BMC Gastroenterology

RESEARCH

Open Access

Biological age prediction and NAFLD risk assessment: a machine learning model based on a multicenter population in Nanchang, Jiangxi, China

Lianrui Deng[1†], Jing Huang[2†], Hang Yuan[3], Qiangdong Liu[4,5], Weiming Lou[5], Pengfei Yu[6], Xiaohong Xie[7], Xuyu Chen[8], Yang Yang[8], Li Song[4,5*] and Libin Deng[2,5*]

Abstract

Background The objective was to develop a biological age prediction model (NC-BA) for the Chinese population to enrich the relevant studies in this population. And to investigate the association between accelerated age and NAFLD.

Methods On the basis of the physical examination data of people without noninfectious chronic diseases (PWNCDs) in Nanchang, Jiangxi, China, the biological age measurement method was developed via three feature selection methods (all-subset regression, LASSO regression (LR), and recursive feature elimination) and three machine learning algorithms (generalized linear model (GLM), support vector machine, and deep generalized linear model (deep GLM)). Model performance was evaluated by the coefficient of determination (R^2) and mean absolute error (MAE). National Health and Nutrition Examination Survey (NHANES) data were used to verify the model's generalizability. The standardized age deviation (SAD) was calculated to explore the associations between age acceleration and the risk of morbidity and mortality from NAFLD.

Results The physical examination data of 26,356 PWNCDs were collected in Nanchang. Among the 26 biomarkers, 26 and 24 biomarkers were associated with chronological age in the male and female groups, respectively ($P < 0.05$). The model combining the LR and deep GLM algorithms provided the most accurate measurement of chronological age ($r = 0.58$, MAE = 5.33) and was named the Nanchang-biological age (NC-BA) model. The generalizability of the NC-BA model was verified in the NHANES dataset ($r = 0.57$, MAE = 7.12). There was a significant correlation between NC-BA and existing biological age indicators (Klemera-Doubal method biological age (KDM-BA), PhenoAge, and homeostatic dysregulation (HD), $r = 0.42$–0.66, $P < 0.05$). The physical examination data of 1,663 and 1,445 patients with NAFLD from the Nanchang population and NHANES, respectively, were obtained. The SAD values of NAFLD patients were

†Lianrui Deng and Jing Huang co-first author, these authors contributed equally to this work.

*Correspondence:
Li Song
ndefy91009@ncu.edu.cn
Libin Deng
lbdeng@ncu.edu.cn

Full list of author information is available at the end of the article

significantly greater than those of PWNCDs ($P < 0.001$). The SAD values of NAFLD patients with younger chronological ages were greater ($P < 0.001$). Higher SAD values were associated with a greater risk of all-cause mortality (HR = 1.73, $P = 0.005$).

Conclusions This study provides a new model for biological age measurement in the Chinese population. There is a clear link between NAFLD and age acceleration.

Keywords Biological age, NAFLD, Aging, Machine learning

Introduction

Aging is a natural phenomenon in living organisms that manifests as increased health risks and decreased physiological functions over time [1]. In the context of global aging, the use of medical big data to explore characterization methods, evaluation indicators, and influencing factors is a key step in exploring aging [2]. Biological age (BA), an indicator of aging that is independent of chronological age (CA), is strongly associated with health characteristics such as physical function, cognition, morbidity, and mortality [3]. Compared with CA, BA reflects the rate of aging associated with functional decline more accurately [4, 5].

In recent years, machine learning (ML) algorithms have made significant progress in BA estimation and healthy aging surveillance. Effective BA construction methods can help deepen our understanding of aging and enable more accurate disease risk stratification [6, 7]. Recently, a variety of ML methods have been proposed to quantify BA in populations from different regions, such as the United States [8], Italy [7], and Singapore [9]. However, Cao et al. reported that these findings may not be generalizable to various populations due to differences in genetic and socioenvironmental factors [10]. More importantly, most ML-based BA studies have focused on European and American populations [8, 11], with relatively few studies on Chinese populations [10, 12]. To enrich the research on biological aging in the Chinese population, we selected several machine learning models suitable for analyzing complex biological data. The generalized linear model (GLM) was chosen for its interpretability in elucidating the impact of biomarkers on biological aging and disease risk. The deep generalized linear model (deep GLM) integrates the strengths of deep learning and GLM, capturing complex nonlinear relationships in the data [13]. The support vector machine (SVM) excels in handling high-dimensional data and is routinely used in biomarker development [14, 15].

Nonalcoholic fatty liver disease (NAFLD) is currently the most common chronic liver disease, affecting at least a quarter of the adult population worldwide [16]. Notably, the prevalence of NAFLD continues to increase in China [17]. Studies have shown that biological aging is positively associated with mortality risk in NAFLD patients, especially in young adults, and that cellular senescence markers are significantly increased in NAFLD patients [18]. Researchers have also revealed a significant association between biological aging and an elevated risk of NAFLD [19], with advanced biological aging specifically linked to nonalcoholic steatohepatitis [20]. However, studies on aging in Chinese patients with NAFLD are still scarce.

This study aims to establish a biological age prediction model applicable to the Chinese population, further explore the validity and application potential of BA, and fill the gap in aging research for Chinese NAFLD patients. Here, we established a BA measure, Nanchang-biological age (NC-BA), based on data from Nanchang and the National Health and Nutrition Examination Survey (NHANES). The generalizability of the model was assessed by comparing NC-BA with BA indicators in the literature, including the Klemera-Doubal method (KDM) biological age [21], PhenoAge [22], and homeostatic dysregulation (HD) [23]. In addition, we assessed age acceleration in the NAFLD population by standardized age deviation (SAD), which is linked to all-cause mortality. It is expected to provide new methods and empirical support for the assessment of aging in Chinese NAFLD patients, and to aid in the development of early interventions and precision medicine.

Materials and methods

Study population

This multicenter study included individuals from 2 hospitals in Nanchang between 2017 and 2022 and individuals who participated in the NHANES between 1999 and 2018 (ethics numbers (SFYYXLL-PJ-2022-KY037, 2024-95) for both hospitals). The NHANES is a survey conducted by the United States Centers for Disease Control and Prevention (CDC) to assess the health and nutritional status of noninstitutionalized civilians in the USA. A more detailed description of the NHANES study design and data is publicly available at https://www.cdc.gov/nchs/nhanes/.

Subjects aged 20 to 80 years were included in this study. Eighty years was chosen as the upper age limit because 85 and 80 years were used as the coding cutoff ages in the 1999–2006 and 2007–2018 NHANES data, respectively. In addition, owing to the large number of missing

indicators for subjects under 20 years of age, we decided to exclude subjects in this age group.

Data preprocessing strategy

Both the hospitals and the NHANES collected laboratory data on biomarkers. Biomarkers that were common to the three datasets were included (Table 1). In addition, the largest dataset with the least amount of missing data was defined, accounting for the distribution of the number of available variables for the respondents. Individuals with more than 20% missing values were excluded from the database.

Given the large number of missing values (≤ 20%) in the data, the chi-square binning method was used to assess the ages participants of different sexes. That is, the ages were segmented, and the missing items of the age groups corresponding to the medians of each age group of males and females were used for filling. This not only facilitates the comprehension of the model's logic but also ensures computational efficiency when dealing with large-scale datasets.

