Cardiovascular Diabetology

RESEARCH Open Access

Association between atherogenicity indices and prediabetes: a 5-year retrospective cohort study in a general Chinese physical examination population

Xianli Qiu[1], Yong Han[2,3], Changchun Cao[4], Yuheng Liao[5,6,7] and Haofei Hu[5,6*]

Abstract

Background and Objective Atherogenicity indices have emerged as promising markers for cardiometabolic disorders, yet their relationship with prediabetes risk remains unclear. This study aimed to comprehensively evaluate the associations between six atherogenicity indices and prediabetes risk in a Chinese population, and explore the predictive value of these atherosclerotic parameters for prediabetes.

Methods This retrospective cohort study included 97,151 participants from 32 healthcare centers across China, with a median follow-up of 2.99 (2.13, 3.95) years. Six atherogenicity indices were calculated: Castelli's Risk Index-I (CRI-I), Castelli's Risk Index-II (CRI-II), Atherogenic Index of Plasma (AIP), Atherogenic Index (AI), Lipoprotein Combine Index (LCI), and Cholesterol Index (CHOLINDEX). To address the natural relationships between the atherogenicity indices and risk of prediabetes, we applied Cox proportional hazards regression with cubic spline functions and smooth curve fitting, using a recursive algorithm to calculate inflection points. Machine learning approach (XGBoost and Boruta methods) to address the high collinearity among indices and assess their relative importance, combined with time-dependent ROC analysis to evaluate the predictive performance at 3-, 4-, and 5-year follow-up.

Results During follow-up, 11,199 participants developed prediabetes (incidence rate: 3.71 per 100 person-years). Significant nonlinear associations were observed between all atherogenicity indices and prediabetes risk. Through Z-score standardization of atherogenicity indices and comprehensive Cox proportional hazards regression and advanced machine learning techniques, we identified AIP as the most significant predictor of prediabetes [HR = 1.057 (95% CI 1.035–1.080, $P < 0.0001$)], with LCI emerging as a secondary important marker [HR = 1.020 (95% CI 1.002–1.038, $P = 0.0267$)]. Our innovative XGBoost and Boruta analysis uniquely validated these findings, providing robust evidence of AIP and LCI's critical role in prediabetes risk assessment. Time-dependent ROC analysis further validated these findings, with LCI and AIP demonstrating comparable discrimination, with overlapping AUC ranges of 0.5952–0.6082. Notably, the combined indices model achieved enhanced predictive performance (AUC: 0.6753) compared to individual indices, suggesting the potential benefit of using multiple atherogenicity indices for prediabetes risk prediction.

*Correspondence:
Haofei Hu
huhaofei0319@126.com

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

Conclusion This study identifies statistically significant associations between atherogenicity indices and prediabetes risk, highlighting their nonlinear relationships and combined effects. While the predictive performance of these indices is modest (AUC 0.55–0.68), these findings may contribute to improved risk stratification when incorporated into comprehensive assessment strategies.

Keywords Atherogenicity indices, Prediabetes, Nonlinear relationship, Risk prediction, Machine learning, Cohort study

Introduction

Prediabetes has emerged as an increasingly prominent public health challenge worldwide. Recent epidemiological data indicate that while global diabetes cases exceeded 536.7 million in 2021, the number of individuals with prediabetes is projected to surpass 587 million by 2030 [1]. In China, according to the 2018 American Diabetes Association (ADA) diagnostic criteria, the age-standardized prevalence of prediabetes among adults has reached 35.2% [2]. Without timely intervention, approximately 5–10% of individuals with prediabetes progress to diabetes annually [3]. This alarming prevalence poses substantial public health challenges, as prediabetes not only increases the risk of progression to diabetes but also independently associates with various cardiovascular complications [4–6].

Early identification and intervention in prediabetes are crucial for preventing its progression to diabetes and associated complications [7, 8]. Recent evidence suggests that lipid metabolism disorders often precede glucose dysregulation, making lipid-related parameters potentially valuable early markers for prediabetes risk assessment [9, 10]. Among various lipid parameters, atherogenicity indices, which integrate multiple lipid components, have emerged as promising indicators for cardiovascular and metabolic disorders [11, 12].

Atherogenicity indices, as novel composite biomarkers, including Castelli's Risk Index-I (CRI-I), Castelli's Risk Index-II (CRI-II), Atherogenic Index of Plasma (AIP), Atherogenic Index (AI), Lipoprotein Combine Index (LCI), and Cholesterol Index (CHOLINDEX), have shown associations with various cardiometabolic conditions [13–16]. Compared to traditional single lipid parameters, these composite indices better reflect the balance between pro-atherogenic and anti-atherogenic lipoproteins [17, 18]. Recent studies have demonstrated that these indices are not only closely associated with cardiovascular events but also reflect insulin resistance status, serving as important predictors of glucose metabolism disorders [12, 19–23]. Notably, AIP has been confirmed as an independent predictor of prediabetes risk, showing superior predictive performance compared to traditional single lipid parameters [24, 25]. Furthermore, novel indices such as LCI have shown promising potential in identifying early atherosclerosis and predicting metabolic abnormalities [25–27].

However, current research on the relationship between atherogenicity indices and prediabetes has several limitations. Most studies have focused on individual lipid parameters, lacking comprehensive evaluations of multiple indices [28, 29]. Additionally, the predictive value of atherogenicity indices for prediabetes, particularly in Chinese populations, remains unclear. Recent studies have suggested that unconventional lipid parameters, such as AIP and LCI, may outperform traditional lipid parameters in predicting prediabetes [25, 29]. However, these studies are often limited by cross-sectional designs, small sample sizes, and insufficient consideration of confounding factors. Furthermore, the nonlinear relationships between atherogenicity indices and prediabetes risk, as well as the critical thresholds for these indices, remain poorly understood.

To address these gaps, we designed this retrospective cohort study to comprehensively evaluate and compare the predictive value of various atherogenicity indices (CRI-I, CRI-II, AIP, AI, LCI, and CHOLINDEX) for prediabetes in a Chinese health examination population. This study aims to: (1) investigate the associations between different atherogenicity indices and prediabetes risk; (2) compare the predictive performance of these indices; and (3) identify the optimal index or combination of indices for prediabetes risk assessment in the Chinese population. Additionally, the study will explore the nonlinear associations and critical thresholds of these indices, providing valuable insights for early risk stratification and personalized prevention strategies.

Methods

Study design

This investigation was structured as a retrospective cohort analysis examining the relationship between baseline atherogenicity indices (CRI-I, CRI-II, AIP, AI, LCI, and CHOLINDEX) and subsequent prediabetes development. We defined the atherogenicity indices at baseline as the exposure variables and incident prediabetes during follow-up as the outcome variable (coded as binary: 0 = non-prediabetes, 1 = prediabetes).

Data source

The research utilized health examination records from the Rich Healthcare Group database, accessible through the DATADRYAD repository (https://datadryad.org/st ash/dataset/doi:https://doi.org/10.5061/dryad.ft8750v). The Dryad repository grants academic researchers' noncommercial access to this dataset, with full rights for adaptation and derivative work creation, contingent upon appropriate source and authorship acknowledgment [30].

Study population

This retrospective cohort study utilized data from the Rich Healthcare Group database in China, which contains comprehensive health examination records from 32 healthcare centers across 11 major Chinese cities (Beijing, Suzhou, Nanjing, Shanghai, Changzhou, Shenzhen, Nantong, Chengdu, Hefei, Guangzhou, and Wuhan) between January 2010 and December 2016. To minimize selection bias, participants were consecutively and non-selectively enrolled from routine health examinations. The database includes standardized physical examinations, laboratory tests, and questionnaire data collected by trained healthcare professionals following standardized procedures.

The initial study population comprised 685,277 adults (age ≥ 18 years) who underwent routine health examinations. Through a comprehensive screening process, we systematically excluded individuals based on multiple criteria. Missing baseline data for fasting blood glucose (FPG), weight, gender, or height resulted in the removal of 135,317 cases. A follow-up period of less than 2 years led to the exclusion of 324,233 individuals. We removed cases with extreme body mass index (BMI) values (<15 kg/m^2 or >55 kg/m^2; n = 152) and those with undetermined diabetes status at follow-up (n = 6,630). Further refinement of the cohort involved eliminating subjects with baseline diabetes mellitus (DM) (n = 7,112), along with those who either self-reported DM or showed FPG ≥ 6.9 mmol/L during follow-up (n = 4,524). Additionally, we excluded individuals with baseline FPG ≥ 5.6 mmol/L (n = 23,121). Cases lacking complete lipid profiles, including total cholesterol (TC), triglyceride (TG), low-density lipoprotein cholesterol (LDL-C), and high-density lipoprotein cholesterol (HDL-C), were also removed (n = 83,923). Lastly, we excluded subjects with atherogenicity indices beyond three standard deviations from the mean (n = 3,114). After applying all exclusion criteria, the final analytical cohort consisted of 97,151 participants (detailed participant flow diagram available in Fig. 1).

Variables

Exposure variables

The primary exposure variables were six atherogenicity indices calculated from fasting lipid profiles:

CRI-I = Total Cholesterol/HDL-C [31].

CRI-II = LDL-C/HDL-C [31].

AIP = Log10(TG/HDL-C) [32].

AI = (TC—HDL-C)/HDL-C [33].

LCI = Total Cholesterol × Triglycerides × LDL-C/HDL-C [34].

CHOLINDEX = LDL-C-HDL-C (TG < 400 mg/dL), LDL-C-HDL-C + 1/5 of TG (TG ≥ 400 mg/dL) [35].

All lipid parameters were measured from fasting blood samples using standardized automated biochemical analyzers (Beckman 5800) within 24 h of collection. Quality control measures included regular calibration of instruments, use of standard operating procedures, and participation in external quality assessment programs [30].

Outcome variable

The primary outcome was incident prediabetes during the 5-year follow-up period, defined according to the ADA criteria (FPG 5.6–6.9 mmol/L) [36]. We censored participants at the time of pre-diabetes diagnosis or the last visit, whichever came first.

Covariates

Covariates included: (1) Demographic factors: age, gender; (2) Clinical measurements: BMI, systolic blood pressure (SBP), diastolic blood pressure (DBP); (3) Laboratory parameters: FPG, blood urea nitrogen (BUN), alanine aminotransferase (ALT), aspartate aminotransferase (AST), serum creatinine (Scr); (4) Lifestyle factors: smoking status (never/ever/current), drinking status (never/ever/current); (5) Medical history: family history of diabetes.

