union* |
| I. Leito | *Univ Tartu, Estonia* |
| M. Sega | *INRIM, Italy* |
| O. Levbarg | *Ukrmetrteststandart, Ukraine* |
| O. Pellegrino | *IPQ/DMET, Portugal* |
| P. Pereira | *IPST, Portugal* |
| R. Kaus | *Eurachem Germany* |
| S. Lardy-Fontan | *LNE, France* |
| W. Wegscheider | *Montanuniversitaet Leoben, Austria* |
| ***CITAC members*** | |
| A. Botha | *NMISA, South Africa* |
| F. Lourenço | *Univ. São Paulo, Brazil* |

*At time of document approval

This publication should be cited* as:
“R Bettencourt da Silva and S L R Ellison (eds.) Eurachem/CITAC Guide: Assessment of performance and uncertainty in qualitative chemical analysis. First Edition, Eurachem (2021). ISBN 978-0-948926-39-6. Available from https://www.eurachem.org”

* *Subject to journal requirements*

## Acknowledgements

This document has been produced by a joint Eurachem/CITAC Working Group with the composition shown (right). The editors are grateful to all these individuals and organisations and to others who have contributed comments, advice and assistance.

Production of this Guide was in part supported by Fundação para a Ciência e a Tecnologia, Portugal.

Assessment of performance and uncertainty in qualitative chemical analysis

English edition

First Edition (2021)

# Contents

INTENTIONALLY BLANK

# Foreword

The problem of evaluating and expressing uncertainty in qualitative chemical analysis has received much less coverage in the literature than that afforded to uncertainty for quantitative analysis (i.e., measurements) [1]. While some authors have addressed this area [2] – [11], general guidance on the assessment of performance in qualitative analysis or the evaluation and reporting of qualitative analysis uncertainty is scarce.

Accredited laboratories are not currently expected to evaluate or report uncertainties associated with qualitative analysis results [12]. However, ISO/IEC 17025 [13] and ISO 15189 [14] both require laboratories to ensure that they can achieve valid qualitative and quantitative analysis results. It is also crucial for laboratories to be aware of the reliability of qualitative analysis results; this enables them, where necessary, to warn of limitations in the interpretation of results and respond accurately to customer queries about reliability. A quantitative assessment of the reliability of a qualitative analysis result is particularly useful when false results are more likely. This Guide is intended for use when a quantitative assessment of the reliability of a qualitative analysis result is desirable.

This Guide relies on experiences from several of the analytical fields where qualitative analysis is frequently used, e.g., in the forensic [15] and clinical fields [16] – [18], and extensive general guidance [7].

# Scope

This Guide is intended to assist laboratories in setting and implementing appropriate methodologies for assessing the performance of qualitative analysis methods and evaluating uncertainties in qualitative chemical analysis.

In this Guide, qualitative analysis is defined as "*Classification according to specified criteria*". For analytical chemistry and related disciplines, the 'criteria' are understood to relate, in general, to information for the determination of chemical composition, properties and/or structure of analysed items.

The following types of criteria are considered in this Guide:

- Quantitative criteria in which a numerical result is used to categorise a test item as belonging to a pre-established class;
- Qualitative criteria such as the presence or absence of a particular feature, colour change on a test, etc.

This Guide is not exhaustive when describing available tools for the performance assessment of qualitative analysis methods and the uncertainty of qualitative analytical results. The performance characteristics presented in this Guide are based on measured or estimated false result rates and do not consider, for example, measures of agreement between qualitative methods or treatment of classification on ordinal scales[1] other than as a correct or incorrect classification.

[1] An ordinal scale is a scale of natural ordered categories where the distance between the categories is not known. The Mohs scale is an ordinal scale for mineral hardness.

# Abbreviations and symbols

The following abbreviations and symbols occur in this Guide. The symbols used in this document are not harmonised in all fields of science where they apply. For example, in the medical laboratory, *FN* and *FNR* are the abbreviations for “number of false negative results” and *“*false negative ratio”, respectively.

| | |
|---|---|
| $A$ | Ion abundance of a mass spectrum |
| $\bar{A}$ | Mean ion abundance of a mass spectrum |
| $AR$ | Abundance ratio of mass spectrum ions |
| $c$ | Measured concentration (or any other quantity) of the analysed item |
| CI | Confidence interval |
| $c_{\mathrm{max}}$ | Maximum admissible concentration |
| $c_{\mathrm{min}}$ | Minimum admissible concentration |
| $DOR$ | Diagnostic odds ratio |
| $E$ | Efficiency |
| $fn$ | Number of false negative results |
| $FN$ | False negative rate referenced to positive cases |
| $fp$ | Number of false positive results |
| $FP$ | False positive rate referenced to negative cases |
| GC-MS | Gas chromatography-mass spectrometry |
| GC-MS/MS | Gas chromatography-tandem mass spectrometry |
| GUM | Guide to the Expression of Uncertainty in Measurement |
| $HL_{RR.95}$ | High limit of 95 % confidence interval for result rate *RR* (e.g., *SS*) |
| LC-MS | Liquid chromatography-mass spectrometry |
| $LL_{RR.95}$ | Low limit of 95 % confidence interval for result rate *RR* (e.g., *SS*) |
| $LL_{RR.95}^{tg}$ | Target or minimum value for the $LL_{RR.95}$ |
| LOD | Limit of detection |
| LOQ | Limit of quantification |
| $LR$ | Likelihood ratio |
| $LR(+)$ | Likelihood ratio of positive results |
| $LR(-)$ | Likelihood ratio of negative results |
| $n$ | Number of negative results |
| $nc$ | Number of negative cases |
| $NPV$ | Negative predictive value |
| $O(\cdot)$ | Odds in favour of an event, *e. g.* $O(A)$ denotes odds for event *A* |
| $p$ | Number of positive results |
| $P(\cdot)$ | Probability of an event; *e. g.* $P(A)$ is the probability of event *A* |
| $P(+)$ | Prior probability of positive case |
| $P(-)$ | Priori probability of negative case |
| $pc$ | Number of positive cases |
| $PN$ | Posterior probability of negative case (see Annex A) |
| $PP$ | Posterior probability of positive case (see Annex A) |
| $PPV$ | Positive predictive value |
| qPCR | Quantitative polymerase chain reaction |
| RT-PCR | Reverse transcription polymerase chain reaction |
| $RA$ | Relative abundance |
| Rn | Normalized reporter |
| $RR$ | Result rate |
| $s_A$ | Ion abundance standard deviation |
| $SP$ | Specificity |
| $SS$ | Sensitivity |
| $s_{t\mathrm{R}i}$ | Retention time standard deviation |
| $tn$ | Number of true negative results |
| $TN$ | True negative rate referenced to negative cases |
| $tp$ | Number of true positive results |
| $TP$ | True positive rate referenced to positive cases |
| $t_{\mathrm{R}}$ | Retention time |
| $\bar{t}_{\mathrm{R}i}$ | Mean retention time |
| $u(c)$ | Standard uncertainty of *c* |
| $w$ | Mass fraction |
| $Y$ | Youden Index |
| ΔRn | Normalised reporter value minus the baseline response |
| $\rho$ | Spearman’s correlation coefficient between pairs of ion abundances |

# 1 Introduction

Many relevant socio-economic or individual interests, such as industrial productivity and health condition, depend on chemical analysis. Some of these analyses are exclusively qualitative or involve a subsequent quantification of the identified chemical entity. The interests intended to be protected by these analyses are only preserved if the analytical quality is fit for the intended use.

In some publications, the term ‘examination’ [1], ‘examination of a nominal property’ [19], or ‘test’ and ‘testing’ are used for ‘qualitative analysis’. The international standard for accrediting medical laboratories uses the term ‘examination’ both for quantitative and qualitative analysis [14]. Therefore, since consensus has not been reached about these terms amongst various relevant international communities, this Guide uses the term ‘qualitative analysis’ for the determination of nominal (qualitative) properties in chemical analysis.

Broadly stated, a qualitative analytical result is a simple statement or categorisation of a test item or material, i.e., a classification. Decisions are invariably taken based on the categorisation; for example, whether to issue a batch of fertiliser, whether water is fit to drink, whether a person is in the possession of a controlled substance or not, or whether a newly synthesised material has the correct structure based on the requirements. Incorrect classifications – such as ‘accepting’ a product when it is unfit for use – carry risks to all parties. To control these risks, professionals involved in analysis work hard to ensure that their procedures lead to acceptably low incorrect classification risks.

It follows that, at some point in the development of any such test procedure, an evaluation must be made of the risk of incorrect classification. Therefore, for most such procedures, it is reasonable to expect a laboratory to establish, or have access to, information on the risks of incorrect results. An important exception is the use of standardised test procedures, established by groups outside the laboratory, as fit for the intended purpose [20] −[21][22] [23]. The laboratory may well have limited or even no access to performance data of such test procedures. However, these procedures invariably specify the test with relevant detail and the laboratory will generally be expected to show that relevant factors within its control do indeed meet the test procedure's requirements. That, in turn, may involve demonstrating that the uncertainty of controlled parameters and test performance is adequate in relation to the test's purpose.

Evaluating uncertainties associated with quantitative parameters or analysis results has been the subject of considerable efforts since the publication of the “Guide to the Expression of Uncertainty in Measurement” (GUM), which is available as ISO Guide 98 [24] as well as a JCGM document [25]. On the other hand, uncertainties in qualitative analysis have received far less attention. After the publication of the first edition of ISO/IEC 17025 [26], the interest in uncertainties of qualitative analysis has increased. The challenges in establishing the uncertainty associated with qualitative analysis, such as ‘pass/fail’, identity or comparative identity analyses have accordingly gained more attention, particularly in fields where the impact of false qualitative analysis results are extremely relevant, e.g., in forensics or doping analysis.

There is a wide variety of metrics for expressing uncertainty in qualitative results [7]. However, there is limited consensus about which metrics to use. Exceptions are the areas of epidemiology and in the clinical laboratory, where the concepts ‘clinical sensitivity’ and ‘clinical specificity’ are consistently used as clinical accuracy parameters [27].

Quantitative and qualitative analyses differ substantially in how the results and associated uncertainties are reported. While quantitative results are reported as an interval that includes the ‘true value’ of the measurand with a defined confidence level, nominal properties are reported as a classification with metrics that express the chance of correct or incorrect classification. That ‘chance’ can be described by a probability, likelihood, odds, or other metrics estimated from the interpretation of input information. The quality of reported metrics depends on the number and diversity of cases studied. The determination of these metrics allows for the identification of cases where procedures should be improved to reduce the likelihood of producing false results.

This Guide describes the general principles for the performance assessment of qualitative analysis for reporting the uncertainty of qualitative analytical results, and presents application examples of the described theory. The Guide does not discuss the test item's ability to represent a group of identical items or a larger object; that is, it does not discuss the impact of sampling in these assessments.

Ordinal results can be reduced to binary (yes/no) outcomes and treated using the methods in this guide by assigning ordinal classification results as 'correct' or 'incorrect'. Other methods for treating ordinal scales are outside the scope of this guide.

# 2 Types of qualitative analysis

As mentioned in the scope of this Guide, Qualitative Analysis[2] is defined as *"Classification according to specified criteria"* [28]. Table 1 lists some examples. Although all these cases appear very different, they share one common characteristic; once the criteria are specified, the performance of the classification methodology is relatively simple to describe in terms of its success or failure rate. These success and failure rates form the basis of most qualitative analysis performance metrics.

Qualitative analyses covered in the main text are divided into two categories based on different types of classification criteria, i.e., qualitative or quantitative. Table 1 presents examples of each. Section 3 describes strategies for assessing performance for different types of classification criteria. For qualitative analysis where the true or false response rates depend on a quantitative property, such as the presence of a banned substance whose detection depends on the amount present, a limit of detection is also considered (see section 3.4).

Conformity assessment of the value of a quantitative property of an item with a limit value or interval can sometimes be considered as a conversion of a measurement result into a qualitative result ('conforming' or 'non-conforming'). The use of measured values and their measurement uncertainties for conformity assessment is covered in detail in another Eurachem/CITAC guide [29], and is accordingly not considered in detail in the present Guide. However, Annex B discusses how some of the metrics used to assess the performance or uncertainty of qualitative analysis can be determined for quantitative conformity assessment.

**Table 1.** Types of qualitative analysis based on different types of classification criteria.

| Classification criterion | Qualitative analysis example |
|---|---|
| Qualitative | **(1)** Detection of aliphatic aldehydes in a solution by colour change after the addition of Schiff's reagent.<br>**(2)** Identification of the crystalline form of material by observation.<br>**(3)** Identification of the brand and year of wine by sensory analysis.<br>**(4)** Identification of a biological species by determining or detecting a particular DNA sequence.<br>**(5)** Identification of human blood type by observation of agglutination. |
| Quantitative | **(1)** Identification of a pesticide residue in fruit using measured fragment masses and relative fragment abundances in GC-MS.<br>**(2)** Determination of infrared spectral equivalence between a new and a previously accepted industrial raw material, using wavelength and intensity criteria.<br>**(3)** Identification of a diuretic in urine from an athlete using retention time and measured fragment masses in GC-MS.<br>**(4)** Identification of a drug in blood using retention time and measured fragment masses in LC-MS.<br>**(5)** Detection of a virus in a clinical sample based on fluorescence intensity in quantitative real-time polymerase chain reaction (qPCR). |

[2] 'Qualitative testing' is in principle a wider field than 'qualitative analysis', simply because chemical analysis, frequently referred to as analytical work, is one particular activity among many fields of testing. However, in this Guide for analytical chemists and those in related disciplines, the terms are used synonymously.

# 3 Performance assessment for qualitative analysis

## 3.1 General considerations

This section provides guidance on the assessment and expression of performance for procedures intended to provide simple classification into two classes (‘binary classification’). The ‘Classes’ are labelled here as ‘positive’ or ‘negative’ to denote ‘member of the class of interest’ or ‘non-member’. This classification covers most practical situations, including ‘above a limit’, ‘acceptable’, ‘unacceptable’, ‘identity as’, or ‘presence of a particular species’.

Classes are assumed to be comprehensive and exclusive to allow calculation of unambiguous false response rates. This implies that no test item may be classified as a member of a third class. This can generally be achieved by careful specification of classification criteria. However, it remains possible that a result may not provide sufficient confidence in classification. Under these circumstances, it is entirely reasonable for the analyst to report a test result as ‘inconclusive’ in the sense of insufficiently certain. Inconclusive results require further study to report results as ‘conclusive’. These results are recognised in medical laboratories to be from a ‘grey zone’ or ‘equivocal zone’.

Some of the concepts described can, in principle, be extended to more classes, such as in an ordinal scale classification, by assessing correct and incorrect classification rates for all classes. A useful extension is to treat identification of structure or identity (formally a multi-class problem) as either ‘correct’ or ‘incorrect’, and that approach is assumed here. A detailed treatment of the multi-class problem, which may involve multiple simultaneous assignments or assignment to several classes, is, however, beyond the scope of this Guide.

Qualitative analysis involves various stages namely, (1) problem description, (2) method development and (3) validation, (4) tests on unknown items checked through quality control, and (5) reporting of results (Figure 1). Unambiguous specification of the property to be determined and assessing the fitness of the analysis for the intended use are critical. The reporting of a qualitative analytical result must be supported by valid procedures and adequate quality control of the test. How results are reported depends on the purpose of the analysis and the report recipient. This Guide does not detail how the method should be developed or how quality control should be designed.

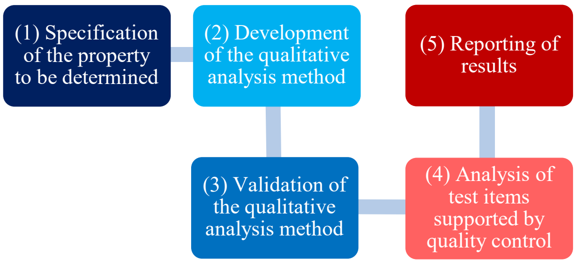
(1) Specification of the property to be determined
(2) Development of the qualitative analysis method
(5) Reporting of results
(3) Validation of the qualitative analysis method
(4) Analysis of test items supported by quality control

**Figure 1.** Qualitative analysis process from problem description to the reporting of results.

## 3.2 Quantification of qualitative analysis performance

### *3.2.1 Defining the basis for performance assessment*

The most basic way of quantifying the performance of a qualitative analysis method is by calculating false result rates. With ‘positive’ or ‘negative’ results, it is useful to report ‘true positive’ and ‘false positive’ or ‘true negative’ and ‘false negative’ rates, respectively. However, these rates can be referenced to either the total number of a specific type of case or result, or the total number of possible cases or results.

