Forensic Chemistry 10 (2018) 15–26

Contents lists available at ScienceDirect

Forensic Chemistry

journal homepage: www.elsevier.com/locate/forc

Comparison of measured and recommended acceptance criteria for the analysis of seized drugs using Gas Chromatography–Mass Spectrometry (GC–MS)

J. Tyler Davidson [a], Benny J. Lum [b], Gina Nano [c], Glen P. Jackson [a,d,*]

[a] Department of Forensic and Investigative Science, West Virginia University, Morgantown, WV 26506-6121, USA
[b] Broward Sheriff's Office Crime Lab, Fort Lauderdale, FL 33301, USA
[c] University of Massachusetts Medical School, Drugs of Abuse Laboratory, USA
[d] C. Eugene Bennett Department of Chemistry, West Virginia University, Morgantown, WV 26506, USA

ARTICLE INFO

Article history:
Received 29 March 2018
Received in revised form 4 July 2018
Accepted 4 July 2018
Available online 5 July 2018

ABSTRACT

Gas Chromatography–Mass Spectrometry (GC–MS) is a widely used analytical technique that has become a critical tool in many industries, including forensic science. Many governing bodies provide recommendations for the uncertainty of measurement for chemical substance identification, but existing guidelines often do not provide the numerical data to support the basis for their acceptance criteria. The guidelines therefore appear anecdotal and, if they are not continually updated, may not reflect modern instrument capabilities. This study provides data, with detailed interpretations, to assess the magnitudes and sources of measurement uncertainties of typical GC–MS data, as commonly practiced.

Data analysis was conducted using 13 different drug standards from three different laboratories using five different GC–MS setups. The laboratories were not prescribed a set of instrumental parameters, but rather were asked to submit the parameters ordinarily practiced within their respective laboratories. An expanded uncertainty of two times the relative standard deviation (2σ) of replicate measurements was used to report the uncertainty of measurement for the retention time and relative ion abundance measurements made on each set-up.

The retention time acceptance criteria currently recommended by many agencies are on the magnitude of ±2%, but such criteria are much wider than the measured within-week or within-month 2σ values, which are actually on the magnitude of ±0.20%. The measured uncertainties of relative ion abundances are similar to recommended acceptance criteria, but a careful assessment shows that ion abundances are not independently variable within a spectrum. Some ion abundances correlate with correlation coefficients (R squared) that exceed 0.9. Acceptance criteria for the GC retention time measurements should therefore be stricter than most of the current recommended guidelines. The application of tighter acceptance criteria would provide: 1) an evidenced-based, statistical reason for making drug identifications; and 2) fewer type I errors (false positives) in seized drug analyses than provided by existing standards.

1. Introduction

One of the most difficult tasks that crime laboratories face today is the assessment and communication of the uncertainty of measurements [1]. Whereas Gas Chromatography–Mass Spectrometry (GC–MS) is widely accepted as the gold standard of forensic drug analysis, as part of an analytical scheme [2–4], the uncertainty-of-measurement recommendations associated with this technique vary between organizations [5]. The uncertainty of a measurement is the interval, on the measurement scale, that encompasses the true value of the measurement with a specified probability, after all sources of error have been taken in to account [6]. The sources of error can be divided into two categories: random and systematic. Systematic error involves reproducible inaccuracies that occur consistently in the same direction, whereas random error involves irreproducible deviations, such as would be measured by replicate measurements of a sample [7]. To provide accurate, unbiased results, an analyst must be able to correct for any systematic errors in an analysis. To correct for systematic error, it is desirable—though not always essential—to know the source(s) of systematic error. Oftentimes, one cannot correct for random error, so the magnitude of the random error must simply be reported.

* Corresponding author at: Department of Forensic and Investigative Science, West Virginia University, Morgantown, WV 26506-6121, USA.
E-mail address: forensicchemistry@mail.wvu.edu (G.P. Jackson).

https://doi.org/10.1016/j.forc.2018.07.001
2468-1709/

In seized drug analyses, samples are often analyzed on a gas chromatograph–mass spectrometer (GC–MS) [2] as part of an analytical scheme for seized drug identification [2–4], so analysts must be aware of the uncertainty of the retention time of the gas chromatograph and the relative ion abundances of the mass spectrometer. The retention time describes the length of time that an analyte of interest takes to elute through the separation column. The relative ion abundance is the intensity of each ion present in the mass spectrum relative to that of the base peak, the most intense ion. To account for the uncertainty of measurements, many organizations have developed acceptance criteria, or tolerance windows, to guide analysts in their interpretations. Such guidelines allow for some acceptable level of random and systematic variation that occurs between analyses. However, different agencies have adopted different guidelines, and the data on which their acceptance criteria are based are rarely evident. The present study attempts to overcome this gap in the literature by providing an analysis of replicate data from three different laboratories; spreadsheets containing the extracted raw data are provided in the Supplemental material.

The uncertainty of measurement for GC retention time data can be described in either absolute time or relative time units where relative time is a percentage of the known reference standard retention time [8]. Organizations such as the European Commission (EC) and the United States Department of Agriculture (USDA) recommend that, for an unknown analyte and a reference material to be not significantly different, their retention times must agree to within ±0.1 min [9,10]. Other agencies, such as the United Nations Office on Drugs and Crime (UNODC), the German Society for Toxicological and Forensic Chemistry (GTFCh), and the North Carolina Department of Justice recommend that retention times should agree to within ±2% [11–13]. The World Anti-Doping Agency (WADA), Federal Drug Administration (FDA), Clinical and Laboratory Standards Institute (CLSI), and the Association of Official Racing Chemists (AORC) use an either/or system, so analysts can adopt a percentage value or an absolute value in minutes (e.g. ±2% or ±0.1 min) [14–17]. On one extreme, the American Society for Testing and Materials (ASTM) endorses a retention time uncertainty of ±6% [18].

Acceptance criteria for the uncertainty of mass spectral comparisons are similarly diverse. Whereas attempts to develop a general set of rules for the interpretation of mass spectra and functional group specific fragmentation pathways do exist [19,20], issues regarding inconsistencies in substance identification have been raised [21]. The ASMS Measurements and Standards Committee describe a "fit for purpose" approach to uncertainty wherein the definitions for the acceptable degree of the measurement uncertainty and identification certainty should be set to meet the needs of the application [22]. This committee recognized that uncertainty was inherent in mass spectrometric methods, and that the tolerance required by the application should drive the selection process. Core concepts of mass spectrometric acceptance were also proposed, as was the use of false positives and false negatives to express uncertainty. While the use of false negatives and false positives to express the uncertainty of analytical results has been accepted in many forensic laboratories [23], other issues, such as reference database searching and non-uniform acceptance criteria (e.g. relative and absolute abundances), have been critiqued [24].

McLafferty et al. performed extensive work on quality of evaluations of databases for unknown identifications using electron ionization mass spectrometry (EI–MS) data. The conclusion was that the size of the database was much more important than the number of peaks in the query spectrum [25]. Demuth demonstrated the value of structural similarity searches to return structurally similar compounds in a database search, even if the query compound isn't in the reference library [26]. Database searches have made use of a variety of different reference library algorithms. For example, the use of dot products as the search algorithm resulted in the correct top hits approximately 75% of the time [27]. Other work has included Bayesian statistics to assess the confidence generated by similarity scores [28], random projections in binary space between the query and reference spectra [29], calculated properties in juncture with the mass spectrum information [30], and power normalization to systematically alter weighting of different peaks [31]. Whereas reference database algorithms have been explored throughout the literature, their effectiveness has also been questioned [32]. The use of computational methods for mass spectral data analysis is a growing field, but other mechanisms for comparison of spectra outside of a reference library exist [33], including an unequal-variance *t*-test [34,35].

Regarding the identification of drugs using EI-MS spectra, some organizations define acceptable uncertainties of peak intensities on a relative scale, some on an absolute scale, and others combine relative and absolute scales, depending on the intensity of an ion relative to the base peak [14,17,36]. Organizations such as the USDA, UNODC, and Scientific Working Group for Forensic Toxicology (SWGTOX) use a ±20% absolute uncertainty of the relative ion abundance [12,37,38]. The International Food Safety Training Laboratory (IFSTL) and the European Commission report from 2015 suggest a ±30% absolute value of the relative ion abundance [39,40]. The Environmental Protection Agency (EPA), FDA, CLSI, and ASTM recommend relative uncertainty values of 20–30%, such that a relative uncertainty of ±20% relative to an ion with an absolute intensity of 50% (relative to the base peak) would produce a range of ±10% relative to the base peak; i.e. an acceptable range from 40% to 60% of the base peak [16,18,41,42].

Whereas different agencies must meet the fit-for-purpose needs of their disciplines [22], the drug analysis communities could benefit from either a uniform set of criteria, some transparency on how these recommendations were derived, or some guidance on how a laboratory could establish its own acceptance criteria [1]. To help rectify these problems, Kelly and Bell recently provided some measures of uncertainty for retention indices and mass spectral ion abundances [43]. We hereby provide additional results of GC–MS measurements of 13 different drug standards measured on five different instruments in three different laboratories. The measurements were collected using parameters specific to each laboratory, rather than prescribed parameters, to attempt to quantify the uncertainty of measurement in GC–MS measurements across a variety of instrumental parameters. Replicate measurements were conducted multiple times a week over multiple months and included routine maintenance protocols like septum changes and vacuum pump maintenance. An expanded uncertainty of two times the relative standard deviation of the mean (2σ) was used to estimate the 95% confidence interval of the retention time and relative ion abundance measurements [44]. The results show that many agencies propose very conservative (i.e. large) acceptance criteria for the uncertainty of GC retention times, with the undesirable result of elevating the risk of false positive identifications, or type I errors. Acceptance criteria for the comparison of electron ionization mass spectra, such as those provided by the USDA, FDA, and UNODC [10,12,15], are also conservative, but they are considerably closer to the measured uncertainty (2σ) of replicate measurements of drug standards. An example of a type I error which occurred due to the uncertainty of relative ion abundances was recently described by Valdez et al. [45].