Trained healthcare professionals conducted comprehensive baseline assessments following standardized protocols. Height and weight were measured by trained staff using standardized equipment, with participants wearing light clothing and no shoes. Height was measured to 0.1 cm precision using a stadiometer, while weight was recorded to 0.1 kg using calibrated electronic scales. Body mass index was computed as weight (kg)/height(m^2). Blood pressure was measured using calibrated mercury sphygmomanometers after 5–10 min of rest in a seated position.

All laboratory tests were performed after at least 10 h of fasting. The assessed parameters included BUN, Scr, AST, FPG, and ALT.

Missing data processing

The dataset contained varying proportions of missing values across different variables. Physiological parameters showed minimal missing data: blood pressure measurements (n = 12, 0.012%), Scr (n = 1185, 1.220%), ALT (n = 365, 0.376%), and BUN (n = 2216, 2.281%). More substantial missing data were observed for liver function (AST: n = 56,386, 58.040%), and lifestyle factors (smoking and drinking status: both n = 70,542, 72.61%). To address potential bias and optimize data utilization, we

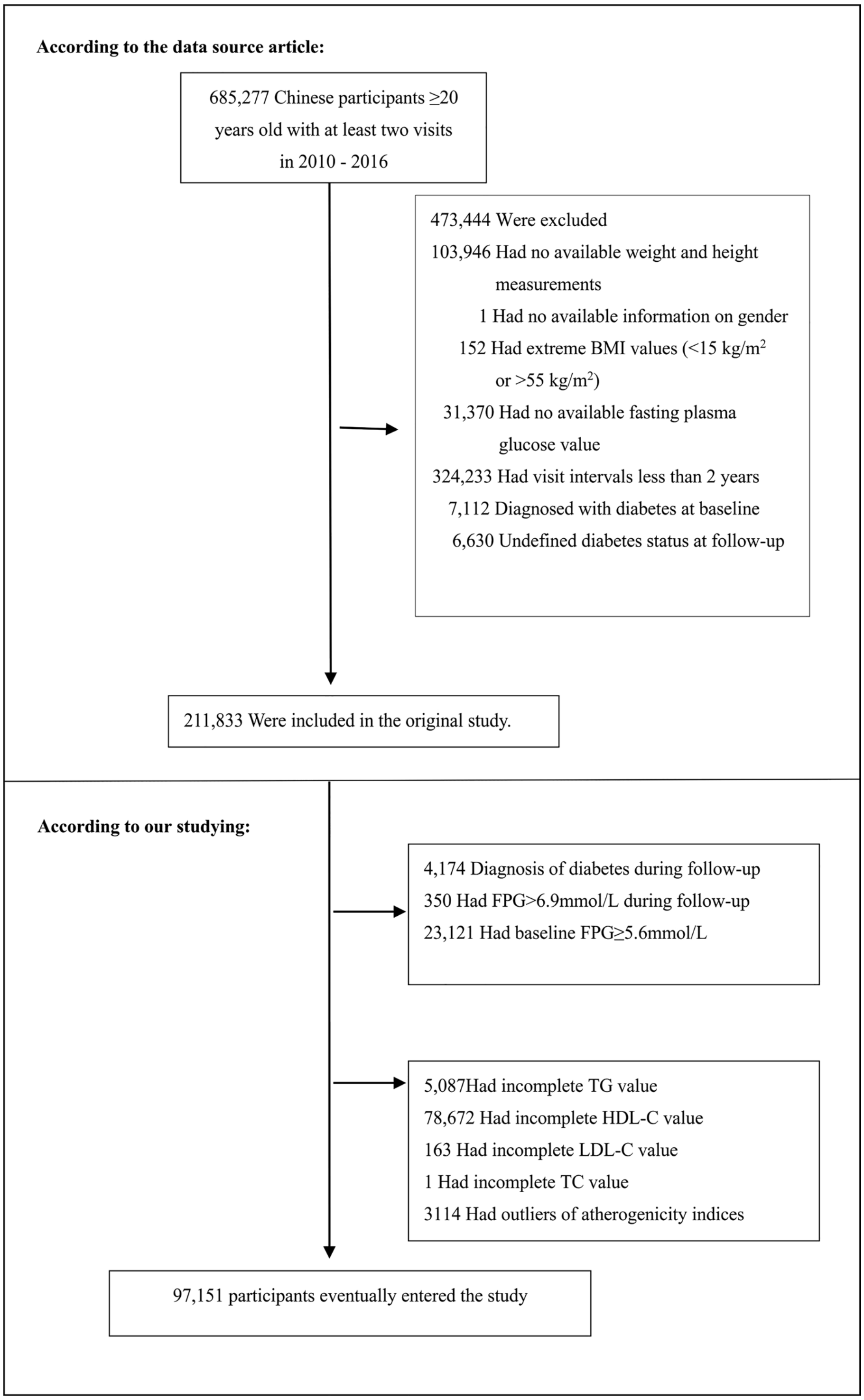
According to the data source article:
685,277 Chinese participants ≥20 years old with at least two visits in 2010 - 2016
473,444 Were excluded
103,946 Had no available weight and height measurements
1 Had no available information on gender
152 Had extreme BMI values (<15 kg/m2 or >55 kg/m2)
31,370 Had no available fasting plasma glucose value
324,233 Had visit intervals less than 2 years
7,112 Diagnosed with diabetes at baseline
6,630 Undefined diabetes status at follow-up
211,833 Were included in the original study.
According to our studying:
4,174 Diagnosis of diabetes during follow-up
350 Had FPG>6.9mmol/L during follow-up
23,121 Had baseline FPG≥5.6mmol/L
5,087Had incomplete TG value
78,672 Had incomplete HDL-C value
163 Had incomplete LDL-C value
1 Had incomplete TC value
3114 Had outliers of atherogenicity indices
97,151 participants eventually entered the study

Fig. 1 Study flow diagram of participant selection 211,833 participants were assessed for eligibility in the original study. We further excluded 114682participants. The final analysis included 97,151 subjects in the present study

implemented multiple imputation by chained equations [37]. This imputation included variables such as BMI, SBP, age, gender, Scr, DBP, ALT, BUN, TC, HDL-c, FPG, TG, family history of diabetes, LDL-c, AST, and drinking and smoking status. The missing data were analyzed based on the assumption that they were missing at random (MAR) [38].

Ethics statement

This study was approved by the Ethics Committee of the Rich Healthcare Group Review Board [30]. As this was a retrospective study using anonymized data, the requirement for informed consent was waived [30, 39]. All procedures were conducted in accordance with the Declaration of Helsinki and relevant regulations. Data security and participant privacy were protected through strict confidentiality protocols and data encryption measures. The study was reported following the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cohort studies [40].

Statistical analysis

Continuous variables were expressed as mean ± standard deviation (for normal distributions) or median with interquartile range [M (Q1, Q3)] for skewed data. Categorical variables were summarized as frequencies (n) or percentages (%). Differences between groups (the grouping variable was determined by the likelihood of progression to prediabetes) were compared using the χ^2 test for categorical variables, the student's t-test for normally distributed continuous variables, or the Mann–Whitney U test for non-normally distributed continuous variables. The Kolmogorov–Smirnov test was employed to assess the normality of continuous variables. In addition, the study cohort was stratified into four quartiles based on the atherogenicity indices. Statistical comparisons among groups utilized One-Way ANOVA for normally distributed parameters, Kruskal–Wallis H tests for non-normally distributed variables, and χ^2 test for categorical data analysis.

To analyze the association between atherogenicity indices (CRI-I, CRI-II, AIP, AI, LCI, and CHOLINDEX) and the development of prediabetes, we followed several analytical steps:

Step 1 Univariate and Multivariate Cox Proportional-hazards Regression.

We applied Z-score standardization to atherogenicity indices and then used univariate and multivariate Cox proportional-hazards regression models to evaluate their standardized impact on prediabetes.

Three models were constructed:

Model I No covariates were adjusted.

Model II Adjusted for demographic characteristics (age, gender, BMI) and clinical indicators (blood pressure, family history of diabetes, lifestyle factors).

Model III Further adjusted for metabolic parameters (FPG, BUN, Scr, ALT, AST) based on Model 2.

Covariate selection was guided by previous literature [25, 29, 30], and collinearity assessment. No variable was excluded from the multivariate analysis due to demonstrated collinearity (Table S1). The proportional hazards assumptions were validated using Schoenfeld residuals and log-minus-log plots. These models were designed to assess changes in effect estimates under different adjustment strategies and to evaluate the robustness of the results.

Step 2 Nonlinearity and Two-Piecewise Regression Analysis.

To explore potential nonlinearity in the relationship between atherogenicity indices and the development of prediabetes, a Cox proportional hazards regression model with cubic spline functions and smooth curve fitting (penalized spline method) was employed. If nonlinearity was detected, the inflection point was identified using a recursive algorithm, and a two-piecewise linear regression model was constructed on either side of the inflection point. For sensitivity analysis, the standard linear regression model was compared to the two-piecewise linear model, with the likelihood ratio test employed to determine the best fit for explaining the association between atherogenicity indices and the risk of prediabetes.

Step 3 Statistical Methods for Time-Dependent ROC Analysis and Machine Learning-Based Feature Importance Evaluation of Atherogenicity Indices in Prediabetes Risk Prediction.

Time-dependent ROC analysis

Time-dependent receiver operating characteristic (ROC) curve analysis was conducted to assess the predictive performance of each atherogenicity index, following the methodology described by Heagerty et al. [41]. The area under the curve (AUC) values were calculated for three time points (3-, 4-, and 5-year) to evaluate the discrimination ability of each index. For each atherogenicity index, optimal thresholds were determined using the Youden index method (maximum value of sensitivity + specificity—1) at each time point, as recommended by Pencina et al. [42].

Combined model development

The purpose of constructing our combined model is to explore the incremental predictive value of a comprehensive multi-index assessment for prediabetes, which helps to comprehensively understand the synergistic effects among different indices. Given the high correlation among the six atherogenicity indices, traditional Cox regression analysis was not suitable for combining these parameters to predict prediabetes risk due to multicollinearity issues. Therefore, we employed a machine learning approach to select and integrate these indices for developing a combined prediction model. The eXtreme Gradient Boosting (XGBoost) algorithm was chosen for its ability to handle correlated features and capture nonlinear relationships between predictors [43].