For instance, the false positive rate can be defined as:

i) The fraction of negative cases that are falsely reported as positive ($fp/nc$), where $fp$ and $nc$ are the numbers of false positive results and negative cases, respectively. Figure 2 graphically represents the overlapping of different types of cases and results. The ($fp/nc$) is represented by the ratio of the areas of the intersection (∩) of positive results, ‘$p$’, with ‘$nc$’ ($p \cap nc = fp$) and the area of ‘$nc$’. The *FP* of Table 2 presents this determination.

ii) The fraction of positive results that are falsely reported as positive ($fp/p$), where $p$ is the number of positive results. In Figure 2,

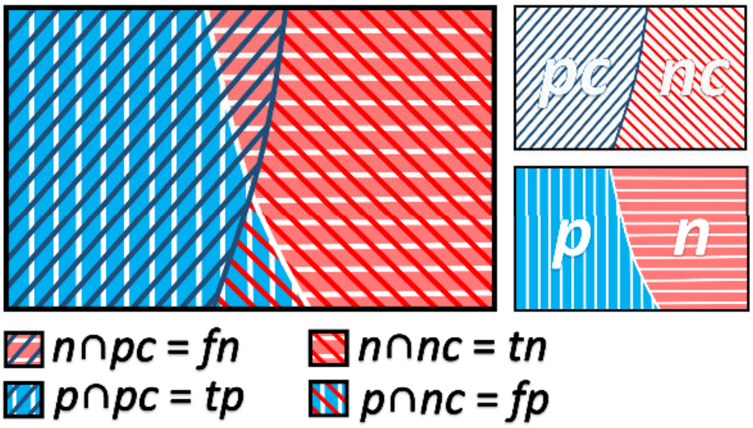

**Figure 2.** Graphical representation of an example of the overlapping of the number of positive, *pc*, or negative, *nc*, cases with the number of positive, *p*, or negative, *n*, results. The symbol "∩" represents the intersection of groups; for example *n*∩*pc*, here, denotes the set of negative results from positive cases. The '*n*∩*pc*', '*p*∩*pc*', '*n*∩*nc*' and '*p*∩*nc*' define *fn*, *tp*, *tn* and *fp*, respectively.

this rate is represented by the ratio between the areas (*p*∩*nc* = *fp*) and '*p*'.

iii) The fraction of the total number of cases or results that are falsely reported as positive ($fp/(pc + nc) = fp/(p + n)$), where *pc* and *n* represent the number of positive cases and negative results, respectively. In Figure 2, this rate is represented by the ratio between the area labelled (*p*∩*nc* = *fp*) and the figure's total area.

The difference between these definitions is crucial. In case i), the rate does not vary with the proportion of 'negative cases' in the population, i.e., $nc/(nc+pc)$, since $FP = fp/nc$. However, for cases ii) and iii), the false positive rate depends on $nc/(nc+pc)$ since more *fp* are observed in populations containing more *nc*. Therefore, these definitions characterise the performance of the qualitative analysis in different ways and, hence, involve different interpretations of their values.

The true positive, *TP*, (*tp*/*pc*) and the true negative, *TN*, (*tn*/*nc*) rates referenced to a relevant number of cases are known in clinical chemistry as qualitative analysis 'sensitivity' and 'specificity', respectively [7] (Table 2). The determination of clinical sensitivity and specificity requires the proper determination of studied cases by a conclusive clinical diagnosis. For quantitative analysis, the term 'sensitivity' [1] or 'analytical sensitivity' [30] has a different meaning.[3]

The true positive rate referenced to positive cases (*tp*/*p*) is also known as the qualitative analysis

**Table 2.** Alternative performance characteristics for expressing the quality of qualitative analytical results.

| Performance characteristics | Expression |
|---|---|
| True positive rate, *TP* (Sensitivity, *SS*) | $tp/pc = tp/(tp + fn) = 1 - FN$ |
| False positive rate, *FP* | $fp/nc = fp/(tn + fp) = 1 - TN$ |
| True negative rate, *TN* (Specificity, *SP*) | $tn/nc = tn/(tn + fp) = 1 - FP$ |
| False negative rate, *FN* | $fn/pc = fn/(tp + fn) = 1 - TP$ |
| 'Precision' or 'Positive predictive value', *PPV* | $tp/p = tp/(tp + fp)$ |
| 'Negative predictive value', *NPV* | $tn/n = tn/(tn + fn)$ |
| Efficiency, *E* | $(tp + tn)/(p + n)$ |
| Youden Index, *Y* | $SS(\%) + SP(\%) - 100$ |
| Likelihood ratio of positive results, $LR(+)$ | *TP*/*FP* |
| Likelihood ratio of negative results, $LR(-)$ | *TN*/*FN* |
| Posterior probability | See Annex A |

*tp* – number of true positive results; *fp* – number of false positive results; *tn* – number of true negative results; *fn* – number of false negative results; *p* – number of positive results (*tp* + *fp*); *n* – number of negative results (*tn* + *fn*); *pc* – number of positive cases and *nc* – number of negative cases.

[3] According to the International Vocabulary of Metrology [1], the 'sensitivity of a measuring system' is the "quotient of the change in an indication of a measuring system and the corresponding change in a value of a quantity being measured".

‘precision’ or ‘positive predictive value’, *PPV* [30]. The term ‘negative predictive value’, *NPV*, is used for the true negative rate referenced to the total number of negative results (i.e., *tn*/*n*). The efficiency of the qualitative analysis is defined as the fraction of any type of correct results given all results (i.e., (*tp* + *tn*)/(*p* + *n*)). The Youden Index is an alternative way of quantifying the success of qualitative analysis (Table 2) [31].

Although the metrics referenced to the number of positive or negative cases do not depend on the prevalence of the case types, these numbers alone cannot provide the probability that a specific result is correct. To estimate the probability of a result being correct, a relevant result rate and prevalence of cases also need to be considered. This, and other metrics for confidence in qualitative results, is discussed in Section 4.

### *3.2.2 Defining performance assessment reference*

The metrics used to quantify qualitative analysis performance can have additional peculiarities. The positive and negative cases can be established in different ways. Some cases or samples used as references can be known to be 'positive' for a characteristic because of their origin, or through formulation. Others might be, as defined by AOAC International, cases where results from “a confirmatory technique and another analytical technique are both positive” [32]. Some examples of adequate origins of positive cases can be patients diagnosed with a particular disease or soil known to be contaminated. Positive test items can be formulated by adding the species to be identified in a matrix equivalent to the analysed items, such as a pesticide in a food product confirmed or not confirmed as having native levels of the pesticide. If identification performance varies significantly with a quantitative property (for example, the concentration of the substance to be identified or detected) the formulation should allow for the determination of that level. A negative case can also be interpreted as a case known to be negative from its origin, formulation or defined as negative since a “confirmatory technique and another analytical technique are both negative”. The AOAC International definitions of positive and negative cases have a more comprehensive application since it is the only approach applicable to the analysis of complex items that are challenging to reproduce from formulation. However, it relies on the quality of the output of the analytical techniques used. In some fields, it is difficult to artificially prepare items with the studied analyte and possible interferents for analysis performance testing, because the matrices of the items are unknown and unpredictable.

The positive and negative cases can also be provided as reference data, such as spectra known to be from a specific compound. After defining identification criteria, the probability of reporting the composition match correctly or incorrectly compared to these criteria can be determined. For instance, in mass spectrometry, the identification can be based on assessing the presence or presence and abundance of characteristic ions. The chance of spectroscopic match can be predicted by binomial or hypergeometric statistics as discussed in Examples E1 and E2.

### *3.2.3 Method performance reporting*

#### *3.2.3.1 Contingency tables*

A very convenient way of reporting the performance of a qualitative method of analysis that does not vary significantly within the analytical scope is through a contingency table. Table 3 presents an example of such a table. In this example, the *TP*, *FP*, *TN* and *FN* are 97.8 % (228/233), 0.33 % (1/301), 99.7 % (300/301) and 2.1 % (5/233), respectively.

Typically, the analytical scope can involve different levels of the studied species or property and various matrices of the analysed item. This can require separate contingency tables for different parts of the analytical scope.

**Table 3.** A specific example of a contingency table that describes the performance of a qualitative analytical method that should be approximately constant within the analytical scope.

| | | **Case** | | |
|---|---|---|---|---|
| | | **Positive (*pc*)** | **Negative (*nc*)** | **Result totals** |
| **Result** | **Positive (*p*)** | *tp* = 228 | *fp* = 1 | *p* = 229 |
| | **Negative (*n*)** | *fn* = 5 | *tn* = 300 | *n* = 305 |
| | **Case totals** | *pc* = 233 | *nc* = 301 | |

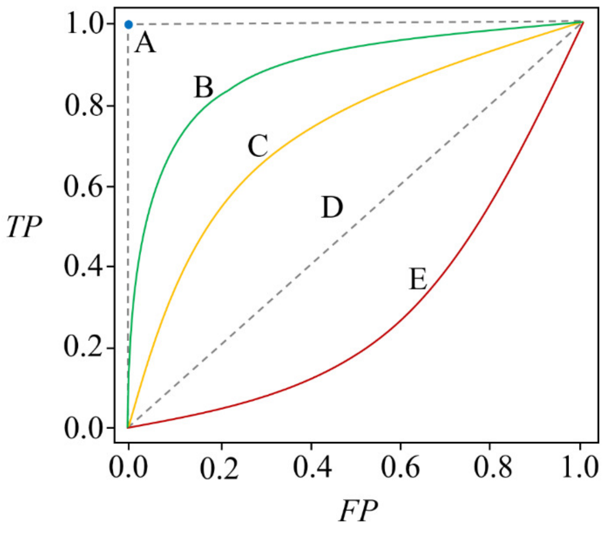

**Figure 3.** Five examples of ROC curves where the variation of *TP* and *FP* with the variation of quantitative identification criteria is represented. Curve A (blue point) presents a perfect test where identification criteria do not affect results, and *TP* and *FP* are 100 % and 0 %, respectively. Curve B and C represent suitable methods, where $TP \geq FP$. Among these three, Method B is preferable to method C. Curve D represents the chance diagonal where $TP = FP$ for all decision thresholds; this would not be a useful classifier. Curve E, at first sight, seems to be very poor classifier, consistently showing a false positive rate larger than the true positive rate. However, a simple switch of reported outcome would generate an ROC curve near that of C; the classifier might then prove useful.

*3.2.3.2 Receiver Operating Characteristic (ROC) curve*

For qualitative analysis based on the assessment of a quantitative characteristic, the selection of the classification criteria that balances the true and false result rates, typically *TP* and *FP*, can be performed using Receiver Operating Characteristic (ROC) curves, which plot the pair (*TP*, *FP*) as a classification criterion (i.e., a discrimination threshold) varies. These curves can also be used to compare different qualitative analysis procedures [31]. Although the detailed description of these curves is beyond the scope of this Guide, Figure 3 presents five schematic examples of ROC curves. Each curve shows how the true positive rate and false positive rate vary as the identification criterion varies from a stricter to a less stringent identification of positive cases associated with a low or high *TP*, respectively. In addition to providing a visual illustration of performance, the area under the curve (often abbreviated “AUC”) can be used as a summary of classifier performance [31].

## 3.3 Evaluating false positive and false negative rates

*3.3.1 Method scope and validation detail*

The validation of a qualitative analysis method involves defining performance requirements and checking if they are met [30].

Before this performance assessment, the analysis scope should be clearly defined in terms of the type of classification (e.g., presence of pentachlorophenol above 1 mg kg$^{-1}$) and analysed items (e.g., leather products). The classification method should also be specified, namely, the analytical technique (e.g., GC-MS/MS), how this technique is used (e.g., sample preparation and instrumental conditions) and the classification criteria. The classification criteria must be clearly described to guarantee that collected performance data will apply to subsequent analyses.

In some qualitative analysis, driven by efficiency considerations, the analytical method is divided into two stages: a preliminary faster and cheaper screening method that, whenever required, is followed by a more time consuming and expensive confirmatory method. Confirmation is performed when the first assessment produces results contrary to those expected or can have a relevant impact on an individual or collective interest. However, it is essential to assess the false negative and positive rates for the entire procedure that includes the screening and confirmatory tests. For instance, when only positive results are subject to confirmation, it is essential to check if the screening stage's false negative rate is adequately low.

Regarding the method validation detail, for methods applicable to a diversity of items (e.g., different food products), performance should be tested for a representative set of types of items. The types and number of tested items depend on the impact of the analysed matrix on the performance. In some cases, understanding the classification principles can allow groups of items associated with equivalent qualitative analysis performance to be anticipated, from which a representative item can be selected and studied. The performance of the analysis of the representative item can then be extrapolated to the group of items associated with equivalent qualitative analysis performance. If the performance of the classification technique allows, the decision can be taken to study the performance of the analysis of items and/or property values where the false result rates reach the highest values. The laboratory should manage the thoroughness of the performance assessment while keeping in mind the available time and resources for this assessment. In some cases, it can be acceptable to execute an on-going validation strategy where every time an item that is new to the laboratory is tested, additional and specific controls of the quality of the analysis are performed.

### *3.3.2 Using information from the literature*

For commonly used qualitative analysis procedures, performance information might be expected to be in the public domain. Before embarking on studying the performance of a well-established analytical procedure, an appropriate study of the relevant literature in the field should be performed to gather independent information on its fitness for the intended use. However, published false response rates should be used with caution; they could have been obtained using specific equipment, reagents, and personnel and refer to specific sample matrices and characteristic levels, so the analyst must consider whether his/her situation is equivalent. For instance, if the items studied in the literature have characteristic levels far from the thresholds used to distinguish between classes and if their matrices are relatively free from interferences, the determined identification performance can be too optimistic compared to the "real" analytical problems experienced by the laboratory. Therefore, true and false result rates depend heavily on available data.

In some cases, it is possible to anticipate if performance observed in the literature will be better or worse than the performance observed for qualitative analysis in the laboratory. If it is concluded that the qualitative analysis procedure is valid for worst-case scenarios, i.e., can produce results fit for their intended purpose, the procedure can be used to analyse unknown items with no restrictions.

Section 4 discusses how criteria for deciding whether an analysis is fit for the intended use can be set.

### *3.3.3 Assessment exclusively from experimentation*

Regardless of the type of qualitative analysis mentioned in section 2, the *FP* and *FN* can be estimated directly from the number of false results from a set of analysis. In qualitative analysis based exclusively on qualitative inputs (Table 1), this is the only way of estimating qualitative analysis uncertainty. However, if false responses are unlikely, this approach requires a large number of tests.

Given that the number of false responses should ideally be low, the problem arises of how many samples to test to be reasonably sure of finding a non-zero number of false responses.

From published information (see, for example, Ferrara et al. [33]), it is evident that false positive or negative rates can be as low as 0.5 % and in some cases even lower [6, 8, 9]. For a range of false result probabilities, Table 4 shows the number of samples that would need to be analysed in order to be certain, to at least within the confidence levels indicated, of finding one or more false result. Table 4 uses the binomial distribution and shows that for a 95 % chance of detecting one or more false results, the number of tests that need to be performed will be three times more than the number of tests that produce an average of one false result. For instance, for a method with a 1 % false positive rate, it is found that (on average) for each 100 analyses of negative cases, one positive result is observed. However, to be "95 % sure" that a false positive result is observed 299 (about

**Table 4.** The minimum number of analyses to find one or more false (positive or negative) result(s).

| | Confidence level | |
|---|---|---|
| False result rate | 95 % | 99 % |
| 0.5 % | 598 | 919 |
| 1 % | 299 | 459 |
| 5 % | 59 | 90 |

3 × 100) tests on negative cases must be performed.

The values of Table 4 are not sufficient for a good estimate of false result rates or to compare different methods. Even an approximate estimate would usually require five to ten times the minimum number of observations given in Table 4. This table can also be interpreted as the minimum number of analyses necessary to check compliance with different acceptable false result rates, as discussed below.

In an attempt to determine false result rates directly from experimentation for a new procedure, the analyst is frequently faced with a dilemma. On the one hand, for a given procedure, the false response rate of interest is unknown, and therefore any performed classifications can be unreliable. On the other hand, merely analysing until the first false response occurs would not necessarily give a true picture of the false response rate. To address this problem, it is suggested that the analyst decides in advance on tolerable levels for the two false response rates. For a chosen confidence level, the binomial distribution may be used to estimate the number of experiments needed to find one or more false response(s) with enough confidence. This approach is not guaranteed to produce an exact figure for the false response rate, but it will place a bound on it. For example, suppose the analyst decides that a *FP* of 5 % is acceptable, and after performing 59 experiments (Table 4), covering the likely range of matrices, no false positives are found. In that case, it may be concluded that the *FP* is not greater than 5 %. As a quality control measure of the validated procedure, it is further recommended that the samples be interspersed with blanks and reference materials containing the target characteristic (e.g., analyte) at relevant characteristic levels. It should always be remembered that false result rates depend very much upon the vagaries and/or specificities of the population being sampled and upon this population's sampling strategy.

Table 4 shows that, for low false response rates, it may be impractical to analyse a sufficient number of samples to detect a false response. Accordingly, if a test is inexpensive and/or is intended to be widely used, e.g., as a drugs screening test, it can be acceptable to first establish that the false response rate does not exceed an upper limit, say 5 %, by experiment, and then to refine this figure in the light of experience with further samples.

Where sample numbers are likely to be relatively low and/or the test is expensive to apply, all tests should be run in parallel with a confirmatory test and, from time to time, the false response rates should be recalculated.

The mathematical processing of available information can be used to overcome some limitations of the experimental determination of false response rates (see sections 3.3.4 and 3.3.5).

### *3.3.4 Assessment from a database*

An alternative to determining false results' rates from experimentation is chance mismatch studies in reference databases, such as mass spectra or infrared spectra databases. In some cases, this allows the equivalent of many thousands of experiments. However, though informative and powerful, a current limitation is that such databases are often quite unrepresentative of the testing population; for example, while the prevalence of different materials in general use varies widely, a typical reference database will only contain one of each. This may lead to significantly biased probability estimates; again, the values obtained are unlikely to be better than order-of-magnitude estimates. Examples E1 and E2 illustrated the use of this methodology for assessing qualitative analysis performance.

