Data-driven clustering approach to identify novel phenotypes using multiple biomarkers in acute ischaemic stroke: A retrospective, multicentre cohort study

Lingling Ding,[a,b,c] Ravikiran Mane,[d] Zhenzhou Wu,[d] Yong Jiang,[a,b,c] Xia Meng,[a,b] Jing Jing,[a,b,c] Weike Ou,[d] Xueyun Wang,[d] Yu Liu,[d] Jinxi Lin,[a,b] Xingquan Zhao,[a,b,c] Hao Li,[a,b] Yongjun Wang,[a,b,c**] and Zixiao Li[a,b,c,e*]

[a]Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China
[b]China National Clinical Research Center for Neurological Diseases, Beijing, China
[c]Research Unit of Artificial Intelligence in Cerebrovascular Disease, Chinese Academy of Medical Sciences, Beijing, China
[d]CNCRC-Hanalytics Artificial Intelligence Research Centre for Neurological Disorders
[e]Chinese Institute for Brain Research, Beijing, China

eClinicalMedicine 2022;53: 101639
Published online xxx
https://doi.org/10.1016/j.eclinm.2022.101639

Summary

Background Acute ischaemic stroke (AIS) is a highly heterogeneous disorder and warrants further investigation to stratify patients with different outcomes and treatment responses. Using a large-scale stroke registry cohort, we applied data-driven approach to identify novel phenotypes based on multiple biomarkers.

Methods In a nationwide, prospective, 201-hospital registry study taking place in China between August 01, 2015 and March 31, 2018, the patients with AIS who were over 18 years of age and admitted to the hospital within 7 days from symptom onset were included. 92 biomarkers were included in the analysis. In the derivation cohort (n=9539), an unsupervised Gaussian mixture model was applied to categorize patients into distinct phenotypes. A classifier was developed using the most important biomarkers and was applied to categorize patients into their corresponding phenotypes in an validation cohort (n=2496). The differences in biological features, clinical outcomes, and treatment response were compared across the phenotypes.

Findings We identified four phenotypes with distinct characteristics in 9288 patients with non-cardioembolic ischaemic stroke. Phenotype 1 was associated with abnormal glucose and lipid metabolism. Phenotype 2 was characterized by inflammation and abnormal renal function. Phenotype 3 had the least laboratory abnormalities and small infarct lesions. Phenotype 4 was characterized by disturbance in homocysteine metabolism. Findings were replicated in the validation cohort. In comparison with phenotype 3, the risk of stroke recurrence (adjusted hazard ratio [aHR] 2.02, 95% confidence intervals [CI] 1.04-3.94), and mortality (aHR 18.14, 95%CI 6.62-49.71) at 3-month post-stroke were highest in phenotype 2, followed by phenotype 4 and phenotype 1, after adjustment for age, gender, smoking, drinking, history of stroke, hypertension, diabetes mellitus, dyslipidemia, and coronary heart disease. The Monte Carlo simulation showed that the patients with phenotype 2 could benefit from high-intensity statin therapy.

Interpretation A data-driven approach could aid in the identification of patients at a higher risk of adverse clinical outcomes following non-cardioembolic ischaemic stroke. These phenotypes, based on different pathophysiology, can suggest individualized treatment plans.

Funding Beijing Natural Science Foundation (grant number Z200016), Beijing Municipal Committee of Science and Technology (grant number Z201100005620010), National Natural Science Foundation of China (grant number 82101360, 92046016, 82171270), Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (grant number 2019-I2M-5-029).

Keywords: Acute ischemic stroke; Machine Learning; Biomarkers; Phenotypes; Clinical outcome

*Corresponding author at: Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, No 119 S 4th Ring W Rd, Fengtai District, Beijing 100070, China.
**Corresponding author at: Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, No. 119 South 4th Ring West Road, Fengtai District, Beijing, 100070, China.
E-mail addresses: yongjunwang@ncrcnd.org.cn (Y. Wang), lizixiao2008@hotmail.com (Z. Li).

Research in context

Evidence before this study

Acute ischaemic stroke (AIS) is a highly heterogeneous disorder with high risk of stroke recurrence, disability, and mortality. We searched PubMed using the terms ("ischaemic stroke" or "cerebrovascular disease"), ("biomarker" or "molecular" or "phenotype" or "subtype" or "subgroup"), and ("machine learning" or "artificial intelligence" "data-driven" or "clustering" or "non-supervised" or "classify") for articles published up to May 1, 2022 and found no study of clustering analysis based on biomarkers in patients with ischaemic stroke. Although several studies have confirmed the associations of biomarkers with pathogenesis and prognosis in patients with ischaemic stroke, most of them just focused on a single biomarker and neglected the interaction effects of multiple biomarkers. Therefore, it's necessary to identify novel phenotypes of AIS using unsupervised clustering analysis and further investigate their relationships with treatment responses and clinical outcomes.

Added value of this study

To the best of our knowledge, this is the first to identify four phenotypes with specific characteristics using 92 biomarkers from a large-scale, multi-centre cohort of patients with AIS. We adopted the Gaussian mixture model and light gradient boosted machine (LightGBM) model to identify the novel phenotypes. We described that the biological features, clinical outcomes, and treatment response varied across phenotypes. We found that phenotype 2, which was characterized by inflammation and abnormal renal function, had the highest risk of stroke recurrence, disability, and mortality, and was associated with a good response to the high-intensity statin therapy. Besides, we revealed that phenotype 1(abnormal glucose and lipid metabolism) and phenotype 4(disturbance in homocysteine metabolism) were also associated with adverse clinical outcomes.

Implications of all the available evidence

This study provides evidence of biological heterogeneity for AIS, that may help gain a deeper insight into the potential pathogenesis in ischaemic stroke. In addition, we provide a new risk stratification approach for supporting clinical decision making.

Introduction

Acute ischaemic stroke (AIS) is a highly heterogeneous disorder that is associated with considerably high morbidity, disability, and mortality.[1,2] Antiplatelet and lipid-lowering drugs are recommended for the prevention of non-cardioembolic ischaemic stroke.[3] Despite strict adherence to current guideline recommendations for the prevention of stroke recurrence, some patients have been observed still to be at a high risk of recurrent stroke.[4,5] This suggests a need for a reassessment of the presumptions regarding the pathophysiology of ischaemic stroke and potential therapeutic targets. Besides, the traditional stroke subtypes based on the Trial of Org 10 172 in Acute Stroke Treatment (TOAST) and Causative Classification of Stroke (CCS) criteria need a comprehensive and systematic evaluation of intracranial or extracranial arteries, as well as cardiac examination, which makes it difficult to intervene early in patients with AIS.[6,7] Therefore, stratification of the heterogeneity among patients based on the ensemble of multiple biomarkers can enhance the understanding of acute ischaemic stroke and enable more personalized treatment planning.

Many recent works have shown that rather than relying on expert clinicians' knowledge, data-driven approaches like unsupervised machine learning, can be used to discover novel phenotypes of patients in various diseases including diabetes,[8] sepsis,[9] dilated cardiomyopathy,[10] pulmonary arterial hypertension,[11] and heart failure,[12] that may help in understanding mechanisms of diseases and treatment effects. Therefore, with the availability of a large amount of biomarker data, a comprehensive, data-driven assessment of the heterogeneity using machine learning methods may provide new opportunities to understand AIS, which previously has not been done.

This study aims to develop and evaluate novel phenotypes of acute non-cardioembolic ischaemic stroke based on 92 biomarkers using a large-scale multi-centre dataset. Through a machine learning-based unsupervised clustering approach, we aim to identify different phenotypes of patients that share similar pathophysiological characteristics, treatment responses and clinical outcomes.

Methods

Study design and population

This study retrospectively analysed the data from the Third China National Stroke Registry (CNSR-III), which is a nationwide, multi-centre, prospective, observational registry study of 15,166 patients with AIS or transient ischaemic attack (TIA) enrolled at 201 hospitals in China between August 01, 2015 and March 31, 2018. The patients participating in the CNSR-III study were over 18 years of age and were admitted to the hospital within 7 days of AIS or TIA onset. Further details about the CNSR-III study design and methodology have been described elsewhere.[13] This study was approved by the Institutional Review Boards (IRB) of Beijing Tiantan Hospital. Written informed consent was obtained from all included patients or their representatives. The data were reported in adherence to the Strengthening the Reporting of

OBservational studies in Epidemiology (STROBE) reporting guidelines.

From the CNSR-III dataset, patients with acute non-cardiac ischaemic stroke were included in this analysis. To reduce the heterogeneity of populations, patients who experienced TIA or a stroke of other determined etiology (OE) were excluded from the analysis. As we did not collect the cardiac-specific biomarkers such as cardiac troponin-T (cTnT), cardiac troponin-I (cTnI) and B-type natriuretic peptide (BNP), we excluded the patients diagnosed with cardioembolic stroke in the analysis. Also, the patients who presented with cancer or infection within 2 weeks before stroke onset were excluded. The baseline characteristics between the included and excluded patients are presented in Supplementary Table 1. A temporal split was applied to the included patients to divide them into the derivation cohort (~75% data, admitted before August 2017) and the validation cohort (~25% data, admitted after August 2017) (Figure 1).

