Printed in the Republic of Korea

ANALYTICAL SCIENCE & TECHNOLOGY

Vol. 39 No. 4, 256-265, 2026

https://doi.org/10.5806/AST.2026.39.4.256

Automated R-based data processing workflow for GC-MS quantitative confirmation of methamphetamine, 3,4-methylenedioxymethamphetamine, and their metabolites in hair

Seon Yeong Kim, Jae Chul Cheong, and Jin Young Kim★

Forensic Genetics and Chemistry Division, Supreme Prosecutors' Office, Seoul 06590, Republic of Korea

(Received June 16, 2026; Revised July 15, 2026; Accepted July 18, 2026)

Abstract: In forensic toxicology, the reliable quantitative confirmation of illicit drugs in hair is critical; however, conventional manual data evaluation using proprietary instrument software introduces operational bottlenecks and risks of human error. This study developed a fully automated, vendor-neutral data processing workflow for the quantitative analysis of methamphetamine, amphetamine, 3,4-methylenedioxymethamphetamine, and 3,4-methylenedioxyamphetamine by gas chromatography–mass spectrometry (GC-MS). To eliminate vendor lock-in, proprietary raw data from multiple platforms were standardized into the open-source mzML format. A customized R script was subsequently employed to perform automated base-to-base peak integration, unweighted linear regression, and concentration back-calculation without manual intervention. The automated evaluation demonstrated robust analytical performance, yielding excellent linearity ($R^2 > 0.999$) over a dynamic range of 0.1–5.0 ng/mg. The automatically calculated accuracies and precisions satisfied forensic bioanalytical acceptance criteria (±15 %, and ±20 % at the lower limit of quantification). Application to authentic hair samples confirmed that the R script successfully enforced definitive identification parameters, including retention time consistency (±2 %) and quantifier-to-qualifier ion ratios (±20 %), and generated consolidated quantitative reports. By systematically replacing manual spreadsheet manipulations, this open-source workflow enhances analytical throughput, ensures rigorous data integrity, and facilitates long-term data archiving, establishing a scalable foundation for broader multi-vendor forensic applications.

Key words: Automated data processing, mzML, Hair analysis, Amphetamine type stimulants, GC-MS

1. Introduction

The increasing prevalence of drug-related crimes and addiction to illicit substances has established forensic drug testing as an essential component of criminal investigations and judicial proceedings. Among the biological matrices available for forensic drug testing, hair offers a distinct advantage over urine and blood in that it provides a retrospective record of long-term drug exposure, making it particularly

★ Corresponding author
Phone : +82-(0)2-3480-2152 Fax : +82-(0)2-535-4175
E-mail : paxus@spo.go.kr

valuable as legal evidence.[1] Methamphetamine (MA) and 3,4-methylenedioxymethamphetamine (MDMA), which represent the most frequently examined substances in forensic drug analysis in South Korea, are internationally analyzed in hair by gas chromatography–mass spectrometry (GC-MS). This approach involves the simultaneous quantification of both the parent drugs and their respective metabolites, amphetamine (AP) and 3,4-methylenedioxyamphetamine (MDA), following extraction from the hair matrix.[2] Quantitative analysis is performed by calculating the peak area ratio of each target compound to its corresponding isotopically labeled internal standard (IS), from which the drug concentration is derived. The IS method effectively compensates for variability in extraction recovery, matrix effects, and fluctuations in injection volume and ionization efficiency; therefore, it is essential for ensuring the reproducibility and accuracy of the analytical results.[3]

Qualitative identification of illicit drugs in forensic toxicology requires more than the presence of a chromatographic peak. To minimize false-positive results and maintain result reliability, at least three GC-MS selected ion monitoring (SIM) traces per compound—comprising one quantifier ion and two qualifier ions—must be simultaneously confirmed. Additionally, the retention time (RT) of each target compound must remain within a defined tolerance, and the ion ratios between the quantifier and qualifier ions must satisfy guideline-specified acceptance criteria (e.g., ±20 %) in both quality control (QC) and authentic hair samples.[4,5] This multi-ion verification process reduces the risk of spectral interference from co-eluting substances and is a fundamental requirement for supporting the admissibility of forensic evidence in legal proceedings.

