Electronic Supplementary Information

Integration of ultra-high-pressure liquid chromatography-tandem mass spectrometry with machine learning for identifying fatty acid metabolite biomarkers of ischemic stroke

Lijian Zhang,[‡a] Fei Ma,[‡b] Ao Qi,[a] Lulu Liu,[a] Junjie Zhang,[a] Simin Xu,[a] Qisheng Zhong,[c] Yusen Chen,[*a] Chun-yang Zhang,[*b] Chun Cai[*a]

[a] Analytical center, Neurology Department of Affiliated Hospital, Institute of Neurology, Guangdong Medical University, Zhanjiang, Guangdong 524023, China.

[b] Collaborative Innovation Center of Functionalized Probes for Chemical Imaging in Universities of Shandong, Key Laboratory of Molecular and Nano Probes, Ministry of Education, Shandong Provincial Key Laboratory of Clean Production of Fine Chemicals, College of Chemistry, Chemical Engineering and Materials Science, Shandong Normal University, Jinan 250014, China.

[C] Shimadzu Global COE for Application& Technical Development, Guangzhou, Guangdong, 510010, China

* Correspondence author. E-mail: caichun2006@tom.com; cyzhang@sdnu.edu.cn; chenyusen925@163.com

‡ These authors contributed equally.

MATERIALS AND METHODS

Study population

This research was conducted at Department of Neurology, Guangdong Medical University Affiliated Hospital, and the protocols were approved by the Ethics Committee of Guangdong Medical University Affiliated Hospital. The consents of all subjects were obtained in accordance with the Declaration of Helsinki. From June 2015 to September 2016, a total of 218 patients were recruited. The subjects were required to be >40 years of age with first-ever ischemic stroke (confirmed by computed tomography or magnetic resonance imaging of the brain with 48 h of symptom onset). Besides, patients with recurrent stroke (n = 60) who had a history of stroke within previous 3 years (evidence of an acute disturbance of focal neurological functions, with symptoms lasting more than 24 hours) were admitted. Patients were excluded if they had severe heart failure, acute myocardial infarction, unstable angina, atrial fibrillation, aortic dissection, cerebrovascular stenosis. Patients who were in a deep coma or were treated with intravenous thrombolytic therapy were excluded as well. Additionally, a total of 204 healthy volunteers were recruited from the physical examination center of Guangdong Medical University Affiliated Hospital. The inclusion criteria for healthy volunteers were as follow: (1) normal hemodynamics of cerebral arteries, (2) no infarct region in brain magnetic resonance imaging or computed tomography examination, (3) age >40 years old, (4) no carotid artery wall lesions by neck vascular color doppler ultrasound; (5) without comorbidities (e.g., malignant tumors and severe organ dysfunction). In addition, patients and healthy volunteers were excluded if blood samples were not available or hemolyzed during collection. Finally, a total of 155 patients with first-ever ischemic stroke, 42 patients with recurrent ischemic stroke and 122 healthy volunteers were recruited in this

research. Baseline data about demographic characteristics, lifestyle risk factors, medical history, clinical features, and imaging data were collected at the time of enrollment. The National Institutes of Health Stroke Scale was used to evaluate the stroke severity at baseline by the trained neurologists. The level of triglyceride, total cholesterol, high-density lipoprotein, low-density lipoprotein, blood glucose and uric acid were collected. Patients were defined as overweight or obese if body mass index was ≥24.

Reagents and solutions preparation

Methanol (chromatographic grade, purity > 99.995%) and formic acid were purchased from Merck (Darmstadt, Germany). Purified water was obtained using a Milli-Q water system (Millipore, Massachusetts, USA). Isotope internal standards, which were purchased from Cayman Chemical (USA), consisted of 16 kind fatty acids and their metabolites including prostaglandin (PG) E2-d4, PGD2-d4, PGF2α-d4, 6-keto-PGF1α-d4, leukotriene C4-d5, leukotriene 4-d4, tetranor-PGEM-d6, 15-hydroxy-eicosatetraenoic acid (HETE)-d8, 5-HETE-d8, thromboxane B2-d4, 12-HETE-d8, platelet activating factor-d4, oleoylethanolamine-d4, eicosapentaenoic acid-d5, docosahexaenoic acid-d5 and arachidonic acid-d8. All internal standard solutions were mixed by methanol with a final concentration of 100 ng/mL.

