

An Interpretable Machine Learning Framework for Predicting Disease Progression in Early-Stage Cardiovascular-Kidney-Metabolic(CKM) Syndrome: A Prospective Cohort Study

Xuhui Tang

Hunan University of Chinese Medicine

Yu Liu

The Second People's Hospital of Hunan Province (Brain Hospital of Hunan Province)

E Zhu

The Second People's Hospital of Hunan Province (Brain Hospital of Hunan Province)

Xiaosong Huang

Huangxs-0819@163.com

Hunan University of Chinese Medicine

Research Article

Keywords: cardiovascular-kidney-metabolic syndrome, machine learning, Random Forest, disease progression, SHAP, CHARLS, calibration

Posted Date: August 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10440906/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Cardiovascular Disorders on October 12th, 2026. See the published version at <https://doi.org/10.1186/s12872-026-06780-y>.

Abstract

Objective: Cardiovascular-kidney-metabolic (CKM) syndrome constitutes a continuum of metabolic, renal, and cardiovascular dysfunction, yet practical tools for stratifying progression risk in its early, modifiable stages remain absent. We aimed to develop an interpretable machine learning (ML) framework to estimate stage-advancement risk among individuals with early-stage CKM syndrome, emphasizing calibration stability and subgroup heterogeneity.

Methods: This prospective cohort study used data from the China Health and Retirement Longitudinal Study (CHARLS). A total of 3,773 participants aged ≥ 45 years with baseline (2011) CKM stages 0–2 and complete follow-up through 2015 were included. Progression was defined as an increase of at least one CKM stage. Six ML models were evaluated through a three-stage selection process that sequentially prioritized discrimination, calibration stability (calibration slope and Integrated Calibration Index [ICI]), and clinical utility (Decision Curve Analysis). SHapley Additive exPlanations (SHAP) were applied to interpret model predictions. Subgroup analyses were stratified by baseline CKM stage.

Results: Over the 4-year period, 1,560 participants (41.3%) progressed. This rate substantially exceeded the 15–34% stage-transition frequencies documented in Western cohorts, reflecting marked dynamic instability of early CKM in Chinese middle-aged and older adults. LASSO retained eight predictors: age, BMI, SBP, HbA1c, eGFR, baseline CKM stage, sex, and education. CatBoost yielded the highest discrimination (test AUC = 0.772) but was poorly calibrated (calibration slope = 1.305; ICI = 0.060). Random Forest (RF) was selected as the optimal model given its calibration stability (slope = 1.056; ICI = 0.026) and highest net benefit. Global SHAP analysis identified baseline CKM stage, HbA1c, and SBP as the leading predictors and uncovered a non-monotonic inverse stage effect. Subgroup analyses by baseline stage revealed a predictive ceiling effect in Stages 0–1 (AUCs ~ 0.60), attributable to progression rates exceeding 70%, while calibration remained acceptable across all stages.

Conclusions: We present a well-calibrated, interpretable RF framework for risk stratification. The ceiling effect observed in early CKM stages indicates that individuals in Stages 0–1 may benefit more from immediate preventive care than from further conventional risk stratification.

1. Introduction

Cardiovascular-kidney-metabolic (CKM) syndrome, formally defined by the American Heart Association (AHA) in October 2023, represents a paradigm shift in understanding the interplay among metabolic risk factors, chronic kidney disease (CKD), and cardiovascular disease (CVD) as a unified systemic disorder[1]. The AHA proposed a five-stage framework (Stage 0–4) to characterize the progressive nature of CKM syndrome. Early stages (0–2) represent a reversible "golden window" for intervention, during which underlying pathophysiological changes—such as insulin resistance, subclinical inflammation, endothelial dysfunction, and early renal hyperfiltration—have not yet resulted in irreversible structural organ damage[2,3]. Chronic low-grade inflammation and lipid toxicity are recognized as common

pathophysiological pathways linking metabolic dysfunction to multisystem damage in the heart, kidneys, and vasculature[2].

Although early intervention is important, the early-stage CKM population is large and heterogeneous. U.S. epidemiological data indicate that approximately 90% of adults meet criteria for at least Stage 1 CKM, with 49% at Stage 2 and 15% at advanced stages[4]. In China, CKM prevalence rose from 77.1% in 2010 to 83.7% in 2019, corresponding to an estimated 875 million affected individuals; advanced stages increased by 60.3% over this period[5]. The Kailuan study demonstrated a stage-dependent increase in all-cause mortality: hazard ratios rose across Stages 1 to 4, reaching 3.73 for Stage 4 (reference: Stage 0)[6]. CKM progression is dynamic—in a Korean nationwide cohort, 15.3% of participants progressed over 1–2 years, and progression was associated with a 7.1% higher risk of adverse clinical outcomes[7,8]. Trajectories of disease progression and the associated risk predictors within early-stage CKM, however, remain incompletely characterized.

In the present study, over 4-year follow-up, 1,560 participants (41.3%) experienced CKM progression—a rate substantially exceeding the 34% progression rate reported among stage 1 individuals in Western cohorts[9]. This discrepancy reflects the marked dynamic instability of early CKM in Chinese middle-aged and older adults and prompts the need for reliable risk stratification tools[9,10].

Current risk stratification tools focus on predicting 10-year CVD risk in later stages but lack the granularity to identify which individuals in early stages are most likely to advance rapidly. Traditional statistical approaches have limited capacity to capture complex, non-linear interactions among the multidimensional risk factors characterizing CKM syndrome.

Machine learning (ML) offers an alternative by automatically modeling high-dimensional, non-linear associations[11]. Combined with SHapley Additive exPlanations (SHAP)—a game-theoretic approach providing the unique additive feature importance measure that satisfies local accuracy, missingness, and consistency[12]—ML predictions can be decomposed to enhance clinical transparency.

Current ML-based prediction studies frequently face two major limitations. First, the prevailing "AUC-chasing" paradigm prioritizes discrimination over calibration stability. Calibration has been labeled the "Achilles heel" of predictive analytics[13], as poorly calibrated algorithms can mislead clinical decision-making. The Hosmer-Lemeshow (H-L) test, although widely used, exhibits excessive sensitivity to sample size—in cohorts of 50,000 patients, a miscalibration of merely 0.4% yields significant H-L test results in all models[14]. Modern standards endorse calibration metrics insensitive to sample size, such as calibration slope, calibration intercept, and the Integrated Calibration Index (ICI), which quantify overfitting and the average magnitude of miscalibration, respectively[15,16]. Second, most prediction models assume uniform performance across populations, overlooking subgroup heterogeneity, including stage-specific prediction challenges. The CKM framework explicitly acknowledges that social determinants of health and demographic factors contribute to heterogeneity within CKM syndrome[17].

Using data from the China Health and Retirement Longitudinal Study (CHARLS), a nationally representative prospective cohort of middle-aged and older adults[18], we aimed to develop an explainable ML framework

to predict 4-year disease progression among individuals with early-stage CKM syndrome. A hierarchical model selection process—prioritizing calibration stability over marginal gains in AUC—was integrated with comprehensive subgroup analyses to provide a clinically actionable, equitable precision risk stratification tool.

2. Methods

2.1 Study Design and Population

This prospective cohort study used data from CHARLS, a nationally representative longitudinal survey of Chinese adults aged ≥ 45 years. The CHARLS study design, multistage stratified sampling strategy, and data collection procedures are described elsewhere[18]. CHARLS used a four-stage stratified probability-proportional-to-size sampling design to recruit participants from 150 counties/districts across 28 provinces. The baseline survey was conducted in 2011, with follow-up assessments in 2013, 2015, 2018, and 2020. Ethical approval was obtained from the Biomedical Ethics Review Committee of Peking University (IRB00001052-11015), and all participants provided written informed consent.

For this analysis, we used baseline data from 2011 and follow-up data from 2015 to define a 4-year disease progression window. Of 17708 respondents in the 2011 baseline survey, we excluded those with: (1) missing age or age < 45 years ($n = 777$); (2) a self-reported cancer history at baseline ($n = 391$; to avoid competing mortality and metabolic derangements related to malignancy); (3) missing data for CKM staging ($n = 11,110$); (4) baseline CKM Stage 3 or 4 ($n = 1,509$; to retain early-stage individuals); and (5) missing covariates ($n = 148$). After exclusions, 3773 participants with early-stage CKM (Stages 0–2) at baseline constituted the analytical sample. This study is reported in accordance with the TRIPOD Statement[19], with the completed checklist provided as a supplementary file.

Figure 1 Flowchart of the study population selection from the CHARLS cohort

Note: CHARLS, China Health and Retirement Longitudinal Study; CKM, cardiovascular-kidney-metabolic syndrome. Participants were excluded sequentially based on predefined criteria. The final analytic cohort included 3773 individuals with complete baseline and follow-up data.