Although most commercially available instrument software provides automated quantification, ion ratio verification, RT and QC assessment, they have structural limitations in that it does not readily facilitate the integrated comparison of blank, QC, and real samples within a single page. Specifically, cross-referencing data among blank, QC, and real samples often requires the user to manually toggle between multiple data-viewing windows, which disrupts the continuity of high-throughput routine analysis. Furthermore, these platforms generally lack the capacity to simultaneously verify the complex identification criteria described above or to automatically generate reports in the format required for routine laboratory reporting. To address these limitations, many analysts currently manually transfer raw data to spreadsheet software and perform the verification of peak area ratios, ion ratios, and concentration back-calculations. Such manual procedures not only introduce operational bottlenecks but also substantially increase the risk of transcription and calculation errors. The absence of systematic, end-to-end documentation of the data processing steps compromises traceability and represents a potential threat to data integrity—a primary requirement of international quality standards such as ISO/IEC 17025.[6] Previous efforts to address these issues through laboratory information management systems (LIMS) or Excel-based macros (Visual Basic for Applications, VBA) have been reported; however, these approaches involve considerable implementation costs and often encounter compatibility issues following instrument software updates, making long-term maintenance challenging.[7]

A further practical consideration is that forensic laboratories commonly operate instruments from multiple manufacturers, each requiring its own proprietary software, which places a considerable training burden on analysts. Overcoming this vendor dependency and consolidating the data processing workflow into a unified system necessitates the adoption of a universal, open-format data standard. The mzML format, developed by the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI), is an XML-based open standard capable of losslessly encoding raw data from virtually all mass spectrometry platforms. Because it preserves metadata and spectral information irrespective of the instrument manufacturer, mzML provides a consistent basis for data extraction and is well-suited to long-term archiving.[8]

In the present study, the R programming environment together with the mass spectrometry-oriented packages

xcms and mzR were employed to extract the relevant data from converted mzML files and to automate the analytical data processing workflow in accordance with laboratory-specific requirements. These open-source packages enable algorithm-based, automated base-to-base peak integration and area calculation across large chromatographic datasets without manual operator input.[9] Although a degree of programming proficiency is required and initial parameter optimization presents a learning threshold, the transparency of data processing criteria makes this approach well-suited to the rigorous demands of forensic bioanalytical validation. Once a script-based workflow has been established, it can substantially improve data processing efficiency by automating repetitive manual operations.[10]

The objective of this study was therefore to develop a script-based, customized automated reporting workflow for the quantitative confirmation of MA, AP, MDMA, and MDA in hair—the most frequently requested analyses in forensic drug testing in South Korea, thereby eliminating the manual computational steps that have represented a bottleneck in multi-instrument data processing. The automated reporting system integrates the following functions into a single workflow: (1) automated identification of parent drug–metabolite pairs, (2) verification of ion ratios and retention times against guideline-specified acceptance criteria, (3) baseline-corrected peak integration and QC assessment, and (4) generation of consolidated quantitative output reports. By automating the entire process from raw data extraction to final report generation, the system eliminates the possibility of arbitrary data manipulation or manual entry errors, ensures processing reproducibility, and maintains data integrity throughout. This approach is expected to reduce the software training burden associated with multi-vendor instrument operation and to contribute to the speed, objectivity, and reliability of forensic toxicological results within the criminal justice system.

2. Experimental

2.1. Chemicals and reagents

The reference standards for amphetamine (AP), methamphetamine (MA), 3,4-methylenedioxyamphetamine (MDA), and 3,4-methylenedioxymethamphetamine (MDMA) were obtained from Cerilliant (Austin, TX, USA) as 250 μg/mL methanolic solutions. Deuterated internal standards (IS) —namely AP-d_8, MA-d_{11}, MDA-d_5, and MDMA-d_5—were also purchased from Cerilliant at 100 μg/mL. Heptafluorobutyric anhydride (HFBA), used as the derivatization reagent, was supplied by Acros Organics (Geel, Belgium). Solvents, including water, methanol, ethyl acetate, and acetone, were of HPLC-grade and purchased from Fisher Chemical (Fair Lawn, NJ, USA). Working standard solutions (0.02, 0.1, 0.4, and 1.0 μg/mL) and a combined IS solution (0.3 μg/mL) were prepared by serial dilution with methanol.