Clinical information about the patients was collected through in-person interviews by trained research coordinators. Stroke severity was assessed within 24 hours of hospital admission using the National Institutes of Health Stroke Scale (NIHSS) score. Stroke etiology was classified into 5 major categories: large artery atherosclerosis (LAA), cardioembolism (CE), small-vessel occlusion (SVO), stroke of other determined etiology (OE) and stroke of undetermined cause (UE), according to the TOAST and CCS criteria.[6,7]

Blood biomarker

The blood samples were collected on the day of the hospital enrollment. All the specimens were stored at -80°C until the testing was performed. The measurement of blood biomarkers was performed at the central laboratory at Tiantan Hospital, Beijing, China by laboratory staff who were blinded to the patients' characteristics and clinical outcomes. A total of 83 blood biomarkers involved in this study, including blood constituents (n=14), coagulation function (n=6), liver function (n=13), renal function (n=5), inflammation (n=12), electrolyte (n=5), lipid metabolism (n=15), homocysteine metabolism (n=4), glucose metabolism (n=2), and gut microbial metabolites (n=7).

Imaging data

Brain magnetic resonance imaging (MRI) and vascular assessment for intracranial arteries and extracranial arteries were collected from 13,012 patients in the Digital Imaging and Communications in Medicine (DICOM) format for the extraction of neuroimaging features. The patients were assessed for the presence of symptomatic intracranial atherosclerotic stenosis

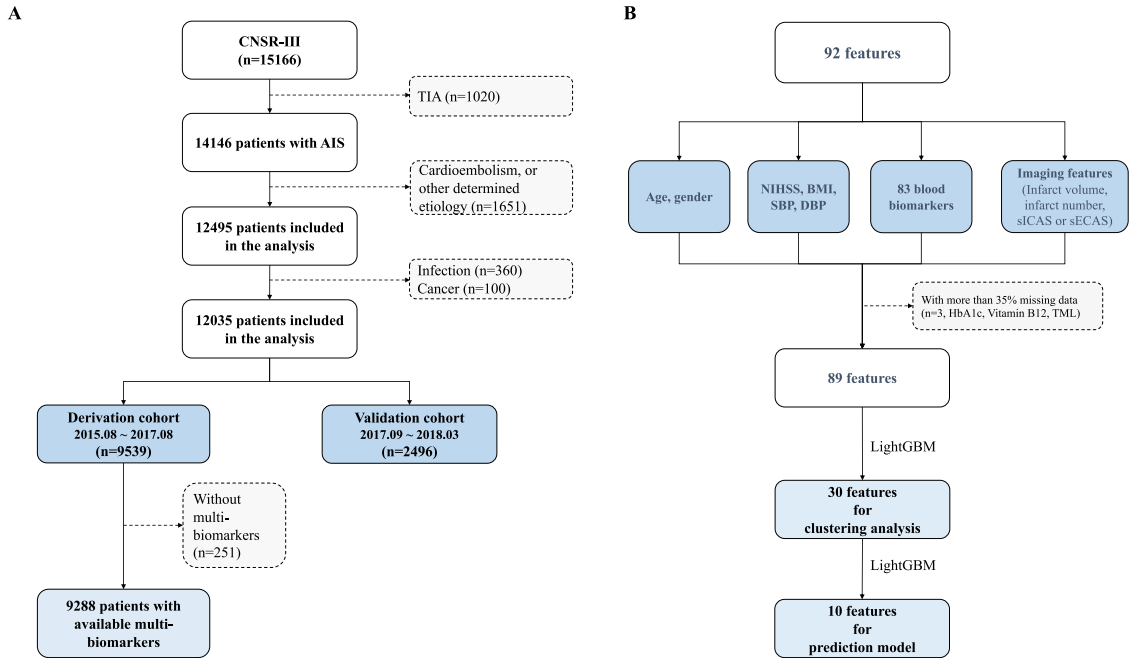

Figure 1. Study flow chart. A. Patient selection. B. Feature selection.

Abbreviations: CNSR-III, Third China National Stroke Registry; TIA, transient ischaemic attack; AIS, acute ischaemic stroke; NIHSS, National Institutes of Health Stroke Scale; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; sICAS, symptomatic intracranial atherosclerotic stenosis; sECAS, symptomatic extracranial atherosclerotic stenosis; HbA1c, Hemoglobin A1c; TML, trimethyllysine; LightGBM, light gradient boosted machine.

(sICAS) and extracranial atherosclerotic stenosis (sECAS). sICAS and sECAS were defined as severe (50%-99%) stenosis or occlusion of clinically relevant intracranial and extracranial arteries, respectively. sICAS judgement was based on the Warfarin-Aspirin Symptomatic Intracranial Disease (WASID) criteria.[14] The North American Symptomatic Carotid Endarterectomy Trial (NASCET) criteria was adopted to adjust the assessment of sECAS.[15,16] The brain tissue damage caused by the acute ischaemic stroke, measured as the volume of the ischaemic lesions, was calculated from the diffusion-weighted image (DWI) and apparent diffusion coefficient (ADC) scans using a deep learning segmentation model.[17]

Clinical outcomes

Recurrent stroke at 3-, 6-, and 12-months post-stroke were the primary clinical outcomes of this study. The onset of a composite vascular event (stroke, myocardial infarction, or vascular death), all-cause mortality at 3-, 6-, and 12-months post-stroke, and poor functional outcome (defined as modified Rankin Scale [mRS] score of 3-6) at 3-months post-stroke were the secondary clinical outcomes. Patients were followed up via in-person interview at 3 months, and via telephonic interview at 6 and 12 months by trained interviewers based on a standardized interview protocol to collect the clinical outcomes.[13] For patients who were enrolled in this study, 141 patients were lost to follow-up, of which 97 individuals were in the derivation cohort and 44 individuals were in the validation cohort.

Data pre-processing

All the included patients in the study were assessed for the presence of clinical features. In this study, we tried to use features that were objective and can be automatically extracted. A total of 92 biomarkers were included in the analysis (Supplementary Table 2). In the derivation cohort, we excluded 251 patients without multiple biomarkers. The features which were missing in more than 35% of the patients were excluded from the clustering analysis (Figure 1). The missing value of the features were shown in Supplementary Figure 2. Then, the missing values were imputed with the mode of the data for categorical features and with the median of the data for numerical features. The patients were divided into six subgroups according to their age (≤60, 61-70 and >70) and gender (male and female), and the missing values were imputed based on the mode and the median of the respective subgroup.

To focus only on the clinically important features and remove irrelevant features from the analysis, a feature selection was performed using the light gradient boosted machine (LightGBM), a gradient boosting decision tree (GBDT) algorithm.[18] Using the data from the derivation cohort, the features were ranked according to their importance in the prediction of stroke recurrence, and the 30 most important features were selected for further clustering analysis (Figure 2A). SHapley Additive exPlanations (SHAP) values were calculated for these 30 features (Figure 2B-C). The selected features were standard normalized to zero mean and unit standard deviation. Supplementary Figure 3 shows a heatmap representing the correlation between 30 biomarkers.

Unsupervised clustering analysis

To identify phenotypes of patients with similar clinical characteristics, an unsupervised clustering analysis on the data from the derivation cohort was performed. To extract more generalizable and robust phenotypes, an unsupervised Gaussian mixture model (GMM) clustering method was used. GMM is a probabilistic model that uses a soft clustering approach to group patients into discrete phenotypes, and it assumes that all data samples X are generated by a mixture of K multivariate Gaussian distributions. Here, each phenotype is modeled as a gaussian multivariate mixture with a mean and covariance that describes the shape of each phenotype.[19] In our analysis, the GMM model was trained using an iterative expectation-maximization algorithm for 1000 epochs. Also, the number of phenotypes that can optimally describe the derivation cohort data was determined using the Calinski Harabasz (CH) Score[20] and Davies Bouldin (DB) Score.[21] Once the phenotypes were determined, patterns of biomarkers were visualized using chord plots[9] and an unsupervised hierarchical clustering heat map.[22]

Simplified supervised patient stratification model

The unique phenotypes identified in the clustering analysis were based on 30 features. To further reduce the dependence on multiple features and to simplify the stratification of the patients, we employed a light gradient boosted machine (LightGBM) model to classify the patients into the identified phenotypes with a reduced number of features. We first identified the 10 most important features of the LightGBM model using the information gain criteria. Next, a LightGBM prediction model using these features was developed with the data in the derivation cohort. The performance of the proposed prediction model in assigning the patients to the correct phenotype was assessed in a 10-fold cross-validation analysis using the area under the receiver operating characteristics curve (AUC). Finally, the prediction model was used to stratify the patients from the validation cohort. The clinical characteristics and outcomes in the sub-groups of the validation cohort were analysed to validate the generalizability of the proposed phenotypes.

