**Title:** **Biomarker discovery for MASLD utilizing Mendelian Randomization, Machine Learning, and external validation**

***Short Title*:** Biomarker discovery for MASLD

**Authors:**

Gong Feng1, Giovanni Targher2,3, Christopher D. Byrne4, Na He5, Man Mi6, Yi Liu7, Hongbin Zhu8, Ming-Hua Zheng9,10, Feng Ye1\*

1. Department of Infectious Diseases, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China;
2. Department of Medicine, University of Verona, Verona, Italy;
3. Metabolic Diseases Research Unit, IRCSS Sacro Cuore - Don Calabria Hospital, Negrar di Valpolicella, Italy;
4. Southampton National Institute for Health and Care Research Biomedical Research Centre, University Hospital Southampton and University of Southampton, Southampton General Hospital, Southampton, UK;
5. Department of Gastroenterology, The First Affiliated Hospital of Xi'an Medical University, Xi'an, China;
6. Xi'an Medical University, Xi'an, China;
7. Department of Traditional Chinese Medicine, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China;
8. Department of Gastroenterology, the 983rd Hospital of the Joint Logistics Support Force of the Chinese People's Liberation Army, Tianjin, China;
9. MAFLD Research Center, Department of Hepatology, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China;
10. Key Laboratory of Diagnosis and Treatment for The Development of Chronic Liver Disease in Zhejiang Province, Wenzhou, Zhejiang, China.

**\*Corresponding authors:**

Dr Feng Ye, Department of Infectious Disease, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China. Email: yefeng.jiaotong@163.com.

**Number of figures/supplementary figures:** 6/17

**Number of supplementary tables:** 5

**Author contributions to this manuscript:**

FY, GF, and NH were involved in study design, data interpretation, and verification. GF, and NH performed data analysis and collection. GF, FY and NH wrote the manuscript. MHZ, GT, CDB, YL, HBZ, MM and FY conducted a critical revision and contributed to writing the manuscript. All authors reviewed and commented on the manuscript and approved the final version.

**Grants Support:**

This work was supported by grants from Key Research and Development Program of the Shaanxi Provincial Department of Science and Technology (2025GH-YBXM-032), Project of Shaanxi Provincial Administration of Traditional Chinese Medicine (2022-SLRH-YQ-011), Tianjin Key Medical Discipline (Specialty) Construction Project (TJYXZDXK-077D), and National Inheritance Studio Project of Zheng Qinglian (Document No. [2022] 270).

**Conflicts of Interest:**

CDB has received grant support from Echosens. Other authors have no conflicts of interest to declare.

**Data availability statement**

The data supporting the findings of the study are available from the corresponding author upon reasonable request.

**Abbreviation list**

ALT, alanine aminotransferase; ApoA, apolipoprotein A; ApoB, apolipoprotein B; ApoE, apolipoprotein E; AST, aspartate aminotransferase; CNPY4, canopy FGF signaling regulator 4; ENTPD6, ectonucleoside triphosphate diphosphohydrolase 6; eGFR, estimated glomerular filtration rate; FDR, false discovery rate; GGT, gamma-glutamyl transferase; GWAS, genome-wide association study; HDL-C, high-density lipoprotein cholesterol; IVs, instrumental variables; HLA-A, human leukocyte antigen A; IGF-1, insulin-like growth factor-1; IVW, inverse variance weighting; LDL, low-density lipoprotein; MASLD, metabolic dysfunction-associated steatotic liver disease; MASH, metabolic dysfunction-associated steatohepatitis; MR, Mendelian randomization; NHANES, National Health and Nutrition Examination Survey; OR, odds ratio; *P*fdr, false discovery rate-adjusted P value; PPV, positive predictive value; ROC, receiver operating characteristic; SCG3, secretogranin III; SNP, single-nucleotide polymorphism; TCGA, The Cancer Genome Atlas; TOR1AIP1, torsin 1A interacting protein 1;T2DM, type 2 diabetes mellitus; UKB, UK Biobank; VLDL, very low-density lipoprotein.

**ABSTRACT**

**Background and Aims:** The causal biomarkers for metabolic dysfunction-associated steatotic liver disease (MASLD) and their clinical value remain unclear. We aimed to identify biomarkers for MASLD and evaluate their diagnostic and prognostic significance.

**Methods:** We conducted a Mendelian randomization (MR) analysis to assess the causal effects of 2,925 molecular biomarkers (derived from proteomics data) and 35 clinical biomarkers on MASLD. Mediation analysis was performed to determine whether clinical biomarkers mediated the effects of molecular biomarkers. The association between key clinical biomarkers and MASLD was validated externally in a hospital-based cohort (n = 415). A machine-learning diagnostic model for MASLD was developed and validated using the identified molecular biomarkers. Prognostic significance was assessed for both molecular and clinical biomarkers.

**Results:** Six molecular biomarkers—including CNPY4, ENTPD6, and HLA-A—and eight clinical biomarkers (e.g., serum total protein [STP]) were identified as causally related to MASLD. STP partially mediated the effect of HLA-A on MASLD (19.13%) and was associated with MASLD in the external cohort (OR = 1.080, 95% CI 1.011–1.155). A random forest model demonstrated high diagnostic performance (AUC = 0.941 in training; 0.875 in validation). High expression levels of CNPY4 and ENTPD6 correlated with the development of and poorer survival from hepatocellular carcinoma. Low STP (< 60 g/L) predicted all-cause mortality (HR = 2.50, 95% CI 1.22-5.09).

**Conclusions:** This study identified six causal molecular biomarkers (e.g., CNPY4, ENTPD6, HLA-A) and eight clinical biomarkers for MASLD. Notably, STP mediated the effect of HLA-A on MASLD and was associated with all-cause mortality.

**Keywords:** Metabolic dysfunction-associated fatty liver disease; Mendelian randomization; Machine learning; Proteomics; Mediation analysis; Non-invasive diagnosis; Prognosis; Causal biomarkers.