2.2. Hair sample preparation

Hair samples were washed sequentially with methanol and acetone to remove exogenous contaminants. After drying, approximately 10 mg of the dried hair was pulverized in a 2-mL polypropylene tube (Eppendorf, Hamburg, Germany) containing steel beads using a Qiagen TissueLyser II (Retsch, Haan, Germany). Then, 1.0 mL of methanol and 50 μL of the IS solution were added to each sample. Following sonication and centrifugation, the supernatant of the hair extract was filtered and evaporated under a stream of nitrogen. The residue was then derivatized with 50 μL of HFBA and 50 μL of acetone at 50 °C for 30 min to facilitate GC-MS analysis. The resulting derivatives were reconstituted with 50 μL of ethyl acetate and subjected to instrumental analysis. The detailed sample preparation procedure, including hair pulverization and extraction conditions, was performed according to the method previously reported by Kim *et al*..[11]

2.3. GC–MS analysis and data acquisition

The derivatized extracts were analyzed using an Agilent 8890 GC system coupled with a 5977B MSD (Palo Alto, CA, USA) operated in selected ion monitoring (SIM) mode. Chromatographic separation was achieved on Agilent DB-5MS UI column (30 m × 0.25 mm i.d., 0.25 μm film thickness) with a

temperature program optimized for the resolution of the target analytes and their respective ISs. The injector and interface temperatures were maintained at 250 °C and 280 °C, respectively. Helium was used as the carrier gas at a constant flow rate. These instrumental operating conditions, including the temperature program and SIM parameters, were configured according to the established protocol described by Kim *et al.*.[11]

2.4. Data conversion and R script

To overcome the limitations of vendor-specific software and establish a universally compatible analytical workflow, the proprietary raw data files acquired from the GC-MS system were converted into the open-source, XML-based mzML format. This conversion was batch-processed using the MSConvert tool from the ProteoWizard software suite, ensuring that all mass spectrometric metadata were comprehensively preserved without loss of quantitative fidelity (*Fig.* 1).[12]

The standardized mzML files were processed using a custom-built script developed in the R programming environment (version 4.6.0). This script utilized the mzR package for efficient, low-level access to the XML structures of the mzML files, facilitating rapid parsing of chromatographic metadata. To standardize the quantitative confirmation process, the xcms and MSnbase packages were integrated to perform consistent peak detection and base-to-base integration (*Fig.* 2). The automated workflow was configured to execute the following steps: (1) identification of target compounds based on user-defined RT windows, (2) extraction of peak areas and heights for the quantifier and qualifier ions, (3) calculation of ion ratios for qualitative confirmation, and (4) regression analysis for quantitative determination using weight-corrected peak area ratios relative to the corresponding IS. All analytical parameters, including baseline correction settings and integration thresholds, were programmatically defined to ensure consistent data processing across multi-vendor raw data files (Supplementary material).

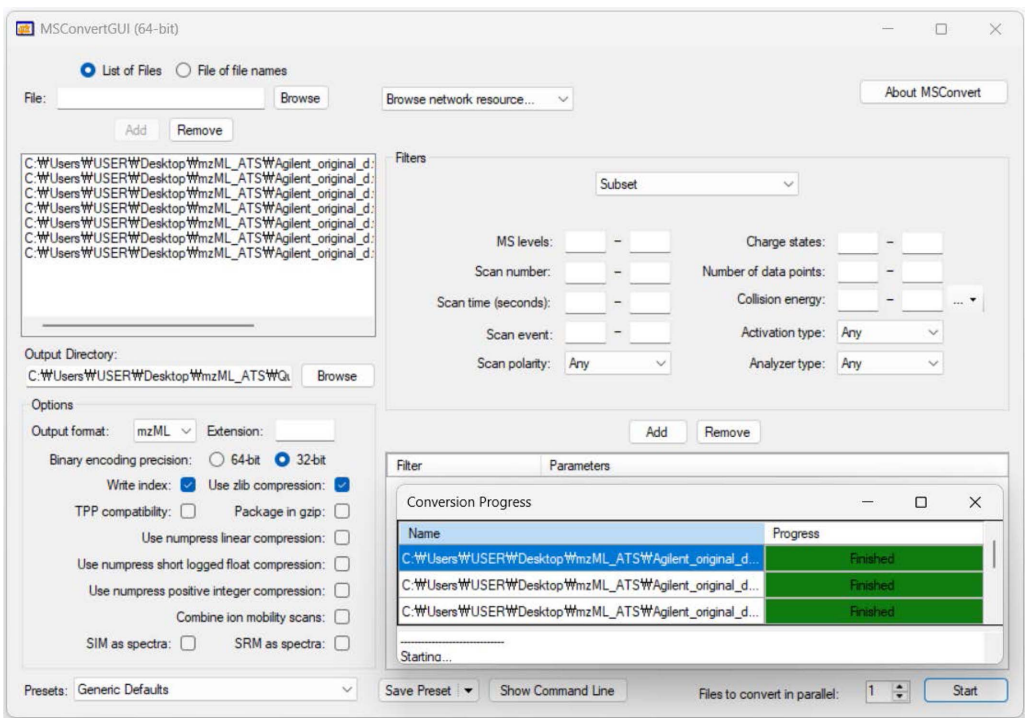

Fig. 1. Graphical user interface of MSConvert used for data conversion.

