Development and validation of machine learning models for MASLD: based on multiple potential screening indicators

Hao Chen[1†], Jingjing Zhang[1†], Xueqin Chen[1†], Ling Luo[1], Wenjiao Dong[1], Yongjie Wang[1], Jiyu Zhou[1], Canjin Chen[1], Wenhao Wang[1], Wenbin Zhang[1], Zhiyi Zhang[1], Yongguang Cai[2*], Danli Kong[1*] and Yuanlin Ding[1*]

[1]Department of Epidemiology and Medical Statistics School of Public Health, Guangdong Medical University, Dongguan, Guangdong, China, [2]Department of Medical Oncology, Central Hospital of Guangdong Nongken, Zhanjiang, Guangdong, China

Background: Multifaceted factors play a crucial role in the prevention and treatment of metabolic dysfunction-associated steatotic liver disease (MASLD). This study aimed to utilize multifaceted indicators to construct MASLD risk prediction machine learning models and explore the core factors within these models.

Methods: MASLD risk prediction models were constructed based on seven machine learning algorithms using all variables, insulin-related variables, demographic characteristics variables, and other indicators, respectively. Subsequently, the partial dependence plot(PDP) method and SHapley Additive exPlanations (SHAP) were utilized to explain the roles of important variables in the model to filter out the optimal indicators for constructing the MASLD risk model.

Results: Ranking the feature importance of the Random Forest (RF) model and eXtreme Gradient Boosting (XGBoost) model constructed using all variables found that both homeostasis model assessment of insulin resistance (HOMA-IR) and triglyceride glucose-waist circumference (TyG-WC) were the first and second most important variables. The MASLD risk prediction model constructed using the variables with top 10 importance was superior to the previous model. The PDP and SHAP methods were further utilized to screen the best indicators (including HOMA-IR, TyG-WC, age, aspartate aminotransferase (AST), and ethnicity) for constructing the model, and the mean area under the curve value of the models was 0.960.

Conclusions: HOMA-IR and TyG-WC are core factors in predicting MASLD risk. Ultimately, our study constructed the optimal MASLD risk prediction model using HOMA-IR, TyG-WC, age, AST, and ethnicity.

KEYWORDS

metabolic dysfunction-associated steatotic liver disease, machine learning, insulin resistance, triglyceride glucose, risk prediction model

Introduction

In June 2023, the international consensus group introduced the term " Steatotic Liver Disease (SLD)" as an inclusive term covering all different etiologies of hepatic steatosis, including metabolic, alcohol-related, drug-induced, and cryptogenic causes (1). Metabolic dysfunction-associated steatotic liver disease (MASLD), previously known as Non-alcoholic fatty liver disease (NAFLD), is one of the most common chronic liver diseases worldwide, affecting approximately 30% of the global population (2). MASLD typically progresses over time, potentially leading to hepatic inflammation (metabolic dysfunction-associated steatohepatitis, MASH), liver fibrosis, and ultimately, the development of cirrhosis or even hepatocellular carcinoma (3). With the continued increase in obesity and diabetes mellitus (DM), the prevalence of NAFLD and associated healthcare costs are expected to rise, significantly impacting global public health. MASLD is also considered a hepatic manifestation of metabolic syndrome, as it is closely associated with metabolic disorders such as obesity, dyslipidemia, and DM. Early screening and effective intervention measures help reduce and delay the occurrence of adverse prognostic events associated with MASLD. Liver biopsy has long been considered the gold standard for histological assessment, diagnosis, and prognosis determination of liver fibrosis (4). However, its invasive nature, potential risk of bleeding, and sampling errors due to uneven distribution of liver parenchymal lesions make it difficult to be widely used in clinical practice (5), resulting in a large number of MASLD patients missing the optimal timing for diagnosis and treatment. Therefore, exploring accurate and non-invasive biomarkers for the diagnosis of MASLD is crucial to reduce the need for invasive liver biopsy and identify patients at high risk of liver and metabolic complications early on.

Machine learning (ML) is a branch of artificial intelligence with the capability to handle large, complex, and entirely different

datasets, creating complex analytical models based on learning frameworks, thereby improving and optimizing prediction accuracy (6). Therefore, an increasing number of researchers are beginning to develop disease risk prediction models through ML (7–11). Numerous studies have utilized clinical data and machine learning algorithms to construct NAFLD risk prediction models. Zhou et al. (11) developed a NAFLD risk prediction model based on obese children, which demonstrated good clinical discriminative ability, with an area under the curve (AUC) value of 0.821 for the receiver operating characteristic (ROC) curve. Liu et al. (12) constructed a NAFLD risk prediction model based on a population undergoing health checkups, using machine learning algorithms. eXtreme Gradient Boosting (XGBoost) demonstrated excellent clinical predictive value, with an AUC value of 0.926 for the ROC curve. Huang et al. (13) developed a risk prediction model for NAFLD within a population based on prospective cohort studies, utilizing various machine learning algorithms. The best model, Categorical Boost (CatBoost) achieved a predictive performance of AUC = 0.810.
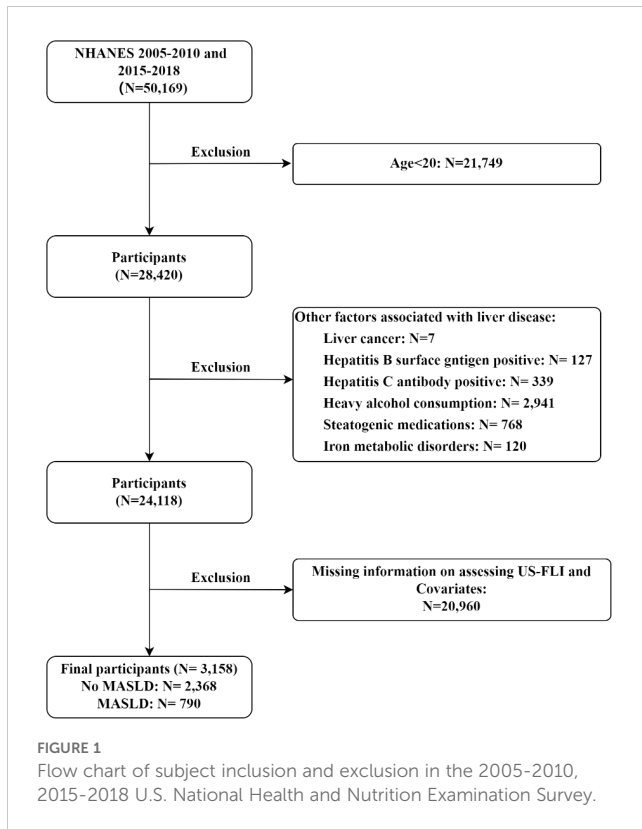
However, available MASLD risk prediction models are scarce currently, and previous studies have seldom explored the roles of various indicators in NAFLD risk prediction models, making it difficult to assess which indicators play a core role in predicting MASLD risk. Therefore, in this study, we aimed to utilize multifaceted indicators from the National Health and Nutrition Examination Survey (NHANES) data from 2005-2010 and 2015-2018 to construct MASLD risk prediction machine learning models and explore the core factors within the models.