### *3.3.5 Assessment from quantitative data modelling*

The assessment of the performance of highly selective and time-consuming and/or expensive qualitative analyses exclusively from experimentation performed in a single laboratory is not feasible.

In qualitative analysis based on a quantitative classification criterion for quantitative results (such as an instrumental method of analysis), models of the dispersion of results can be used to estimate true and false result rates. Annex B gives further details. For instance, if the relevant instrumental signal, such as analyte retention time in a chromatographic method, is normally distributed the chance of an interfering component having a retention time within the acceptance retention time interval for the analyte can be predicted (See Quick reference 1, page 13).

However, the modelling relies on the validity of the model assumption and input variable values. For instance, since relative retention times can be not normally distributed, the assumption of normality can underestimate false result rates. The simulation

**Quick reference 1 – Signal modelling example**

If for deltamethrin identification in olive oil by GC-MS, the estimated standard deviation of the retention time repeatability, $s_{t\mathrm{R}i}$, is 0.022 min with $\nu = 32$ degrees of freedom, the retention time tolerance for identifying this compound in a sample can be: ($t_\mathrm{R} \pm t \cdot s_{t\mathrm{R}i}$) = ($t_\mathrm{R}$ ± 2.04 × 0.022) = ($t_\mathrm{R}$ ± 0.045) min; where $t_\mathrm{R}$ is the retention time observed from a single daily injection of a standard solution and *t* the 2-sided 95 % critical value for the *t*-distribution with 32 degrees of freedom. Therefore, for $t_\mathrm{R}$ of 36.055 min, the acceptance interval for a sample peak would be (36.055 ± 0.045) min. Assuming an interferent has a retention time 0.05 min less than for deltamethrin and the precision of both retention times is equivalent, the probability of the interferent having a retention time within the acceptance interval would be 1.5 %. This value is estimated by the cumulative *t*-distribution for a *t*-value of (-0.05/0.022) and *ν* (MS-Excel formula: T.DIST(-0.05/0.022;32;TRUE)).

of instrumental signals by the Monte Carlo method is a convenient way of estimating *FP* and *FN* from non-normally distributed parameters [8, 9]. Example E5 illustrates the modelling of instrumental signal dispersion for estimating the false result rate of highly selective GC-MS/MS identifications.

### 3.3.6 *Assessment of qualitative test performance dependent on a continuous variable*

Many confirmatory or detection tests show a strong dependence on detection probability or false response rate on some continuous variable. For example, detection rates often depend on the concentration or the number of particles of the material sought. It may then be valuable to model the dependence of false response rate on the continuous (or other) variables.

Logistic regression and probit regression [34, 35] are commonly applied to such problems and have been suggested (with examples) for the performance assessment of qualitative methods of analysis [36]. Logistic regression has been demonstrated in low copy number DNA detection [37]. The procedure is well documented in textbooks and available in essentially all statistical software packages, therefore, it is not presented in detail here. Simple logistic regression models the probability of a binary response as a function of some continuous variable. The model is:

$$p = \frac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)} \qquad (1a)$$

Or, equivalently:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x \qquad (1b)$$

where *p* is the probability of interest (for example, probability of a positive result), *x* the continuous variable (usually concentration of the analyte) and $b_0$ and $b_1$ the regression coefficients. Most statistical packages will provide a fitting method either from raw data (concentration/qualitative result pairs) or from proportions calculated from the number of results. Note that the former only requires a sequence of yes/no (or 1/0) values; it does not require proportions. This makes it possible to apply the method to a range of test samples with different (known or independently measured) concentrations which are subjected to the qualitative test procedure only once each.

Once a relationship is established, it becomes possible to estimate detection limits (see below) from the fitted relationship between concentration and probability of detection, simply by choosing an appropriate limit for the probability of detection which corresponds to the definition of the detection capability in use.

Example E4 provides a practical example of logistic regression.

### 3.3.7 *Expert judgement*

When no data on the performance of the analytical method is available from a third party, and it is not possible to assess performance exclusively from experimentation (section 3.3.3) or modelling (sections 3.3.4 and 3.3.5), the analyst can use his/her practical experience in the classification technique, for the studied or similar items, to decide if the method is fit for the intended use.

Whenever possible, the decision on the fitness for purpose of a method for an intended use should be supported by objective evidence.

The process of formulation of expert judgements is the topic of several studies. The judgements are influenced by many factors leading to corresponding estimates of the uncertainty of the result [38].

## 3.4 Limit of detection and selectivity

### *3.4.1 Limit of detection*

The Limit of detection (LOD) typically describes the lowest concentration of a substance that leads to reliable detection. For tests where the classification involves the assessment of a quantitative characteristic, and the value of this characteristic affects the qualitative results, the 'limit of detection' (LOD) and/or the 'limit of quantification' (LOQ) considered in the qualitative and/or quantitative analysis should be checked in relation to qualitative analysis performance [30]. The qualitative analysis result should be fit for the intended use at that level(s).

NOTE: The Commission Regulations (EU) No 589/2014 [39] and No 152/2009 [40] define an LOQ as "the lowest content of the analyte that can be measured with reasonable statistical certainty, fulfilling the identification criteria as described in internationally recognised standards" [41].

For exclusively qualitative analysis, the LOD can be found by applying the procedure to items containing progressively smaller levels of the characteristic until the likelihood of producing false results reaches a pre-established criterion. Logistic and probit regression can also be used in this type of assessment (Section 3.3.4).

### *3.4.2 Selectivity*

Selectivity, in the sense in which this term is usually employed in analytical chemistry, refers to "the extent to which a particular method can be used to determine analytes under given conditions in the presence of other components of similar behaviour" [42]. The International Vocabulary of Metrology (VIM) defines this term equivalently as a measuring system property [1].

NOTE. The term 'specificity', in the context of quantitative analysis, is used for perfectly selective analysis [42, 43], which can only be claimed in chemistry on very rare occasions. However, there is a clear alternative use of the term 'specificity' in the context of qualitative analysis (see Table 2). In this Guide, the term 'selectivity' is used in a general sense and the term 'specificity' reserved for the purpose noted in Table 2.

Selectivity can be assessed by analysing one or more test items having known or likely interfering characteristics, that is, characteristics that are not the target of the analysis but might be considered likely to generate the test response.

It is sometimes possible to identify interfering species or scenarios which are particularly likely to generate false positive results. For example, tests for ammonia might reasonably be expected to respond to primary amines, and tests for specific bacterial strains might be expected to respond to any bacteria of the same general species.

If the qualitative analysis performs relatively well in worst-case situations, it can be concluded that the procedure is valid for all types of items.

Although the false response rate can be measured for each different material or each interfering species present, selectivity studies are unlikely to generate a single definitive value for selectivity. This is because the response depends on the potential cross-reacting species included in the study and on the level of these species. Therefore, selectivity studies are best considered as providing a broad indication of the adequacy of the qualitative analysis method when faced with different challenges.

# 4 Expressions of confidence in qualitative analysis

## 4.1 General considerations

While statements of measurement uncertainty in quantitative analysis typically result in a range of values, like an expanded uncertainty interval or a minimum purity, classification statements cannot usually be associated with a range. In general, one cannot report that the material is 90 % of a ‘pass’, that an analyte is 99 % present or that a chemical species is within some contiguous sequence. Instead, the typical form of uncertainty information is probabilistic in nature. That is, one gives an indication of the probability of a given classification being correct, or of typical probabilities of misclassification of items whose correct class is known.

The performance figures that can be obtained from validation studies can be reported with the qualitative test result. In general, however, they rarely give direct information about (for example) the probability that a qualitative result is correct. In this section, two metrics that have been proposed for this purpose are described, with a view to aid understanding and improve the state of the art in expressing uncertainty for qualitative analysis results. The metrics presented here use variants of Bayes’ rule [4] (See Annex A). These can be used to give a) an indication of the strength of evidence, provided by one or more qualitative result(s), in favour of one possible classification over another; b) in conjunction with sound information about the probabilities of encountering different (true) values of qualitative characteristics in a population, an indication of the probability that a particular classification is true given a particular qualitative analysis result.

## 4.2 Likelihood ratio

The most familiar and widely used form of reporting qualitative analysis performance is false result rates, particularly $FP$ and $FN$ or their complementary rates, $TN$ and $TP$, respectively (e.g., $TN = 1 - FP$). Two of these rates can be conveniently combined into the same performance characteristic: the likelihood ratio, *LR*.

If a positive result is reported, the *LR*(+) is estimated by Eq. (2):

$$LR(+) = TP/FP \qquad (2)$$

The *LR*(+) is a ratio of two probabilities; the probability of reporting a positive result if the case is positive divided by the probability of reporting a positive result if the case is negative. Broadly, the likelihood ratio gives a measure of the change in the probability that the sample is genuinely positive, after seeing a positive test result. Mathematically, the likelihood ratio is the change

**Quick reference 2 – Interpretation of likelihood ratio**

If a positive result is reported, the probability of the case being, in fact, positive, $PP$, is calculated by Eq. (Q2.1) (see below). This equation is the well-known Bayes’ theorem (Annex A), substituting true and false positive rates for conditional probabilities.

$$PP = \frac{P(+)TP}{P(+)TP + P(-)FP} \qquad \text{(Q2.1)}$$

where $P(+)$ is the probability of the case being positive prior to the test. This can also be expressed in “odds” form (see Annex A):

$$\frac{\mathrm{PP}}{1 - \mathrm{PP}} = \frac{P(+)TP}{P(-)FP} \qquad \text{(Q2.2)}$$

In Q2.2, the ratio $P(+)/P(-)$ represents the odds in favour of a positive case before applying the qualitative test; that is, the 'prior odds'. The ratio $TP/FP$ is the calculated likelihood ratio $LR(+)$.

The likelihood ratio therefore describes how the probability (represented by the odds) changes after a positive test result; it can be thought of as a measure of the additional information provided by the test.

In the special case where $P(+) = P(-) = 0.5$, so that the prior odds are 1.0; the *LR*(+) then represents the ratio of posterior probabilities of the case being positive or negative. For instance, with equal prevalence (or assumed prevalence) of positive and negative cases, a positive result with a $LR(+)$ of 1000 would indicate that the posterior probability of the case being genuinely positive is 1000 times larger than the probability of the case being negative.

in probability expressed as “odds” (see Annex A). A high likelihood ratio from a test indicates that the test item is more likely to be positive than could be said before carrying out the test. Sometimes, this is interpreted as a ‘weight of evidence’, contributed by the positive test result, in favour of the test item being genuinely positive.

In the special case where both positive and negative cases are equally likely ($P(+) = P(-) = 0.5$; where $P(+)$ and $P(-)$ are the prevalence of positive and negative cases, respectively), the *LR*(+) can be understood to indicate how much more the reported positive result is likely to be true rather than false (see Quick reference 2).

For example, if positive and negative cases are considered equally likely prior to the test, a positive result associated with a $LR(+)$ of 7300 means that the positive result is 7300 times more likely to be true than false.

If a negative result is reported, the $LR(-)$ is:

$$LR(-) = \frac{TN}{FN} \tag{3}$$

For the special case of equally probable positive or negative cases prior to examination, the $LR(-)$ represents how much more a negative result is likely to be true rather than false.

Some authors combine both likelihood ratios in the parameter ‘Diagnostic odds ratio’, $DOR$ ($DOR = LR(+)/LR(-)$) [30].

One of the most useful features of the likelihood ratio ($LR(+)$ or $LR(-)$) is that if the classification depends on two independent pieces of evidence (i.e., the result is only reported when two independent analyses from independent procedures confirms it), the $LR_{(1\&2)}$ of the outcome of both analyses is estimated by multiplying the *LR* that quantifies the uncertainty of each piece of evidence ($LR_{(1)}$ and $LR_{(2)}$):

$$LR_{(1\&2)} = LR_{(1)} \cdot LR_{(2)} \tag{4}$$

For instance, if the presence of a contaminant in a food product, determined by GC-MS, is based on retention time with a $LR(+)$ of 99.9 and mass spectrum data with a $LR(+)$ of 490, the $LR(+)$ of identifications based on both these tools becomes $4.9 \times 10^4$ (i.e., 99.9×490). The Eq. (4) results from the fact that the probability of the convergence of two independent results is estimated by multiplying the respective individual probabilities.

If *m* independent pieces of evidence are considered (*i* = 1 to *m*) to report a positive or a negative result, i.e., a result is only reported if indicated by the *m* pieces of evidence, the *LR* from the combined pieces of evidence is estimated by Eq. (5).

$$LR = \prod_{i=1}^{m} LR_{(i)} \tag{5}$$

where Π denotes the product of a sequence of variables and $LR_{(i)}$ is the likelihood ratio from the *i*-th qualitative analysis ($LR_{(i)}(+)$ or $LR_{(i)}(-)$).

When the pieces of evidence are not independent, Eq. (5) will underestimate or overestimate the joint probability. Quick reference 3 shows how non-independent probabilities can be combined.

Likelihood ratios can be challenging to interpret, especially for non-specialists. For forensic applications, the scale in Table 5 has been recommended [44] to give a verbal indication of the strength of evidence. According to this table, collected evidence is considered “extremely strong” only if the *LR* is larger than $10^6$. In principle, this kind of approach can be adapted for other circumstances if a general indication of the strength of evidence is required. For instance, for identifying the polymer type of microplastics collected from sediments in environmental

**Quick reference 3 – Probability for non-independent pieces of evidence**

The probability of two independent events A and B, $P(\mathrm{A}\cap\mathrm{B})$, occurring is estimated by Eq. (Q3.1).

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A})P(\mathrm{B}) \tag{Q3.1}$$

where $P(\mathrm{A})$ and $P(\mathrm{B})$ are the probabilities of events A and B occurring, for instance, producing a positive result from analysing a positive case (i.e., a *TP*).

However, if $P(\mathrm{A})$ and $P(\mathrm{B})$ are associated, probability of the coincidence of both events is determined by Eq. (Q3.2), which involves the conditional probability of event B occurring given that event A has occurred:

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A})P(\mathrm{B}|\mathrm{A}) \tag{Q3.2a}$$

Or, equivalently:

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{B})P(\mathrm{A}|\mathrm{B}) \tag{Q3.2b}$$

For direct or inverse correlations, associated with $r_{\mathrm{AB}} > 0$ or $r_{\mathrm{AB}} < 0$, respectively, $P(\mathrm{A} \cap \mathrm{B})$ will be respectively greater than or smaller than for cases where A and B are independent.

monitoring, criteria presented in Table 5 are too strict. For these analyses, results associated with a $LR(+)$ greater than 19 should be adequate (i.e., with a $TP \geq 95\,\%$ and $FP \leq 5\,\%$) since contamination is determined after identifying many particles from several samples [45].

Although the determination of a binary property can only produce one of two results, if the most likely result (e.g., yes or no) is associated with a low *LR*, the decision can be made to report a result as inconclusive instead of reporting the verbal equivalent of Table 5. For instance, the decision can be made to report a positive or negative result if the respective *LR* is larger than (for example) $10^5$, with lower *LR* reported as inconclusive. This "grey zone" for the *LR* value can be set below $10^5$, 19, or any other values depending on the purpose of the analysis. The reporting of a result as inconclusive is useful if both false positive and false negative results have a relevant impact. When testing for doping substances in an athlete's urine, false positives are much more serious than false negative results suggesting that if no evidence of doping is observed, the result can be reported as negative (i.e., no evidence of doping) [9, 46]. However, both false positive and false negative results can be a problem for maternity or paternity identification, suggesting that a positive match with a low *LR* should not be reported as a conclusive "no-match" [47].

## 4.3 Posterior probability

If there is reliable information about the prevalence of a particular characteristic (e.g., a population with a well documented prevalence of a particular disease), the *LR*(+) associated with a test result can be converted into the probability *PP* that the tested item is positive, given the positive test result. This is known as a posterior probability, and is estimated using Bayes' theorem (Annex A). One form of this, using the likelihood ratio, is:

$$PP = \frac{\frac{P(+)}{P(-)}LR(+)}{\frac{P(+)}{P(-)}LR(+) + 1} \qquad (6)$$

Here, $P(+)$ and $P(-)$ are prior probabilities, i.e., information available before the test, and *PP* and *PN* are posterior probabilities.

Taking the previous example of the analysis of a contaminant in a food product by GC-MS, where a positive result is associated with a $LR(+)$ of $4.9\times10^4$, assuming $P(+) = P(-) = 0.5$, the *PP* becomes 99.998 % ($PP = 4.9\times10^4/(4.9\times10^4+1)$).

If a negative result is reported, the posterior probability of the sample being genuinely negative, *PN*, is estimated by:

$$PN = \frac{\frac{P(-)}{P(+)}LR(-)}{\frac{P(-)}{P(+)}LR(-) + 1} \qquad (7)$$

This equation is similar to Eq. (6). The $P(+)$ and $P(-)$ express the prevalence of positive or negative cases.

Broadly, since the posterior probability relates to a reported classification, the posterior probability can be thought of as a measure of the probability that the reported value is correct.

Posterior probabilities can be difficult to apply in practice. Sometimes, sufficiently relevant and reliable prior probabilities are not available. Although some authors have suggested that this concern can be overcome by assuming positive and negative results are equally probable, so that $P(+)/P(-) = 1$, this is not always sensible. Sometimes, particularly in forensic work, it may be inappropriate to infer prior probabilities for a particular case from knowledge of unrelated cases. In such cases, a likelihood ratio (section 4.2) can provides a useful summary of the confidence provided by a test result, without the need to determine prior probabilities.