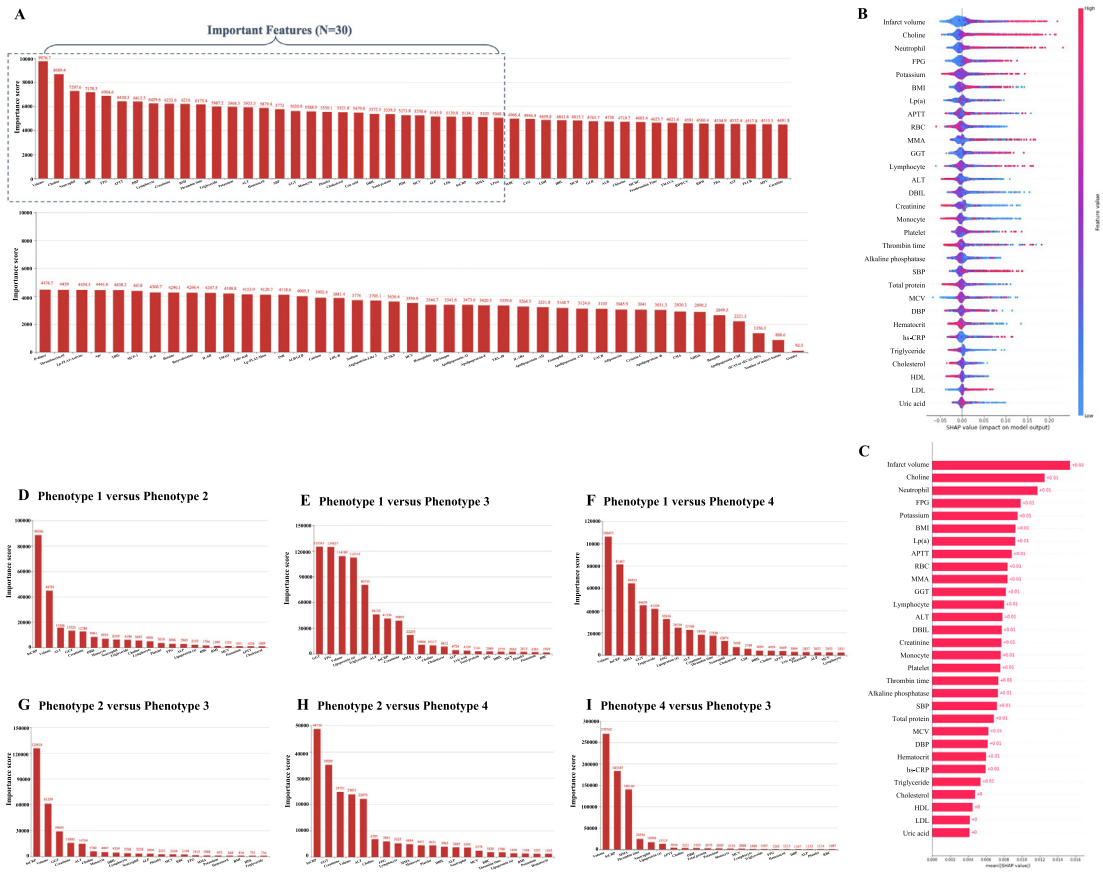

Figure 2. Importance ranking of features. A. Importance ranking of 89 features according to light gradient boosted machine model. B-C. SHapley Additive exPlanations (SHAP) values for 30 features. D-I. Importance of features for phenotypes according to light gradient boosted machine models.

Abbreviations: RBC, red blood cell; FPG, fasting plasma glucose; APTT, activated partial thromboplastin time; DBP, diastolic blood pressure; BMI, body mass index; ALT, alanine aminotransferase; SBP, systolic blood pressure; GGT, γ-Glutamyl transpeptidase; DBIL, Direct bilirubin; TBIL, total bilirubin; HDL, high density lipoprotein; MCV, mean corpuscular volume; ALP, alkaline phosphatase; LDL, low density lipoprotein; hs-CRP, high-sensitivity C-reactive protein; MMA, methylmalonic aciduria; LP(a), Lipoprotein (a); WBC, white blood cell; CO2, Carbon dioxide combining power; LDH, lactate dehydrogenase; IBIL, indirect bilirubin; MCH, mean corpuscular hemoglobin; GLB, globulin; ALB, albumin; MCHC, mean corpuscular hemoglobin concentration; TMAVA, N,N,N-trimethyl-5-aminovaleric acid; RDWCV, coefficient of variation of RBC distribution width; RDW, RBC distribution width; TBA, total bile acid; AST, aspertate aminotransferase; PLCR, Platelet large cell ratio; MPV, mean platelet volume; TBIL, total bilirubin; MCP-1, monocyte chemoattractant protein-1; IL-6, interleukin-6; IL-6R, interleukin-6 receptor; TMAO, trimethylamine-N-oxide; INR, international normalized ratio; LDL-R, low density lipoprotein-receptor; PCSK9, proprotein convertase subtilisin/Kexin type 9; HCY, homocysteinemia; YKL-40, chitinase-3-like protein 1; IL-1Ra, Interleukin-1 receptor antagonist; UACR, urea albumin creatinine ratio; UMA, urine microalbumin; NIHSS, National Institutes of Health Stroke Scale; sICAS, symptomatic intracranial atherosclerotic stenosis; sECAS, symptomatic extracranial atherosclerotic stenosis.

For each phenotype, we also drew radar plots based on 10 key features, using z-values of each feature.[23]

Monte-Carlo simulation for stratified treatment effect

We used Monte-Carlo simulations to explore the heterogeneity of the treatment effects to the frequency distributions of these phenotypes. High-intensity statin treatment can provide more clinical benefits compared with standard statin in patients with high-risk atherosclerotic cardiovascular disease. In this study, we assessed how the benefits of high-intensity statin therapy(atorvastatin 40-80 mg/day, or rosuvastatin 20-40 mg/day)[24] during hospitalization in reducing the probability of a recurrent stroke at 3 months could change with the alteration in the relative distribution of the identified phenotypes. (Supplementary Methods)

Statistical analysis

Continuous variables were expressed as means and standard deviations (SD), ranges, or medians and interquartile ranges (IQR). Categorical variables were expressed as frequencies and percentages. Univariate comparisons were done with the Kruskal-Wallis H test for continuous variables and with the chi-square test for categorical data. Spearman's correlation coefficients were calculated for associations between features and were rearranged with hierarchical clustering. Hazard ratios (HRs) and 95% confidence intervals (CIs) for stroke recurrence, composite vascular events, and all-cause mortality were estimated for every phenotype by the Cox regression model. Covariates known to be predictive of outcomes in ischaemic stroke such as age, gender, smoking, drinking, history of stroke, hypertension, diabetes mellitus, dyslipidemia, and coronary heart disease, were adjusted in the multivariable models. Crude and multivariable-adjusted odds ratios (ORs) and 95% CIs for poor functional outcomes at 3 months were obtained from a logistic regression model. All data were analysed with the SAS version 9.4 software (SAS Institute Inc, Cary, NC) or python 3.7. The level of significance was defined as $p < 0.05$ (2-sided).

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all data and final responsibility to submit for publication.

Results

12035 patients with acute non-cardioembolic ischaemic stroke were included in this study. Among these patients, 9539 patients were assigned to the derivation cohort and the remaining 2496 patients were assigned to the validation cohort. 251 patients in the derivation cohort without multiple biomarkers were excluded. (Figure 1 and Supplementary Figure 1) The details of all the biomarkers, along with the demographic details of all patients are presented in Supplementary Tables 2–4.

Comparison of clinical characteristics among phenotypes

In the clustering analysis, based on the DB score and CH score, 4 phenotypes were observed to be most optimal to represent the derivation cohort data (Supplementary Figure 4). Thus, we identified four phenotypes with distinctive patterns of clinical features (Supplementary Figure 5), and the summary statistics of these phenotypes are presented in Supplementary Tables 2–3. Figure 3 and Supplementary Figure 6 showed patterns of abnormal features that were characteristics of the observed phenotypes. Post-hoc analysis of the phenotypes indicated that phenotype 1, which included 2475 (26.65%) patients in the derivation cohort, was characterized by a low level of adiponectin, and abnormal lipid metabolism, with an increased level of low-density lipoprotein (LDL), triglycerides, lipoprotein (a) (Lp [a]), and impaired fasting plasma glucose (FPG). Phenotype 2, including 507 (5.46%) patients, was characterized by circulating inflammation, manifested as an increased level of neutrophil, high-sensitivity C-reactive protein (hs-CRP), interleukin-6 (IL-6), chitinase-3-like Protein 1 (YKL40), and interleukin-1 receptor antagonist (IL-1RA); abnormality of renal function, with an increased level of creatinine and cystatin C, urine microalbumin (UMA), and urea albumin creatinine ratio (UACR); and increased level of proprotein convertase subtilisin/Kexin type 9 (PCSK9) and angiopoietin-Like 3 (ANGPTL3). The 4392 (47.29%) patients in phenotype 3 were associated with minimum abnormalities in biomarkers of liver and renal function indexes, inflammation, and glucose metabolism. Phenotype 3 had the highest HDL level and smaller infarct volume than patients in other phenotypes. Also, the incidence of sICAS or sECAS (16.6%) was the lowest in phenotype 3. Phenotype 4, including 1914 (20.61%) patients, was characterized by disturbance in homocysteine metabolism, with a high level of homocysteine (HCY), methylmalonic acid (MMA), and low levels of vitamin B12. The incidence of sICAS or sECAS (45.5%) was the highest in phenotype 4. Medical treatments and adherence did not differ substantially in the four phenotypes (Supplementary Figure 7). However, the patients in phenotype 2 and phenotype 4 were more likely to receive reperfusion therapy than others (Supplementary Table 3).

We analysed the relationship between the newly identified phenotypes and the traditional stroke subtypes based on the TOAST and CCS criteria. The comparison of the novel phenotypes and CCS classification showed that phenotype 2 and phenotype 4 were marked by large artery atherosclerosis (46.4% and 52.1%, respectively), and phenotype 3 was marked by small artery occlusion (37.4%). The results indicated that the observed phenotypes were significantly different from the traditional ways of stroke stratification (Figure 4, Supplementary Figure 8).

Supervised prediction model

To further simply the characterization of the identified phenotypes, the features in the clustering model were evaluated for their importance in clustering decisions, and these feature importance scores are presented in Figure 2D-I. Here, infarct volume, alanine aminotransferase (ALT), hs-CRP, γ-Glutamyl transpeptidase

(GGT), neutrophil counts, FPG, creatinine, triglyceride, methylmalonic aciduria (MMA), and Lp(a) were observed to be the 10 most important features. Using these 10 model-derived, routinely collected, important biomarkers, a prediction model that can classify patients into one of the four phenotypes was developed. In the 10-cross validation analysis on the development dataset, the supervised prediction model achieved a 4-class micro-average AUC of 0.983 (95% CI 0.980-0.986) and a macro-average AUC of 0.974 (95% CI 0.969-0.979) (Individual phenotype AUC: Phenotype 1: AUC 0.975, 95% CI 0.971- 0.978; Phenotype 2: AUC 0.954, 95% CI 0.944-0.963; Phenotype 3: AUC 0.986, 95% CI 0.983-0.989; Phenotype 4: AUC 0.976, 95% CI 0.969-0.982) (Supplementary Figure 9). Using the same prediction model, the patients from the validation cohort were assigned to one of the four phenotypes. The phenotypes in the validation cohort were observed to have similar clinical characteristics as that of the derivation cohort (Figure 3 and Supplementary Table 4). Radar plots represent profiles of the four phenotypes based on 10 key features (Supplementary Figure 10).