A combined model incorporating all six atherogenicity indices was developed to evaluate the potential improvement in predictive performance. The time-dependent AUC values for the combined model were calculated using the same methodology as individual indices to ensure direct comparability [41, 42].

Feature importance analysis

To enhance the robustness and reliability of our findings, we employed two complementary machine learning approaches—XGBoost algorithm and Boruta method—to independently quantify and cross-validate the relative importance of atherogenicity indices in predicting prediabetes risk [43]. The model was trained using fivefold cross-validation, and feature importance was calculated based on gain metrics, which measure each variable's contribution to the model. The importance scores were normalized to a 0–1 scale, allowing direct comparison between indices. This analysis provided insights into the hierarchical contribution of different atherogenicity indices, as visualized in the relative importance plot.

Statistical software

All statistical analyses were performed using R software (http://www.R-project.org, The R Foundation) and EmpowerStats software (X&Y Solutions, Inc; http://www.empowerstats.com). Statistical significance was defined as two-sided $P<0.05$.

Results

Characteristics of participants

In this retrospective cohort study, we analyzed data from 97,151 participants with a mean age of 42.8 ± 12.4 years. The study population demonstrated a near-equal gender distribution (51.2% male, n = 49,698; 48.8% female, n = 47,453). Metabolic parameters revealed a mean fasting plasma glucose of 4.8 ± 0.5 mmol/L, with a BMI of 23.0 ± 3.2 kg/m^2.

Based on progression to prediabetes during the study period, participants were categorized into non-progression (n = 85,952) and progression groups (n = 11,199). The group that progressed to prediabetes demonstrated significantly higher mean age (48.9 ± 13.7 vs 42.0 ± 12.0 years, $P<0.001$) and a greater proportion of male participants (61.4% vs 49.8%, $p = 0.001$) compared to those who did not progress. Statistical analysis revealed significant differences (all $P<0.001$) across multiple clinical parameters, with the progression group consistently showing elevated values in BMI, blood pressure, FPG, lipid profiles, atherogenicity indices, and hepatic and renal function markers. Furthermore, the progression group exhibited higher rates of smoking (20.4% vs 14.8%) and alcohol consumption (2.9% vs 1.7%), although family history distribution showed no significant variation between groups (Table 1).

Across the stratification of multiple metabolic risk indices (CRI-I, CRI-I, AIP, AI, LCI, and Cholesterol Index), consistent significant trends were observed: progressive increases in age, BMI, blood pressure, lipid levels (TC, TG, LDL-C), and liver enzyme levels, accompanied by decreasing HDL-C. The stratifications revealed systematic changes in gender composition and lifestyle factors, with higher proportions of males and increased smoking and drinking rates in higher-risk quartiles. These findings suggest a robust interconnection between metabolic risk indices and comprehensive health parameters, highlighting their potential as valuable tools for holistic health risk assessment (Table S2–S7).

Inter-relationships among atherogenicity indices and their distributions

Comprehensive correlation analysis demonstrated statistically significant associations among all cardiovascular indices examined ($P<0.001$). CRI-I exhibited a perfect positive correlation with AI (r = 1.00, $P<0.001$), indicating complete linear dependence between these parameters. The CRI-II demonstrated robust positive correlations with both AI (r = 0.94, $P<0.001$) and the cholesterol index (r = 0.91, $P<0.001$), suggesting strong physiological interconnections. The LCI showed substantial correlations with other cardiovascular parameters, with correlation coefficients ranging from 0.68 to 0.75 (all $P<0.001$). The AIP demonstrated variable associations with other indices, exhibiting its strongest correlation with AI (r = 0.67, $P<0.001$) and weakest with the cholesterol index (r = 0.44, $P<0.001$). Distribution analysis revealed predominantly normal distributions across most indices, with the exception of LCI, which displayed mild positive skewness (Fig. 2).

The incidence rate of pre-diabetes

During the median 2.99 (2.13,3.95)-year follow-up, 11,199 participants developed prediabetes (11.53%, 95% CI: 11.33%-11.73%), yielding an overall cumulative incidence rate of 3.71 per 100 person-years.

The incidence rate of prediabetes demonstrated a significant age-dependent increase across all age groups ($p<0.001$). The overall incidence rates progressively increased from 5.2% in individuals younger than 30 years

Table 1 The baseline characteristics of participants

Characteristics	Total population (N = 97,151)	Non-pre-diabetes (N = 85,952)	pre-diabetes (N = 11,199)	*P* value
Age, years	42.8 ± 12.4	42.0 ± 12.0	48.9 ± 13.7	< 0.001
Gender, n (%)				0.001
Male	49,698 (51.2%)	42,817 (49.8%)	6,881 (61.4%)	
Female	47,453 (48.8%)	43,135 (50.2%)	4,318 (38.6%)	
BMI (kg/m^2)	23.0 ± 3.2	22.9 ± 3.2	24.3 ± 3.2	< 0.001
SBP (mmHg)	117.9 ± 16.0	117.0 ± 15.6	125.0 ± 17.5	< 0.001
DBP (mmHg)	73.6 ± 10.7	73.1 ± 10.5	77.3 ± 11.3	< 0.001
FPG (mmol/L)	4.8 ± 0.5	4.8 ± 0.5	5.0 ± 0.4	< 0.001
TC (mmol/L)	4.7 ± 0.8	4.7 ± 0.8	4.9 ± 0.9	< 0.001
TG (mmol/L)	1.0 (0.7–1.5)	1.0 (0.7–1.5)	1.3 (0.9–1.8)	< 0.001
HDL-C (mmol/L)	1.4 ± 0.3	1.4 ± 0.3	1.4 ± 0.3	< 0.001
LDL-C (mmol/L)	2.7 ± 0.6	2.7 ± 0.6	2.8 ± 0.6	< 0.001
ALT (U/L)	17.3 (12.6–26.0)	17.0 (12.3–25.5)	20.0 (14.2–30.0)	< 0.001
AST (U/L)	22.0 (18.0–27.1)	21.9 (17.8–27.0)	23.1 (19.0–28.8)	< 0.001
BUN (mmol/L)	4.6 ± 1.2	4.6 ± 1.2	4.8 ± 1.2	< 0.001
Scr (umol/L)	69.7 ± 15.7	69.2 ± 15.6	73.0 ± 15.7	< 0.001
CRI-I	3.5 ± 0.8	3.5 ± 0.8	3.7 ± 0.8	< 0.001
CRI-II	2.0 ± 0.6	2.0 ± 0.6	2.1 ± 0.6	< 0.001
AIP	−0.1 ± 0.3	−0.1 ± 0.3	−0.0 ± 0.3	< 0.001
AI	2.5 ± 0.8	2.5 ± 0.8	2.7 ± 0.8	< 0.001
LCI	9.4 (5.3–17.1)	9.0 (5.2–16.4)	12.8 (7.2–22.0)	< 0.001
CHOLINDEX	1.3 ± 0.7	1.3 ± 0.6	1.5 ± 0.6	< 0.001
Smoking status, n (%)				0.001
Current smoker	14,986 (15.4%)	12,698 (14.8%)	2,288 (20.4%)	
Ever smoker	3,147 (3.2%)	2,708 (3.2%)	439 (3.9%)	
Never smoker	79,018 (81.3%)	70,546 (82.1%)	8,472 (75.6%)	
Drinking status, n (%)				0.001
Current drinker	1,815 (1.9%)	1,487 (1.7%)	328 (2.9%)	
Ever drinker	13,316 (13.7%)	11,490 (13.4%)	1,826 (16.3%)	
Never drinker	82,020 (84.4%)	72,975 (84.9%)	9,045 (80.8%)	
Family history of diabetes, n (%)				0.104
Yes	2,128 (2.2%)	1,859 (2.2%)	269 (2.4%)	
No	95,023 (97.8%)	84,093 (97.8%)	10,930 (97.6%)	

Values are n(%), mean ± SD or medians (quartiles)

BMI body mass index, *FPG* fasting plasma glucose, *DBP* diastolic blood pressure, *TC* total cholesterol, *SBP* systolic blood pressure, *TG* triglyceride, *ALT* alanine aminotransferase, *LDL-c* low-density lipid cholesterol, *AST* aspartate aminotransferase, *HDL-c* high-density lipoprotein cholesterol, *BUN* blood urea nitrogen, *Scr* serum creatinine, *CRI-I* Castelli's Risk Index I, *CRI-II* Castelli's Risk Index II, *AIP* Atherogenic Index of Plasma, *AI* Atherogenic Index, *LCI* Lipoprotein Combine Index, *CHOLINDEX* Cholesterol Index

to 26.2% in those aged over 70 years. A consistent gender disparity was observed across all age groups, with males showing higher incidence rates compared to females (male vs. female: 27.4% vs. 24.5% in > 70 age group, $p < 0.001$). The most pronounced increase in incidence rate occurred between the age groups of 40–50 and 50–60 years, where the overall rate increased from 11.5 to 17.0% (Δ = 5.5%). This gender-specific age-related pattern suggests that both age and male sex are risk factors for prediabetes development (Figure S1).

Our survival analysis revealed distinct patterns of association between atherogenicity indices and the risk of prediabetes development. For AIP, LCI, and CHOLINDEX, a clear dose-dependent relationship was observed: as biomarker levels increased from Q1 to Q4, the probability of maintaining prediabetes-free survival progressively decreased. In contrast, CRI-I, CRI-II, and AI exhibited a different pattern: rather than showing a linear increase in risk with higher levels, the Q3 group demonstrated the lowest prediabetes-free survival rates. This manifested as the highest risk of prediabetes development in the Q3 group, while Q1, Q2, and Q4 groups showed relatively lower incidence rates. The survival differences across all indicators became apparent after 2–3 years of follow-up and reached maximum separation at the 5-year endpoint (Figure S2).