2. Methods

2.1. Chemicals and reagents

Laboratory A used a standard drug mixture comprised of methamphetamine (1500 ppm), cocaine (1500 ppm), and hydromorphone (2000 ppm). Methamphetamine and hydromorphone were supplied by Sigma-Aldrich, cocaine was supplied by Mallinckrodt, and the methanol solvent was supplied by Alfa Aesar.

Laboratory B used a standard drug mixture comprised of ecgonine methyl ester (5050 ppm), cocaine (5050 ppm), 6-monoacetylmorphine (6-MAM) (5700 ppm), diacetylmorphine (DAM) (5700 ppm), and fentanyl (5200 ppm). The ecgonine methyl ester and cocaine were supplied by Sigma-Aldrich, the 6-MAM, DAM, and fentanyl were supplied by Lipomed, and the methanol solvent was supplied by Fisher Scientific.

Methods C1, C2, and C3 in Laboratory C each used 12 different 2,5-dimethoxy-N-(N-methoxy-benzyl)phenethylamine (NBOMe) isomer solutions. The ortho, meta, and para isomers of 25C-NBOMe and 25I-NBOMe were analyzed at concentrations of 125 ppm and 1250 ppm. All drug standards were provided via the DEA Special Testing Laboratory. The methanol solvent was supplied by Fisher Scientific.

2.2. Data collection/extraction

Data from two operational crime labs were collected during the processing of routine casework. The standard operating procedures (SOPs) in both crime laboratories require data from drug standards to be collected on regular intervals between casework samples. These standards are used as true positives to ensure the instruments are working properly. All samples were comprised of commercially-available drug standards; no casework samples were used. Even though several instruments involved in this study were manufactured by Agilent, the retention time locking (RTL) feature was not used during the collection of this data set. However, if RTL were to have been used, it would have only increased the disparity between the uncertainty of measurement identified in this study with the recommended retention time acceptance criteria [46,47]. Table 1 is a summary of the drug standards analyzed within each laboratory as well as the frequency of measurement. The total number of measurements column displays the actual number of data points collected within each laboratory over the duration of time the drug standards were analyzed. In all, 13 drug standards were analyzed between all five laboratories (with cocaine being analyzed in two labs and six NBOMe isomers analyzed by three laboratories).

Unprocessed raw data files were transferred from the crime laboratories to the PI's laboratory for data processing and analysis. All data extraction and analysis was performed using either MSD ChemStation Version C.01.01 or TurboMass Version 6.1.0. All retention times were extracted via auto-integration with constant parameters to eliminate analyst bias. The retention time was extracted at the peak apex in all cases in both ChemStation and TurboMass software. The instrument conditions represented the realistic conditions observed in different operational laboratories. The instruments included split and splitless modes of sample introduction, conventional, narrow, and wide-bore capillary columns, and temperature gradients that include unilinear and non-unilinear increases in oven temperature.

2.3. Instrumentation

Laboratory A used an Agilent Technologies 7890 GC-5977 MS with a HP-5 (5% phenyl-methylpolysiloxane) 12 m × 200 μm × 0.33 μm column manufactured by Agilent J&W Columns. The GC–MS parameters were as follows: injection volume was 1 μL; injection temperature was 220 °C; split ratio was 100:1. The initial oven temperature was set to 80 °C for 1.5 min, which was ramped to 270 °C at 50 °C/min, then held for 1.67 min. A second ramp to 290 °C, at 35 °C/min was held for 2.7 min. The carrier gas (helium) flow rate was set to 1 mL/min and the transfer line temperature was set to 290 °C. The mass spectrometer was scanned from m/z 30 to 650 after a solvent delay of 0.80 min. The scan rate was 2852 Da/sec. The source and quadrupole temperatures were 230 °C and 150 °C, respectively.

Laboratory B used an Agilent Technologies 7890 GC-5977 MS with a DB-5MS (5% phenyl-methylpolysiloxane) 30 m × 250 μm × 0.25 μm column manufactured by Agilent J&W Columns. The GC–MS parameters were as follows: injection volume was 0.2 μL; injection temperature was 280 °C; split ratio was 20:1. The initial oven temperature was 80 °C, which was ramped to 300 °C at 30 °C/min, then held for 9 min. The carrier gas (helium) flow rate was set to 0.684 mL/min and the transfer line temperature was set to 280 °C. The mass spectrometer was scanned from m/z 40 to 500 after a solvent delay of 2 min. The scan rate was 1472 Da/sec. The source and quadrupole temperatures were 230 °C and 150 °C, respectively.

Laboratory C1 used an Agilent Technologies 7890 GC-5977 MS with a VF-5MS (5% phenylmethyl-polysiloxane) 10 m × 150 μm × 0.15 μm column manufactured by Agilent J&W Columns. The GC–MS parameters were as follows: injection volume was 1 μL; injection temperature was 250 °C; split ratio was 100:1. The initial oven temperature was 150 °C, which was ramped to 280 °C at

Table 1
Summary of drug standards and the frequency of measurements.

Lab	Drugs analyzed	Concentration (ppm)	Solvent	Duration (weeks)	Total #measurements
A	Methamphetamine	1500	Methanol	25	396
	Cocaine	1500			
	Hydromorphone	2000			
B	Ecgonine methyl ester	5050	Methanol	22	117
	Cocaine	5050			
	6-Monoacetylmorphine	5700			
	Diacetylmorphine	5700			
	Fentanyl	5200			
C1	25C-[$]NBOMe (ortho, meta, para) and 25I-NBOMe (ortho, meta, and para) isomers	[#]125 & 1250	Methanol	5	348
C2	25C-NBOMe (ortho, meta, para) and 25I-NBOMe (ortho, meta, and para) isomers	[#]125 & 1250	Methanol	7	348
C3	25C-NBOMe (ortho, meta, para) and 25I-NBOMe (ortho, meta, and para) isomers	[#]125 & 1250	Methanol	8	325

[$] NBOMe is an abbreviation for 2,5-dimethoxy-N-(N-methoxy-benzyl)phenethylamine.
[#] Each isomer was analyzed at both concentrations.

25 °C/min, then held for 1 min. The carrier gas (helium) flow rate was set to 1 mL/min and the transfer line temperature was set to 280 °C. The mass spectrometer was scanned from m/z 25 to 500 after a solvent delay of 0.5 min. The scan rate was 1500 Da/sec. The source and quadrupole temperatures were 250 °C and 200 °C, respectively.

Laboratory C2 used an Agilent Technologies 7890 GC-5977 MS with a HP-5 (5% phenyl-methylpolysiloxane) 30 m × 250 μm × 0.25 μm column manufactured by Agilent J&W Columns. The GC–MS parameters were as follows: injection volume was 1 μL; injection temperature was 250 °C; split ratio was 40:1. The initial oven temperature was 150 °C, which was ramped to 280 °C at 15 °C/min, then held for 3 min. The carrier gas (helium) flow rate was set to 1 mL/min and the transfer line temperature was set to 280 °C. The mass spectrometer was scanned from m/z 25 to 500 after a solvent delay of 2 min. The scan rate was 1500 Da/sec. The source and quadrupole temperatures were 250 °C and 200 °C, respectively.

Laboratory C3 used a PerkinElmer Clarus 680 GC-SQ8S MS with a ZB-5MS ((5% phenyl)-dimethylpolysiloxane) 20 m × 180 μm × 0.18 μm column manufactured by Phenomenex. The GC–MS parameters were as follows: injection volume was 1 μL; injection temperature was 250 °C, splitless mode. The initial oven temperature was 150 °C, which was ramped to 280 °C at 15 °C/min, then held for 3 min. The carrier gas (helium) flow rate was set to 1 mL/min and the transfer line temperature was set to 280 °C. The mass spectrometer was scanned from m/z 25 to 500 after a solvent delay of 2 min. The scan rate was 1800 Da/sec. The source and quadrupole temperatures were 250 °C and 200 °C, respectively.

2.4. Data analysis

All data analysis was performed in Microsoft Excel Version 14, Microsoft Excel Version 16, and SPSS Version 24. The relative standard deviations and averages were used to generate an expanded uncertainty of two times the relative standard deviation of the mean (2σ). The average 2σ value for the sample mean was determined on a within-week and within-month basis. An average across all substances within a single laboratory was calculated so that a comparison of the within-week and within-month uncertainty of measurement values within each laboratory could be performed.

The 2σ criterion was used for retention time and relative ion abundance data. For each mass spectrum, auto-integration parameters were used to obtain representative mass spectrum from the apex of each drug standard in a chromatogram. For each drug standard, the signal intensities of 12–15 of the most abundant ions were extracted and normalized to the base peak (defined as 100) to produce relative ion abundances. Because the base peak, by definition, has no uncertainty, the relative abundance of 11–14 peaks after the base peak were examined in more detail.