Table 1 Baseline characteristics of the study population in Jiangxi

	PWNCDs *N* = 39,833	NAFLD patients *N* = 1,663	*P* value
Sex (%):			< 0.001
Male	23,557 (59.1)	1,499 (90.1)	
Female	16,276 (40.9)	164 (9.9)	
AGE (mean (SD))	40.99 (13.38)	46.27 (12.59)	< 0.001
ALT (mean (SD))	19.22 (9.96)	30.68 (21.76)	< 0.001
AST (mean (SD))	19.92 (5.09)	26.28 (10.78)	< 0.001
BASO# (mean (SD))	0.01 (0.02)	0.03 (0.02)	< 0.001
BASO% (mean (SD))	0.19 (0.28)	0.55 (0.29)	< 0.001
CHOL (mean (SD))	4.74 (0.81)	5.21 (1.04)	< 0.001
EO# (mean (SD))	0.11 (0.07)	0.16 (0.12)	< 0.001
EO% (mean (SD))	1.93 (1.14)	2.58 (1.82)	< 0.001
GLU (mean (SD))	4.95 (0.45)	5.33 (1.32)	< 0.001
HCT (mean (SD))	42.79 (3.63)	45.69 (3.57)	< 0.001
HDL_C (mean (SD))	1.37 (0.30)	1.21 (0.31)	< 0.001
HGB (mean (SD))	143.74 (13.34)	152.55 (13.10)	< 0.001
PLT (mean (SD))	222.32 (46.77)	234.89 (56.40)	< 0.001
RDW (mean (SD))	12.61 (0.60)	12.59 (0.97)	0.208
TBIL (mean (SD))	14.65 (4.30)	15.20 (5.63)	< 0.001
TGL (mean (SD))	1.28 (0.60)	2.05 (1.99)	< 0.001
URCA (mean (SD))	347.33 (83.71)	408.88 (84.54)	< 0.001
LYM (mean (SD))	2.05 (0.50)	2.04 (0.59)	0.770
LYM% (mean (SD))	35.25 (6.76)	33.43 (7.05)	< 0.001
MCH (mean (SD))	30.47 (1.27)	30.50 (2.14)	0.440
MCHC (mean (SD))	335.91 (9.93)	333.84 (11.58)	< 0.001
RBC (mean (SD))	4.72 (0.44)	5.02 (0.48)	< 0.001
MONO (mean (SD))	0.39 (0.12)	0.43 (0.13)	< 0.001
MONO% (mean (SD))	6.75 (1.61)	7.00 (1.61)	< 0.001
MPV (mean (SD))	10.64 (0.89)	10.78 (1.03)	< 0.001
PCT (mean (SD))	0.24 (0.04)	0.26 (0.06)	< 0.001
MCV (mean (SD))	90.77 (3.54)	91.31 (5.32)	< 0.001

Notes: PWNCDs = people without noninfectious chronic diseases; NAFLD = nonalcoholic fatty liver disease; ALT = alanine aminotransferase; AST = aspartate aminotransferase; BASO# = absolute value of basophils; BASO% = percentage of basophils; CHOL = total cholesterol; EO# = absolute value of eosinophils; EO% = percentage of eosinophils; GLU = glucose; HCT = hematocrit; HDL_C = high-density lipoprotein; HGB = hemoglobin; PLT = platelet; RDW = red blood cell distribution width; TBIL = total bilirubin; TGL = triglyceride; URCA = uric acid; LYM = percentage of eosinophils; LYM% = lymphocyte ratio; MCH = mean hemoglobin of red blood cells; MCHC = average hemoglobin concentration; RBC = red blood cell; MONO = monocyte count; MONO% = percentage of monocytes; MPV = average platelet volume; PCT = platelet volume; MCV = average red blood cell volume

Definitions of PWNCDs and NAFLD patients

Owing to inconsistent information across the three datasets, the definitions of PWNCDs and NAFLD patients differed.

PWNCDs: For the Center 1 and Center 2 datasets, we first excluded participants with the following chronic diseases associated with aging: hypertension, type 2 diabetes, cancer (excluding minor skin cancer), chronic lung disease, heart problems (heart attack, coronary heart disease, angina, congestive heart failure), and stroke [24]. We further excluded patients with extreme or outlier values. After this screening process, the remaining subjects were defined as PWNCDs.

When processing the NHANES data, we adopted similar exclusion criteria but considered a broader range of chronic systemic diseases, including diseases of the digestive, cardiovascular, metabolic, visual, urogenital, respiratory, and immune systems; musculoskeletal diseases; and neoplasms [25]. We first identified and excluded individuals with these disorders and then similarly excluded those with extreme or outlier values. After this series of screening steps, the final population comprised the PWNCDs included in the NHANES database.

NAFLD patients: NAFLD was diagnosed at Nanchang Hospital on the basis of the Guidelines for the Prevention and Treatment of Nonalcoholic Fatty Liver Disease (2018 Edition) [26]: (1) the absence of a history of alcohol consumption or the consumption of less than 210 g of alcohol per week in men (less than 140 g per week in women); (2) the exclusion of diseases that can lead to fatty liver, such as infection with hepatitis C virus (HCV) genotype 3, Wilson's disease, autoimmune hepatitis, total parenteral nutrition, etc.; and (3) when evaluating hepatic steatosis with abdominal ultrasound, two of the following three criteria were met: anterior field echo enhancement ("bright liver"), a liver echo greater than the kidney, far-field echo attenuation, and an unclear display of the intrahepatic duct structure. Ultrasound has high sensitivity (85%) for detecting moderate to severe fatty liver, but its accuracy decreases in obese patients or those with comorbid kidney disease, and its ability to detect mild fatty liver is limited [27].

In the NHANES 2017–2018 cycle, liver fat was quantified by the controlled attenuation parameter (CAP) [27]. In this study, fatty liver disease (FLD) was diagnosed when the median CAP score was ≥ 285 dB/m, with an optimal sensitivity of 80% and a specificity of 77% for detecting hepatic steatosis [28]. In the absence of other chronic liver diseases or excessive alcohol consumption (< 20 g/day for men and < 10 g/day for women), individuals with FLD were identified as having NAFLD [29–31]. CAP outperforms ultrasound in detecting and grading hepatic steatosis, but it is influenced by factors such as diabetes and BMI, and it cannot effectively distinguish between adjacent grades of steatosis [27].

Machine learning for BA

Data standardization

To reduce the impact of different data ranges, the data were standardized via the Z score method before modeling so that they could be used for subsequent modeling and analysis.

Feature selection

We used all-subset regression, LASSO regression [32], and recursive feature elimination to select biomarkers for men and women. The study employed all-subset regression to comprehensively evaluate the optimal combination of features, utilized the automatic feature compression of LASSO with L1 regularization for dimensionality reduction, and applied recursive feature elimination to balance computational efficiency and the accuracy of feature selection.

Best model selection

In this study, the generalized linear model (GLM), deep generalized linear model (deep GLM), and support vector machine (SVM) methods were used to construct BA prediction models. The Center 1 dataset was divided into training and test sets at a ratio of 8:2, and CA was used as the target value to train the algorithm. Using the training dataset, a grid-search exploration of hyperparameters with a tenfold cross-validation was performed for each model. Different machine learning methods were compared according to the coefficient of determination (R^2) and mean absolute error (MAE) [33].

Estimation of age acceleration

To accurately estimate age acceleration and exclude interference from physiological factors, we introduced the standardized age deviation (SAD):

$$SAD = \frac{\Delta\text{age} - \overline{X}_{\Delta\text{age}}}{\text{Std}_{\Delta age}} \tag{1.1}$$

$$\Delta age = BA - CA \tag{1.2}$$

where Δage represents the difference between theBA and the CA, $\overline{X}$ represents the mean of Δage within a ± 5-year interval of the CA, and Std represents the standard deviation of Δage within a ± 5-year interval of the CA.

KDM, phenoage, and HD algorithms

We used the _nhanes function of the BioAge software package to calculate 3 bioaging measurements: the KDM-BA, PhenoAge, and HD.

Statistical analysis

Baseline measured traits are expressed as the mean (standard deviation [SD]) or number (percentage), and the correlations between traits and targets are expressed as Pearson correlation coefficients. We compared the SAD in NAFLD patients and PWNCDs and different disease states via t tests and analysis of variance, respectively.