Results from a multivariate Cox proportional-hazards regression model

Through Z-score standardization of atherogenicity indices and comprehensive Cox proportional hazards regression analysis, we systematically evaluated the differential impact of these indices on prediabetes risk. The AIP (Z-score) demonstrated a consistent positive association across three progressive models, with HR declining from 1.343 (95% CI 1.320–1.367, $P < 0.0001$) in Model I to 1.096 (95% CI 1.073–1.119, $P < 0.0001$) in Model II, and 1.057 (95% CI 1.035–1.080, $P < 0.0001$) in the final model. Despite the reduction in HR, AIP maintained statistically significant positive associations, highlighting its robust predictive potential for prediabetes risk. The LCI (Z-score) exhibited a similar pattern of stable positive correlation, with HR marginally decreasing from 1.264 (95% CI 1.245–1.283, $P < 0.0001$) in Model I to 1.055 in Model II and 1.020 in Model III, while retaining statistical significance. In contrast, Castelli's Risk Indices (CRI-I and CRI-II) demonstrated more pronounced variations: CRI-I (Z-score) decreased from 1.113 (95% CI 1.094–1.132, $P < 0.0001$) to 0.919 (Model II) and 0.878 (Model III), while CRI-II (Z-score) declined from 1.120 (95% CI 1.101–1.139, $P < 0.0001$) to 0.953 (Model II) and 0.916 (Model III), both maintaining statistical significance. The AI mirrored CRI1's trajectory exactly. The CHOLINDEX(Z-score) exhibited the most unique

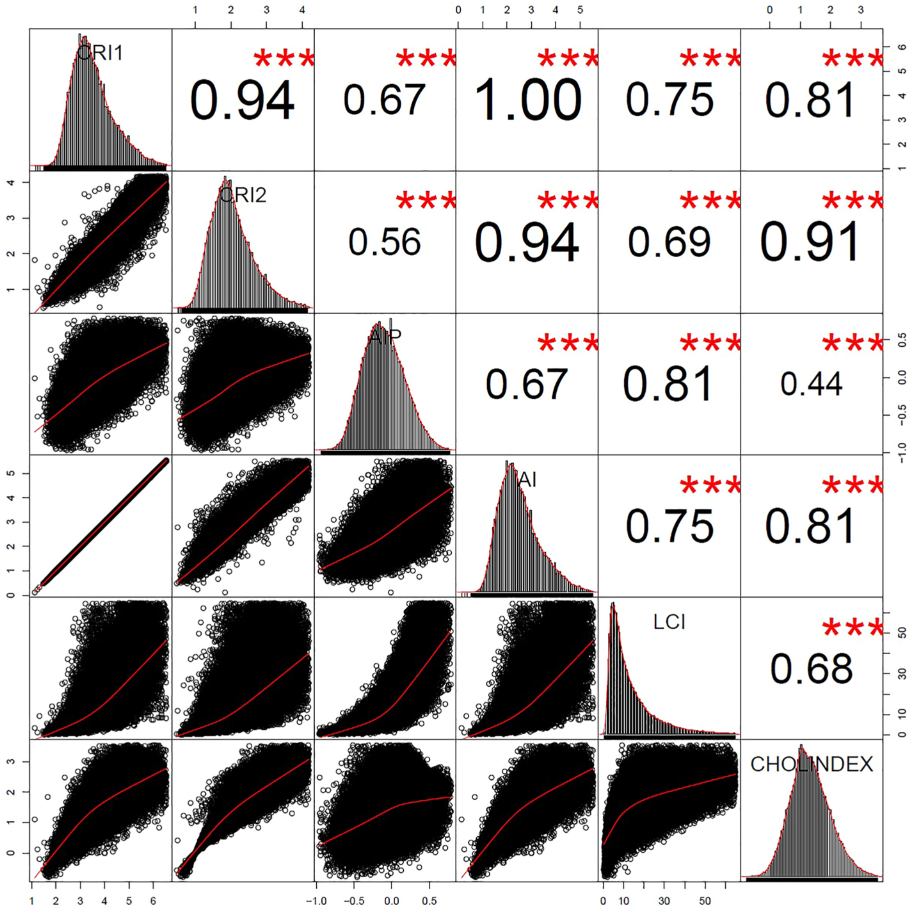

Fig. 2 Correlation matrix analysis of atherogenicity indices. Correlation matrix showing relationships between the six atherogenicity indices (CRI-I, CRI-II, AIP, AI, LCI, and CHOLINDEX) in the study cohort. The diagonal panels display histogram distributions for each variable. Lower triangular panels show scatter plots with fitted curves (red lines) illustrating bivariate relationships. Upper triangular panels present Pearson correlation coefficients, with asterisks (***) indicating statistical significance at $p < 0.001$. The correlation matrix reveals strong positive associations between most indices, with correlation coefficients ranging from 0.44 to 1.00

Table 2 Relationship between atherogenicity indices and the incident prediabetes in different models

Variables	Model I HR (95% CI)	*P* value	Model II HR (95% CI)	*P* value	Model III HR (95% CI)	*P* value
CRI-I(Z-score)	1.113(1.094, 1.132)	< 0.0001	0.919(0.901, 0.936)	< 0.0001	0.878(0.861, 0.895)	< 0.0001
CRI-II(Z-score)	1.120 (1.101, 1.139)	< 0.0001	0.953(0.935, 0.971)	< 0.0001	0.916(0.898, 0.933)	< 0.0001
AIP(Z-score)	1.343 (1.320, 1.367)	< 0.0001	1.096(1.073, 1.119)	< 0.0001	1.057(1.035, 1.080)	< 0.0001
AI(Z-score)	1.113 (1.094, 1.132)	< 0.0001	0.919(0.901, 0.936)	< 0.0001	0.878(0.861, 0.895)	< 0.0001
LCI(Z-score)	1.264 (1.245, 1.283)	< 0.0001	1.055(1.037, 1.074)	< 0.0001	1.020(1.002, 1.038)	0.0267
CHOLINDEX(Z-score)	1.195 (1.174, 1.216)	< 0.0001	1.015(0.996, 1.035)	0.1163	0.974(0.955, 0.992)	0.0063

Model I: we did not adjust other covariates

Model II: we adjust age, gender, BMI, SBP, DBP, smoking status, drinking, family history of diabetes

Model III: we adjust variables in Adjust I + FPG, ALT, AST, BUN, Scr

HR Hazard Ratio, *CI* Confidence Interval, *CRI-I* Castelli Risk Index I, *CRI-II* Castelli Risk Index II, *AIP* Atherogenic Index of Plasma, *AI* Atherogenic Index, *LCI* Lipoprotein Combine Index, *CHOLINDEX* Cholesterol Index

pattern, transitioning from 1.195 (95% CI 1.174–1.216, $P < 0.0001$) in the initial model to 1.015 ($P = 0.1163$, losing significance) in Model II, and ultimately to 0.974 (95% CI 0.955–0.992, $P = 0.0063$) in the final model (Table 2).

The non-linearity addressed by Cox proportional hazards regression model with cubic spline functions

The present study first systematically revealed the complex nonlinear associations between six atherogenicity indices and prediabetes risk (Fig. 3). The likelihood ratio tests (all $P < 0.001$) provided strong statistical evidence

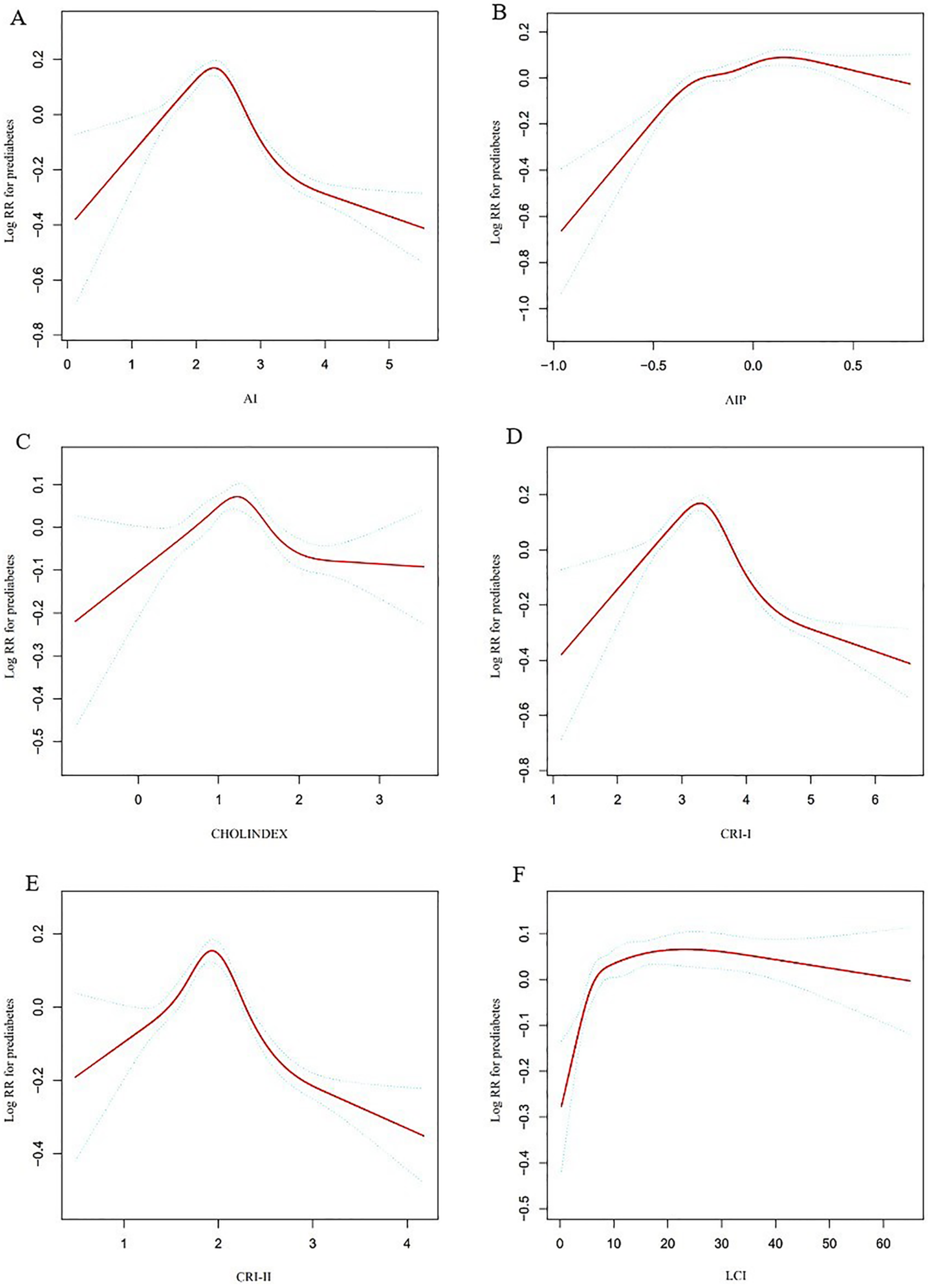

Fig. 3 Non-linear associations between atherogenicity indices and risk of prediabetes. Restricted cubic spline models showing the relationship between various atherogenicity indices and log-transformed relative risk (Log RR) for adverse outcomes. The solid red line represents the estimated Log RR, and the green shaded areas indicate the 95% confidence intervals. The reference values were set at the median of each index. **A** AI, **B** AIP, **C** CHOLINDEX, **D** CRI-I, **E** CRI-II, and **F** LCI. *P* values for non-linearity were <0.001 for all indices

for the nonlinear associations of all indices, significantly breaking through the limitations of traditional linear risk assessment models.

