

Assessment of Performance and Uncertainty in Qualitative Chemical Analysis

by Brynn Hibbert, Teo Tang Lin, Ricardo Bettencourt, Stephen R. Ellison, Elvar Theodorsson, Paulo Pereira, Melissa Kennedy, and Wayne Dimech

An online workshop organized in January 2022 was brought together by like-minded people who recognize the importance of the long-awaited Guide [1]. While the Guide is published in an open-access format, proactive dissemination and communication are fundamental to achieve the desired impact, which encompasses continued implementation and application to a wide field such as analytical chemistry, forensics, and laboratory medicine. The idea of a workshop was thus conceived. To connect individuals from various communities into the common sphere of interest, we leverage on the websites of the event's sponsor, being the Eurachem, the Cooperation on International Traceability in Analytical Chemistry (CITAC), the Health Science Authority (HSA), Singapore and IUPAC and push messaging through pre-established networks with professional bodies.

The 12-hour event which spread across four days, created an inclusive community of participants centered around the core issue of how to express the validity of qualitative analytical results. The fundamental concepts of terminologies and content of the Eurachem/CITAC were brought to the participants in the very first session and formed the basis of subsequent discussions. Bite-size examples linked to historical stories, personal experiences or published works supplemented by the speakers enhanced the meanings and know-how for the use of the guide.

This workshop achieved a very active participation, with over 500 unique individuals from about 75 countries/territories around the world (Diagram 1). The participants have interacted with the authors of the Eurachem/CITAC Guide directly and with the workshop speakers via chat channels available during the online workshop and emails. The post workshop feedback revealed improved knowledge on the topic after the workshop and the ideality of the duration of each session (three hours). Individuals participated in the workshop mostly for work-related reasons or personal upgrading of knowledge. Over 75 % of the respondents saw medium to high impact of the Guide on their immediate goals, while about 74 % of them (out of 68) are likely to apply the Eurachem/CITAC Guide. Participants valued the e-Certificate of Attendances,

which were automatically issued to attendees who stayed through most of a session. Acknowledging challenges posed by the different time zones and other commitments of individuals, the recordings of the sessions were promptly shared through the websites of Eurachem [2], CITAC [3] and HAS [4]. The proactive model of communications and dissemination of the Guide (Diagram 2) may encourage others to embrace the use of digital tools to increase the visibility of valuable and useful references.

First session covering terminology and content of the Eurachem/CITAC Guide

The first day of the workshop started with a presentation from Gunnar Nordin from EQUALIS AB, Sweden, on terminology for the management of qualitative analysis entitled "The VIM4 approach to nominal properties." Dr Nordin discussed the differences between the determination of a quantity, an ordinal quantity or a nominal property.

The second communication of the day was delivered by one of the editors of the Eurachem/CITAC guide on the Assessment of the Performance and Uncertainty of Qualitative Analysis, Ricardo Bettencourt da Silva from University of Lisbon, Portugal. Dr Silva is the current chair of the Qualitative Analysis Working Group that produced this guide. This presentation discussed the socio-economic relevance of qualitative analysis, the quantitative or qualitative nature of the information considered, and the difficulty of assessing the performance of highly selective methods exclusively from experimentation. For instance, reliable quantification of a 1 % false positive rate requires performing at least 1500 tests on negative cases. Dr Silva mentioned that an alternative to the experimental assessment of the performance of some analytical methods is testing the classification through database search or by modelling instrumental signals. Usually, database searching only allows an initial assessment of performance since the diversity of items on the database is often not representative of the studied population of unknown items. Database searches are frequently used in qualitative analysis using the optic of mass spectrometric techniques. Monte Carlo simulation of instrumental signals of positive and negative cases can be used to quantify the probability of true and false decisions on the used classification criteria. Dr Silva also presented the list of contents of the Eurachem/CITAC Guide.