We used mortality data from the 2019 Public Mortality File of the NHANES database and merged them with data from the NHANES database on the basis of respondents' SEQNs. In conducting the analysis, we considered the complex sampling design and weights of the data. We set the median SAD of males, females, and all patients as cutoff points and then divided the groups into a higher-SAD group and a lower-SAD group. We then evaluated the relationship between SAD and all-cause mortality in patients with NAFLD by conducting CA-adjusted weighted Cox proportional hazards regression model analysis, using restricted cubic spline (RCS) to visualize the potential nonlinear association between SAD and all-cause mortality in patients with NAFLD.

All the statistical analyses were performed via R version 4.3.2. $P < 0.05$ (two-tailed) was considered statistically significant.

Results

Baseline characteristics of the population

A total of 39,774 PWNCDs were obtained among 100,063 individuals in two centers in Nanchang. As shown in Fig. 1, we used 26,356 PWNCDs in Center 1 to construct a BA prediction model and validated the model with 13,418 individuals from Center 2. In addition, 1,663 patients with NAFLD were screened.

Comparison of feature selection methods and machine learning algorithms

After performing quality control, we obtained 26 validated biological traits and the basic characteristics of PWNCDs and NAFLD patients in Nanchang (Table 1). On the basis of the correlation analysis, 26 and 24 biological traits were identified as candidate traits for CA in the male and female groups, respectively ($P < 0.05$, Fig. 2A,

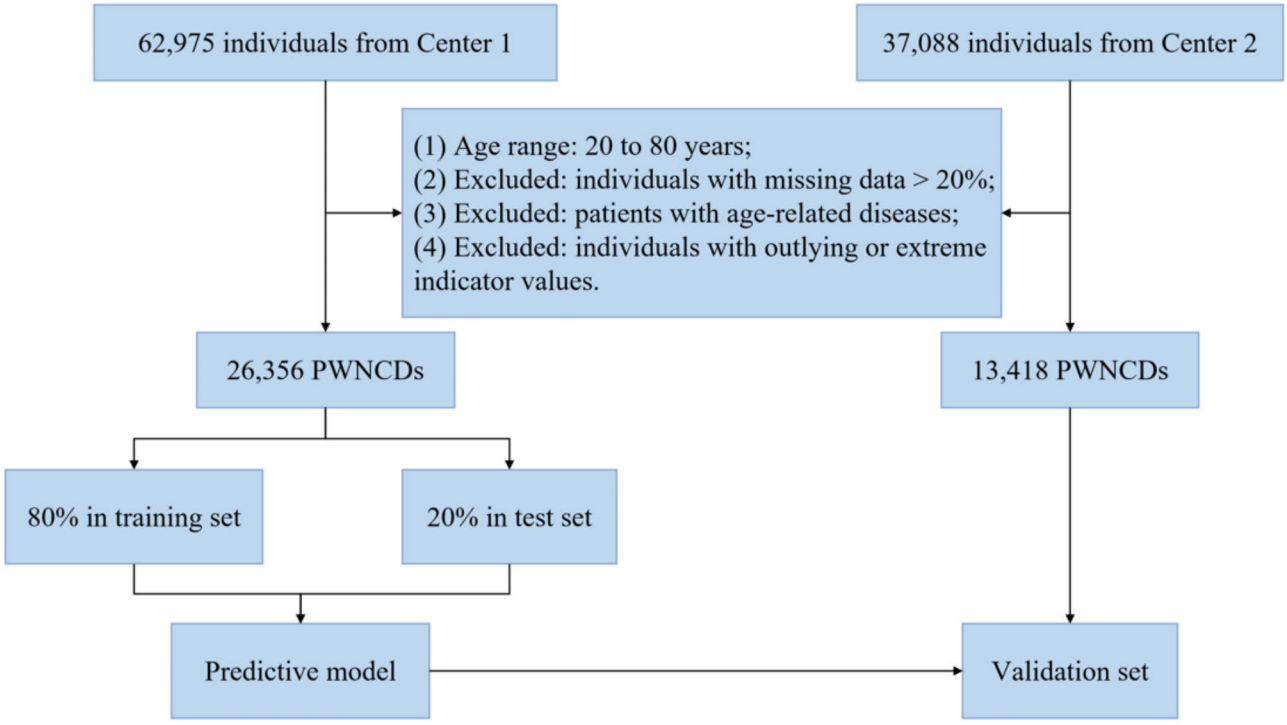

Fig. 1 Screening flow chart of Nanchang participants. PWNCDs, people without noninfectious chronic diseases

B). As shown in Fig. 2C, there was a certain correlation among these features, suggesting that the features needed to be screened during the modeling process. We compared the effects of three feature selection methods (all-subset regression, LASSO regression, and recursive feature elimination) on the performance of various ML algorithms (GLM, deep GLM, and SVM). Compared with the ML model without feature selection, the LASSO regression method can improve the performance of the model (Fig. 2D, E).

Further analysis revealed that the combination of the Lasso feature selection method and the deep GLM algorithm had the best performance in predicting CA. The R^2 values of this combination were 0.60 and 0.67 for males and females, respectively. As shown in Supplementary Fig. 1, the combined method also performed best in terms of the MAE, with values of 6.78 for males and 5.78 for females. Supplementary Fig. 2 illustrates in detail the optimization process of lambda parameters in LASSO regression and its influence on feature selection. The most important features identified by Lasso for males were glucose (GLU) and red cell distribution width (RDW), whereas for females, total cholesterol (CHOL) and triglycerides (TGL) were highlighted as the most significant (Supplementary Table).

Evaluating the performance of different models on the center 2 dataset

To verify the performance of different ML algorithms in predicting CA, the optimal feature selection method (LASSO) was applied in PWNCDs from Center 2. We assessed the different models by R^2 and the MAE (Fig. 3A, B). The results show that the deep GLM has the best performance in predicting CA. Subsequently, this model was named the NC-BA model, and the Pearson correlation coefficient between CA and NC-BA was 0.58 (Fig. 3C). These results indicated that the BA predicted by the deep GLM had a high correlation with the CA in external datasets.

Generalization of the NC-BA in the NHANES

We screened 101,316 individuals from the NHANES dataset and identified 11,447 PWNCDs (Fig. 4A). Additionally, we identified 1,445 patients with NAFLD from the NHANES dataset. The baseline characteristics of the PWNCDs and NAFLD patients are shown in Table 2.

We calculated the NC-BA of PWNCDs in the NHANES dataset. As shown in Fig. 4B, NC-BA was strongly correlated with CA in the NHANES dataset ($r = 0.57$, MAE = 7.12). This result suggested that the NC-BA has potential for use in different populations. To further validate the generalizability of the NC-BA, we calculated three BA measurements (KDM-BA, PhenoAge, and HD) via the _nhanes function of the BioAge package [34]. We detected a significant correlation