1. **Introduction**

Metabolic dysfunction-associated steatotic liver disease (MASLD) has attracted increasing attention, with a global prevalence of MASLD around 30-35%.1 A recent epidemiological model prediction suggests that by 2030, the global prevalence of MASLD will increase to 36.8%, corresponding to approximately 101.2 million people. By 2050, the prevalence of MASLD is expected to further increase to 41.4%, corresponding to ~122 million people.2 With the increasing prevalence of MASLD, the disease burden of MASLD-related liver cancer has also become increasingly significant.3 Globally, the incident cases, deaths, and disability-adjusted life years of MASLD-related liver cancer in 2019 increased by 205%, 195%, and 166%, respectively, compared with 1990.4 In the United States, MASLD has become the main pathogenic cause of liver cancer in liver transplant candidates.5

Despite growing awareness of MASLD, the causal biomarkers of this condition remain unclear.6 Identifying causal biomarkers of MASLD has become an urgent research priority. As an effective tool for causal inference, the Mendelian randomization (MR) method can reduce confounding and reverse causation biases, inherent in observational studies by utilizing genetic variations as instrumental variables, thereby providing more reliable evidence for the causal relationships between diseases and related factors.7,8 Although some studies have employed the MR method to examine genetic susceptibility and potential biomarkers for MASLD, most published studies have focused on a limited number of biomarkers, lacking a systematic and comprehensive approach.9,10

**The biomarkers related to MASLD include not only traditional clinical biomarkers but also** **novel molecular biomarkers identified through multi-omics.**11 **Clinical biomarkers,** as standard medical indicators, are widely used for disease diagnosis, prognostic assessment, and monitoring therapeutic responses because of their accessibility and established clinical utility.11 In contrast, **molecular biomarkers** reflect the underlying molecular mechanisms of disease development and progression, typically derived from multi-omics platforms such as transcriptomics, and proteomics.11,12 **Although many biomarkers associated with MASLD have been identified, their clinical significance remains insufficiently characterized.13 These biomarkers have the potential to enhance the accuracy of non-invasive MASLD diagnosis when used collectively, thereby diminishing the necessity for liver biopsies.14,15 Moreover, they may serve as prognostic indicators, exemplified by the novel acMASH index, correlating with all-cause mortality risk and aiding in risk stratification.16** Therefore, understanding the clinical and translational relevance of biomarkers with a confirmed causal relationship to MASLD is of critical importance.

In this study, we aimed to identify causal molecular biomarkers (based on proteomics data) and clinical biomarkers associated with MASLD, as well as evaluate their diagnostic and prognostic significance. First, we conducted MR analysis to assess the causal effects of 2,925 molecular biomarkers (based on proteomics data) and 35 clinical biomarkers on MASLD, which can reveal the casual relationship between molecular and clinical biomarkers and MASLD. Mediation analysis was performed to determine whether key clinical biomarkers could mediate the effects of molecular biomarkers (exposures) on MASLD (outcome). The association between key clinical biomarkers and MASLD was also validated externally in a hospital-based cohort. Second, we applied six machine-learning algorithms to develop and validate a novel noninvasive diagnostic model for MASLD based on the identified molecular biomarkers. Third, we explored the prognostic value of molecular biomarkers for the development of and poorer survival from hepatocellular carcinoma (HCC) by analyzing data from The Cancer Genome Atlas (TCGA). We also determined the prognostic value of key clinical biomarkers for all-cause and mortality and cause-specific mortality by analyzing prospective data from the National Health and Nutrition Examination Survey (NHANES).

**2. Methods**

***2.1 Data Collection of*** ***molecular and clinical biomarkers and MASLD for MR analysis***

In this study, the molecular biomarkers were derived from proteomics data obtained from the FinnGen database.17 This database included 619 samples encompassing 2,925 molecules, such as apolipoprotein E (APOE), FGF signaling regulator 4 (CNPY4), ectonucleoside triphosphate di-phosphohydrolase 6 (ENTPD6), major histocompatibility complex, class I, A (HLA-A), secretogranin III (SCG3) and torsin 1A interacting protein 1 (TOR1AIP1), which were publicly released in April 2024.17 Our clinical biomarkers comprised 35 blood and urine biomarkers from the UK Biobank dataset, which included 363,228 individuals.18 These 35 biomarkers are extensively utilized in clinical diagnostics and play a crucial role in assessing various physiological functions, including serum albumin, apolipoprotein E (ApoE), gamma-glutamyl transferase (GGT), high-density lipoprotein cholesterol (HDL-C), insulin-like growth factor 1 (IGF-1), total protein, triglycerides, and urinary sodium. Data on MASLD included a meta-analysis of genome-wide association studies (GWAS) involving four cohorts of electronic health record-documented MASLD among participants of European ancestry (8,434 cases and 770,180 controls).19

***2.2 Instrumental variable screening and MR analysis***

We referred to prior literature for the screening criteria of instrumental variables (IVs) (**Supplementary Method 1**).20 For example, in the selection of IVs for 35 clinical biomarkers, single-nucleotide polymorphisms (SNPs) were selected based on genome-wide significance (*P* < 5 × 10⁻⁸) and were subsequently clumped to ensure independence using a linkage disequilibrium threshold of r² < 0.001 within a 10,000-kb distance. A two-step, two-sample MR approach was employed to assess the causal relationships between molecular biomarkers and MASLD, as well as between clinical biomarkers and MASLD. This two-step, two-sample MR approach is a genetics-based method for causal inference, carried out through two independent sample datasets in two stages. In the first step, GWAS is used to identify genetic variants (such as SNPs) that are significantly associated with the exposure factor in the initial sample, thereby establishing a strong link between these variants and the exposure.20 In the second step, these instrumental variables are tested for association with disease outcomes in a second independent sample using statistical models, such as the inverse-variance weighted method (IVW), while conducting heterogeneity and pleiotropy tests to validate the plausibility of the hypothesis.20 To ensure the reliability of the MR analysis results, we conducted heterogeneity, pleiotropy, and sensitivity analyses (**Supplementary Method 2**). All MR analyses employed the IVW method as the primary causal association test method, and the results were verified using the MR-Egger, weighted median, weighted mode, and simple mode methods. In addition, since all MR analyses used the IVW method as the main causal association test method, we corrected the results of the IVW using the false discovery rate (FDR) method, and *P*fdr <0.05 was considered statistically significant. For other statistical analyses without FDR adjustment, a *P*-value < 0.05 was considered indicative of statistical significance.

***2.3 Mediation effect analysis to identify key clinical biomarkers with a mediating role***

Mediation effect analysis was used to determine the mediating role of key clinical biomarkers in the causal relationship between **molecular biomarkers** and MASLD. The formulas for calculating each effect size are as follows: total effect = C × 100%, indirect effect = A × B × 100%, direct effect (C') = [C – (A × B) × 100%], and the proportion of the mediation effect = (A × B) / C × 100%. Among them, A is the β1 value of the MR analysis between the key protein and the clinical marker, B is the β2 value of the MR analysis between the clinical marker and MASLD, and C is the β value of the MR analysis between the key protein and MASLD.