Stephen R. Ellison from LGC, United Kingdom presented the third communication. Dr Ellison is another editor of the guide and was the former chair of the Qualitative Analysis working group. His

communication discussed alternative metrics for quantifying the performance of qualitative analysis, such as the posterior probability of a case given the observation of respective evidence and the likelihood calculated from Bayes' Theorem. Dr Ellison presented a comprehensive and graphical description of the role of information of the prevalence of cases known prior to the test, on the probability that the most probable case is correct. He highlighted that the Guide provides tools for quantifying and expressing the uncertainty of qualitative analysis results but does not make a recommendation as to whether such uncertainties should, or should not, be reported to customers. This is currently left to laboratories' discretion, in the light of regulation and accreditation body requirements. Where the laboratory does choose to report information on confidence in a qualitative analysis, however, the Guide does recommend caution to ensure that the information is presented clearly and in a manner that avoids misinterpretation.

The last two presentations of the first day of the workshop discussed examples of applications of the theory described in the Guide. The identification of trace levels of compounds in foodstuffs by GC-MS-MS was discussed and the identification of drugs of abuse in urine by enzyme multiplied immunoassay (EMIT) by Dr Silva and Dr Ellison, respectively. The first example illustrated the simulation of the instrumental signal to quantify the performance of highly selective identification methods. The second example presented and compared different metrics to express confidence in reported results.

Second session covering qualitative chemical analysis

The second workshop session focused on applications of performance and uncertainty evaluation in chemical analysis. This included application to identification of unknown materials, such as controlled drugs, and environmental applications. The methods used included spectroscopies (infrared, Raman, uv-visible, fluorescence), and combinations of mass spectrometric and chromatographic information.

Beginning the session, Dr Ellison discussed the assessment of infrared chance match probabilities. These correspond to false positive rates and are a key input to any assessment of the performance of a spectroscopic identification procedure. The response rates can be approached either practically, using databases or other information, or from theoretical considerations. In the examples shown, library match rates showed consistently higher chance match probabilities

than the simplest theoretical approaches, suggesting that theoretical approaches might be best used for comparing approaches rather than for assessing confidence in practice. A practical limitation of library matching, however, was that spectral libraries are not generally intended to represent a possible target population; rather, they intentionally include only one example of each material. This, too, leads to chance match probabilities that can differ from performance in the field, where a single material may appear disproportionately often.

Dr Silva discussed the identification of microplastics in environmental applications, using attenuated total reflectance FT-IR. This is increasingly important because of the increasing quantity of plastics now entering the environment each year. The process includes isolation of particles, micro spectroscopy, and automated or manual identification. Automated identification, in particular, relied heavily on establishing clear matching criteria, together with consistent signal processing and other data treatment, including rejection of spectra showing excessive biofilm contamination or insufficient signal strength. False response rates were evaluated by a bootstrapping method, and an optimal automated identification process chosen based on the best performance. Likelihood ratios were derived from the true and false response rates to give an indication of confidence in the results; for the purpose of research, a likelihood ratio of 19 (equivalent to a probability of about 95 %) or more was considered sufficient for characterising microplastic burdens in the environment.

Brynn Hibbert from University of New South Wales, Australia described some examples of the combination of evidence from different analytical methods, applied to the identification of the origin of oil spills in accidental environmental releases. Several analytical techniques had been used to examine the test samples, including UPLC with UV and with fluorescence detection, GC-MS, and isotope ratio determination by GCMS. The key tool for combining evidence from the different techniques was Bayes' theorem, which provides a natural approach to updating probabilities as evidence accumulated. This made it possible to obtain a single probability of matching each possible source, providing a direct indication of confidence in the conclusions.

The session concluded with a lively discussion session, addressing some of the challenges in quantifying performance of qualitative chemical analysis. The issue of metrological traceability was clearly still important for qualitative analysis, where the identification methods used measurements of quantities. In addition,