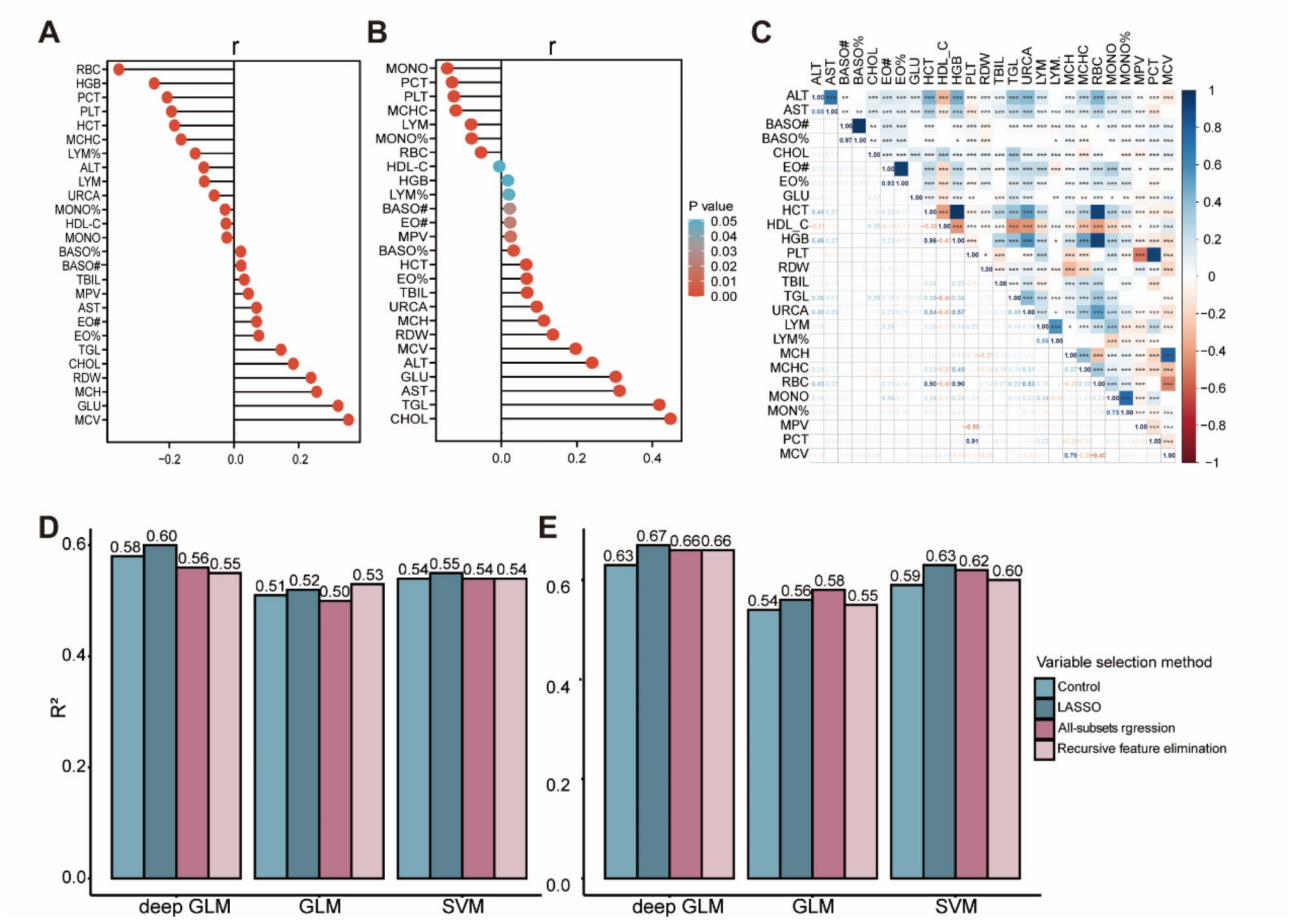

Fig. 2 Comparison of different feature selection methods and different classes of ML models. (**A**) and (**B**) explore the correlation between biological traits and CA by sex via data from PWNCDs from Center 1. The color of the dot represents the P value, and the length of the horizontal line represents the Pearson correlation coefficient. (**A**) Correlations among males and (**B**) correlations among females. (**C**) Correlation heatmap of biomarkers in Center 1. (**D**) and (**E**) show the performance of different feature selection methods and models on the male and female datasets, respectively, to estimate BA. The control was no feature selection. The LASSO feature screening method and the deep GLM achieved the best performance

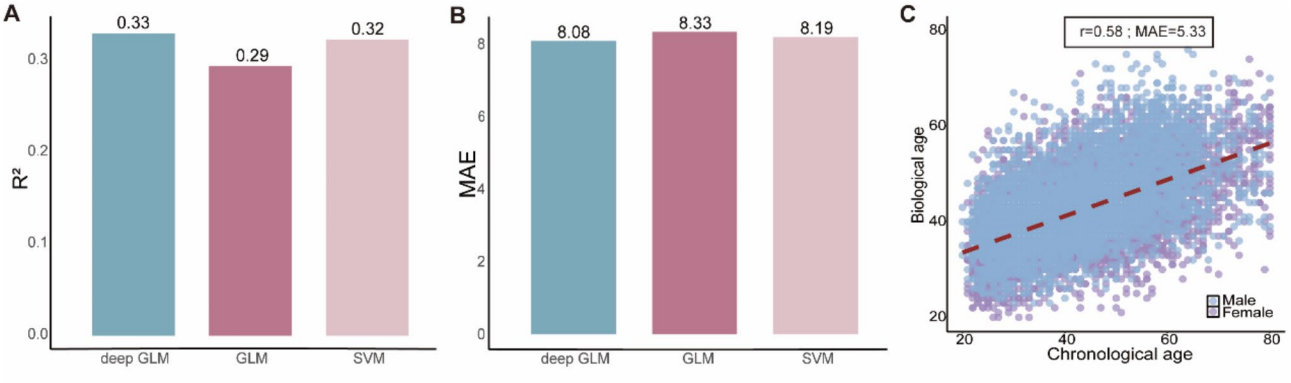

Fig. 3 Performance of different models in predicting CA via the Center 2 dataset. (**A**) and (**B**) CA was estimated with three models, and the performance of the final implementation was compared on this external dataset. (**A**) Performance of the model for males and (**B**) performance of the model for females. (**C**) Correlation of CA with NC-BA calculated in the best machine learning mode (deep GLM)