CRI-I and AI exhibited highly consistent nonlinear risk patterns. Both indices showed significant differences in predictive efficacy before and after different inflection points (CRI-I at 3.12, AI at 2.12). Below the inflection point, risk increased significantly (HR: 1.341, 95% CI 1.214–1.482, $P<0.0001$), suggesting that minor index changes in specific metabolic states might lead to a significant increase in prediabetes risk. Above the inflection point, risk rapidly declined (HR: 0.786, 95% CI 0.763–0.810, $P<0.0001$), implying potential protective metabolic compensation mechanisms.

CRI-II demonstrated a slightly different nonlinear association characteristic. Its inflection point was 1.884, with significantly increasing risk below the inflection point (HR: 1.348, 95% CI 1.218–1.492, $P<0.0001$) and more significantly declining risk above the inflection point (HR: 0.742, 95% CI 0.708–0.777, $P<0.0001$), reflecting a more dramatic risk transformation process.

AIP presented the most unique nonlinear association. At an inflection point near zero, AIP exhibited extremely high-risk prediction capabilities (HR: 1.673, 95% CI 1.453–1.927, $P<0.0001$); beyond the inflection point, predictive efficacy significantly disappeared (HR: 0.892, 95% CI 0.773–1.030, $P=0.1192$), revealing the fine-tuned regulatory mechanism of lipid metabolism in prediabetes development.

Table 3 Nonlinear associations between atherogenicity indices and prediabetes risk

Variables	Model I	Model II (Nonlinear effects)		
	HR (95% CI)	*P* value	Inflection point	Nonlinear effects HR (95% CI) *P* value
CRI-I	0.855 (0.836, 0.875)	< 0.0001	3.12	< Inflection point: 1.341 (1.214, 1.482), *P* < 0.0001 > Inflection point: 0.786 (0.763, 0.810), *P* < 0.0001 Likelihood ratio test: *P* < 0.001
CRI-II	0.862 (0.835, 0.890)	< 0.0001	1.884	< Inflection Point: 1.348 (1.218, 1.492), *P* < 0.0001 > Inflection point: 0.742 (0.708, 0.777), *P* < 0.0001 Likelihood ratio test: *P* < 0.001
AIP	1.226 (1.134, 1.324)	< 0.0001	0	< Inflection Point: 1.673 (1.453, 1.927), *P* < 0.0001 > Inflection point: 0.892 (0.773, 1.030), *P* = 0.1192 Likelihood Ratio Test: *P* < 0.001
AI	0.855 (0.836, 0.875)	< 0.0001	2.12	< Inflection point: 1.341 (1.214, 1.482), *P* < 0.0001 > Inflection point: 0.786 (0.763, 0.810), *P* < 0.0001 Likelihood ratio test: *P* < 0.001
LCI	1.002 (1.000, 1.003)	0.0267	3.625	< Inflection point: 1.213 (1.122, 1.312), *P* < 0.0001 > Inflection point: 1.001 (0.999, 1.003), *P* = 0.1762 Likelihood ratio test: *P* < 0.001
CHOLINDEX	0.960 (0.932, 0.988)	0.0063	1.11	< Inflection point: 1.179 (1.077, 1.290), *P* = 0.0003 > Inflection point: 0.894 (0.857, 0.932), *P* < 0.0001 Likelihood ratio test: *P* < 0.001

HR Hazard Ratio, *CI* Confidence Interval, *CRI-I* Castelli Risk Index I, *CRI-II* Castelli Risk Index II, *AIP* Atherogenic Index of Plasma, *AI* Atherogenic Index, *LCI* Lipoprotein Combine Index, *CHOLINDEX* Cholesterol Index

Adjusted Variables: age, gender, BMI, SBP, DBP, smoking status, drinking, family history of diabetes, FPG, ALT, AST, BUN, Scr

LCI showed a relatively moderate nonlinear association pattern. Its inflection point was 3.625, with significant risk below the inflection point (HR: 1.213, 95% CI 1.122–1.312, $P<0.0001$) and near-neutral risk above the inflection point (HR: 1.001, 95% CI 0.999–1.003, $P=0.1762$), reflecting the gradual transformation characteristics of metabolic indices.

CHOLINDEX displayed a typical n-shaped association pattern, with an inflection point at 1.11. Below the inflection point, risk progressively increased (HR: 1.179, 95% CI 1.077–1.290, $P=0.0003$); above the inflection point, risk rapidly declined (HR: 0.894, 95% CI 0.857–0.932, $P<0.0001$), reflecting the complex regulatory network of lipid metabolism (Table 3).

Comparative analysis of atherogenicity indices' predictive capability using XGBoost algorithm and Boruta method

The XGBoost algorithm and Boruta method were implemented to quantify the relative importance of each atherogenicity index in predicting prediabetes risk (Fig. 4). The result of XGBoost algorithm showed that among the included indices in our model, LCI showed the highest relative importance score (0.49), followed by AIP, CRI-I, CRI-II, and with moderate importance scores ranging from 0.10 to 0.15. CHOLINDEX exhibited the lowest relative importance (0.05).

Similarly, Boruta feature importance analysis revealed a hierarchical pattern of predictive significance among atherogenicity indices. AIP, CRI-II and LCI emerged as the most influential predictors, with importance scores approaching 70, significantly outperforming other markers. AI demonstrated moderate predictive potential, while CHOLINDEX and CRI-I exhibited relatively lower discriminative capabilities.

Although the results from XGBoost and Boruta methods were not entirely congruent, their convergent findings consistently underscore the potential significance of AIP and LCI as informative markers. These complementary analytical approaches suggest that AIP and LCI capture distinctive and potentially critical metabolic signatures uniquely associated with prediabetes risk, highlighting their robust predictive utility in our comprehensive model. However, it's important to note that these relative importance scores are model-dependent and require validation in independent cohorts before drawing definitive conclusions about their clinical significance.

Time-dependent receiver operating characteristic curves

Table 4 presents the best thresholds and areas under the time-dependent ROC curves for various atherogenicity

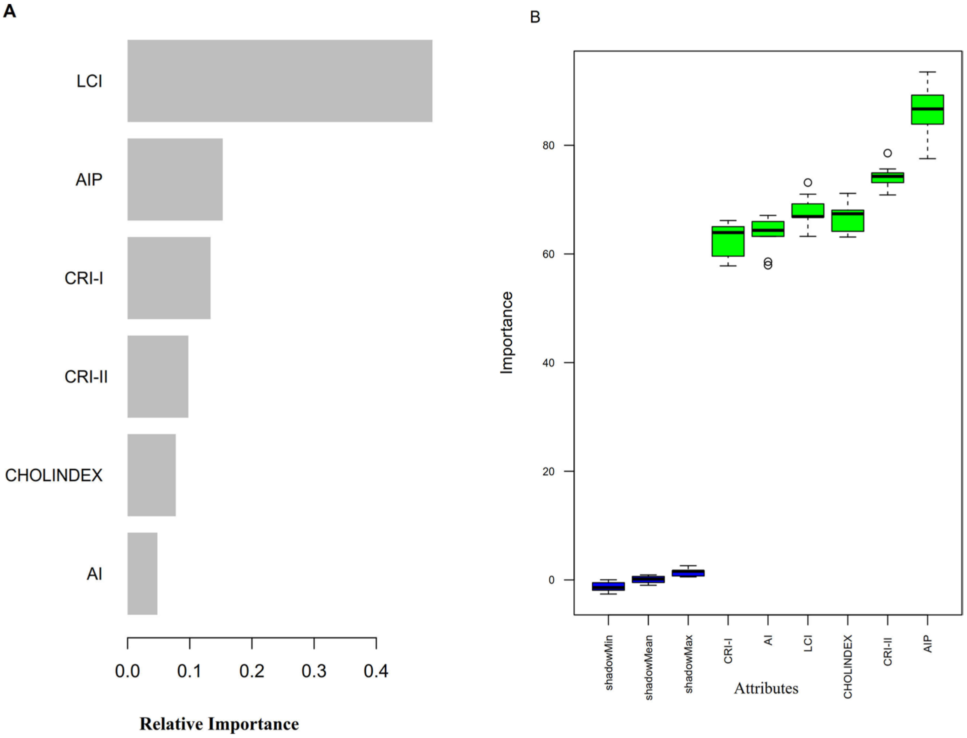

Fig. 4 Relative importance of different atherogenicity indices in predicting incident prediabetes. **A**. Relative importance of atherogenicity indices in predicting prediabetes risk, determined by XGBoost algorithm. The x-axis represents the normalized relative importance, highlighting the predictive potential of each index. **B**. Boruta feature importance analysis of atherogenicity indices, showing the statistical significance and predictive power of different markers. The y-axis represents the importance score, with higher values indicating greater predictive relevance for prediabetes risk

Table 4 Best threshold and areas under the time-dependent receiver operating characteristic curves for each atherogenicity indices predicting future prediabetes risk

Lipid parameter	3-Year AUC (best threshold)	4-Year AUC (best threshold)	5-Year AUC (best threshold)
CRI-I	0.5514 (3.0063)	0.5671 (3.2478)	0.5789 (3.2478)
CRI-II	0.5514 (1.7251)	0.5577 (1.8703)	0.5664 (1.8765)
AIP	0.5952 (−0.0712)	0.6014 (−0.0798)	0.6030 (−0.0875)
AI	0.5514 (2.0063)	0.5671 (2.2478)	0.5789 (2.2478)
LCI	0.5985 (9.3887)	0.6071 (10.3095)	0.6082 (10.2597)
CHOLINDEX	0.5597 (1.0650)	0.5638 (1.0850)	0.5671 (1.0950)
Combined indicators	0.6753	0.6590	0.6618

AUC area under the curve, *CRI-I* Castelli Risk Index I, *CRI-II* Castelli Risk Index II, *AIP* Atherogenic Index of Plasma, *AI* Atherogenic Index, *LCI* Lipoprotein Combine Index, *CHOLINDEX* Cholesterol Index

indices in predicting future prediabetes risk over three, four, and five years. The predictive accuracy remained stable across different time points, with minimal variation between 3-year, 4-year, and 5-year follow-up periods, suggesting robust long-term prognostic value of these indicators. Among the evaluated atherogenicity indices, the Combined Indicators exhibited the highest AUC values, with scores of 0.6753, 0.6590, and 0.6618 for the 3-year, 4-year, and 5-year timeframes, respectively, indicating modest predictive capability that, while statistically significant, reflects limited discriminative ability. The AIP showed AUC values of 0.5952, 0.6014, and 0.6030 across the respective years, while the LCI exhibited comparable AUC values of 0.5985, 0.6071, and 0.6082. These values, though statistically significant, indicate limited discriminative ability for clinical applications. Other indices, such as the CRI-I and II, displayed lower AUC values ranging from 0.5514 to 0.5789. The combined indicators model showed improved performance compared to individual indices, suggesting potential value in multifactorial assessments. However, even the combined model achieved only moderate discrimination ($AUC < 0.7$), indicating limitations for standalone clinical application (Fig. 5).