In some fields, such as medical sciences, the consideration of the prevalence of a condition or characteristic in decisions on qualitative analytical results can be crucial for diagnosis. The diagnosis of a disease or clinical situation based on clinical analytical results will also rely on additional information such as mucosal colour, location and intensity of pains, age and gender, risk behaviour, etc. The way this information contributes to the final decision on the observed result can be illustrated by calculating a *PP* or *PN*, although clinicians do not routinely perform these calculations; rather, they are expected to be aware of the general importance of prevalence when making a diagnosis based on a test result.

More details about these metrics are presented in the bibliography [4, 8, 9].

## 4.4 Reliability of metrics

The reliability of the calculated *LR*, *PP* or *PN* depends on the reliability of the considered result rate and, for posterior probabilities, on the reliability of any prior probability used. Table 4 presents the number of tests required for reliable detection of one or more false responses at different probabilities of false response. Reliable *quantitation* of such a rate generally requires many more (see section 3.3.3). The number of studied cases may need to be further increased to cover the complete scope of the test method; for example, in testing food, it may be necessary to examine multiple different food matrices. The modelling of an instrumental signal considered in qualitative analysis can make the quantification of low false result rates feasible but depends on the adequacy of the input data and the modelling algorithm.

For the determination of *PP* or *PN* from very discrepant $P(+)$ and $P(-)$, Table 4 can be used to define the number of cases (from the target population) that should be studied.

The input data quality for estimating these metrics is even more critical when various pieces of evidence are combined, and metrics quantifying the strength of the combined information calculated.

Therefore, the presented metrics should be used with caution, keeping in mind relevant details of the input data, how the reported result is used, and the respective consequences. Over-interpretation of qualitative analysis performance data can be as harmful as ignoring the limitations of a particular qualitative analysis.

## 4.5 Uncertainty of proportions

The statistical quality of the estimated result rate that depends on the number of tests used for their determination can be expressed as a confidence interval, CI, for the calculated rate. This confidence interval is also known as "condition uncertainty" (4.4.6 of [18]), being typically calculated for the 95 % confidence level (95 % CI).

For instance, a wide 95 % CI for sensitivity *SS* indicates that the "true" value of the *SS* could be very different from the estimate. The same logic can be applied to other result rates, such as *SP*. Since the result rates are not estimated from any prior knowledge of the population of cases, these intervals only characterise the estimated analytical performance quality.

The interpretation of 95 % CI is, to some degree, similar to what happens with the expanded measurement uncertainty [1]. For a 95 % CI, there is a 5 % probability that the "true" value of the result rate is outside the CI limits. Similarly, the 95 % CI for an experimentally determined rate provides the statistical uncertainty for the calculated rate.

For example, if the ability of a qualitative analysis method to correctly identify positive cases is tested from the analysis of 400 of such cases and all 400 results are positive, the estimated *SS* of 100 % is associated with a 95 % CI bounded between 99 % and 100 %; i.e., the true value of the *SS* varies between 99 % and 100 % with 95 % confidence. If the method is tested with only 5 positive cases, the 95 % CI of the *SS* will be limited by 57 % and 100 %. The 95 % CI allows the quality of the analytical method performance parameters to be expressed, which is required for their sound interpretation. In the examples above, both *SS* of 100 % are reported, but the *SS* estimate is much more reliable in the first case. The calculation of 95 % CI for *SS* and *SP* is a standard practice in the clinical laboratory (10.1.3 of [27]).

Several models have been published to compute the CI [48] –[49, 50, 51, 52, 53, 54]. The Wilson score interval [54] was used for simplicity and applicability to small counts. Equations 8 and 9 can be used to calculate the low, $LL_{SS.95}$, and high, $HL_{SS.95}$, limits of 95 % CI for the *SS* or *TP*.

$$LL_{SS.95} = \frac{A_1 - A_2}{A_3}\,100 \qquad (8)$$

$$HL_{SS.95} = \frac{A_1 + A_2}{A_3}\,100 \qquad (9)$$

where:

$$A_1 = 2\,tp + 1.96^2$$
$$\mathrm{A}_2 = 1.96\,(1.96^2 + 4\,tp \cdot fn \,/\, (tp + fn))^{1/2}$$
$$A_3 = 2\,(tp + fn + 1.96^2).$$

Equations 10 and 11 are used to calculate the low, $LL_{SP.95}$, and high, $HL_{SP.95}$, limits of 95 % CI for the *SP* or *TN*.

$$LL_{SP.95} = \frac{B_1 - B_2}{B_3}\,100 \qquad (10)$$

$$HL_{SP.95} = \frac{B_1 + B_2}{B_3}\,100 \qquad (11)$$

where:

$$B_1 = 2\,tn + 1.96^2$$
$$B_2 = 1.96\,(1.96^2 + 4\,fp \cdot tn \,/\, (fp + tn))^{1/2}$$

$B_3 = 2\ (fp + tn + 1.96^2)$.

A target or minimum value of $LL_{SS.95}$ or $LL_{SP.95}$ (i.e., $LL^{tg}_{SS.95}$ or $LL^{tg}_{SP.95}$) should be defined according to the purpose of the analysis. The target is particularly critical when the impact of false results is critical. For example, for blood components used in transfusions, the screening for infectious diseases should be performed with tests associated with a $LL_{SS.95}$ close to 100 % which can only be confirmed if many positive cases are tested during validation.

When the result rate is compared to a target minimum value or when either an increase or decrease in the parameter is being investigated, a one-tailed assessment should be performed. For a 95 % confidence test, the factor 1.96 should be changed to 1.64.

# 5 Reporting the qualitative analytical result

Currently, accredited laboratories are not required to report qualitative analysis results with uncertainty. The examples in this section are accordingly intended to suggest possible reporting approaches when a laboratory chooses to do so to assist a customer.

A positive result can be reported with the $TP$ and $FP$, $LR(+)$ or $PP$ and a negative result with the $TN$ and $FN$, $LR(-)$ or $PN$. The other metrics presented in Table 2 can also be used to report confidence in the result. These metrics typically provide information about an individual test result. However, for cases, where the value of a metric is constant for the analytical scope, such parameters can be interpreted as characterising the analytical method.

The following four examples illustrate how qualitative results can be reported with the respective performance or uncertainty.

**Example 1** (*the italic text mentions the qualitative analysis uncertainty*)**:**

Mrs A. B. is infected with SARS-CoV-2 virus.

*(test with a sensitivity of* 90 % *and a specificity of* 99 %*)*

**Example 2** (*the italic text mentions the qualitative analysis uncertainty*)**:**

The urine of Mr C. D. contains canrenone residues

*(identification with a likelihood ratio of* $4.9\times10^4$*)*

**Example 3** (*the italic text mentions the qualitative analysis uncertainty*)**:**

Cocaine is present in sample 123

*(identification with a likelihood ratio of* $4.9\times10^4$ *and considered 'very strong' evidence of analyte presence)*

**Example 4** (*the italic text mentions the qualitative analysis uncertainty*)**:**

Gasoline residues were identified in the fire debris with sample code 456

*(identification with a posterior probability of* 99.998 %*, estimated from signal model simulation and assuming analyte presence or absence are equally probable)*

# 6 Conclusions and recommendation

It is important for laboratories to check at least the most critical false response rate. For some metrics, both the false positive and false negative rates must be established.

It is realistic to expect that most laboratories have the relevant parameters of their qualitative analysis procedures (i.e., conditions of analysis) under adequate control. Evidence of that will typically involve:

- clear evidence of the adequate metrological traceability of values of parameters subject to control due to their relevance for the test;
- evidence that uncertainties of these parameters are sufficiently small for the purpose.

It is reasonable to expect laboratories to be following published codes of best practice in qualitative analysis where they are available, including the use of appropriate reference data and materials.

Quantitative (i.e., numerical) reports of uncertainties in qualitative test results should be presented in a way that avoids misinterpretation.

Whenever it is concluded that the obtained analytical result is associated with too low and too high true or false result rates, respectively, it is entirely reasonable to report the test result as ‘inconclusive’ in the sense of insufficiently certain.

# 7 Examples

Examples are described after their scope is listed.

## 7.1 E1: Identification of compounds by low-resolution mass spectrometry using database searching or the presence of characteristic ions

### *7.1.1 Introduction*

This example is divided into Case A or B, where different procedures are used to identify compounds in complex matrices by low-resolution mass spectrometry. The parallel presentation of the two cases highlights the alternative nature of the identification options.

Note in practice identification usually used multiple criteria such as a combination of mass spectra match and chromatographic retention time. This example focuses only on the mass spectrometry component. Example E5 gives an example of multiple criteria.

| **Scope:** |
|---|
| **Type of qualitative analysis:** Analysis based on quantitative criteria |
| **Item/matrix:** A) Meat products and B) forensic or environmental samples |
| **Parameter/analyte:** A) Diethylstilboestrol, DES (forbidden growth hormone for beef and poultry meat) or B) Heroin, DES and dichlorodiphenyltrichloroethane (DDT) |
| **Type of classification criterion:** 1) Identification based on tolerances for the relative abundances (*RA*) of specific ions of the mass spectrum; 2) Identification based on the presence of specific ions of the mass spectrum regardless of the *RA* values. |
| **Technique/instrumentation:** Gas-chromatography hyphenated with low-resolution mass spectrometry using electron impact ionisation (GC-MS) |
| **Type of results reporting:** Likelihood ratio |

This example describes the evaluation of the uncertainty for the identification of compounds by GC-MS using different identification criteria (sections 7.1.1 and 7.1.2). The examples present results for the identification of three compounds (i.e., DES, Heroin and DDT) in two types of samples (meat products and forensic or environmental samples).

Mass spectrometry, particularly in combination with a chromatographic separation stage, is a powerful tool that can help identify unknown compounds. For most purposes, low-resolution mass spectrometry using electron impact (EI) ionisation is the method of choice when identification, instead of quantification, is required. A mass spectrum can contain many ions, not all of which are useful for diagnostic purposes. This raises the question of whether there is a minimum number of ions that would be sufficient to ensure an unequivocal identification.

NOTE: In some analytical fields, the minimum number of ions required is defined for identifying compounds [20] − [21][22][23].

### *7.1.2 Identification based on the relative abundance of characteristic ions*

Sphon [55] investigated the minimum number of ions that need to be monitored to produce an unambiguous identification of diethylstilboestrol[4] (DES) in meat products. Data related to a subsequent study [56], based on a commercial mass spectral library containing about 270 000 entries, is presented in Table E1.1.

Table E1.1 shows the number of spectra in the library used that match specified criteria for the relative abundance, *RA*, of one or more ions. The *RA* is estimated by dividing the abundance of the studied ion by the abundance of the most abundant ion (i.e., the base peak). This normalisation aims at producing an identification parameter less dependent on analyte level (e.g., analyte concentration). It can be observed in Table E1.1 that when more ions and narrower abundance ranges are considered, the number of matches occurring in the

[4] DES was used as growth hormone for cattle and poultry, and subsequently banned after its carcinogenic properties were proven.

**Table E1.1.** Number of spectra of a Wiley library, with 270 000 entries, matching specific criteria for the relative abundances of some ions.

| Identification criterion | | | Matches |
|---|---|---|---|
| # | Ion (*m*/*z*) | $RA$ (%) acceptance interval | |
| 1 | 268 | 1 – 100 | 9 995 |
| 2 | 268<br>239 | 1 – 100<br>1 – 100 | 5 536 |
| 3 | 268<br>239 | 90 – 100<br>10 – 90 | 46 |
| 4 | 268<br>239 | 90 – 100<br>50 – 70 | 9 |
| 5 | 268<br>239<br>145 | 90 – 100<br>50 – 90<br>5 – 90 | 15 |
| 6 | 268<br>239<br>145 | 90 – 100<br>50 – 70<br>45 – 65 | 1 (DES)[a] |

$RA$: Relative abundance (percentage of the most abundant ion, the base peak)

[a] The single match corresponds to the mass spectrum of DES

database is reduced dramatically. The identification criteria set #6 isolates the mass spectrum of DES, leading to a single match.

For comparison with an alternative library, Table E1.2 presents the number of matches from a publicly available reference library, then containing 62 235 spectra, considering tolerances for the relative abundance ($RA$) of one or more ions [57]. As the tolerances associated with the $RA$ of more ions become narrower, the mass spectra of fewer compounds are isolated. Table E1.2 presents the number of spectra matching three different target compounds, namely DES, heroin and DDT. Heroin and DDT are relevant for the analysis of some forensic and environmental samples, respectively.

The comparison of the isolation of the mass spectrum of DES in both libraries, described in Tables E1.1 and E1.2, allows for the conclusion that, as expected, the number of matches depends on the number of spectra in the library (see identification criteria #1 and #2 in Tables E1.1 and E1.2). If the number of matches is divided by the total number of entries, the differences observed in Tables E1.1, and E1.2 reduces.

Table E1.3 summarises the information collected in Tables E1.1 and E1.2 from the most selective identifications. Table E1.3 converts the collected information into $TP$ and $FP$, further combined into a $LR(+)$ that estimates the uncertainty of a positive result (i.e., reporting analyte presence).

The estimated $TP$ (i.e., approximately 100 %) assumes that the defined tolerances for the $RA$ of ions take their variability into account. Ideally, the tolerances should be set from signal variability models built from replicate spectra of sample solutions with relevant analyte concentrations [8, 9] (Example E5).

The $FP$ presented in Table E1.3 assumes possible interferents are all compounds whose mass spectrum is available in the used library. Many compounds present in the sample solutions will not be detectable by GC-MS or eliminated in sample preparation. On the other hand, many compounds whose spectrum is in the library are not likely to be in analysed samples due to chemical incompatibility or independence of sources or origins. A worst-case $FP$ equivalent to one per total number of spectra, $N$, minus 1 ($FP = 1/(N-1)$) is estimated since it is known that $FP$ will not be zero. The described limitations of how $FP$ was evaluated should be considered when using this value. The $FP$ can be alternatively estimated from models of signal noise, as discussed in Example E5 [8, 9].

From the data in Table E1.3, it can be seen that by using the same library, all the analytes with a single match have the same $TP$, $FP$ and $LR(+)$, depending only on the number of library entries. According to the criteria defined by the European Network of Forensic Science Institutes (Table 5), the identification of an analyte by

**Table E1.2.** The number of spectra in a publicly available library, with 62 235 entries, matching specific criteria for the relative abundances of some ions.

| Identification criterion | | | |
|---|---|---|---|
| # | Ion (*m*/*z*) | *RA* (%) acceptance interval | Matches |
| 1 | 268 | 1 – 100 | 3597 |
| 2 | 268<br>239 | 1 – 100<br>1 – 100 | 1597 |
| 3 | 268 | 55 – 95 | 83 |
| 4 | 268<br>239 | 55 – 95<br>30 – 70 | 4 |
| 5 | 268<br>239<br>145 | 55 – 95<br>30 – 70<br>60 – 100 | 1 (DES) |
| 6 | 369 | 1 – 100 | 1672 |
| 7 | 369<br>327 | 1 – 100<br>1 – 100 | 526 |
| 8 | 369 | 45 – 85 | 43 |
| 9 | 369<br>327 | 45 – 85<br>60 – 100 | 1 (Heroin) |
| 10 | 352 | 1 – 100 | 1242 |
| 11 | 352<br>235 | 1 – 100<br>1 – 100 | 234 |
| 12 | 352 | 1 – 40 | 1140 |
| 13 | 352<br>235 | 1 – 40<br>60 – 100 | 1 (DDT) |
| 14 | 352<br>235<br>237 | 1 – 40<br>1 – 100<br>48 – 88 | 7 |
| 15 | 352<br>235<br>237 | 1 – 40<br>60 – 100<br>48 – 88 | 1 (DDT) |

*RA*: Relative Abundance (percentage of the most abundant ion, the base peak)

[a] The single match corresponds to the mass spectrum of DES, heroin or DDT

mass spectrometry, using a described identification procedure, produces 'Very strong' evidence of analyte presence ($LR(+)$ between $10^4$ and $10^6$). Suppose the identification is also supported by the analyte's retention time in the chromatographic system (i.e., in the GC) and the retention time window is adequate. In that case, the $LR(+)$ of identification can increase (Example E5).