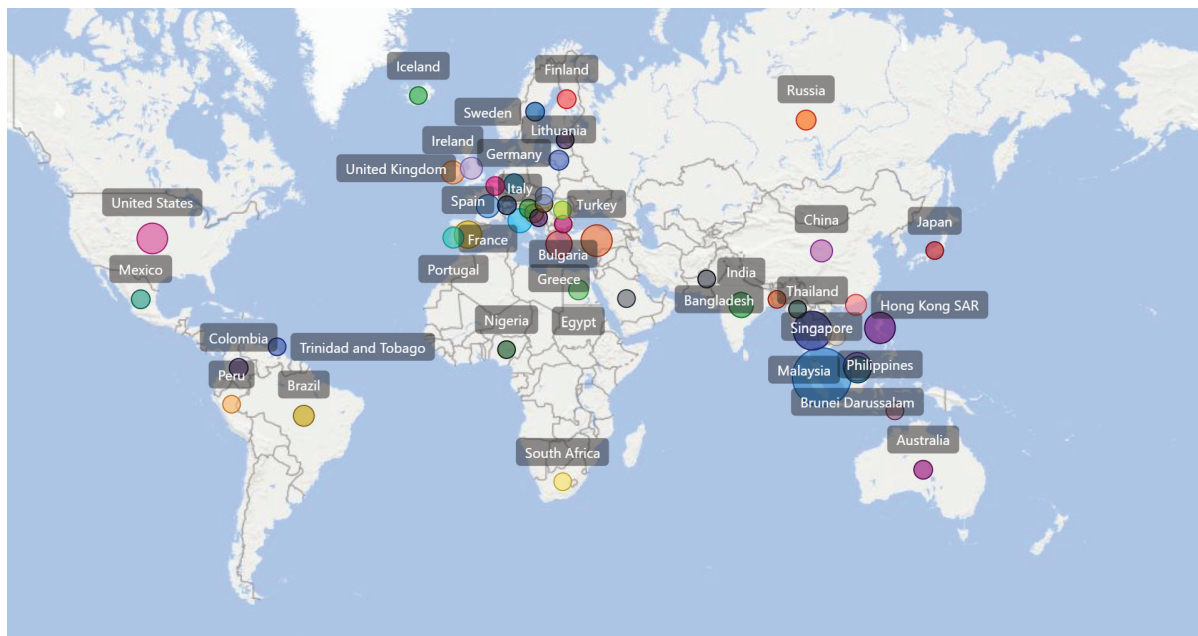
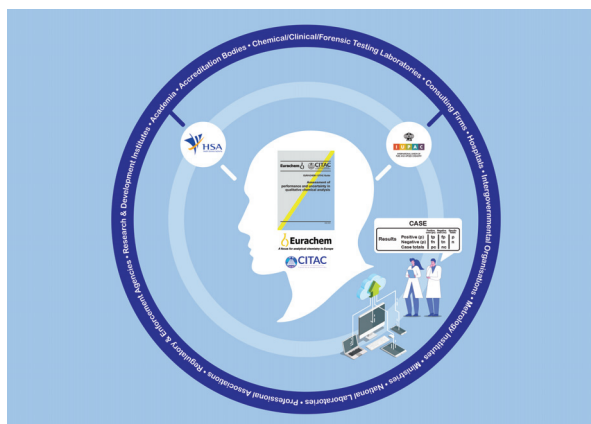


Diagram 1: A map to illustrate the degree of outreach of the online workshop. *Only a subset of the countries/territories from which the participants of the workshop come from is plotted on the map due to limited space*

Diagram 2: Illustration of model used to disseminate the Eurachem/CITAC Guide to various stakeholders and users



issues of authenticity and provenance (sometimes also described in terms of documentary or other traceability) were often very important; comparison with a reference specimen was only useful if the authenticity of the reference specimen could be demonstrated. These issues were particularly important for certifying reference materials for qualitative analysis. Similarly, the reliability of spectral libraries, reference data etc. needed to be demonstrated. In addition, it was important to ensure that reference data were acquired under conditions that matched the individual laboratories' procedure, underlining the need for good documentary evidence of the origin of reference data.

A different challenge was the difficulty of 'propagating' uncertainties through a qualitative process. For example, a reference spectrum might have an associated uncertainty that could be handled using Bayes' theorem for purely qualitative information. Quantitative

measurement uncertainty information could also be used to estimate performance in qualitative methods, particularly based on decision thresholds, but it would still be difficult to incorporate information on (for example) variation in field sampling. In practice, this relied on extensive field tests; a good example was the requirement for comprehensive field testing for approval of Covid-19 test kits.