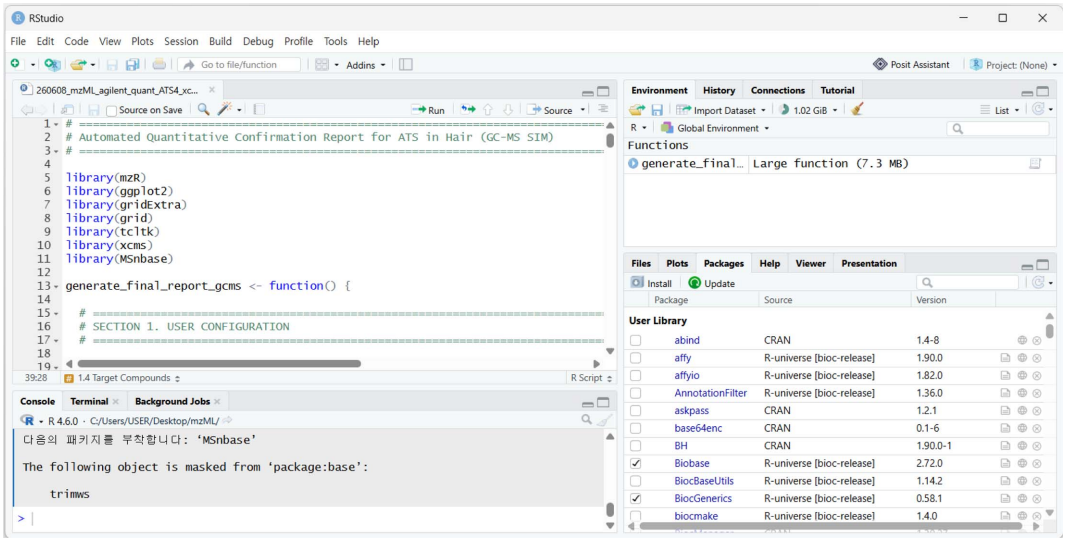

Fig. 2. Representative screenshot of the custom R script developed for automated GC-MS data processing, showing the loaded libraries (mzR, xcms, MSnbase, and ggplot2) and the primary function structure used in this study.

2.5. Automated quantitative confirmation

For the quantitative confirmation of MA, AP, MDMA, and MDA, the algorithmic workflow was designed to perform precise peak integration and concentration calculations directly from the standardized mzML files. Initially, the script parsed the chromatogram header metadata, employing a custom search function to identify and extract the specific SIM traces—comprising one quantifier ion and two qualifier ions per target analyte—within a mass-to-charge tolerance of ±0.5 *m/z* to account for minor instrumental mass drifts.

Upon targeting the correct SIM channels, the algorithm isolated the chromatographic data within a predefined RT window symmetrically centered around the expected RT for each compound. Within this localized window, background baseline noise was programmatically subtracted, and peak areas and heights were computed using an automated base-to-base integration algorithm. To compensate for matrix effects and variations during sample preparation inherent to solid hair matrices, the script performed internal standardization by calculating the target-to-IS peak area ratios.

To quantify the analytes, the software automatically parsed the four-point calibration data (0.1, 0.5, 2.0, and 5.0 ng/mg). An unweighted ordinary least squares linear regression was applied via the lm() function in R, without incorporating statistical weighting factors, as the variance remained homoscedastic within this defined concentration range. Once the slope and y-intercept were obtained, the script inverted the linear regression equation ($y = ax + b$) to calculate the concentrations of the target analytes in authentic hair samples.

Furthermore, to ensure compliance with international forensic guidelines—such as those of the Society of Hair Testing (SoHT)—the R script applied rigorous tolerances for definitive identification. Specifically, it verified that the RTs and the quantifier-to-qualifier ion ratios of the authentic samples remained within the acceptable variance limits (±2 % for RT and ±20 % for ion ratio) relative to the QC samples.

To maximize the efficiency of data review and minimize human error, the final confirmation report was generated as a consolidated PDF. The report was configured to sequentially display the linear calibration curves, followed by the integrated chromatograms and calculated quantitative tables for the authentic hair, QC, and blank samples. This arrangement enabled immediate visual verification of peak integration, quantification results, and potential contamination.

By algorithmically enforcing these established bioanalytical parameters, the script ensures that the routine evaluation of authentic samples remains standardized, thereby supporting the evidentiary integrity, throughput, and reproducibility of the final forensic reports.