3. Results and discussion

3.1. Retention time analysis

After gathering retention time data for repeated measurements of drug standards from the five different instrumental setups, the first course of action was a manual inspection of the retention time data. Regions affected by different systematic and random errors were apparent in each instrumental setup, and regions of discrete changes in retention times were often attributable to laboratory notes regarding scheduled maintenance and adverse events. Such events are typical in real laboratory settings.

Fig. 1 demonstrates the visual differences observed between systematic and random error present in the GC retention times of methamphetamine (Fig. 1a) and cocaine (Fig. 1b) on one instrument in one laboratory represented as raw data. Cocaine and methamphetamine were analyzed from the same drug standard mixture (1500 ppm), in the same injections on the same days. One should therefore expect that any systematic factors, such as a slightly faster oven temperature ramp in a given run, would influence each drug similarly. However, random deviations such as divergence from the linear temperature ramp, might affect the drugs separately, especially because methamphetamine may have eluted before the factor continues to influence cocaine. Fig. 2, which is a plot of the methamphetamine (Fig. 2a) and cocaine (Fig. 2b) residuals is plotted with both absolute (mins) and relative (%) axes to help convey the relative magnitudes of the residuals. There are three distinct color-coded regions of behavior in Fig. 1. Each region contains a linear change in retention time as a function of collection day. The regions of discontinuity correspond with discrete systematic occurrences, including: 1) a power cut and vacuum pump maintenance, 2) shortening of the column, and 3) replacement of the septum in the GC injection port, respectively. Each linear region contains random and non-random error, which are visualized by the spread of the data (i.e. the residuals) around the linear-regression lines of best fit (Fig. 2a and b). This data set contains at least three different sources of uncertainty in the retention times of methamphetamine and cocaine: 1) uncertainty due to unexpected instrument performance, as demonstrated by short non-random behavior before—and a discrete change in retention time after—the vacuum system was vented for maintenance and the column was trimmed; 2) uncertainty due to column bleed over repeated temperature cycles of the GC oven, as demonstrated by a systematic reduction in retention time as a function of analysis day; and 3) uncertainty due to random variance in column head pressure, column flow, oven temperature and vacuum outlet pressure, as demonstrated by the random distribution of residuals about the linear regression lines of best fit (Fig. 2a and b). For this data set, there are three long-term regions (e.g. >30 days) over which systematic and random error can be analyzed for the determination of uncertainty of measurement, represented by the blue, orange, and gray subsets in Fig. 2a and b.

If WADA acceptance criteria (±2%) were applied to the retention time data in Fig. 1a and b, the permissible range of retention times—for an unknown analyte to be considered not significantly different from the standard—would be 2.72 ± 0.05 min for methamphetamine and 5.63 ± 0.11 min for cocaine. The acceptable ranges proposed by WADA are therefore 2.67–2.77 min for methamphetamine and 5.52–5.74 min for cocaine. This range assumes that the unknown and reference material are measured on the same day, but the acceptance criteria are so large that all the retention times for methamphetamine from day 1 through day 104—which includes the first few data points in the orange series—fall within the acceptance criteria of the original data point collected on day one. For cocaine, all the retention times collected over the entire 6-month duration of the study fall within the acceptance window proposed by WADA. Clearly, the acceptance window of ±2% is extremely conservative—in terms of avoiding type II errors—and covers more variation than is typical for random error on this particular instrument.

The acceptance window of ±2% also covers more variation than can be explained by systematic decreases in retention time caused by column bleed over a three-month period. The acceptance window of ±2% also covers more variation than can be explained by instrument maintenance, such as repair of the turbo pump or shortening of the column. Confidence intervals or acceptance windows are supposed to capture the random errors that occur in a measurement, not the systematic errors. Yet the acceptance

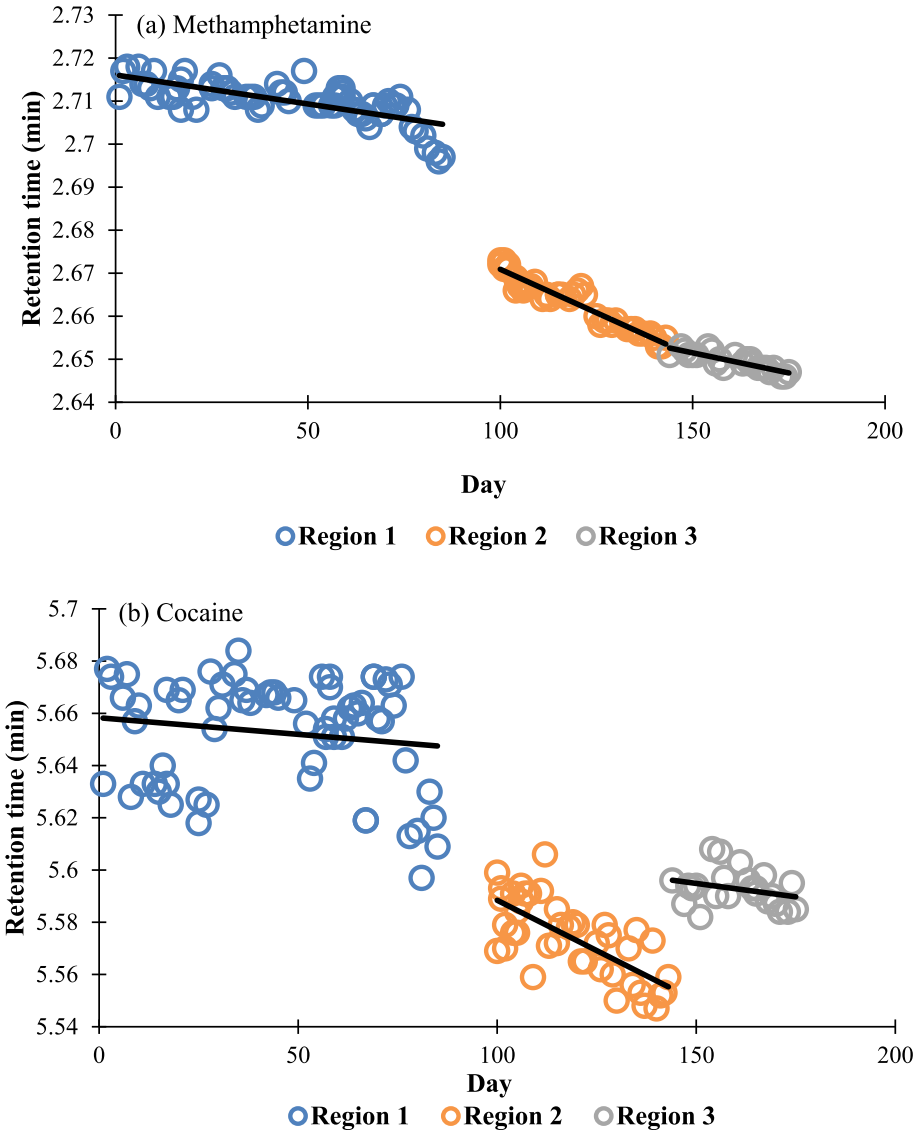

Fig. 1. Retention time data for (a) methamphetamine and (b) cocaine measured from the same injections and the same chromatograms over an extended period. The retention times show different regions of systematic and random sources of deviation.

criterion of ±2% spans such a wide range that it would not adequately reject measurements that are obviously outside the normal random error of repeated measurements.

The line of best fit through the first 85 days of methamphetamine and cocaine data both provide slopes of 0.00013 min/-day, or 0.008 s/day. The decrease in retention times caused by column bleed is therefore a very small effect, but one that adds up to systematic change of 0.013 min or 0.8 s over 3 months of normal daily use. The correlation coefficient (R^2) of the line of best fit for this first region is 0.53 for methamphetamine and 0.02 for cocaine. A plot of the residuals from the linear regression line for methamphetamine (Fig. 2a) shows that there are many areas where the residuals are not randomly distributed. Instead, the residuals show some periods where measurements on consecutive days show systematic deviations from the regression line. As an example, the last 12 days of the first region (blue data points) shows a systematic decrease of 0.012 min in the retention time relative to the linear regression line. The effect is on the magnitude of 0.001 min/day. A similar behavior in observed in the plot of the residuals for cocaine (Fig. 2b), except at a rate of approximately 0.004 min/day. Anecdotally, the turbo pump was failing during this time period leading to a turbo pump replacement and the 15 day delay between regions 1 and 2 in Figs. 1 and 2. However, based on mathematical calculations of the conductance changes expected by a large change in the GC outlet pressure, a failing turbo pump would not be expected to provide the observed magnitude of variance. The short term non-random behavior may be a mere coincidence or due to factor(s) outside the consideration of this study. The GC injection liner and septum were also replaced at this time, but septa were changed on a weekly basis with no observed effect. After the turbo pump repair, there is an obvious decrease in retention time, due to the column being clipped by a few inches during the maintenance.