### *7.1.3 Identification based on the presence of characteristic ions – estimated chance match probabilities*

Suppose, instead of identifying the analyte by using tolerances for the *RA* of specific ions, the identification is based on the simple presence of three selective ions. In that case, the following mathematics can be used for a rough evaluation of the identification uncertainty. Suppose a low-resolution mass spectrometer is used that can only distinguish *m/z* units (i.e., mass-to-charge ratios units) and ions must have *m/z* values between 180 *m/z* and 480 *m/z*. In that case, approximately 300 possible *m*/*z* values can be observed in a spectrum. Therefore, since the number of combinations of 300 objects taken three at a time is $300!/[3! \cdot (300-3)!]$ = 4 455 100, for three peaks, the probability of three random peaks matching the three chosen ions by chance would, assuming that all *m/z* ratios are equally likely, be $1/4.6\times10^6$, or approximately $2.2\times10^{-7}$. This, however, does not allow for the fact that most mass spectra typically have many more than three ions in the region of interest;

**Table E1.3.** Uncertainty of identifying several analytes in different matrices, by GC-MS, estimated from the number of matches of spectra from the library used, taking tolerances for the relative abundances of two or three ions into account.

| Analyte | Analysed item | Number of spectra of the library, *N* | Number of matches [a] | *TP* (%) [b] | *FP* (%) [c] | *LR*(+) (*TP*/*FP*) |
|---|---|---|---|---|---|---|
| DES | Meat products | 270 000 | 1 in *N* | ~100 | $3.7\times10^{-4}$ | $2.7\times10^{5}$ |
| DES | F&E | 62 235 | 1 in *N* | ~100 | $1.6\times10^{-3}$ | $6.2\times10^{4}$ |
| Heroin | F&E | 62 235 | 1 in *N* | ~100 | $1.6\times10^{-3}$ | $6.2\times10^{4}$ |
| DDT | F&E | 62 235 | 1 in *N* | ~100 | $1.6\times10^{-3}$ | $6.2\times10^{4}$ |

DES: diethylstilboestrol; F&E: Forensic and environmental samples; *TP*: True positive rate; *FP*: False positive rate; *LR*(+): Likelihood ratio (*TP*/*FP*) (see Table 2).

[a] Number of matches considering the defined tolerances for the *RA* of specific ions.

[b] Optimistic estimation of *TP* (approx. 100 %).

[c] Estimated as 1/(*N*-1) (worst-case scenario from collected information).

this increases the probability that a mass spectrum containing *n* ions might match the chosen three by a factor of $n!/[3!\cdot(n-3)!]$. Taking ten as a typical number of peaks, the chance match probability increases by $10!/3!\cdot 7! = 120$. The estimated chance match probability is accordingly $120\times2.2\times10^{-7}$ or approximately $2.6\times10^{-5}$. This provides an approximate false positive rate, *FP*.

Assuming contaminant levels are high enough to provide reliable mass spectra, it can be considered that *TP* is approximately 1, or 100 %.

The estimated *TP* and *FP* above can be combined to give an *LR*(+) of $1/2.6\times10^{-5} = 3.8\times10^{5}$ that indicates an upper bound for the likelihood ratio of for identifications based on the described procedure.

NOTE. This estimate assumes all ion combinations are equally likely and possible; and that their appearances are independent; this is known to be a rough approximation (see Example 7.2). The effective chance match probability is therefore likely to be very much higher than the figure calculated, leading to a lower likelihood ratio.

### *7.1.4 Final remarks*

The methodologies for estimating the *LR*(+) of identifications by GC-MS presented in this example tend to be over optimistic about the validity of qualitative analytical results (see Example 7.2 for a comparison with actual match probabilities). Therefore, these calculations should only be used as an initial assessment of identification validity. Example E5 discusses alternative and more realistic determinations of the uncertainty of identifications performed by GC-MS/MS.

Although the methodologies for estimating chance match probabilities and likelihood ratios presented here may be optimistic, it is still possible to conclude that, for example, the identification of heroin in samples from a crime scene based solely on the presence of ions with *m/z* of 369 and 327 would be inadequate since it is associated with an *LR*(+) of 118 ($118 = TP/\{FP\} = 100/\{[526/(62235\text{-}1)]\cdot100\}$) (Methodology of section 7.1.1: case 7 of Table E1.2) or $4.5\times10^{4}$ ($4.5\times10^{4} = 100/\{[1/(44850\text{-}1)]\cdot100\}$; where $44850 = 300!/[2!\cdot(300-2)!]$) (Methodology of section 7.1.2) depending on the approach used for evaluating confidence in the results. In practice, this would indicate that additional criteria, or further confirmatory tests, would be required to provide adequate confidence.

Similarly, although Sphon [55] and others suggested that veterinary drug residues in cattle can be identified by taking three mass spectrum ions, for the official monitoring of unauthorised substances, the European Union (EU) requires the collection of additional evidence of the presence of these compounds [20]. For instance, if it is only possible to monitor two characteristic ions by GC-MS at appropriate levels of the analyte, two independent chromatographic runs must be considered based on electron impact or chemical ionisation to confirm the presence of the analyte [20].

## 7.2 E2: Identification of purified compounds by infrared spectrometry

| **Scope:** |
|---|
| **Type of qualitative analysis:** Analysis based on quantitative criteria |
| **Item/matrix:** Purified chemical compound |
| **Parameter/analyte:** A compound from the available infrared spectra database |
| **Type of classification criterion:** Match of the wavenumber of three or six bands of the infrared spectra in the interval [500, 1800] cm$^{-1}$ |
| **Technique/instrumentation:** Infrared spectrometry |
| **Type of results reporting:** Likelihood ratio |

Several authors have investigated the use of database statistics in evaluating criteria for qualitative analysis. De Ruig et al. [58] proposed criteria to be met to identify veterinary drug residues in meat products (see section 7.1.2). The authors give indicative values of chance match probabilities based on a simple binomial model. Ellison et al. have shown that a hypergeometric distribution was a more appropriate model for chance matching in spectra because it allows for a small number of matching peaks between two spectra, both of which contain a larger number of peaks [5]. The latter authors focused on chance matches when an infrared spectrum is compared to a spectral library.

Ellison et al. [5] studied the reliability of the identification of purified compounds by comparing the obtained infrared spectrum with spectra from a library. The library used by Ellison et al. was the Sadtler library containing spectra from 59 626 different materials. A random subset of thirty compounds were selected from this library and the number of bands, *m*, in the interval [500, 1800] cm$^{-1}$ noted for each compound. It was determined that the average number of bands per spectrum in the interval [500, 1800] cm$^{-1}$, *M*, was 16. The spectral resolution available was 4 cm$^{-1,}$ and this implied the existence of 1300/4 = 325, *N*, discrete peak positions in the [500, 1800] cm$^{-1}$ interval. For each different spectrum in the chosen subset, the entire database was searched twice – first for a minimum of $n = 3$ matching peaks and the second time for a minimum of $n = 6$ matching peaks.

For $n \geq 3$, the number of observed matches was about twice the number predicted by the hypergeometric distribution. For $n \geq 6$, although the number of matches was considerably lower, as would be expected, the

**Table E2.1.** Chance matches against six bands in an infrared database.

| Compound | No. of bands *m* in range | Chance match probability [a] | Predicted matches [b] | Observed matches | *LR*(+) |
|---|---|---|---|---|---|
| 1-Chloro-3-(1-napthyloxy)-2-propanol | 23 | $3.19 \times 10^{-4}$ | 19 | 192 | 311 |
| α-Cyano cinnamic acid, methyl ester | 17 | $5.03 \times 10^{-5}$ | 3 | 29 | 2056 |
| Phenyl ν-triazolo-[1,5-α]-pyridin-3-yl ketone | 24 | $4.19 \times 10^{-4}$ | 25 | 190 | 314 |
| Benzo-β-thiophene-6-acrylic acid | 20 | $1.34 \times 10^{-4}$ | 8 | 52 | 1147 |
| 3-((Dipropylamino)methyl)1-5-nitroindole | 17 | $5.03 \times 10^{-5}$ | 3 | 29 | 2056 |
| 2-Mesityl-5-phenyl-oxazole | 22 | $2.52 \times 10^{-4}$ | 15 | 99 | 602 |
| *p*-Hydroxy-benzoic acid | 18 | $6.71 \times 10^{-5}$ | 4 | 44 | 1355 |
| Caproic acid, isobutyl ester | 8 | $1.36 \times 10^{-7}$ | 0 | 1 | 59626 |
| 1-Bromoadamantane | 10 | $9.64 \times 10^{-7}$ | 0 | 1 | 59626 |
| Phenyl propyl ether | 17 | $5.03 \times 10^{-5}$ | 3 | 47 | 1269 |

[a] The chance match probability is the probability that $n = 6$ peaks match by chance across two spectra of *m* bands, assuming random incidence of bands across the spectrum. The probability was calculated using the hypergeometric distribution (ref [5]).

[b] The predicted number of matches is the chance match probability multiplied by the number of spectra in the database

observed matches exceeded the predicted matches by a factor of ten. Part of the data for six peaks matched is presented in Table E2.1.

The calculated chance match probabilities for six-peak matches were in the interval [$10^{-7}$, $10^{-5}$]. The chance match probability for a compound, when multiplied by the number of entries in the database, estimates the number of compounds fitting the search criteria. In the case of two of the compounds in Table E2.1, *viz.* caproic acid isobutyl ester and 1-bromoadamantane, the search criteria produce a single match and hence would appear to be adequate if these compounds are suspected. Many more matches are produced for the remaining compounds, which indicates a requirement for more stringent criteria.

Assuming that $TP$ is approximately 100 % (because IR spectra of pure compounds reliably match their own reference spectra) and taking $FP$ as the ratio between the number of observed matches and the total number of spectra of the library (i.e., 59 626), it is possible to estimate the $LR(+)$ of the identification. The last column of Table E2.1 presents the estimated $LR(+)$ (calculated as $TP/FP$ , with $TP = 1.0$). The reported $LR(+)$ are much lower than the $10^6$ minimum value considered to classify evidence as ‘Extremely strong’ (Table 5), suggesting that a simple six-peak match on wavelength alone might not provide sufficient confidence without additional criteria. Additional criteria (such as additional peak intensity matching, absence of peaks not present in the target compound, matching chromatographic retention time, or a requirement for close visual match to complete spectra) might be needed to provide sufficiently unambiguous identification.

This example stresses that reference databases, of which spectral libraries are one type, can only obtain indicative information on false response rates. The response rates are only strictly reliable for populations similar to the population of samples expected. It is also the analyst's responsibility to decide which, if any, of a set of matches corresponds to an unknown.

## 7.3 E3: Identification of drugs of abuse in urine by the enzyme multiplied immunoassay technique (EMIT) and an alternative technique

**Scope:**

| |
|---|
| **Type of qualitative analysis:** Analysis based on quantitative criteria (studied using qualitative information) |
| **Item/matrix:** Urine |
| **Parameter/analyte:** Cocaine, methadone or opiates |
| **Type of classification criterion:** Not specified |
| **Technique/instrumentation:** Enzyme multiplied immunoassay technique, EMIT, and an alternative, proprietary, technique. |
| **Type of results reporting:** Likelihood ratio and the probability of the positive result being correct |

Although EMIT determinations involve the processing of an instrumental signal, this example assesses the performance of this qualitative analysis from experimentally determined rates of false positive and false negative results. Therefore, this example illustrates the determination of the quality of a qualitative analysis based on qualitative information.

The use of sample result databases for obtaining the relevant probabilities for a Bayesian assessment of qualitative analysis performance has been reported in the literature. To test for drugs of abuse in urine, Ferrara *et al.* [33] assembled a database containing information on drug types, analytical techniques, false response rates for the techniques, and prevalence of the drugs. For the cited authors' laboratory, Table E3.1 summarises part of this data for EMIT. The table also presents the estimated posterior probability of the test sample being genuinely positive, *PP*, as described in Eq. (6). For the calculation, the prevalence of negative results is taken as (1 – *P*(+)), the *LR*(+) = *TP*/*FP* and the *TP* = 1 – *FN*.

**Table E3.1.** Probabilities for EMIT detection of drugs of abuse in urine [33].

| | Values of performance characteristics for different drugs or drug classes | | |
|---|---|---|---|
| Probability | Opiates | Methadone | Cocaine |
| *P*(+) | 0.44 | 0.26 | 0.20 |
| *FP* | 0.028 | 0.004 | 0.009 |
| *FN* | 0.069 | 0.018 | 0.056 |
| *PP* | 0.963 | 0.988 | 0.963 |

*P*(+) - Prevalence of positive results.

For instance, for identifying methadone, *PP* is evaluated by Eq. (E3.1) as 0.988 (*PP* is the posterior probability determined from the Bayes' theorem). In other words, the analyst could be over 98 % certain that a positive response for methadone genuinely indicates this drug's presence. However, note that this depends in part on the high observed prevalence in the particular population sampled. *PP* might be very much lower in the general population, i.e., a population where drug use is not so frequent.

$$PP = \frac{\frac{P(+)}{P(-)}LR(+)}{\frac{P(+)}{P(-)}LR(+) + 1} = \frac{\left(\frac{0.26}{1-0.26}\right)\left(\frac{1-0.018}{0.004}\right)}{\left(\frac{0.26}{1-0.26}\right)\left(\frac{1-0.018}{0.004}\right) + 1} = 0.988 \qquad \text{(E3.1)}$$

Table E3.2 shows similar data for a different, non-immunochemical technique. Note that the *FP* for cocaine by this technique is reported as zero. However, it is debatable whether the false response rates for such screening tests can indeed be zero. In this case, no false positives were found, but had more samples been analysed, one or more false positives could have appeared. An estimated false response rate has accordingly been used in calculating the posterior probability.

**Table E3.2.** Probabilities for the detection of drugs of abuse in urine by the proprietary technique [33].

| | Values of performance characteristics for different drugs or drug classes | | |
|---|---|---|---|
| Probability | Opiates | Methadone | Cocaine |
| $P(+)$ | 0.44 | 0.26 | 0.20 |
| $FP$ | 0.038 | 0.012 | 0.000 |
| $FN$ | 0.276 | 0.179 | 0.247 |
| $PP$ | 0.937 | 0.960 | 0.995 § |

§ *PP* calculated using an estimated worst-case *FP* value equal to 0.001 (one in 1000 tests).

Considering methadone again, the *PP* is 0.960. This is a reasonably high probability, though slightly less convincing than that produced by the EMIT test.

To illustrate the effect of combining data, suppose that both screening tests were performed. If a positive response is obtained in each case, then the combined *PP* becomes 0.9998 (see Eq. (E3.2) based on the combination of Eq. (5) and (6).

$$PP = \frac{\frac{P(+)}{P(-)}LR(+)}{\frac{P(+)}{P(-)}LR(+)+1} = \frac{\left(\frac{0.26}{1-0.26}\right)\left(\frac{1-0.018}{0.004}\right)\left(\frac{1-0.179}{0.012}\right)}{\left(\frac{0.26}{1-0.26}\right)\left(\frac{1-0.018}{0.004}\right)\left(\frac{1-0.179}{0.012}\right)+1} = 0.9998 \quad \text{(E3.2)}$$

In this example, reliable prevalence values (i.e., prior probabilities) are available. Had these not been at hand, or if the analyst had preferred not to use them, likelihood ratios could have been used instead; the corresponding values being 246 (EMIT) and ~68 (proprietary), giving a combined likelihood ratio of approximately 17 000 (Eq. (E3.3)) that, according to Table 5, corresponds to a ‘Very strong’ evidence of the presence of methadone.

$$LR(+) = \left(\frac{1-0.018}{0.004}\right)\left(\frac{1-0.179}{0.012}\right) = 1.7 \times 10^4 \quad \text{(E3.3)}$$

In all cases, GC-MS was used as a reference technique to establish false result rates. The particular database referred to here is quite comprehensive for the studied analytes and has been designed to provide a representative collection, permitting a Bayesian analysis of the data. There are inevitably some missing values but, as more data is added, these should be reduced in number and the accuracy of predictions further improved.

A further advantage of a database set up to record representative data for several different techniques is the information it provides to enable one to optimise analytical performance. For example, selecting a screening method with a low false positive rate should minimise the costs of expensive confirmatory analyses. However, other factors also need to be considered, such as the limit of detection of the technique, its false negative rate, and the speed and cost of analysis.

## 7.4 E4: Identification of human SRY gene in biological material by qPCR

| **Scope:** |
|---|
| **Type of qualitative analysis:** Analysis based on quantitative criteria (fluorescence exceeding threshold) |
| **Item/matrix:** Biological material |
| **Parameter/analyte:** SRY gene (sex-determining region Y) |
| **Type of classification criterion:** Not specified |
| **Technique/instrumentation:** Quantitative polymerase chain reaction (qPCR) |
| **Type of results reporting:** True positive rate. |

Although the interpretation of qPCR involves the processing of an instrumental signal, in this example, performance is assessed from determining *TP* at different DNA concentrations.

Figure E4.1 shows some experimental data from a study of human SRY gene detection, by quantitative PCR (qPCR), in biological material [37]. The data were generated from a 3-plate assay using a 5'-nuclease assay with a dual labelled fluorogenic 'TaqMan' probe directed towards the human genome's male-specific SRY region. A result was considered 'positive' if the observed fluorescence exceeded a predetermined threshold within 55 amplification cycles. Though different preparative methods were used as part of the validation study, statistical tests showed no significant differences, so the data were treated as a single data set.

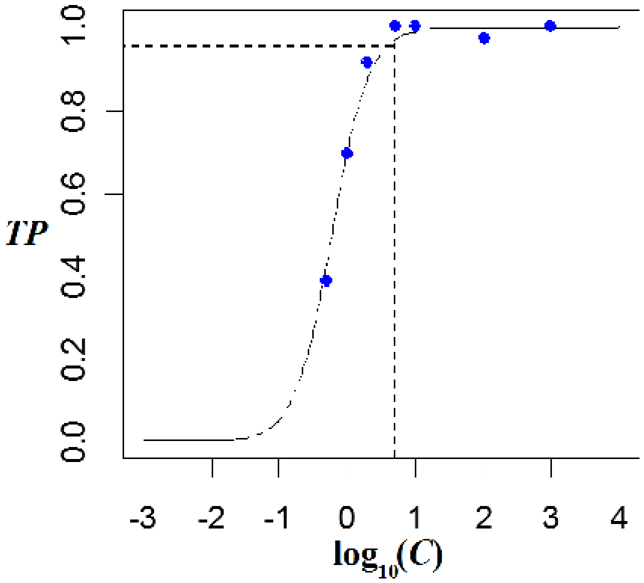

**Figure E4.1.** DNA detection data processed by logistic regression. The *TP* is plotted as a function of $\log_{10}(C)$, where *C* is the copy number of the gene, in a study of DNA detection and classification capability. Points show the proportion of positive results from a total of 36 replicates at each copy number. The solid line shows the logistic regression fit with $b_0 = 0.85$ and $b_1 = 3.75$ (see Section 3.3.6). The dashed line shows $TP = 0.95$ and the corresponding log concentration of $\log_{10}(C) = 0.56$, for $C = 3.6$ copies.