Discussion

This study investigated the association between six atherogenicity indices (CRI-I, CRI-II, AIP, AI, LCI, and CHOLINDEX) and the risk of prediabetes development. In this large-scale multicenter retrospective cohort study involving 97,151 participants from 32 healthcare centers with standardized data collection procedures over a 5-year follow-up period, we found significant nonlinear associations between all atherogenicity indices and

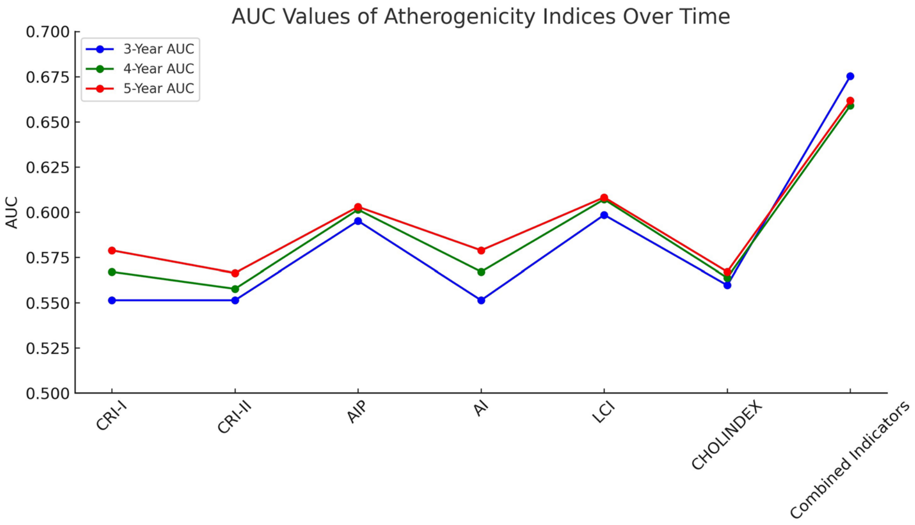

Fig. 5 Temporal comparison of predictive performance for atherogenicity indices. This figure shows the area under the receiver operating characteristic curve values for various atherogenicity indices assessed over different time periods (3-year, 4-year, and 5-year follow-up). The x-axis represents different lipid parameters including CRI-I, CRI-II, AIP, AI, LCI, CHOLINDEX, and Combined Indicators. The y-axis shows AUC values ranging from 0.500 to 0.700. The Combined Indicators demonstrated superior predictive performance compared to individual indices

prediabetes risk. Through comprehensive Cox regression and advanced machine learning techniques, we identified AIP as the most significant predictor of prediabetes, with LCI emerging as a secondary important marker. Our innovative XGBoost and Boruta analysis uniquely validated these findings, providing robust evidence of AIP and LCI's critical role in prediabetes risk assessment. Time-dependent ROC analysis further validated these findings, with LCI and AIP demonstrating comparable discrimination abilities (LCI AUC: 0.5985–0.6082 vs. AIP AUC: 0.5952–0.6030), while the combined indices model demonstrated enhanced predictive performance (AUC: 0.6753).

During a median follow-up of 2.99 (2.13,3.95) years, we observed an overall prediabetes incidence of 3.71/100 person-years (11.53%, 95% CI 11.33%-11.73%). This incidence rate is comparable to previous findings in Asian populations [44]. The growing obesity epidemic in the United States has been inextricably linked with a surge in rates of pre-diabetes. Pre-diabetes has been called “America’s largest healthcare epidemic,” and current estimates indicate that about 35% of adults in the United States have pre-diabetes, or approximately 79 million people [45]. Differences in the prevalence of prediabetes between East and West may be attributed to multiple factors: First, dietary patterns significantly influence glucose metabolism, with East Asian populations potentially benefiting from lower saturated fatty acid intake and higher dietary fiber consumption. Second, genetic background plays a crucial role, as studies have identified Asia-specific genetic polymorphisms that may affect insulin sensitivity and secretion [46]. Notably, we observed significant age-dependent increases and gender differences in incidence rates. The incidence increased progressively from 5.2% in those under 30 years to 26.2% in those over 70 years, with the most pronounced increase occurring between the 40–50 and 50–60 age groups (from 11.5 to 17.0%). This age-related pattern likely reflects the cumulative impact of metabolic disorders and organ function deterioration with advancing age [47].

Our multivariate Cox regression analysis revealed that AIP maintained significant positive associations with prediabetes risk after adjusting for multiple confounders, consistent with recent findings by Yang et al. [28] and Zou et al. [29]. The study by Zou et al. based on the CHARLS cohort supported the positive association between AIP and prediabetes progression [29]. Their study, including 2939 middle-aged and elderly prediabetic participants, found that cumulative AIP exposure significantly correlated with prediabetes progression. However, unlike our study, Zou et al. did not include CRI-I, CRI-II, AI, LCI, and CHOLINDEX as exposure variables, potentially limiting their comprehensive assessment of various atherogenicity indices. Additionally, their study focused primarily on middle-aged and elderly populations, while our study covered a broader age range (≥ 18 years), enhancing the generalizability of our findings across different age groups. Similarly, LCI demonstrated stable predictive value, potentially due to its characteristic of comprehensively reflecting multiple lipid components [48]. Despite the overall consistency with previous studies, some differences were observed. For instance, Li et al. observational study found that the relationship between CRI-I and CRI-II (Castelli’s Risk Index I and II) and prediabetes risk was not linear but showed a complex U-shaped curve [25]. Their study, including 100,309 participants, found that prediabetes risk was highest in the middle range (Q3 group) of CRI-I and CRI-II, with lower risks in both low (Q1 group) and high (Q4 group) ranges. While this non-linear relationship aligns with our observations, Li et al.'s study did not further explore the mechanisms underlying this non-linearity. In contrast, our study employed two-piecewise regression models and recursive algorithms to determine inflection points and validate the statistical significance of these non-linear relationships, providing more precise guidance for clinical risk stratification. In addition, the study by Li et al. used logistic regression analysis in analyzing the relationship between CRI-I and CRI-II and prediabetes, and therefore could not better explore the possible causal relationship between CRI-I and CRI-II and the risk of developing prediabetes compared to the Cox proportional risk regression model used in our study.

Through cubic spline analysis, we revealed non-linear associations between six atherosclerosis indices and prediabetes risk, enriching traditional linear risk assessment models. Despite unique characteristics of individual indices, a common pattern emerged: metabolic parameters in early stages influencing prediabetes risk, with potential metabolic compensation mechanisms observed in later stages. This risk transformation pattern reflects the complexity of lipid metabolism indices, indicating they do not follow simple linear relationships.

The non-linear patterns across CRI-I, AI, CRI-II, LCI, AIP, and CHOLINDEX suggest underlying metabolic regulatory mechanisms, potentially involving interactions between lipid metabolism, insulin sensitivity, and glucose homeostasis. Our findings demonstrate that metabolic risk is a dynamic process with critical inflection points, where subtle biochemical changes may impact disease risk. Different indices capture unique metabolic state information, reflecting the body's compensatory and adaptive characteristics [23, 25, 49–53]. From a clinical perspective, these findings provide a new lens for risk assessment. The study suggests that prediabetes risk evaluation should consider the non-linear features of indices, rather than relying solely on linear judgments. Clinicians can use these insights to more comprehensively assess individual metabolic risks and develop more targeted intervention strategies.

Our study innovatively employed machine learning approaches to address the high collinearity among atherogenicity indices in predicting prediabetes risk. Traditional Cox regression analysis struggled to handle the strong correlations among CRI-I, CRI-II, AIP, AI, LCI, and CHOLINDEX (correlation coefficients ranging from 0.44 to 1.00), which aligns with findings from Liu et al.[24]. While analyzing 1066 patients with non-alcoholic fatty liver disease from the NHANES database (2017–2020), Liu et al. encountered similar collinearity issues. However, unlike Liu et al.'s single-index approach, we successfully constructed a joint prediction model incorporating multiple collinear indices using the XGBoost algorithm and Boruta method.

Time-dependent ROC curve analysis demonstrated that the combined prediction model achieved AUC values of 0.6753, 0.6590, and 0.6618 for 3-year, 4-year, and 5-year follow-up periods, respectively, significantly outperforming individual indices. While Li et al. [25] also evaluated the predictive capability of atherogenicity indices for prediabetes risk in their cross-sectional studies of 100,309 Chinese adults, they focused on individual indices rather than a combined model [25]. Their study found that among single indices, LCI and AIP showed the strongest predictive performance for prediabetes. However, both studies were limited by their cross-sectional design and inability to assess long-term predictive value. In contrast, our study not only developed a machine learning-based combined prediction model that integrates multiple indices, but also confirmed its long-term stability through time-dependent ROC analysis, representing a significant methodological advancement in this field.

Feature importance analysis using XGBoost and Boruta methods revealed the relatively moderate performance of AIP and LCI in predicting prediabetes risk. Among the six atherogenicity indices, AIP and LCI demonstrated relatively strong predictive capabilities, a finding validated through multiple machine learning and statistical approaches. This aligns with Li et al. [25] retrospective study of 100,309 Chinese adults [25], which identified LCI and AIP's superior predictive capability through traditional statistical methods. XGBoost analysis showed that AIP and LCI were not only relatively prominent in univariate analysis but also maintained their predictive advantages in multivariate models that considered complex interactions and collinearity (correlation coefficients 0.44–1.00) among indices. The Boruta method further confirmed the importance of these two indices, providing a more robust feature selection result. Compared to traditional single statistical methods, the machine learning approach could simultaneously evaluate multiple indices and capture subtle interaction effects between them. This method overcomes the limitations of traditional analysis and provides more comprehensive and precise insights for clinical risk assessment. Our large-sample study (n = 97,151), by integrating XGBoost and Boruta methods, offers compelling scientific evidence for the role of AIP and LCI in prediabetes risk prediction.