2.2 CKM Syndrome Definition and Staging

CKM syndrome stages were defined based on the 2023 AHA criteria[20], adapted for the data available in CHARLS. The staging was operationalized as follows: Stage 0: No CKM risk factors. Participants did not meet criteria for overweight/obesity ($\text{BMI} < 25 \text{ kg/m}^2$ and waist circumference $< 90 \text{ cm}$ in males, $< 85 \text{ cm}$ in females), hypertension ($\text{SBP} < 130 \text{ mmHg}$, $\text{DBP} < 80 \text{ mmHg}$, no antihypertensive medication), dyslipidemia ($\text{TG} < 150 \text{ mg/dL}$, $\text{TC} < 200 \text{ mg/dL}$, no lipid-lowering medication), hyperglycemia ($\text{FPG} < 100 \text{ mg/dL}$, $\text{HbA1c} < 5.7\%$, no glucose-lowering medication), or CKD ($\text{eGFR} \geq 60 \text{ mL/min/1.73 m}^2$). Stage 1: Excess or dysfunctional adiposity ($\text{BMI} \geq 25 \text{ kg/m}^2$ or elevated waist circumference) without other metabolic abnormalities. Stage 2: Metabolic risk factors, including hypertension, hypertriglyceridemia, hyperglycemia, or moderate-risk CKD ($\text{eGFR} 45\text{--}59 \text{ mL/min/1.73 m}^2$). Stage 3: Subclinical CVD or high-risk CKD ($\text{eGFR} < 45$

mL/min/1.73 m²). Given the absence of coronary artery calcium (CAC) and urine albumin-to-creatinine ratio (UACR) data in CHARLS, Stage 3 was defined primarily by eGFR <45. Stage 4: Clinical CVD events (myocardial infarction, angina, stroke, or heart failure). The primary outcome was CKM progression, defined as an increase of at least one CKM stage from baseline (2011) to follow-up (2015).

2.3 Data Collection and Measurements

Anthropometric measurements were performed by trained staff. Blood pressure was measured three times on the right arm after 5 minutes of rest using an automated sphygmomanometer (Omron HEM-7200); the mean was used for analysis. Overnight fasting blood samples were collected, stored at -20°C, and analyzed at the China CDC laboratory. Biomarkers included FPG, HbA1c, lipid profiles, serum creatinine, hs-CRP, and uric acid. eGFR was calculated with the CKD-EPI 2021 equation[21]. Standardized questionnaires collected sociodemographic data, health behaviors, and medical history.

In total, 37 candidate predictors were collected, spanning sociodemographic variables (age, sex, education), lifestyle factors (smoking, alcohol use), psychosocial factors (social participation, pain, sleep, CESD-10 score, self-reported health), comorbidities, physical examination measures (BMI, waist circumference, SBP, DBP), and laboratory biomarkers (lipids, WBC, MCV, HCT, platelet count, BUN, CRP, uric acid, hemoglobin, fasting glucose, HbA1c, eGFR).

2.4 Missing Data Handling

Participants with missing data for CKM staging or outcome were excluded. Within the final analytic cohort (n = 3773), missing data among the 37 predictors were minimal (all <5%). A complete-case analysis was performed, because multiple imputation when the missingness mechanism is unknown may introduce noise and reduce model robustness, given the minimal missing data. Following Luo et al.[22], we limited the number of candidate predictors to maintain at least 10 events per variable, reducing the risk of overfitting[23].

2.5 Feature Selection

The cohort was randomly split 7:3 into training (n = 2641) and test (n = 1132) sets, stratified by outcome. LASSO regression was applied only to the training set to mitigate overfitting[24]. The optimal regularization parameter (λ) was selected via 10-fold cross-validation using the 1-SE rule to favor a parsimonious model. At $\lambda_{1se} = 0.0187$, eight predictors with non-zero coefficients were retained: age, BMI, SBP, HbA1c, eGFR, baseline CKM stage, sex, and education level. Multicollinearity was assessed using the Variance Inflation Factor (VIF); all VIFs were <5, indicating no substantial collinearity. These features represent a parsimonious yet comprehensive multidimensional profile of metabolic, hemodynamic, and renal function.

Figure 2: Feature selection using LASSO regression and collinearity assessment

Note: LASSO, least absolute shrinkage and selection operator; VIF, variance inflation factor; λ , regularization parameter; SD, standard deviation. (A) Cross-validation curve showing mean deviance versus $\log(\lambda)$; vertical dashed lines indicate λ_{min} and λ_{1se} . (B) Coefficient profiles across λ values for all

candidate predictors. (C) VIF assessment for multicollinearity; variables with VIF >5 were reviewed for redundancy prior to model training. All selected features had VIF <5.

2.6 Model Development and Hierarchical Evaluation

Six ML algorithms were implemented: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), XGBoost, CatBoost, and Multilayer Perceptron (MLP). Hyperparameters were tuned via 10-fold cross-validation grid search in the training set.

Model selection in the independent test set followed a rigorous, three-stage hierarchical framework, consistent with the ABCD validation framework proposed by Steyerberg and Vergouwe[25], which recommends four key performance measures: (A) calibration-in-the-large, (B) calibration slope, (C) discrimination (C-statistic), and (D) clinical usefulness (Decision Curve Analysis).

In the first stage, discrimination was evaluated with AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Brier score. Calibration stability (second stage) was assessed using sample-size-independent metrics: calibration intercept, calibration slope, and the Integrated Calibration Index (ICI). The ICI, defined as the weighted average absolute difference between the loess-based calibration curve and the diagonal line of perfect calibration[16], quantified the average magnitude of miscalibration across the risk spectrum. The calibration slope measured the degree of overfitting (slope >1 indicates overfitting/extreme predictions). The Hosmer–Lemeshow test was reported for reference but was excluded as a primary selection criterion given its known limitations in large cohorts[14,26]. Clinical utility (third stage) was evaluated with Decision Curve Analysis (DCA); the mean net benefit was calculated over threshold probabilities of 0.20–0.40.

2.7 Model Interpretability and Subgroup Analysis

SHAP values were computed for the final model to examine global feature importance, effect direction (beeswarm plots), non-linear dose-response associations (dependence plots), and individual-level predictions (force plots). SHAP satisfies the properties of local accuracy, missingness, and consistency, yielding a unique additive feature importance measure[12]. Subgroup analyses stratified by baseline CKM stage (0, 1, 2) assessed model discrimination (AUC) and calibration (slope and ICI) within each early-stage stratum. The DeLong test was used to compare AUCs between subgroups.

2.8 Statistical Analysis

Statistical analyses were performed using R (version 4.4.3). For baseline comparisons between progression and non-progression groups, continuous variables were first assessed for normality within each group using the Shapiro–Wilk test ($\alpha = 0.05$). Variables satisfying normality in all groups were described as mean $\pm$ standard deviation (SD); otherwise, they were presented as median with interquartile range [M (Q1–Q3)]. For normally distributed variables, Levene's test was used to evaluate homogeneity of variances. Between-group comparisons were conducted as follows: Student's t-test for normally distributed data with equal variances; Welch's t-test for normally distributed data with unequal variances; Wilcoxon rank-sum test for non-normally distributed continuous variables or ordinal categorical variables.

For unordered categorical variables, the Pearson chi-square test was applied when all expected cell counts were ≥ 5 ; otherwise, Fisher's exact test was used. All tests were two-sided, and $P < 0.05$ was considered statistically significant. Categorical variables were described as frequencies (percentages).

Details of model development, hyperparameter tuning, and performance evaluation are provided in Section 2.6. The model selection framework prioritized calibration stability and clinical utility over marginal gains in discrimination.

3. Results

3.1 Baseline Characteristics

A total of 3,773 participants with early-stage CKM at baseline were included. The median age was 57.0 years (IQR: 51.0–63.0), and 52.5% were female. Over the 4-year follow-up, 1560 participants (41.3%) progressed. This rate, substantially higher than the 34% stage-transition frequency reported in Western cohorts and consistent with the dynamic transitions observed in East Asian populations[7,9,10], reflects the marked instability of early CKM in Chinese middle-aged and older adults and indicates a need for reliable risk stratification[6].

Baseline characteristics by progression status are summarized in Table 1. Baseline CKM stage distribution differed markedly between groups ($P < 0.001$). Progressors were disproportionately in Stage 0 (34.9%) and Stage 1 (24.7%), whereas non-progressors were predominantly Stage 2 (85.1%). Because of this imbalance, unadjusted comparisons of metabolic parameters showed a classic epidemiological paradox. Progressors had lower unadjusted SBP (122.0 vs 125.0 mmHg), fasting glucose (95.9 vs 102.8 mg/dL), and triglycerides than non-progressors. These apparent protective associations reflected the overrepresentation of metabolically healthier Stage 0 and 1 participants among progressors and were reversed after multivariable adjustment (see SHAP results).