Extraction of fatty acid metabolites from plasma

The 0.1 mL of plasma was mixed with 0.1 mL of methanol and 10 μL of internal standard solution, and the mixture was vortexed at 4 ℃ for 10 min and centrifuged at a speed of 13000 rpm for 5 min. The supernatant was mixed with 0.1% formic acid before extraction using strata-x polymerized solid reverse phase extraction columns (Phenomenex). Extraction program was as follows: column was washed by 1 mL of methanol, followed by adding 1 mL of 0.1% formic acid,

extracting sample, washing by the addition of 1 mL of 0.1% formic acid to remove the non-specific binding metabolites, and finally eluted into 0.2 mL of methanol. The samples were stored at -80℃ to prevent the degradation of metabolites.

Ultra-high-pressure liquid chromatography tandem mass spectrometry analysis

Analysis of fatty acids and their metabolites was carried out by using 8045 series ultra-high-pressure liquid chromatography tandem mass spectrometry (Shimadzu, Japan). Ultra-high-pressure liquid chromatography was performed with Phenomenex C8 (2.1 mm × 150 mm × 2.6 μm). The sample cooler and the column temperature were set at 5 ℃ and 40 ℃, respectively. The injection volume was 5 μL. Mobile phase consisted of water (containing 0.1% formic acid) and acetonitrile at a flow rate of 0.4 mL/min, and gradient elution program of acetonitrile was listed as follow: 10% (0 min) – 25% (5 min) – 35% (10 min) – 75% (20 min) – 95% (20.1 min) – 95% (25 min) – 10% (25.1 min) – 10% (30 min). The autosampler was programmed to aspirate 5 μL of sample, 1 μL of air, and 15 μL of water. Mass spectrometric measurements were performed with an electrospray ionization (ESI) source operating in both positive and negative mode using nitrogen as the nebulizer gas. The parameters of mass spectrometry were set as follows: nebulizer gas flow of 3 L/min, drying gas of 10 L/min, heat block temperature of 400 ℃, desolvation line temperature of 250 ℃, collision induced dissociation pressure of 230 kPa. For the quantitative analysis of the identified fatty acid metabolites, the internal standard method was employed based on the internal standard corresponding to the same fatty acid metabolite class.

Determination of conventional clinical biochemical index

The levels of triglyceride, total cholesterol, high-density lipoprotein, low-density lipoprotein,

blood glucose and uric acid were measured with AU2700 series automatic biochemical analyzer (Olympus Corporation, Japan).

Machine learning algorithms

The data sets were split into training set (75%) and validation set (25%). We randomly split data sets at the patient level so that patients cannot appear in both the training set and the validation set. Logistic regression model using stepwise variable selection with backward elimination was used to find biomarkers classifier that can discriminate health person from ischemic stroke patients as well as first-ever ischemic stroke patients from recurrent ischemic stroke patients, and the discriminant analytes were determined by using 10-fold cross-validation. The random forest machine learning algorithms were used to investigate the performance of the classifier in validation set.

Statistical analysis

All statistical tests and analyses were performed with R software, including ggplot2, caret, pROC and MASS packages. All subsequent analyses were performed on log-transformed to achieve normal distribution. Association of individual analytes with disease status was tested using t-test. For multiple testing, the p values from the plasma fatty acid metabolite analysis were further adjusted for the Benjamini–Hochberg false discovery rate, and p adjust value of < 0.05 was considered as significant difference.