Third session covering qualitative forensic analysis

The Forensic Science session was organised and chaired by Melissa Kennedy, ANSI National Accreditation Board/ANAB, United States. After an introduction by Ms Kennedy four very different speakers (academic lawyer, academic analytical chemist, forensic DNA specialist and a forensic scientist who deals with standards) gave their views of the state of

forensic science in the light of the Eurachem/CITAC Guide. There was a lively discussion following the talks, joined by Dr Ellison who was involved with forensic matters at LGC, UK, before the forensic laboratories were moved elsewhere.

It could be said that a theme running through the presentations was that although forensic science is the *sine qua non* of qualitative analysis (either the defendant is guilty or not guilty), the issues that affect the usefulness of scientific evidence go far beyond the uncertainty of a qualitative result. The first speaker, Law Professor Gary Edmond spelled out the “very low bar” that the legal system presents in terms of admissibility. Despite efforts to set reliability standards for scientific evidence, the expert who simply declares “that there is a match”, or “the defendant handled the weapon”, or “based on my 25 years of experience the similarities are clear”, is more often than not allowed to say his piece without challenge. There has been some attempt at progress. For example, in the USA, ballistic experts used to say that a bullet came from a gun; now they must say “to a reasonable degree of ballistic certainty.” But is this any better, and what does it actually mean? Professor Edmond’s advice to the forensic scientist was to be “driven by the science.”

Two examples of uncontrolled experts were given by Emeritus Professor Hibbert who recounted two Australian murder trials about 25 years apart. In the first a geologist was allowed to call a match between dirt on the jeans of a body and dirt in the boot (trunk) of the defendant’s car. No validation, no likelihood ratios, and no uncertainties of either the quantitative or qualitative results. In the second case in 2020, evidence of lead isotope ratios by Multi-Collector Inductively Coupled Plasma Mass Spectrometry, which has a fantastic repeatability (relative ~0.01%), was used to declare that bullets in a body had come from a box of bullets in the possession of a defendant. As before there was no validation, but this time the judge determined the evidence was not admissible. There was also no attempt to calculate likelihood ratios, not true positive rate/ false positive rate (see Guide p15), but the wider question of how many other identical bullets had been sold was not addressed. Even accepting the match with no uncertainty, there needed to be only one more box of bullets from the same pig of lead (batch) in circulation as a possible source of the murder bullets to reduce the likelihood ratio ($P(E|HP)/P(E|HD)$), to 1, that is the evidence favours neither prosecution or defence.

The third talk was in the safer realm of DNA analysis. Associate Professor Michael Coble from University of North Texas Health Science Center in Fort Worth,

United States, gave a clear exposition on the use of probabilistic software for genotyping. Just 10 years ago some 75 % of US laboratories were using combined probability of inclusion (CPI). The CPI refers to the proportion of a given population that would be expected to be included as a potential contributor to an observed DNA mixture. In a landmark paper (of which Dr Coble was an author) [5] the use of likelihood ratios was strongly recommended, and now in the US over 60 % of laboratories are using probabilistic software. In one of his examples using the modified random match probability approximately 1 in 400 trillion individuals would also be included in the mixture. He explained how this can be turned into a likelihood statement (“the evidence is 400 trillion times more likely if the stain came from the person of interest, than if it came from an unknown, unrelated individual”). In response to a question from the audience Dr Coble explained how different levels of uncertainties can be taken account of.

Finally, Agnes D. Winokur who chairs the National Institute of Standards and Technology (NIST) Organization of Scientific Area Committees for Forensic Science (OSAC) Seized Drugs Sub-committee, discussed the current status of consensus-based standard development efforts in statistical reporting for qualitative analysis (identification) of seized drugs. There is no need to sell the utility of developing standards, but the talk put into perspective how difficult it is to produce a workable standard with the buy-in of all stakeholders. Taking estimation of ‘error rates’ as an example, three approaches: evaluating competency and proficiency tests, evaluating quality control samples, and evaluating re-analysis of casework were described. ASTM has published three standards for microcrystal tests for amphetamines and other drugs. The problems with using interlaboratory studies to develop standards occupied three slides and the practicalities of organising laboratories with limited budgets and time, different equipment, not to mention how to make test samples resemble real forensic cases, or even how to ship drug samples around the country made the point that a lot of effort must be put into making standards.