Table 1
Baseline Characteristics Stratified by CKM Progression Status

Characteristic	Overall (N = 3773)	non-progression (n = 2213)	progression (n = 1560)	P
baseline CKM stage				< 0.001***
0	740 (19.6%)	196 (8.9%)	544 (34.9%)	
1	519 (13.8%)	133 (6.0%)	386 (24.7%)	
2	2514 (66.6%)	1884 (85.1%)	630 (40.4%)	
sex				0.116
male	1794 (47.5%)	1076 (48.6%)	718 (46.0%)	
female	1979 (52.5%)	1137 (51.4%)	842 (54.0%)	
age	57.0 (51.0–63.0)	57.0 (51.0–63.0)	57.0 (51.0–64.0)	0.226
education level				0.080
primary and lower	3462 (91.8%)	2016 (91.1%)	1446 (92.7%)	
junior high school	284 (7.5%)	180 (8.1%)	104 (6.7%)	
high school and above	27 (0.7%)	17 (0.8%)	10 (0.6%)	
smoke				0.838
no	2564 (68.0%)	1501 (67.8%)	1063 (68.1%)	
yes	1209 (32.0%)	712 (32.2%)	497 (31.9%)	
drink				0.003**
no	2465 (65.3%)	1403 (63.4%)	1062 (68.1%)	
yes	1308 (34.7%)	810 (36.6%)	498 (31.9%)	
residence				0.326
urban	1170 (31.0%)	700 (31.6%)	470 (30.1%)	
rural	2603 (69.0%)	1513 (68.4%)	1090 (69.9%)	
Note: CKM, cardiovascular-kidney-metabolic; BMI, body mass index; CESD-10, 10-item Center for Epidemiologic Studies Depression Scale; SBP, systolic blood pressure; DBP, diastolic blood pressure; TG, triglycerides; HDL, high-density lipoprotein cholesterol; TC, total cholesterol; LDL, low-density lipoprotein cholesterol; WBC, white blood cell count; MCV, mean corpuscular volume; HCT, hematocrit; BUN, blood urea nitrogen; CRP, C-reactive protein; UA, uric acid; HbA1c, glycated hemoglobin; eGFR, estimated glomerular filtration rate. Continuous variables are presented as median (interquartile range) and compared using the Mann-Whitney U test; categorical variables are presented as number (percentage) and compared using the χ^2 test. *P < 0.05; **P < 0.01; ***P < 0.001.				

Characteristic	Overall (N = 3773)	non-progression (n = 2213)	progression (n = 1560)	P
medical_insurance				0.024*
no	207 (5.5%)	137 (6.2%)	70 (4.5%)	
yes	3566 (94.5%)	2076 (93.8%)	1490 (95.5%)	
social				0.676
no	2042 (54.1%)	1204 (54.4%)	838 (53.7%)	
yes	1731 (45.9%)	1009 (45.6%)	722 (46.3%)	
pain				0.016*
yes	1219 (32.3%)	681 (30.8%)	538 (34.5%)	
no	2554 (67.7%)	1532 (69.2%)	1022 (65.5%)	
sleep_hours	7.0 (5.0–8.0)	7.0 (5.0–8.0)	7.0 (5.0–8.0)	0.660
CESD10	7.0 (3.0–12.0)	7.0 (3.0–12.0)	7.0 (3.0–12.0)	0.863
self_report_of_health				0.651
Very good	235 (6.2%)	136 (6.1%)	99 (6.3%)	
Good	695 (18.4%)	410 (18.5%)	285 (18.3%)	
Fair	2029 (53.8%)	1197 (54.1%)	832 (53.3%)	
Poor	707 (18.7%)	416 (18.8%)	291 (18.7%)	
Very Poor	107 (2.8%)	54 (2.4%)	53 (3.4%)	
lung_disease				0.052
no	3388 (89.8%)	2005 (90.6%)	1383 (88.7%)	
yes	385 (10.2%)	208 (9.4%)	177 (11.3%)	
arthre				0.627
no	2438 (64.6%)	1437 (64.9%)	1001 (64.2%)	
yes	1335 (35.4%)	776 (35.1%)	559 (35.8%)	
Note: CKM, cardiovascular-kidney-metabolic; BMI, body mass index; CESD-10, 10-item Center for Epidemiologic Studies Depression Scale; SBP, systolic blood pressure; DBP, diastolic blood pressure; TG, triglycerides; HDL, high-density lipoprotein cholesterol; TC, total cholesterol; LDL, low-density lipoprotein cholesterol; WBC, white blood cell count; MCV, mean corpuscular volume; HCT, hematocrit; BUN, blood urea nitrogen; CRP, C-reactive protein; UA, uric acid; HbA1c, glycated hemoglobin; eGFR, estimated glomerular filtration rate. Continuous variables are presented as median (interquartile range) and compared using the Mann-Whitney U test; categorical variables are presented as number (percentage) and compared using the χ^2 test. *P < 0.05; **P < 0.01; ***P < 0.001.				

Characteristic	Overall (N = 3773)	non-progression (n = 2213)	progression (n = 1560)	P
liver_disease				0.979
no	3650 (96.7%)	2141 (96.7%)	1509 (96.7%)	
yes	123 (3.3%)	72 (3.3%)	51 (3.3%)	
digestive_disease				0.092
no	2856 (75.7%)	1697 (76.7%)	1159 (74.3%)	
yes	917 (24.3%)	516 (23.3%)	401 (25.7%)	
asthmae				0.338
no	3613 (95.8%)	2125 (96.0%)	1488 (95.4%)	
yes	160 (4.2%)	88 (4.0%)	72 (4.6%)	
BMI	23.0 (20.8–25.4)	23.0 (20.9–25.5)	22.8 (20.7–25.4)	0.324
waist	83.8 (77.0–91.0)	84.0 (77.5–91.0)	83.0 (76.6–90.3)	0.054
systolic	123.5 (112.0–138.0)	125.0 (113.0–141.5)	122.0 (110.5–134.0)	< 0.001***
diastolic	73.5 (66.0–81.5)	74.0 (67.0–83.5)	72.0 (65.5–80.0)	< 0.001***
TG	100.0 (72.6–144.3)	101.8 (73.5–151.3)	98.2 (70.8–135.4)	< 0.001***
HDL	50.3 (41.4–60.7)	49.5 (41.0–60.7)	51.4 (42.4–61.1)	0.016*
TC	189.4 (166.2–213.0)	189.8 (167.4–214.2)	189.0 (165.5–211.9)	0.159
WBC	5.9 (4.9–7.2)	5.9 (4.9–7.2)	5.9 (4.9–7.2)	0.704
MCV	91.4 (87.0–95.7)	91.3 (87.0–95.5)	91.6 (87.0–95.8)	0.848
HCT	41.6 (38.0–45.1)	41.8 (38.1–45.2)	41.2 (37.8–45.0)	0.019*

Note: CKM, cardiovascular-kidney-metabolic; BMI, body mass index; CESD-10, 10-item Center for Epidemiologic Studies Depression Scale; SBP, systolic blood pressure; DBP, diastolic blood pressure; TG, triglycerides; HDL, high-density lipoprotein cholesterol; TC, total cholesterol; LDL, low-density lipoprotein cholesterol; WBC, white blood cell count; MCV, mean corpuscular volume; HCT, hematocrit; BUN, blood urea nitrogen; CRP, C-reactive protein; UA, uric acid; HbA1c, glycated hemoglobin; eGFR, estimated glomerular filtration rate. Continuous variables are presented as median (interquartile range) and compared using the Mann-Whitney U test; categorical variables are presented as number (percentage) and compared using the χ^2 test. *P < 0.05; **P < 0.01; ***P < 0.001.

Characteristic	Overall (N = 3773)	non-progression (n = 2213)	progression (n = 1560)	P
platelet	207.0 (162.0-254.0)	205.0 (161.0-251.0)	209.0 (164.0-257.0)	0.136
BUN	15.0 (12.5–17.9)	15.2 (12.6–18.2)	14.8 (12.3–17.7)	0.019*
LDL	114.0 (93.6-136.1)	113.7 (93.6-136.5)	114.4 (93.2-135.3)	0.981
CRP	0.9 (0.5–1.9)	0.9 (0.5–1.9)	0.9 (0.5–1.9)	0.812
UA	4.2 (3.5-5.0)	4.2 (3.5-5.0)	4.2 (3.5-5.0)	0.203
Hemoglobin	14.2 (13.0-15.5)	14.3 (13.1–15.6)	14.1 (13.0-15.3)	0.012*
fasting_glucose	100.1 (93.2-107.3)	102.8 (95.8-108.9)	95.9 (90.9–103.0)	< 0.001***
HbA1c	5.1 (4.8–5.3)	5.1 (4.8–5.3)	5.1 (4.9–5.4)	0.202
eGFR	100.6 (91.7-106.6)	100.7 (91.9-106.6)	100.3 (91.2-106.8)	0.410
Note: CKM, cardiovascular-kidney-metabolic; BMI, body mass index; CESD-10, 10-item Center for Epidemiologic Studies Depression Scale; SBP, systolic blood pressure; DBP, diastolic blood pressure; TG, triglycerides; HDL, high-density lipoprotein cholesterol; TC, total cholesterol; LDL, low-density lipoprotein cholesterol; WBC, white blood cell count; MCV, mean corpuscular volume; HCT, hematocrit; BUN, blood urea nitrogen; CRP, C-reactive protein; UA, uric acid; HbA1c, glycated hemoglobin; eGFR, estimated glomerular filtration rate. Continuous variables are presented as median (interquartile range) and compared using the Mann-Whitney U test; categorical variables are presented as number (percentage) and compared using the χ^2 test. *P < 0.05; **P < 0.01; ***P < 0.001.				

3.2 Feature Selection

LASSO regression (1-SE criterion, $\lambda = 0.0187$) retained eight predictors: baseline CKM stage, HbA1c, SBP, BMI, age, eGFR, sex, and education level. VIF analysis confirmed that all eight predictors had VIF values well below 5, indicating the absence of problematic multicollinearity. The selected predictors cover metabolic, hemodynamic, and renal domains.