SUPPLEMENTARY RESULTS

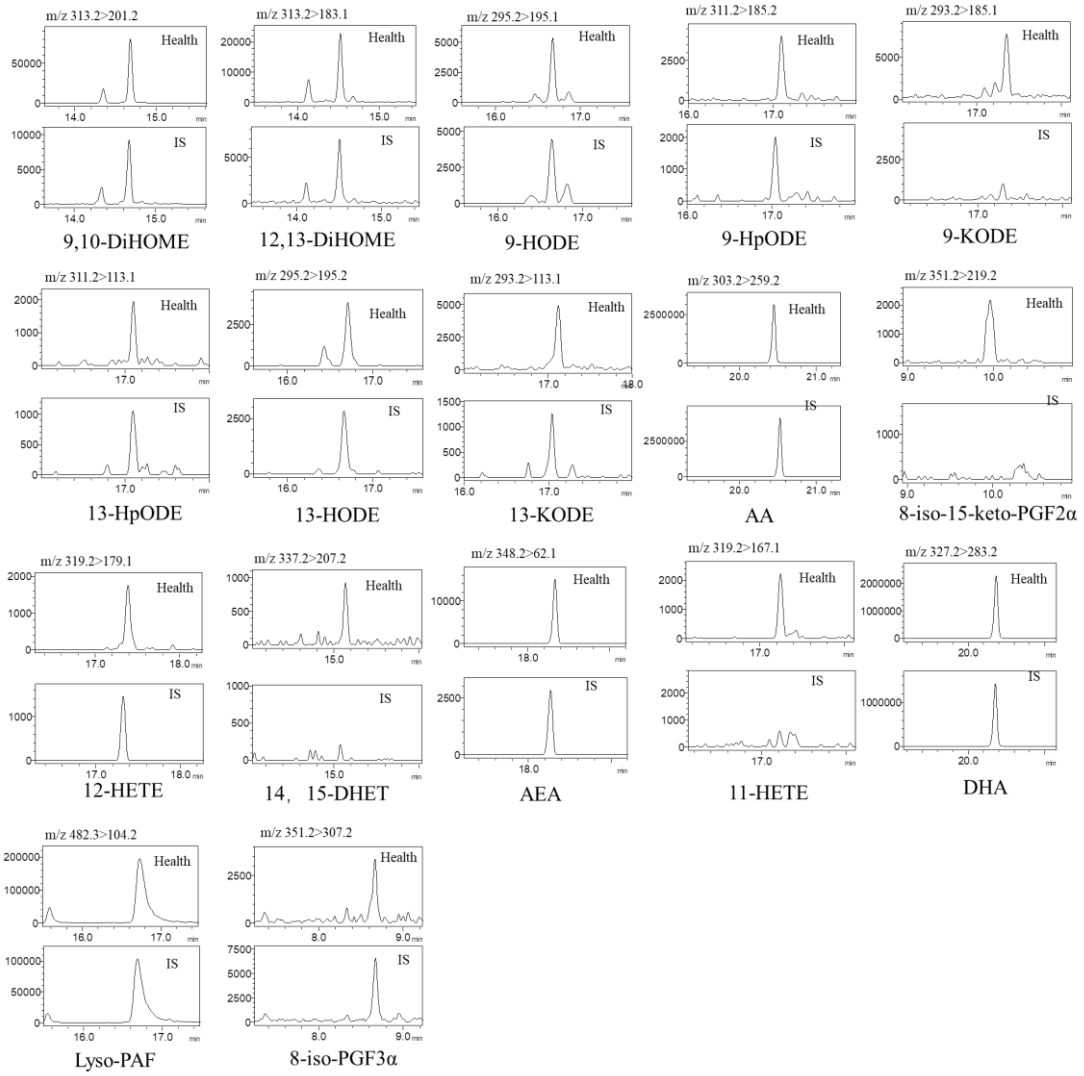


Fig. S1 Multiple reaction monitoring ion chromatogram of fatty acid metabolites between the ischemic stroke patients (IS) and the healthy volunteers.