The potential clinical applications of this study should be considered within the context of its limitations. First, our machine learning approach addressed the high collinearity among atherogenicity indices, potentially contributing to risk assessment methodology. Our combined prediction model showed modest improvements in discriminative ability compared to individual indices, but these values remain at the lower end of what would typically be considered clinically useful for prediction models. Second, this study uniquely quantified the relative importance of each atherogenicity index through XGBoost algorithm and Boruta method, identifying LCI and AIP as the relative important predictors, establishing new benchmarks for clinical risk stratification.

Drawing from our large-scale cohort study involving 97,151 participants, we believe these indices should be applied cautiously and precisely in clinical practice, particularly for individuals aged 40–60 with potential metabolic syndrome risk factors, including those with a family history of diabetes, who are overweight or obese, or exhibit lipid metabolism abnormalities. Clinicians should not treat these as standard universal screening tools, but rather integrate them with conventional clinical indicators as a precise, personalized risk assessment approach. For instance, when a patient's LCI or AIP index exceeds the predictive ROC curve critical value or approaches the inflection point of a non-linear curve, it signals the optimal timing for initiating proactive intervention strategies. Such intervention extends far beyond pharmaceutical treatment, emphasizing comprehensive lifestyle modifications, including rational dietary adjustments, regular physical exercise, weight management, and psychological regulation.

Our study demonstrates several notable strengths. First, in terms of study design, we included a large cohort of 97,151 participants with a 5-year follow-up period, which not only enhanced statistical power but also improved external validity. Second, we systematically evaluated the relationship between six atherogenicity indices and prediabetes risk for the first time. Regarding statistical analysis strategies, our study innovatively applied the machine learning algorithm to address the high collinearity among indices and assessed the long-term stability of prediction models through time-dependent ROC curve analysis. Furthermore, we employed two-piecewise regression models to explore non-linear relationships and identify critical inflection points, providing more precise guidance for clinical risk stratification. In handling missing data, we utilized multiple imputation methods to maximize the retention of valid information. Finally, by establishing a combined prediction model and quantifying the relative importance of each index, we not only overcame the limitations of traditional Cox regression but also provided a more comprehensive risk assessment tool for clinical practice. These methodological innovations enabled our study to more accurately reveal the complex relationships between atherogenicity indices and prediabetes risk.

Several limitations should be noted in this study. First, as an observational study, we can only establish associations between atherogenicity indices and prediabetes risk, rather than causal relationships. Second, although our study included data from multiple medical centers, the study population was limited to urban residents aged 18–75 years, potentially limiting the generalizability of our findings to these special populations. Third, our study primarily focused on Chinese Han population; considering the genetic and lifestyle differences among different ethnic groups, caution should be exercised when extrapolating these findings to other populations. Furthermore, while we adjusted for multiple known confounding factors, including age, gender, BMI, blood pressure, and lifestyle factors, there might still be unmeasured or unmeasurable confounding factors, such as genetic factors and detailed dietary information. Additionally, several methodological limitations should be acknowledged. Despite using advanced machine learning techniques, our models may still be subject to overfitting, particularly given the high collinearity among atherogenicity indices. The XGBoost algorithm, while powerful for handling complex data, can sometimes produce results that are difficult to interpret clinically. Furthermore, our prediction model lacks external validation in independent populations, which limits the generalizability of our findings. An important limitation of this study is the modest predictive performance of the atherogenicity indices, with AUC values ranging from approximately 0.55 to 0.68. In the context of clinical prediction models, these values indicate poor to moderate discriminative ability. Even our combined model achieved AUC values below 0.7, suggesting limited utility as standalone predictors in clinical practice. These findings indicate that atherogenicity indices might be more valuable when incorporated into more comprehensive risk assessment approaches that include additional established risk factors for prediabetes.

In conclusion, this study demonstrates statistically significant associations between atherogenicity indices and prediabetes risk, highlighting the importance of considering their nonlinear relationships and combined effects. While the predictive performance of these indices is modest, with AUC values at the lower end of clinical utility, these findings advance our understanding of prediabetes pathophysiology and may contribute to risk assessment when used in conjunction with other established predictors.

Abbreviations

CRI-I	Castelli's Risk Index-I
ADA	American Diabetes Association
AI	Atherogenic Index
AIP	Atherogenic Index of Plasma
AUC	Area under the curve
BMI	Body Mass Index
CHOLINDEX	Cholesterol Index

CI	Confidence Interval
CRI-II	Castelli's Risk Index-II
FPG	Fasting plasma glucose
HDL-C	High-density lipoprotein cholesterol
HR	Hazard ratio
IQR	Interquartile range
LCI	Lipoprotein Combine Index
LDL-C	Low-density lipoprotein cholesterol
ROC	Receiver operating characteristic
SD	Standard deviation
TC	Total cholesterol
TG	Triglycerides
XGBoost	Extreme gradient boosting
DM	Diabetes mellitus
SBP	Systolic blood pressure
AST	Aspartate aminotransferase
VIF	Variance inflation factor
BUN	Blood urea nitrogen
DBP	Diastolic blood pressure
Scr	Serum creatinine
Ref	Reference
ALT	Alanine aminotransferase
MAR	Missing-at-random

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12933-025-02768-8.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements

The data and methodology presented in this study are primarily based on the research conducted by Chen Y, Zhang XP, Yuan J, et al., which was published in BMJ Open in 2018 (Sep 28;8(9):e021768. doi: https://doi.org/10.1136/bmjopen-2018-021768) and titled "Association of body mass index and age with incident diabetes in Chinese adults: a population-based cohort study." In conducting this secondary analysis, the authors of the current study would like to acknowledge and express their sincere gratitude to the authors of the original publication for making their valuable data and research available.

Author contributions

XL Qiu conceived and designed the study, performed the data analysis, and drafted the manuscript. Y Han participated in data collection and interpretation. CC Cao contributed to the statistical analysis and methodology. YH Liao assisted with data processing and visualization. HF Hu supervised the entire project, provided critical revision of the manuscript for important intellectual content, and took responsibility for the integrity of the work. All authors read and approved the final manuscript.The final manuscript was read and approved by all authors.

Funding

This study was supported by the Discipline Construction Ability Enhancement Project of the Shenzhen Municipal Health Commission (SZXJ2017031), the Shenzhen Key Medical Discipline Construction Fund (SZXK009), and Sanming Project of Medicine in Shenzhen (SZSM202211013).

Availability of data and materials

The'DATADRYAD' database (https://datadryad.org/stash) offers access to the data.

Declarations

Ethical approval and consent to participate

The studies involving human participants were reviewed and approved by the Rich Healthcare Group Review Board. The data were retrospectively collected, and the information gathered is anonymized. Given the observational nature of the study, the Rich Healthcare Group Review Board waived the requirement for informed consent, as previously reported [30, 39].

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Fuwei Community Health Service Station, Shenzhen Baoan District Fuyong People's Hospital, Shenzhen 518000, Guangdong, China
[2]Department of Emergency, Shenzhen Second People's Hospital, Shenzhen 518000, Guangdong, China
[3]Department of Emergency, The First Affiliated Hospital of Shenzhen University, Shenzhen 518000, Guangdong, China
[4]Department of Rehabilitation, Longgang E.N.T Hospital & Shenzhen Key Laboratory of E.N.T, Institute of Ear Nose Throat (E.N.T), Shenzhen 518000, Guangdong, China
[5]Department of Nephrology, Shenzhen Second People's Hospital, No.3002 Sungang Road, Futian, Shenzhen 518000, Guangdong, China
[6]Department of Nephrology, The First Affiliated Hospital of Shenzhen University, Shenzhen 518000, Guangdong, China
[7]Department of Nephrology, Shenzhen University Health Science Center, Shenzhen 518000, Guangdong, China

Received: 29 January 2025 / Accepted: 1 May 2025
Published online: 21 May 2025

References

1. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, Stein C, Basit A, Chan J, Mbanya JC, et al. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Res Clin Pract. 2022;183:109119.
2. Li Y, Teng D, Shi X, Qin G, Qin Y, Quan H, Shi B, Sun H, Ba J, Chen B, et al. Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American diabetes association: national cross sectional study. BMJ-Brit Med J. 2020;369:m997.
3. Hostalek U. Global epidemiology of prediabetes–present and future perspectives. Clin Diabetes Endocrinol. 2019;5:5.
4. Wallace AS, Rooney MR, Fang M, Echouffo-Tcheugui JB, Grams M, Selvin E. Natural history of prediabetes and long-term risk of clinical outcomes in middle-aged adults: the atherosclerosis risk in communities (ARIC) study. Diabetes Care. 2023;46(2):e67–8.
5. Shang Y, Marseglia A, Fratiglioni L, Welmer AK, Wang R, Wang HX, Xu W. Natural history of prediabetes in older adults from a population-based longitudinal study. J Intern Med. 2019;286(3):326–40.
6. Schlesinger S, Neuenschwander M, Barbaresko J, Lang A, Maalmi H, Rathmann W, Roden M, Herder C. Prediabetes and risk of mortality, diabetes-related complications and comorbidities: umbrella review of meta-analyses of prospective studies. Diabetologia. 2022;65(2):275–85.
7. Galaviz KI, Weber MB, Suvada KB, Gujral UP, Wei J, Merchant R, Dharanendra S, Haw JS, Narayan K, Ali MK. Interventions for reversing prediabetes: a systematic review and meta-analysis. Am J Prev Med. 2022;62(4):614–25.
8. Chakkalakal RJ, Galaviz KI, Thirunavukkarasu S, Shah MK, Narayan K. Test and treat for prediabetes: a review of the health effects of prediabetes and the role of screening and prevention. Annu Rev Publ Health. 2024;45(1):151–67.
9. Vergès B. Pathophysiology of diabetic dyslipidaemia: where are we? Diabetologia. 2015;58(5):886–99.
10. Östgren CJ, Otten J, Festin K, Angerås O, Bergström G, Cederlund K, Engström G, Eriksson MJ, Eriksson M, Fall T, et al. Prevalence of atherosclerosis in individuals with prediabetes and diabetes compared to normoglycaemic individuals-a Swedish population-based study. Cardiovasc Diabetol. 2023;22(1):261.
11. Shi Y, Wen M. Sex-specific differences in the effect of the atherogenic index of plasma on prediabetes and diabetes in the NHANES 2011–2018 population. CARDIOVASC DIABETOL. 2023;22(1):19.
12. Yin B, Wu Z, Xia Y, Xiao S, Chen L, Li Y. Non-linear association of atherogenic index of plasma with insulin resistance and type 2 diabetes: a cross-sectional study. Cardiovasc Diabetol. 2023;22(1):157.