Materials and methods

Study design and population

Study data is from NHANES and includes 5 cycles from 2005-2010 and 2015-2018. The exclusion criteria adopted in this study were as follows (1) participants younger than 20 years of age; (2) participants with liver disease caused by other factors, including 1) iron metabolic disorders, indicated by ferritin concentration exceeding 200ug/L; 2) alcohol-related liver disease, characterized by heavy drinking (≥3 drinks per day for females and ≥4 drinks per day for males) or binge drinking (≥5 drinks on a single occasion); 3) hepatitis virus infection, identified by the presence of hepatitis B surface antigen or hepatitis C confirmation antibody; 4) self-reported liver cancer; 5) taking steatogenic medications for at least 6 months(including amiodarone, methotrexate, tamoxifen, aspirin, ibuprofen, nucleoside reverse transcriptase inhibitors, protease inhibitors, valproic acid, carbamazepine, glucocorticoids); (3) participants who lacked the information used to assess MASLD; and (4) participants who were missing other covariates. The detailed flowchart is shown in Figure 1. 3,158 participants were finally included, consisting of 2,368 non-MASLD patients and 790 MASLD patients.

FIGURE 1

Flow chart of subject inclusion and exclusion in the 2005-2010, 2015-2018 U.S. National Health and Nutrition Examination Survey.

Definition of MASLD

The diagnosis of MAFLD typically relies on techniques such as abdominal ultrasonography, magnetic resonance imaging, and other imaging modalities aimed at identifying liver fat accumulation. Further confirmation may necessitate a liver biopsy. However, the latter is not commonly employed due to its high operator dependence, cost considerations, and the requirement for significant steatosis, typically exceeding 20-30% of liver cells, for detection. Consequently, alternative approaches have been developed to address these limitations. One such method is the United States Fatty Liver Index (US-FLI), pioneered by CERuhl, designed specifically for assessing fatty liver disease within the U.S. population (14). The specific calculation formula for the United States fatty liver index (US-FLI) was as follows:

$$US-FLI = \frac{e^{-0.8073 * non-Hispanic\ black + 0.3458 * Mexican\ American + 0.0093 * age + 0.6151 * \log_e(GGT) + 0.0249 * waist\ circumference + 1.1792 * \log_e(insulin) + 0.8242 * \log_e(glu\cos e) - 14.7812}}{1 + e^{-0.8073 * non-Hispanic\ black + 0.3458 * Mexican\ American + 0.0093 * age + 0.6151 * \log_e(GGT) + 0.0249 * waist\ circumference + 1.1792 * \log_e(insulin) + 0.8242 * \log_e(glu\cos e) - 14.7812}} * 100$$

In the exclusion of other liver diseases associated with the aforementioned factors, when US-FLI≥30, we considered the participant to have MASLD.

Study covariates

In this study, we considered several covariates that could potentially confound the outcomes: 1) Demographic variables: age, gender, ethnicity, education level, military status, marital status, sleep status, smoke status, the family income poverty ratio (PIR), physical activity (PA); 2) Examination variables: waist circumference (WC), waist to height ratio (WtHR), and body mass index (BMI); 3) Laboratory variables: low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), total cholesterol (TC), triglyceride (TG), alanine transaminase (ALT), aspartate aminotransferase (AST), bilirubin, albumin, fasting blood glucose (FBG), triglyceride glucose (TyG) index–related parameters, including the TyG, TyG–BMI, TyG-WtHR, and TyG-WC, Homeostasis Model Assessment of Insulin Resistance (HOMA-IR); 4) Medical history: diabetes mellitus (DM), hypertension, cardiovascular disease (CVD); 5) Other indexes: dietary inflammatory index (DII), systemic immune-inflammation index (SII), oxidative balance score (OBS), composite dietary antioxidant index (CDAI), visceral adiposity index (VAI), lipid accumulation product (LAP).

Statistical analyses

In this study, data were analyzed and visualized using R 4.0.1 software. Continuous variables were described using mean ± standard deviation (SD) or median (IQR), while categorical variables were expressed as percentages. The Shapiro-Wilk test was employed to assess the normality of continuous variables, and the independent sample t-test or Mann-Whitney test was used for between-group comparisons of continuous variables. Between-group comparisons of categorical variables were conducted using the chi-square test or Fisher's exact test.

Construction of the machine learning model

Research data was split between 70% as training sets (n = 2,211) and 30% as testing sets (n =947).

In this study, seven different machine learning algorithms, namely Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbor (KNN), eXtreme Gradient Boosting (XGBoost), Naïve Bayes Model (NBM), Backpropagation Neural Network (BPNN), and Logistic Regression (LG), were used to evaluate the predictive effectiveness of various variable characteristics on MASLD, each with its own features. RF can handle high-dimensional data and has a strong resilience to noise. It is robust against overfitting when hyperparameters are optimized, but has high time complexity with large datasets (15). SVM is effective for both linear and non-linear classification tasks and performs well on high-dimensional datasets. It is less affected by outliers, but its time complexity also can be high when working with large sample sizes (16, 17). KNN is simple to implement and highly accurate, making it ideal for data without strict assumptions. It handles outliers well

but can be computationally expensive and slow when the dataset is large ([18], [19]). Known for its high accuracy and efficiency, XGBoost is well-suited for complex, high-dimensional datasets. It includes options for regularization to prevent overfitting, although it can be computationally intensive if not properly tuned ([20]). NBM is efficient and works well with small datasets, especially when the assumption of feature independence is approximately met. However, it may perform poorly with data where features are heavily correlated or when feature distribution assumptions are violated ([21], [22]). BPNN is powerful for capturing complex patterns in large datasets and is capable of handling non-linear relationships. However, it requires significant computational resources and training time, and it is prone to overfitting without proper regularization or dropout methods ([23–25]). LR is interpretable and effective for binary classification tasks with linear relationships. It has low time complexity, making it efficient for large datasets, but it may underperform with non-linear or complex patterns ([26], [27]). To evaluate the predictive effect of various types of variables on MASLD, we constructed models utilizing 1) all variables; 2) insulin-related indexes (HOMA-IR, TyG index-related parameters); 3) demographic characteristics variables; and 4) other indexes, respectively. Subsequently, to develop a more accurate and parsimonious risk prediction model, we selected the feature variables from the best-performing model among the four evaluated models, ranking them by importance. The top 10 variables were chosen based on their importance rankings and used to construct a refined MASLD risk prediction model. To avoid overfitting the models, we performed hyperparameter optimization for each model. In addition, considering the robustness and generalizability of the models, we integrated multiple evaluation metrics of seven machine learning algorithms and performed 10-fold cross-validation for each model.