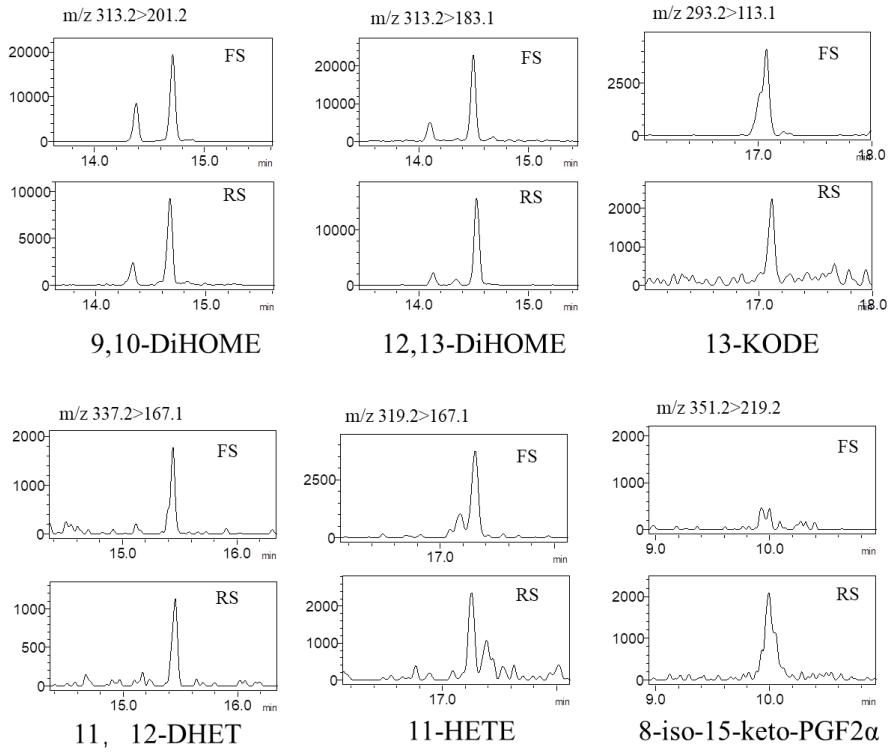


Fig. S2 Multiple reaction monitoring ion chromatogram of fatty acid metabolites between the first-ever (FS) patients and the recurrent ischemic stroke (RS) patients.

Using similar strategy, we tried to find the classifier that can discriminate first-ever and recurrent ischemic stroke. After the measurement of fatty acid metabolite levels by UHPLC-MS/MS (Fig. S2), difference between first ever and recurrent ischemic stroke were statistically analyzed. As shown in Fig. S3A, in the training set, eight fatty acid metabolites display significant different levels in first-ever and recurrent ischemic stroke. Specifically, the levels of three LA metabolites decreased in recurrent ischemic stroke compare with the first-ever ischemic stroke including two CYP450 products (i.e., 9, 10-DiHOME and 12, 13- DiHOME), and one 15-LOX pathway product (i.e, 13-KODE). The levels of two AA metabolites including one CYP450 pathway product (i.e., 11, 12-DHET) and one 11-LOX pathway product (i.e., 11-HETE) are downregulated in recurrent ischemic stroke, while the level of another AA metabolite produced

from COX and NE pathway (i.e., 8-iso-15-ketoPGF2α) was upregulated in recurrent ischemic stroke. Moreover, 8-iso-PGF3α increased and Lyso-PAF deceased in recurrent ischemic stroke, respectively. Besides fatty acid metabolites, two clinical biochemical parameters including LDL and TC significantly decreased in recurrent ischemic stroke compared with those in first-ever stroke (Fig. S3B). These results indicate that fatty acid metabolites (i.e., 9, 10-DiHOME, 12, 13-DiHOME, 13-KODE, 11, 12-DHET, 11-HETE, 8-iso-15-ketoPGF2α, 8-iso-PGF3α, and Lyso-PAF) and clinical biochemical parameters (i.e., LDL and TC) may serve as the potential biomarkers for discriminating recurrent ischemic from first-ever stroke.