13. Millán J, Pintó X, Muñoz A, Zúñiga M, Rubiés-Prat J, Pallardo LF, Masana L, Mangas A, Hernández-Mijares A, González-Santos P, et al. Lipoprotein ratios: physiological significance and clinical usefulness in cardiovascular prevention. Vasc Health Risk Man. 2009;5:757–65.
14. Wu TT, Gao Y, Zheng YY, Ma YT, Xie X. Atherogenic index of plasma (AIP): a novel predictive indicator for the coronary artery disease in postmenopausal women. Lipids Health Dis. 2018;17(1):197.
15. Zakerkish M, Hoseinian A, Alipour M, Payami SP. The Association between Cardio-metabolic and hepatic indices and anthropometric measures with metabolically obesity phenotypes: a cross-sectional study from the Hoveyzeh Cohort Study. BMC Endocr Disord. 2023. https://doi.org/10.1186/s12902-023-01372-9.
16. Li YW, Kao TW, Chang PK, Chen WL, Wu LW. Atherogenic index of plasma as predictors for metabolic syndrome, hypertension and diabetes mellitus in Taiwan citizens: a 9-year longitudinal study. Sci Rep-UK. 2021;11(1):9900.
17. Zhu XW, Deng FY, Lei SF. Meta-analysis of atherogenic index of plasma and other lipid parameters in relation to risk of type 2 diabetes mellitus. Prim Care Diabetes. 2015;9(1):60–7.
18. Wen J, Zhong Y, Kuang C, Liao J, Chen Z, Yang Q. Lipoprotein ratios are better than conventional lipid parameters in predicting arterial stiffness in young men. J Clin Hypertens. 2017;19(8):771–6.
19. Babic N, Valjevac A, Zaciragic A, Avdagic N, Zukic S, Hasic S. The triglyceride/HDL ratio and triglyceride glucose index as predictors of glycemic control in patients with diabetes mellitus type 2. Med Arch. 2019;73(3):163–8.
20. Mo Z, Han Y, Cao C, Huang Q, Hu Y, Yu Z, Hu H. Association between non-high-density lipoprotein to high-density lipoprotein ratio and reversion to normoglycemia in people with impaired fasting glucose: a 5-year retrospective cohort study. Diabetol Metab Syndr. 2023;15(1):259.
21. Zeng Q, Zhong Q, Zhao L, An Z, Li S. Combined effect of triglyceride-glucose index and atherogenic index of plasma on cardiovascular disease: a national cohort study. Sci Rep-UK. 2024;14(1):31092.
22. Liu J, Zhao L, Zhang Y, Wang L, Feng Q, Cui J, Zhang W, Zheng J, Wang D, Zhao F, et al. A higher non-HDL-C/HDL-C ratio was associated with an increased risk of progression of nonculprit coronary lesion in patients with acute coronary syndrome undergoing percutaneous coronary intervention. Clin Cardiol. 2024;47(2):e24243.
23. Mahdavi-Roshan M, Mozafarihashjin M, Shoaibinobarian N, Ghorbani Z, Salari A, Savarrakhsh A, Hekmatdoost A. Evaluating the use of novel atherogenicity indices and insulin resistance surrogate markers in predicting the risk of coronary artery disease: a case‒control investigation with comparison to traditional biomarkers. Lipids Health Dis. 2022. https://doi.org/10.1186/s12944-022-01732-9.
24. Liu J, Fu Q, Su R, Liu R, Wu S, Li K, Wu J, Zhang N. Association between non-traditional lipid parameters and the risk of type 2 diabetes and prediabetes in patients with nonalcoholic fatty liver disease: from the national health and nutrition examination survey 2017–2020. Front Endocrinol. 2024;15:1460280.
25. Li M, Zhang W, Zhang M, Li L, Wang D, Yan G, Qiao Y, Tang C. Nonlinear relationship between untraditional lipid parameters and the risk of prediabetes: a large retrospective study based on Chinese adults. Cardiovasc Diabetol. 2024. https://doi.org/10.1186/s12933-023-02103-z.
26. Zhou Y, Yang G, Qu C, Chen J, Qian Y, Yuan L, Mao T, Xu Y, Li X, Zhen S, et al. Predictive performance of lipid parameters in identifying undiagnosed diabetes and prediabetes: a cross-sectional study in eastern China. BMC Endocr Disord. 2022;22(1):76.
27. Yu S, Yan L, Yan J, Sun X, Fan M, Liu H, Li Y, Guo M. The predictive value of nontraditional lipid parameters for intracranial and extracranial atherosclerotic stenosis: a hospital-based observational study in China. Lipids Health Dis. 2023;22(1):16.
28. Yang H, Kuang M, Yang R, Xie G, Sheng G, Zou Y. Evaluation of the role of atherogenic index of plasma in the reversion from prediabetes to normoglycemia or progression to diabetes: a multi-center retrospective cohort study. Cardiovasc Diabetol. 2024;23(1):17.
29. Zou Y, Lu S, Li D, Huang X, Wang C, Xie G, Duan L, Yang H. Exposure of cumulative atherogenic index of plasma and the development of prediabetes in middle-aged and elderly individuals: evidence from the CHARLS cohort study. Cardiovasc Diabetol. 2024;23(1):355.
30. Chen Y, Zhang XP, Yuan J, Cai B, Wang XL, Wu XL, Zhang YH, Zhang XY, Yin T, Zhu XH, et al. Association of body mass index and age with incident diabetes in Chinese adults: a population-based cohort study. BMJ Open. 2018;8(9):e21768.
31. Castelli WP, Abbott RD, McNamara PM. Summary estimates of cholesterol used to predict coronary heart disease. Circulation. 1983;67(4):730–4.
32. Fernández-Macías JC, Ochoa-Martínez AC, Varela-Silva JA, Pérez-Maldonado IN. Atherogenic index of plasma: novel predictive biomarker for cardiovascular illnesses. Arch Med Res. 2019;50(5):285–94.
33. Olamoyegun MA, Oluyombo R, Asaolu SO. Evaluation of dyslipidemia, lipid ratios, and atherogenic index as cardiovascular risk factors among semi-urban dwellers in Nigeria. Ann Afr Med. 2016;15(4):194–9.
34. Si Y, Liu J, Han C, Wang R, Liu T, Sun L. The correlation of retinol-binding protein-4 and lipoprotein combine index with the prevalence and diagnosis of acute coronary syndrome. Heart Vessels. 2020;35(11):1494–501.
35. Akpınar O, Bozkurt A, Acartürk E, Seydaoğlu G. A new index (CHOLINDEX) in detecting coronary artery disease risk. Anadolu Kardiyol Derg. 2013;13(4):315–9.
36. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2018. *Diabetes Care* 2018, 41(Suppl 1):S13-S27.
37. Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. Can Med Assoc J. 2012;184(11):1265–9.
38. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377–99.
39. Geleris J, Sun Y, Platt J, Zucker J, Baldwin M, Hripcsak G, Labella A, Manson DK, Kubin C, Barr RG, et al. Observational study of hydroxychloroquine in hospitalized patients with covid-19. New Engl J Med. 2020;382(25):2411–8.
40. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Plos Med. 2007;4(10):e297.
41. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005;61(1):92–105.
42. Pencina MJ, D'Agostino RBS, D'Agostino RBJ, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. STAT MED. 2008;27(2):157–72.
43. Chen T, Guestrin C: XGBoost: A scalable tree boosting system. *ACM* 2016.
44. Jayawardena R, Ranasinghe P, Byrne NM, Soares MJ, Katulanda P, Hills AP. Prevalence and trends of the diabetes epidemic in South Asia: a systematic review and meta-analysis. BMC Public Health. 2012;12:380.
45. Garber AJ, Handelsman Y, Einhorn D, Bergman DA, Bloomgarden ZT, Fonseca V, Garvey WT, Gavin JR, Grunberger G, Horton ES, et al. Diagnosis and management of prediabetes in the continuum of hyperglycemia: when do the risks of diabetes begin? A consensus statement from the American college of endocrinology and the American association of clinical endocrinologists. Endocr Pract. 2008;14(7):933–46.
46. Qiao J, Wu Y, Zhang S, Xu Y, Zhang J, Zeng P, Wang T. Evaluating significance of European-associated index SNPs in the East Asian population for 31 complex phenotypes. BMC Genomics. 2023;24(1):324.
47. Wang Z, Zhu H, Xiong W. Metabolism and metabolomics in senescence, aging, and age-related diseases: a multiscale perspective. Front Med. 2025. https://doi.org/10.1007/s11684-024-1116-0.
48. Samuel VT, Shulman GI. The pathogenesis of insulin resistance: integrating signaling pathways and substrate flux. J Clin Invest. 2016;126(1):12–22.
49. Han Y, He X, Jin M, Sun H, Kwon T. Lipophagy: a potential therapeutic target for nonalcoholic and alcoholic fatty liver disease. Biochem Bioph Res Commun. 2023;672:36–44.
50. Yang C, Liu Z, Zhang L, Gao J. The association between blood glucose levels and lipids in the general adult population: results from NHANES (2005–2016). J Health Popul Nutr. 2024;43(1):163.
51. Zheng X, Zhang X, Han Y, Hu H, Cao C. Nonlinear relationship between atherogenic index of plasma and the risk of prediabetes: a retrospective study based on Chinese adults. Cardiovasc Diabetol. 2023;22(1):205.
52. He J, Chen L. Perspective from NHANES data: synergistic effects of visceral adiposity index and lipid accumulation products on diabetes risk. Sci Rep-UK. 2025;15(1):258.
53. Qiu J, Huang X, Kuang M, Yang R, Li J, Sheng G, Zou Y. Lipoprotein combine index as a better marker for NAFLD identification than traditional lipid parameters. Diabet Metab Synd Ob. 2024;17:2583–95.

Publisher's Note