**2.4 *Validation of the association between key clinical biomarkers and MASLD***

Through mediation analyses, we identified key clinical markers that mediate the effects of molecular biomarkers on MASLD. To further validate externally the association between these key clinical markers and MASLD, we utilized a follow-up cohort of MASLD patients from the First Affiliated Hospital of Xi'an Medical University who had undergone vibration controlled transient elastography (VCTE) examinations.20 Patients with a CAP value ≥ 248 dB/m, as assessed by FibroScan, were considered to have hepatic steatosis.21 (**Supplementary Method 3**).

***2.5 Machine learning algorithms for MASLD based on the identified*** ***molecular biomarkers***

We employed six widely used machine learning (ML) algorithms—including Extreme Gradient Boosting (XGBoost), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Light Gradient Boosting Machine (LightGBM)—to develop a non-invasive diagnostic model for MASLD. Expression profiles of the identified molecular biomarkerswere obtained from the Gene Expression Omnibus (GEO) database, with GSE89632 (n = 63) serving as the training set and GSE48452 (n = 73) as the external validation cohort.

***2.6 Evaluation of the prognostic significance of the identified molecular and clinical biomarkers***

We further evaluated the prognostic value of identified molecular biomarkers, particularly their impact on hepatocellular carcinoma (HCC) development and patients' overall survival **(****Supplementary Method 4).** To investigate the associations between clinical biomarkers and the risk of all-cause and cause-specific mortality, we utilized prospective data from the NHANES, collected between 1999 and 2006 **(Supplementary Method 5)**. The study design is illustrated in **Figure 1.**

1. **Results**

**3.1** ***Molecular biomarkers causally related to MASLD***

The IVW algorithm indicated that among the 2,925 molecular biomarkers, only six exhibited a significant relationship with MASLD (**Figure 2A**). The mean F-statistics for the selected IVs were as follows: APOE (F = 36.919), CNPY4 (F = 26.539), ENTPD6 (F = 43.279), HLA-A (F = 38.635), SCG3 (F = 30.938), and TOR1AIP1 (F = 48.678). These mean F-statistics were well above the threshold of 10, indicating that all the IVs are strong instruments (all F > 10). The IVW algorithm showed that the odds ratio (OR) for APOE was 1.057 (95%CI 1.031-1.083, *P*fdr = 4.62E-03), the OR for CNPY4 was 1.054 (95%CI 1.029-1.081, *P*fdr = 8.38E-03), the OR for ENTPD6 was 1.031 (95%CI 1.016-1.045, *P*fdr = 5.75E-03), the OR for HLA-A was 0.969 (95%CI 0.960-0.979, *P*fdr = 1.81E-06), the OR for SCG3 was 0.956 (95%CI 0.937-0.975, *P*fdr = 4.62E-03), and the OR for TOR1AIP1 was 0.964 (95% CI 0.950-0.979, *P*fdr = 8.13E-04), respectively. Results from the other four algorithms are shown in **Table S1**. The Cochran's Q test analysis revealed no heterogeneity in the results for APOE, CNPY4, ENTPD6, HLA-A, SCG3, and TOR1AIP1(**Table S2**). Subsequently, the MR-Egger intercept test was conducted to evaluate the presence of pleiotropy among IVs (**Table S2**). The results showed no horizontal pleiotropy between APOE, CNPY4, ENTPD6, HLA-A, SCG3, TOR1AIP1, and MASLD (**Table S2**). The leave-one-out sensitivity analysis confirmed that the relationship between TOR1AIP1 and MASLD remained stable (**Figure 2B**). From the scatter plots, the relationship between TOR1AIP1 and MASLD showed a consistent trend (the OR values were all greater than 1) among the five MR algorithms, further indicating the robustness of our results (**Figure 2C**). The scatter plots and leave-one-out sensitivity analyses for the remaining five proteins are detailed in **Figures S1-S5**. The reverse MR analysis found no significant association between MASLD and the aforementioned six proteins.

***3.2 Clinical biomarkers causally related to MASLD***

The IVW algorithm indicated that, among the 35 clinical biomarkers, only eight exhibited a significant relationship with MASLD (**Figure 3A**). The mean F-statistics for the selected IVs were as follows: albumin (F = 73.095), ApoA (F = 135.563), GGT (F = 129.355), HDL-C (F = 145.250), IGF-1 (F = 98.572), urinary sodium (F = 40.886), total protein (F = 63.070), and triglycerides (F = 147.741). All these values exceeded the threshold of 10, indicating that the IVs are strong instruments. The IVW algorithm revealed that the OR for albumin was 1.373 (95%CI: 1.140 - 1.654, *P*fdr = 4.11E-03), the OR for ApoA was 0.811 (95% CI: 0.695 - 0.946, *P*fdr = 0.026), the OR for GGT was 1.281 (95%CI: 1.167 - 1.406, *P*fdr = 1.25E-06), the OR for HDL-C was 0.792 (*P*fdr = 7.53E-05), the OR for IGF-1 was 0.870 (95%CI: 0.784 - 0.964, *P*fdr = 0.027), the OR for urinary sodium was 2.583 (95%CI: 1.407 - 4.743, *P*fdr = 8.548E-03), the OR for total protein was 1.248 (95%CI: 1.083 - 1.438, *P*fdr = 8.488E-03), and the OR for triglycerides was 1.392 (95%CI: 1.239 - 1.563, *P*fdr = 2.15E-07), respectively. The MR-Egger intercept test demonstrated that there was no horizontal pleiotropy between albumin, ApoA, GGT, HDL-C, IGF-1, urinary sodium, total protein, triglycerides, and MASLD. The scatter plots indicated that the relationship between total protein and MASLD exhibited a consistent trend (the OR values were all greater than 1) among the five MR algorithms, further establishing the robustness of our results (**Figure 3B**). We also used a Manhattan plot to display the distribution characteristics of SNPs of total protein after removing linkage disequilibrium (**Figure 3C**). The scatter plots of the other seven clinical biomarkers are shown in **Figures S6-S9.**

***3.3 Total protein demonstrated a significant mediating effect in the relationship between HLA-A and MASLD***

Before conducting the mediation analysis, we also used an MR approach to explore the causal relationships between six molecular biomarkers (exposures) and eight clinical biomarkers (outcomes), which serves as a prerequisite for the mediation analysis. This analysis revealed that the OR between HLA-A and total protein was 0.967 (95% CI: 0.948 - 0.987, *P*fdr = 1.15E-02). The results confirmed no horizontal pleiotropy between HLA-A and MASLD (MR-Egger intercept = -0.011, *P*fdr = 0.556). The scatter plots and leave-one-out sensitivity analyses between HLA-A and total protein are presented in **Figure S10.** Further mediation analysis showed that total protein exhibited a significant mediating effect in the association between HLA-A and MASLD (**Figure 4A, 4B**). Specifically, the total effect was 0.969 (95% CI: 0.960–0.979, *P*fdr = 1.81 × 10⁻⁶), with total protein mediating 19.13% of the relationship between HLA-A and MASLD (OR = 0.993, *P*fdr < 0.05).

