

Development and preliminary validation of a screening model for metabolic syndrome using routine indicators from the CHARLS cohort

Wei Liu

Yueyang Second People's Hospital

Shangfu Li

Yueyang Second People's Hospital

Yiqun Zeng

People's Hospital of Wangcheng District Changsha

Wei Dai

Yueyang Second People's Hospital

Jing Peng

Yueyang Second People's Hospital

Xiaofang Wang

Yueyang Second People's Hospital

Qingping Chen

Yueyang Second People's Hospital

Wei Li

Yueyang Second People's Hospital

Duanlin Du

Yueyang Second People's Hospital

Sen Lu

Yueyang Second People's Hospital

Lingge Yang

lingeryang93@foxmail.com

Yueyang Second People's Hospital

Research Article

Keywords: Metabolic syndrome (MetS), Screening model, Machine learning, Support vector machine (SVM), China Health and Retirement Longitudinal Study (CHARLS)

Posted Date: September 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10772120/v1>

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

Additional Declarations: No competing interests reported.

Abstract

Metabolic syndrome (MetS) elevates risks of type 2 diabetes and cardiovascular disease in middle-aged and older adults. Most existing predictive models for MetS derived from the China Health and Retirement Longitudinal Study (CHARLS) rely on conventional metabolic indicators, overlooking low-cost hematological and C-reactive protein (CRP) markers. We developed an interpretable support vector machine (SVM) model for MetS screening targeting primary-care triage. Using cross-sectional CHARLS data from 3,115 participants, we integrated blood parameters, CRP, demographics and comorbidities to construct logistic regression, least absolute shrinkage and selection operator (LASSO), and SVM models with SHAP interpretation. A 1:1 matched inpatient cohort (n = 64) served for preliminary external feasibility testing. The SVM achieved favorable discrimination, with a training area under the curve (AUC) of 0.767 and an external AUC of 0.860, and yielded superior net benefit, with CRP and cardiovascular histories as dominant predictors. Mild calibration deviation was acceptable for risk stratification. This low-cost tool supports preliminary MetS screening in resource-limited primary settings, though further multi-center community validation is required.

Introduction

Metabolic syndrome (MetS) is a cluster of metabolic abnormalities, including central obesity, impaired glucose metabolism, dyslipidemia, and elevated blood pressure, and is recognized as an important upstream condition for cardiovascular disease, type 2 diabetes, and other chronic disorders¹. Given the growing prevalence of obesity and related metabolic disturbances in aging populations, identifying individuals with undiagnosed MetS in routine clinical settings has become an increasingly pressing need^{2,3}. In Chinese adults aged 45 years and older, studies based on the China Health and Retirement Longitudinal Study (CHARLS) have shown a substantial burden of MetS, with clear heterogeneity by sex, age, and region^{4,5}. Growing evidence also suggests that chronic low-grade inflammation contributes to the development and progression of MetS by affecting insulin sensitivity, endothelial function, and adipose tissue homeostasis^{6,7}. More recently, adipokine-driven metabolic inflammation has been proposed as a central mechanistic framework linking excess adiposity to systemic immune dysregulation, insulin resistance, and downstream cardiometabolic complications. These mechanistic insights suggest that routinely available inflammatory markers, such as C-reactive protein (CRP) and white blood cell count (WBC), might serve as accessible proxies for metabolic risk in clinical practice.

Despite its clinical importance, MetS risk assessment in routine practice still relies largely on fixed diagnostic thresholds and a limited number of traditional metabolic indicators, which may not adequately capture the continuum and heterogeneity of individual risk^{1,8}. Recent evidence further supports the concept that cardiometabolic risk is biologically heterogeneous even among individuals with similar conventional obesity- or syndrome-based classifications, highlighting the need for more precise risk stratification strategies⁹. Previous prediction studies have often been based on relatively homogeneous populations or have relied on questionnaire-based or specialized measures that may limit

applicability in community and primary care settings. In addition, routinely available inflammatory and hematological indicators, such as CRP, WBC, hemoglobin (HGB), and mean corpuscular volume (MCV), have been associated with prevalent MetS but have not been sufficiently integrated into multidimensional screening models¹⁰. More broadly, recent high-dimensional risk modeling studies have suggested that integrating diverse biological features may improve individualized disease prediction beyond conventional single-domain markers¹¹.

In this study, we used routinely available clinical and laboratory data to develop and compare screening models for MetS based on multivariable logistic regression, least absolute shrinkage and selection operator (LASSO) regression, and support vector machine (SVM). Using data from the CHARLS cohort, we evaluated model performance in terms of discrimination, calibration, and decision-analytic utility, and conducted preliminary external feasibility assessment to examine model discriminative capacity. Our aim was to develop a practical risk stratification tool for identifying individuals with a high likelihood of MetS in primary-care and community settings¹². However, in many of these settings, complete metabolic evaluations, including waist circumference, fasting glucose, and lipid profiles, are not routinely performed without specific metabolic indications, and waist circumference measurements are often not recorded in electronic health records even when obtained. This gap highlights the need for a preliminary triage tool that uses routinely accessible, non-fasting indicators to identify individuals who may benefit from subsequent confirmatory diagnostic testing.

Patients and Methods

Data collection and processing

The training cohort was derived from the 2011 and 2015 waves of the CHARLS¹³. Participants with more than 50% missing data or non-fasting biochemical measurements were excluded. For duplicated participant IDs, only the first record was retained. Baseline characteristics were compared between included and excluded participants to evaluate potential selection bias from the complete-case approach (**Table S1**). Candidate variables were initially screened using univariable logistic regression, and participants with remaining missing values in the selected variables were subsequently excluded to generate a complete-case training dataset for model development. Given that missingness appeared consistent with a missing-not-at-random pattern, complete-case analysis was chosen as a more conservative approach than multiple imputation, which would rely on untestable assumptions under this condition. A total of 3,115 participants were ultimately included, comprising 578 individuals with MetS and 2,537 without MetS. MetS was defined according to the harmonized criteria proposed by the joint interim statement, using waist circumference cut-off values applicable to the Chinese population¹⁴. Nineteen candidate variables were included, covering demographic characteristics, comorbidities, lifestyle factors, and laboratory indicators.

A separate small inpatient cohort was obtained from the inpatient information system of Yueyang People's Hospital of Hunan Normal University for exploratory external feasibility testing. Eligible patients meeting the identical inclusion and exclusion criteria were randomly sampled at a balanced 1:1 case-control ratio, resulting in 32 prevalent MetS inpatients and 32 controls without MetS (total $n = 64$). This small single-center inpatient cohort was used only for preliminary external feasibility testing rather than formal confirmatory validation; inpatient populations differ substantially from community-based CHARLS participants in comorbidity burden, inflammatory status, and medication exposure, which may introduce inherent selection bias.

Multivariable logistic regression analysis

All statistical analyses were performed using R software (version 4.5.0). Variables with a P value < 0.05 in univariable analysis were entered into the multivariable logistic regression model to identify factors independently associated with MetS. Multicollinearity was assessed using the variance inflation factor (VIF) by rms package (version 8.1-0). A logistic regression risk score (LRRS) was then constructed based on the regression coefficients.

LASSO regression analysis

LASSO regression was performed using the glmnet¹⁵ package (version 4.1–10) with a binomial family setting ($\alpha = 1$). Ten-fold cross-validation was applied, and variables were selected according to the lambda value corresponding to the minimum binomial deviance. A LASSO-based risk score (LARS) was subsequently calculated.

SVM modeling

Variables selected from the logistic regression and LASSO analyses were combined for SVM modeling. After standardization of input variables, a C-support vector classification model with a radial basis function kernel was developed by e1071¹⁶ package (version 1.7–16). SHapley Additive exPlanations (SHAP) were used to assess variable contributions to the final model by kernelshap¹⁷ package (version 0.9.1). All sequential variable screening and final SVM modeling were conducted within a single training dataset. Without nested cross-validation across all selection steps, the internal performance estimates may be subject to some degree of optimistic bias due to the high fitting flexibility of SVM.

Model performance evaluation

Model performance was evaluated in terms of discrimination, calibration, and potential clinical utility. Discrimination was assessed by pROC package (version 1.19.0.1) using receiver operating characteristic (ROC) curves and the area under the curve (AUC), and differences between models were compared.

