Supplementary 2

The whole CHARLS cohort we downloaded comprised five waves of follow-up interviews every 2-3 years and is detailed below: baseline data from Wave 1 in 2011-2012, Wave 2 in 2013, Wave 3 in 2015, Wave 4 in 2018 and Wave 5 in 2020. More details of CHARLS have been disclosed previously. In our investigation, a total of 17708 participants were included in the initial survey, and some exclusion criteria were applied in Wave 1. Complete laboratory examination data from blood samples and the diagnosis of stroke were required. The exclusion criteria were as follows: (1) had a history of stroke until 2011; (2) did not have records of follow-up since 2011; (3) were aged < 18 years; and (4) had chronic kidney diseases or cancers. Finally, 9516 eligible participants in our study were categorized into four groups. Moreover, 4968 respondents whose complete data were collected in both 2011/2012 and 2015 were included in the analysis to explore the prediction performance of changes in METS-IR. By way of standard questionnaires, the baseline demographic characteristics of the participants were gathered via household interviews (age, sex, marital status, smoking history, alcohol consumption history, hypertension, diabetes and heart problems). NHANES collects data on the health and nutritional status of the U.S. population every two years. Participants in this analysis came from seven two-year survey cycles conducted between 2005 and 2018. For CHARLS, the diagnosis of stroke was defined as “yes” to the following question: “Have you been diagnosed with stroke by a doctor since the last follow-up visit?”. The time of stroke was confirmed in line with the responses to the following questions: “When was the stroke first diagnosed or known by yourself?” and “When was your most recent stroke?”.

Logistic Regression is a supervised learning technique in machine learning with high generality and robustness, mainly used to solve classification tasks. It is widely applied in scientific and medical research to address various classification problems. During training, the LR model learns from the training set, capturing the patterns necessary for making classification decisions. After training, the model applies the learned knowledge to classify new data points in the test set and obtains various features related to the decision-making process.

Extreme Gradient Boosting is an algorithm mainly used for regression and classification problems, with the core being the use of gradient boosting methods to improve model performance and gradually correct previous errors. Its advantages include: efficient handling of nonlinear relationships, strong resistance to overfitting, and high flexibility.

Random Forest is an ensemble learning method commonly used for classification or regression tasks. We can think of it as a "team of decision trees," where each decision tree is a mini-expert skilled at making its own judgments based on a subset of data. The randomness includes random data and random features. Its advantages include high robustness, resistance to overfitting, and suitability for handling multidimensional data.

For LR, we implemented L2 regularization (C=1.0) to prevent overfitting, with feature importance assessed via coefficient magnitude. XGBoost was optimized using a grid search over 50 iterations for n_estimators (50–200), learning_rate (0.01–0.1), and max_depth (3–6), combined with early stopping (10 iterations) to avoid overfitting. RF models utilized 100 decision trees with random feature sampling (max_features='sqrt') and out-of-bag error estimation for internal validation. Cross-validation was performed using 10-fold stratified sampling within the training set, with performance metrics (AUC-ROC, accuracy, F1 score) averaged across folds. Model discrimination was evaluated via the area under the ROC curve, while calibration was assessed using Hosmer-Lemeshow tests. In external validation, the external validation cohort underwent identical preprocessing, and model performance was compared using the DeLong test for AUC differences. To address class imbalance (stroke event rate <15%), we applied weighted cross-entropy loss in XGBoost and class_weight='balanced' in LR, while RF naturally handles imbalance via bootstrap sampling.

In this study, the performance of machine learning models was evaluated using key evaluation metrics: accuracy, precision and F1 score. Accuracy represents the ratio of correctly predicted instances to the total number of instances in the model overall. Precision quantifies the proportion of true positive results among all predicted

positive cases. The F1 score is the harmonic mean of precision and recall, providing a balanced assessment of model accuracy. A higher F1 score indicates better model performance. All continuous features (e.g., METS-IR, BMI) were standardized to z-scores to mitigate scale bias, while categorical variables (e.g., hypertension status) were one-hot encoded.

Accuracy = TP+TN/TP+TN+FP+FN

Precision = TP/TP+FP

Recall = TP/TP+FN

TP True Positive

FP False Positive

FN False Negative

TN True Negative