***3.4 Validation of the relationship between serum total protein levels and MASLD***

The independent validation cohort included 330 patients with MASLD confirmed by VCTE and 85 controls. **Table S3** shows the baseline characteristics of the participants. In the subsequent multivariable logistic regression analysis **(Table S4)**, serum total protein levels showed a positive association with the risk of MASLD that remained statistically significant across all regression models. Specifically, in the age- and sex-adjusted model, the OR for MASLD was 1.103 (95% CI: 1.064–1.143). In multivariable model 1, which adjusted for age, sex, body mass index (BMI), diabetes, and hypertension, the OR was 1.092 (95% CI: 1.052–1.134, *P* < 0.001). Even after additional adjustment for serum liver enzymes, lipids, creatinine, HbA1c, and platelet count (adjusted model 2), the association between total protein and MASLD remained significant, with an adjusted OR of 1.080 (*P* = 0.023).

**3.5 *Performance of multiple machine learning algorithms for MASLD based on six molecular biomarkers***

We developed a non-invasive model for diagnosing MASLD using the six identified molecular biomarkers and six commonly used supervised machine-learning algorithms. As shown in **Figure 5A**, the RF model demonstrated the best performance in the training set, achieving an AUC of 0.941 (95% CI: 0.829–1.000), followed by KNN (AUC = 0.885, 95% CI: 0.732–1.000) and XGBoost (AUC = 0.834, 95% CI: 0.662–0.975). Conversely, the MLP model performed poorly, with an AUC of 0.536 (95% CI: 0.244–0.827). Additionally, we applied the Shapley Additive exPlanations (SHAP) method to analyze the random forest model. As illustrated in **Figure 5B**, the analysis revealed that CNPY4, SCG3, TOR1AIP1, and ENTPD6 were the top four molecular features influencing the model output. Notably, CNPY4 exhibited the highest mean SHAP value, indicating its pivotal role in distinguishing MASLD from non-MASLD individuals. In the validation dataset (GSE48452) (**Figure 5C**), the RF model maintained superior discriminative performance with an AUC of 0.875, demonstrating robust generalization and discriminative capability. KNN and XGBoost also showed consistent performance, while MLP continued to exhibit poor performance in the validation set (GSE48452). To assess the robustness and potential overfitting of the RF model, we performed 500 bootstrap resampling iterations separately in both the training and validation sets. The resulting ROC curves with 95% confidence intervals are shown in **Figure S11**.

* 1. ***Prognostic molecular and clinical biomarkers associated with MASLD***

We further analyzed the prognostic significance of key molecular biomarkers closely associated with MASLD, particularly their impact on the development of HCC and overall survival. We found that CNPY4 was significantly upregulated in HCC **(Figure S12)**. Across multiple cancer types, elevated CNPY4 expression demonstrated varying degrees of association with overall survival, with a particularly strong link in LIHC, where high CNPY4 expression was significantly associated with poor prognosis (HR = 1.753, **Figure S13**). Notably, in LIHC, CNPY4 expression was positively correlated with several immune cell types, especially macrophages, dendritic cells, and T helper cells **(Figure S14)**, suggesting that CNPY4 may contribute to tumor progression by modulating the tumor immune microenvironment. Similarly, we observed that ENTPD6 was significantly overexpressed in LIHC (*P* < 0.001, **Figure S15**). Survival analysis indicated that high ENTPD6 expression was significantly associated with poorer overall survival in patients with (HR 1.483, *P* < 0.05, **Figure S16**), highlighting ENTPD6 as another potential adverse prognostic biomarker. In LIHC, ENTPD6 expression also exhibited significant positive associations with various immune cell populations, including dendritic cells, macrophages, CD8+ T cells, T helper cells, and regulatory T cells (*P* < 0.05, **Figure S17).**

Given that serum total protein level mediates the effect of the molecular biomarker HLA-A on MASLD, we regarded it as an important clinical biomarker. Therefore, we further investigated its prognostic significance using data from the NHANES study. A total of 41,474 individuals were initially identified from the NHANES database. After excluding participants with missing data for serum total protein levels and mortality data, 3,540 individuals were included in the final analysis. Kaplan–Meier survival curves are shown in **Figure 6**. Individuals with total protein levels <60 g/L had a significantly higher risk of all-cause mortality compared to those with total protein levels ≥60 g/L (HR, 2.180; 95% CI 1.188-4.002) in the age- and sex-adjusted model, with an HR of 2.769 (95% CI 1.441-5.319) in regression model 1 adjusted for age, sex, marital status, hypertension, diabetes, and BMI, and an HR of 2.495 (95% CI 1.224-5.087) in regression model 2, which was additionally adjusted for serum GGT, ALT, total cholesterol, triglycerides, and platelet count (**Table S5**).

**4. Discussion**

In this study, we identified several molecular and clinical biomarkers that are causally associated with MASLD and explored their diagnostic significance for MASLD as well as prognostic relevance for mortality outcomes. Notably, total serum protein levels were found to partially mediate the effect of HLA-A on the risk of MASLD, revealing a novel immuno-metabolic causal pathway. To translate these findings into a practical clinical tool, we developed a non-invasive diagnostic model based on the six MR-identified proteins using multiple machine learning algorithms. Among these, the RF model demonstrated excellent performance (AUC = 0.941 in the training set and 0.875 in the validation set), underscoring its potential utility in early MASLD screening for diagnosis. Additionally, we demonstrated that higher expression levels of CNPY4 and ENTPD6 were associated with poorer overall survival in HCC, while lower serum total protein levels were linked to an increased all-cause mortality in the general population. These findings suggest that certain MASLD-related biomarkers could be relevant to prognosis in broader clinical settings.