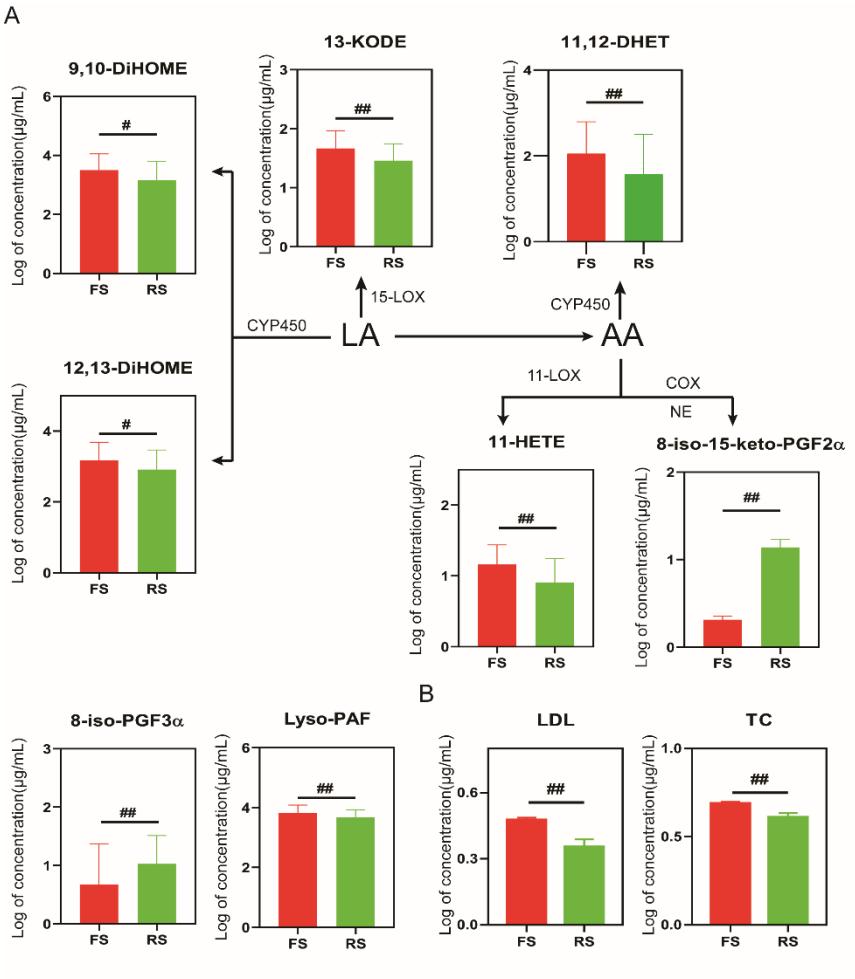


Fig. S3 Comparison of fatty acid metabolites (A) and clinical biochemical parameters (B) levels between first-ever (FS) and recurrent ischemic stroke (RS) (# indicates p value < 0.05, ## p

indicates value < 0.01).

We further used the stepwise logistic regression to generate an optimal classifier for distinguishing between first-ever and recurrent ischemic stroke. As shown in Fig. S4, two models including model D and model E were generated. Model D comprised of two fatty acid metabolites including 11-HETE and 8-iso-15-keto-PGF2α; Model E comprised of two clinical biochemical parameters including TC and LDL. The odd ratios of both models were greater than 0, indicating that these models may act as the efficient classifiers to distinguishing first-ever from recurrent ischemic stroke.