Interpretable methods pipeline of prediction models

We applied interpretability techniques, specifically Partial Dependence Plots (PDPs) and SHapley Additive exPlanations (SHAP), to better understand each variable's contribution to MASLD risk prediction. PDPs were used to illustrate the marginal effect of each individual variable on MASLD risk by showing how the predicted risk changes across a range of values for each variable while keeping other features constant. This helps isolate the impact of each feature on the outcome and provides a clearer picture of its direction and strength in influencing MASLD risk.

In addition, SHAP was employed to quantify and visualize the contributions of each variable across different predictions, offering insights into feature importance and interaction effects. SHAP values reveal how much each variable pushes the prediction toward or away from higher MASLD risk in individual cases. By integrating insights from both PDP and SHAP analyses, we identified the most influential features and refined our model to retain only those with substantive predictive power, thereby constructing an optimized final model.

Evaluation of machine learning model

The performance of the model was assessed by various evaluation metrics including receiver operating characteristic curve (ROC), area under the receiver operator curve (AUC), accuracy, sensitivity/recall, specificity, false positive rate (FPR), false negative rate (FNR), positive predictive value (PPV), negative predictive value (NPV) and the F_1 score. The AUC was mainly used as an evaluation metric for performance comparison between models. The variance inflation factor (VIF) was used to evaluate multicollinearity in the multivariate analysis based on the intercorrelation between variables.

Results

Characteristics of the study population

Table 1 displays the general characteristics of the study participants. A total of 3,158 U.S. adults were included in the study, with a greater proportion of MASLD patients being male, non-Hispanic whites. Hypertension, DM, and CVD may all be risk factors for MASLD. In addition, compared with participants without MASLD, participants with MASLD have higher levels of HOMA-IR, TyG index-related parameters, DII, SII, VAI, and LAP index relative to participants without MASLD.

Evaluation and comparison of the predictive models

Figure 2 depicts the ROC curves for each of the seven models constructed using different variables. The results of other evaluation metrics of the models can be viewed in Supplementary Tables S1-S4. In the test set, the overall effect of the models constructed using insulin-related indexes was slightly better than that of the models constructed using all variables, but the effect of the overall ROC curve of the seven models constructed using all variables (average AUC: 0.942) was slightly higher than that of the models constructed using only insulin-related indexes (average AUC: 0.941). In addition, in the analysis of multicollinearity between the two models, it was found that there was multicollinearity between several variables in both models (VIF>10).

Predictive models constructed with top 10 variables

To further explore the optimal solution for constructing the MASLD risk prediction model, we selected the two predictive models (RF and XGBoost) with the highest AUC values among the seven models constructed with all variables and filtered out the top 10 importance variables in the two predictive models. Figure 3 illustrates the top 10 variables of importance for the two models. Focusing on the top 10 variables, which contribute most

TABLE 1 Baseline characteristics of participants.

	No MASLD (N=2,368)	MASLD (N=790)	P
Age (mean ± SD)	48.79 ± 17.13	55.41 ± 15.50	<0.001
PIR (mean ± SD)	2.94 ± 1.63	2.75 ± 1.64	<0.05
TC (mean ± SD)	5.08 ± 1.03	5.08 ± 1.06	0.995
HDL (mean ± SD)	1.49 ± 0.41	1.20 ± 0.30	<0.001
LDL (mean ± SD)	3.04 ± 0.91	3.06 ± 0.94	0.651
TG (median, IQR)	1.05 (0.73)	1.61 (1.11)	<0.001
BMI (mean ± SD)	26.73 ± 5.26	33.33 ± 5.94	<0.001
WC (mean ± SD)	93.06 ± 12.96	111.66 ± 13.20	<0.001
FBG (median, IQR)	5.38 (0.72)	6.00 (1.22)	<0.001
ALT (median, IQR)	20.00 (10.00)	26.00 (14.00)	<0.001
AST (median, IQR)	23.00 (7.00)	24.00 (8.00)	<0.001
Bilirubin (mean ± SD)	13.77 ± 5.19	13.76 ± 5.18	0.951
Albumin (mean ± SD)	42.45 ± 3.07	41.88 ± 2.92	<0.001
PA (median, IQR)	1,280.00 (3300.00)	1250.00 (2980.50)	0.173
TyG (mean ± SD)	8.45 ± 0.55	9.03 ± 0.52	<0.001
HOMA-IR (median, IQR)	1.78 (1.41)	5.05 (3.61)	<0.001
WtHR (mean ± SD)	0.55 ± 0.08	0.66 ± 0.08	<0.001
TyG-BMI (mean ± SD)	226.59 ± 49.97	300.77 ± 55.96	<0.001
TyG-WC (mean ± SD)	788.83 ± 135.39	1,008.07 ± 134.32	<0.001
TyG-WtHR (mean ± SD)	4.69 ± 0.81	5.95 ± 0.79	<0.001
DII (mean ± SD)	1.33 ± 1.90	1.51 ± 1.80	<0.05
SII (mean ± SD)	519.20 ± 307.72	561.82 ± 295.45	0.001
OBS (mean ± SD)	20.08 ± 8.60	18.84 ± 8.06	<0.001
CDAI (mean ± SD)	0.82 ± 4.08	0.52 ± 3.76	0.061
VAI (median, IQR)	1.16 (1.11)	2.25 (1.93)	<0.001
LAP (median, IQR)	32.58 (18.50)	77.42 (54.69)	<0.001
Gender (N, %)			<0.001
Female	1,222 (51.6)	287 (36.3)	
Male	1,146 (48.4)	503 (63.7)	
Ethnicity (N, %)			<0.001
Non-Hispanic White	1,233 (52.1)	436 (55.2)	
Non-Hispanic Blacks	534 (22.6)	86 (10.9)	
Mexican	270 (11.4)	166 (21.0)	
Other	331 (14.0)	102 (12.9)	
Education (N, %)			<0.001
Elementary and secondary education	429 (18.1)	202 (25.6)	

(Continued)

TABLE 1 Continued

	No MASLD (N=2,368)	MASLD (N=790)	P
High school	539 (22.8)	193 (24.4)	
Bachelor degree or higher	1,400 (59.1)	395 (49.9)	
Military (N, %)			<0.001
No	2,073 (87.5)	632 (80.0)	
Yes	295 (12.5)	158 (20.0)	
Marital (N, %)			<0.001
Married or Living with partner	1,508 (63.7)	552 (69.9)	
Widowed, Divorced, or Separated	476 (20.1)	170 (21.5)	
Never married	384 (16.2)	68 (8.6)	
Sleep status (N, %)			0.473
Short	898 (37.9)	311 (39.4)	
Normal	1,305 (55.1)	433 (54.7)	
Long	165 (7.0)	46 (5.8)	
Smoke status (N, %)			<0.001
Never	1,385 (58.5)	421 (53.3)	
Former	583 (24.6)	267 (33.8)	
Now	400 (16.9)	102 (12.9)	
Hypertension (N, %)			<0.001
No	1,569 (66.3)	321 (40.6)	
Yes	799 (33.7)	469 (59.4)	
DM (N, %)			<0.001
No	2,087 (88.1)	530 (67.1)	
Yes	281 (11.9)	260 (32.9)	
CVD (N, %)			<0.001
No	2,208 (93.2)	671 (84.9)	
Yes	160 (6.8)	119 (15.1)	

significantly to model output, allows us to provide clearer insights into the key factors influencing MASLD risk without overwhelming the interpretation with too many variables.