In our study, six proteins (molecular biomarkers) were identified as having a significant causal relationship with MASLD. ApoE is an essential component of chylomicrons, and studies have shown that polymorphisms in the ApoE gene are closely related to the development of MASLD.22 Amzolini and colleagues reported that the frequency of HLA-A25 in patients with MASLD was significantly lower than that in the healthy control group.23 Shin et al. found that the deletion of the TOR1AIP1 gene induced the development of MASLD.24 Currently, there are very few reports on the associations of CNPY4, ENTPD6, and SCG3 with MASLD, and these proteins may represent key targets for future research on MASLD. CNPY4 is localized in the endoplasmic reticulum and assists in protein maturation and lipid synthesis. Hotta et al. found that the rs3764220 variant in the SCG3 gene was associated with metabolic syndrome.25 ENTPD6 belongs to the Ectonucleoside Triphosphate Diphydrolase family, whose encoded proteins hydrolyze nucleoside triphosphates and nucleoside diphosphates, playing a crucial role in cell signaling and energy metabolism.26 A recent MR study identified potential targets for abdominal obesity and found that ENTPD6 may be a novel biomarker for interventions targeting visceral adipose tissue.27

Among the 35 blood and urine biomarkers, eight were identified as closely associated with MASLD. A meta-analysis incorporating 12 studies revealed that circulating levels of IGF-1 were significantly lower in individuals with MASLD than healthy controls.28 GGT in human serum primarily originates from the liver and biliary system. Serum GGT level not only serves as a conventional liver function marker but also plays a broader role in metabolic health.29 Additionally, serum GGT level is included in the equation for the fatty liver index used to identify hepatic steatosis.30 Furthermore, urinary sodium concentration was closely associated with MASLD (OR = 2.48, 95% CI: 1.52–4.06).31

HLA-A is a member of the major histocompatibility complex class I family and plays a central role in antigen presentation and immune surveillance. Genetic variations in HLA-A have previously been associated with metabolic and inflammatory diseases.32-34 This study revealed that total protein functions as a (partial) mediator in the causal pathway from HLA-A to MASLD, accounting for 19.13% of the total effect. Our mediation analysis indicates that downregulation of HLA-A may lead to increased total protein levels, which in turn could elevate the risk of MASLD. The liver is enriched with CD4+ T helper (Th) cells, CD8+ cytotoxic T cells, and B lymphocytes, all of which contribute to persistent inflammation and tissue remodeling.35,36 Mechanistically, HLA-A, a core component of MHC class I molecules, is responsible for presenting endogenous antigens to CD8⁺ T cells, thereby maintaining immune surveillance and homeostasis. When HLA-A expression is reduced, impaired antigen presentation leads to dysfunctional CD8⁺ T-cell activation and inefficient clearance of endogenous antigens derived from apoptotic cells or lipotoxic stress.37,38 The persistence of these antigens may induce chronic antigenic stimulation, activate B cells, and promote polyclonal immunoglobulin production, resulting in elevated serum globulin and thus increased total protein.39 The consequent elevation in globulin levels contributes to increased total protein concentrations, indicative of immune activation, which plays a crucial role in the pathogenesis and progression of MASLD.40 These findings suggest that HLA-A downregulation may indirectly contribute to MASLD by promoting immune pathways.

In this study, six ML algorithms were used to construct a MASLD noninvasive diagnostic model. In recent years, the application of ML technology in MASLD has made remarkable progress. The utilization of ML in MASLD has shown notable advancements in recent years, positioning it as a pivotal tool for precise diagnosis and treatment.41 According to different training methods, ML mainly includes supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.42 In our analysis, we used supervised ML and found that RF has the best discriminative power, better than other algorithms. RF is an ensemble learning algorithm based on decision trees, which significantly improves the accuracy of the model by integrating the prediction results of multiple decision trees.43 Compared with other algorithms, RF has a better tolerance for noise and outliers, because its prediction results are based on the synthesis of multiple decision trees.44 Furthermore, RF demonstrates resistance to overfitting and strong generalization abilities through the incorporation of a random construction approach.43

Previous studies have shown that patients with MASLD are at an increased risk of developing HCC.**45** In our study, we found that CNPY4 and ENTPD6 not only contribute to the pathogenesis of MASLD but may also play a role in the development of HCC, further supporting their significance and utility as molecular biomarkers related to liver disease. Moreover, both CNPY4 and ENTPD6 were associated with poor prognosis in HCC, suggesting their potential as prognostic biomarkers in liver diseases. CNPY4 is closely associated with immune regulation and exhibits carcinogenic effects across various tumors.46 ENTPD6 is involved in extracellular purine metabolism and may contribute to tumor metabolic reprogramming by modulating mitochondrial function. Previous studies have consistently demonstrated that the rs738409 variant in the PNPLA3 gene, which encodes patatin-like phospholipase domain-containing protein 3, significantly increases the risk of both MASH and MASLD-related HCC.47,48 This variant impairs triglyceride hydrolysis in hepatocytes, promoting lipid accumulation and resulting in more than a twofold increase in MASH risk (compared to healthy controls) and a 2.2-fold increased risk of MASLD-related HCC (compared to MASLD patients without the variant).47,48 However, as a static genetic susceptibility variant, PNPLA3 is primarily useful for long-term risk prediction and does not reflect dynamic biological changes during disease progression. In contrast, CNPY4 and ENTPD6 are expression-based protein biomarkers that can be quantitatively monitored and may offer greater potential for clinical translation, particularly in dynamic risk stratification, progression surveillance, and therapeutic response assessment in MASLD-HCC.

Additionally, based on prospective data from the NHANES study, we found that individuals with serum total protein levels below 60 g/L, a marker identified as a potential mediator of mortality, exhibited an increased risk of all-cause mortality. This may be due to the fact that hypo-proteinemia often reflects inadequate protein intake or synthesis, leading to malnutrition, impaired organ function, delayed tissue repair, and reduced resilience to illness; hypo-proteinemia can also result in edema, ascites, hypovolemia, and tissue hypoperfusion, potentially triggering complications such as renal dysfunction and hypotension, all of which may contribute to increased all-cause mortality.**49**

Although this study employed the MR method, it still has several limitations. First, the GWAS data lacks extensive validation across different ethnic groups. Therefore, future studies must validate these findings in various ethnic populations to confirm the causal relationships. Second, while this study has identified the causal proteins and biomarkers associated with MASLD, the specific biological mechanisms underlying these findings remain unexplored. Third, MASLD is a metabolically driven liver disease with complex and multifactorial etiologies. Other potential mediators—particularly those related to systemic inflammation or immune-metabolic interactions—may exist but were not included in the current analysis due to limitations in the available datasets. Future studies incorporating more comprehensive inflammatory markers, multi-omics data, and immune phenotyping are warranted to refine and expand the mediation pathway models. Fourth, the hospital-based cohort used in this study was derived from a tertiary medical center and had a relatively limited sample size, which may introduce a risk of selection bias. Compared with the general MASLD population, patients treated at tertiary centers are more likely to have complex disease presentations and a higher burden of metabolic comorbidities. Such differences in population characteristics may affect the generalizability of our findings.