The next 47 days of data (orange data points) fit linear regression lines with slopes of 0.0040 min/day for methamphetamine and 0.008 min/day for cocaine. The R^2 values were 0.92 for methamphetamine and 0.51 for cocaine, which both show marked increases in linearity relative to the period before turbo pump maintenance. The residuals show that the deviations from the line of best fit are still non-random, but that the magnitude of the deviations is much smaller than before the pump maintenance. The non-random distribution of residuals indicates that there is still

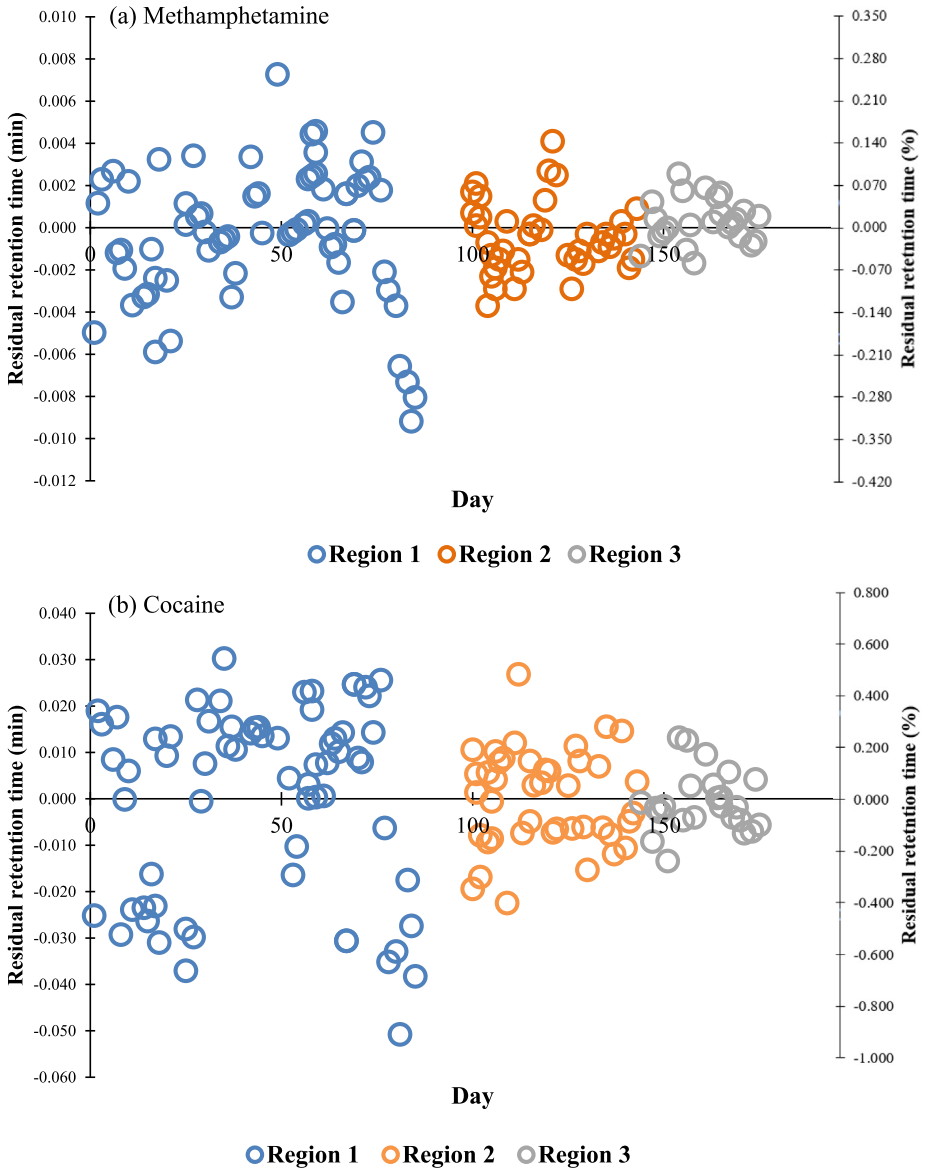

Fig. 2. Residuals from the linear regression lines for (a) methamphetamine and (b) cocaine from lab A showing both absolute and relative units.

at least one other unknown systematic factor that influences the retention time of these drugs. However, the factor is not so systematic that it can be easily modelled or corrected for.

On day 144, the GC septum was replaced again, but with no discernable systematic change in the retention time for methamphetamine. However, the retention time for cocaine increased by 0.037 min (0.7%), despite the fact that these retention times are extracted from the same chromatograms. The systematic decrease in retention times over the next 30 days is slightly different than before the septum change, with slopes of 0.002 min/day for both methamphetamine and cocaine and coefficients of determination (R^2) of 0.73 and 0.08, respectively.

A bivariate plot of the z-score normalized cocaine retention time versus the z-score normalized methamphetamine retention time (Fig. S1) showed that the correlation between the two retention times was quite weak; the first and third regions gave coefficients of determination (R^2) of 0.1, and the second region gave a coefficient of determination of 0.4. The slopes for the z-score normalized ranged from 0.3 to 0.6 but, once denormalized, ranged from 1.1 to 1.6. These correlations indicate that the retention time for cocaine is influenced to a greater extent than methamphetamine, and that a decrease in retention time for methamphetamine leads to a significantly larger decrease in retention time for cocaine. A second bivariate plot of the z-score normalized cocaine retention time versus the z-score normalized hydromorphone retention time (Fig. S2) showed that the correlation between the two retention times was very strong. The R^2 values for all three regions were 0.90 or above. The conclusion about these relationships is that compounds that spend a longer time on the column experience more of the random and systematic deviations from the instrumental parameters. Specifically, compounds that have similar retention times have correlating deviations from an expected (e.g. mean) value within a run. Additionally, because methamphetamine does not appear to be affected the same as cocaine and hydromorphone by the instrumental parameter(s) causing the fluctuations, it is likely that the variance does not arise from any process in the injection port, and is most likely caused by slight deviations in oven temperature programming.

Overall, the retention time behavior of repeated measurements of these two drug standards (methamphetamine and cocaine) in

one instrument shows a variety of sources of error. Some sources, like column bleed, lead to long-term and small systematic effects that can be modelled and predicted, with systematic decreases on the magnitude of 0.002 min/day. Other sources of error, like a failing turbo pump, can lead to short periods of non-random behavior that results in deviations that are also on the magnitude of 0.001 min/day. A septum change on day 144 did not cause any systematic change in retention time for methamphetamine, but did lead to a small increase (0.037 min) in retention time for cocaine. The remaining, random deviations, discussed in more detail below, are on the order of magnitude of 0.002 min.

The retention time analysis described for methamphetamine and cocaine was also conducted for other drug standards on all five instruments in all three laboratories. The results are provided in a spreadsheet in the supplemental material. One obvious take-home message from this comprehensive analysis is that the standard retention time of a drug reference material should be re-established after instrument maintenance to account for changes in retention time caused by systematic deviations. A particular example of this is the septum change on day 144, which caused a systematic deviation on the magnitude of 0.04 min or 0.7% for the cocaine standard retention time (Fig. 1b). An explanation for why this particular septum change caused a systematic deviation of a larger magnitude than other septum changes is that the septum wasn't aligned properly, which caused a small leak leading to decreased flow and ultimately, a longer retention time. The impact of trimming the column must also be accounted for, as demonstrated by the 0.04 min decrease from region 1 to region 2 of Fig. 1a.

Fig. 3 is a plot of the average retention time per week over the blue data set (day 1–85) from the methamphetamine data set. The error bars show the ~95% confidence interval of the measurements, which is defined as the expanded uncertainty of 2σ. Note that the error bars do not show the confidence interval of the means ($2\sigma/\sqrt{N}$). The number of measurement varied from N = 4 to N = 8 per week. The mean weekly retention time, shown by the height of the blue bars in Fig. 3, emphasizes the general decrease in retention time as a function of week, which is caused by slow column bleed. The plot also shows that the within-week 2σ confidence interval is relatively constant over the 12-week period of the study. One exception is week eight, wherein the drug standard was only analyzed four times during the week with no measurable change in the retention time (measured to 0.001 min). The 2σ value in week eight is considered unreliable and should be replaced by a pooled 2σ value of 0.005 min. Fig. 3 also shows that the error bars overlap in every pair-wise comparison,

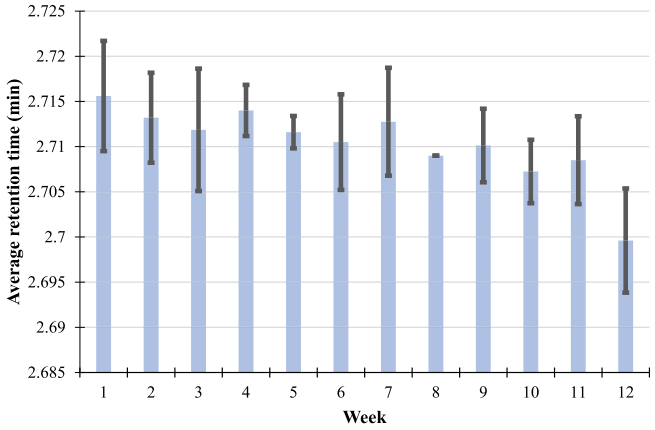

Fig. 3. Methamphetamine average retention time per week over the first 85 days of lab A.