Therefore, we constructed disease prediction models for MASLD using the top10 variables in importance in RF and XGBoost, respectively (Hereafter referred to as RF top10 models and XGBoost top10 models). It was found that the mean AUC values of the RF top 10 models and the XGBoost top 10 models did not differ much (slightly higher in the XGBoost top 10 models), but both were higher than the model constructed using only the insulin-related indexes. Figure 4 illustrates the results of the ROC curves for the models, and the results for the other evaluation metrics are shown in Supplementary Tables S5, S6.

Interpretation of ML models

Since the difference in the overall predictive performance between the RF top 10 models and the XGBoost top 10 models is very small, to further explore the models, we chose to generate PDPs for the Random Forest model selected from both the RF top 10 models and the XGBoost top 10 models to interpret the predictive models. As presented in Figures 5, 6, the application of PDPs allowed for a broader interpretation of model performance, which displayed the relationship between the features and MASLD. The PDP analysis of the RF top 10 models indicates that increases in nine continuous variables included in the model are associated with elevated risk predictions for MASLD. Specifically, HOMA-IR, TyG-WC, and TyG-WtHR exhibit the most significant effects. More specifically, as the levels of HOMA-IR range from approximately 0 to 10, TyG-WC range from approximately 818 to 1091, and TyG-WtHR range from approximately 5 to 6, the risk prediction values for MASLD show an increasing trend. The PDP analysis results of

the XGBoost top 10 models are generally consistent with those of the RF top 10 models. It is noteworthy that, unlike HOMA-IR and TyG-WC, which only impact the risk prediction values for MASLD within specific ranges, the risk prediction values for MASLD increase with age across the entire age range.

PDPs can also provide insights into the interaction performance of a model. Supplementary Figures S1, 3 show the overall interaction strength between variables in the two models, respectively. Supplementary Figures S2, 4 display the synergistic effects of variables' levels on MASLD risk. The results still indicate that HOMA-IR, TyG-WC, and TyG-WtHR dominate in predicting MASLD in the RF top 10 models; In the XGBoost top 10 models, the predicted values of MASLD are mainly determined by HOMA-IR, TyG-WC, and age.

Due to the limitations of PDPs, we could not assess the role of ethnicity in the model. Meanwhile, we observed that variables such as ALT and AST seem to play a role in predicting MASLD risk. To further explore the optimal model configuration, we conducted a

FIGURE 2
Receiver Operating Characteristic (ROC) curves for seven machine learning models are constructed using different variables. **(A)** all variables; **(B)** insulin-related indexes; **(C)** demographic characteristics variables; **(D)** other indexes.

FIGURE 3
The contribution of the top 10 variables in predictive models. **(A)** The contribution of the top 10 variables in the RF predictive model; **(B)** The contribution of the top 10 variables in the XGBoost predictive model.

SHAP analysis on the two models mentioned above (Supplementary Figures S5, S6). The results from the SHAP dependency plots indicate that SHAP values vary across different ethnic groups, suggesting that ethnicity might be a potential risk factor for MASLD and that the effect of TyG-WC may vary across ethnicities. From the color mapping in the SHAP plots, it can be observed that AST is not influenced by HOMA-IR in the model, while ALT is significantly influenced by TyG-WC. In the HOMA-IR plot, where color mapping represents Age, data points with lighter colors (representing older age) are distributed in regions with higher HOMA-IR values and SHAP values, indicating that HOMA-IR may have a greater impact on MASLD risk in older populations.

Construction and evaluation of the optimal MASLD prediction model

Taking into account the effects of the variables in both the PDPs and SHAP results on MASLD risk prediction performance as well as interactions and covariances, we further screened the variables used to construct the model. Eventually, we utilized 5 factors (including HOMA-IR, TyG-WC age, AST, and ethnicity) to construct the optimal MASLD risk prediction model (Figure 7, average AUC=0.960). We also calculated the VIF of all the variables in this model and proved that there is no multicollinearity between them (VIF<10).

FIGURE 4
ROC curves for seven machine learning models are constructed using the top 10 variables of importance. **(A)** model uses the top 10 variables of importance in the RF model; **(B)** model uses the top 10 variables of importance in the XGBoost model.

FIGURE 5
Partial dependence plots (PDPs) of the RF top 10 model **(A)** HOMA-IR' PDP of the RF top 10 model; **(B)** TyG-WC' PDP of the RF top 10 model; **(C)** TyG-WtHR' PDP of the RF top 10 model; **(D)** WC' PDP of the RF top 10 model; **(E)** WtHR' PDP of the RF top 10 model; **(F)** LAP' PDP of the RF top 10 model; **(G)** TyG-BMI' PDP of the RF top 10 model; **(H)** BMI' PDP of the RF top 10 model; **(I)** FBG' PDP of the RF top 10 model.