3.3 Discrimination Performance

Table 2
Discrimination Metrics of Candidate Models in the Test Set

model	ROC-AUC	sensitivity	specificity	PPV	NPV	brier score
CatB	0.772 (0.743–0.800)	0.618	0.837	0.728	0.756	0.189
XGB	0.770 (0.742–0.799)	0.643	0.815	0.710	0.764	0.187
RF	0.761 (0.731–0.791)	0.714	0.699	0.625	0.776	0.188
MLP	0.760 (0.729–0.789)	0.630	0.810	0.701	0.757	0.201
SVM	0.758 (0.729–0.786)	0.652	0.788	0.684	0.762	0.191
LR	0.756 (0.725–0.787)	0.609	0.837	0.725	0.752	0.191

Note: AUC, area under the receiver operating characteristic curve; CI, confidence interval; LR, logistic regression; RF, random forest; SVM, support vector machine; CatB, CatBoost; XGB, XGBoost; Sens, sensitivity; Spec, specificity; PPV, positive predictive value; NPV, negative predictive value; F1, F1-score. All metrics were evaluated in the independent test set (n = 1132). 95% CIs for AUC were calculated via bootstrap resampling with 1000 iterations. Sensitivity and specificity were calculated at the optimal threshold determined by Youden's index.

All six models demonstrated acceptable discrimination on the independent test set, with AUCs ranging from 0.756 to 0.772. CatBoost achieved the highest AUC (0.772), followed closely by XGBoost (0.770) and RF (0.761). Differences among the top models were small ($\Delta\text{AUC} < 0.02$); discrimination alone was insufficient for model selection.

Figure 3: Performance evaluation of the six machine learning models in the test set

Note ROC, receiver operating characteristic; AUC, area under the ROC curve; DCA, decision curve analysis; CI, confidence interval; LR, logistic regression; RF, random forest; SVM, support vector machine; MLP, multilayer perceptron; CatB, CatBoost; XGB, XGBoost. (A) ROC curves for all six models in the test set. The legend displays the AUC values with 95% CIs in parentheses. (B) DCA comparing the net benefit across clinically relevant threshold probabilities (approx. 0.20–0.40). RF demonstrated superior clinical utility in this range. (C) Calibration plots showing the agreement between predicted probabilities and observed outcomes. The solid lines represent the model performance, shaded areas indicate the 95% CIs, and the dashed diagonal line represents perfect prediction.

3.4 Calibration Stability and Clinical Utility

In the second stage, calibration was assessed. CatBoost, despite having the highest AUC, was poorly calibrated (calibration slope = 1.305; ICI = 0.060), indicating overfitting and extreme predictions. MLP showed even greater overfitting (slope = 2.068). In contrast, Random Forest calibrated well: slope = 1.056 (95% CI 0.917–1.200), intercept = 0.059, ICI = 0.026 (the lowest among models). The H-L test was non-

significant for RF ($P = 0.614$), but more importantly, the slope and ICI confirmed its robust reliability across the entire risk spectrum.

In the final clinical utility assessment, Decision Curve Analysis revealed that RF achieved the highest arithmetic mean net benefit (0.22) across the clinically relevant threshold range (0.20–0.40), outperforming CatBoost (0.2193), XGBoost (0.2169), and the "treat-all" strategy (0.1557) (Table 4).

Table 3
Calibration Metrics of Candidate Models in the Test Set

model	Intercept	Intercept 95%CI	Slope	Slope 95%CI	ICI	HL <i>P</i> -value
LR	0.031	(-0.112, 0.176)	0.961	(0.829, 1.098)	0.039	0.002
RF	0.059	(-0.084, 0.205)	1.056	(0.917, 1.200)	0.026	0.614
XGB	-0.267	(-0.403, -0.131)	1.009	(0.880, 1.142)	0.055	0.003
CatB	0.447	(0.277, 0.623)	1.305	(1.141, 1.475)	0.06	< 0.001
SVM	0.001	(-0.143, 0.147)	0.981	(0.848, 1.119)	0.035	< 0.001
MLP	0.003	(-0.139, 0.146)	2.068	(1.797, 2.347)	0.115	< 0.001
Note: ICI, integrated calibration index; LR, logistic regression; RF, random forest; SVM, support vector machine; KNN, k-nearest neighbors; CatB, CatBoost; XGB, XGBoost; HL, Hosmer–Lemeshow goodness-of-fit test. ICI quantifies overall calibration error (lower = better); calibration slope = 1 indicates perfect calibration. All metrics derived from the test set. ICI calculated as the mean absolute difference between observed and predicted probabilities across deciles.						

Table 4
Mean Net Benefit Across Threshold Probabilities 0.20–0.40

model	RF	LR	CatB	XGB	SVM	MLP	treat all	treat none
mean net benefit	0.22	0.2196	0.2193	0.2169	0.2125	0.1886	0.1557	0
Note: LR, logistic regression; RF, random forest; SVM, support vector machine; KNN, k-nearest neighbors; CatB, CatBoost; XGB, XGBoost; NB, net benefit. Mean NB averaged across threshold probabilities 0.20–0.40, reflecting clinically relevant risk ranges for CKM progression intervention. $NB = (TP/N) - (FP/N) \times [pt/(1 - pt)]$, where pt is threshold probability. Higher NB indicates greater clinical utility. All values from test set.								

Random Forest was therefore selected: its calibration and net benefit outweighed the small gain in AUC. This choice is consistent with the calibration hierarchy of Van Calster et al.[15], who distinguish four levels

of calibration. Moderate calibration—in which predicted risks match observed event rates across the risk spectrum—supports clinically safe decision-making, whereas weak calibration does not.

3.5 SHAP-Based Model Interpretability

Global SHAP analysis identified the top three drivers of CKM progression as: (1) baseline CKM stage (mean $|\text{SHAP}| \approx 0.19$), (2) HbA1c (≈ 0.03), and (3) SBP (≈ 0.02).

Figure 4: SHAP-based interpretability analysis of the optimal Random Forest model

Note SHAP, SHapley Additive exPlanations; HbA1c, glycated hemoglobin; BMI, body mass index; CKM, cardiovascular-kidney-metabolic; $f(x)$, predicted probability for an individual; $E[f(x)]$, expected model output (base value). (A) Feature importance ranking based on the mean absolute SHAP value. (B) SHAP summary plot (beeswarm plot) showing the impact of top features on the model output. Each point represents a patient; color indicates the feature value (yellow: high, purple: low). Points to the right indicate increased risk of CKM progression. (C) SHAP dependence plots illustrating the relationship between specific features (baseline_ckm_stage, HbA1c, systolic blood pressure, age) and their SHAP values. In the first three plots, points are colored by HbA1c levels to highlight interaction effects. (D) Force plot visualizing the prediction for a single sample. Red bars indicate features pushing the prediction higher (increasing risk), while blue bars indicate features pushing the prediction lower (decreasing risk).

The beeswarm plot revealed a non-monotonic "inverse stage effect": Stage 0 increased predicted risk (positive SHAP values), whereas Stage 2 decreased it (negative SHAP values). Higher HbA1c and SBP were associated with positive SHAP values, consistent with their independent contributions after multivariable adjustment; this resolves the unadjusted confounding apparent in Table 1.

Dependence plots identified non-linear thresholds: progression risk rose sharply when HbA1c passed ~ 5.0 – 5.5% and when SBP passed ~ 140 mmHg. The age dependence plot showed a weak positive association across older ages. These thresholds may inform clinical intervention. An individual force plot illustrated the prediction for a 79-year-old male with Stage 2, low BMI, and low SBP: the model predicted a low progression probability ($f(x) = 0.189$ vs. baseline 0.413).

3.6 Subgroup Heterogeneity: The Predictive Ceiling Effect Across CKM Stages

Subgroup analyses stratified by baseline CKM stage (0, 1, 2) evaluated model performance in early stages. Random Forest showed poor discrimination in all three stages, with AUCs of 0.599 (Stage 0), 0.618 (Stage 1), and 0.588 (Stage 2). This predictive "ceiling effect" is largely attributable to the high progression rates in these subgroups (72.7% Stage 0, 78.7% Stage 1), where clinical features have limited capacity to distinguish the minority of non-progressors from the majority of progressors.

Despite limited discrimination, calibration remained acceptable across all stages. In Table 5 and Fig. 5B, calibration curves approached the diagonal reference line; ICI values were 0.042 (Stage 0), 0.096 (Stage 1),

and 0.032 (Stage 2), suggesting that predicted probabilities were reliable despite the limited discrimination.

Table 5
Subgroup Performance and Calibration of the Random Forest Model by Baseline CKM Stage

baseline_CKM_stage	n	Events	Intercept(95%CI)	Slope(95%CI)	ICI
0	213	155	0.328 (-0.336, 1.007)	0.751 (0.052, 1.471)	0.042
1	146	115	0.374 (-0.545, 1.309)	1.134 (0.105, 2.25)	0.096
2	773	198	-0.444 (-0.802, -0.09)	0.61 (0.291, 0.933)	0.032

Note: CKM, cardiovascular-kidney-metabolic syndrome;n, number of patients in the test set for the subgroup; Events, number of patients with CKM progression events; Intercept (95% CI), calibration intercept (ideal value = 0, indicating perfect calibration); Slope (95% CI), calibration slope (ideal value = 1, indicating optimal predictive strength); ICI, integrated calibration index (lower values indicate better calibration agreement between predicted and observed risks). Subgroups: Stage 0 ,Stage 1 and Stage 2. AUC and ICI with 95% CIs reported for each subgroup. All metrics from test set.