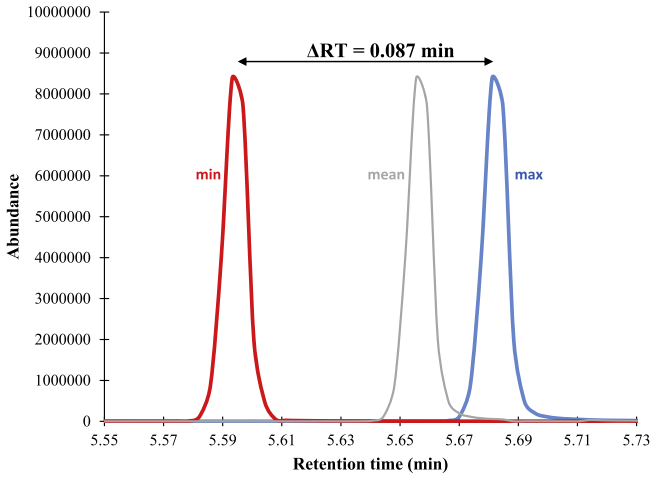

Fig. 4. Extreme retention time values observed for a cocaine standard analyzed on one instrument in one laboratory over the course of three months. Blue = day 35 at 5.68 min, red = day 81 at 5.59 min. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

except for week 12, where the long-term systematic error and the short term non-random behavior of the failing turbo pump take effect. The overlap in error bars is an indication that the retention time measurements are not significantly different over 11 weeks of study, but that week 12—when the turbo pump was failing, was significantly different from weeks 1–11. One-way analysis of variance (ANOVA) was used through the SPSS software package and it was determined that there were not consecutive weeks were significant differences existed except in pairwise comparisons involving weeks 8 and 12, which is explained by the observations described above.

Once the linear regions were identified, the maximum and minimum retention times for each linear region were extracted. Fig. 4 shows the extreme retention times observed for cocaine over the course of three months from lab A. The blue chromatogram with a retention time (RT) of cocaine at 5.68 min was collected on day 35. The red chromatogram at RT = 5.59 min was collected on day 81. The total difference in retention time is 0.087 min, which is still smaller than the ±2% within-day criterion used by many agencies. For cocaine, the ±2% window would result in an absolute acceptance window of ~0.113 min. However, the pooled within-month variability of retention times for cocaine is 0.033 min or 0.6% (pooled 2σ based on an average of N = 19 data points per month). This means that, when considering the average cocaine retention time over three months (5.65 min), the extreme data point at 5.59 min would fall outside the lower limit of the 95% confidence interval of 5.62 min. Extreme data points are expected to fall outside the 95% confidence interval about 5 out of every 100 analyses.

Fig. 5 is a simulated overlay of chromatograms to visualize how the measured retention times for 6-monacetylmorhine (6-MAM) from lab B compare with the recommended acceptance criteria from different agencies. In Fig. 5, the acceptance criteria are applied relative to the average retention time for all measurements for 6-MAM made over the three-month period, shown as the black peak. The green peaks show the positions of the measured maximum and minimum retention times for 6-MAM. The measured range for 6-MAM was ±0.016 min, and the measured 95% confidence interval (2σ) over the three-month period was ±0.015 min. The red peaks show the hypothetical acceptance limits of ±0.10 min recommended by WADA. The blue peaks show the hypothetical acceptance limits of ±2% (0.17 min) recommended by UNODC. The measured three-month confidence interval of N = 16 measurements shows that the recommendations of WADA and UNODC are

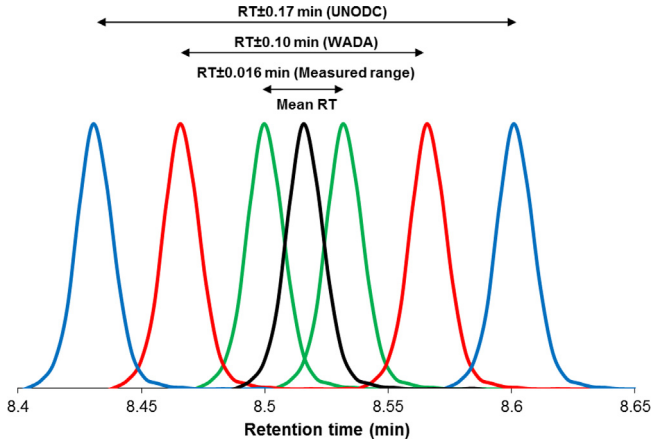

Fig. 5. Comparison between the extreme retention time values—which exceed the 95% confidence interval—for 6-monoacetylmorphine (6-MAM) collected over three months from lab B (green) and the ±0.1 min (red) and ±2% (blue) retention time acceptance criteria of WADA and UNODC, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

extremely conservative and fall way outside the realistic uncertainty.

Table 2 demonstrates that the average within-week and within-month 2σ values for each drug analyzed in each laboratory. In every case, the average within-week and within-month 2σ values are considerably smaller than the currently-recommended acceptance criteria. With the exception of lab C2, the average within-instrument, within-week 2σ values are very similar to the respective within-month 2σ values. The aberration for instrument C2 is likely due to an imbalance in sample size for the average within-month 2σ calculation (see raw data in supplemental material). Additionally, a pooled 2σ for the entire data set has been calculated to display an overall 2σ. However, it must be noted that because absolute retention time has an effect on the 2σ values, a pooled 2σ for the entire data set is skewed by compounds with longer retention times. For the record, the within-*day* 2σ values, which are of most relevance to practitioners, will be even smaller (i.e. better) than the within-week 2σ values provided here, which are themselves already smaller than the recommended acceptance criteria.

Table 2
Comparison of the average within-week and within-month variability of measured retention times and the acceptance criteria of different agencies. Measured values are expressed as 2σ in percent (and absolute in mins).

Lab	Average within-week 2σ of RT[#]	Average within-month 2σ of RT
A	±0.43% (0.021 min)	±0.53% (0.026 min)
B	±0.17% (0.014 min)	±0.17% (0.014 min)
C1[$]	±0.05% (0.002 min)	±0.054% (0.003 min)
C2	±0.14% (0.015 min)	±0.33% (0.036 min)
C3	±0.14% (0.013 min)	±0.18% (0.017 min)
Combined	±0.23% (0.018 min)	±0.29% (0.024 min)
Organization	**Acceptance Criteria**	
USDA (2017)	±0.1 min	
AORC (2016)	greater of ±1% or ±6 s	
NCDOJ (2016)	±2%	
EC (2015/2013/2002)	±0.1 min	
FDA (2015)	±0.2 min or ±2.5%	
WADA (2015/2010)	smaller of ±0.1 min or ±2%	
ASTM (2014)	±6%	
GTFCh (2009)	±2%	
UNODC (2009)	±2%	
CLSI (2002)	Smaller of ±0.2 min or ±1%	

[$] Fast GC method using a 10 m microbore (0.15 mm ID) column.
[#] Average calculated as the geometric mean (pooled 2σ), not the arithmetic mean.

Instrument C1 provided the most reproducible retention times, and this instrument was a conventional Agilent GC, operated with a narrow-bore capillary column; i.e. fast GC mode. The within-week 2σ interval was on the magnitude of 0.05%, which is 40 times smaller than the widely-accepted rule-of-thumb of ±2%. When the Agilent GC, and other GCs, were operated in conventional mode, the within-week 2σ intervals were generally less than 0.20%, which is in close agreement with the value of <0.35% provided by Kelly and Bell [43], and an order of magnitude smaller than the commonly used limit of ±2%.

The above discussion makes the case that the acceptance criteria recommended by most agencies are unreasonably conservative and do not seem to be based on realistic uncertainties. One might argue that casework samples could include matrix effects and concentration differences that could provide greater deviations in retention times. For example, when the conjugate acid form of methamphetamine (e.g. the hydrochloride salt) is dissolved in methanol for GC–MS analysis, it is common to observe erratic peak broadening and retention time behavior, which is why many crime labs require amphetamines to be converted to their basic forms before GC analysis. However, with the exception of amphetamines, GC retention times are widely-demonstrated to be independent of matrix effects and other analytes. In contrast, concentration is known to influence retention times of substances to a very modest degree [48]. The effect of concentration has been reported to cause systematic changes in peak retention times no larger than 0.02 min (0.1%) for a two-orders-of-magnitude change in concentration. Therefore, as long as the concentration of a questioned sample and a reference sample are within two orders of magnitude, the effect of concentration on retention time will be negligible. This assumption assumes that the laboratory standard operating procedures dictate appropriate steps to handle column overloading, including dilution, reanalysis, and appropriate documentation [11,49].

In this study, laboratories C1, C2, and C3 analyzed six NBOMe isomers at two different concentrations (125 ppm and 1250 ppm). The effect of concentration over a single-order-of-magnitude for the within-week 2σ ranges between ±0.004 and 0.022% (±0.0002–0.002 min). For the within-month 2σ, the effect of concentration over a factor of ten is between ±0.008 and 0.024% (±0.0004–0.003 min), which is considerably less variance than reported for the two-order-of-magnitude variance of ±0.1% reported for selected FAMEs [48]. For additional evidence that concentration has a negligible effect on retention times, consider that the retention index for compounds in the NIST database are reported independent of concentration, rather than as a range of values spanning a range of analyte concentration. The implied assumption is that retention index and retention time are practically independent of concentration.