Association of phenotypes with clinical outcomes

The clinical outcomes in all the identified phenotypes were analysed and the results of this analysis are presented in Figure 5, Table 1, and Supplementary Table 5. In the derivation cohort, phenotype 3 was observed to have the best clinical outcomes with the lowest stroke recurrence rate (5.16%), combined vascular events (5.23%), and all-cause mortality (0.38%) at 3-month follow-up. At 3-month follow-up, compared to phenotype 3, patients in phenotype 2 experienced significantly worse outcomes in terms of stroke recurrence (adjusted HR 1.89, 95% CI 1.38-2.57, p<0.0001), combined vascular events (adjusted HR 1.98, 95% CI 1.46-2.68, p<0.0001), and all-cause mortality (adjusted HR 12.92, 95% CI 6.95-24.02, p<0.0001). Also, the adjusted risk of poor functional outcome was 3 times higher in phenotype 2 compared to phenotype 3 (adjusted OR 3.61, 95% CI 2.96-4.39, p<0.0001). The participants in phenotype 4 (vs. phenotype 3) were observed to have a significantly higher risk of all adverse clinical events including stroke recurrence (adjusted HR 1.77, 95% CI 1.45-2.16, p<0.0001), combined vascular events (adjusted HR 1.79, 95% CI 1.47-2.18, p<0.0001), all-cause mortality (adjusted HR 4.18, 95% CI 2.32-7.55, p<0.0001), and poor functional outcome (adjusted OR 2.31, 95% CI 2.04-2.61, p<0.0001) at 3-month follow-up.

A similar pattern was repeated in the validation cohort, the patients in phenotype 3 were observed to have the best clinical outcomes. Whereas, phenotype 2 had the highest risk of stroke recurrence (adjusted HR 2.02, 95% CI 1.04-3.94, p=0.038), all-cause mortality (adjusted HR 18.14, 95% CI 6.62-49.71, p<0.001), and poor functional outcome (adjusted OR 5.62, 95% CI 3.67-8.60, p<0.0001) at 3-month follow-up compared to phenotype 3. Phenotype 1 (adjusted HR 3.44, 95% CI 1.17-10.09, p=0.024) and phenotype 4 (adjusted HR 4.65, 95% CI 1.81-11.93, p=0.0014) were associated with a significantly higher risk of all-cause mortality at 3-month follow-up (Figure 5, Table 1 and Supplementary Table 5).

At one-year follow-up, patients in phenotype 2 had the highest risk of combined vascular events (adjusted HR 1.79, 95% CI 1.02-3.14 p=0.041), and all-cause mortality (adjusted HR 8.94, 95% CI 4.76-16.77, p<0.0001). Patients in phenotype 4 had a higher risk of stroke recurrence (adjusted HR 1.56, 95% CI 1.13-2.16, p=0.0072), combined vascular events (adjusted HR 1.60, 95% CI 1.17-2.21, p=0.0038), and all-cause mortality (adjusted HR 2.16, 95% CI 1.24-3.74, p=0.0062) (Figure 5, Table 1 and Supplementary Table 5).

Differential estimated therapy effects by phenotypes distributions

A Monte-Carlo simulation was performed to analyse the effect of high-intensity statin therapy by varying the proportion of phenotypes and the results of this analysis are presented in Figure 6. In the baseline phenotype distribution, the use of high-intensity statin therapy had a 0.01% chance of a benefit, a 76.69% chance of producing no significant effect, and a 23.30% chance of harm for stroke recurrence at 3 months. The chance of finding benefit increased to 6.10% when phenotype 2 represented the majority of the population, and the risk of high-intensity statin therapy being harmful reduced to 0.22%. A similar pattern was observed in the validation cohort. In the validation cohort, with the baseline phenotypes distribution, the high-intensity statin therapy had a 0.35% chance of a benefit, a 96.92% chance of producing no significant effect, and a 2.73% chance of harm for stroke recurrence at 3 months. With phenotype 2 representing the majority of the population, the chance of finding benefit increased to 87.51%, and the chance of a harmful effect reduced to 0.00%. The results of the Monte-Carlo simulation showed that changing the proportion of phenotype 1, phenotype 3, or phenotype 4 did not significantly benefit from high-intensity statin therapy. (Figure 6, Supplementary Figure 11).

Discussion

In this multi-centre study analysing 12035 patients with acute non-cardioembolic ischaemic stroke, we proposed a novel stratification of patients into four biomarker-based phenotypes with unique clinical characteristics, possibly unique disease pathophysiology, and significantly different clinical outcomes. The proposed stratification of patients may provide information about

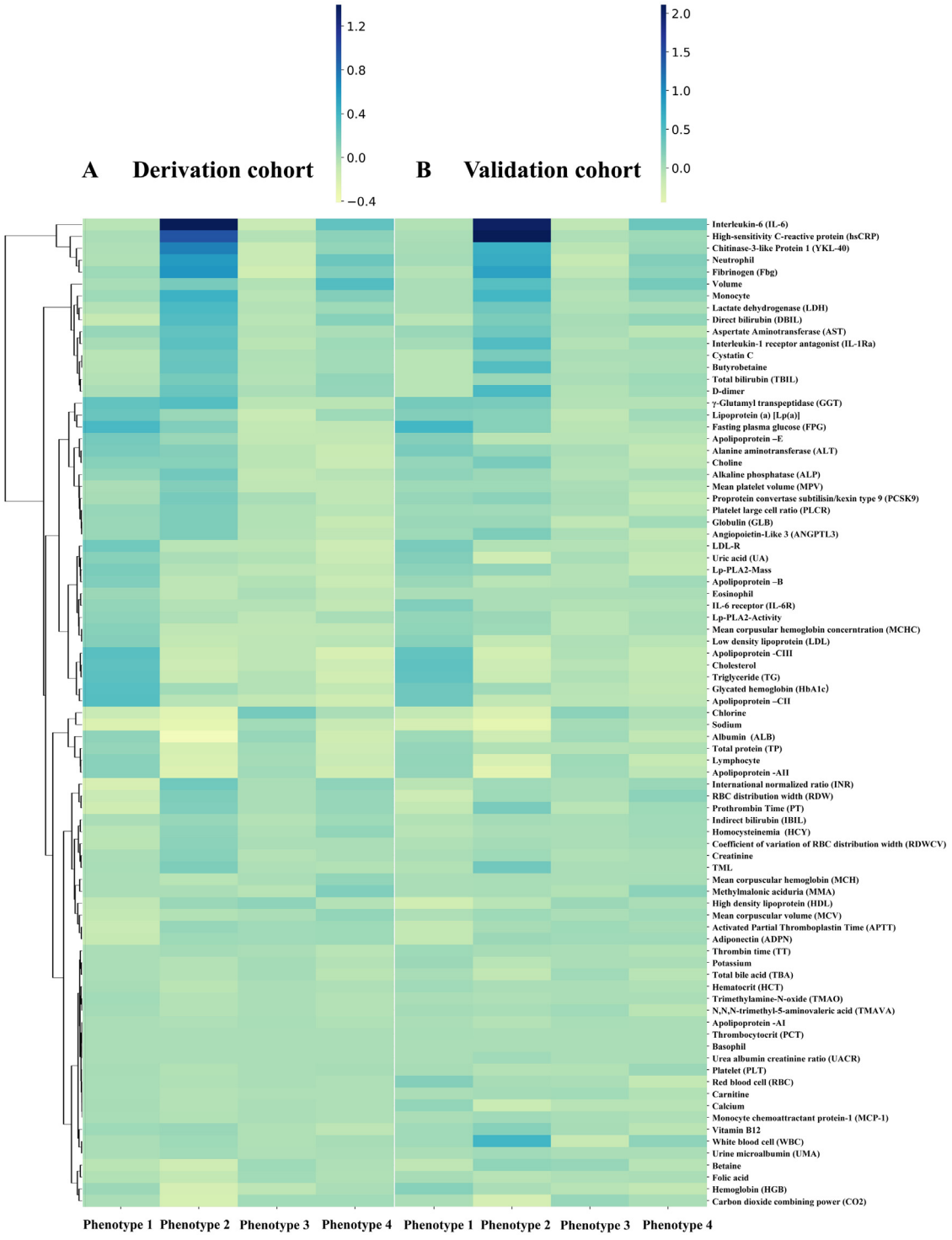

Figure 3. Dendrogram and heat map for unsupervised hierarchical clustering. Dendrogram and heat map for unsupervised hierarchical clustering in 4 phenotypes based on all the biomarkers in the derivation cohort (A) and validation cohort (B).