3. Results and Discussion

3.1. MS data conversion

To establish a universal data processing workflow, the proprietary binary data acquired from the Agilent GC-MS system were converted into a universally compatible format. The MSConvert utility was used for this conversion process. MSConvert offers a practical advantage in laboratory workflows owing to its batch-processing capability. Because Agilent instruments output raw data as structured .D directories rather than single files, the batch-queuing capability of MSConvert allowed entire folders of individual samples to be loaded and converted in a single automated operation, thereby streamlining the data preparation phase.

The output format was standardized to the mzML format, which is ratified by the HUPO-PSI. The adoption of this open-source standard eliminates vendor lock-in and provides a consistent data structure suitable for automated downstream processing. In implementing this conversion, the configuration was optimized to minimize the storage footprint while preserving data integrity; specifically, 64-bit binary encoding precision was coupled with zlib compression, effectively reducing file sizes without sacrificing quantitative fidelity.

Additionally, precise data filtering during the conversion phase was necessary for subsequent automated analysis. Because the R algorithms rely on discrete centroid data for objective peak determination, a peak-picking filter was integrated into the MSConvert pipeline. By selecting the vendor algorithm with the MS level designated as 1, precise centroiding was achieved across the entire SIM acquisition range. Upon execution, these parameters yielded standardized, compressed mzML files in the designated local directory, establishing a reproducible, vendor-neutral foundation for the automated data processing workflow.

3.2. Quantitative linearity and accuracy assessment

The calibration curves constructed for all target analytes displayed excellent linearity within the specified concentration range of 0.1–5.0 ng/mg. The unweighted ordinary least squares linear regression analysis yielded coefficients of determination (R^2) exceeding 0.999 for MA, AP, MDMA, and MDA, confirming a reliable concentration–response relationship. The accuracy and precision of the calculated concentrations were programmatically evaluated against the established acceptance criteria. At the lower limit of quantification (LLOQ, 0.1 ng/mg), the intra- and inter-day accuracies remained within ±20 % of the nominal values, while the precision (expressed as relative standard deviation, %RSD) did not exceed 15 %. For the QC and calibration points, both accuracy and precision were within the acceptable ±15 % threshold. These data demonstrate that the automated R-based regression algorithm provides sufficient quantitative rigor and back-calculation accuracy for routine forensic applications.

3.3. Automated quantitative confirmation reporting

Following the validation of the algorithm, the workflow was applied to authentic hair samples to demonstrate its practical utility in routine forensic toxicology. The target compounds—specifically, MA and its primary metabolite AP as well as MDMA and its metabolite MDA—were successfully evaluated. Among these, MA and AP were selected as representative examples to illustrate the automated quantitative confirmation reporting owing to their high prevalence and detection frequency in forensic cases in South Korea.

For definitive positive identification, the R script programmatically verified two critical parameters: RT and ion ratios. Target analytes were required to elute within a strict tolerance (e.g., ±2 %) relative to