Figure 5: Subgroup analysis by baseline CKM stage

Note AUC, area under the receiver operating characteristic curve; CI, confidence interval; CKM, cardiovascular-kidney-metabolic syndrome. Subgroups defined by baseline CKM stage (0 ,1, 2). (A) AUC comparison with 95% CIs for the Random Forest model across subgroups. (B) Calibration plots showing observed versus predicted probabilities.

4. Discussion

4.1 Principal Findings and the Dynamic Instability of Early CKM

In this prospective longitudinal study of 3773 adults with early-stage CKM syndrome, we developed and validated an interpretable ML framework for predicting 4-year disease progression. Overall, 41.3% of participants progressed over 4 years, a rate exceeding the 34% reported among stage 1 individuals in Western cohorts[9], reflecting the marked instability of early CKM in Chinese middle-aged and older adults. This progression rate aligns with findings from the Kailuan study, where all-cause mortality increased across CKM stages, and Stage 2 carried a 72% higher risk than Stage 0[6]. Longitudinal data from East Asian cohorts show that CKM stage progression is common: 24% of Korean adults worsened over 4 years, and progression from Stage 0 to 1 was associated with a measurable increase in mortality risk[7]. Using a hierarchical model selection process, we selected Random Forest, prioritizing calibration stability and net benefit over marginal gains in AUC. SHAP analysis identified baseline CKM stage, HbA1c, and SBP as the top predictors; subgroup analyses revealed a predictive "ceiling effect" in early stages with high progression rates.

4.2 Methodological Implications: The Primacy of Calibration

A methodological contribution of this study is the departure from the "AUC-chasing" paradigm. While CatBoost yielded a marginally higher AUC than RF (0.772 vs. 0.761), it was poorly calibrated (slope = 1.305; ICI = 0.060). A slope > 1.0 indicates overconfident predictions, which could lead to overtreatment and patient anxiety.

This distinction is consistent with the calibration hierarchy of Van Calster et al.[15], who distinguish four levels of calibration: mean, weak, moderate, and strong. Weak calibration does not guarantee moderate calibration; moderate calibration—where predicted risks match observed event rates across the risk spectrum—supports safe clinical decision-making. Strong calibration is unrealistic in empirical research. RF was selected because of its favorable calibration slope (1.056), lowest ICI (0.026), and highest net benefit, indicating that calibration and utility outweighed small improvements in AUC.

Van Calster and colleagues noted, calibration is the "Achilles heel" of predictive analytics[13]—the C-statistic alone can mislead. The TRIPOD guidelines recommend reporting on calibration performance[19], yet calibration is assessed far less often than discrimination in practice. Our focus on calibration addresses this gap and aligns with the recognition that poorly calibrated models can mislead clinical decisions. Van Calster et al.[13] recommend that external validation studies assess calibration using calibration curves with pointwise confidence intervals, the calibration intercept, and the calibration slope, and suggest a minimum of 200 events and 200 non-events for reliable calibration assessment. Our test set (1,132 participants, 468 events) exceeds these recommendations, supporting the robustness of our calibration evaluations.

4.3 The "Inverse Stage Effect": Unadjusted Confounding and Adjusted Stage-Specific Risk

The unadjusted baseline comparisons revealed a counterintuitive epidemiological paradox: progressors had lower blood pressure and glucose. This was fully attributable to the marked imbalance in baseline CKM stage distribution (confounding by indication/disease severity). After adjustment, SHAP values showed that higher SBP and HbA1c were associated with increased risk.

In addition, SHAP revealed an "inverse stage effect": Stage 2 was associated with lower predicted risk than Stage 0/1. Statistically, this reflects a ceiling effect; Stage 0 individuals have multiple upward trajectories (e.g., developing obesity, then hypertension), whereas Stage 2 participants have fewer transitions to cross to be classified as progressors. Pathophysiologically, it may reflect the dynamic instability of early metabolic disruption: Stage 0/1 individuals are more prone to rapid transitions driven by aging and lifestyle, whereas Stage 2 may represent a more "treated or exhausted" phenotype with blunted short-term progression velocity. Longitudinal studies of metabolic syndrome show that prevalence peaks in midlife and may decline in older age, consistent with differential stage-specific dynamics[10]. This interpretation is consistent with the CKM framework, in which the syndrome often begins early in life with excess and dysfunctional adipose tissue, leading to progressive metabolic deterioration[1,2].

4.4 Clinical Implications of the Predictive Ceiling Effect in Early CKM Stages

Subgroup analysis provided a clinically relevant insight: a predictive ceiling effect was evident in Stages 0 and 1. With progression rates > 70%, AUCs around 0.60 (Fig. 5A) suggest that conventional biomarkers are limited in distinguishing progressors from non-progressors in these high-risk groups.

This observation aligns with the high population prevalence of CKM risk factors: ~90% of U.S. and > 80% of Chinese adults meet criteria for Stage 1 or higher[4,5], suggesting that conventional markers may be near a predictive ceiling in such settings. From a clinical standpoint, individuals already identified as Stage 0 or 1 (e.g., isolated obesity) do not need complex predictive modeling to determine progression risk. These individuals need immediate preventive interventions (e.g., intensive lifestyle modification) regardless of predicted risk. The Look AHEAD trial showed that intensive lifestyle intervention achieved sustained weight loss (6.0% at 9.6 years) and improved CKM-related risk factors, including HbA1c and fitness, although cardiovascular events were not reduced in the overall cohort[27]. The model is most useful for stratifying the large Stage 2 population, where progression risk is more heterogeneous, though performance is still modest. However, the model calibrated well across stages (Table 5), so estimated risks are reliable even if discrimination is limited. This is consistent with the AHA's emphasis on early detection and prevention to halt or reverse CKM progression[1,28].

4.5 Limitations

Several limitations should be noted. First, the definition of Stage 3 was simplified due to lack of UACR and CAC scoring in CHARLS. Second, the 4-year follow-up may be insufficient to capture hard CVD endpoints (Stage 4). Third, time-varying covariates and medication changes were not considered. Fourth, unmeasured confounders (diet quality, psychosocial stress) could bias results. Fifth, validation was limited to a single internal split of the CHARLS data; external validation in independent, multi-ethnic cohorts is needed.

5. Conclusions

We developed and validated a well-calibrated, interpretable Random Forest framework for predicting 4-year disease progression in early-stage CKM syndrome. A hierarchical evaluation framework showed that calibration stability and net benefit should outweigh marginal gains in AUC. The model showed a predictive ceiling effect in high-risk early stages (Stage 0–1), indicating that these individuals may benefit more from immediate preventive care than from further risk stratification based solely on conventional features.

Abbreviations

AHA
American Heart Association
AUC
Area under the curve

BMI
Body mass index
BUN
Blood urea nitrogen
CAC
Coronary artery calcium
CatB
CatBoost
CHARLS
China Health and Retirement Longitudinal Study
CKD
Chronic kidney disease
CKD-EPI
Chronic Kidney Disease Epidemiology Collaboration
CRP
C-reactive protein
CVD
Cardiovascular disease
DBP
Diastolic blood pressure
DCA
Decision curve analysis
eGFR
Estimated glomerular filtration rate
FPG
Fasting plasma glucose
HbA1c
Glycated hemoglobin
HCT
Hematocrit
HDL
High-density lipoprotein
H-L
Hosmer-Lemeshow
ICI
Integrated Calibration Index
LASSO
Least absolute shrinkage and selection operator
LDL
Low-density lipoprotein
LR

Logistic Regression
MCV
Mean corpuscular volume
ML
Machine learning
MLP
Multilayer Perceptron
NPV
Negative predictive value
PPV
Positive predictive value
RF
Random Forest
ROC
Receiver operating characteristic
SBP
Systolic blood pressure
SHAP
SHapley Additive exPlanations
SVM
Support Vector Machine
TC
Total cholesterol
TG
Triglycerides
TRIPOD
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
UA
Uric acid
UACR
Urine albumin-to-creatinine ratio
VIF
Variance inflation factor
WBC
White blood cell
XGB
XGBoost.

Declarations

Acknowledgements

We thank Peking University for providing the open-access CHARLS data resources and all investigators and participants who contributed to the CHARLS study.

Authors' contributions

X.T. conceived the study, performed data curation and machine learning modeling, and wrote the main manuscript text. Y.L. assisted with data curation, data validation, and figure preparation. E.Z. provided guidance on writing and figure layout. X.H. supervised the study, contributed to the conceptualization, and critically revised the manuscript. All authors read and approved the final manuscript.

Funding

This study received no specific funding from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials

The datasets generated and/or analyzed during the current study are publicly available in the CHARLS repository (<http://charls.pku.edu.cn>). The data that support the findings of this study are openly accessible upon registration and request at the CHARLS website.