3.2. Relative ion abundance analysis

Fig. 6a and b show the relationship between the average within-week 2σ and the average relative abundance of different fragment ions in different drugs. The 2σ values are reported here as 2×(%RSD). For clarification, ion abundances are reported on a relative scale in this study, as with other studies. For the sake of clarifying the accepted terminologies, we remind the reader that the percent relative ion abundance is typically considered an absolute scale [18]. Therefore, an ion with 50% relative abundance relative to the base peak could have an uncertainty of ±10% on the absolute scale and ±20% on the relative scale

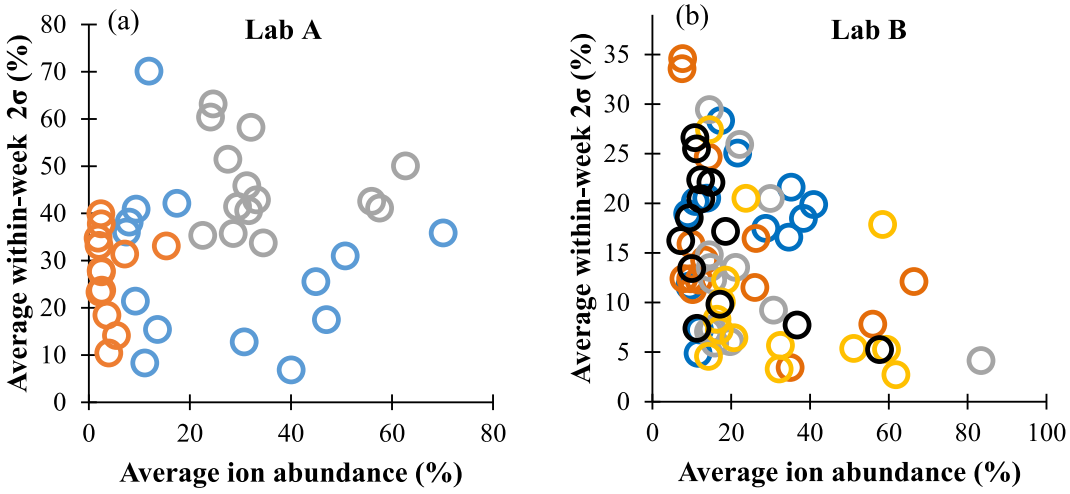

Fig. 6. Relationship between the average relative ion abundance and the average within-week 2σ for labs A & B expressed as a percentage of the relative ion abundance. Markers are color coded by drug. Lab A: blue = cocaine, orange = methamphetamine, and gray = hydromorphone. Lab B: blue = cocaine, orange = ecgonine methyl ester, gray = 6-MAM, yellow = DAM, and black = fentanyl. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(10% abs/50% abs × 100% rel. = 20% rel.). In both cases, the range would be 40–60% relative ion abundance.

Fig. 6a shows data for three substances analyzed in lab A over the course of six months and Fig. 6b shows data for five substances analyzed in lab B over the course of three months. Neither plot shows a strong correlation between the average within-week 2σ and relative ion abundance; the regression lines had R^2 values smaller than 0.22 for all data sets. However, there does appear to be a relationship between the *largest* observable within-week uncertainty and the average ion abundance. This trend is indicated by the lack of data points in the upper right-hand side of Fig. 6b. These results indicate that there *is* a weak experimental justification for establishing different acceptance criteria for different relative ion abundances, such as in guidelines set forth by WADA [14], but that analysts should be aware that the rule-of-thumb uncertainties provided by WADA do not apply to all low-abundant fragment ions. Some low-abundance fragment ions display very tight tolerances, such as within-week expanded uncertainties (2σ) smaller than or 10% relative or ±1% absolute.

The low-abundance ion at *m/z* 146 for fentanyl has within-week 2σ values of 5% (relative), whereas other low abundance peaks, like *m/z* 57 for fentanyl have within-week 2σ values of 19% (relative). The largest measured variance was for the fragment ion of fentanyl at *m/z* 207, which had an average ion abundance of 6.7%. The measured within-week expanded uncertainty (2σ) was ±84% (relative, or ±6% absolute). Several of the ions with largest uncertainties were of fragment ions that overlap with common background ions (e.g. *m/z* 44 for CO_2 and *m/z* 207 for column bleed). The uncertainties are therefore likely to be confounded by the uncertainty in the instantaneous contribution from residual background ions at the same *m/z* values.

To provide a visual tool for the comparison of the measured 2σ uncertainty of relative ion abundance and the recommended guidelines, Fig. 7 compares the variation observed from the 25C-NBOMe ortho isomer at *m/z* 150 from instruments C1-C3 with some of the currently applied acceptance criteria. The blue, red, and green box-and-whisker plots show the interquartile range and maximum variation observed for the 25C-NBOMe ortho isomer at *m/z* 150 from instruments C1, C2, and C3. In comparison, the gray and yellow box-and-whisker plots are the relative ion abundance acceptance criteria for ASTM (±20% relative) and WADA (±10% absolute), respectively, and are based on the mean data from instrument C3. The variation observed from all three instruments

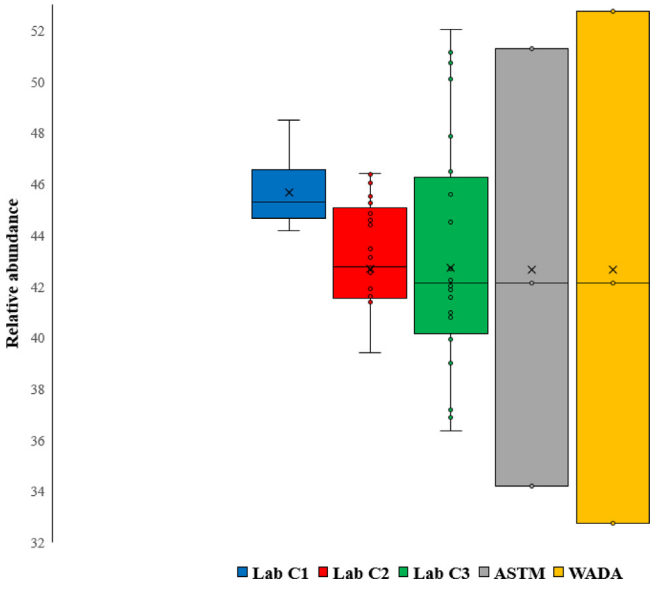

Fig. 7. Comparison between the measured relative ion abundance extremes observed in 25C-Ortho NBOMe at *m/z* 150 for instruments C1-C3 and the acceptance criteria for relative ion abundance from ASTM (±20% relative) and WADA (±10% absolute) based on the results of lab C3.

is smaller than that of the acceptance criteria recommended by ASTM and WADA [14,18], particularly for instruments C1 and C2. This observation indicates that the acceptance criteria are wider than analytically necessary—as was observed for the retention time criteria—and therefore minimizes the risk of type II errors (false negatives). Four outliers were cut off from lab C2 and C3 due to expansion of the y-axis. All four values fell outside of the ASTM and WADA acceptance criteria, but this occurred only four times out of 86 total measurements, or 4.7%, which is around the expected 5% error for a 95% confidence interval.

To determine the uncertainty of measurement for the relative ion abundance data, the average within-week and within-month 2σ values were calculated for the 12–14 most abundant fragment ions of every drug standard in every laboratory (Table 3). Table 3 also shows some commonly-used acceptance criteria as a comparison. The average within-week and within-month 2σ are similar in

Table 3
Comparison of the average within-week and within-month 2σ relative ion abundances with the different agency-recommended acceptance criteria. For reference; ±20% in the absolute uncertainty scale (e.g. USDA standard) is the same as: 1) ±40% relative uncertainty for a peak at 50% absolute abundance; and 2) ±80% relative error for a peak at 25% absolute abundance.

Lab	Average within-week 2σ (relative %)	Average within-month 2σ (relative %)
A	±32	±38
B	±19	±21
C1	±19	±20
C2	±21	±23
C3	±34	±40
Organization	Acceptance Criteria (within day)	
	Absolute (%)	Relative (%)
USDA (2017)	±20	
AORC (2016)	±10	
EC (2015)	±30	
IFSTL (2005)	±30	
FDA (2015)		±20
ASTM (2014)		±20
UNODC (2009)		±20
CLSI (2002)		±20

Table 4
Comparison of the relative ion abundance binning results and the WADA acceptance criteria.

Relative abundance range	Lab	Average within-week 2σ (relative %)	Average within-month 2σ (relative %)
>50%	Combined	±20	±22
	A	±37	±41
	B	±8	±10
	C1	±6	±6
	C2	±6	±9
	C3	±17	±17
	WADA	±10% (absolute) or ± 10–20% (relative)	
25–50%	Combined	±30	±33
	A	±35	±39
	B	±14	±15
	C1	±7	±7
	C2	±18	±17
	C3	±24	±23
	WADA	±20% (relative)	
<25%	Combined	±39	±46
	A	±41	±48
	B	±17	±20
	C1	±8	±8
	C2	±16	±18
	C3	±33	±37
	WADA	±5% (absolute) or ±20–100% (relative)	

magnitude for all laboratories, with two of the labs performing within than the criteria and three labs performing outside the criteria. The average within-week and within-month 2σ values are relatively constant throughout the study for each lab, with three instruments providing uncertainties (2σ) in the range of 19–23% (relative) and two labs in the range 30–40% (relative). Some instruments simply provide superior reproducibility over time. Again, it is important to remember that the within-day 2σ values would be smaller than the within-week 2σ values. The crime laboratories in this study did not typically analyze multiple reference materials each day, so the within-day reproducibility could not be assessed.