In conclusion, we identified six causal molecular biomarkers (e.g., CNPY4, ENTPD6, HLA-A) and eight clinical biomarkers (e.g., serum total protein) for MASLD across various independent cohorts. Serum total protein levels partially mediated the effect of HLA-A on MASLD, highlighting a novel immune-metabolic pathway. Based on these findings, we also developed a random forest model that demonstrated high accuracy in identifying MASLD. Additionally, CNPY4 and ENTPD6 were associated with poor survival in HCC, while low serum total protein levels predicted higher all-cause mortality. These findings support a multi-omics framework for biomarker-driven diagnosis and risk prediction in MASLD.

**ACKNOWLEDGMENTS**

We would like to acknowledge all participants and investigators involved in the FinnGen study. CDB is supported in part by the Southampton National Institute for Health and Care Research, Biomedical Research Centre (NIHR203319).

**References**

[1] Younossi ZM, Golabi P, Paik JM, Henry A, Van Dongen C, Henry L. The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): a systematic review. Hepatology 2023;77(4):1335-1347. doi: 10.1097/hep.0000000000000004. PMID: 36626630

[2] Le P, Tatar M, Dasarathy S, Alkhouri N, Herman WH, Taksler GB*, et al.* Estimated Burden of Metabolic Dysfunction-Associated Steatotic Liver Disease in US Adults, 2020 to 2050. JAMA Netw Open 2025;8(1):e2454707. doi: 10.1001/jamanetworkopen.2024.54707. PMID: 39821400

[3] Ioannou GN. Epidemiology and risk-stratification of NAFLD-associated HCC. Journal of hepatology 2021;75(6):1476-1484. doi: 10.1016/j.jhep.2021.08.012. PMID: 34453963

[4] Huang H, Liu Z, Xu M, Chen Y, Xu C, Wang L, Wang N. Global burden trends of MAFLD-related liver cancer from 1990 to 2019. Portal Hypertension & Cirrhosis 2023;02(04):157-164. doi: doi:10.1002/poh2.63. PMID:

[5] Polyzos SA, Chrysavgis L, Vachliotis ID, Chartampilas E, Cholongitas E. Nonalcoholic fatty liver disease and hepatocellular carcinoma:Insights in epidemiology, pathogenesis, imaging, prevention and therapy. Seminars in cancer biology 2023;93:20-35. doi: 10.1016/j.semcancer.2023.04.010. PMID: 37149203

[6] Huang DQ, Wong VWS, Rinella ME, Boursier J, Lazarus JV, Yki-Järvinen H, Loomba R. Metabolic dysfunction-associated steatotic liver disease in adults. Nat Rev Dis Primers 2025;11(1):14. doi: 10.1038/s41572-025-00599-1. PMID: 40050362

[7] Xia T, Du M, Li H, Wang Y, Zha J, Wu T, Ju S. Association between Liver MRI Proton Density Fat Fraction and Liver Disease Risk. Radiology 2023;309(1):e231007. doi: 10.1148/radiol.231007. PMID: 37874242

[8] Yuan M, Hu X, Yao L, Chen P, Wang Z, Liu P*, et al.* Causal Relationship Between Gut Microbiota and Liver Cirrhosis: 16S rRNA Sequencing and Mendelian Randomization Analyses. J Clin Transl Hepatol 2024;12(2):123-133. doi: 10.14218/jcth.2023.00259. PMID: 38343609

[9] Xu S, Liu L, Li C, Ren Y, Zhang M, Xiang L*, et al.* Correlation Among Psoriasis, Iridocyclitis, and Non-alcoholic Fatty Liver Disease: Insights from Mendelian Randomization and Mediation Analysis. International journal of medical sciences 2025;22(1):121-131. doi: 10.7150/ijms.102369. PMID: 39744174

[10] Yan X, Huang S, Li H, Feng Z, Kong J, Liu J. The causal effect of mTORC1-dependent circulating protein levels on nonalcoholic fatty liver disease: A Mendelian randomization study. Dig Liver Dis 2024;56(4):559-564. doi: 10.1016/j.dld.2023.09.017. PMID: 37778897

[11] Di Mauro S, Scamporrino A, Filippello A, Di Pino A, Scicali R, Malaguarnera R*, et al.* Clinical and Molecular Biomarkers for Diagnosis and Staging of NAFLD. Int J Mol Sci 2021;22(21) doi: 10.3390/ijms222111905. PMID: 34769333

[12] Du Y, Li R, Fu D, Zhang B, Cui A, Shao Y*, et al.* Multi-omics technologies and molecular biomarkers in brain tumor-related epilepsy. CNS Neurosci Ther 2024;30(4):e14717. doi: 10.1111/cns.14717. PMID: 38641945

[13] Trinks J, Mascardi MF, Gadano A, Marciano S. Omics-based biomarkers as useful tools in metabolic dysfunction-associated steatotic liver disease clinical practice: How far are we? World J Gastroenterol 2024;30(14):1982-1989. doi: 10.3748/wjg.v30.i14.1982. PMID: 38681130

[14] Feng G, Zhang X, Zhang L, Liu WY, Geng S, Yuan HY*, et al.* Novel urinary protein panels for the non-invasive diagnosis of non-alcoholic fatty liver disease and fibrosis stages. Liver Int 2023;43(6):1234-1246. doi: 10.1111/liv.15565. PMID: 36924436

[15] Wu XX, Zheng KI, Boursier J, Chan WK, Yilmaz Y, Romero-Gomez M*, et al.* acNASH index to diagnose nonalcoholic steatohepatitis: a prospective derivation and global validation study. EClinicalMedicine 2021;41:101145. doi: 10.1016/j.eclinm.2021.101145. PMID: 34646997

[16] Feng G, Mozes FE, Ji D, Treeprasertsuk S, Okanoue T, Shima T*, et al.* acFibroMASH Index for the Diagnosis of Fibrotic MASH and Prediction of Liver-related Events: An International Multicenter Study. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association 2025;23(5):785-796. doi: 10.1016/j.cgh.2024.07.045. PMID: 39362618

[17] Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM*, et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. Nature 2023;613(7944):508-518. doi: 10.1038/s41586-022-05473-8. PMID: 36653562

[18] Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M*, et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat Genet 2021;53(2):185-194. doi: 10.1038/s41588-020-00757-z. PMID: 33462484