Calibration was examined using bootstrap resampling. Decision curve analysis (DCA) was performed to assess net benefit across a range of threshold probabilities by *rmda*¹⁸ package (version 1.6).

Preliminary external feasibility testing

The best-performing model identified in the training cohort was applied to the inpatient cohort for preliminary external feasibility testing. Risk scores were calculated for each participant, and model performance was evaluated using AUC, sensitivity, specificity, and confusion matrix–based classification metrics to examine performance in an independent population.

Clinical correlation analysis

To further explore the clinical relevance of the final model, associations between the model-derived risk score and variables not included in the model were analyzed. For categorical variables, risk scores were compared across groups; for continuous variables, correlations with the risk score were assessed.

Statistical analysis

All data in this study were statistically analyzed and visualized using R 4.5.0 and GraphPad Prism 10.6.0 (San Diego, California, USA). Continuous variables were compared using the *t* test or Mann–Whitney *U* test, as appropriate according to data distribution, and categorical variables were compared using the chi-square test. Correlations were assessed using Spearman's method. All tests were two-sided, and $P < 0.05$ was considered statistically significant. Model development and reporting were conducted in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement¹⁹, with methodological quality assessed with reference to the Prediction model Risk of Bias Assessment Tool (PROBAST)²⁰. No artificial intelligence (AI) tools were adopted for raw data processing, model construction, statistical analysis, or manuscript original writing. AI language tools were only used for sentence polishing during revision.

Results

Baseline characteristics: participants with MetS had slightly older age, higher comorbidity burden, and distinct hematological profiles

As shown in Table S1, comparison between included ($N = 3,115$) and excluded ($N = 20,996$) participants revealed that included participants were older and had higher prevalences of heart disease, lung disease, liver disease, kidney disease, digestive disorders, smoking, and drinking (all $P < 0.0001$), whereas sex distribution was comparable ($P = 0.5806$). Laboratory parameters and MetS prevalence were not compared due to extensive missing values in the excluded sample.

A total of 3,115 participants were included in the training cohort, including 578 individuals with MetS (18.56%) and 2,537 without MetS (81.44%). Comparisons of the 19 candidate variables showed that participants with MetS differed significantly from those without MetS across several demographic, comorbidity, and laboratory characteristics (**Table S2**). Participants with MetS were slightly older than those without MetS (60.45 ± 9.40 vs. 59.68 ± 9.94 years, $P = 0.0359$), whereas sex distribution did not differ significantly between the two groups ($P = 0.6207$).

With respect to comorbidities, histories of heart disease, stroke, and memory disorder were more common among participants with MetS than among those without MetS (heart disease and stroke: both $P < 0.0001$; memory disorder: $P = 0.0080$). No significant between-group differences were observed for psychosis, lung disease, liver disease, kidney disease, digestive disorders, smoking history, or drinking history. In terms of laboratory findings, the MetS group had higher levels of WBC, platelet count (PLT), CRP, hematocrit (HCT), and HGB, whereas MCV was lower in the MetS group ($P = 0.0021$). Creatinine levels did not differ significantly between groups ($P = 0.6401$). Detailed baseline characteristics are presented in **Table S2**, and the overall study workflow is shown in Fig. 1.

Multivariable logistic regression identified six independent correlates of prevalent MetS: stroke history, heart disease history, CRP, HGB, WBC, and MCV

Univariable logistic regression analysis identified 9 of the 19 candidate variables as significantly associated with MetS; the corresponding odds ratios (ORs), 95% confidence intervals (CIs), and P values are shown in Fig. 2A. These variables were subsequently entered into the multivariable logistic regression model, from which 6 variables were retained for final model construction (Fig. 2B). Multicollinearity analysis showed that all VIF values were below 2.5, indicating no evidence of substantial collinearity among the selected predictors (Fig. 2C).

Among the retained variables, history of stroke (Fig. 2D, $P < 0.0001$), history of heart disease (Fig. 2E, $P < 0.0001$), higher CRP (Fig. 2F, $P < 0.0001$), higher HGB (Fig. 2G, $P < 0.0001$), and higher WBC (Fig. 2H, $P < 0.0001$) were positively associated with MetS, whereas MCV (Fig. 2I, $P = 0.0021$) was inversely associated with MetS. A logistic regression model was then developed based on these six variables.

LASSO regression retained nine MetS-associated variables, including memory disorder, PLT, and HCT in addition to the logistic regression-derived variables

All 19 candidate variables were entered into the LASSO regression model with 10-fold cross-validation. At the lambda value corresponding to the minimum binomial deviance ($\lambda_{\min}$), 9 variables with non-zero coefficients were retained for model construction (Fig. 3A-B). The absolute values of the regression coefficients are shown in Fig. 3C and were used to reflect the relative contribution of the selected variables.

The 9 retained variables included all 6 variables selected by multivariable logistic regression. Their directions of association were consistent across the two approaches. In addition, memory disorder

(Fig. 3D, $P = 0.0080$), PLT (Fig. 3E, $P = 0.0250$), and HCT (Fig. 3F, $P < 0.0001$) were retained in the LASSO model. A LASSO-based screening model was subsequently developed using these 9 variables.

SHAP analysis for SVM model interpretation: CRP, heart disease history, and stroke history were the most influential predictors

All variables retained in the multivariable logistic regression model were also included in the LASSO model, so the combined predictor set used for SVM modeling consisted of 9 variables. SHAP analysis showed that CRP had the highest mean absolute SHAP value, indicating the greatest contribution to the SVM model, followed by history of heart disease and history of stroke, whereas memory disorder had the smallest contribution (Fig. 4A).

The SHAP summary plot further illustrated the relationships between predictor values and their contributions to the model output (Fig. 4B). Higher CRP values were generally associated with higher SHAP values, indicating a greater contribution to MetS screening output, whereas lower MCV values were associated with higher estimated risk.

Performance comparison of the three screening models: distinct trade-offs between discrimination and calibration

The predictive performance of the SVM, LASSO, and logistic regression models was subsequently compared. As shown by the ROC curves (Fig. 4C), the SVM model achieved the highest discrimination, with an AUC of 0.767 (95% CI: 0.743–0.789), which was significantly higher than that of the LASSO model (AUC = 0.652, $P < 0.0001$) and the logistic regression model (AUC = 0.647, $P < 0.0001$). Decision curve analysis (Fig. 4D) further showed that the SVM model provided greater standardized net benefit than the other two models across most threshold probabilities and also outperformed the “treat-all” and “treat-none” strategies.

Confusion matrix heatmaps were generated using the caret package (Fig. 5A–C). The logistic regression model showed a sensitivity of 57.1%, specificity of 66.9%, and accuracy of 65.1%; the LASSO model showed a sensitivity of 64.5%, specificity of 60.6%, and accuracy of 61.3%; and the SVM model showed a sensitivity of 62.5%, specificity of 84.4%, and accuracy of 80.4%. Compared with the other two models, the SVM model showed higher specificity and overall accuracy.

Calibration analysis is shown in Fig. 5D–F. Although the mean absolute error (MAE) of the SVM model (0.106) was higher than that of the logistic regression and LASSO models (both 0.026), this indicates that the SVM model exhibited some degree of calibration deviation when estimating absolute probabilities. However, because the primary objective of this SVM model is to separate participants into high- and low-risk subgroups for clinical triage rather than to calculate precise individual absolute MetS probability, this mild miscalibration has limited interference with risk ranking performance. Taken together, the three models exhibited distinct performance profiles. The SVM model achieved superior discrimination, specificity, overall accuracy, and decision-analytic net benefit, but its calibration

performance was less favorable compared with the logistic regression and LASSO models. Given that the model is intended as a pre-screening and risk triage tool where accurate risk ranking and high specificity are prioritized, the SVM model was selected as the final model for subsequent feasibility testing.

External feasibility testing: SVM model maintained discriminative performance in the small inpatient test cohort

To further assess external performance, the SVM model developed in the training cohort was applied to the independent inpatient test cohort. As shown in Fig. 6A, the model achieved an AUC of 0.860 (95% CI: 0.748–0.936), suggesting that the model retained discriminative capacity in this small independent inpatient sample. Using the optimal cutoff value, a confusion matrix was generated (Fig. 6B), yielding a sensitivity of 71.9%, specificity of 75.0%, and accuracy of 73.4%. Notably, these classification metrics were derived from a 1:1 case-control balanced cohort; metrics strongly influenced by underlying disease prevalence, including positive predictive value and negative predictive value, are not generalizable to community populations with natural MetS prevalence.