Ethics approval and consent to participate

This retrospective cohort study utilized data from the China Health and Retirement Longitudinal Study (CHARLS). The original CHARLS study was approved by the Biomedical Ethics Review Committee of Peking University (IRB00001052-11015), and all participants provided written informed consent at the time of enrollment. The current analysis involved de-identified data from the CHARLS public database; therefore, additional ethical approval and individual informed consent were not required. This study was conducted in accordance with the principles of the Declaration of Helsinki.

Consent for publication

Not applicable. This manuscript does not contain any individual person's data in any form (including individual details, images, or videos).

Competing interests

The authors declare that they have no competing interests.

References

1. Ndumele CE, Rangaswami J, Chow SL, Neeland IJ, Tuttle KR, Khan SS, et al. Cardiovascular-Kidney-Metabolic Health: A Presidential Advisory From the American Heart Association. *Circulation*. 2023;148:1606–35. <https://doi.org/10.1161/CIR.0000000000001184>

2. Xu Z, Yang S, Tan Y, Zhang Q, Wang H, Tao J, et al. Inflammation in cardiovascular-kidney-metabolic syndrome: key roles and underlying mechanisms—a comprehensive review. *Mol Cell Biochem*. 2025;480:6039–75. <https://doi.org/10.1007/s11010-025-05379-9>
3. D’Elia JA, Weinrauch LA. Lipid Toxicity in the Cardiovascular-Kidney-Metabolic Syndrome (CKMS). *Biomedicines*. 2024;12:978. <https://doi.org/10.3390/biomedicines12050978>
4. Aggarwal R, Ostrominski JW, Vaduganathan M. Prevalence of Cardiovascular-Kidney-Metabolic Syndrome Stages in US Adults, 2011-2020. *JAMA*. 2024;331:1858–60. <https://doi.org/10.1001/jama.2024.6892>
5. He D, Zhang X, Li C, Yang C, Yu M, Wang J, et al. Rising Prevalence of Cardiovascular-Kidney-Metabolic Syndrome in China, 2010-2019: National Cross-Sectional Surveys. *J Am Coll Cardiol*. 2025;86:213–6. <https://doi.org/10.1016/j.jacc.2025.05.030>
6. Li N, Li Y, Cui L, Shu R, Song H, Wang J, et al. Association between different stages of cardiovascular-kidney-metabolic syndrome and the risk of all-cause mortality. *Atherosclerosis* [Internet]. Elsevier; 2024 [cited 2026 June 4];397. <https://doi.org/10.1016/j.atherosclerosis.2024.118585>
7. Park S, Kang MW, Lee J, Oh J, Kim M, Koh JH, et al. Stage-Specific Progression of Cardiovascular-Kidney-Metabolic Health and Mortality in a Nationwide Cohort of East Asian Adults. *Clin J Am Soc Nephrol*. 2026; <https://doi.org/10.2215/CJN.0000001104>
8. Kim BS, Kim H-J, Kim H, Lee J, Shin J-H, Sung K-C. Longitudinal Changes in Cardiovascular-Kidney-Metabolic Syndrome Stages and Their Impact on Outcomes: A Nationwide Cohort Study. *J Clin Med*. 2025;14:3888. <https://doi.org/10.3390/jcm14113888>
9. Gunnarsson S, Vito O, Unwin RJ. Cardiovascular-kidney-metabolic syndrome: prevalence, risks, disease trajectories, and early-stage management. *Am J Physiol, Cell Physiol*. 2026;330:C1–8. <https://doi.org/10.1152/ajpcell.00499.2025>
10. Kuzuya M, Ando F, Iguchi A, Shimokata H. Age-specific change of prevalence of metabolic syndrome: longitudinal observation of large Japanese cohort. *Atherosclerosis*. 2007;191:305–12. <https://doi.org/10.1016/j.atherosclerosis.2006.05.043>
11. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15:20170387. <https://doi.org/10.1098/rsif.2017.0387>
12. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc.; 2017 [cited 2025 Dec 17]. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html. Accessed 17 Dec 2025
13. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. <https://doi.org/10.1186/s12916-019-1466-7>
14. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med*. 2007;35:2052–6. <https://doi.org/10.1097/01.CCM.0000275267.64078.B0>

15. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol.* 2016;74:167–76. <https://doi.org/10.1016/j.jclinepi.2015.12.005>
16. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med.* 2019;38:4051–65. <https://doi.org/10.1002/sim.8281>
17. Li J, Lei L, Wang W, Ding W, Yu Y, Pu B, et al. Social Risk Profile and Cardiovascular-Kidney-Metabolic Syndrome in US Adults. *J Am Heart Assoc.* 2024;13:e034996. <https://doi.org/10.1161/JAHA.124.034996>
18. Zhao Y, Hu Y, Smith JP, Strauss J, Yang G. Cohort Profile: The China Health and Retirement Longitudinal Study (CHARLS). *Int J Epidemiol.* 2014;43:61–8. <https://doi.org/10.1093/ije/dys203>
19. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med.* American College of Physicians; 2015;162:55–63. <https://doi.org/10.7326/M14-0697>
20. Ndumele CE, Neeland IJ, Tuttle KR, Chow SL, Mathew RO, Khan SS, et al. A synopsis of the evidence for the science and clinical management of cardiovascular-kidney-metabolic (CKM) syndrome: a scientific statement from the American heart association. *Circulation.* 2023;148:1636–64. <https://doi.org/10.1161/CIR.0000000000001186>
21. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, et al. New creatinine- and cystatin C-based equations to estimate GFR without race. *N Engl J Med.* 2021;385:1737–49. <https://doi.org/10.1056/NEJMoa2102953>
22. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *Journal of Medical Internet Research.* JMIR Publications Inc., Toronto, Canada; 2016;18:e5870. <https://doi.org/10.2196/jmir.5870>
23. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol.* 1996;49:1373–9. [https://doi.org/10.1016/s0895-4356\(96\)00236-3](https://doi.org/10.1016/s0895-4356(96)00236-3)
24. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B.* 1996;58:267–88. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
25. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35:1925–31. <https://doi.org/10.1093/eurheartj/ehu207>
26. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med.* 1997;16:965–80. [https://doi.org/10.1002/\(sici\)1097-0258\(19970515\)16:9%3C965::aid-sim509%3E3.0.co;2-o](https://doi.org/10.1002/(sici)1097-0258(19970515)16:9%3C965::aid-sim509%3E3.0.co;2-o)
27. Cardiovascular Effects of Intensive Lifestyle Intervention in Type 2 Diabetes. *N Engl J Med.* 2013;369:145–54. <https://doi.org/10.1056/NEJMoa1212914>
28. Soliman GA. Cardiovascular–Kidney–Metabolic (CKM) Syndrome Staging and Relevance to Precision Nutrition. *Nutrients.* 2026;18:1430. <https://doi.org/10.3390/nu18091430>