Discussion

MASLD, the most common chronic liver disease worldwide, is mediated by various factors including genetic susceptibility, dietary habits, obesity, insulin resistance, and the endocrine effects of many diseases. Consequently, extensive research is underway to explore non-invasive, practical, and reliable disease prediction models to identify and manage individuals at high risk of MASLD, ultimately alleviating the disease burden. In this large cross-sectional study, our primary aim is to investigate how to construct the optimal MASLD prediction model and the role played b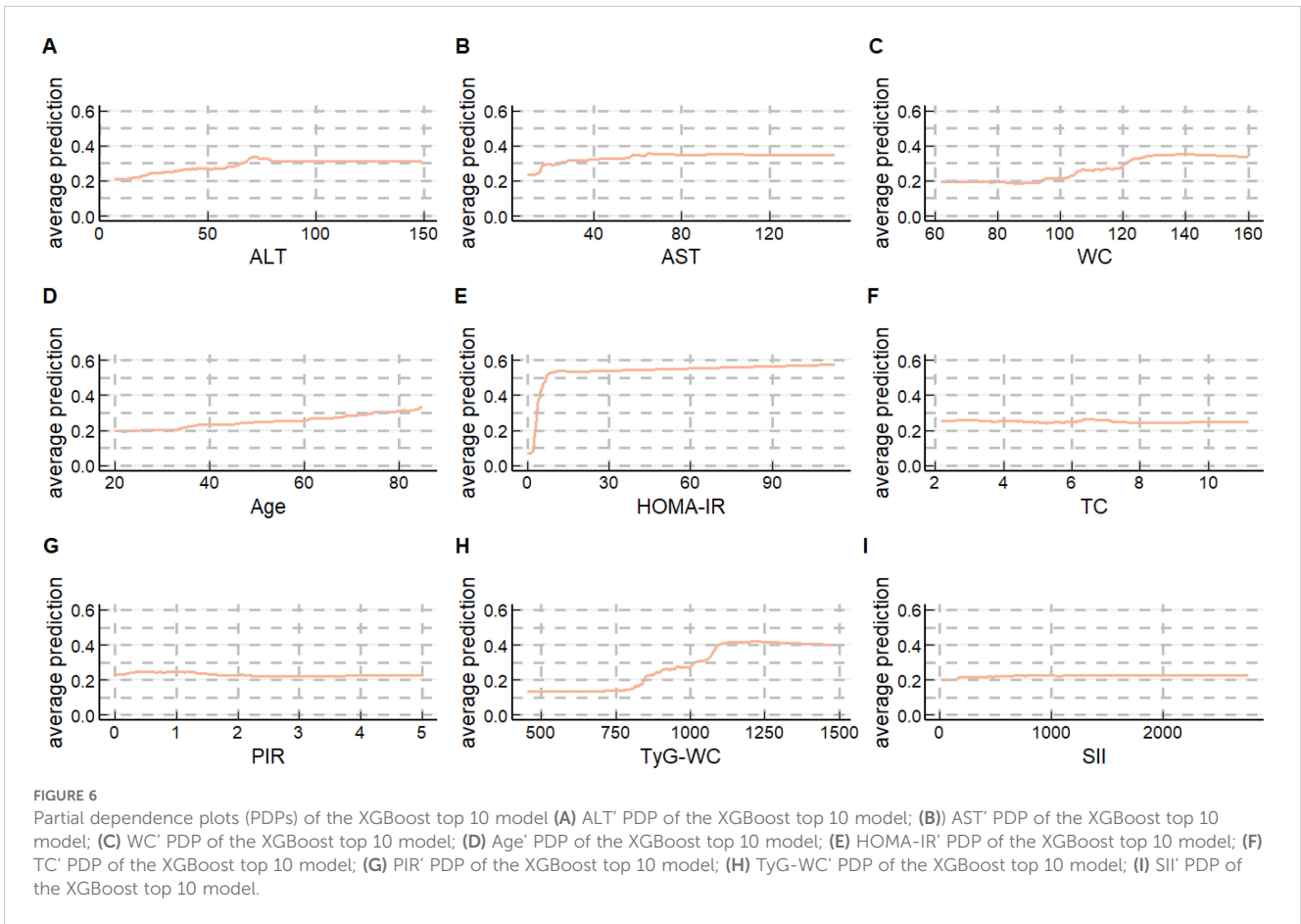y indicators in the model. Despite NAFLD being eventually renamed as MASLD, studies have shown excellent consistency between the definitions of NAFLD and MASLD, with approximately 99% of NAFLD patients meeting MASLD criteria (28). Therefore, while this study focuses on MASLD, it also discusses NAFLD and MAFLD simultaneously.

Previous studies have already indicated that obesity is an independent risk factor for NAFLD (29). The new nomenclature and definition more intuitively reflect that metabolism (obesity and DM) is a key etiology of fatty liver disease. Our study results also found that the average BMI and HOMA-IR of MASLD patients were 33.33 kg/m2 and 5.06, respectively. Additionally, the proportion of participants with diabetes mellitus (DM) was much

higher in MASLD patients compared to those without MASLD. Furthermore, our study found that MASLD patients had higher SII and DII compared to participants without MASLD, consistent with the findings of Yan et al. (30–33).

Inflammation plays a crucial role in the progression of MASLD to MASH, liver fibrosis, and even liver cancer (34). In addition to the direct effects of immune-inflammatory factors, these cytokines also promote the development of IR and T2DM through activation of intracellular pathways, thereby influencing MASLD (35). Studies have also found that higher levels of OBS have a protective effect on MASLD, and Tan et al. (36) suggested that OBS may influence MASLD by reducing insulin resistance levels. Previous studies consistently demonstrate a negative correlation between PA and NAFLD, suggesting that even mild exercise can prevent and treat NAFLD to some extent (37, 38). However, this study did not observe an association between PA and MASLD, speculating that most MASLD patients in the early stages of liver disease prefer to improve their disease burden by adjusting dietary habits. Additionally, MASLD patients are mostly obese and may be less inclined to choose increased physical activity as a means to improve their condition. VAI and LAP are also positively correlated with MASLD, consistent with the findings of previous studies by Vural et al (39, 40). Additionally, Peng et al. (41) suggested that LAP may

FIGURE 6
Partial dependence plots (PDPs) of the XGBoost top 10 model **(A)** ALT' PDP of the XGBoost top 10 model; **(B)**) AST' PDP of the XGBoost top 10 model; **(C)** WC' PDP of the XGBoost top 10 model; **(D)** Age' PDP of the XGBoost top 10 model; **(E)** HOMA-IR' PDP of the XGBoost top 10 model; **(F)** TC' PDP of the XGBoost top 10 model; **(G)** PIR' PDP of the XGBoost top 10 model; **(H)** TyG-WC' PDP of the XGBoost top 10 model; **(I)** SII' PDP of the XGBoost top 10 model.

have better predictive performance than VAI, a viewpoint supported by our results. However, studies by Boden et al. (42–44) suggest that increased visceral adipose tissue may be achieved through pathways such as oxidative stress, inflammation, and IR. Despite the multifactorial etiology of MASLD, IR, and obesity remain core driving factors in its development. Therefore, for research aiming to predict and assess MASLD risk, it is imperative to delve deeper into these core driving factors and explore their role in MASLD risk assessment through model construction.

HOMA-IR, as a commonly used index for measuring IR, has been widely applied. However, its requirement for fasting insulin