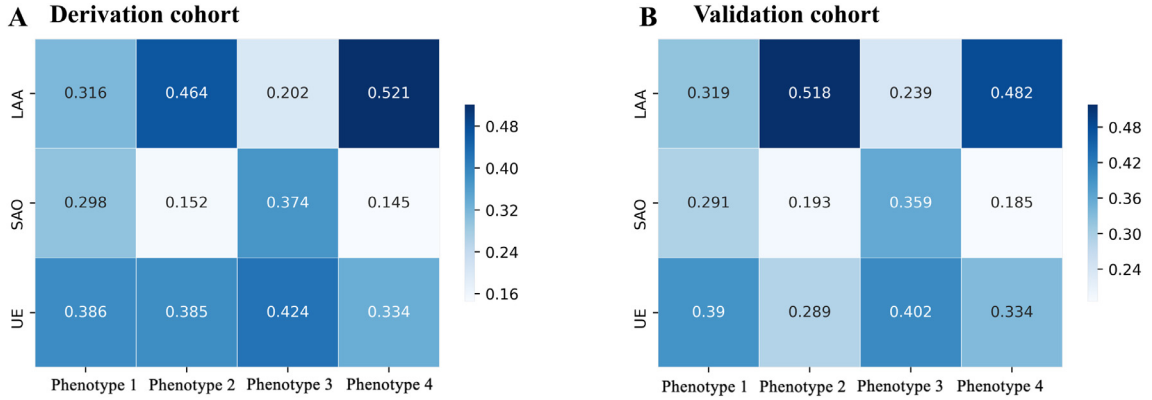

Figure 4. Comparison with traditional stroke subtypes. A. Comparison with CCS classification in the derivation cohort. B. Comparison with CCS in the validation cohort.

Abbreviations: CCS, causative classification of stroke; LAA, large artery atherosclerosis; UE, undetermined etiology; SAO, small artery occlusion.

underlying disease mechanisms, and aid in guiding the choice of post-stroke therapy. To the best of our knowledge, this is the first study that provides a novel stratification of acute non-cardioembolic ischaemic stroke patients based on 92 biomarkers, including blood constituents, coagulation function, liver function, renal function, inflammation, lipid metabolism, homocysteine metabolism, glucose metabolism, gut microbial metabolites, and neuro-imaging features, and it is the first study that applies machine learning techniques to resolve heterogeneity in AIS using dense phenotypic data.

In this study, we have derived phenotypes to facilitate the early detection of patients with a high risk of unfavorable clinical outcomes. These defined phenotypes can be identified at the time of hospital admission, and thus could aid in early treatment planning and enrollment of patients in experimental clinical trials. Furthermore, with the use of feature importance analysis and predictive modeling, we showed that the patients can be uniquely assigned to the identified phenotypes using only ten biomarkers which are routinely acquired even in resource-limited settings. This ensures that the proposed method can be made available in remote healthcare centres.

Phenotype 1 was most strongly characterized by abnormal values of glucose and lipid metabolism as well as clinical features associated with liver dysfunction. The results showed that the patients in phenotype 1 had a low level of adiponectin. Adiponectin is being recognized as a protective adipokine in insulin resistance and liver diseases. Previous studies indicate that the decreased levels of adiponectin might play a key role in the development of atherosclerosis and cardiovascular diseases.[25,26] Changes in gut microbiota-related metabolites represented by increased levels of TMAO and its precursors, choline, were observed in phenotype 1. Alterations in the gut microbiota composition are known to drive activation of lipopolysaccharide, which might result in hepatic steatosis, adipose tissue macrophages infiltration, dyslipidemia, hyperglycemia, hyperinsulinemia, and obesity.[27,28] In particular, TMAO has been shown to directly influence the propensity of macrophages to accumulate cholesterol and form foam cells in atherosclerotic lesions, as well as to alter cholesterol and sterol metabolism within multiple compartments including the liver and intestines. [29,30] The evidence from this study may provide opportunities for the development of new diagnostic tests and therapeutic approaches for the individuals that are classified as phenotype 1.

The participants in phenotype 2 were observed to be at the highest risk of recurrent stroke, combined vascular events, poor functional outcomes, and all-cause mortality. This phenotype was primarily characterized by elevated levels of inflammation and a high incidence of sICAS/sECAS. The serum level of ANGPTL3, lp(a), and PCSK9 were also observed to be increased in phenotype 2. Atherosclerosis is a chronic inflammatory process of the vascular wall that is initiated by excessive LDL-C and is mediated by activated macrophages. Hyperlipidemia elicits a profound enrichment of a pro-inflammatory subset of monocytes. These pro-inflammatory monocytes, home to atherosclerotic lesions, give rise to macrophages, which in the arterial intima form foam cells, and stimulate the innate immune response by expressing high levels of pro-inflammatory cytokines.[31,32] Progress in understanding the basic biology of inflammation in atherosclerosis will help to identify potential novel strategies for modulating inflammation in stroke prevention. Phenotype 2 was also characterized by an abnormal renal function index. Inflammation is highly prevalent in patients with chronic kidney disease (CKD) and is consistently associated with cardiovascular events

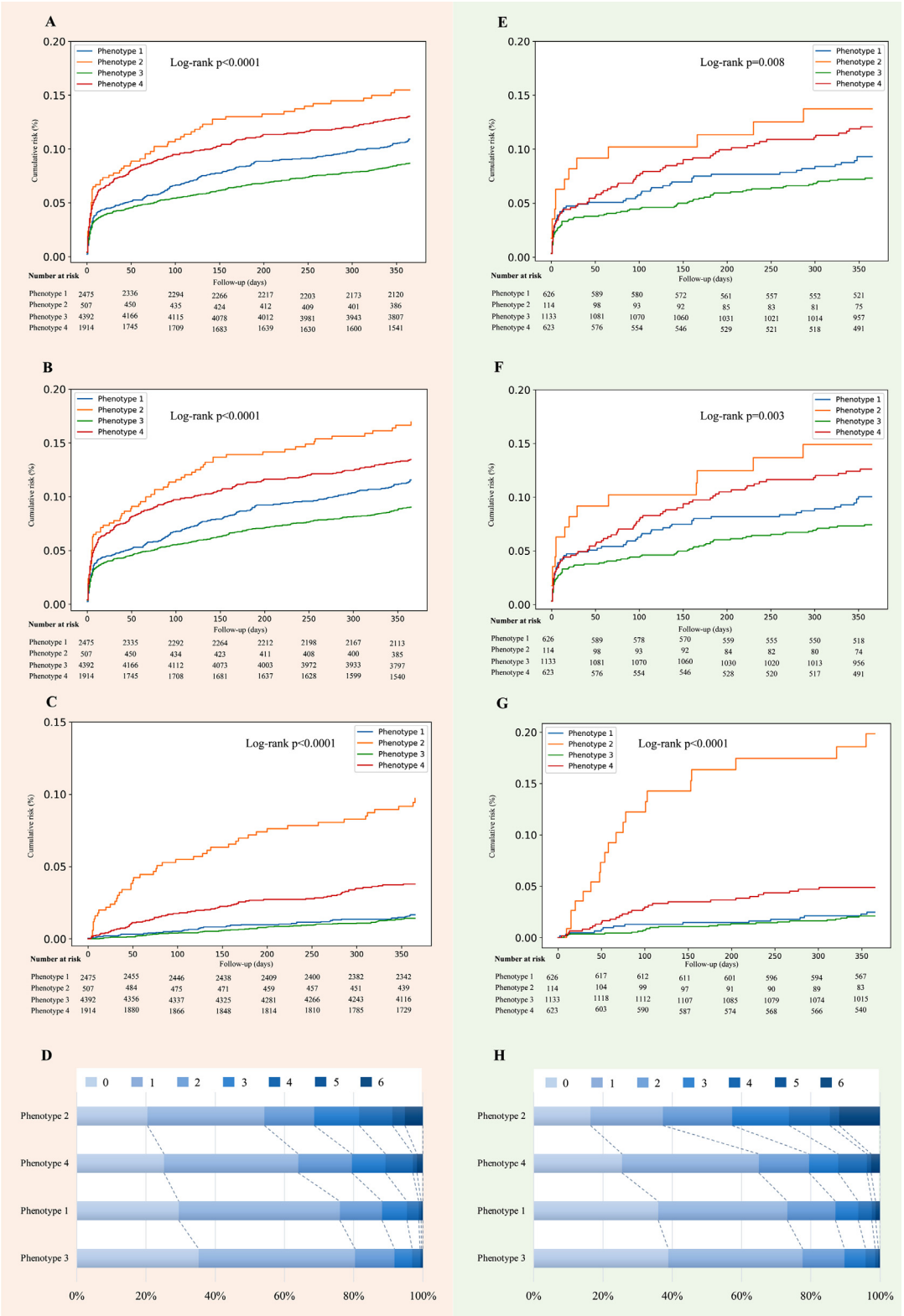

Figure 5. Clinical outcomes stratified by the identified phenotypes. Kaplan-Meier curves of time to stroke recurrence (A), combined vascular events (B), and all-cause mortality (C) within one year after stroke in derivation cohort. D. The distribution of the modified Rankin Scale (mRS) score 90 days after stroke in derivation cohort. Kaplan-Meier curves of time to stroke recurrence (E), combined vascular events (F), and all-cause mortality (G) within one year after stroke in validation cohort. H. The distribution of the mRS score 90 days after stroke in validation cohort.