Figures

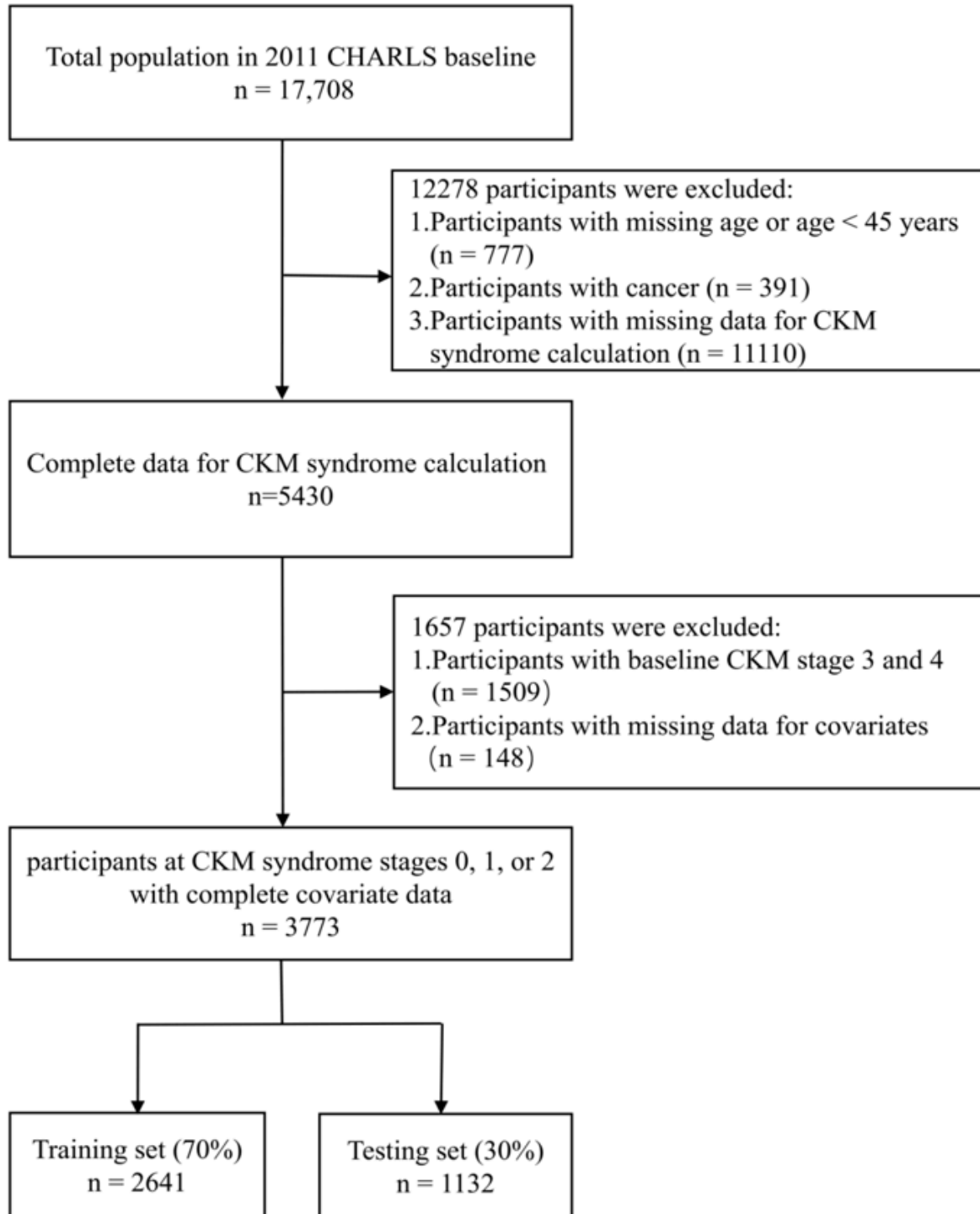


Figure 1

Flowchart of the study population selection from the CHARLS cohort

Note: CHARLS, China Health and Retirement Longitudinal Study; CKM, cardiovascular-kidney-metabolic syndrome. Participants were excluded sequentially based on predefined criteria. The final analytic cohort included 3773 individuals with complete baseline and follow-up data.

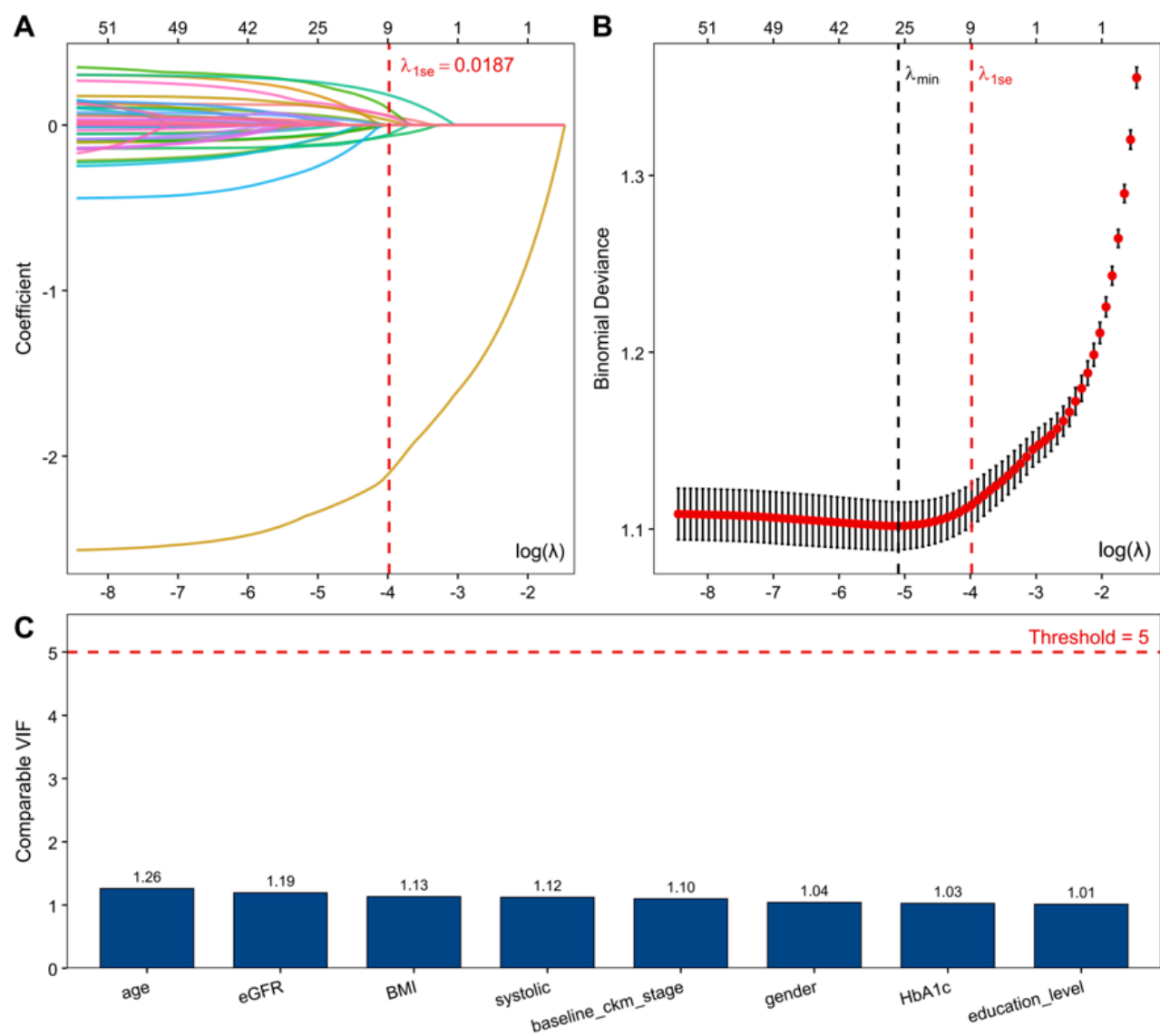


Figure 2

Feature selection using LASSO regression and collinearity assessment

Note: LASSO, least absolute shrinkage and selection operator; VIF, variance inflation factor; λ , regularization parameter; SD, standard deviation. (A) Cross-validation curve showing mean deviance versus $\log(\lambda)$; vertical dashed lines indicate λ_{min} and λ_{1se} . (B) Coefficient profiles across λ values for all candidate predictors. (C) VIF assessment for multicollinearity; variables with $VIF > 5$ were reviewed for redundancy prior to model training. All selected features had $VIF < 5$.