Baseline characteristics differed in several indicators between the training and feasibility testing cohorts (**Table S3**). Although the SVM model retained discriminative performance across these distinct populations, this observation offers only weak preliminary clues about cross-population applicability. These findings should be interpreted cautiously given the small sample size of the feasibility testing cohort, as hospitalized patients carry distinct clinical profiles that may introduce selection bias relative to community participants.

Clinical correlation of SVM-based risk score: higher scores were associated with female sex, smoking history, and lung disease, with a screening nomogram developed

To further explore the practical relevance of the model-derived risk score, we examined the associations between the SVM-based risk score (SVMRS) and baseline characteristics. Correlation analysis for continuous variables showed a weak positive correlation between age and SVMRS ($R = 0.054$, $P = 0.0026$) (Fig. 6C). Comparisons across categorical variables showed that SVMRS was significantly higher in females ($P = 0.0001$; Fig. 6D), participants with a smoking history ($P = 0.0002$; Fig. 6E), and those with coexisting lung disease ($P = 0.0105$; Fig. 6F).

In addition, the nomogram (Fig. 6G) provided a quantitative mapping between SVMRS, total points, and the predicted probability of MetS, offering a more intuitive way to interpret the risk score in practice.

Discussion

Conventional MetS diagnosis requires complete measurement of waist circumference, fasting glucose and lipid profiles. In many primary-care and community screening settings, however, these assessments are not routinely performed in the absence of specific metabolic indications, and waist circumference is frequently missing from available records. Our model is not intended to replace formal MetS diagnosis; instead, it provides preliminary risk stratification using ubiquitously available hematological and inflammatory markers to triage high-risk individuals who warrant subsequent full metabolic work-up. Previous screening models for MetS have largely been built using conventional metabolic indicators together with demographic or lifestyle variables. However, several challenges remain in real-world application. First, some models rely on relatively complex predictor sets or data collection procedures, which may reduce feasibility in primary care and community settings. Second, heterogeneity in study populations, suboptimal handling of missing data, and limited external validation may restrict model transportability, particularly in populations with multimorbidity. Third, although machine learning approaches may improve predictive performance, limited interpretability and inconsistent reporting quality can reduce model credibility and reproducibility^{12,19,20}. These challenges are particularly relevant because current cardiometabolic prediction research is increasingly moving toward multidimensional and interpretable artificial intelligence frameworks, with recent studies in MetS classification highlighting the potential value of explainable models for individualized risk assessment²¹.

Importantly, complete blood count and CRP cannot replace fasting lipid and glucose testing for definitive MetS diagnosis; their value lies in preliminary triage when full metabolic testing data are unavailable, rather than superior discriminative power. It is important to clarify the intended use and conceptual basis of our model. The target outcome of the model remains prevalent MetS itself, rather than downstream cardiometabolic or inflammatory disease burden; the selected clinical and laboratory markers serve as readily available proxies strongly correlated with undiagnosed MetS in the general population. It was not designed as a causal prediction tool to identify early precursors of future MetS, but rather as a cross-sectional risk triage tool to identify individuals likely to already have MetS using routinely available indicators²². In this context, comorbidities and inflammatory markers serve as risk signals reflecting underlying metabolic burden, rather than as causal antecedents²³. The model is best viewed as a pre-screening tool to prioritize individuals for confirmatory diagnostic testing²⁴. Against this background, we developed and compared multiple screening models using routinely available clinical and laboratory indicators and identified an SVM model based on only 9 variables. This model achieved the most favorable discriminative and decision-analytic performance among the models evaluated while maintaining relative parsimony and a low data acquisition burden. In addition, the model-derived risk score was further explored in relation to baseline characteristics, which may help identify population subgroups with elevated estimated risk. Taken together, these findings provide preliminary proof-of-concept evidence that a relatively simple model based on accessible routine indicators has the potential to support MetS risk stratification in middle-aged and older adults. However, consistent with prior SVM-based MetS screening research using low-cost routine indicators, small-scale single-center validation is insufficient to confirm real-world clinical utility, and the current model remains at an early developmental stage rather than a deployable clinical screening tool²⁵.

In this study, age, history of heart disease, and history of stroke were retained as important risk indicators of MetS. These findings are broadly consistent with previous evidence showing that MetS is closely linked to cardiovascular and cerebrovascular risk in middle-aged and older adults^{1,26}. In addition, contemporary cardiovascular prevention frameworks increasingly emphasize obesity and related metabolic dysfunction as upstream drivers of cardiovascular risk across the life course, providing further context for the clinical relevance of these indicators²⁷. Age-related decline in metabolic regulation, together with chronic low-grade inflammation, may contribute to the accumulation of adverse metabolic profiles over time²⁸. Previous studies have also shown that MetS is associated with poorer cerebrovascular outcomes, including recurrent stroke and mortality²⁶. Integrating MetS screening into broader cardiometabolic risk management may help identify high-risk individuals in clinical practice. The additional analyses of the SVM-based risk score showed that higher estimated risk was observed in females, participants with a smoking history, and those with lung disease. The association with smoking history is consistent with prior evidence linking smoking to systemic inflammation, oxidative stress, and increased risk of MetS³⁴. Likewise, the association with lung disease may reflect the contribution of chronic inflammation or hypoxic burden to metabolic dysregulation and is in line with previous reports of a relatively high burden of MetS in patients with chronic pulmonary conditions³⁵. These findings suggest that the model may help identify population subgroups with less favorable metabolic risk profiles. The nomogram provided a quantitative translation of the model-derived risk score into estimated probability of MetS, which may facilitate interpretation of model output in practice; however, its utility in real screening pathways requires further evaluation.

Inflammatory markers appeared to contribute meaningfully to model performance. In particular, WBC and CRP were retained in the final model, supporting previous evidence that chronic low-grade inflammation is closely linked to MetS^{6,7,10}. Chronic inflammatory activation may contribute to insulin resistance, endothelial dysfunction, and broader metabolic imbalance, thereby promoting adverse cardiometabolic profiles^{6,7}. Recent mechanistic work has further emphasized the role of adipokine-mediated metabolic inflammation in orchestrating immune-metabolic crosstalk across adipose tissue and other metabolically active organs, which may partly explain why inflammatory markers provide additional predictive information beyond conventional metabolic indicators alone⁴¹. However, CRP and WBC are non-specific acute-phase reactants that can be elevated by diverse conditions unrelated to metabolic dysfunction, including infections, autoimmune diseases, chronic lung disease, medication use, and stress^{42,43}. This issue is particularly relevant in the context of MetS screening: elevated CRP or WBC levels in a given individual may reflect non-metabolic inflammatory states rather than underlying MetS-related adipose-tissue inflammation. Such non-specificity introduces potential classification bias, as the model may assign elevated MetS risk scores to individuals whose inflammatory marker elevations are driven by unrelated conditions, leading to false-positive risk classifications^{44,45}. Although we adjusted for major comorbidities in our multivariable models, residual confounding from unmeasured inflammatory conditions cannot be fully excluded. Furthermore, because our external feasibility cohort comprised hospitalized patients with inherently higher inflammatory burden, this limitation may be particularly

pronounced in the testing setting. Because these markers are routinely measured and relatively inexpensive, their inclusion theoretically supports potential applicability for community-based screening and risk stratification. However, rather than being interpreted as specific drivers of MetS pathophysiology, CRP and WBC in our model should be viewed as non-specific risk signals that reflect a composite of metabolic and non-metabolic inflammatory states, and their screening contribution should be interpreted in conjunction with clinical context^{42,44}. Future studies incorporating more specific inflammatory biomarkers, such as adipokines and cytokines, may help disentangle MetS-related inflammation from other sources and improve screening specificity. At the same time, the present study does not establish whether these markers are mechanistically involved in disease development or simply reflect broader metabolic and inflammatory burden. Their role in screening should therefore be interpreted primarily from a practical rather than a causal perspective.