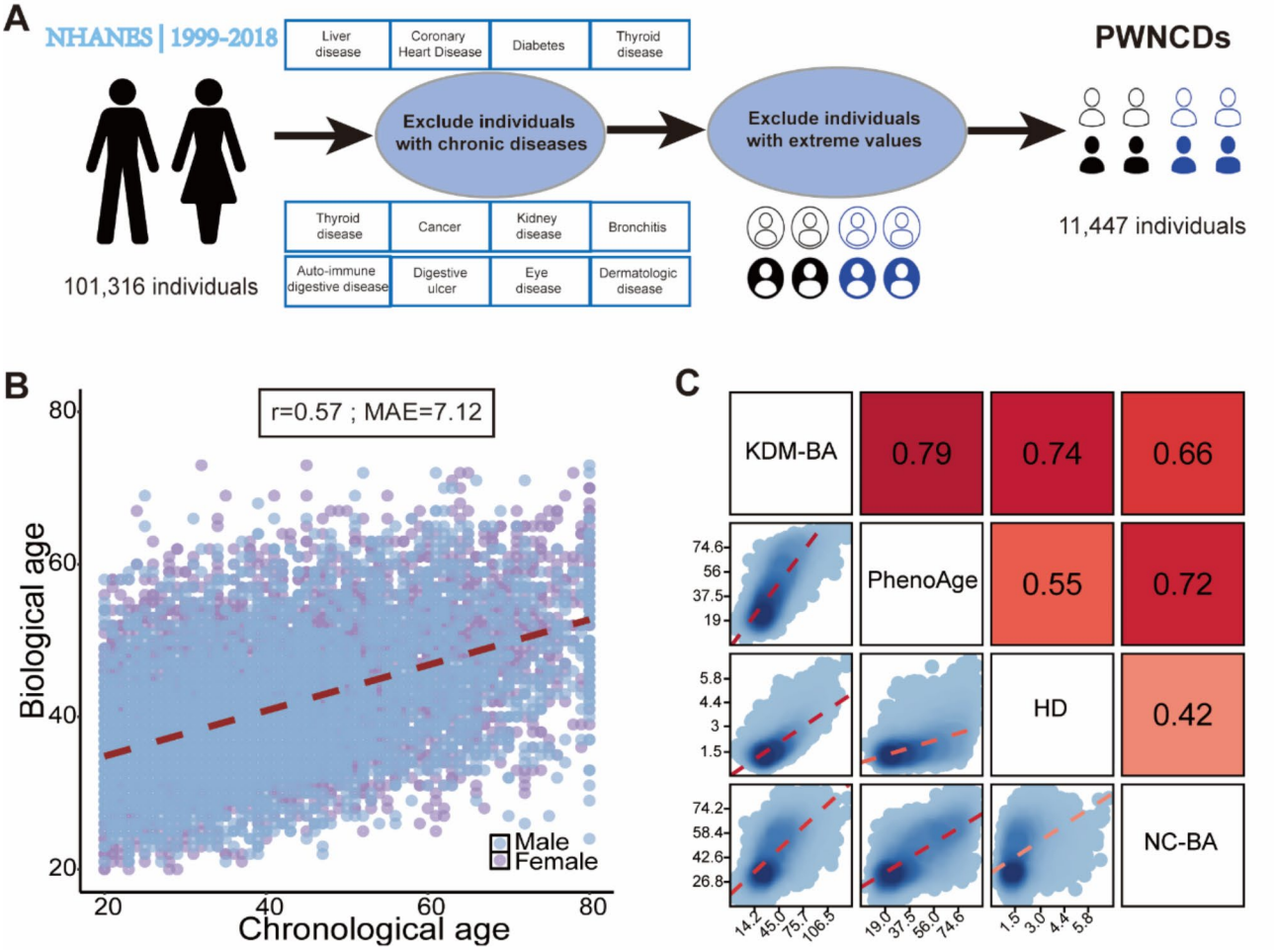

Fig. 4 Generalizability of the NC-BA in NHANES data. (**A**) Flow chart of screening PWNCDs in the NHANES database. All data from the 1999–2018 NHANES cycle were collected, and individuals with chronic diseases and individuals with extreme values or outliers were excluded. As a result, there were 11,447 PWNCDs in the NHANES database. (**B**) Correlation of CA with the NC-BA in PWNCDs from the NHANES database. (**C**) Correlation of the BA determined via different algorithms. We estimated the correlations among KDM-BA, PhenoAge, HD and NC-BA. The color depth indicates the magnitude of the correlation coefficient

between the NC-BA and these measurements ($P<0.05$), and the correlation coefficients ranged from 0.42 to 0.72 (Fig. 4C). These results suggest that the NC-BA has high accuracy and reliability in predicting BA.

Evaluating age acceleration in NAFLD patients from Nanchang

To evaluate the relationship between age acceleration and NAFLD patients in Nanchang, we obtained a measure of age acceleration (SAD) based on the NC-BA (Fig. 5). The average SAD of NAFLD patients was significantly greater than that of PWNCDs, indicating that NAFLD patients experienced a greater degree of aging ($P<0.001$, Fig. 5A). As shown in Fig. 5B, the SAD of NAFLD patients tended to be on the older side. In addition, there was a negative correlation between the SAD and CA (r=-0.36, $P<0.001$), suggesting a greater degree of aging in younger patients with NAFLD.

In this study, patients with NAFLD were followed up for a short period of one year. A total of 743 patients with NAFLD were followed up, and we observed the recovery periods of 552 patients. On this basis, we divided NAFLD patients into two subgroups: nonrecovered NAFLD patients and recovered NAFLD patients. By comparing the SAD at the time of NAFLD diagnosis, we found that the median SAD was the highest in the nonrecovered NAFLD group (0.50), followed by the recovered NAFLD group (0.32) (Fig. 5C). This result further confirmed the strong association between age acceleration and the pathological outcome of NAFLD.

Age acceleration in NAFLD patients from the NHANES

For the 1,445 patients with NAFLD from the NHANES, the SAD was significantly greater than that in the PWNCDs, and the distribution of the SAD tended to increase with age (Fig. 6A, B). Additionally, the SAD was

Table 2 Baseline characteristics of the NHANES study population

	PWNCDs *N* = 11,447	NAFLD patients *N* = 1,445	*P* value
Sex (%):			0.298
Male	6,088 (53.2)	790 (54.7)	
Female	5,359 (46.8)	655 (45.3)	
AGE (mean (SD))	39.94 (14.78)	53.91 (15.93)	< 0.001
ALT (mean (SD))	21.34 (8.28)	26.53 (18.59)	< 0.001
AST (mean (SD))	22.13 (5.01)	22.80 (12.39)	< 0.001
BASO# (mean (SD))	0.04 (0.05)	0.06 (0.05)	< 0.001
BASO% (mean (SD))	0.67 (0.42)	0.78 (0.32)	< 0.001
CHOL (mean (SD))	4.93 (0.91)	4.93 (1.09)	0.913
EO# (mean (SD))	0.16 (0.10)	0.22 (0.17)	< 0.001
EO% (mean (SD))	2.35 (1.31)	2.86 (1.99)	< 0.001
GLU (mean (SD))	4.98 (0.59)	6.24 (2.43)	< 0.001
HCT (mean (SD))	42.33 (3.92)	42.34 (4.11)	0.968
HDL_C (mean (SD))	1.38 (0.35)	1.22 (0.32)	< 0.001
HGB (mean (SD))	14.38 (1.34)	14.24 (1.51)	< 0.001
PLT (mean (SD))	248.90 (53.02)	245.56 (64.97)	0.028
RDW (mean (SD))	12.73 (0.78)	13.91 (1.28)	< 0.001
TBIL (mean (SD))	11.42 (4.20)	7.66 (4.29)	< 0.001
TGL (mean (SD))	1.32 (0.71)	2.07 (1.50)	< 0.001
URCA (mean (SD))	306.57 (76.59)	348.75 (87.92)	< 0.001
LYM (mean (SD))	2.08 (0.55)	2.35 (0.77)	< 0.001
LYM% (mean (SD))	31.04 (7.46)	31.30 (8.67)	0.209
MCH (mean (SD))	30.50 (1.60)	29.52 (2.33)	< 0.001
MCHC (mean (SD))	339.66 (8.16)	336.06 (9.25)	< 0.001
RBC (mean (SD))	4.72 (0.46)	4.84 (0.50)	< 0.001
MONO (mean (SD))	0.52 (0.15)	0.61 (0.26)	< 0.001
MONO% (mean (SD))	7.70 (1.88)	8.01 (2.36)	< 0.001
MPV (mean (SD))	8.16 (0.82)	8.27 (0.90)	< 0.001
PCT (mean (SD))	0.20 (0.04)	0.20 (0.05)	0.632
MCV (mean (SD))	89.78 (4.05)	87.76 (5.76)	< 0.001

Notes: PWNCDs = people without noninfectious chronic diseases; NAFLD = nonalcoholic fatty liver disease; ALT = alanine aminotransferase; AST = aspartate aminotransferase; BASO# = absolute value of basophils; BASO% = percentage of basophils; CHOL = total cholesterol; EO# = absolute value of eosinophils; EO% = percentage of eosinophils; GLU = glucose; HCT = hematocrit; HDL_C = high-density lipoprotein; HGB = hemoglobin; PLT = platelet; RDW = red blood cell distribution width; TBIL = total bilirubin; TGL = triglyceride; URCA = uric acid; LYM = percentage of eosinophils; LYM% = lymphocyte ratio; MCH = mean hemoglobin of red blood cells; MCHC = average hemoglobin concentration; RBC = red blood cell; MONO = monocyte count; MONO% = percentage of monocytes; MPV = average platelet volume; PCT = platelet volume; MCV = average red blood cell volume

significantly greater in younger NAFLD patients (r=-0.46, $P < 0.001$; Fig. 6B).