FIGURE 7
ROC curves for seven machine learning models are constructed using HOMA-IR, TyG-WC, age, AST, and ethnicity.
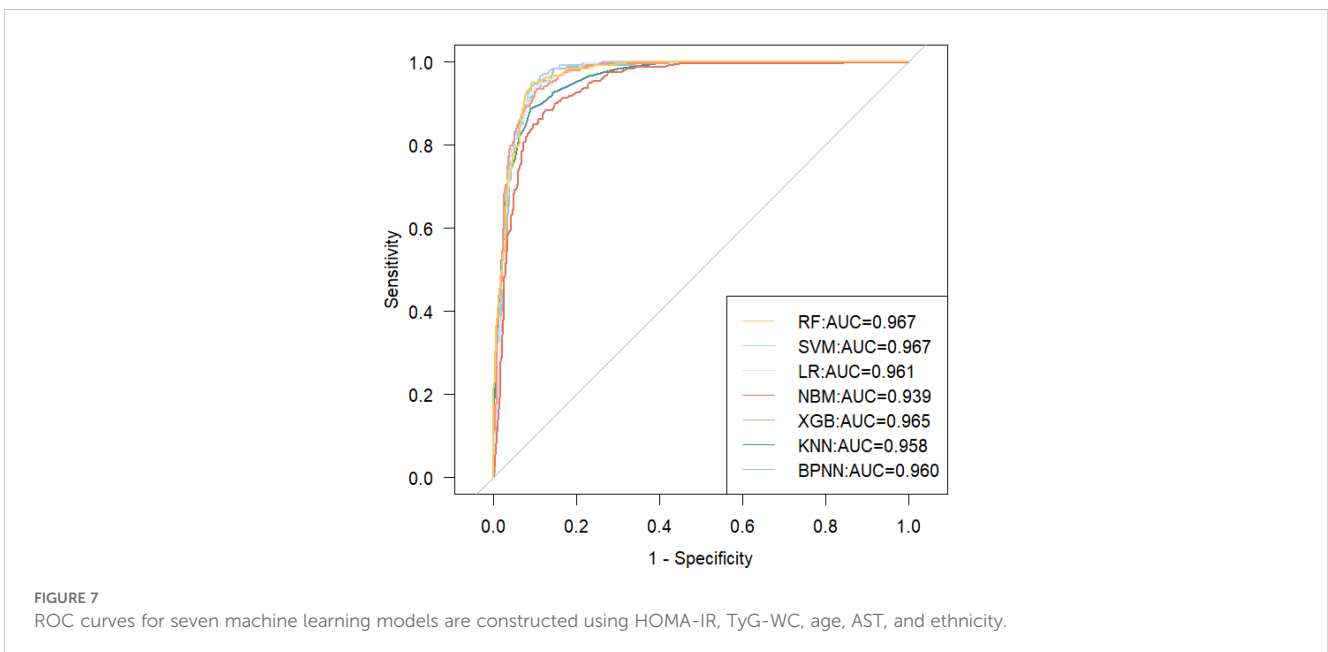
levels to some extent limits its practicality in clinical settings. In contrast, the TyG index addresses this limitation, becoming a simple, reproducible, and reliable indicator for assessing IR. Previous studies have demonstrated that TyG-related indices have good predictive performance for MAFLD (45). Xue et al. (46) used logistic regression to explore the association between insulin-related indexes and MAFLD, evaluating the predictive performance of individual indicators on MAFLD. They found that TyG-WC had the best predictive performance for MAFLD (AUC=0.832). Similar results were obtained by Peng et al. (41). Therefore, this study integrated insulin-related indexes to construct models for predicting MASLD risk. The performance of the LR model (AUC=0.957) was far superior to the models constructed by Peng (41) and Xue (46) et al. using single indicators. This significant performance difference emphasizes the predictive capability of the composite index model. To further obtain indicators that more comprehensively reflect MASLD risk and construct the optimal MASLD risk prediction model, we selected the top 10 variables of importance in the RF model and XGBoost model constructed using all variables. We found that HOMA-IR and TyG-WC were ranked first and second, respectively, fully reflecting the importance of insulin resistance in MASLD. Additionally, ethnicity and WC also appeared in the top 10 of importance in both models. The genotypes and lifestyle habits of different ethnic groups may be major factors influencing MASLD. WC, as an important indicator of central obesity, can also reflect the risk of MASLD to some extent. However, its importance is much lower than that of TyG-WC, possibly because TyG-WC integrates indicators of IR and central obesity, making it more prominent in predicting MASLD risk.

We further constructed MASLD risk prediction models using the top 10 variables of importance separately (the RF top 10 models and the XGBoost top 10 models mentioned in the results section) and found that the overall predictive performance of these models was higher than those constructed using only insulin-related indexes. Subsequently, we used the PDP method and SHAP method to interpret the RF top 10 models and the XGBoost top 10 models separately, further confirming the central role of HOMA-IR and TyG-WC in predicting MASLD risk. It is worth noting that in the XGBoost top 10 models, age consistently influences the occurrence and development of MASLD across the entire range.

Obviously, with increasing age, the body's metabolism, liver function, and fat metabolism abilities weaken, particularly the decreased ability of liver cells to metabolize fat, leading to fat accumulation in the liver and thereby increasing the risk of MASLD. Additionally, as age increases, the incidence of chronic diseases such as hypertension, DM, and CVD rises. These chronic diseases have a mutually influencing relationship with MASLD, and they may worsen due to the influence of MASLD, further exacerbating the disease burden of MASLD, and forming a vicious cycle (47, 48). Furthermore, AST can reflect the degree of liver cell damage, liver inflammation, and fibrosis (49). Therefore, it plays a supplementary role in predicting MASLD risk, thereby optimizing the effectiveness of MASLD risk prediction models. Although the variables included in the two types of models differ significantly, and there are also differences in the ranking of importance, the overall predictive performance of the RF top 10 models and the XGBoost top 10 models does not differ significantly. We believe the main reason is that HOMA-IR and TyG-WC play a primary role in predicting MASLD risk, while other indicators further complement aspects not captured by these two indicators. Considering that TyG-related parameters simultaneously enter the model, there may be multicollinearity issues, leading to the model repetitively capturing the biological information or metabolic status reflected by TyG-related indicators and neglecting other potential indexes. The study also found that MASLD incidence risk varies significantly across different ethnicities, which may be attributed to differences in metabolic characteristics between ethnic groups, particularly in insulin resistance, lipid metabolism, and fat distribution. Mexican Americans tend to exhibit higher levels of insulin resistance and abdominal obesity (50), while Non-Hispanic Blacks individuals typically have a higher proportion of subcutaneous fat, with relatively less visceral fat, which may confer some protection to liver health (51). Additionally, factors such as diet and lifestyle, as well as socioeconomic status, may also contribute to the differences in MASLD incidence risk among ethnicities.

Based on the PDP and SHAP results of the RF top 10 models and the XGBoost top 10 models, we further screened the variables and constructed the optimal MASLD risk prediction model using HOMA-IR, TyG-WC, age, AST, and ethnicity. This model not only has the largest average AUC value, but also other evaluation metrics are overall better than the previously constructed model. Therefore, we believe that although insulin resistance and obesity (especially central obesity) are core factors in developing MASLD, considering other factors that reflect liver function simultaneously helps improve the effectiveness of constructing clinical prediction models for MASLD risk.

Although the predictive performance of the final models constructed in this study is superior to that of previous research, there are still some limitations. Firstly, this study relied solely on the NHANES database, which includes data from the American population, without incorporating external datasets to further validate and optimize the model's performance. Moreover, the cross-sectional nature of the NHANES data limits the ability to establish causal relationships between variables and MASLD, which should be acknowledged as a constraint on causal inference in this study. In addition, there may be other factors affecting the prediction of MASLD risk that were not considered in this study or were not collected in the NHANES database, such as indicators of fibrosis, other indicators of inflammation, and so on, which may play a role in the diagnosis and prediction of MASLD.