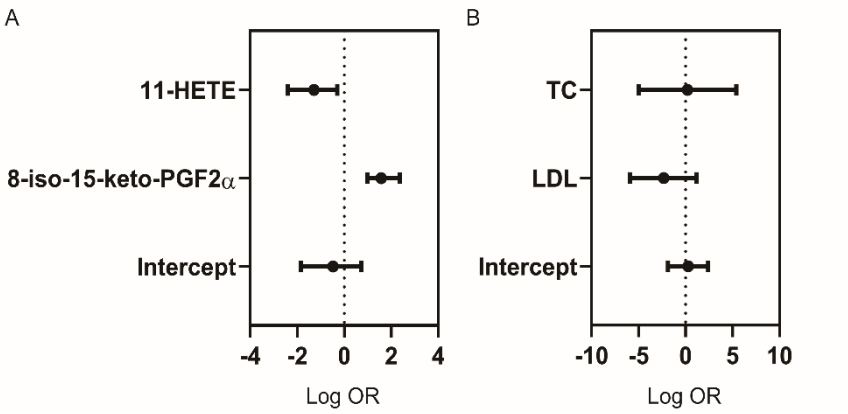
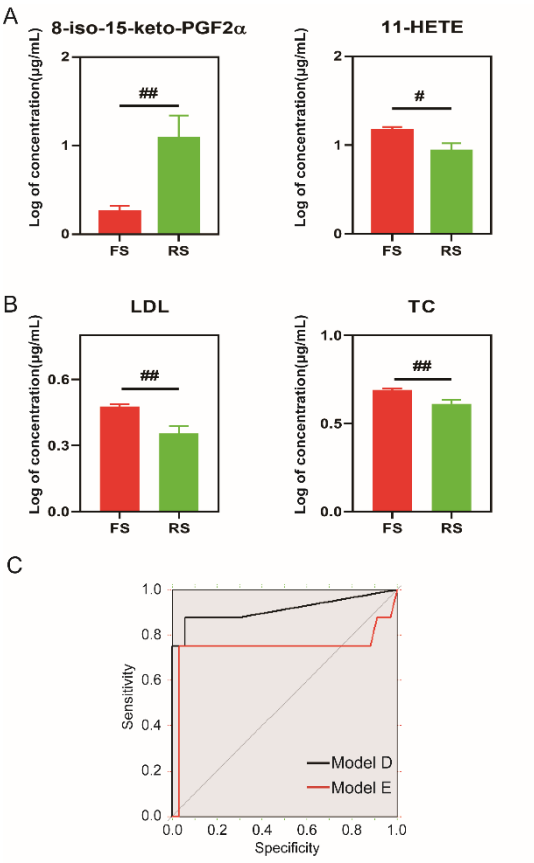


Fig. S4 Predictive models for discriminating first-ever from recurrent ischemic stroke in training set using the stepwise logistic regression. (A) A predictive model (model D) comprising of 11-HETE and 8-iso-15-keto-PGF2α. (B) A predictive model (model E) comprising of TC and LDL. Odd ratio (OR) reflects the relationship between analyte and recurrent ischemic stroke (> 0: positive association; = 0: non-association; < 0: negative association).

We further evaluated the practice diagnosis performance of models D and E in validation set.

As shown in Fig. S5, the associated biomarkers including 8-iso-15-keto-PGF2α, 11-HETE, LDL, and TC involved in the two models exhibited similar changes in the validation set compared with training set, i.e., higher level of 8-iso-15-keto-PGF2α was observed in recurrent ischemic stroke group, and higher levels of 11-HETE, LDL, and TC were observed in first-ever stroke group. The receiver operating characteristic curves of two models were generated. As shown in Fig. S5, models D and E can identify first-ever and recurrent ischemic stroke with 94.12% sensitivity and 87.50% specificity, and 91.18% sensitivity and 75% specificity, respectively, and the AUC value of model D (0.9188) was much greater than that of model E (0.7721). These results suggest that mode D is a more suitable classifier for discriminating first-ever ischemic stroke from recurrent ischemic stroke.



	Accuracy (95%CI)	AUC (95%CI)	Sensitivity (%)	Specificity (%)
Model D	0.9286 (0.8052~0.985)	0.9118 (0.7463~1)	94.12	87.50
Model E	0.881 (0.7437~0.9602)	0.7721 (0.5112~1)	91.18	75.00

Fig. S5 Evaluation of predictive models for discriminating first-ever ischemic stroke from recurrent ischemic stroke in validation set. (A) Comparison of 8-iso-15-keto-PGF2α and 11-HETE levels in validation set (# indicates p value < 0.05; ## indicates p value < 0.01). (B) Comparison of LDL and TC levels in validation set (# indicates p value < 0.05; ## indicates p value < 0.01). (C) Comparison of the performance of different models using the receiver operating characteristic (ROC) curve.