During a median follow-up of 25 months (interquartile range (IQR) 18–32 months), 26 of the 1439 patients with NAFLD (1.81%) died. We assessed the estimated hazard ratios (HRs) for NAFLD under different SAD stratifications (Fig. 6C). The risk of all-cause mortality was significantly greater in the high-SAD group than in the low-SAD group (HR 6.25, 95% CI 1.44–27.05, $P = 0.014$), and the same was observed in men (HR 13.78, 95% CI 2.30–22.55, $P = 0.004$). RCS analysis revealed that the SAD was positively and linearly associated with all-cause mortality in all patients with NAFLD (nonlinear $P = 0.287$) and in male patients (nonlinear $P = 0.275$) (Fig. 6D, F). According to the CA-adjusted weighted Cox regression analysis, the risk of all-cause mortality increased by 106% in men (HR 2.06, 95% CI 1.51–2.82, $P < 0.0001$) for each unit of increase in the SAD in males and increased by 73% in all patients (HR 1.73, 95% CI 1.18–2.55, $P = 0.005$).

Discussion

In this study, we developed a measurement of BA (NC-BA) that is applicable to the Chinese population. By comprehensively comparing the combined methods of three feature selection and three ML algorithms, we found that the deep GLM algorithm using the LASSO method had the best prediction performance. Notably, LASSO regression identified key biomarkers from males and females. In males, GLU and erythrocyte-related markers were most prominent, highlighting the importance of glucose metabolism and erythrocyte status in biological aging. In females, lipid metabolism markers were particularly influential. These findings emphasize the gender differences in the physiological processes that promote biological aging. NC-BA-based SAD data revealed a statistically significant difference between PWNCDs and NAFLD patients. Additionally, the SAD was related to the risk of death from NAFLD.

The generalizability of the NC-BA was evaluated in two independent populations. The correlation between BA and CA decreased after the model was applied to the NHANES dataset, which is consistent with the findings of Mamoshina et al. [10, 35]. This phenomenon may have arisen because of racial factors when the model was evaluated in different populations. Importantly, the influence of factors such as economic status, cultural and lifestyle on the aging process varies from country to country [36]. For example, dietary patterns, physical activity levels, and healthcare access differ between these populations, and these factors might influence both biological age and disease progression. In addition, slight operational biases between different laboratories may also have an impact on the results [35].

On the basis of NHANES data, three published methods for determining BA (KDM-BA, PhenoAge, and HD) were obtained and showed a moderate correlation with the NC-BA. The differences among BAs are largely due to the differences in algorithms/methods from which they are derived. In addition, biomarkers representing different physiological functions or systemic domains also affect the assessment of the aging process. Previous studies revealed that immunity, metabolism, liver dysregulation and kidney dysregulation are associated with aging [37–41]. Therefore, we selected biomarkers reflecting

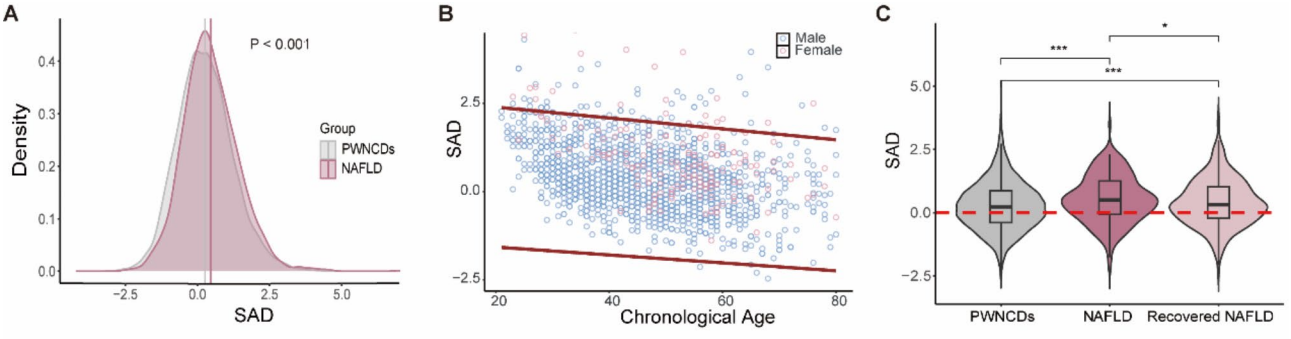

Fig. 5 Age acceleration among PWNCDs and NAFLD patients in Nanchang. (**A**) Distribution of the SAD in PWNCDs and NAFLD patients from Nanchang. The distribution of the SAD is shown in gray for PWNCDs and in red for NAFLD patients. The straight line is expressed as the mean of the two groups of people. (**B**) Scatter plot of SAD and CA in NAFLD patients, with blue dots representing males and red dots representing females. The upper red line is the 97.5th percentile fitted line for the SAD in PWNCDs with different CAs, and the lower red line is the 2.5th percentile fitted line for the SAD in PWNCDs with different CAs. (**C**) Comparative analysis of the SAD among PWNCDs, nonrecovered NAFLD patients, and recovered NAFLD patients. ∗ indicates $P < 0.05$, ∗∗∗indicates $P < 0.001$

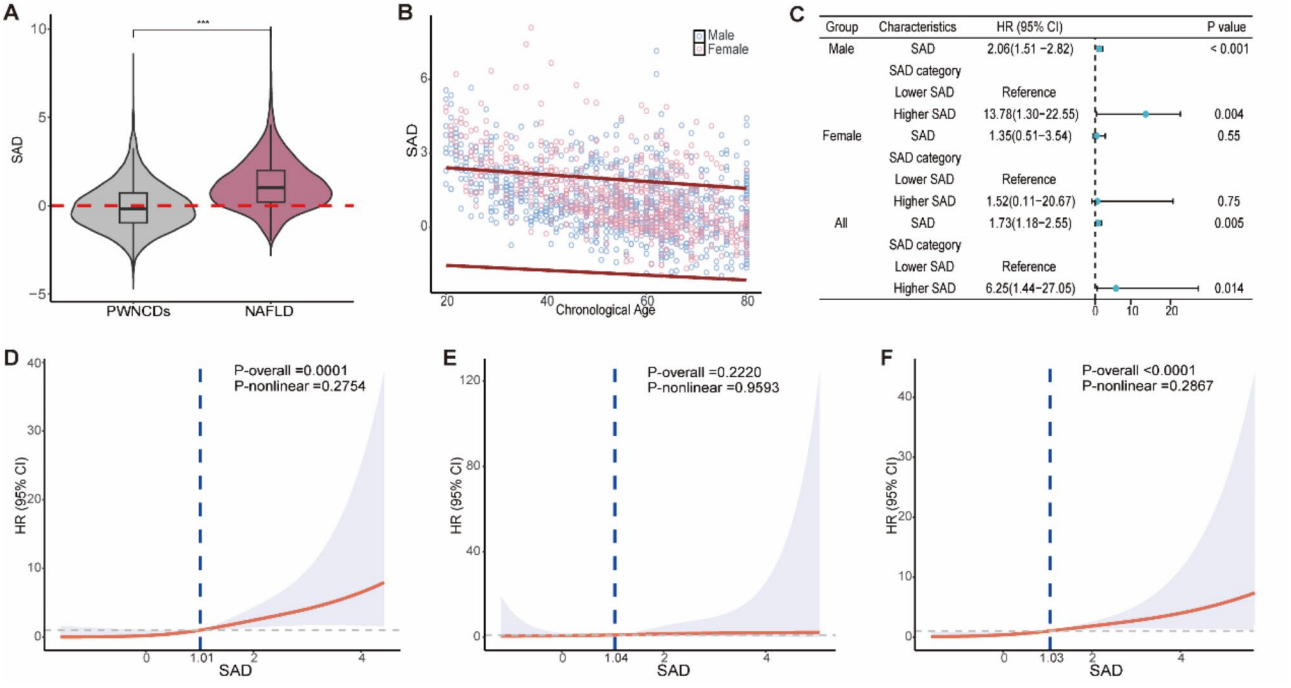

Fig. 6 Assessing age acceleration in NAFLD patients from the NHANES via the SAD. (**A**) Violin plot showing the SAD distributions for PWNCDs and NAFLD patients. The horizontal bars within boxes denote medians, and the tops and bottoms of the boxes represent the 25th and 75th percentiles, respectively. (**B**) Scatter plot of the SAD and CA in NAFLD patients from the NHANES, with blue dots representing males and red dots representing females. The upper red line is the 97.5th percentile fitted line for the SAD in PWNCDs with different CAs, and the lower red line is the 2.5th percentile fitted line for the SAD in PWNCDs with different CAs. (**C**) The forest plot shows stratification by sex, with the median of each group's SAD indicators divided into a lower SAD group and a higher SAD group, and the hazard ratios adjusted for CA. (**D-F**) The association between SAD and all-cause mortality. The limiting triplicate spline shows the relationship between the SAD and all-cause mortality in male NAFLD patients (**D**), female NAFLD patients (**E**), and all NAFLD patients (**F**), and the median of the SAD indicators is the dashed line according to the CA-adjusted hazard ratios. All P values for nonlinearity were > 0.05