Notice that the point at $\log_{10}(C) = 2$ does not show 100 % *TP*, though points either side do, making it hard to accurately assess the classification capability. However, the regression curve allows a reasonably precise location of the effective limit of detection, as shown in the figure; using a continuous model has effectively smoothed the random count data. This is a considerable advantage; it allows for the analyst to study large numbers of concentrations with relatively few replicates per level, instead of requiring many replicates at few levels, and still obtain relatively reliable probability estimates.

This data set illustrates an important caveat in modelling. The concentration data is plotted and modelled in the log domain; a common practice in working with experimental DNA concentrations or microbial counts. For a dependent variable, transformation is often dictated by the error distribution. However, there is no compelling reason to choose log-transformation for the independent variable in this instance; the choice is essentially arbitrary. So, too, is the logistic model choice; other models may also fit the data reasonably well. Where different, but smooth, models fit the data similarly, interpolation is not strongly sensitive to the choice of model. However, extreme probabilities can be very sensitive indeed to the choice of model. It follows that even where a model provides a good description of the data and relatively reliable probabilities and limits of detection within that range, it is very unsafe to extrapolate probability estimates much beyond the range studied without substantial evidence of model validity.

## 7.5 E5: Identification of pesticide residues in foodstuffs by GC-MS/MS based on retention time and ion abundance ratio

| **Scope:** |
|---|
| **Type of qualitative analysis:** Analysis based on quantitative criteria |
| **Item/matrix:** Foodstuffs of vegetable origin |
| **Parameter/analyte:** Chlorpyrifos-methyl and malathion |
| **Type of classification criterion:** Intervals for retention time and abundance ratios of characteristic ions |
| **Technique/instrumentation:** GC-MS/MS |
| **Type of results reporting:** Likelihood ratio and the probability of the positive result being correct |

This example discusses the estimation of *FP* of highly selective determinations of pesticide residues in foodstuffs by GC-MS/MS through modelling the instrumental signal using the Monte Carlo Method. Monte Carlo Method simulations were performed in an MS-Excel spreadsheet.

The identification of analytes is based on the retention time, $t_{\mathrm{R}}$, in the chromatographic system and on the abundance ratio, $AR = A_1/A_2$, of two characteristic ions of the analyte's mass spectrum. Although the $t_{\mathrm{R}}$ is approximately normally distributed, the *AR* can deviate significantly from normality. The ratio of correlated variables is known not to be normally distributed, particularly if the variable with lower precision (i.e., larger dispersion of values) is in the denominator [8].

The development and validation of a procedure for the identification of trace levels of the analytes in the foodstuffs by GC-MS/MS starts with the definition of the qualitative analysis method, including the sample preparation and GC-MS/MS conditions. The specification of the GC-MS/MS conditions includes selecting characteristic ions of the analyte's mass spectrum (chlorpyrifos-methyl ions: 208 *m/z* and 271 *m/z*; malathion ions: 99 *m/z* and 127 *m/z*). Afterwards, replicate injections of analyte stock solutions and foodstuff extracts are performed. The injections of analyte stock solutions are used to study the dispersion of $t_{\mathrm{R}}$ and ions abundances, $A_1$ and $A_2$, and the correlation of ion abundances of each analyte. Performance data were collected at various analyte concentrations since the value and dispersion of $A_1$ and $A_2$ varies with the concentration. Table E5.1 presents a summary of $t_{\mathrm{R}}$, $A_1$ and $A_2$ performance parameters. The replicate analysis of extracts without detectable analyte levels was used to define models of signal noise dispersion in the retention time window (Table E5.1). Signals of extracts of food products representative of the nutritional patterns of foodstuffs of vegetable origin with high water content were studied.

From the data in Table E5.1, models of $t_{\mathrm{R}}$ and *AR* variability were developed. $t_{\mathrm{R}}$ models were built from confidence intervals based on Student's t distribution ($\bar{t}_{\mathrm{R}i} \pm ts_{t\mathrm{R}i}$; where $\bar{t}_{\mathrm{R}i}$ and $\mathrm{s}_{t\mathrm{R}i}$ are the mean and standard deviation of the $t_{\mathrm{R}}$, and $t$ the $t$-value of the $t$-distribution for the defined confidence level and degrees of freedom of $\bar{t}_{\mathrm{R}i}$ and $\mathrm{s}_{t\mathrm{R}i}$). The *AR* models were built from Monte Carlo Method simulations. From the observed dispersion of the various $A_1$ and $A_2$ of the analyte, ratios of correlated $A_1$ and $A_2$ (i.e., *AR*) were simulated. From blank extracts, the signal noise and, subsequently, the *AR* in blanks were simulated. The signal noise was modelled as a normal distribution truncated at zero since chromatographic peaks do not have negative areas. Table E5.2 presents the estimated dispersion of $t_R$ and *AR*. This table also presents the MS-Excel formulas used in the simulation of $A_1$ and $A_2$. The confidence limits for the $t_{\mathrm{R}}$ and *AR* were set for a confidence level of 99.9 % or 98 %, respectively, that corresponds to the *TP*.

The *FP* from identifications based on $t_{\mathrm{R}}$ was set at 10 % based on the experience of the analyst. This *FP* represents the probability of a peak not confirmed to be from the analyte, being observed within the defined retention time window for the analyte. The *FP* from *AR* was estimated from Monte Carlo simulations of signal noise and from determining how many times the simulated noise produces *AR* within the acceptance interval for this parameter. Since the *FP* can be extremely large for low analyte levels, it was determined at different analyte mass fractions by defining a minimum abundance of each ion. Table E5.3 presents the estimated *TP*, *FP*, and their combination in $LR(+)$. In the last column of the table, the performance of identifications based on both the $t_{\mathrm{R}}$ and *AR* is reported. Table E5.3 also presents the posterior probability *PP* that a test item is positive, assuming that positive or negative results are equally likely (i.e., $P(+) = P(-) = 0.5$).

**Table E5.1.** Performance parameters relevant for the identification of chlorpyrifos-methyl and malathion in extracts of vegetable origin by GC-MS/MS. All parameters were estimated with 11 degrees of freedom.

| Analyte: Chlorpyrifos-methyl | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Abundance | | | | |
| | | Retention time, $t_{\mathrm{R}}$ | | Ion: 208 *m/z* | | Ion: 271 *m/z* | | |
| | $w$ | $\bar{t}_{\mathrm{R}i}$ | $s_{t\mathrm{R}i}$ | $\bar{A}$ | $s_A$ | $\bar{A}$ | $s_A$ | $\rho$ |
| Extract | (mg kg$^{-1}$) | (min) | (min) | (a.u.) | (a.u.) | (a.u.) | (a.u.) | |
| E § | 3.33 | 17.24 | 0.024 | 105668 | 13.3 | 138678 | 6.14 | 0.9956 |
| E § | 0.33 | 17.24 | 0.024 | 10163 | 10.5 | 15025 | 8.10 | 0.6151 |
| E § | 0.083 | 17.24 | 0.024 | 4366 | 21.4 | 5790 | 15.1 | 0.3965 |
| G | < 0.04 | - | - | 372 | 892 | 372 | 892 | - |
| O | < 0.04 | - | - | 372 | 892 | 372 | 892 | - |
| I | < 0.04 | - | - | 372 | 892 | 372 | 892 | - |

| Analyte: Malathion | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Abundance | | | | |
| | | Retention time, $t_R$ | | Ion: 99 *m/z* | | Ion: 127 *m/z* | | |
| | $w$ | $\bar{t}_{\mathrm{R}i}$ | $s_{t\mathrm{R}i}$ | $\bar{A}$ | $s_A$ | $\bar{A}$ | $s_A$ | $\rho$ |
| Extract | (mg kg$^{-1}$) | (min) | (min) | (a.u.) | (a.u.) | (a.u.) | (a.u.) | |
| E § | 3.33 | 19.45 | 0.070 | 226592 | 7.85 | 226765 | 10.3 | 0.9988 |
| E § | 0.33 | 19.45 | 0.070 | 22354 | 17.4 | 22969 | 15.6 | 0.9672 |
| E § | 0.083 | 19.45 | 0.070 | 5882 | 30.7 | 6345 | 28.0 | 0.7677 |
| G | < 0.11 | - | - | 372 | 892 | 372 | 892 | - |
| O | < 0.11 | - | - | 372 | 892 | 372 | 892 | - |
| I | < 0.11 | - | - | 372 | 892 | 372 | 892 | - |

§ – The dispersion of ion abundances was estimated by combining signals from the analyte in a pure solvent with signals from vegetable extracts. Extract matrix: G – Ginger, O – Spring onion, I – Irish moss seaweed; E – unspecified matrix.

$w$ – mass fraction of the analyte (mg kg$^{-1}$); $\bar{t}_{\mathrm{R}i}$ – mean retention time (min) (this parameter can vary with the day of the injection); $s_{t\mathrm{R}i}$ – standard deviation of the retention time estimated under repeatability conditions (min); $\bar{A}$ – mean ion abundances (arbitrary units, a.u.); $s_A$ – standard deviation of the ion abundance; $\rho$ – Spearman's correlation coefficient.

**Table E5.2.** Acceptance intervals for the retention time and ion abundance ratio.

| Analyte | Extract | Mass fraction interval, $w$ (mg kg$^{-1}$) | Maximum $t_{\mathrm{R}i}$ difference (min) (*c.l.* 99.9 %) § | $AR$ interval (*c.l.* 98 %) † |
|---|---|---|---|---|
| Chlorpyrifos-methyl | E | 0.04 – 3.33 | 0.18 | 0.439 – 1.18 |
| Malathion | E | 0.11 – 3.33 | 0.54 | 0.467 – 1.54 |

E – unspecified matrix; *c.l.* – confidence level;

§ – Maximum difference between the retention time of the analyte in a standard solution and in the analysed sample ($\sqrt{2}ts_{t\mathrm{R}i}$);

† – MS-Excel formula used in the simulation: First ion: $A_1 = \bar{A}_1 + s_{A1} * TINV(R1, \nu_1)$ and Second ion: $A_2 = \bar{A}_2 + s_{A2} * (TINV(R1, \nu_2) * \rho + TINV(R2, \nu_2) * (1 - \rho\char`^2)\char`^0.5)$, where $\nu_{\mathrm{i}}$ are the degrees of freedom associated with $A_i$ and $s_{Ai}$, and R1 and R2 two independent random value generators U(0,1) (Excel formula RAND()).

According to Table E5.3, only when identifications are performed at or above the Limit of Quantification (0.14 mg kg$^{-1}$ or 0.38 mg kg$^{-1}$ for chlorpyrifos-methyl and malathion, respectively), identifications are supported by 'very strong' pieces of evidence (i.e., $10^5 < LR(+) < 10^6$). All performance characteristics presented in Table E5.3 ($TP$ and $FP$, $LR(+)$ and $PP$) are valid alternatives for reporting performance or uncertainty of qualitative analysis at different analyte mass fractions. Reporting the $LR(+)$ has the advantage of combining $TP$ and $FP$ in a single metric, and of not requiring the assumption of a prevalence of the pesticide in analysed samples.

**Table E5.3.** Performance characteristics of the identification of chlorpyrifos-methyl and malathion by GC-MS/MS.

| | | | Performance characteristics at different analyte levels, *w* | | |
|---|---|---|---|---|---|
| Analyte | | *w* (mg kg$^{-1}$) | $t_{\mathrm{R}}$ | *AR* | $t_{\mathrm{R}}$ & *AR* |
| Chlorpyrifos-methyl | *TP* (%) | ≥ 0.04 | 99.9 | 98 | 97.9 |
| | *FP* (%) | 0.04 § | 10 | 30.2 | 3.02 |
| | | 0.08 | 10 | 0.2 | 0.02 |
| | | 0.14 † | 10 | 0.001 | 0.0001 |
| | | 0.27 | 10 | 0.001 | 0.0001 |
| | $LR(+)$ | 0.04 § | 9.99 | 3.24 | 32.4 |
| | | 0.08 | 9.99 | 490 | 4895 |
| | | 0.14 † | 9.99 | 98000 | 979020 |
| | | 0.27 | 9.99 | 98000 | 929020 |
| | *PP* (%) | 0.04 § | 90.9 | 76.4 | 97.0 |
| | | 0.08 | 90.9 | 97.8 | 99.98 |
| | | 0.14 † | 90.9 | 99.999 | 99.9999 |
| | | 0.27 | 90.9 | 99.999 | 99.9999 |
| Malathion | *TP* (%) | ≥ 0.11 | 99.9 | 98 | 97.9 |
| | *FP* (%) | 0.11 § | 10 | 29.8 | 2.98 |
| | | 0.23 | 10 | 0.001 | 0.0001 |
| | | 0.38 † | 10 | 0.001 | 0.0001 |
| | | 0.77 | 10 | 0.001 | 0.0001 |
| | $LR(+)$ | 0.11 § | 9.99 | 3.29 | 32.9 |
| | | 0.23 | 9.99 | 98000 | 979020 |
| | | 0.38 † | 9.99 | 98000 | 979020 |
| | | 0.77 | 9.99 | 98000 | 979020 |
| | *PP* (%) | 0.11 § | 90.9 | 76.68 | 97.0 |
| | | 0.23 | 90.9 | 99.999 | 99.9999 |
| | | 0.38 † | 90.9 | 99.999 | 99.9999 |
| | | 0.77 | 90.9 | 99.999 | 99.9999 |

§ – Limit of detection; † – Limit of quantification; $t_{\mathrm{R}}$ – analyte retention time; *AR* – abundance ratio of two characteristic ions of the analyte's mass spectrum.

This example illustrates how the Monte Carlo simulation of signals can overcome the difficulty of experimental determination for false positive rates of highly selective identifications.

## 7.6 E6: Identification of SARS-CoV-2 RNA by nucleic acid amplification testing

| **Scope:** |
|---|
| **Type of qualitative analysis:** Analysis based on quantitative criteria |
| **Item/matrix:** Nasal swabs, nasopharyngeal and oropharyngeal swabs |
| **Parameter/analyte:** SARS-CoV-2 RNA |
| **Type of classification criterion:** Cycle threshold, Ct, values equal or lower than the Ct cutoff are classified as positives; higher values are classified as negatives. |
| **Technique/instrumentation:** Reverse transcription polymerase chain reaction (RT-PCR) |
| **Type of results reporting:** True positive rate and true negative rate. |

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the virus that causes the disease COVID-19, the respiratory illness responsible for the COVID-19 pandemic. One of the screening tests for this virus's presence in nasal swabs involves a reverse-transcription polymerase chain reaction (RT-PCR), a type of nucleic acid amplification testing (NAAT). The “in-house” validation of this test involves determining the cycle threshold cut-off, LOD, (clinical) sensitivity (*SS*) and specificity (*SP*) (Table 2), with other performance parameters (not presented in this example). In clinical analysis, the *SS* and *SP* are known as clinical accuracy estimators. Clinical accuracy is limited by *FN* and *FP* and by epidemiological prevalence, types and subtypes of agents, mutations, and other biological factors.

### *7.6.1 Cycle threshold*

The “number of cycles needed for an amplicon to become detectable above background” is defined as the cycle threshold (Ct) [59] – the number of cycles needed to amplify viral RNA to reach a detectable level. Some variables should be recognized to understand the application of the Ct cut-off. Rn (normalized reporter value) is the magnitude of the signal generated by the given set of PCR conditions. ΔRn (Figure E6.1) is the normalised reporter value minus the baseline response. The threshold is the signal level that reflects a statistically significant increase over the computed baseline signal (see Figure E6.1). This decision line is established to distinguish relevant amplification signals from the background. In the example, the software sets the threshold to 10 times the standard deviation of the baseline fluorescence value. The limit is defined in the region related to an exponential growth of the PCR product. Figure E6.1 illustrates the positive classification (less than, or equal to, the Ct cut-off of 32) of a human sample. Fluorescence results higher than 32 are classified as negatives. Note that only sigmoidal amplification curves are indicative of true amplification (see Figure E6.1).