[19] Ghodsian N, Abner E, Emdin CA, Gobeil É, Taba N, Haas ME*, et al.* Electronic health record-based genome-wide meta-analysis provides insights on the genetic architecture of non-alcoholic fatty liver disease. Cell Rep Med 2021;2(11):100437. doi: 10.1016/j.xcrm.2021.100437. PMID: 34841290

[20] Feng G, He N, Gao J, Li XC, Zhang FN, Liu CC*, et al.* Causal relationship between key genes and metabolic dysfunction-associated fatty liver disease risk mediated by immune cells: A Mendelian randomization and mediation analysis. Diabetes Obes Metab 2024 Published Online First: 20240904. doi: 10.1111/dom.15925. PMID: 39228284

[21] Lin H, Lee HW, Yip TC, Tsochatzis E, Petta S, Bugianesi E*, et al.* Vibration-Controlled Transient Elastography Scores to Predict Liver-Related Events in Steatotic Liver Disease. JAMA 2024;331(15):1287-1297. doi: 10.1001/jama.2024.1447. PMID: 38512249

[22] Liu S, Weng R, Gu X, Li L, Zhong Z. Association between apolipoprotein E gene polymorphism and nonalcoholic fatty liver disease in Southern China: A case-control study. Journal of clinical laboratory analysis 2021;35(12):e24061. doi: 10.1002/jcla.24061. PMID: 34664321

[23] Amzolini AM, Fortofoiu M, Tudorica-Micu SE, Fortofoiu MC, Neagoe D, Popescu M*, et al.* Genetic Factors Involved in the Development and Progression of Nonalcoholic Fatty Liver Disease. Curr Health Sci J 2015;41(4):297-301. doi: 10.12865/chsj.41.04.01. PMID: 30538833

[24] Shin JY, Hernandez-Ono A, Fedotova T, Östlund C, Lee MJ, Gibeley SB*, et al.* Nuclear envelope-localized torsinA-LAP1 complex regulates hepatic VLDL secretion and steatosis. J Clin Invest 2019;129(11):4885-4900. doi: 10.1172/jci129769. PMID: 31408437

[25] Hotta K, Kitamoto T, Kitamoto A, Mizusawa S, Matsuo T, Nakata Y*, et al.* Association of variations in the FTO, SCG3 and MTMR9 genes with metabolic syndrome in a Japanese population. Journal of human genetics 2011;56(9):647-651. doi: 10.1038/jhg.2011.74. PMID: 21796137

[26] Zimmermann H. Ectonucleoside triphosphate diphosphohydrolases and ecto-5'-nucleotidase in purinergic signaling: how the field developed and where we are now. Purinergic Signal 2021;17(1):117-125. doi: 10.1007/s11302-020-09755-6. PMID: 33336318

[27] Zhang G, Han B, Chen Y, Jiang W, Fu J, Xu X*, et al.* Genetic insights into visceral obesity with health conditions, from disease susceptibility to therapeutic intervention. Postgraduate medical journal 2025 Published Online First: 20250121. doi: 10.1093/postmj/qgaf004. PMID: 39835424

[28] Yao Y, Miao X, Zhu D, Li D, Zhang Y, Song C, Liu K. Insulin-like growth factor-1 and non-alcoholic fatty liver disease: a systemic review and meta-analysis. Endocrine 2019;65(2):227-237. doi: 10.1007/s12020-019-01982-1. PMID: 31243652

[29] Kwak J, Seo IH, Lee YJ. Serum γ-glutamyltransferase level and incidence risk of metabolic syndrome in community dwelling adults: longitudinal findings over 12 years. Diabetology & metabolic syndrome 2023;15(1):29. doi: 10.1186/s13098-023-01000-5. PMID: 36823659

[30] Chung GE, Jeong SM, Cho EJ, Yoo JJ, Cho Y, Lee KN*, et al.* Association of fatty liver index with all-cause and disease-specific mortality: A nationwide cohort study. Metabolism 2022;133:155222. doi: 10.1016/j.metabol.2022.155222. PMID: 35636583

[31] Shojaei-Zarghani S, Safarpour AR, Fattahi MR, Keshtkar A. Sodium in relation with nonalcoholic fatty liver disease: A systematic review and meta-analysis of observational studies. Food Sci Nutr 2022;10(5):1579-1591. doi: 10.1002/fsn3.2781. PMID: 35592291

[32] Zhang J, Zhao L, Wang B, Gao J, Wang L, Li L*, et al.* HLA-A\*33-DR3 and A\*33-DR9 haplotypes enhance the risk of type 1 diabetes in Han Chinese. Journal of diabetes investigation 2016;7(4):514-521. doi: 10.1111/jdi.12462. PMID: 27181214

[33] Jin H, Kim YA, Lee Y, Kwon SH, Do AR, Seo S*, et al.* Identification of genetic variants associated with diabetic kidney disease in multiple Korean cohorts via a genome-wide association study mega-analysis. BMC Med 2023;21(1):16. doi: 10.1186/s12916-022-02723-4. PMID: 36627639

[34] Karrar A, Hariharan S, Fazel Y, Moosvi A, Houry M, Younoszai Z*, et al.* Analysis of human leukocyte antigen allele polymorphism in patients with non alcoholic fatty liver disease. Medicine (Baltimore) 2019;98(32):e16704. doi: 10.1097/md.0000000000016704. PMID: 31393374

[35] Li Z, Wang S, Xu Q, Su X, Wang Y, Wang L, Zhang Y. The double roles of T cell-mediated immune response in the progression of MASLD. Biomed Pharmacother 2024;173:116333. doi: 10.1016/j.biopha.2024.116333. PMID: 38479177

[36] Li H, Xia N. The multifaceted roles of B lymphocytes in metabolic dysfunction-associated steatotic liver disease. Front Immunol 2024;15:1447391. doi: 10.3389/fimmu.2024.1447391. PMID: 39372417

[37] Dhatchinamoorthy K, Colbert JD, Rock KL. Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation. Front Immunol 2021;12:636568. doi: 10.3389/fimmu.2021.636568. PMID: 33767702

[38] Sari G, Rock KL. Tumor immune evasion through loss of MHC class-I antigen presentation. Current opinion in immunology 2023;83:102329. doi: 10.1016/j.coi.2023.102329. PMID: 37130455

[39] Ahmed R, Ford ML, Sanz I. Regulation of T and B cell responses to chronic antigenic stimulation during Infection, autoimmunity and transplantation. Immunol Rev 2019;292(1):5-8. doi: 10.1111/imr.12836. PMID: 31883175