Common hematological parameters also showed associations with model output. HGB was positively associated with MetS, whereas MCV showed an inverse association. Previous population-based studies have similarly suggested that HGB and MCV may be related to metabolic risk⁴⁷. These indicators may capture aspects of hypoxic stress, oxidative stress, hemorheological changes, or underlying metabolic disturbances that are not fully reflected by conventional metabolic variables alone^{48,49}. Nevertheless, these findings should be interpreted cautiously. In the present study, the effect size of HGB was small, suggesting that its contribution may reflect a modest gradient at the population level rather than a strong independent signal. Moreover, both HGB and MCV may be influenced by other clinical conditions, including hydration status, chronic pulmonary disease, renal dysfunction, or abnormalities in iron metabolism^{48,49}. Therefore, these variables are better regarded as readily available risk markers than as direct mechanistic indicators. Future studies incorporating more specific markers of iron metabolism and longitudinal designs may help clarify their role in MetS risk stratification.

This study has several limitations. First, the cross-sectional design precludes causal inference. Because the predictor and outcome were measured concurrently, we cannot establish temporal precedence. Consequently, the model should not be viewed as a prognostic tool for incident MetS, but rather as a screening aid to identify individuals with a high probability of currently having MetS, who may then be prioritized for confirmatory testing²⁴. The associations reported herein are strictly correlational. Second, residual confounding and missing-data bias may affect internal validity. Despite including a broad range of variables, we could not adjust for unmeasured confounders such as dietary patterns, physical activity, socioeconomic status, or medication use^{20,51}. More importantly, missingness in the CHARLS cohort was not random: participants retained for complete-case analysis were older and had higher comorbidity burdens and different lifestyle profiles than those excluded, whereas sex distribution was similar (Table S1). Laboratory parameters and MetS prevalence could not be compared due to extensive missingness in the excluded group. We deliberately chose complete-case analysis over multiple imputation because the missing pattern appeared more consistent with missing-not-at-random (MNAR) than with missing-at-random (MAR); under MNAR, imputation would rely on untestable assumptions and could introduce less discernible bias⁵². Nevertheless, this analytical decision inevitably narrows the study population and may

compromise generalizability to the broader older Chinese population. Third, the model's external generalizability is constrained by both the test cohort's design and the target population's scope. The independent test set comprised only 64 hospitalized patients (32 case-control pairs) from a single center, with an artificial 1:1 prevalence ratio. Such a small, balanced sample renders all performance metrics, including AUC, sensitivity, and specificity, statistically fragile, and precludes reliable estimation of prevalence-dependent indices such as positive/negative predictive values, population-level calibration, and real-world net benefit. Thus, the numerically higher AUC observed in this inpatient cohort should be interpreted as a preliminary feasibility signal rather than as robust evidence of transportability⁵³. Moreover, the training and test populations differed fundamentally: the former came from community-dwelling middle-aged and older adults, whereas the latter consisted of hospitalized patients with heavier comorbidity, inflammatory, and medication burdens. This demographic shift introduces inherent selection bias and distributional divergence, making it difficult to disentangle genuine model stability from chance fluctuations. Beyond this, our findings derive exclusively from Chinese adults aged 45 years or older; whether the model performs similarly in other geographic regions, healthcare systems, or ethnic groups remains unknown. Future large-scale, prospective evaluation in unselected community cohorts with natural disease prevalence is essential to determine whether this model can offer meaningful clinical utility⁵⁴. Fourth, although the SVM model showed superior discrimination, its calibration was suboptimal compared with logistic regression and LASSO. Poorly calibrated probabilities may limit the model's utility when precise absolute risk estimates are needed⁵⁵. However, because we intend the model for pre-screening risk triage rather than for making dichotomous clinical decisions based on exact probability cut-offs, the practical impact of miscalibration may be partially mitigated⁵⁶. Nonetheless, before any real-world deployment, external recalibration of the SVM model is strongly recommended to align predicted probabilities with observed outcomes in the target population. Finally, the inflammatory markers included (CRP and WBC) lack pathophysiological specificity. Elevations in these acute-phase reactants can result from infections, autoimmune diseases, chronic respiratory disorders, medication effects, or psychological stress, conditions not necessarily reflective of metabolic inflammation^{42–45}. Therefore, the model may assign falsely elevated risk scores to individuals whose inflammatory markers are driven by non-metabolic causes, potentially introducing bias in heterogeneous populations. Future studies incorporating more specific adipokines or cytokines may help disentangle metabolic from non-metabolic inflammation and improve screening accuracy. Taken together, these limitations underscore that our current model remains at an exploratory, proof-of-concept stage. While it offers a low-cost, accessible risk triage strategy, it is not yet ready for population-wide screening. Rigorous external validation and recalibration in diverse, representative cohorts are indispensable next steps before any clinical translation can be considered.

Conclusions

Using routinely available clinical and laboratory indicators, we developed and preliminarily evaluated a relatively simple SVM-based screening model for MetS. Age, history of heart disease, history of stroke, WBC, CRP, and HGB were positively associated with MetS, whereas MCV was inversely associated with

estimated risk. Higher model-derived risk scores were also observed in females, participants with a smoking history, and those with lung disease, offering correlational clues that may inform future population risk stratification research. This study provides preliminary proof-of-concept for a low-cost and accessible approach to pre-screening and risk triage, particularly in middle-aged and older adults. However, given the small external test sample, artificially balanced case-control ratio, suboptimal calibration, and currently undefined clinically actionable thresholds, the model is not yet ready for population-level screening or stratified management. Further large-scale validation in representative community populations, along with external recalibration and threshold determination, is required before any practical clinical application can be considered.

Declarations

Acknowledgments

We would like to thank the patients who provided clinical data for this study and the dedicated internists involved in their clinical care. We also sincerely acknowledge the CHARLS research team for generously sharing their rigorously collected high-quality cohort data in the public domain, which enabled the model development, validation and comparative analysis of this work.

Author Contributions

Conceptualization: Wei Liu, Shangfu Li, Sen Lu, Lingge Yang; Methodology: Wei Liu, Shangfu Li, Yiqun Zeng, Wei Dai, Jing Peng, Xiaofang Wang, Qingping Chen, Sen Lu, Lingge Yang; Validation: Wei Liu, Shangfu Li, Yiqun Zeng, Jing Peng, Qingping Chen, Duanlin Du, Sen Lu, Lingge Yang; Formal Analysis: Shangfu Li, Wei Dai, Wei Li, Duanlin Du, Sen Lu, Lingge Yang; Investigation: Wei Liu, Shangfu Li, Yiqun Zeng, Wei Dai, Jing Peng, Xiaofang Wang, Lingge Yang; Resources: Wei Liu, Shangfu Li, Wei Dai, Qingping Chen, Wei Li, Duanlin Du, Sen Lu, Lingge Yang; Data curation: Wei Liu, Yiqun Zeng, Jing Peng, Xiaofang Wang, Wei Li, Duanlin Du; Supervision: Shangfu Li, Sen Lu, Lingge Yang; Project administration: Sen Lu, Lingge Yang. Writing—original draft preparation: All authors; Writing—review and editing: All authors; All authors have read and approved the final version of the manuscript.

Funding

This study did not receive any external financial support.

Data Availability Statement

The CHARLS data used in this study are publicly available from the China Health and Retirement Longitudinal Study website (<https://charls.pku.edu.cn/>). The inpatient data from Yueyang People's Hospital of Hunan Normal University supporting the findings of this study are not publicly available due to patient privacy considerations and institutional restrictions. Access to these data may be granted by the corresponding author upon reasonable request, subject to approval by the participating institution.

Ethics statement

Ethical approval for this study was obtained from the institutional ethics committee of Yueyang People's Hospital of Hunan Normal University (Approval No. 2026003).

Consent for publication

Not applicable.

Competing interests

The authors report no financial or personal conflicts of interest that could have influenced the conduct or interpretation of this study.