the immune system (such as platelets), cardiometabolic system (such as total cholesterol), hepatic system (such as alanine aminotransferase), and renal system (such as uric acid) for estimating the NC-BA. By integrating these multisystem biomarkers, we can provide a more comprehensive picture of the aging process.

In most existing studies, age acceleration is directly calculated by the residual value of the BA returned to the CA [24, 42, 43]. Considering the different rates of aging in individuals at different ages [44], in this study, we calculated the mean and standard deviation of the difference between the BA and CA for each individual on the basis of the population of PWNCDs within a ± 5-year range.

This allowed us to compute the individual SAD, enabling comparisons of the current physiological state with the reference population as closely as possible.

NAFLD is a long-term disease, and monitoring the disease course and predicting mortality risk are effective indicators. BA, serving as such an indicator, can also capture physiological alterations earlier than specific phenotypes [12]. In the Nanchang study, patients who did not recover from NAFLD during subsequent follow-up showed signs of age acceleration at baseline. In addition, the mean SAD values of recovered NAFLD patients were between those of PWNCDs and those of nonrecovered patients. These findings suggest that SAD has the potential to be a useful tool for evaluating the severity of disease in patients with NAFLD. Moreover, we revealed that SAD was significantly associated with increased all-cause mortality risk among people with NAFLD, which is consistent with the findings of a previous study [45].

In addition, we found that the SAD was significantly greater in younger NAFLD patients. This is likely due to the earlier exposure of the younger generation to various risk factors and environmental insults, such as a poor diet, low physical activity, poor mental health, and environmental stress [46]. Considering that BA is modifiable [47, 48], monitoring and treatment for these young patients may be highly important. Notably, in our study, the SAD of NAFLD patients from the NHANES data exhibited a more pronounced age trend. This is primarily due to differences in diagnostic criteria, which result in a higher CA and disease severity for this patient population.

Some limitations may be mentioned in the present study. First, instead of directly calculating the calibration and predictive discriminatory power of the all-cause mortality model, we used the SAD which indirectly reflects the effect of accelerated age on the health status of patients with NAFLD. Future research should validate the correlation of SAD with long-term mortality in NAFLD patients through longitudinal cohorts and improve its predictive accuracy via model calibration to enhance its clinical value for personalized health assessment and risk stratification. Second, socioeconomic factors (e.g., education, income) may confound the results by interacting with genetic factors and lifestyle choices (diet, exercise) to influence biological aging and NAFLD. Future studies should control for these variables to better assess their impact.

In summary, we developed a valid measure of biological aging for the Chinese population and demonstrated that SAD indicators are associated with the severity of NAFLD as well as the risk of death. This study not only validates the application of machine learning in gerontology but also deepens our insight into the multifaceted nature of aging.

Conclusion

We successfully developed a BA prediction model, the NC-BA, suitable for the Chinese population. We also showed that NC-BA is generalizable across populations and is correlated with existing biological age indicators. Furthermore, we used SAD to assess age acceleration in NAFLD patients and found an association between SAD and the severity of NAFLD. Additionally, we found that SAD was associated with the risk of all-cause mortality in these patients.

Abbreviations

NAFLD	Nonalcoholic fatty liver disease
PWNCDs	People without noninfectious chronic diseases
LR	LASSO regression
GLM	Generalized linear model
deep GLM	Deep generalized linear model
MAE	Mean absolute error
SAD	Standard age deviation
NC-BA	Nanchang-biological age
KDM-BA	Klemera-Doubal method biological age
HD	Homeostatic dysregulation
BA	Biological age
CA	Chronological age
ML	Machine learning
HCV	Hepatitis C virus
VCTE	Vibration-controlled transient elastography
CAP	Controlled attenuation parameter
FLD	Fatty liver disease
SVM	Support vector machine
SD	Standard deviation
RCS	Restricted cubic spline

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12876-025-03752-y .

Supplementary Material 1

Acknowledgements

We gratefully acknowledge all the people who helped in the establishment of the data set.

Author contributions

LD and LS performed study concept, supervision, project administration and design; LD, JH and HY performed original article; LD, QL, WL and PY performed development of methodology, formal analysis, data curation and writing, review and revision of the paper; JH, XX and XC provided acquisition, analysis and interpretation of data, and statistical analysis; HY, and YY are responsible for the investigation and conceptualization. LD provided funding acquisition. All authors reviewed the manuscript.

Funding

This work was supported by the National Natural Sciences Foundation of China (82060112).

Data availability

The datasets generated and/or analysed during the current study are available in the NHANES, https://www.cdc.gov/nchs/nhanes/. In addition, the availability of data that support the findings of this study are available from the Second Affiliated Hospital and the Fourth Affiliated Hospital of Nanchang University but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Second Affiliated Hospital and the Fourth Affiliated Hospital of Nanchang University.

Declarations

Ethics approval and consent to participate
The study protocol was approved by the Fourth Affiliated Hospital of Nanchang University Ethics Committee (SFYYXLL-PJ-2022-KY037) and the Second Affiliated Hospital of Nanchang University Ethics Committee (2024-95). All the data and methods used in this study are in line with the relevant regulations of ethical review. Owing to the retrospective nature of the study, informed consent from patients was waived by the Second Affiliated Hospital of Nanchang University and the Fourth Affiliated Hospital of Nanchang University in accordance with national legislation and institutional requirements. Approval for consent was waived, as NHANES data are publicly available.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Affiliated Rehabilitation Hospital of Nanchang University, Nanchang, China
[2]School of Public Health, Jiangxi Medical College, Jiangxi Provincial Key Laboratory of Disease Prevention and Public Health, Nanchang University, Nanchang, China
[3]Chaisang District Center for Disease Control and Prevention, Jiujiang, China
[4]Center of Stomatology, The Second Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang, China
[5]The Institute of Periodontal Disease, JXHC Key Laboratory of Periodontology (The Second Affiliated Hospital of Nanchang University), Nanchang University, Nanchang, China
[6]Big Data Research Center, The Second Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang, China
[7]Sanming City Shaxian District General Hospital, Nanchang, China
[8]Physical Examination Center, The Second Affiliated Hospital of Nanchang University, Nanchang, China

Received: 1 November 2024 / Accepted: 3 March 2025
Published online: 13 March 2025