		Phenotype	Derivation cohort						Validation cohort					
			Total	Events, n (%)	HR (95% CI)	*P* value	Adjust HR (95% CI)	*P* value	Total	Events, n (%)	HR (95% CI)	*P* value	Adjust HR (95% CI)	*P* value
Stroke recurrence	3 months	Phenotype 1	2475	148 (5.97%)	1.16 (0.94-1.43)	0.161	1.10 (0.89-1.36)	0.385	626	33 (5.27%)	1.23 (0.79-1.91)	0.365	1.22 (0.78-1.19)	0.385
		Phenotype 2	507	49 (9.6%)	1.93 (1.41-2.62)	<0.0001	1.89 (1.38-2.57)	<0.0001	114	11 (9.64%)	2.32 (1.21-4.45)	0.011	2.02 (1.04-3.94)	0.038
		Phenotype 3	4392	227 (5.16%)	-	-	-	-	1133	49 (4.32%)	-	-	-	-
		Phenotype 4	1914	169 (8.82%)	1.74 (1.43-2.12)	<0.0001	1.77 (1.45-2.16)	<0.0001	623	41 (6.58%)	1.54 (1.02-2.33)	0.041	1.47 (0.97-2.23)	0.069
	6 months	Phenotype 1	2475	199 (8.04%)	1.26 (1.05-1.51)	0.013	1.21 (1.01-1.46)	0.041	626	45 (7.18%)	1.35 (0.92-1.98)	0.128	1.37 (0.92-2.04)	0.116
		Phenotype 2	507	61 (12.03%)	1.95 (1.48-2.58)	<0.0001	1.92 (1.46-2.54)	<0.0001	114	12 (10.52%)	2.08 (1.12-3.86)	0.020	1.77 (0.94-3.33)	0.075
		Phenotype 3	4392	282 (6.42%)	-	-	-	-	1133	61 (5.38%)	-	-	-	-
		Phenotype 4	1914	199 (10.39%)	1.66 (1.38-1.99)	<0.0001	1.67 (1.39-2.00)	<0.0001	623	54 (8.66%)	1.65 (1.14-2.38)	0.0075	1.57 (1.08-2.26)	0.016
	12 months	Phenotype 1	2475	253 (10.22%)	1.25 (1.07-1.47)	0.0059	1.227 (1.04-1.45)	0.014	626	55 (8.78%)	1.27 (0.90-1.80)	0.168	1.31 (0.92-1.86)	0.137
		Phenotype 2	507	71 (14.00%)	1.81 (1.40-2.33)	<0.0001	1.77 (1.37-2.28)	<0.0001	114	14 (12.28%)	1.93 (1.10-3.42)	0.022	1.72 (0.96-3.06)	0.066
		Phenotype 3	4392	361 (8.21%)	-	-	-	-	1133	79 (6.97%)	-	-	-	-
		Phenotype 4	1914	231 (12.06%)	1.52 (1.29-1.79)	<0.0001	1.52 (1.29-1.80)	<0.0001	623	69 (11.07%)	1.64 (1.19 -2.27)	0.0026	1.56 (1.13-2.16)	0.0072
Combined vascular events	3 months	Phenotype 1	2475	151 (6.10%)	1.17 (0.95-1.43)	0.138	1.10 (0.89-1.36)	0.362	626	36 (5.75%)	1.34 (0.87-2.06)	0.184	1.36 (0.87-2.11)	0.179
		Phenotype 2	507	52 (10.25%)	2.02 (1.49-2.73)	<0.0001	1.98 (1.46-2.68)	<0.0001	114	11 (9.64%)	2.32 (1.21-4.45)	0.011	2.02 (1.04-3.94)	0.038
		Phenotype 3	4392	230 (5.23%)	-	-	-	-	1133	49 (4.32%)	-	-	-	-
		Phenotype 4	1914	173 (9.03%)	1.76 (1.44-2.14)	<0.0001	1.79 (1.47-2.18)	<0.0001	623	42 (6.74%)	1.58 (1.05-2.39)	0.029	1.51 (1.00- 2.28)	0.051
	6 months	Phenotype 1	2475	206 (8.32%)	1.25 (1.05-1.50)	0.012	1.21 (1.01-1.45)	0.043	626	48 (7.66%)	1.42 (0.97-2.06)	0.070	1.46 (0.99-2.15)	0.056
		Phenotype 2	507	65 (12.82%)	2.01 (1.54-2.63)	<0.0001	1.98 (1.51-2.60)	<0.0001	114	13 (11.40%)	2.23 (1.22-4.05)	0.0088	1.86 (1.01 -3.42)	0.046
		Phenotype 3	4392	293 (6.67%)	-	-	-	-	1133	62 (5.47%)	-	-	-	-
		Phenotype 4	1914	204 (10.65%)	1.64 (1.37-1.96)	<0.0001	1.65 (1.38-1.98)	<0.0001	623	57 (9.14%)	1.71 (1.20-2.46)	0.0033	1.62 (1.13-2.33)	0.0085
	12 months	Phenotype 1	2475	267 (10.78%)	1.27 (1.09 -1.49)	0.0025	1.24 (1.06-1.46)	0.0081	626	59 (9.42%)	1.35 (0.97-1.89)	0.079	1.40 (0.99-1.99)	0.054
		Phenotype 2	507	77 (15.018%)	1.89 (1.48 -2.42)	<0.0001	1.85 (1.45-2.37)	<0.0001	114	15 (13.15%)	2.06 (1.18-3.57)	0.010	1.79 (1.02-3.14)	0.041
		Phenotype 3	4392	375 (8.53%)	-	-	-	-	1133	80 (7.06%)	-	-	-	-
		Phenotype 4	1914	238 (12.43%)	1.51 (1.28 -1.77)	<0.0001	1.51 (1.28-1.78)	<0.0001	623	72 (11.55%)	1.69 (1.23-2.33)	0.0011	1.60 (1.17-2.21)	0.0038
Mortality	3 months	Phenotype 1	2475	12 (0.48%)	1.25 (0.60-2.62)	0.551	1.26 (0.59 -2.70)	0.548	626	8 (1.27%)	2.41 (0.84-6.95)	0.102	3.44 (1.17-10.09)	0.024
		Phenotype 2	507	26 (5.12%)	13.63 (7.39-25.11)	<0.0001	12.92 (6.95-24.02)	<0.0001	114	13 (11.40%)	22.64 (8.61-59.59)	<0.0001	18.14 (6.62-49.71)	<0.0001
		Phenotype 3	4392	17 (0.38%)	-	-	-	-	1133	6 (0.52%)	-	-	-	-
		Phenotype 4	1914	32 (1.67%)	4.35 (2.42-7.83)	<0.0001	4.18 (2.32-7.55)	<0.0001	623	16 (2.56%)	4.92 (1.92-12.57)	<0.0001	4.65 (1.81-11.93)	0.0014
	6 months	Phenotype 1	2475	24 (0.96%)	1.37 (0.81-2.34)	0.243	1.49 (0.86 -2.57)	0.156	626	9 (1.43%)	1.25 (0.54-2.93)	0.602	1.60 (0.67-3.81)	0.285
		Phenotype 2	507	35 (6.90%)	10.16 (6.27-16.48)	<0.0001	9.69 (5.93-15.84)	<0.0001	114	17 (14.91%)	14.11 (6.85-29.06)	<0.0001	12.33 (5.73-26.51)	<0.0001
		Phenotype 3	4392	31 (0.71%)	-	-	-	-	1133	13 (1.14%)	-	-	-	-
		Phenotype 4	1914	50 (2.61%)	3.75 (2.39-5.86)	<0.0001	3.60 (2.30-5.64)	<0.0001	623	22 (3.53%)	3.15 (1.58-6.24)	0.0011	2.96 (1.48-5.90)	0.0021

Table 1 (Continued)

	Phenotype	Derivation cohort					Validation cohort					
		Total	Events, n (%)	HR (95% CI)	*P* value	Adjust HR (95% CI)	*P* value	Total	Events, n (%)	HR (95% CI)	*P* value	Adjust HR (95% CI)	*P* value
12 months	Phenotype 1	2475	40 (1.61%)	1.16 (0.78-1.73)	0.458	1.31 (0.87-1.97)	0.199	626	15 (2.39%)	1.18 (0.62-2.26)	0.619	1.38 (0.71-2.68)	0.348
	Phenotype 2	507	46 (9.07%)	6.90 (4.70-10.11)	<0.0001	6.38 (4.32-9.41)	<0.0001	114	20 (17.54%)	9.72 (5.33-17.70)	<0.0001	8.94 (4.76-16.77)	<0.0001
	Phenotype 3	4392	61 (1.38%)	-	-	-	-	1133	23 (2.03%)	-	-	-	-
	Phenotype 4	1914	70 (3.65%)	2.68 (1.90-3.78)	<0.0001	2.53 (1.80-3.58)	<0.0001	623	29 (4.65%)	2.36 (1.36-4.08)	0.0021	2.16 (1.24-3.74)	0.0062

Table 1: **Clinical outcomes in the derivation cohort and validation cohort by phenotypes.**
Adjust for age, gender, smoking, drinking, history of stroke, hypertension, diabetes mellitus, dyslipidemia, and coronary heart disease.
Abbreviations: HR, Hazard ratios; CI, confidence intervals.

and mortality, supporting the role of kidney dysfunction in the systemic process.[33,34]

Phenotype 2 had the highest level of inflammatory markers, which may explain the correlation between the new classification and outcomes. In recent years, inflammation has been increasingly recognized as an important contributor to the fate of the ischaemic brain and the survival of people after ischaemic stroke.[35] The concentrations of various inflammatory markers like neutrophils, high-sensitive C-reactive protein (hs-CRP), and interleukin-6 (IL-6) could reflect a systemic stress response to injury, which have been associated with a high risk of cerebrovascular events.[36–39] Therefore, anti-inflammatory therapy has been proposed as a potential treatment for preventing stroke recurrence and other vascular events after ischaemic stroke or TIA.[40–42] The Colchicine Cardiovascular Outcomes Trial (COLCOT) trial has shown that anti-inflammatory therapy with colchicine can reduce the occurrence of vascular events.[43] Also, recent evidence indicates that statins, in addition to their lipid-lowering properties, can have anti-inflammatory and immunomodulatory effects and these additional effects may play a vital role in the prevention of vascular events.[44] In this study, Monte Carlo simulation revealed that patients in phenotype 2 could benefit from high-intensity statin therapy. The results of An Intervention Trial Evaluating Rosuvastatin (JUPITER) trial showed that rosuvastatin (20 mg daily) effectively reduced the incidence of major cardiovascular events as compared with placebo (P <0.001) in 17,802 healthy individuals without hyperlipidemia, but with high hs-CRP levels of >2 mg/L, and the level of hs-CRP and LDL concentrations were reduced by 37% and 50%, respectively.[45] Furthermore, in vitro and in vivo experiments have shown that statins can modulate the NLRP3 inflammasome and pro-inflammatory cytokine release such as IL-6.[46,47] The concept of statin pleiotropy has provided a window of opportunity to test and target other nonlipid-lowering signaling pathways that may affect cardiovascular disease.[48,49] Future prospective intervention studies are needed to explore the therapeutic effect of interventions targeting inflammation in patients with phenotype 2.