References

1. Neeland IJ, et al. Metabolic syndrome. *Nat reviews Disease primers*. 2024;10(1):77.
2. Lingvay I, Cohen RV, Roux CWL, Sumithran P. Obesity in adults. *Lancet*. 2024;404(10456):972–87.
3. Zhang H, et al. Global burden of metabolic diseases, 1990–2021. *Metabolism*. 2024;160:155999.
4. Liu B, Chen G, Zhao R, Huang D, Tao L. Temporal trends in the prevalence of metabolic syndrome among middle-aged and elderly adults from 2011 to 2015 in china: the china health and retirement longitudinal study (charls). *BMC Public Health*. 2021;21(1):1045.
5. Ge H, et al. The prevalence and associated factors of metabolic syndrome in chinese aging population. *Sci Rep*. 2020;10(1):20034.
6. Rohm TV, Meier DT, Olefsky JM, Donath MY. Inflammation in obesity, diabetes, and related disorders. *Immunity*. 2022;55(1):31–55.
7. Soták M, Clark M, Suur BE, Börgeson E. Inflammation and resolution in obesity. *Nat Rev Endocrinol*. 2025;21(1):45–61.
8. Yang S, et al. Development and validation of an age-sex-ethnicity-specific metabolic syndrome score in the chinese adults. *Nat Commun*. 2023;14(1):6988.
9. Coral DE, et al. Subclassification of obesity for precision prediction of cardiometabolic diseases. *Nat Med*. 2025;31(2):534–43.
10. Xue Q, et al. Association between baseline and changes in high-sensitive c-reactive protein and metabolic syndrome: a nationwide cohort study and meta-analysis. *Nutr Metab*. 2022;19(1):2.
11. Buergel T, et al. Metabolomic profiles predict individual multidisease outcomes. *Nat Med*. 2022;28(11):2309–20.
12. Efthimiou O, et al. Developing clinical prediction models: a step-by-step guide. *BMJ*. 2024;386:e78276.
13. 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(1):61–8.

14. Alberti KGMM, et al. Harmonizing the metabolic syndrome: a joint interim statement of the international diabetes federation task force on epidemiology and prevention; National heart, lung, and blood institute; American heart association; World heart federation; International atherosclerosis society; And international association for the study of obesity. *Circulation*. 2009;120(16):1640–5.
15. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
16. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–97.
17. Amann J, et al. To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems. *PLOS Digit health*. 2022;1(2):e16.
18. Sadatsafavi M, et al. Moving beyond auc: decision curve analysis for quantifying net benefit of risk prediction models. *Eur Respir J*. 2021;58(5):2101186.
19. Collins GS, et al. Tripod + ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e78378.
20. Wolff RF, et al. Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51–8.
21. Lee TK, Kim SY, Choi HJ, Choe EK, Sohn K. Vision transformer based interpretable metabolic syndrome classification using retinal images. *NPJ Digit Med*. 2025;8(1):205.
22. Li Z, Wu W, Kang H. Machine learning-driven metabolic syndrome prediction: an international cohort validation study. *Healthc (Basel Switzerland)*. 2024;12(24):2527.
23. Behkamal B, et al. A machine learning-based framework for predicting metabolic syndrome using serum liver function tests and high-sensitivity c-reactive protein. *Sci Rep*. 2025;15(1):21725.
24. He Y, Chen Z, Tang Z, Qin Y, Wang F. Machine learning with the body roundness index and associated indicators: a new approach to predicting metabolic syndrome. *BMC Public Health*. 2025;25(1):2608.
25. Mohseni-Takaloo S, Mozaffari-Khosravi H, Mohseni H, Mirzaei M, Hosseinzadeh M. Metabolic syndrome prediction using non-invasive and dietary parameters based on a support vector machine. *Nutr metabolism Cardiovasc Dis*. 2024;34(1):126–35.
26. Zhang F, et al. Association of metabolic syndrome and its components with risk of stroke recurrence and mortality: a meta-analysis. *Neurology*. 2021;97(7):e695–705.
27. Powell-Wiley TM, et al. Obesity and cardiovascular disease: a scientific statement from the american heart association. *Circulation*. 2021;143(21):e984–1010.
28. Franceschi C, Garagnani P, Parini P, Giuliani C, Santoro A. Inflammaging: a new immune-metabolic viewpoint for age-related diseases. *Nat Rev Endocrinol*. 2018;14(10):576–90.
29. Arshadi M, et al. The association between alcohol and metabolic syndrome: a systematic review and meta-analysis. *Diabetol Metab Syndr*. 2025;17(1):250.
30. Alrabbaie H, et al. The prevalence of metabolic syndrome in chronic obstructive pulmonary disease: a systematic review and meta-analysis. *Chronic Respir Dis*. 2025;22:201196782.

31. Tilg H, Ianiro G, Gasbarrini A, Adolph TE. Adipokines: masterminds of metabolic inflammation. *Nat Rev Immunol*. 2025;25(4):250–65.
32. Pepys MB, Hirschfield GM. C-reactive protein: a critical update. *J Clin Investig*. 2003;111(12):1805–12.
33. Emerging RFC, et al. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *Lancet*. 2010;375(9709):132–40.
34. Ridker PM. C-reactive protein and the prediction of cardiovascular events among those at intermediate risk: moving an inflammatory hypothesis toward consensus. *J Am Coll Cardiol*. 2007;49(21):2129–38.
35. Oda E, Kawai R. Comparison between high-sensitivity c-reactive protein (hs-crp) and white blood cell count (wbc) as an inflammatory component of metabolic syndrome in japanese. *Intern Med*. 2010;49(2):117–24.
36. Tanaka M, et al. Combined effect of hemoglobin and mean corpuscular volume levels on incident metabolic syndrome: a population-based cohort study. *Clin Nutr ESPEN*. 2020;40:314–9.
37. Harrison AV, Lorenzo FR, McClain DA. Iron and the pathophysiology of diabetes. *Annu Rev Physiol*. 2023;85:339–62.
38. Hilton C, Sabaratnam R, Drakesmith H, Karpe F. Iron, glucose and fat metabolism and obesity: an intertwined relationship. *Int J Obes*. 2023;47(7):554–63.
39. van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. *Int J Epidemiol*. 2020;49(1):338–47.
40. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med*. 2010;29(28):2920–31.
41. Riley RD, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ*. 2024;384:e74821.
42. Park J, Jeong I, Ko G, Jeong S, Lee H. Development of a predictive model for metabolic syndrome using noninvasive data and its cardiovascular disease risk assessments: multicohort validation study. *J Med Internet Res*. 2025;27:e67525.
43. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Mak*. 2015;35(2):162–9.
44. Van Calster B, et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167–76.