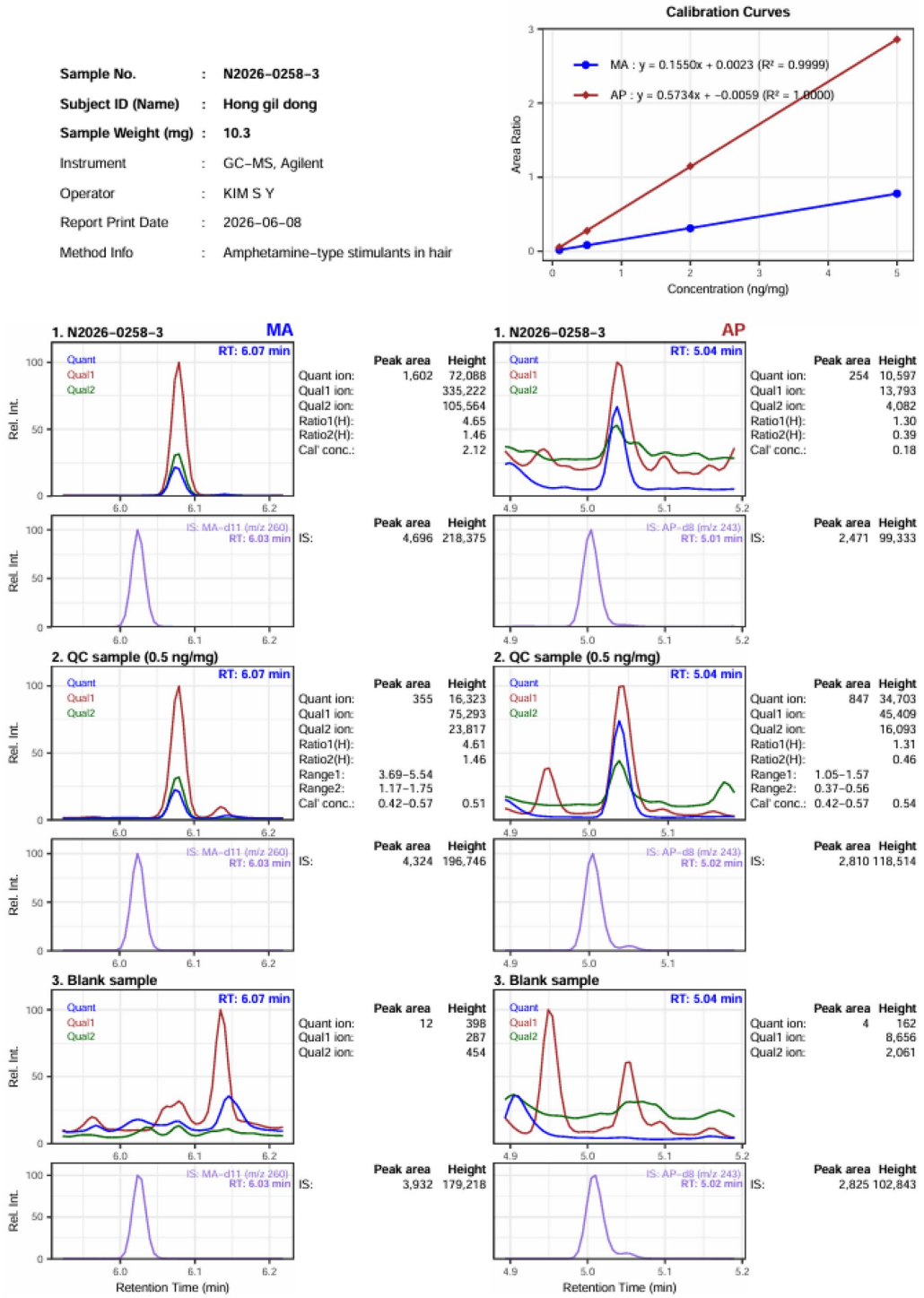

Fig. 3. Automated data analysis report for the identification and quantitative confirmation of methamphetamine (MA), amphetamine (AP), 3,4-methylenedioxymethamphetamine (MDMA), and 3,4-methylenedioxyamphetamine (MDA) obtained via the automated R script using standardized mzML files.