In the present analysis, out of the 4 identified phenotypes, the patients in phenotype 3 presented with the least amount of laboratory abnormalities. Also, patients in phenotype 3 had significantly smaller infarct lesions and small artery occlusion was the prominent cause of stroke in this phenotype. Consequently, the patients in this group were observed to be at the lowest risk of recurrent stroke and had relatively better clinical outcomes.

The risks of all adverse clinical events were observed to be significantly higher in phenotype 4. Phenotype 4 was characterized by a low level of vitamin B12 and a high level of MMA. Vitamin B12 has shown efficacy as a nitric oxide scavenger. Accumulating pieces of evidence

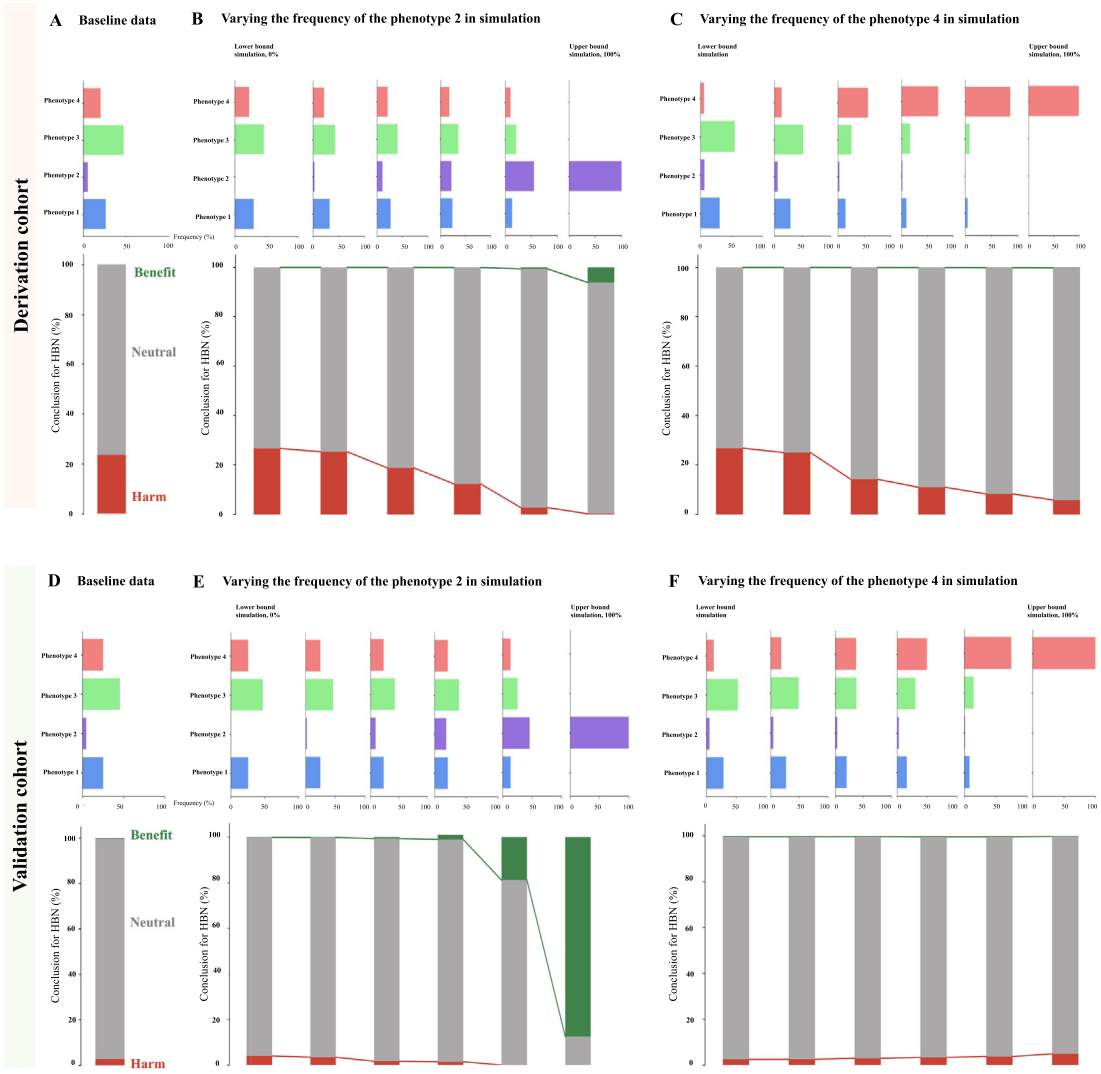

Figure 6. Monte-Carlo simulation of response to high-intensity statin therapy with a different relative frequency of phenotypes. In the derivation cohort, the actual distribution of the data in the given phenotypes and the associated results of the harm, benefit, and neutral effect analysis using Monte-Carlo simulation analysis are presented in panel A. Each simulation was conducted with 100000 iterations using sampling with replacement. The results of the same analysis by changing the phenotype distributions of the data are presented in panels B and C. In panel B, the distribution of phenotype 2 was gradually increased whereas panel C presents the results associated with the gradually increasing the distribution of phenotype 4. Panels D to F present similar results for the validation cohort data.

Abbreviations: HBN: harm, benefit, or neutral.

suggest that Cbl-associated metabolites, MMA and HCY, may promote atherogenesis through its toxic effects on the vascular endothelium, which is likely mediated through oxidative stress.[50] Besides, MMA accumulation reflects the decreased activity of a mitochondrial Cbl-dependent enzyme, which is more sensitive to oxidative damage.[51 52] Further studies are needed to establish specific therapy for the patients in phenotype 4, and vitamin supplements or antioxidant therapy may prove to be beneficial to these patients.

The results presented in this study have significant implications for understanding the mechanisms of AIS. First, the phenotypes presented in this study can be used to prospectively stratify patients in future clinical research, paving the way to a personalized precision medicine approach in the management of AIS. Second, our findings in this study advance the understanding of circulating biomarker profiles in AIS and suggest that multi-biomarker approaches can be implemented for achieving better risk stratification. More importantly,

owing to the multi-biomarker approach in this study, we were able to shed light on more complex pathophysiological pathways associated with AIS, which couldn't have been possible with single-biomarker analysis methods. Lastly, in this analysis, the phenotypes were derived from a large observational cohort and their generalizability was ensured using a large validation cohort.

Some potential limitations of the study should be noted. First, despite all the efforts, the imputation of missing values may affect the results of the study. To reduce missing data, we derived a machine learning model based on 10 key biomarkers, which are basic variables from clinical practice. Second, although the identified phenotypes were found to be generalizable in the validation cohort, further research is needed to determine the utility of these novel phenotypes to optimize clinical care and trial design. Third, the output of machine learning is limited by the limitations of input. Future studies with large biological data that can enable an integrative analysis of multi-omics data (e.g., genomics, transcriptomics, metabolomics) should be conducted to uncover the complex molecular pathways leading to AIS. Fourth, imputing mode or median by subgroup may incur bias. However, considering ischemic stroke is a disease highly correlated with age and gender, we divided the patients into six subgroups according to their age and gender. Other imputation methods such as multiple imputation could also be used to generate accurate estimations of missing values. Fifth, we employed telephonic interviews to collect information about cardiovascular events at 6 and 12 months after AIS, which may potentially influence the appraisal of the clinical outcomes. However, previous studies have indicated the telephonic assessment of recurrent ischaemic strokes to be reliable and creditable.[53,54] Furthermore, the present study was based on participants from China, which may potentially limit the interethnic extrapolation of the findings. Further studies using the independent cohort and other ethnic cohorts are needed to generalize the study's findings.

In conclusion, using data from a nationwide cohort and machine learning methods, we identified four biomarker-based phenotypes that were correlated with specific pathophysiology and clinical outcomes in patients with acute non-cardioembolic ischaemic stroke. With a data-driven approach, this study presents a step towards a more clinically useful stratification of patients, which can play an important role in precision medicine and clinical decision-making in AIS.

Contributors

YW, ZL, and LD designed and implemented the study. RM, ZW, WO, XW, and YL performed the statistical analyses. LD wrote the manuscript. YJ, XM,JJ, JL, XZ, and HL completed the data collection and management. ZL, YW, and YL have accessed and verified the underlying data. All authors contributed to the interpretation of the data and critical revision of the manuscript. All authors approved the final version of the manuscript.

Data sharing statement

The data that support the findings of this study are available from the corresponding author (Prof. YW, yongjunwang@ncrcnd.org.cn) on reasonable request. Interested parties can apply for data access requests from the website of China National Clinical Research Center for Neurological Diseases at https://www.ncrcnd.org.cn.

Declaration of interests

All authors declare no competing interests.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (grant number Z200016), Beijing Municipal Committee of Science and Technology (grant number Z201100005620010), National Natural Science Foundation of China (grant number 82101360, 92046016, 82171270), Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (grant number 2019-I2M-5-029). We acknowledge the significant contribution of the patients, families, researchers, and clinical staff.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.eclinm.2022.101639.