Figures

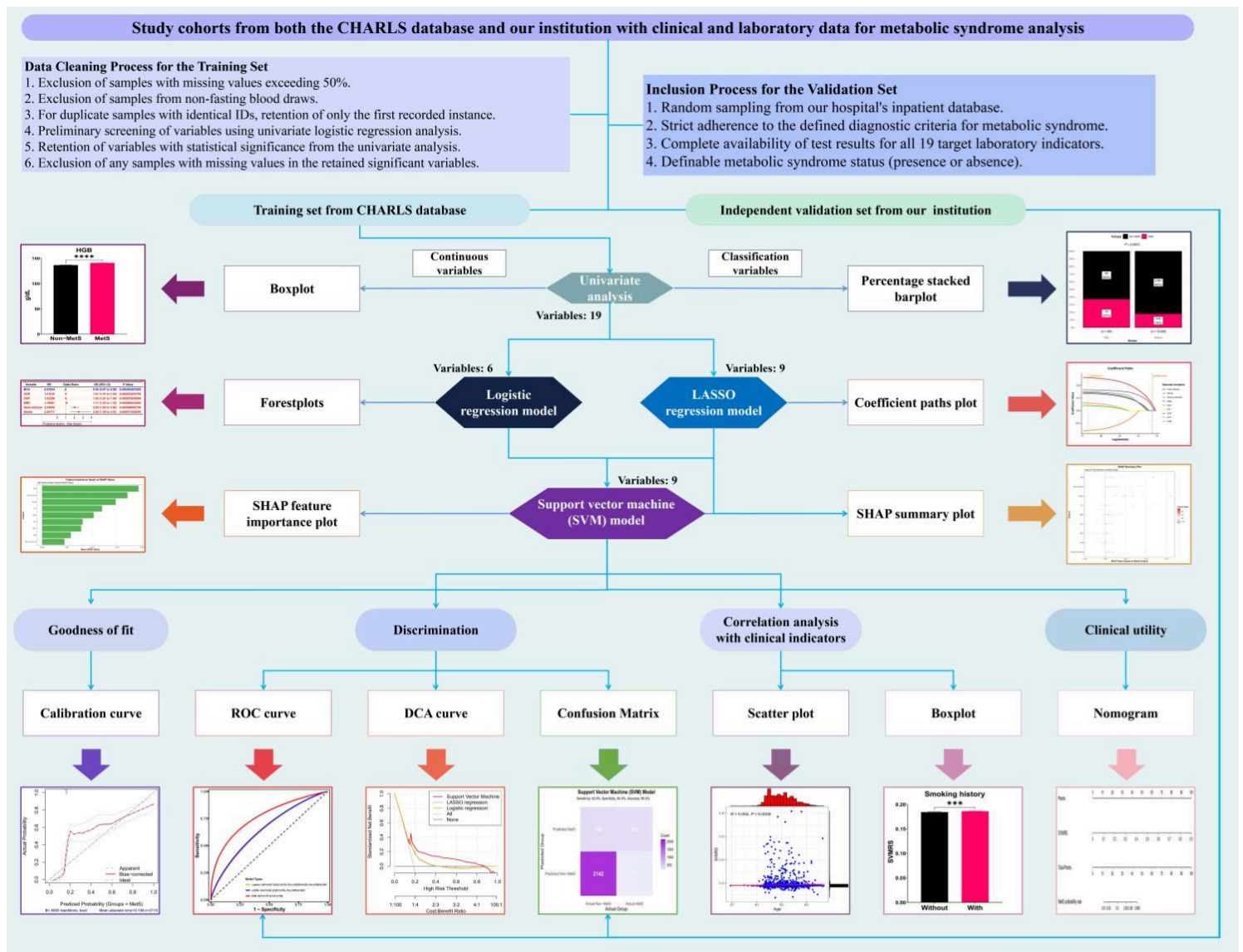


Figure 1

Flowchart of this Study. Data from the CHARLS cohort were used to construct the training set, and an institutional inpatient cohort was used for preliminary external feasibility testing. Candidate variables were screened and used to develop logistic regression, LASSO, and SVM models. Model performance was evaluated by ROC analysis, calibration, and decision curve analysis. SHAP was used to assess model interpretability, and additional analyses included confusion matrix evaluation and clinical correlation analysis.

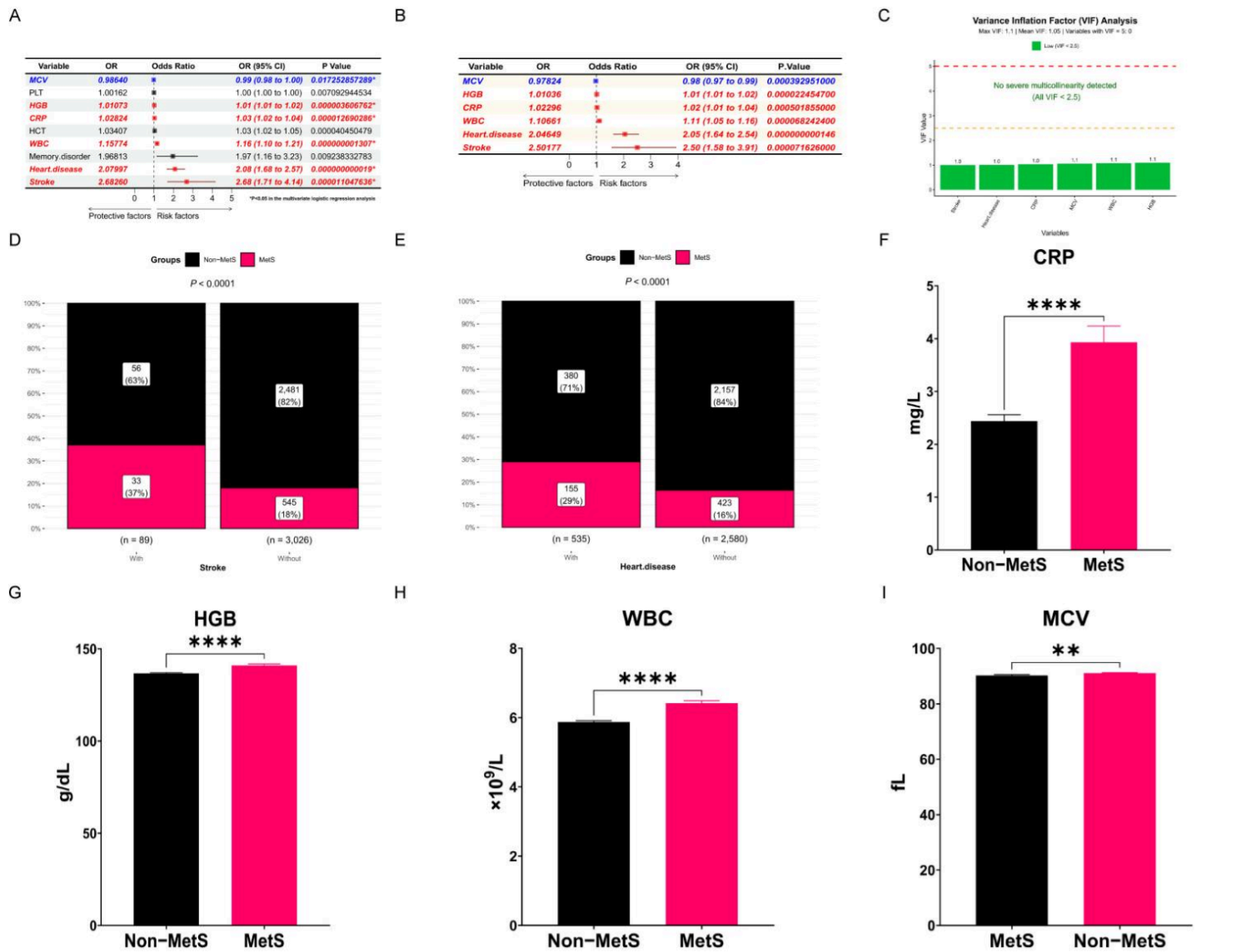


Figure 2

Construction of the logistic regression model. (A) Forest plot of the univariable logistic regression analysis, showing the odds ratios (ORs), 95% confidence intervals (CIs), and P values for the 9 variables significantly associated with MetS among the 19 candidate variables. Red indicates positive associations and blue indicates inverse associations. (B) Forest plot of the multivariable logistic regression analysis, showing the ORs, 95% CIs, and P values for the 6 variables retained in the final model. (C) Bar plot of VIF values for the 6 selected variables, all of which were below 2.5, indicating no substantial multicollinearity. (D–E) Stacked bar plots showing the proportions of MetS and non-MetS participants according to stroke history (D) and heart disease history (E). (F–I) Bar plots comparing CRP (F), HGB (G), WBC (H), and MCV (I) levels between participants with and without MetS. CRP, HGB, and WBC were higher in the MetS group, whereas MCV was lower. ** $P < 0.01$; **** $P < 0.0001$.