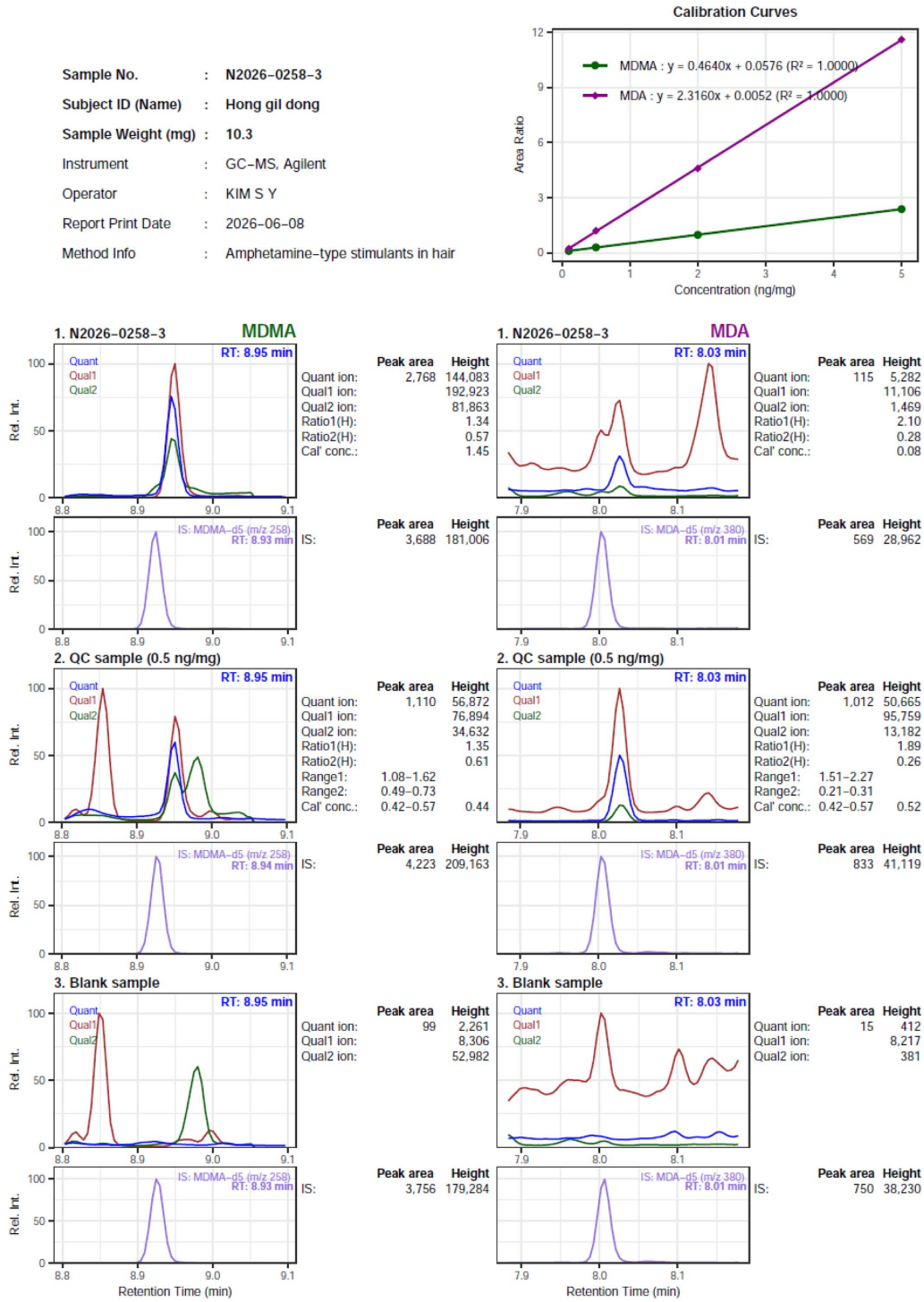

Fig. 3. Continued.

the reference RT established by the QC samples. Furthermore, the ion ratio between the quantifier and qualifier ions, monitored in SIM mode, was required to remain within ±20 % of the batch-specific mean ratio derived from the calibration standards. Any integrated peaks exceeding these tolerances were automatically flagged to minimize the risk of false-positive results.

The generated confirmation report systematically aligned the calibration curve and the SIM chromatograms with the quantitative and identification data of blank, QC, and real samples vertically, while positioning the parent drug and its metabolites horizontally side-by-side. This multi-dimensional layout facilitated rapid visual assessment of peak symmetry, integration baseline, and RT consistency, enabling intuitive determination of drug intake at a glance (*Fig.* 3). Furthermore, the inclusion of QC back-calculation accuracy results within the same report significantly enhanced the reliability of the quantitative analysis.

In a representative positive subject sample, both MA and AP were identified above the quantification limits, with all RTs and quantifier-to-qualifier ion ratios satisfying the established Society of Hair Testing (SoHT) criteria. These results demonstrate that the mzML-based automated workflow can efficiently handle high-throughput hair sample analysis with rigorous analytical transparency.

3.4. Cross-platform applicability

A major methodological advantage of the proposed workflow is its ability to overcome the technical limitations imposed by proprietary mass spectrometry data formats. In practical forensic laboratory environments—which frequently operate heterogeneous analytical platforms such as Agilent, Shimadzu, and Thermo Fisher Scientific GC-MS systems—vendor lock-in presents a significant operational bottleneck. To address this, the conversion of proprietary raw binary data into the open-source mzML standard was utilized. This conversion process ensured that multi-vendor files were standardized without compromising mass spectral metadata, confirming the universal applicability of the data conversion step.

In the present study, the downstream automated evaluation—using the customized R script—was empirically validated using the converted mzML data acquired from the Agilent GC-MS system. The algorithm successfully extracted the SIM metadata and generated definitive quantitative confirmation reports, demonstrating that the workflow maintains rigorous analytical reliability. While the automated reporting script was applied to the Agilent platform in this phase, the core cross-platform applicability of our framework has already been comprehensively validated across various vendor systems in our previous study.[13] Future studies will focus on extending the application of the R script to encompass the converted data from these additional platforms. Implementing this open-source workflow effectively decouples the instrumental data acquisition phase from the subsequent bioanalytical evaluation. Ultimately, our proposed workflow systematically addresses these immediate operational bottlenecks by consolidating multi-sample cross-referencing and regulatory-compliant report generation into a single, automated, and vendor-neutral interface.