[40] Jee YM, Lee JY, Ryu T. Chronic Inflammation and Immune Dysregulation in Metabolic-Dysfunction-Associated Steatotic Liver Disease Progression: From Steatosis to Hepatocellular Carcinoma. Biomedicines 2025;13(5) doi: 10.3390/biomedicines13051260. PMID: 40427086

[41] Zhu G, Song Y, Lu Z, Yi Q, Xu R, Xie Y*, et al.* Machine learning models for predicting metabolic dysfunction-associated steatotic liver disease prevalence using basic demographic and clinical characteristics. J Transl Med 2025;23(1):381. doi: 10.1186/s12967-025-06387-5. PMID: 40155991

[42] Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci 2021;2(3):160. doi: 10.1007/s42979-021-00592-x. PMID: 33778771

[43] Feng G, Zheng KI, Li YY, Rios RS, Zhu PW, Pan XY*, et al.* Machine learning algorithm outperforms fibrosis markers in predicting significant fibrosis in biopsy-confirmed NAFLD. Journal of hepato-biliary-pancreatic sciences 2021;28(7):593-603. doi: 10.1002/jhbp.972. PMID: 33908180

[44] Feng G, He N, Xia HH, Mi M, Wang K, Byrne CD*, et al.* Machine learning algorithms based on proteomic data mining accurately predicting the recurrence of hepatitis B-related hepatocellular carcinoma. Journal of gastroenterology and hepatology 2022;37(11):2145-2153. doi: 10.1111/jgh.15940. PMID: 35816347

[45] Kong Q, Kong D, Li B, Peng W, Chen Z. Impact of Metabolic Dysfunction-Associated Fatty/Steatotic Liver Disease on Hepatocellular Carcinoma Incidence and Long-Term Prognosis Post-Liver Resection: A Systematic Review and Meta-Analysis. Academic radiology 2025 Published Online First: 20250121. doi: 10.1016/j.acra.2025.01.003. PMID: 39843280

[46] Li JW, Huang QR, Mo LG. CNPY4 is a potential promising prognostic-related biomarker and correlated with immune infiltrates in gliomas. Medicine (Baltimore) 2022;101(33):e30044. doi: 10.1097/md.0000000000030044. PMID: 35984129

[47] Anstee QM, Darlay R, Cockell S, Meroni M, Govaere O, Tiniakos D*, et al.* Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort(☆). Journal of hepatology 2020;73(3):505-515. doi: 10.1016/j.jhep.2020.04.003. PMID: 32298765

[48] Liu YL, Patman GL, Leathart JB, Piguet AC, Burt AD, Dufour JF*, et al.* Carriage of the PNPLA3 rs738409 C >G polymorphism confers an increased risk of non-alcoholic fatty liver disease associated hepatocellular carcinoma. Journal of hepatology 2014;61(1):75-81. doi: 10.1016/j.jhep.2014.02.030. PMID: 24607626

[49] Tan Y, Xiang W, Chen Y, Huang J, Sun D. Effect of hypoproteinemia on mortality of elderly male patients with chronic heart failure. Medicine (Baltimore) 2024;103(5):e37078. doi: 10.1097/md.0000000000037078. PMID: 38306508

**FIGURE LEGENDS**

**Figure 1. Flowchart of the study.**

First, we conducted MR analysis to identify six molecular biomarkers, and eight clinical biomarkers causally associated with metabolic dysfunction-associated steatotic liver disease (MASLD). Mediation analysis was performed to determine whether key clinical biomarkers could mediate the effects of molecular biomarkers (exposures) on MASLD (outcome). We found that total protein demonstrated a significant mediating effect in the relationship between HLA-A and MASLD. The association between total protein and MASLD was also validated externally in a hospital-based cohort. Second, we applied six machine-learning algorithms to develop and validate a novel noninvasive diagnostic model for MASLD based on the identified molecular biomarkers. Third, we explored the prognostic value of molecular biomarkers for the development of and poorer survival from hepatocellular carcinoma (HCC) by analyzing data from The Cancer Genome Atlas (TCGA). We also determined the prognostic value of key clinical biomarkers for all-cause and mortality and cause-specific mortality by analyzing prospective data from the National Health and Nutrition Examination Survey (NHANES).

**Figure 2. Causal molecular biomarkers related to MASLD**

1. Forest plot displaying the causal effect estimates for six molecular biomarkers significantly associated with MASLD based on the inverse variance weighted (IVW) method; (B) Leave-one-out sensitivity analysis plot for TOR1AIP1; (C) Scatter plot of SNP effects on TOR1AIP1 (exposure) and MASLD (outcome).

**Figure 3. Mendelian Randomization (MR) analysis revealing causal relationships between clinical biomarkers and MASLD.**

1. Forest plot displaying the causal effect estimates for eight clinical biomarkers significantly associated with MASLD. (B) Scatter plot showing the MR analysis for total protein as exposure and MASLD as the outcome. (C) Manhattan plot illustrating the distribution of SNPs associated with total protein across chromosomes. The -log10(P) values are plotted against chromosomal positions, highlighting genome-wide significant SNPs after linkage disequilibrium clumping.

**Figure 4. Mediation analysis exploring the role of total protein in the association between HLA-A and MASLD.**

1. Conceptual framework of the mediation analysis. Exposure (proteins) influences the outcome (MASLD) through direct and indirect pathways. The indirect effect is mediated by clinical biomarkers (mediators), while the direct effect bypasses the mediators. The total effect (β) is the sum of the direct effect (β₂) and the mediation effect (β₁ × β₂). (B) Proportion of the mediation and direct effects for the relationship between HLA-A, total protein, and MASLD. The indirect effect, mediated by total protein, accounts for 19.13% of the total association, whereas the direct effect contributes 80.87%.

**Figure 5. Performance evaluation and feature importance interpretation of machine learning models.**

1. Receiver operating characteristic (ROC) curves of six machine learning algorithms in the training set. (B) Feature importance analysis using SHapley Additive exPlanations (SHAP). (C) ROC curves of the six machine learning models in the independent validation set.

**Figure 6. Kaplan–Meier survival curves for all-cause and cardiovascular mortality stratified by serum total protein levels (<60 g/L vs. ≥60 g/L).**

(A) Kaplan–Meier survival curve for all-cause mortality stratified by serum total protein levels (<60 g/L vs. ≥60 g/L). (B) Kaplan–Meier survival curve for cardiovascular mortality stratified by serum total protein levels (<60 g/L vs. ≥60 g/L).