References

1. Galkin F, Zhang B, Dmitriev SE, et al. Reversibility of irreversible aging. Ageing Res Rev. 2019;49:104–14.
2. Zhang B, Trapp A, Kerepesi C, et al. Emerging rejuvenation strategies-Reducing the biological age. Aging Cell. 2022;21:e13538.
3. Levine ME. Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? J Gerontol Biol Sci Med Sci. 2013;68:667–74.
4. Hannum G, Guinney J, Zhao L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49:359–67.
5. Jylhävä J, Pedersen NL, S Hägg. Biol Age Predictors. EBioMedicine. 2017;21:29–36.
6. Ashiqur Rahman S, Giacobbi P, Pyles L, et al. Deep learning for biological age Estimation. Brief Bioinform. 2021;22:1767–81.
7. Gialluisi A, A Di Castelnuovo, M B Donati, et al. Machine learning approaches for the Estimation of biological aging: the road ahead for population studies. Front Med (Lausanne). 2019;6:146.
8. Pyrkov TV, Slipensky K, Barg M, et al. Extracting biological age from biomedical data via deep learning: too much of a good thing? Sci Rep. 2018;8:5210.
9. Zhong X, Lu Y, Gao Q, et al. Estimating biological age in the Singapore longitudinal aging study. J Gerontol Biol Sci Med Sci. 2020;75:1913–20.
10. Cao X, Yang G, Jin X, et al. A machine Learning-Based aging measure among Middle-Aged and older Chinese adults: the China health and retirement longitudinal study. Front Med (Lausanne). 2021;8:698851.
11. Lin H, Lunetta KL, Zhao Q, et al. Whole blood gene expression associated with clinical biological age. J Gerontol Biol Sci Med Sci. 2019;74:81–8.
12. Chen L, Zhang Y, Yu C, et al. Modeling biological age using blood biomarkers and physical measurements in Chinese adults. EBioMedicine. 2023;89:104458.
13. Putin E, Mamoshina P, Aliper A, et al. Deep biomarkers of human aging: application of deep neural networks to biomarker development. Aging. 2016;8:1021–33.
14. Libbrecht M W, W S Noble. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16:321–32.
15. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24:1565–7.
16. Huang DQ, El-Serag HB, R Loomba. Global epidemiology of NAFLD-related HCC: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol. 2021;18:223–38.
17. Man S, Deng Y, Ma Y, et al. Prevalence of liver steatosis and fibrosis in the general population and various High-Risk populations: A nationwide study with 5.7 million adults in China. Gastroenterology. 2023;165:1025–40.
18. Baboota RK, Rawshani A, Bonnet L, et al. BMP4 and Gremlin 1 regulate hepatic cell senescence during clinical progression of NAFLD/NASH. Nat Metab. 2022;4:1007–21.
19. Xia M, Li W, Lin H, et al. DNA methylation age acceleration contributes to the development and prediction of non-alcoholic fatty liver disease. Geroscience. 2024;46:3525–42.
20. Loomba R, Gindin Y, Jiang Z et al. DNA methylation signatures reflect aging in patients with nonalcoholic steatohepatitis. JCI Insight. 2018; 3.
21. Klemera P, S Doubal. A new approach to the concept and computation of biological age. Mech Ageing Dev. 2006;127:240–8.
22. Levine ME, Lu AT, Quach A, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging. 2018;10:573–91.
23. Cohen AA, Milot E, Yong J, et al. A novel statistical approach shows evidence for multi-system physiological dysregulation during aging. Mech Ageing Dev. 2013;134:110–7.
24. Graf GH, Crowe CL, Kothari M, et al. Testing Black-White disparities in biological aging among older adults in the united States: analysis of DNA-Methylation and Blood-Chemistry methods. Am J Epidemiol. 2022;191:613–25.
25. Bernard D, Doumard E, Ader I, et al. Explainable machine learning framework to predict personalized physiological aging. Aging Cell. 2023;22:e13872.
26. Fan JG, Wei L, Zhuang H. Guidelines of prevention and treatment of nonalcoholic fatty liver disease (2018, China). J Dig Dis. 2019;20:163–73.
27. Castera L, Friedrich-Rust M, Loomba R. Noninvasive assessment of liver disease in patients with nonalcoholic fatty liver disease. Gastroenterology. 2019;156:1264–e12814.
28. Siddiqui MS, Vuppalanchi R, Van Natta ML, et al. Vibration-Controlled transient elastography to assess fibrosis and steatosis in patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol. 2019;17:156–e1632.
29. Paik JM, Deshpande R, Golabi P, et al. The impact of modifiable risk factors on the long-term outcomes of non-alcoholic fatty liver disease. Aliment Pharmacol Ther. 2020;51:291–304.
30. Younossi ZM, Stepanova M, Negro F, et al. Nonalcoholic fatty liver disease in lean individuals in the united States. Med (Baltim). 2012;91:319–27.
31. Younossi ZM, Paik JM, Al Shabeeb R, et al. Are there outcome differences between NAFLD and metabolic-associated fatty liver disease? Hepatology. 2022;76:1423–37.
32. Tibshirani R. The Lasso method for variable selection in the Cox model. Stat Med. 1997;16:385–95.
33. Liu Z. Development and validation of 2 composite aging measures using routine clinical biomarkers in the Chinese population: analyses from 2 prospective cohort studies. J Gerontol Biol Sci Med Sci. 2021;76:1627–32.
34. Kwon D, DW Belsky. A toolkit for quantification of biological age from blood chemistry and organ function test data: bioage. Geroscience. 2021;43:2795–808.
35. Mamoshina P, Kochetov K, Putin E, et al. Population specific biomarkers of human aging: A big data study using South Korean, Canadian, and Eastern European patient populations. J Gerontol Biol Sci Med Sci. 2018;73:1482–90.
36. Liu Z, Chen X, Gill TM, et al. Associations of genetics, behaviors, and life course circumstances with a novel aging and healthspan measure: evidence from the health and retirement study. PLoS Med. 2019;16:e1002827.
37. DelaRosa O, Pawelec G, Peralbo E, et al. Immunological biomarkers of ageing in man: changes in both innate and adaptive immunity are associated with health and longevity. Biogerontology. 2006;7:471–81.
38. Abdellatif M, Rainer PP, Sedej S, et al. Hallmarks of cardiovascular ageing. Nat Rev Cardiol. 2023;20:754–77.
39. Hommos MS, Glassock RJ. A D rule. Structural and functional changes in human kidneys with healthy aging. J Am Soc Nephrol. 2017;28:2838–44.
40. Bloomer SA, ED Moyer. Hepatic macrophage accumulation with aging: cause for concern? Am J Physiol Gastrointest Liver Physiol. 2021;320:G496–505.

41. Ahadi S, Zhou W, SM Schüssler-Fiorenza Rose, et al. Personal aging markers and ageotypes revealed by deep longitudinal profiling. Nat Med. 2020;26:83–90.
42. Gao X, Geng T, Jiang M, et al. Accelerated biological aging and risk of depression and anxiety: evidence from 424,299 UK biobank participants. Nat Commun. 2023;14:2277.
43. Huang W, Zhang Z, Colucci M, et al. The mixed effect of Endocrine-Disrupting chemicals on biological age acceleration: unveiling the mechanism and potential intervention target. Environ Int. 2024;184:108447.
44. Shen X, Wang C, Zhou X et al. Nonlinear dynamics of multi-omics profiles during human aging. Nat Aging. 2024.
45. Wang H, Liu Z, Fan H, et al. Association between biological aging and the risk of mortality in individuals with non-alcoholic fatty liver disease: A prospective cohort study. Arch Gerontol Geriatr. 2024;124:105477.
46. Accelerated Aging May Increase the Risk of Early-onset Cancers in Younger Generations. Retrieved from https://www.aacr.org/about-the-aacr/newsroom/news-releases/accelerated-aging-may-increase-the-risk-of-early-onset-cancers-in-younger-generations/
47. Fitzgerald KN, Campbell T, Makarem S, et al. Potential reversal of biological age in women following an 8-week methylation-supportive diet and lifestyle program: a case series. Aging. 2023;15:1833–9.
48. Poganik JR, Zhang B, Baht GS, et al. Biological age is increased by stress and restored upon recovery. Cell Metab. 2023;35:807–e8205.

Publisher's note