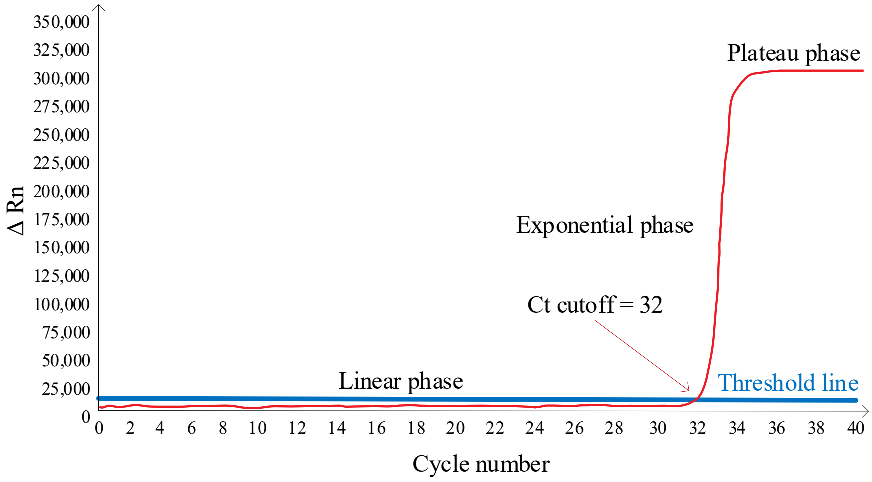

**Figure E6.1.** Detection of SARS-CoV-2 RNA from a true positive sample.

### 7.6.2 LOD estimation

Detection capability near to the Ct cut-off is evaluated using the *LOD*. The *LOD* is defined as the concentration multiple (for example, expressed in number of copies/mL) associated with a *TP* of 95 % ($LOD_{95\%}$). The $LOD_{95\%}$ is estimated by modelling the variation of *TP* with concentration and estimating the concentration where *TP* is 95 % using probit regression [60] – [63]. In the example, the $LOD_{95\%}$ estimated in a series of seven dilutions from a sample with a concentration of 500 copies/mL is 114 copies/mL. More detail about the presented procedure for the determination of the $LOD_{95\%}$ is available in the bibliography (5.5 of [63]).

### 7.6.3 Clinical accuracy

The assessment of the clinical accuracy involves defining target or minimum values for the lower limits of the 95 % CI of *SS* and *SP*, $LL_{SS.95}^{tg}$ and $LL_{SP.95}^{tg}$, and checking if the estimated lower limits are equal or higher than the respective target value (i.e., if $LL_{SS.95} \geq LL_{SS.95}^{tg}$ and $LL_{SP.95} \geq LL_{SP.95}^{tg}$). For this test purpose, $LL_{SS.95}^{tg} = 95$ % and $LL_{SP.95}^{tg} = 90$ %, that is the lower limit of the sensitivity should be greater than 95 % and the lower limit of specificity should be greater than 90 %.

For method validation, 200 nasal swabs, nasopharyngeal and oropharyngeal swab samples were analysed: 100 from individuals known to be infected with SARS-CoV-2 and 100 from individuals confirmed not to be infected with this virus. Table E6.1 shows the contingency table obtained from the 200 tests. This table reveals that no false negatives and three false positives have been reported, so the clinical *SS* and *SP* are 100 % and 97 %, respectively. The limits of the 95 % CI of *SS* and *SP*, calculated from Eq. (8) to (11), are [96.3, 100] and [91.6, 99.0], respectively (Table E6.2). Since $LL_{SS.95}$ and $LL_{SP.95}$ are higher than 95 % and 90 %, respectively, the analytical method is considered valid. *SS* is complemented by seroconversion sensitivity [64].

**Table E6.1.** Contingency table that summarises the performance of the method for detecting SARS-Cov-2 RNA in nasal swabs, nasopharyngeal and oropharyngeal swab samples.

| | | **Case** | | |
|---|---|---|---|---|
| | | **Positive (*pc*)** | **Negative (*nc*)** | **Result totals** |
| **Result** | **Positive (*p*)** | *tp* = 100 | *fp* = 3 | *p* = 103 |
| | **Negative (*n*)** | *fn* = 0 | *tn* = 97 | *n* = 97 |
| | **Case totals** | *pc* = 100 | *nc* = 100 | |

**Table E.6.2.** Clinical accuracy of the method for detecting SARS-Cov-2 RNA in human serum or plasma

| | Clinical sensitivity | |
|---|---|---|
| $SS = 100$ % | $LL_{SS.95} = 96.3$ % | $HL_{SS.95} = 100$ % |
| | Clinical specificity | |
| $SP = 97$ % | $LL_{SP.95} = 91.6$ % | $HL_{SP.95} = 99.0$ % |

INTENTIONALLY BLANK

# Annex A – Bayes' theorem, odds, and the likelihood ratio

## A.1 Bayes' theorem

Bayes' theorem describes how the probability of an event A (such as a test item being genuinely positive) changes with new information E, such as a positive test result. Bayes' theorem is most commonly written for two events, $A$ and $E$, as:

$$P(A|E) = \frac{P(E|A)P(A)}{P(E)} \qquad \text{(A.1)}$$

Here, $P(A)$ and $P(E)$ are the probabilities of the events A and E occurring alone, $P(A|E)$ is the probability of event $A$ given that $E$ has occurred and $P(E|A)$ is the probability of event $E$ given that $A$ has occurred. In statistics, $P(A|E)$ and $P(E|A)$ are usually referred to as "conditional probabilities"; for example, $P(E|A)$ can be referred to as the conditional probability of event $E$ given $A$.

In the context of qualitative analysis, taking the positive case as an example, $P(A)$ can be understood as the probability that a randomly chosen test item is genuinely positive before any tests are performed. $P(E|A)$ is the probability that a genuinely positive test item will generate a positive test result – the true positive rate TP in Table 2. $P(E)$ is the probability of a positive test result irrespective of the state of the test item. Finally, $P(A|E)$ is the probability that the test item is genuinely positive, *after* considering the information added by the positive test result. Since it is calculated after the evidence $E$ becomes available, $P(A|E)$ is usually called the "posterior probability" for $A$. An estimated posterior probability gives a direct indication of the confidence that can be placed in a classification.

It is important to remember that (continuing with the positive case) $P(E)$ includes both true positive results and false positive results, and also that $P(E)$ applies to the complete population of test items. This means that $P(E)$ is sensitive both to the true and false positive rates and to the proportions of genuinely positive and negative test items. Quantitatively, for two cases $A$ and $\neg A$ (denoting "Not-$A$", a genuinely negative test item), $P(E)$ can be written as a weighted sum:

$$P(E) = P(A)P(E|A) + P(\neg A)P(E|\neg A) \qquad \text{(A.2)}$$

Considering E as a positive test result, equation A.2 says that the combined probability of E is the true positive rate times the proportion of genuinely positive samples, plus the false positive rate times the proportion of genuinely negative samples. This is why a high false positive rate reduces confidence in positive test results. Referring to equation A.1, if there is a high probability of positive results from genuinely negative test items, $P(A|E)$ reduces because $P(E)$ increases. This matches intuition; however high the true positive rate, the chance of a large number of false positives should make us less certain that a positive result indicates a genuinely positive test item.

## A.2 Probability and odds.

A probability $P$ is usually expressed as a number between 0 and 1. However, it can also be expressed in the form of "odds"[5], a term perhaps most familiar in sports betting. If the probability of an event $A$ is $P(A)$ and the alternative possibility is simply "Not $A$", the odds $O(A)$ in favour of $A$ can be calculated using:

$$\mathrm{O(A)} = \frac{P(A)}{1 - P(A)} \qquad \text{(A.3)}$$

Unlike probabilities, odds can take any non-negative value; odds of $10^6$ or "a million to one" are possible.

Odds can be converted back to probabilities by rearranging A.3 to give:

$$\mathrm{P(A)} = \frac{\mathrm{O}(A)}{\mathrm{O}(A) + 1} \qquad \text{(A.4)}$$

[5] The term "odds" is generally regarded as plural

## A.3 The odds form of Bayes’ theorem and the likelihood ratio

If there are only two alternative and complementary hypotheses, A and ¬A (that is, “Not $A$”), and some evidence E (such as a test result that is positive for A) is used to update the probabilities of each, Bayes’ theorem gives the posterior probabilities as:

$$\mathrm{P}(A|E) = \frac{P(E|A)P(A)}{P(E)} \qquad \text{(A.5a)}$$

$$\mathrm{P}(\neg A|E) = \frac{P(E|\neg A)P(\neg A)}{P(E)} \qquad \text{(A.5b)}$$

The ratio of their probabilities is then:

$$\frac{P(A|E)}{P(\neg A|E)} = \frac{P(E|A)P(A)}{P(E|\neg A)P(\neg A)}$$

or, separating terms for clarity, (A.6)

$$\frac{P(A|E)}{P(\neg A|E)} = \frac{P(E|A)}{P(E|\neg A)} \times \frac{P(A)}{P(\neg A)}$$

Since there are only two hypotheses, $A$ and $\neg A$, the prior probabilities and the posterior probabilities must sum to 1; that is, $P(\neg\, A) = 1 - P(A)$ and $P(\neg A|E) = 1 - P(A|E)$. This means that the left side of A.6 is equal to $\mathrm{P}(A|E)/[1 - P(A|E)]$. Comparing with A.3, this is just the odds in favour of $A$, given $E$, or $O(A|E)$, the “posterior odds” in favour of hypothesis $A$. Similarly, the prior odds $O(A)$ appear on the right side of A.6 as $P(A)/P(\neg A) = P(A)/[1 - P(A)] = O(A)$. The remaining ratio, $P(E|A)/P(E|\neg A)$, is known as the “likelihood ratio”. For the qualitative analysis case, where $\neg A$ corresponds to a genuinely negative test item, Table 5 gives the (estimated) likelihood ratio as $TP/FP$.

The odds form of Bayes’ theorem can therefore be written as

$$O(A|E) = O(A) \times \frac{P(E|A)}{P(E|\neg A)} \qquad \text{(A.7)}$$

or, schematically,

Posterior odds = Prior odds × Likelihood ratio

The likelihood ratio can therefore be interpreted quantitatively as the change in odds in favour of a particular hypothesis.

# Annex B – Qualitative analysis associated with the assessment of conformity with a quantitative limit

## B.1 Conformity assessment as a qualitative analysis

The assessment of the conformity of the value of a quantitative parameter of the analysed item with a limit value or interval can be considered a qualitative analysis using a single quantitative criterion (Section 2), with the outcomes 'conforming' or 'non-conforming'. Table B.1 presents some examples of these types of analyses.

**Table B.1.** Examples of qualitative analysis based on the assessment of the conformity of the value of a quantitative parameter of the analysed item with a limit value or interval.

| | |
|---|---|
| **(1)** | Assessment of the colour of raw material by comparing absorbance measurements with a threshold. |
| **(2)** | Assessment of the conformity of an alloy with a minimum content for its major component. |
| **(3)** | Assessment of the conformity of a medicine with the specification interval for the concentration of the active substance. |
| **(4)** | Assessment of the conformity of a pesticide residue in fruit given a maximum residue level. |
| **(5)** | Assessment of the health condition of an individual by comparing a measured blood component with an interval of values from healthy individuals. |

The use of decision rules and measurement uncertainty in compliance assessment is discussed in detail in the Eurachem/CITAC guide on "Use of uncertainty information in compliance assessment" [29] ("The compliance guide"). For completeness, however, this annex discusses how uncertainty or performance information for quantitative analysis can be used to derive some of the metrics in Table 2. These can then be used to characterise the performance of qualitative analysis procedures based, wholly or partly, on comparing measurement results with a limit or specification.

When the analysis involves assessing whether a measured property is above, below, or within a specification limit or interval, the measurement uncertainty can be used to quantify conformity assessment reliability.

NOTE. The present Guide does not discuss how the measurement uncertainty should be evaluated. The evaluation of measurement uncertainty is described in detail in the Eurachem/CITAC guide "Quantifying uncertainty in analytical measurement" [65].

The use of measurement uncertainty for conformity decisions described in the Eurachem/CITAC compliance guide [29] involves setting a criterion for deciding if an item conforms or does not conform, with a maximum probability of false conformity decisions of $x$ %. The compliance guide distinguishes "specific" and "global" risks. The "specific risk" quantifies the probability of a false decision on the conformity of a particular item; it is based solely on the distribution associated with the measurement result for that item. In contrast, the "global risk" quantifies the probability of false decisions on the conformity of a randomly chosen future item [66]. Global risk takes account of the distribution of possible values for items measured, such as the distribution of values of items from a production line or an environmental area. For instance, calculation of global producer risk requires the probability of a production line producing products with a value close to the limit value, such that it can be falsely considered as non-conforming. Therefore, for the determination of global risk, the distribution of values for the population of items must be well characterised.

Most frequently, analysts are interested in assessing the conformity of a specific analysed item. In such cases, how can metrics used to quantify the reliability of other types of qualitative analysis be determined? A case study and the formulae used for determining these parameters are presented below

## B.2 Positive and negative conformity assessment results

If the item is considered to conform from a 'positive result', the distribution describing the measurement uncertainty can be used to provide either a likelihood ratio $LR(+)$ in favour of conformity or, under some

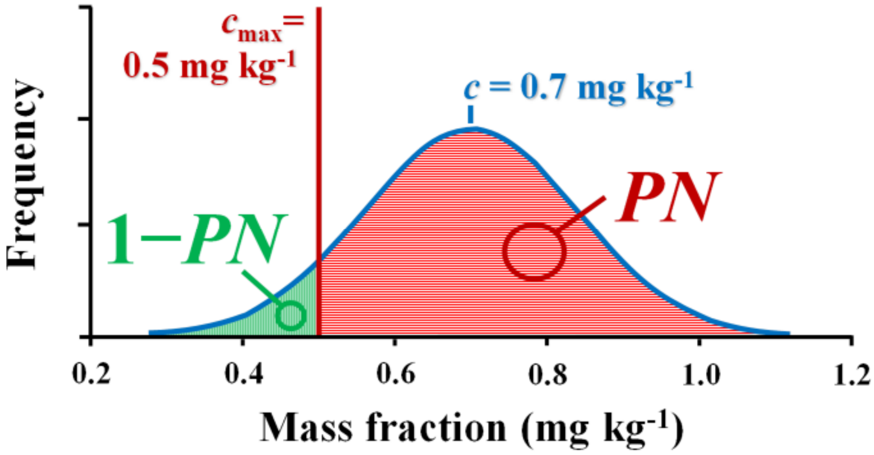

**Figure B.1.** Graphical representation of the probability $PN$ that an analysed item is non-conforming with a maximum limit, $c_{\mathrm{max}}$, from a measured value, $c$, with associated standard uncertainty, $u(c)$ = 0.14 mg kg$^{-1}$, and the corresponding probability $1 - PN$ that it does not conform.

circumstances, the posterior probability $PP$ of the item actually conforming ('a positive case') and the probability of the item being, in fact, non-conforming ('a negative case') $(1 - PP)$. Equivalently, if a result is reported as 'non-conforming' or 'negative' by comparison with a limit or interval, the measurement uncertainty can be used to obtain the corresponding likelihood ratio $LR(-)$ or the posterior probabilities $PN$ and $(1 - PN)$. The example below considers the posterior probabilities.

NOTE. The definition of 'positive results' and 'negative results' as 'conforming' and 'non-conforming', respectively, is arbitrary; the opposite convention can be followed.

## B.3 Example – Conformity assessment for pesticide residue in fruit

Assume the conformity of a sample of grapes is being assessed against a maximum residue level, $c_{\mathrm{max}}$, of 0.5 mg kg$^{-1}$ for acetamiprid [67], and the measured mass fraction in the sample is 0.70 mg kg$^{-1}$, $c$. [6] The measurement result has a normal distribution with a standard uncertainty, $u(c)$, of 0.14 mg kg$^{-1}$ estimated with a very high number of degrees of freedom. Since $c > c_{\mathrm{max}}$, the most likely conclusion about conformity is the 'non-conformity' of the grapes (a 'negative result').

Formally, a posterior probability such as $PP$ or $PN$ requires a prior probability. In this case no information is available on the general distribution of acetamiprid in grapes. However, in cases of suspected contamination, it is often reasonable to assume that the distribution is so broad as to be essentially uninformative in the region of the measurement result. Where that is the case, the measurement uncertainty can be taken as an approximation to the posterior distribution. Taking that approach here, the posterior probability of a negative (non-conforming) test sample, $PN$, is the area under the probability density function of the measurement result to the right of $c_{\mathrm{max}}$, represented in Figure B.1 in red. The area is the upper tail probability of a normal distribution with mean $c$ and standard deviation equal to $u(c)$. This can be calculated from a spreadsheet or statistical package; in Microsoft Excel, for example, the required formula is 1 −NORM.DIST($c_{\mathrm{max}} = 0.5$, $c = 0.7$, $u(c) = 0.14$, TRUE) (see Table B.2). For this example, the area is 0.923, or 92.3 %. The corresponding probability $PP$ that the sample is positive (conforming) is $(1 - PN)$ or 7.6 %.

Where informative prior information is available and it is appropriate to make use of it, the calculations involve integration over the prior distribution. Integrals for normally distributed prior and measurement uncertainty are given in, for example, JCGM 106 [66], together with guidance on other distributions.

[6] The symbol, $c$, for concentration is used in this Guide for cases applicable to various types of quantities such as mass concentration, mass fraction and pH.

NOTE It can be surprisingly hard to establish a genuinely uninformative prior distribution. For example, a simple uniform distribution over the range of Figure B.1 (0.2 – 1.2 mg kg$^{-1}$) would correspond to a 30 % prior probability that a test item complies with the limit, simply because only 30 % of that range is below the limit of 0.5 mg kg$^{-1}$. A complete Bayesian analysis will therefore include a check, typically involving alternative choices of prior distribution, to make sure that the conclusion is not unduly sensitive to the assumed prior distribution.