Questions raised after the talks revolved around the utility of likelihood ratios. If the LR is 400 trillion how can a jury not read this as ‘guilty’? The probability of human error in the forensic process must be greater. Dr Ellison said that in the UK only DNA evidence was given in terms of likelihood ratios and offered the use of verbal equivalents (‘strong support’ etc., see Table 5 in the Guide). The ability of most people who make up juries have little grasp of large numbers (Professor

Edmond suggested >10 according to a study of a colleague from the School of Psychology), and whether a number or verbal equivalent were given, courts would tend to hear ‘support for the prosecution hypothesis’.

As Dr Ellison, remarked: “And I think, you know, worrying about whether [the LR is] 5 or 106 or 1010 pales into insignificance by [the question] “are we telling you the right thing?”. I think that is one of the biggest challenges in conveying forensic evidence of this kind. What does the number actually tell you?

Fourth session covering qualitative analysis in laboratory medicine

Although results expressed on a ratio scale represent most results in laboratory medicine, qualitative results expressed on nominal and ordinal scales are also common and at least as necessary for healthcare practice. Results in pathology, transfusion medicine, immunology, and microbiology are commonly qualitative. Furthermore, examination procedures producing qualitative results are increasingly being made readily and affordably available to the public. Home pregnancy tests have been available since the 1970s, and the recent COVID pandemic has acquainted wide swathes of the public with the use of lateral flow tests for the virus. Therefore, it is highly appropriate to pay attention to the investigation and expression of the performance and uncertainty of qualitative results in the numerous fields where they are practiced, including in medicine.

Paulo Pereira from Portuguese Institute of Blood and Transplantation, Portugal gave a broad overview of the assessment of performance and uncertainty in qualitative tests in the medical laboratory, including how to handle decision/cut-off limits and calculate diagnostic sensitivity and specificity, predictive values, and uncertainty of proportions.

Elvar Theodorsson from Linköping University, Sweden discussed ways of obtaining and maintaining metrological traceability in qualitative analysis in the light of the ISO 17511:2020 and the IFCC-IUPAC Recommendations 2017 (Nordin, G., *et al* (2018). “Vocabulary on nominal property, examination, and related concepts for clinical laboratory sciences.” *Pure Appl. Chem.* 90(5): 913-935).

Wayne Dimech from National Serology Reference

Laboratory, Australia used data from an international external quality control program called QConnect, to estimate the uncertainty of qualitative measurement in infectious disease testing. This method uses the imprecision and the bias of quality control data submitted to the program.

While there is a rapid development of quantitative measurement methods that replace qualitative methods in laboratory medicine, there is also a strong development of numerous qualitative methods that are highly useful for self-diagnosis and monitoring by the public. It is, therefore, crucial to continue the development of harmonized concepts, terminology, and procedures for improving the quality, diagnostic accuracy, and the expression of the performance of qualitative results in the field of medicine. Hopefully, such developments can progress in a coordinated and harmonized manner in all fields of qualitative measurements.

The public has learned how to deal with the pre-examination, examination, and post-examination matters relating to qualitative pregnancy tests. A similar understanding is needed to interpret qualitative results in all fields of chemical analysis, environmental analysis, forensic analysis, and of course in laboratory medicine.

Special acknowledgements to: Gunnar Nordin, Gary Edmond, Michael Coble, and Agnes D. Winokur

References:

1. EURACHEM/CITAC Guide: Assessment of performance and uncertainty in qualitative chemical analysis, 2021. <https://www.eurachem.org/index.php/publications/guides/performance-and-uncertainty-in-qualitative-analysis>
2. www.eurachem.org/index.php/events/workshops/394-wks-aqa2022
3. www.citac.cc/conferences-and-workshops/
4. Chemical metrology events (hsa.gov.sg)
5. Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J. M., Coble, M. D. Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. *BMC Genet* 17, 125 (2016). <https://doi.org/10.1186/s12863-016-0429-7>

<<https://iupac.org/event/performance-and-uncertainty-in-qualitative-chemical-analysis/>>