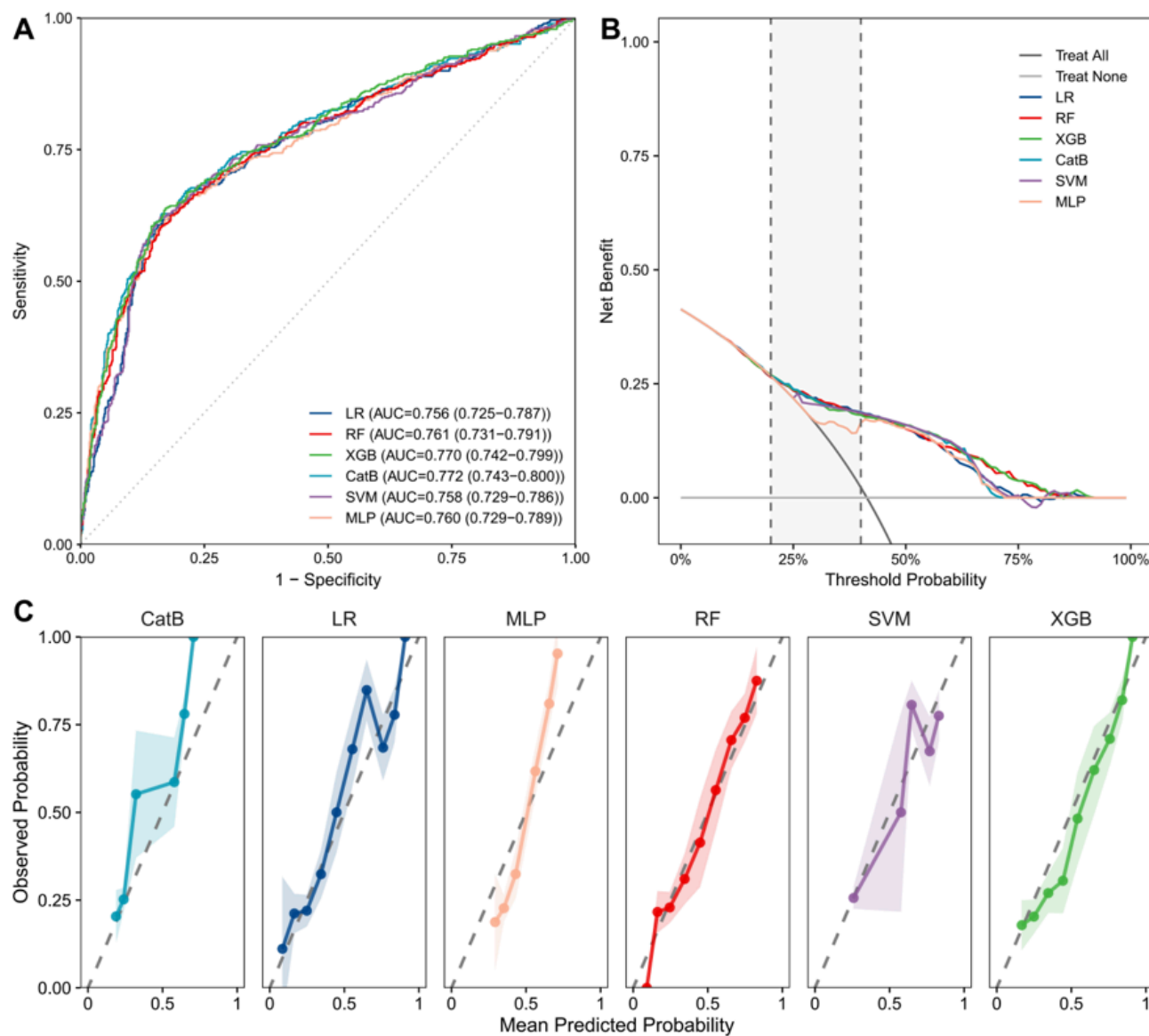


Figure 3

Performance evaluation of the six machine learning models in the test set

Note: ROC, receiver operating characteristic; AUC, area under the ROC curve; DCA, decision curve analysis; CI, confidence interval; LR, logistic regression; RF, random forest; SVM, support vector machine; MLP, multilayer perceptron; CatB, CatBoost; XGB, XGBoost. (A) ROC curves for all six models in the test set. The legend displays the AUC values with 95% CIs in parentheses. (B) DCA comparing the net benefit across clinically relevant threshold probabilities (approx. 0.20–0.40). RF demonstrated superior clinical utility in this range. (C) Calibration plots showing the agreement between predicted probabilities and observed outcomes. The solid lines represent the model performance, shaded areas indicate the 95% CIs, and the dashed diagonal line represents perfect prediction.

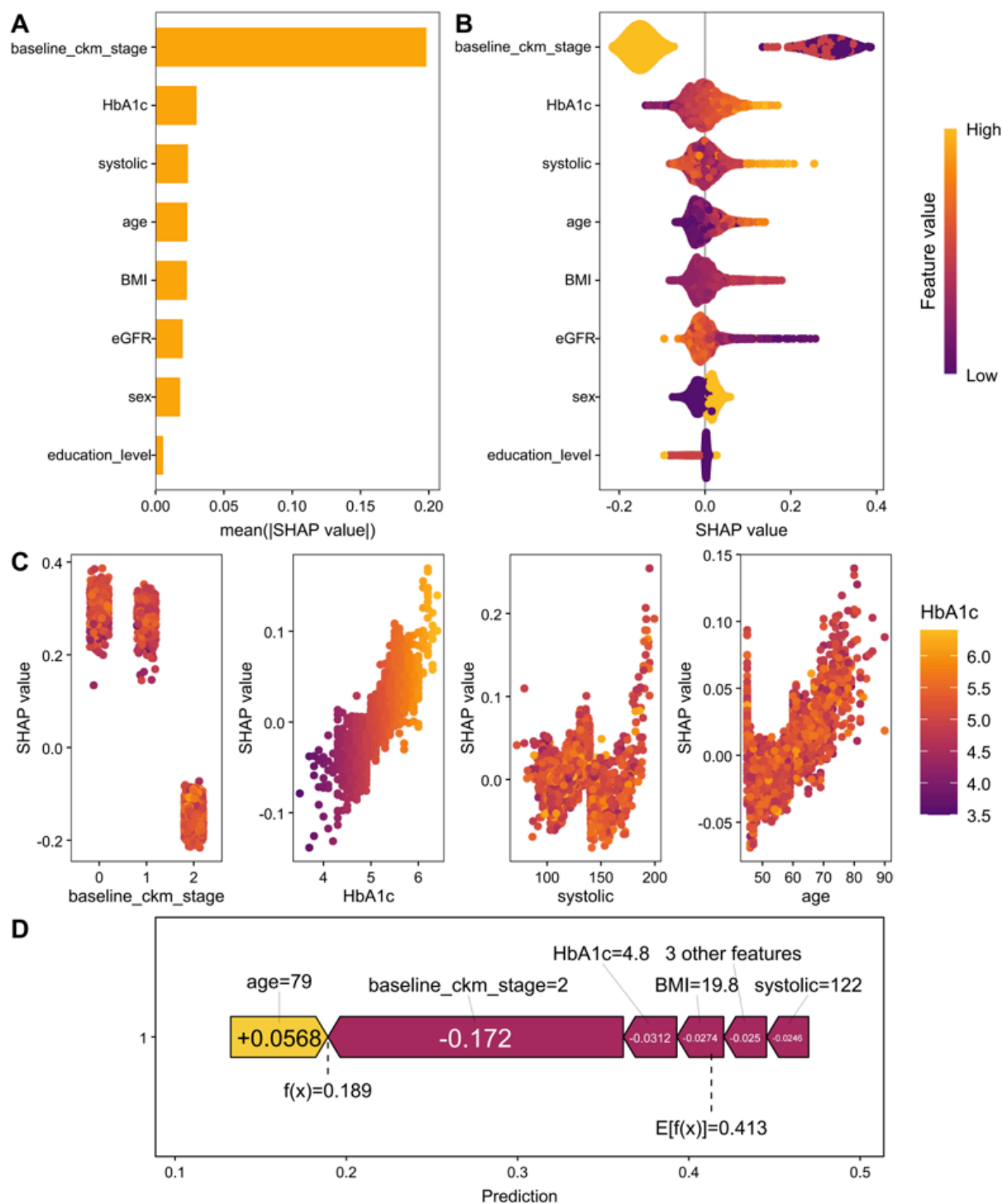


Figure 4

SHAP-based interpretability analysis of the optimal Random Forest model

Note: SHAP, SHapley Additive exPlanations; HbA1c, glycated hemoglobin; BMI, body mass index; CKM, cardiovascular-kidney-metabolic; $f(x)$, predicted probability for an individual; $E[f(x)]$, expected model output (base value). (A) Feature importance ranking based on the mean absolute SHAP value. (B)

SHAP summary plot (beeswarm plot) showing the impact of top features on the model output. Each point represents a patient; color indicates the feature value (yellow: high, purple: low). Points to the right indicate increased risk of CKM progression. (C) SHAP dependence plots illustrating the relationship between specific features (baseline_ckm_stage, HbA1c, systolic blood pressure, age) and their SHAP values. In the first three plots, points are colored by HbA1c levels to highlight interaction effects. (D) Force plot visualizing the prediction for a single sample. Red bars indicate features pushing the prediction higher (increasing risk), while blue bars indicate features pushing the prediction lower (decreasing risk).

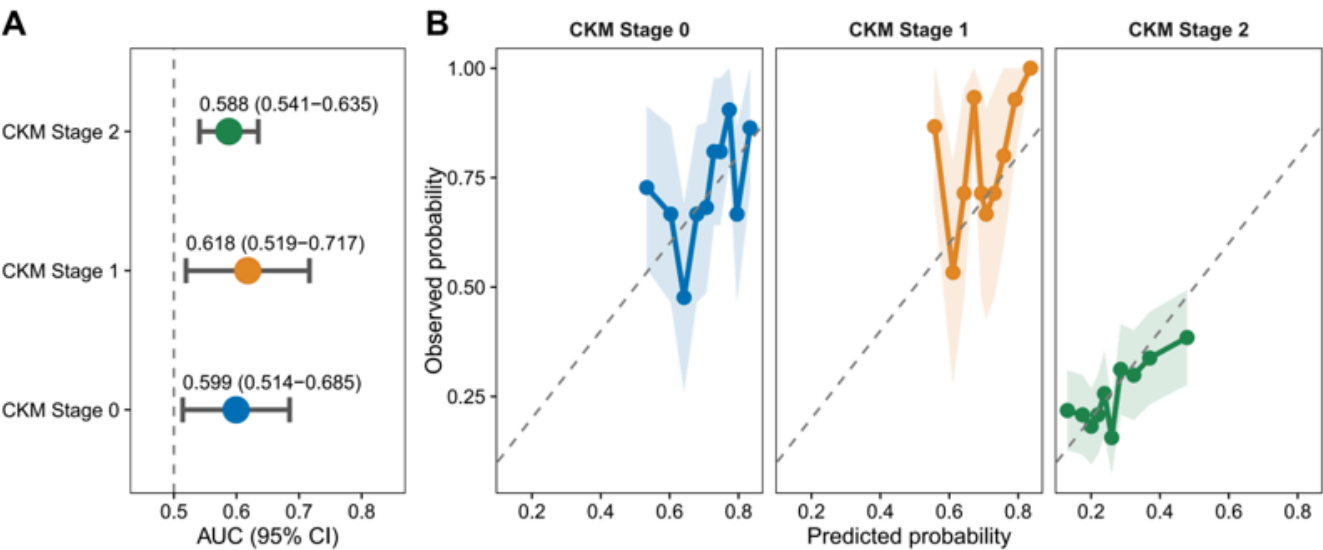


Figure 5

Subgroup analysis by baseline CKM stage

Note: AUC, area under the receiver operating characteristic curve; CI, confidence interval; CKM, cardiovascular-kidney-metabolic syndrome. Subgroups defined by baseline CKM stage (0 ,1, 2). (A) AUC comparison with 95% CIs for the Random Forest model across subgroups. (B) Calibration plots showing observed versus predicted probabilities.