Conclusions

Our study found that HOMA-IR and TyG-WC are core factors in predicting MASLD risk. However, integrating multiple factors can further enhance the model's predictive performance. Ultimately, our study constructed the optimal MASLD risk prediction model using HOMA-IR, TyG-WC, age, AST, and ethnicity.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

HC: Formal analysis, Visualization, Writing – original draft. JjZ: Formal analysis, Visualization, Writing – original draft. XC: Formal analysis, Writing – original draft. LL: Software, Writing – original draft. WD: Data curation, Validation, Writing – original draft. YW: Methodology, Writing – original draft. JyZ: Software, Writing – original draft. CC: Software, Writing – original draft. WW: Validation, Writing – original draft. WZ: Software, Writing – original draft. ZZ: Writing – original draft. YC: Supervision, Validation, Writing – review & editing. DK: Supervision, Validation, Writing – review & editing. YD: Software, Supervision, Validation, Writing – review & editing.

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2024.1449064/full#supplementary-material

References

1. Rinella ME, Lazarus JV, Ratziu V, Francque SM, Sanyal AJ, Kanwal F, et al. A multisociety Delphi consensus statement on new fatty liver disease nomenclature. *J Hepatol*. (2023) 79:1542–56. doi: 10.1016/j.jhep.2023.06.003

2. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*. (2016) 64:73–84. doi: 10.1002/hep.28431

3. Lekakis V, Papatheodoridis GV. Natural history of metabolic dysfunction-associated steatotic liver disease. *Eur J Intern Med*. (2024) 122:3–10. doi: 10.1016/j.ejim.2023.11.005

4. Li W, Alazawi W. Non-alcoholic fatty liver disease. *Clin Med*. (2020) 20:509–12. doi: 10.7861/clinmed.2020-0696

5. Neuberger J, Cain O. The need for alternatives to liver biopsies: non-invasive analytics and diagnostics. *Hepatic Med: Evid Res*. (2021) 13:59–69. doi: 10.2147/HMER.S278076

6. Lynch CJ, Liston C. New machine-learning technologies for computer-aided diagnosis. *Nat Med*. (2018) 24:1304–5. doi: 10.1038/s41591-018-0178-4

7. Nam D, Chapiro J, Paradis V, Seraphin TP, Kather JN. Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction. *JHEPReport*. (2022) 4(4):100443. doi: 10.1016/j.jhepr.2022.100443

8. Pan X, Xie X, Peng H, Cai X, Li H, Hong Q, et al. Risk prediction for non-alcoholic fatty liver disease based on biochemical and dietary variables in a chinese han population. *Front Public Health*. (2020) 8:220. doi: 10.3389/fpubh.2020.00220

9. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, et al. Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. *Cell Metab*. (2017) 25:1054–1062.e5. doi: 10.1016/j.cmet.2017.04.001

10. Taylor-Weiner A, Pokkalla H, Han L, Jia C, Huss R, Chung C, et al. A machine learning approach enables quantitative measurement of liver histology and disease monitoring in NASH. *Hepatology*. (2021) 74:133. doi: 10.1002/hep.31750

11. Zhou X, Lin X, Chen J, Pu J, Wu W, Wu Z, et al. Clinical spectrum transition and prediction model of nonalcoholic fatty liver disease in children with obesity. *Front Endocrinol (Lausanne)*. (2022) 13:986841. doi: 10.3389/fendo.2022.986841

12. Liu YX, Liu X, Cen C, Li X, Liu JM, Ming ZY, et al. Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: An extended study. *Hepatobiliary Pancreat Dis Int*. (2021) 20:409–15. doi: 10.1016/j.hbpd.2021.08.004

13. Huang G, Jin Q, Mao Y. Predicting the 5-year risk of nonalcoholic fatty liver disease using machine learning models: prospective cohort study. *J Med Internet Res*. (2023) 25:e46891. doi: 10.2196/46891

14. Ruhl CE, Everhart JE. Fatty liver indices in the multiethnic United States National Health and Nutrition Examination Survey. *Aliment Pharmacol Ther*. (2015) 41:65–76. doi: 10.1111/apt.13012

15. Han S, Kim H, Lee YS. Double random forest. *Mach Learn*. (2020) 109:1569–86. doi: 10.1007/s10994-020-05889-1

16. Valkenborg D, Rousseau AJ, Geubbelmans M, Burzykowski T. Support vector machines. *Am J Orthod Dentofacial Orthop*. (2023) 164:754–7. doi: 10.1016/j.ajodo.2023.08.003

17. Wang X, Huang F, Cheng Y. Computational performance optimization of support vector machine based on support vectors. *Neurocomputing*. (2016) 211:66–71. doi: 10.1016/j.neucom.2016.04.059

18. Goin JE. Classification bias of the k-nearest neighbor algorithm. *IEEE Trans Pattern Anal Mach Intell*. (1984) 6:379–81. doi: 10.1109/tpami.1984.4767533

19. Xiong Y, Zhu M, Li Y, Huang K, Chen Y, Liao J. Recognition of geothermal surface manifestations: A comparison of machine learning and deep learning. *Energies*. (2022) 15:2913. doi: 10.3390/en15082913

20. Liu B, Lin H, Chen Y, Yang C. Prediction of rock unloading strength based on PSO-XGBoost hybrid models. *Materials*. (2024) 17:4214. doi: 10.3390/ma17174214

21. Wang S, Ren J, Bai R. A regularized attribute weighting framework for naive bayes. *IEEE Access*. (2020) 8:225639–49. doi: 10.1109/ACCESS.2020.3044946

22. Ren J, Jiang X, Yuan J. A chi-squared-transformed subspace of LBP histogram for visual recognition. *IEEE Trans Image Processing*. (2015) 24:1893–904. doi: 10.1109/TIP.2015.2409554

23. Han W, Nan L, Su M, Chen Y, Li R, Zhang X. Research on the prediction method of centrifugal pump performance based on a double hidden layer BP neural network. *Energies*. (2019) 12:2709. doi: 10.3390/en12142709

24. Wan L, Li H, Chen Y, Li C. Rolling bearing fault prediction method based on QPSO-BP neural network and dempster–shafer evidence theory. *Energies*. (2020) 13:1094. doi: 10.3390/en13051094

25. Wang W, Feng J, Xu F. Estimating downward shortwave solar radiation on clear-sky days in heterogeneous surface using LM-BP neural network. *Energies*. (2021) 14:273. doi: 10.3390/en14020273

26. Seo W, Pak W. Real-time network intrusion prevention system based on hybrid machine learning. *IEEE Access*. (2021) 9:46386–97. doi: 10.1109/ACCESS.2021.3066620

27. Sannigrahi M, Thandeeswaran R. Predictive analysis of network-based attacks by hybrid machine learning algorithms utilizing bayesian optimization, logistic regression, and random forest algorithm. *IEEE Access*. (2024) 12:142721–32. doi: 10.1109/ACCESS.2024.3464866