The effect of concentration on the relative ion abundances was also studied through the analysis of 12 NBOMe isomers analyzed in methods C1, C2, and C3 at concentrations of 125 ppm and 1250 ppm. The difference between the within-week 2σ reported as relative percentages were determined to be 6% for method C1, 5% for method C2, and 18% for method C3, while the difference between the within-month 2σ uncertainties (reported relative percentages) were 5% for method C1, 4% for method C2, and 20% for method C3. Concentration effects the reproducibility much less for methods C1 and C2 than method C3, where concentration describes about half of the variance. This discrepancy points to differences in the tune frequency or tune performance of different mass spectrometers and methods.

Several prominent organizations use a binning system to establish acceptance criteria for the uncertainty of replicate measurements. In binning systems, the magnitude of the acceptable range of variance is related to arbitrary ranges (bins) of relative ion abundance. These organizations include the FBI, SWGTOX, the European Commission (2013), and WADA [9,14,36,38]. Table 4 is a comparison of the average within-week and within-month 2σ averaged across all labs with the bins of acceptance criteria provide by WADA [14]. When broken down by individual laboratory, laboratories A and C3 fall outside the WADA recommendations on a within-week and within-month basis, but laboratories B, C1, and C2 perform better than the WADA recommendations for all three relative ion abundance bins. Furthermore, even with laboratories A and C3 falling outside of the WADA acceptance criteria, the combined within-week average for the entire data set is roughly in agreement with the acceptance criteria for all relative abundance bins. The average within-week and within-month 2σ was slightly above the WADA acceptance criteria for the range 25–50% (relative). The average within-week and within-month 2σ of 5% (absolute) for the range of abundances <25% (absolute) is highly dependent on the abundance of the particular ion, ranging from 20 to 100% relative uncertainty (e.g. ±20% relative uncertainty at 25% (absolute) abundance to ±100% relative) uncertainty at ±5% (absolute) abundance. Once more, the within-day 2σ would be even tighter than the calculated within-week 2σ values.

Any conversation about the uncertainty of relative ion abundances would be incomplete without a discussion of the correlation that exists between ions within a mass spectrum. Fig. 8a and b are bivariate plots of two ions from the different cocaine mass spectra collected from lab A over the course of 6 months. These bivariate plots demonstrate coefficient-of-correlation (R^2) values greater than 0.90, which indicates a very strong correlation between these pairs of ions. An assumption made in all the recommendations for uncertainty of measurements (of relative ion abundances) is that the uncertainty in the abundance of one fragment is independent of the uncertainty of all other fragment ion abundances. Fig. 8 indicates that this important assumption is invalid.

For example, Fig. 8a shows that the mean relative ion abundance at m/z 105 is around 45%. All mass spectrometry guidelines suggest comparing ion abundances to an 'exemplar' reference spectrum, which is assumed to represent the mean of a set of reference spectra. However, if, in a given spectrum, the relative ion abundance of the fragment at m/z 77 is at ~30%, the observed correlation in Fig. 8a makes it highly improbable than the relative ion abundance of the ion at m/z 105 would exceed 40%. In other words, the abundance at m/z 77 correlates so strongly with m/z 105 that we can be very confident that when the abundance at m/z 77 is at ~30%, the abundance of the fragment at m/z 105 will be smaller than its mean value of ~45%. In this case, the abundance of the ion at m/z 77—along with its known correlation with the abundance at m/z 105—would be a better predictor of the abundance of the ion at m/z 105 than the mean or exemplar value. For this reason, instead of comparing ion abundances to their mean or exemplar values, we should instead consider the correlation of an ions' abundance with all the other ion abundances in the spectrum. Although computationally difficult, such consideration would minimize the uncertainty—or confidence intervals—of measured ion abundances, and would therefore minimize the possibility of type I errors.

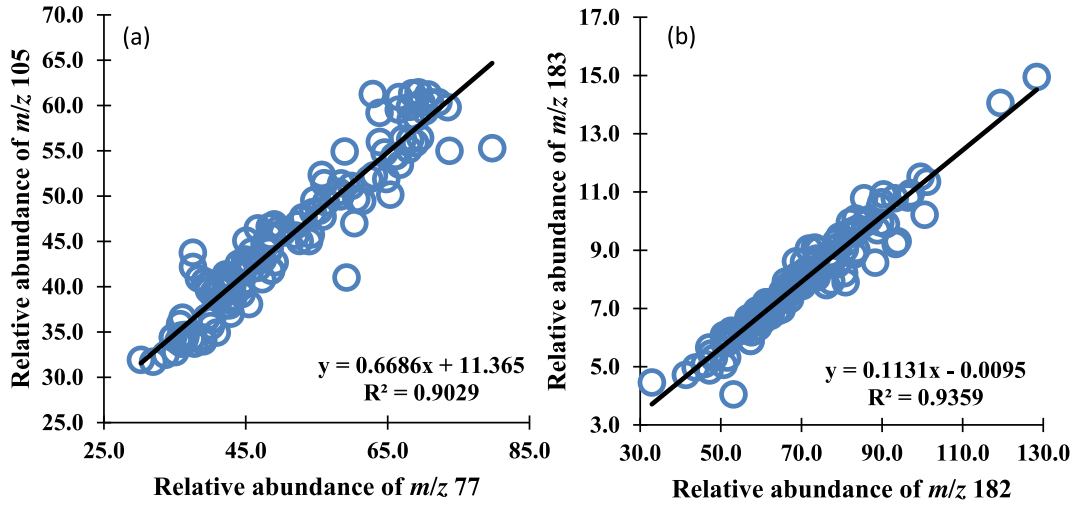

Fig. 8. Bivariate plots demonstrate strong correlations between different fragment ions: (a) *m*/*z* 77 and *m*/*z* 105 and (b) *m*/*z* 182 and *m*/*z* 183 for cocaine analyzed by lab A. This correlation, among others, invalidates the assumption that the abundance at each *m*/*z* value is an independently variable. To the abundance at other *m*/*z* values,

Another interesting use of the regression lines is that, as shown in Fig. 8b, the slope of the regression line represents the typical ratio of abundance between *m*/*z* 183 and *m*/*z* 182. In this case, because there are no known isobars at these *m*/*z* values [50], the slope of the two relative ion abundances is the ratio of the abundance of the ^{13}C isotope [A + 1] to the abundance of the ^{12}C isotope at *m*/*z* 182 [A]. The number of carbon atoms (n_c) in an ion can be estimated from $n_c(1.1\%) = [A + 1]/[A]$ [51]. In this case, the slope of 0.1131 informs us that there are 11.3%/1.1% ≈ 10 carbon atoms in these fragments, as has been demonstrated through isotope labelling and accurate mass measurements [50]. This discussion shows that correlation analysis between different fragment ions has the potential to inform analysts about elemental compositions in addition to the other structural information.

4. Conclusions

This manuscript provides a comparison between the acceptance criteria of many leading governing bodies with actual, measured data representing a variety of instrumental parameters. The data consists of replicate measurements of drug standard mixtures that were analyzed multiple times per week over several months during routine casework in two crime laboratory settings. A third laboratory, in a university setting, analyzed a mixture of drug standards hundreds of times on three different instrument configurations. Retention time data and relative ion abundance measurements are found to contain different sources and magnitudes of random error, but the magnitude of this random uncertainty was found to be smaller than the magnitude of systematic uncertainties, such as those caused by turbo pump conditions, column maintenance and, occasionally, a GC septum replacement. Most GC septum replacements had no effect on retention times.

Careful analysis of the residual retention times within regions of linear behavior revealed that, in some cases, the uncertainties in retention times were short periods of non-random rather than random deviations, especially when the turbopump needed maintenance. Still, the magnitudes of the random, systematic, and short periods of non-random and deviations were all less than 0.50% (2σ), even when combined. The magnitude of the measured retention time uncertainties at the most extreme cases are therefore approximately 4 times smaller than the acceptance criteria currently recommended by most agencies, which are typically in the region of ±2% or ±0.1 min. The use of unreasonably-large acceptance criteria in the currently available acceptance criteria has the consequence of minimizing type II errors (false negatives) and maximizing the possibility of type I errors (false positives), which is widely accepted to be an unfavorable position for a crime laboratory. To optimize their decision-making processes, laboratories should assess the actual uncertainty of retention time measurements on each instrument, and they should use their own measured uncertainties to guide drug identifications that involve GC retention times. Of course, when EI-MS data is used in conjunction with the GC retention times, the EI-MS data should provide the selectivity required to exclude false positives, when GC–MS is being applied as part of an analytical scheme.

The relative ion abundance data from replicate measurements of drug standards for five different instruments and three different laboratories agree quite well with the WADA recommendations based on measured within-week uncertainties. However, some instruments perform better than others. For example, laboratories A and C3 seemed to demonstrate larger variance than the other instruments/labs, possibly because of the natural mass spectral variation on these instruments or because of the selected tuning parameters [43].

The results of this study demonstrate that the retention time acceptance criteria currently recommended by a variety of governing bodies are substantially broader than the average within-week and within-month 2σ values measured in several different laboratories. In contrast, most recommendations for EI-MS fragment ion abundances are much closer to the average uncertainties measured across all instruments and are therefore much better approximations for the typical confidence intervals (2σ) observed in typical seized drug settings. As a general rule, fragment ion abundances within a spectrum are not independently variable, and it is common to find ions separated by more than 2 Da that have correlation coefficients (R) that exceed 0.9. In light of this finding, search algorithms that assume that fragment ions are independent variables should be re-evaluated to test the effect of correlated ion abundances. In the future, the uncertainty of ion abundancies could be accounted for, or minimized, through consideration of the correlations that exists between different fragment ion pairs. Such considerations are the target of future work.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.forc.2018.07.001.