## B.4 Spreadsheet formulae for conformity assessment probabilities

Table B.2 presents the MS-Excel formulas that should be used when different conformity limits or intervals are considered, and the measured value is below, above, within or outside the limit(s).

If the standard uncertainty, $u(c)$, is estimated with a small number of degrees of freedom, $v$, instead of describing the dispersion of the estimate of the measurand by a normal distribution, Student's $t$-distribution should be considered. In that case, the general formula used in Table B.2, NORM.DIST($C$, $c$, $u(c)$, TRUE), should be substituted by TDIST(ABS($C - c$)/$u(c)$,$v$,TRUE).

**Table B.2.** MS Excel formulas used to estimate the probability of conformity, $PP$, or non-conformity, $PN$, decision on the specific analysed item. The formulae can be used to calculate $LR(+) = PP/(1 - PP)$ and $LR(-) = PN/(1 - PN)$.

| S | Limit | **Item conformity** (*result type*) Scenario | Conformity reliability | MS-Excel formula (based on the cumulative normal distribution) |
|---|---|---|---|---|
| 1 | Max. | **Conforming** (*positive*)<br>$c \le c_{\max}$ | $PP$ | NORM.DIST($c_{\max}$, $c$, $u(c)$, TRUE) |
| 2 | Max. | **Non-conforming** (*negative*)<br>$c > c_{\max}$ | $PN$ | 1 −NORM.DIST($c_{\max}$, $c$, $u(c)$, TRUE) |
| 3 | Min. | **Conforming** (*positive*')<br>$c \ge c_{\min}$ | $PP$ | 1 −NORM.DIST($c_{\min}$, $c$, $u(c)$, TRUE) |
| 4 | Min. | **Non-conforming** ('*negative*')<br>$c < c_{\min}$ | $PN$ | NORM.DIST($c_{\min}$, $c$, $u(c)$, TRUE) |
| 5 | Inter. | **Conforming** ('*positive*')<br>$c_{\min} \le c \le c_{\max}$ | $PP$ | NORM.DIST($c_{\max}$, $c$, $u(c)$, TRUE) – NORM.DIST($c_{\min}$, $c$, $u(c)$, TRUE) |
| 6 | Inter. | **Non-conforming** ('*negative*')<br>$c > c_{\max}$ or $c < c_{\min}$ | $PN$ | 1 − NORM.DIST($c_{\max}$, $c$, $u(c)$, TRUE) + NORM.DIST($c_{\min}$, $c$, $u(c)$, TRUE) |

S – Scenario; Inter., Max. or Min. – Interval, maximum or minimum limit; Positive or negative result – a conforming or non-conforming result; $c$ and $u(c)$ – measured concentration and associated standard uncertainty; $c_{\max}$ or $c_{\min}$ – Maximum or minimum admissible concentration

INTENTIONALLY BLANK

# Bibliography

[1] JCGM, International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (3rd edn.) (JCGM 200:2012), Sevres: BIPM, 2012.

[2] W. G. D. Ruig, G. Dijkstra and R. W. Stephany, “Chemometric criteria for assessing the certainty of qualitative analytical methods,” *Anal. Chim. Acta,* pp. 277-282, 1989.

[3] B. L. Milman and L. A. Konopelko, “Identification of Chemical Substances by Testing and Screening of Hypotheses. I. General,” *Fresenius. J. Anal. Chem.,* vol. 367, pp. 621-628, 2000.

[4] S. L. R. Ellison and S. Gregory, “Quantifying uncertainty in qualitative analysis,” *Analyst,* vol. 123, pp. 1155-1161, 1998.

[5] S. L. R. Ellison and S. L. Gregory, “Predicting chance infrared spectroscopic matching frequencies,” *Anal. Chim. Acta,* vol. 370, pp. 181-190, 1998.

[6] S. L. R. Ellison, “Uncertainties in qualitative testing and analysis,” *Accred. Qual. Assur.,* vol. 5, pp. 346-348, 2000.

[7] A. Ríos, D. Barceló, L. Buydens, S. Cárdenas, K. Heydorn, B. Karlberg, K. Klemm, B. Lendl, B. Milman, B. Neidhart, R. W. Stephany, A. Townshend, A. Zschunke and M. Valcárcel, “Quality assurance of qualitative analysis in the framework of the European project ‘MEQUALAN’,” *Accred. Qual. Assur.,* vol. 8, pp. 68-77, 2003.

[8] R. B. Silva, “Evaluation of trace analyte identification in complex matrices by low-resolution gas chromatography - mass spectrometry through signal simulation,” *Talanta,* vol. 150, pp. 553-567, 2016.

[9] J. Narciso, C. Luz and R. B. d. Silva, “Assessment of the Quality of Doping Substances Identification in Urine by GC/MS/MS,” *Anal. Chem.,* vol. 91, no. 10, pp. 6638-6644, 2019.

[10] ISO, Reference materials - Examples of reference materials for qualitative properties (ISO/TR 79:2015), Geneva: ISO, 2015.

[11] P. Pereira, B. Magnusson, E. Theodorsson, J. O. Westgard and P. Encarnação, “Measurement uncertainty as a tool for evaluating the ‘grey zone’ to reduce the false negatives in immunochemical screening of blood donors for infectious diseases,” *Accred. Qual. Assur.,* vol. 21, pp. 25-32, 2016.

[12] ILAC, ILAC Guidelines for Measurement Uncertainty in Testing (ILAC G17:01), Silverwater: ILAC, 2021.

[13] ISO, IEC, General requirements for the competence of testing and calibration laboratories (ISO/IEC 17025), Geneva: ISO, 2017.

[14] ISO, Medical laboratories – Requirements for quality and competence (ISO 15189), Geneva: ISO, 2012.

[15] R. Bramley, A. Brown, S. Ellison, W. Hardcastle and A. Martin, “Qualitative analysis: A guide to best practice - forensic science extension,” *Sci. Justice,* vol. 3, no. 40, pp. 163-170, 2000.

[16] L. Wide and C. A. Gemzell, “An immunological pregnancy test,” *Acta Endocrinol. (Copenh.),* no. 35, pp. 261-267, 1960.

[17] U. Forsum, H. O. Hallander, A. Kallner and D. Karlsson, “The impact of qualitative analysis in laboratory medicine,” *Trends Anal. Chem.,* vol. 6, no. 24, pp. 546-555, 2005.

[18] P. Pereira, Quality control of qualitative tests for medical laboratories, Lisbon: Author-edition, 2019.

[19] G. Nordin, R. Dybkaer, U. Forsum, X. Fuentes-Arderiu and F. Pontet, "Vocabulary on nominal property, examination, and related concepts for clinical laboratory sciences (IFCC-IUPAC Recommendations 2017)," *Pure Appl. Chem.,* vol. 5, no. 90, pp. 913-935, 2018.

[20] EU, Commission decision implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results (2002/657/EC), EU, 2002.

[21] EU, Commission Regulation No 2017/644 of 5 April 2017 laying down methods of sampling and analysis for the control of levels of dioxins, dioxin-like PCBs and non-dioxin-like PCBs in certain foodstuffs and repealing Regulation (EU) No 589/2014, EU, 2017.

[22] SANTE, Guidance document on analytical quality control and method validation procedures for pesticide residues and analysis in food and feed (SANTE/12682/2019), DG SANTE, 2019.

[23] WADA, Technical Document – TD2010IDCR, Identification Criteria for qualitative assays incorporating column chromatography and mass spectrometry, WADA, 2010.

[24] JCGM, Evaluation of measurement data – Guide to the expression of uncertainty in measurement (JCGM 100:2008), Sèvres: BIPM, 2008.

[25] ISO/IEC, Uncertainty of measurement – Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)(ISO/IEC Guide 98-3), Geneva: ISO, 2008.

[26] ISO, IEC, General requirements for the competence of testing and calibration laboratories (ISO/IEC 17025)(superseded), Geneva: ISO, 1999.

[27] Clinical Laboratory and Standards Institute, EP12-A2 User protocol for evaluation of qualitative test performance, 2nd ed., Wayne (PA): CLSI, 2008.

[28] N. R. Campbell, Physics, the elements, Cambridge: Cambridge University Press, 1920.

[29] A. Williams and B. Magnusson, (Eds.), Eurachem/CITAC Guide: Use of uncertainty information in compliance assessment, 2nd ed., Eurachem, 2021.

[30] B. Magnusson and U. Örnemark, (Eds.), Eurachem Guide: The Fitness for Purpose of Analytical Methods – A Laboratory Guide to Method Validation and Related Topics (2nd Edn.), Eurachem, 2014.

[31] A. Agresti, Categorical Data Analysis (3rd Ed.), New Jersey: Wiley, 2012.

[32] AOAC International, Official Methods of Analysis of AOAC International (20th Edn.), Maryland: AOAC International, 2016.

[33] S. D. Ferrara, L. Tedeschi, G. Frison, G. Brusini and F. Castagna, "Drugs-of-Abuse Testing in Urine: Statistical Approach and Experimental Comparison of Immunochemical and Chromatographic Techniques," *J. Anal. Toxicol.,* vol. 18, no. 5, pp. 278-291, 1994.

[34] R. J. Freund and W. J. Wilson, Regression Analysis, San Diego, CA: Academic Press, 1998.

[35] J. Fox, An R and S-Plus companion to applied regression, Thousand Oaks, CA.: Sage Publications Inc., 2002.

[36] S. L. R. Ellison and T. Fearn, "Characterising the performance of qualitative analytical methods: Statistics and terminology," *Trends Anal. Chem. ,* vol. 24, pp. 468-476, 2005.

[37] S. L. R. Ellison, C. A. English, M. J. Burns and J. T. Keer, "Routes to improving the reliability of low level DNA analysis using real-time PCR," *BMC Biotechnology,* vol. 6, pp. 33 (1-11), 2006.

[38] I. Kuselman and F. Pennecchi, “IUPAC/CITAC Guide: Classification, modelling and quantification of human errors in chemical analytical laboratory (IUPAC Technical Report),” *Pure Appl. Chem.,* vol. 88, pp. 477-515, 2016.

[39] EU, Commission Regulation No 589/2014 of 2 June 2014 laying down methods of sampling and analysis for the control of levels of dioxins, dioxin-like PCBs and non-dioxin-like PCBs in certain foodstuffs and repealing Regulation No 252/2012, EU, 2014.

[40] EU, Commission Regulation No 152/2009 of 27 January 2009 laying down the methods of sampling and analysis for the official control of feed, EU, 2009.

[41] T. Wenzl, J. Haedrich, Schaechtele, Alexander, P. Robouch and J. Stroka, Guidance Document on the Estimation of LOD and LOQ for Measurements in the Field of Contaminants in Feed and Food, JRC, 2016.

[42] J. Vessman, R. I. Stefan, J. F. Van Staden, K. Danzer, W. Lindner, D. T. Burns, A. Fajgelj and H. Müller, “Selectivity in Analytical Chemistry (IUPAC Recomendation 2001),” *Pure Appl. Chem.,* vol. 73, no. 8, p. 1381–1386, 2001.

[43] ISO, In vitro diagnostic medical devices – Information supplied by the manufacturer (labelling) – Part 1: Terms, definitions and general requirements (ISO 18113-1), Geneva: ISO, 2009.

[44] European Network of Forensic Science Institutes, ENFSI Guideline for evaluative reporting in forensic science, ENFSI, 2015.

[45] V. Morgado, C. Palma and R. J. N. Bettencourt da Silva, “Microplastics identification by Infrared spectroscopy – Evaluation of identification criteria and uncertainty by the Bootstrap method,” *Talanta,* vol. 224, p. 121814, 2021.

[46] A. J. Nunes, P. Paixão, J. Proença and R. J. N. Bettencourt da Silva, “Early warning of suspected doping from Biological Passport based on multivariate trends,” *Int. J. Sports Med.,* vol. 41, pp. 44-53, 2020.

[47] N. Pinto, M. Magalhães, E. Conde-Sousa, C. Gomes, R. Pereira, C. Alves, L. Gusmão and A. Amorim, “Assessing paternities with inconclusive STR results: The suitability of bi-allelic markers,” *Forensic Sci. Int. Genet.,* vol. 7, pp. 16-21, 2013.

[48] B. Meijer, J. Thijs, J. Kleibeuker, A. van Zwet and R. Berrelkamp, “Evaluation of eight enzyme immunoassays for detection of immunoglobin G against Helicobacter pylori,” *J. Clin. Microbiol,* vol. 35, no. 1, pp. 292-294, 1997.

[49] B. G. Armitage P, Statistical methods in medical research, 3rd ed., Cambridge: Blackwell Science, 1994.

[50] D. Zwillinger and S. Kokoska, Standard probability and statistics tables and formulae, Boca Raton (FL): Chapman & Hall/CRC, 2000.

[51] C. B. Agresti A, “Approximate is better than "exact" for interval estimation of binomial proportions,” *Am. Stat.,* vol. 52, no. 2, pp. 119-126, 1998.

[52] D. Altman, D. Machin, T. Bryant and M. Gardner, Statistics with confidence, 2nd ed., M. D. B. T. G. M. Altman DA, Ed., London: BMJ Books, 2000.

[53] R. Newcombe, “Two-sided confidence intervals for the single proportion: comparison of seven methods,” *Stat. Med.,* vol. 17, no. 8, pp. 857-872, 1998.

[54] E. Wilson, “Probable inference, the law of succession, and statistical inference,” *JASA,* vol. 22, no. 158, pp. 209-212, 1927.

[55] J. A. Sphon, “Use of Mass Spectrometry for Confirmation of Animal Drug Residues,” *J. Assoc. Off. Anal. Chem.,* vol. 61, pp. 1247-1252, 1978.

[56] R. Baldwin, R. Bethem, R. Boyd, W. Budde, T. Cairns, R. Gibbons, J. Henion, M. Kaiser, D. Lewis, J. Matusik, J. Sphon, R. Stephany and R. Trubey, “1996 ASMS FALL WORKSHOP: Limits to Confirmation, Quantitation, and Detection,” *J. Am. Soc. Mass Spectrom.,* vol. 8, pp. 1180-1190, 1997.

[57] K. S. Webb and D. Carter, "GC Report number LGC/VAM/1998/010," LGC Limited, London, 1998.

[58] W. G. De Ruig, R. W. Stephany and G. Dijkstra, "Criteria for the detection of analytes in test samples," *J. Assoc. Off. Anal. Chem.,* no. 72, pp. 487-490, 1989.

[59] Clinical Laboratory and Standards Institute, MM17 - Validation and Verification of Multiplex Nucleic Acid Assays (2nd ed.), Wayne (PA): CLSI, 2018.

[60] J. Gaddum, “Medical Research Council, Special Report Series no. 183,” Br Med J, 1933.

[61] C. I. Bliss, “The method of probits,” *Science,* vol. 79, no. 2037, pp. 38-39, 1934.

[62] D. J. Finney, Probit analysis, Cambridge: Cambridge University Press, 1947.

[63] Clinical Laboratory and Standards Institute, EP-17A2 Evaluation of detection capability for clinical laboratory measurement procedures, 2nd ed., Wayne (PA): CLSI, 2020.

[64] Clinical Laboratory and Standards Institute, MM53-A - Criteria for Laboratory Testing and Diagnosis of Human Immunodeficiency Virus Infection, Wayne (PA): CLSI, 2011.

[65] S. L. R. Ellison and A. Williams, (Eds); Eurachem/CITAC guide: Quantifying Uncertainty in Analytical Measurement, Third edition, 2012.

[66] JCGM, Evaluation of measurement data – The role of measurement uncertainty in conformity assessment (JCGM 106:2012), Sèvres: BIPM, 2012.

[67] I. Kuselman, F. Pennecchi, R. J. N. Bettencourt da Silva and D. B. Hibbert, “IUPAC/CITAC Guide: Evaluation of risks of false decisions in conformity assessment of a multicomponent material or object due to measurement uncertainty (IUPAC Technical Report),” *Pure Appl. Chem.,* vol. 93, no. 1, pp. 113-154, 2021.

[68] Association of Forensic Science Providers, “Standards for the formulation of evaluative forensic science expert opinion,” *Science and Justice,* vol. 2009, pp. 161-164, 2009.

[69] L. A. Currie, “Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995),” *Pure Appl. Chem.,* vol. 67, no. 10, pp. 1699-1723, 1995.

[70] European Food Safety Authority, The 2016 European Union report on pesticide residues in food, EFSA, 2018.

[71] I. Kuselman, F. Pennecchi, R. J. N. Bettencourt da Silva and D. B. Hibbert, “Risk of false decision on conformity of a multicomponent material when test results of the components’ content are correlated,” *Talanta,* no. 174, pp. 789-796, 2017.

[72] R. B. Silva and A. Williams, (Eds.), Eurachem/CITAC Guide: Setting and Using Target Uncertainty in Chemical Measurement, Eurachem, 2015.

[73] EU, Commission Regulation 2017/626 of 31 March 2017, EU, 2017.

[74] D. R. Cox, “The regression analysis of binary sequences (with discussion),” *J R Stat Soc B,* vol. 20, no. 2, pp. 215-242, 1958.

[75] D. J. Finney, Probit analysis, 2nd ed., Cambridge: Cambridge University Press, 1952.

[76] S. L. R. Ellison and A. Williams, (Eds.), Traceability in Chemical Measurement, 2nd ed., UK: Eurachem, 2019.

INTENTIONALLY BLANK

ISBN 978-0-948926-39-6