28. Hagström H, Vessby J, Ekstedt M, Shang Y. 99% of patients with NAFLD meet MASLD criteria and natural history is therefore identical. *J Hepatol*. (2024) 80:e76–7. doi: 10.1016/j.jhep.2023.08.026

29. Li L, Liu DW, Yan HY, Wang ZY, Zhao SH, Wang B. Obesity is an independent risk factor for non-alcoholic fatty liver disease: evidence from a meta-analysis of 21 cohort studies. *Obes Rev*. (2016) 17:510–9. doi: 10.1111/obr.12407

30. Liu K, Tang S, Liu C, Ma J, Cao X, Yang X, et al. Systemic immune-inflammatory biomarkers (SII, NLR, PLR and LMR) linked to non-alcoholic fatty liver disease risk. *Front Immunol*. (2024) 15:1337241. doi: 10.3389/fimmu.2024.1337241

31. Song Y, Guo W, Li Z, Guo D, Li Z, Li Y. Systemic immune-inflammation index is associated with hepatic steatosis: Evidence from NHANES 2015-2018. *Front Immunol*. (2022) 13:1058779. doi: 10.3389/fimmu.2022.1058779

32. Tian T, Zhang J, Xie W, Ni Y, Fang X, Liu M, et al. Dietary quality and relationships with metabolic dysfunction-associated fatty liver disease (MAFLD) among United States adults, results from NHANES 2017-2018. *Nutrients*. (2022) 14:4505. doi: 10.3390/nu14214505

33. Yan J, Zhou J, Ding Y, Tu C. Dietary inflammatory index is associated with metabolic dysfunction-associated fatty liver disease among United States adults. *Front Nutr*. (2024) 11:1340453. doi: 10.3389/fnut.2024.1340453

34. Peiseler M, Schwabe R, Hampe J, Kubes P, Heikenwälder M, Tacke F. Immune mechanisms linking metabolic injury to inflammation and fibrosis in fatty liver disease - novel insights into cellular communication circuits. *J Hepatol*. (2022) 77:1136–60. doi: 10.1016/j.jhep.2022.06.012

35. Shoelson SE, Lee J, Goldfine AB. Inflammation and insulin resistance. *J Clin Invest*. (2006) 116:1793–801. doi: 10.1172/JCI29069

36. Tan Z, Wu Y, Meng Y, Liu C, Deng B, Zhen J, et al. Trends in oxidative balance score and prevalence of metabolic dysfunction-associated steatotic liver disease in the United States: national health and nutrition examination survey 2001 to 2018. *Nutrients*. (2023) 15:4931. doi: 10.3390/nu15234931

37. van Kleef LA, Hofman A, Voortman T, de Knegt RJ. Objectively measured physical activity is inversely associated with nonalcoholic fatty liver disease: the rotterdam study. *Off J Am Coll Gastroenterol ACG*. (2022) 117:311. doi: 10.14309/ajg.0000000000001584

38. Zelber-Sagi S, Nitzan-Kaluski D, Goldsmith R, Webb M, Zvibel I, Goldiner I, et al. Role of leisure-time physical activity in nonalcoholic fatty liver disease: A population-based study. *Hepatology*. (2008) 48:1791–8. doi: 10.1002/hep.22525

39. Vural Keskinler M, Mutlu HH, Sirin A, Erkalma Senates B, Colak Y, Tuncer I, et al. Visceral adiposity index as a practical tool in patients with biopsy-proven nonalcoholic fatty liver disease/nonalcoholic steatohepatitis. *Metab Syndr Relat Disord*. (2021) 19:26–31. doi: 10.1089/met.2020.0054

40. Zhang Y, Li B, Liu N, Wang P, He J. Evaluation of different anthropometric indicators for screening for nonalcoholic fatty liver disease in elderly individuals. *Int J Endocrinol*. (2021) 2021:6678755. doi: 10.1155/2021/6678755

41. Peng H, Pan L, Ran S, Wang M, Huang S, Zhao M, et al. Prediction of MAFLD and NAFLD using different screening indexes: A cross-sectional study in U.S. adults. *Front Endocrinol (Lausanne)*. (2023) 14:1083032. doi: 10.3389/fendo.2023.1083032

42. Boden G, She P, Mozzoli M, Cheung P, Gumireddy K, Reddy P, et al. Free fatty acids produce insulin resistance and activate the proinflammatory nuclear factor-kappaB pathway in rat liver. *Diabetes*. (2005) 54:3458–65. doi: 10.2337/diabetes.54.12.3458

43. Stefan N, Kantartzis K, Häring HU. Causes and metabolic consequences of Fatty liver. *Endocr Rev*. (2008) 29:939–60. doi: 10.1210/er.2008-0009

44. Fontana L, Eagon JC, Trujillo ME, Scherer PE, Klein S. Visceral fat adipokine secretion is associated with systemic inflammation in obese humans. *Diabetes*. (2007) 56:1010–3. doi: 10.2337/db06-1656

45. Zou H, Ma X, Zhang F, Xie Y. Comparison of the diagnostic performance of twelve noninvasive scores of metabolic dysfunction-associated fatty liver disease. *Lipids Health Dis*. (2023) 22:145. doi: 10.1186/s12944-023-01902-3

46. Xue Y, Xu J, Li M, Gao Y. Potential screening indicators for early diagnosis of NAFLD/MAFLD and liver fibrosis: Triglyceride glucose index-related parameters. *Front Endocrinol (Lausanne)*. (2022) 13:951689. doi: 10.3389/fendo.2022.951689

47. Younossi ZM, Golabi P, de Avila L, Paik JM, Srishord M, Fukui N, et al. The global epidemiology of NAFLD and NASH in patients with type 2 diabetes: A systematic review and meta-analysis. *J Hepatol*. (2019) 71:793–801. doi: 10.1016/j.jhep.2019.06.021

48. He Y, Su Y, Duan C, Wang S, He W, Zhang Y, et al. Emerging role of aging in the progression of NAFLD to HCC. *Ageing Res Rev*. (2023) 84:101833. doi: 10.1016/j.arr.2022.101833

49. Mridha AR, Wree A, Robertson AAB, Yeh MM, Johnson CD, Van Rooyen DM, et al. NLRP3 inflammasome blockade reduces liver inflammation and fibrosis in experimental NASH in mice. *J Hepatol*. (2017) 66:1037–46. doi: 10.1016/j.jhep.2017.01.022

50. Sarathy H, Henriquez G, Abramowitz MK, Kramer H, Rosas SE, Johns T, et al. Abdominal obesity, race and chronic kidney disease in young adults: results from NHANES 1999-2010. *PloS One*. (2016) 11:e0153588. doi: 10.1371/journal.pone.0153588

51. Dhaliwal R, Shepherd JA, El Ghormli L, Copeland KC, Geffner ME, Higgins J, et al. Changes in visceral and subcutaneous fat in youth with type 2 diabetes in the TODAY study. *Diabetes Care*. (2019) 42:1549–59. doi: 10.2337/dc18-1935