References

1 Collaborators GBDS. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*. 2019;18(5):439–458.
2 Wang YJ, Li ZX, Gu HQ, et al. China stroke statistics 2019: a report from the National Center for Healthcare Quality Management in Neurological Diseases, China National Clinical Research Center for Neurological Diseases, the Chinese Stroke Association, National Center for Chronic and Non-communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention and Institute for Global Neuroscience and Stroke Collaborations. *Stroke Vasc Neurol*. 2020;5(3):211–239.
3 Lavados PM, Sacks C, Prina L, et al. Incidence, case-fatality rate, and prognosis of ischaemic stroke subtypes in a predominantly Hispanic-Mestizo population in Iquique, Chile (PISCIS project): a community-based incidence study. *Lancet Neurol*. 2007;6(2):140–148.
4 Kaasenbrood L, Boekholdt SM, van der Graaf Y, et al. Distribution of estimated 10-year risk of recurrent vascular events and residual risk in a secondary prevention population. *Circulation*. 2016;134(19):1419–1429.
5 Wang Y. Residual recurrence risk of ischaemic cerebrovascular events: concept, classification and implications. *Stroke Vasc Neurol*. 2021;6(2):155–157.

6 Adams Jr HP, Bendixen BH, Kappelle LJ, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke*. 1993;24(1):35–41.
7 Ay H, Benner T, Arsava EM, et al. A computerized algorithm for etiologic classification of ischemic stroke: the causative classification of stroke system. *Stroke*. 2007;38(11):2979–2984.
8 Ahlqvist E, Storm P, Karajamaki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol*. 2018;6(5):361–369.
9 Seymour CW, Kennedy JN, Wang S, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA*. 2019;321(20):2003–2017.
10 Verdonschot JAJ, Merlo M, Dominguez F, et al. Phenotypic clustering of dilated cardiomyopathy patients highlights important pathophysiological differences. *Eur Heart J*. 2021;42(2):162–174.
11 Sweatt AJ, Hedlin HK, Balasubramanian V, et al. Discovery of distinct immune phenotypes using machine learning in pulmonary arterial hypertension. *Circ Res*. 2019;124(6):904–919.
12 Segar MW, Patel KV, Ayers C, et al. Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis. *Eur J Heart Fail*. 2020;22(1):148–158.
13 Wang Y, Jing J, Meng X, et al. The Third China National Stroke Registry (CNSR-III) for patients with acute ischaemic stroke or transient ischaemic attack: design, rationale and baseline patient characteristics. *Stroke Vasc Neurol*. 2019;4(3):158–164.
14 Warfarin-Aspirin Symptomatic Intracranial Disease Trial I. Design, progress and challenges of a double-blind trial of warfarin versus aspirin for symptomatic intracranial arterial stenosis. *Neuroepidemiology*. 2003;22(2):106–117.
15 Suo Y, Jing J, Meng X, et al. Inconsistent centralised versus non-centralised ischaemic stroke aetiology. *Stroke Vasc Neurol*. 2020;5(4):337–347.
16 Eliasziw M, Rankin RN, Fox AJ, Haynes RB, Barnett HJ. Accuracy and prognostic consequences of ultrasonography in identifying severe carotid artery stenosis. North American Symptomatic Carotid Endarterectomy Trial (NASCET) Group. *Stroke*. 1995;26(10):1747–1752.
17 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI*. Springer, Cham; 2015:234–241.
18 Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017:3146–3154.
19 Xiong L, Xu K, Tian K, et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat Commun*. 2019;10(1):4576.
20 Caliński T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat*. 1974;3(1):1–27.
21 Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;PAMI-1(2):224–227.
22 Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn*. 2003;52:91–118.
23 Saary MJ. Radar plots: a useful way for presenting multivariate health care data. *J Clin Epidemiol*. 2008;61(4):311–317.
24 Grundy SM, Stone NJ, Bailey AL, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. *Circulation*. 2019;139(25):e1082–e1143.
25 Marso SP, Mehta SK, Frutkin A, House JA, McCrary JR, Kulkarni KR. Low adiponectin levels are associated with atherogenic dyslipidemia and lipid-rich plaque in nondiabetic coronary arteries. *Diabetes Care*. 2008;31(5):989–994.
26 Rothenbacher D, Brenner H, Marz W, Koenig W. Adiponectin, risk of coronary heart disease and correlations with cardiovascular risk markers. *Eur Heart J*. 2005;26(16):1640–1646.
27 Sanz Y, Santacruz A, Gauffin P. Gut microbiota in obesity and metabolic disorders. *Proc Nutr Soc*. 2010;69(3):434–441.
28 Cani PD, Amar J, Iglesias MA, et al. Metabolic endotoxemia initiates obesity and insulin resistance. *Diabetes*. 2007;56(7):1761–1772.
29 Wang Z, Klipfell E, Bennett BJ, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*. 2011;472(7341):57–63.
30 Tang WH, Wang Z, Levison BS, et al. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N Engl J Med*. 2013;368(17):1575–1584.
31 Hansson GK, Hermansson A. The immune system in atherosclerosis. *Nat Immunol*. 2011;12(3):204–212.
32 Stoll G, Bendszus M. Inflammation and atherosclerosis: novel insights into plaque formation and destabilization. *Stroke*. 2006;37(7):1923–1932.
33 Yeun JY, Levine RA, Mantadilok V, Kaysen GA. C-Reactive protein predicts all-cause and cardiovascular mortality in hemodialysis patients. *Am J Kidney Dis*. 2000;35(3):469–476.
34 Sarnak MJ, Amann K, Bangalore S, et al. Chronic kidney disease and coronary artery disease: JACC state-of-the-art review. *J Am Coll Cardiol*. 2019;74(14):1823–1838.
35 Macrez R, Ali C, Toutirais O, et al. Stroke and the immune system: from pathophysiology to new therapeutic strategies. *Lancet Neurol*. 2011;10(5):471–480.
36 Stewart RA, White HD, Kirby AC, et al. White blood cell count predicts reduction in coronary heart disease mortality with pravastatin. *Circulation*. 2005;111(14):1756–1762.
37 Soehnlein O. Multiple roles for neutrophils in atherosclerosis. *Circ Res*. 2012;110(6):875–888.
38 Lawler PR, Bhatt DL, Godoy LC, et al. Targeting cardiovascular inflammation: next steps in clinical translation. *Eur Heart J*. 2021;42(1):113–131.
39 Tyrrell DJ, Goldstein DR. Ageing and atherosclerosis: vascular intrinsic and extrinsic factors and potential role of IL-6. *Nat Rev Cardiol*. 2021;18(1):58–68.
40 Iadecola C, Anrather J. The immunology of stroke: from mechanisms to translation. *Nat Med*. 2011;17(7):796–808.
41 Kelly PJ, Murphy S, Coveney S, et al. Anti-inflammatory approaches to ischaemic stroke prevention. *J Neurol Neurosurg Psychiatry*. 2018;89(2):211–218.
42 Coveney S, McCabe JJ, Murphy S, O'Donnell M, Kelly PJ. Anti-inflammatory therapy for preventing stroke and other vascular events after ischaemic stroke or transient ischaemic attack. *Cochrane Database Syst Rev*. 2020;5:CD012825.
43 Tardif JC, Kouz S, Waters DD, et al. Efficacy and safety of low-dose colchicine after myocardial infarction. *N Engl J Med*. 2019;381(26):2497–2505.
44 Parihar SP, Guler R, Brombacher F. Statins: a viable candidate for host-directed therapy against infectious diseases. *Nat Rev Immunol*. 2019;19(2):104–117.
45 Ridker PM, Danielson E, Fonseca FA, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N Engl J Med*. 2008;359(21):2195–2207.
46 Henriksbo BD, Lau TC, Cavallari JF, et al. Fluvastatin causes NLRP3 inflammasome-mediated adipose insulin resistance. *Diabetes*. 2014;63(11):3742–3747.
47 Rezaie-Majd A, Maca T, Bucek RA, et al. Simvastatin reduces expression of cytokines interleukin-6, interleukin-8, and monocyte chemoattractant protein-1 in circulating monocytes from hypercholesterolemic patients. *Arterioscler Thromb Vasc Biol*. 2002;22(7):1194–1199.
48 Oesterle A, Laufs U, Liao JK. Pleiotropic effects of statins on the cardiovascular system. *Circ Res*. 2017;120(1):229–243.
49 Eschenhagen T, Laufs U. Statins do more than lower cholesterol-depending on what you eat? *Circulation*. 2021;143(18):1793–1796.
50 Group VTS. B vitamins in patients with recent transient ischaemic attack or stroke in the VITAmins TO Prevent Stroke (VITATOPS) trial: a randomised, double-blind, parallel, placebo-controlled trial. *Lancet Neurol*. 2010;9(9):855–865.
51 Solomon LR. Disorders of cobalamin (vitamin B12) metabolism: emerging concepts in pathophysiology, diagnosis and treatment. *Blood Rev*. 2007;21(3):113–130.
52 Hansen JM, Go YM, Jones DP. Nuclear and mitochondrial compartmentation of oxidative stress and redox signaling. *Annu Rev Pharmacol Toxicol*. 2006;46:215–234.
53 Merino JG, Lattimore SU, Warach S. Telephone assessment of stroke outcome is reliable. *Stroke*. 2005;36(2):232–233.
54 Moniche F, De La Torre Laviana FJ, Palomino Garcia A, Cayuela Dominguez A, Vigil E, Jimenez MD. Evaluation of telephone assessment in stroke and TIA recurrence. *Neurologia*. 2012;27(2):97–102.