References

[1] G.P. Jackson, Error terror in forensic science: when spectroscopy meets the courts, Spectroscopy 31 (11) (2016) 2–4.
[2] Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) Recommendations Version 7-1, U.S.D.o.J.D.E. Administration, Editor, 2016.
[3] International, A., Standard Practice for Quality Assurance of Laboratories Performing Seized-Drug Analysis, 2015.
[4] International, A., Standard Practice for Identification of Seized Drugs, 2017
[5] K. Sahil, B. Prashant, M. Akanksha, S. Premjeet, R. Devashish, Gas Chromatography–Mass Spectrometry Applications, Int. J. Pharm. Biol. Sci. Arch. 2 (6) (2011) 1544–1560.
[6] Uncertainty of Measurement: Implications of Its Use in Analytical Science. Analyst 120(9), 1995.
[7] R. Allain, Error Analysis: Systematic vs Random. Available from: https://www2.southeastern.edu/Academics/Faculty/rallain/plab193/labinfo/Error_Analysis/05_Random_vs_Systematic.html.
[8] European Communities OJEC 17.8.2002. Official Journal of European Communities, 2002.
[9] European Commission, Health & Consumer Protection Directorate-General: Guidance Document on Analytical Quality Control and Validation Procedures for Pesticide Residues Analysis in Food and Feed, 2013.
[10] S.J. Lehotay, USDA-FSIS: Update of Qualitative Analysis Considerations in GC-MS(/MS).
[11] German Society for Toxicological and Forensic Chemistry: Guideline for Quality Control in Forensic-Toxicological Analyses, 2009.
[12] United Nations, Office on Drugs and Crime: Guidance for the Validation of Analytical Methodology and Calibration of Equipment used for Testing of Illicit Drugs in Seized Materials and Biological Specimens, 2009.
[13] North Carolina Department of Justice: Gas Chromatography-Mass Spectrometry (GC-MS) Data Processing, 2016.
[14] WADA Technical Document: Minimum Criteria for Chromatographic-Mass Spectrometric Confirmation of the Identity of Analytes for Doping Control Purposes, 2015.
[15] FDA Food and and Veterinary Medicine Science and Research Steering Committee: Acceptance Criteria for Confirmation of Identity of Chemical Residues Using Exact Mass Data within the Office of Foods and Veterinary Medicine, 2015.
[16] Clinical and Laboratory Standards Institute: Gas Chromatography/Mass Spectrometry (GC/MS) Confirmation of Drugs; Approved Guideline 22(22), 2002.
[17] Association for Official Racing Chemists: Guidelines for the Minimum Criteria for Identification by Chromatography and Mass Spectrometry, 2016.
[18] ASTMD6420: Standard Test Method for Determination of Gaseous Organic Compounds by Direct Interface Gas Chromatography–Mass Spectrometry, 2014.
[19] M. Holcapek, R. Jirasko, M. Lisa, Basic rules for the interpretation of atmospheric pressure ionization mass spectra of small molecules, J. Chromatogr. A 1217 (25) (2010) 3908–3921.
[20] A. Weissberg, S. Dagan, Interpretation of ESI(+)-MS-MS spectra—Towards the identification of "unknowns", Int. J. Mass Spectrom. 299 (2–3) (2011) 158–168.
[21] R.A. de Zeeuw, Substance identification: the weak link in analytical toxicology, J. Chromatogr. B 811 (1) (2004) 3–12.
[22] R. Bethem, J. Boison, J. Gale, D. Heller, S. Lehotay, J. Loo, S. Musser, P. Price, S. Stein, Establishing the fitness for purpose of mass spectrometric methods, J. Am. Soc. Mass Spectrom. 14 (5) (2003) 528–541.
[23] S.E. Rodriguez-Cruz, R.S. Montreuil, Assessing the quality and reliability of the DEA drug identification process, Forens. Chem. 6 (2017) 36–43.
[24] R.A. de Zeeuw, Letter to the Editor – fitness for purpose of mass spectrometric methods in substance identification, J. Forensic. Sci. 50 (3) (2005).
[25] F.W. McLafferty, D.A. Stauffer, S.Y. Loh, C. Wesdemiotis, Unknown identification using reference mass spectra. Quality evaluation of databases, J. Am. Soc. Mass Spectrom. 10 (1999) 1229–1240.
[26] W. Demuth, M. Karlovits, K. Varmuza, Spectral similarity versus structural similarity: mass spectrometry, Anal. Chim. Acta 516 (1–2) (2004) 75–85.
[27] S.E. Stein, D.R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification, J. Am. Soc. Mass Spectrom. 5 (1994) 859–866.
[28] S. Stein, Mass spectral reference libraries: an ever-expanding resource for chemical identification, Anal. Chem. 84 (17) (2012) 7274–7282.
[29] J. Zhang, X. Wei, C. Zheng, B. Wang, F. Wang, P. Chen, Compound identification using random projection for gas chromatography–mass spectrometry data, Int. J. Mass Spectrom. 407 (2016) 16–21.
[30] E.L. Schymanski, C.M. Gallampois, M. Krauss, M. Meringer, S. Neumann, T. Schulze, S. Wolf, W. Brack, Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties, Anal. Chem. 84 (7) (2012) 3287–3295.
[31] Y.M. Du, Y. Hu, Y. Xia, Z. Ouyang, Power normalization for mass spectrometry data analysis and analytical method assessment, Anal. Chem. 88 (6) (2016) 3156–3163.
[32] E.L. Schymanski, M. Meringer, W. Brack, Matching structures to mass spectra using fragmentation patterns- are the results as good as they look?, Anal Chem. 81 (2009) 3608–3617.
[33] L. Gatto, K.D. Hansen, M.R. Hoopmann, H. Hermjakob, O. Kohlbacher, A. Beyer, Testing and validation of computational methods for mass spectrometry, J. Proteome. Res. 15 (3) (2016) 809–814.
[34] M.A. Bodnar Willard, R. Waddell Smith, V.L. McGuffin, Statistical approach to establish equivalence of unabbreviated mass spectra, Rapid. Commun. Mass Spectrom. 28 (1) (2014) 83–95.
[35] M.A. Bodnar Willard, V.L. McGuffin, R.W. Smith, Statistical comparison of mass spectra for identification of amphetamine-type stimulants, Foren. Sci. Int. 270 (2016) 111–120.
[36] Federal Bureau of Investigation: Guidelines for Comparison of Mass Spectra, 2007.
[37] United States Department of Agriculture Agricultural Marketing Service, Science and Technology Pesticide Data Program: SOP, 2017.
[38] Scientific Working Group for Forensic Toxicology : Standard for Mass Spectral Data Acceptance for Definitive Identification, 2014.
[39] International Food Safety: Considerations in Regulatory Applications of Mass Spectrometry in Food Safety.
[40] European Commission Directorate-General for Health and Food Safety: Guidance Document on Analytical Quality Control and Method Validation Procedures for Pesticides Residues Analysis in Food and Feed, 2015.
[41] Environmental Protection Agency: Method 8260B Volatile Organic Compounds by Gas Chromatography/Mass Spectrometry (GC/MS), 1996.
[42] U.S. Food and Drug Administration: Mass Spectrometry for Confirmation of the Identity of Animal Drug Residues, 2003.
[43] K. Kelly, S. Bell, Evaluation of the reproducibility and repeatability of GCMS retention indices and mass spectra of novel psychoactive substances, Forens. Chem. 7 (2018) 10–18.
[44] Evaluation of measurement data – Guide to the expression of uncertainity in measurement, 2010.
[45] C.A. Valdez, R.N. Leif, S. Hok, A. Alcaraz, Assessing the reliability of the NIST library during routine GC-MS analyses: structure and spectral data corroboration for 5,5-diphenyl-1,3-dioxolan-4-one during a recent OPCW proficiency test, J. Mass Spectrom. (2018).
[46] N. Etxebarria, O. Zuloaga, M. Olivares, L.J. Bartolome, P. Navarro, Retention-time locked methods in gas chromatography, J. Chromatogr. A 1216 (10) (2009) 1624–1629.
[47] I. Rasanen, I. Kontinen, J. Nokua, I. Ojanperä, E. Vuori, Precise gas chromatography with retention time locking in comprehensive toxicological screening for drugs in blood, J. Chromatogr. B 788 (2) (2003) 243–250.
[48] R. Kapeller, Retention time correction in gas chromatography by modeling concentration related effects, applied to the analysis of fatty acid methyl esters, J. Chromatogr. A 1394 (2015) 118–127.
[49] Forensic Toxicology Laboratory Office of Chief Medical Examiner City of New York, Fentanyl by Solid Phase Extraction and Gas Chromatography/Mass Spectrometry, version 08.31.2015, accessed at https://www1.nyc.gov/site/ocme/services/toxicology-technical-manuals.page.
[50] R.M. Smith, J.F. Casale, The mass spectrum of cocaine: deuterium labeling and MS/MS studies, Microgram 7 (1) (2010) 16–41.
[51] F.W. McLafferty, F. Turecek, in: Interpretation of Mass Spectra Fourth Edition, University Science Books, 1993, pp. 19–35.