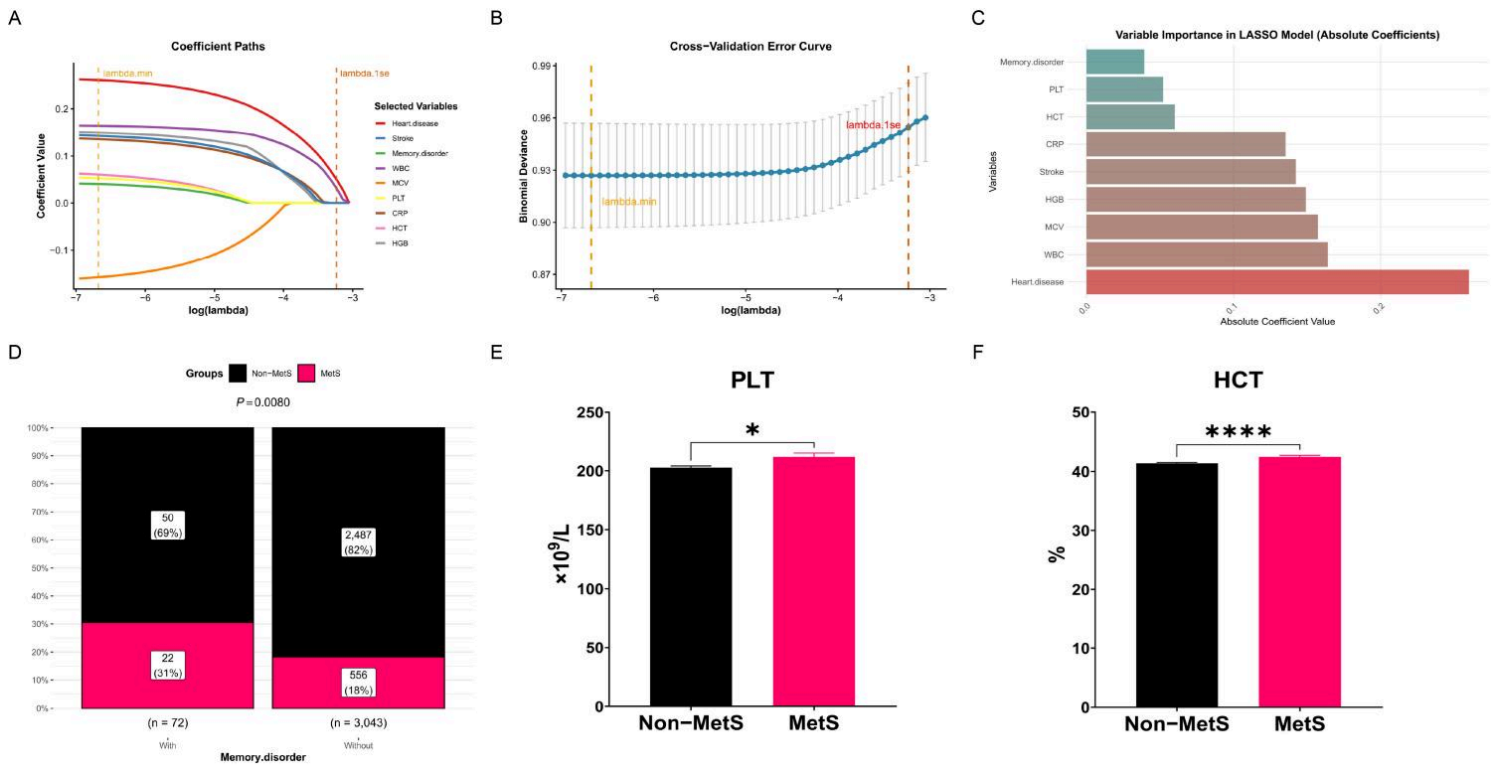


Figure 3

Construction of the LASSO regression model. (A) Coefficient profile plot of the LASSO regression model. The x-axis represents $\log(\lambda)$, and the y-axis represents the coefficient values. As λ increased, the coefficients gradually shrank toward zero; 9 variables with non-zero coefficients were retained at $\lambda_{\min}$. (B) Ten-fold cross-validation plot for selection of the optimal tuning parameter. The x-axis represents $\log(\lambda)$, and the y-axis represents binomial deviance. The yellow dashed line indicates $\lambda_{\min}$, at which the binomial deviance was minimized. (C) Bar plot showing the absolute values of the LASSO coefficients for the retained variables, with larger absolute values indicating greater relative contribution. (D) Stacked bar plot showing the proportions of participants with and without MetS according to memory disorder status. (E-F) Bar plots comparing platelet count (PLT) (E) and hematocrit (HCT) (F) between participants with and without MetS. * $P < 0.05$; **** $P < 0.0001$.

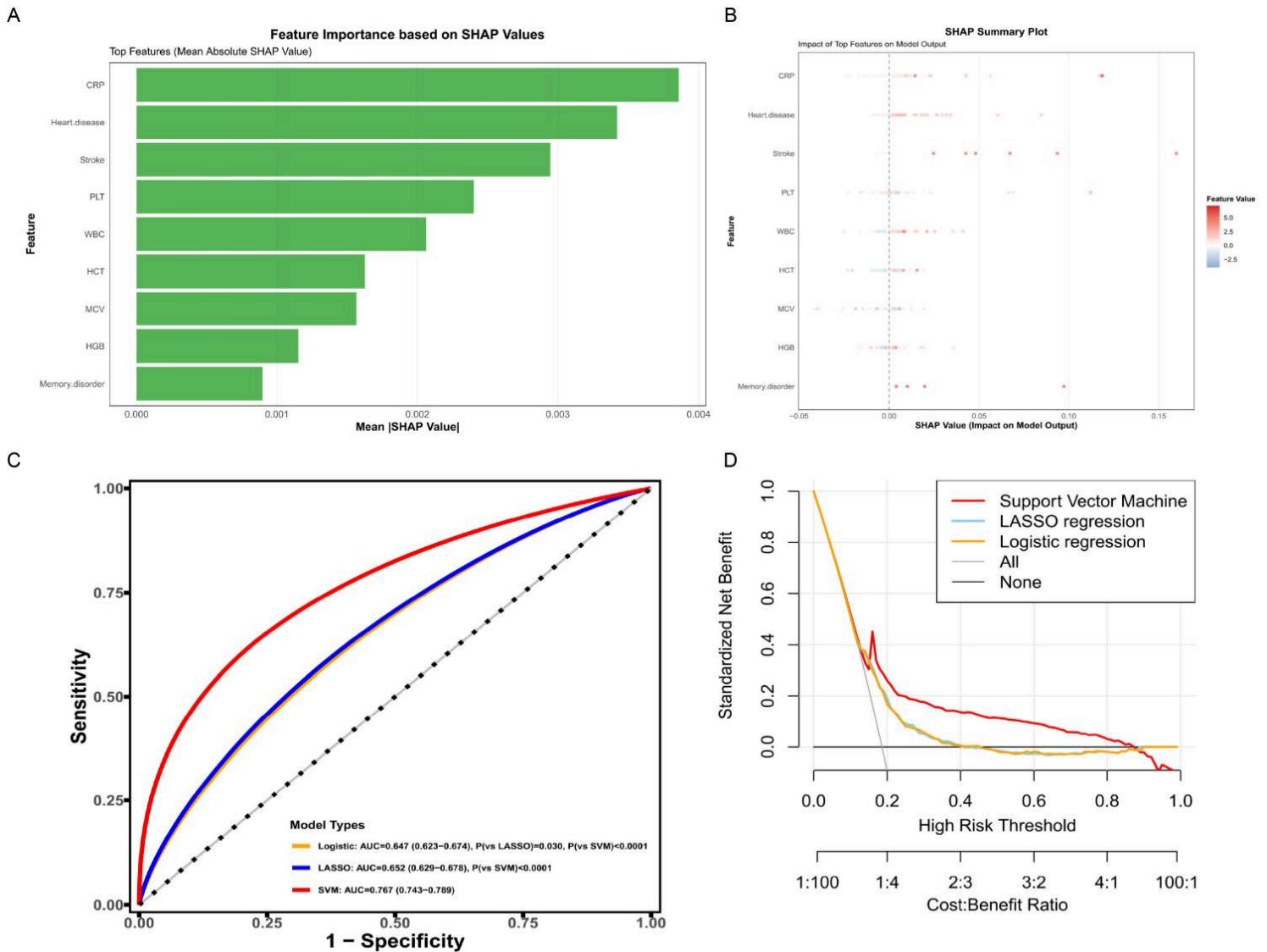


Figure 4

Construction of the support vector machine (SVM) model and performance comparison of the three models. (A) SHAP feature importance plot showing the mean absolute SHAP values of the predictors included in the SVM model. CRP had the largest contribution to the model. (B) SHAP summary plot showing the relationships between feature values and SHAP values. Red indicates higher feature values and blue indicates lower feature values. (C) ROC curves comparing the discrimination of the logistic regression (yellow), LASSO regression (blue), and SVM (red) models. The SVM model achieved the highest area under the curve (AUC = 0.767) and outperformed the other two models ($P < 0.0001$). (D) Decision curve analysis showing the standardized net benefit of the three models across a range of threshold probabilities. The SVM model showed greater net benefit than the logistic regression and LASSO models across most threshold ranges.