Furthermore, forensic toxicology demands the strict, long-term preservation of evidentiary data. Reliance on the vendor-neutral mzML format ensures the secure archiving and seamless re-evaluation of historical analytical results over extended periods. This standardized approach protects critical forensic data from the obsolescence of vendor-specific software versions, ensuring that data integrity is maintained for future judicial and scientific review.

4. Conclusion

In this study, a vendor-natural, fully automated data processing workflow was successfully developed and implemented for the GC-MS quantitative confirmation of MA, MDMA, and their primary metabolites in forensic hair analysis. By standardizing proprietary binary data into the open-source mzML format and employing a customized R script, the proposed workflow effectively decouples downstream data

evaluation from vendor-specific instrument software.

The algorithmic automation of base-to-base peak integration, calibration curve regression, and concentration back-calculation eliminates the need for manual spreadsheet manipulations. This approach enhances high-throughput operational efficiency, thereby mitigating the risk of human error inherent in manual data transcription and calculation. Furthermore, the programmatic enforcement of forensic validation criteria—such as RT tolerances and quantifier-to-qualifier ion ratios—ensures that the generated reports maintain objectivity, transparency, and data integrity, which are essential for evidentiary acceptance in the criminal justice system.

While the current quantitative confirmation reporting script was empirically validated using Agilent GC-MS data for specific amphetamine-type stimulants, the successful mzML standardization of raw data from various vendor systems confirms the cross-platform potential of this workflow. Future research will focus on expanding the application scope of this automated workflow to encompass a broader spectrum of illicit drugs, therapeutic agents, and new psychoactive substances (NPS). The implementation of this standardized analytical workflow is expected to streamline routine laboratory operations, reduce software training burdens, and strengthen the scientific reliability of forensic toxicological investigations.

Supplemental Material

The R script for automated quantitative confirmation of amphetamine (AP), methamphetamine (MA), 3,4-methylenedioxyamphetamine (MDA), and 3,4-methylenedioxymethamphetamine (MDMA) by GC-MS is available at https://github.com/paxus11.

Conflicts of Interest

The authors declare no competing financial interest.

References

1. C. Ferreira, C. Paulino, and A. Quintas, *Chem. Res. Toxicol.*, **32**(12), 2367-2381 (2019). https://doi.org/10.1021/acs.chemrestox.9b00301
2. G. A. A. Cooper, R. Kronstrand, and P. Kintz, *Forensic Sci. Int.*, **218**(1-3), 20-24 (2012). https://doi.org/10.1016/j.forsciint.2011.10.024
3. F. T. Peters, O. H. Drummer, and F. Musshoff, *Forensic Sci. Int.*, **165**(2-3), 216-224 (2007). https://doi.org/10.1016/j.forsciint.2006.05.021
4. S. M. R. Wille, F. T. Peters, V. Di Fazio, and N. Samyn, *Accred. Qual. Assur.*, **16**, 279-292 (2011). https://doi.org/10.1007/s00769-011-0775-0
5. S. P. Elliott, D. W. S. Stephen, and S. Paterson, *Sci. Justice*, **58**(5), 335-345 (2018). https://doi.org/10.1016/j.scijus.2018.05.004
6. K. C. Tsimillis, *Accred. Qual. Assur.*, **24**, 165-171 (2019). https://doi.org/10.1007/s00769-018-1366-0
7. S. Kanza, C. Willoughby, N. Gibbins, R. Whitby, J. G. Frey, J. Erjavec, K. Zupančič, M. Hren, and K. Kovač, *J. Cheminform.*, **9**(1), 31 (2017). https://doi.org/10.1186/s13321-017-0221-3
8. L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Römpp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P. A. Binz, and E. W. Deutsch, *Mol. Cell Proteomics*, **10**(1), R110.000133 (2011). https://doi.org/10.1074/mcp.R110.000133
9. C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, *Anal. Chem.*, **78**(3), 779-787 (2006). https://doi.org/10.1021/ac051437y
10. L. Gatto and A. Christoforou, *Biochim. Biophys. Acta*, **1844**(1), 4251 (2014). https://doi.org/10.1016/j.bbapap.2013.04.032
11. J. Y. Kim, S. H. Shin, J. I. Lee, and M. K. In, *Forensic Sci. Int.*, **196**, 43-50 (2010). https://doi.org/10.1016/j.forsciint.2009.12.045
12. D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick, *Bioinformatics,* **24**(21), 2534-2536 (2008). https://doi.org/10.1093/bioinformatics/btn323
13. N. H. Kown, J. C. Cheong, and J. Y. Kim, *Biomed. Chromatogr.*, **40**(6), e70465 (2026). https://doi.org/10.1002/bmc.70465