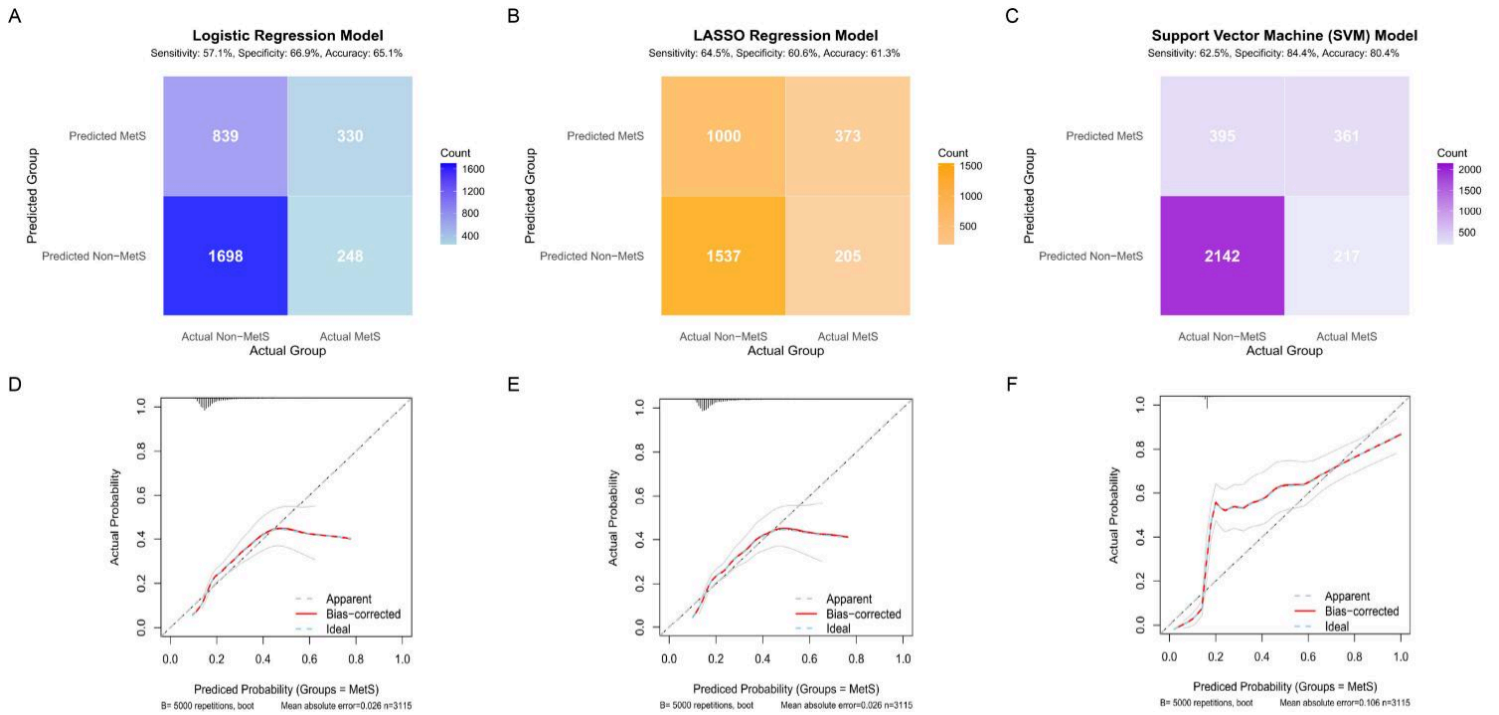


Figure 5

Confusion matrices and calibration plots of the three screening models. (A–C) Confusion matrix heatmaps for the logistic regression model (A), LASSO model (B), and SVM model (C). Color intensity indicates the number of samples classified into each observed and predicted category. The logistic regression model showed a sensitivity of 57.1%, specificity of 66.9%, and accuracy of 65.1%; the LASSO model showed a sensitivity of 64.5%, specificity of 60.6%, and accuracy of 61.3%; and the SVM model showed a sensitivity of 62.5%, specificity of 84.4%, and accuracy of 80.4%. (D–F) Calibration plots for the logistic regression model (D), LASSO model (E), and SVM model (F). The red solid line represents the apparent calibration curve, the dashed line represents the bootstrap-corrected calibration curve (B = 5000), and the gray diagonal line indicates ideal calibration. The mean absolute error (MAE) values were 0.026 for the logistic regression model, 0.026 for the LASSO model, and 0.106 for the SVM model.

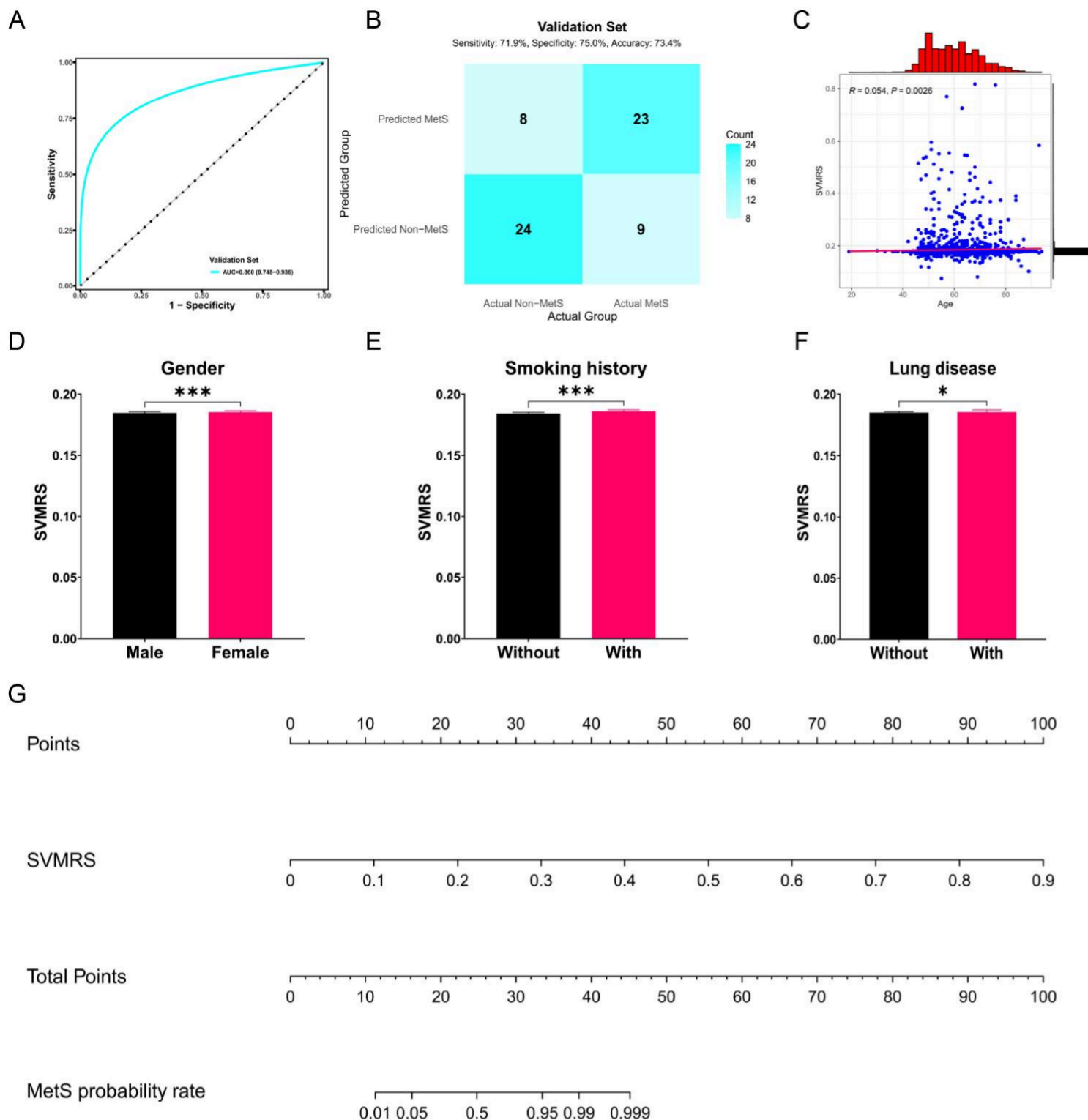


Figure 6

External feasibility testing and additional analyses of the SVM-based screening model for MetS. (A) ROC curve of the SVM model in the independent feasibility testing cohort, showing an area under the curve (AUC) of 0.860 (95% CI: 0.748–0.936). (B) Confusion matrix of the SVM model in the feasibility testing cohort using the optimal cutoff value, with a sensitivity of 71.9%, specificity of 75.0%, and accuracy of 73.4%. (C) Scatter plot showing the correlation between age and the SVMRS, indicating a

weak positive correlation ($R = 0.054$, $P = 0.0026$). (D–F) Comparisons of SVMRS according to sex (D), smoking history (E), and lung disease status (F). SVMRS was higher in females, participants with a smoking history, and those with lung disease. (G) Nomogram showing the mapping of SVMRS to total points and predicted probability of MetS.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial.docx](#)