Diagnostic Accuracy Study

Medicine

OPEN

# Machine learning-based risk assessment for cardiovascular diseases in patients with chronic lung diseases

Huiming Xi, MM[a], Qingxin Kang, MM[a], Xunsheng Jiang, MM[a,*]

**Abstract**

The association between chronic lung diseases (CLDs) and the risk of cardiovascular diseases (CVDs) has been extensively recognized. Nevertheless, conventional approaches for CVD risk evaluation cannot fully capture the risk factors (RFs) related to CLDs. This research sought to construct a CLD-specific CVD risk prediction model based on machine learning models and evaluate the prediction performance. The cross-sectional study design was adopted with data retrieved from Waves 1 and 3 of the China Health and Retirement Longitudinal Study, including 1357 participants. Multiple RFs were integrated into the models, including conventional RFs for CVDs, pulmonary function indicators, physical features, and measures of quality of life and psychological state. Four machine learning algorithms, including extreme gradient boosting (XGBoost), logistic regression, random forest, and support vector machine, were evaluated for prediction performance. The XGBoost model displayed superior performance to machine learning algorithms for predictive accuracy (area under the receiver operating characteristic curve [AUC]: 0.788, accuracy: 0.716, sensitivity: 0.615, specificity: 0.803). This model pinpointed the top 5 RFs for CLD-specific CVD RFs: body mass index, age, C-reactive protein, uric acid, and grip strength. Moreover, the prediction performance of the random forest model (AUC: 0.709, accuracy: 0.633) was higher relative to the logistic regression (AUC: 0.619, accuracy: 0.584) and support vector machine (AUC: 0.584, accuracy: 0.548) models. Nonetheless, these models performed less favorably compared to the XGBoost model. The XGBoost model presented the most accurate predictions for CLD-specific CVD risk. This multidimensional risk assessment approach offers a promising avenue for the establishment of personalized prevention strategies targeting CVD in patients with CLDs.

**Abbreviations:** ADL = activities of daily living, AUC = area under the receiver operating characteristic curve, BMI = body mass index, CESD-10 = 10-item Center for Epidemiologic Studies Depression Scale, CLDs = chronic lung diseases, CRP = C-reactive protein, CVD = cardiovascular disease, DLP = dyslipidemia, DM = diabetes mellitus, GS = grip strength, HTN = hypertension, IADL = instrumental activities of daily living, LR = logistic regression, QoL = quality of life, RF = risk factor, SVM = support vector machine, XGBoost = extreme Gradient boosting.

**Keywords:** cardiovascular disease, chronic lung disease, machine learning, prediction model, risk factors, XGBoost algorithm

## 1. Introduction

Chronic lung disease (CLD) is a group of pulmonary conditions predominantly characterized by sustained airflow limitation or a gradual reduction in pulmonary functions. CLDs pose considerable challenges to public health globally.[1,2] Existing statistics uncover a yearly rise in the prevalence of CLDs, positioning it as the third leading death-related cause on a global scale. CLDs led to 4 million deaths, making a 28.5% elevation from 2017, though the age-standardized mortality has dropped by 41.7%. Notably, chronic obstructive pulmonary disease (COPD) is one of the most widespread subtypes, contributing to an estimated 3.3 million (2.9–3.6) deaths worldwide.[2–4] In China alone, the prevalence of COPD is approximately 8.6% of the population, thereby affecting about 100 million individuals.[5] In addition to pulmonary dysfunction, CLDs frequently present with a range of complications, with cardiovascular diseases (CVDs) being one of the most prevalent and severe comorbidities.[6–8]

HX and QK contributed equally to this work.

All participants in the original CHARLS survey signed informed consent forms, and those unable to write marked the consent forms.

The authors have no funding and conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are publicly available.

This study was conducted using publicly available summary data. Ethical approval was granted by the Institutional Review Board (IRB) of Peking University, with the approval number IRB00001052-11015.

Supplemental Digital Content is available for this article.

[a] Department of Pulmonary and Critical Care Medicine, Nanchang People's Hospital, Nanchang, China.

* Correspondence: Xunsheng Jiang, Department of Pulmonary and Critical Care Medicine, Nanchang People's Hospital, Nanchang 330009, Jiangxi, China (e-mail: 18720991667@163.com).

How to cite this article: Xi H, Kang Q, Jiang X. Machine learning-based risk assessment for cardiovascular diseases in patients with chronic lung diseases. Medicine 2025;104:10(e41672).

Received: 27 October 2024 / Received in final form: 6 February 2025 / Accepted: 7 February 2025

http://dx.doi.org/10.1097/MD.0000000000041672

A growing body of evidence has delineated that individuals with CLDs encounter a markedly higher risk of CVDs than those without the conditions.[9–11] A large epidemiological analysis has documented that the risk of myocardial infarction is 2 to 3 times higher, and that of heart failure is 4 times higher in individuals with COPD relative to the general population.[9] The augmented risk prominently contributes to escalated mortality and compromises the quality of life (QoL), thereby adding substantial economic burdens on healthcare systems. Estimates from the United States point out that the annual direct costs for the COPD-specific CVD population are 135% higher than those without CVDs, and total COPD-related healthcare expenses are 38% higher.[12]

Inflammatory reaction is a primary mechanism in CLDs.[7] CLDs, especially COPD, are commonly accompanied by continuous low-level systemic inflammation.[13] The chronic inflammation extends beyond the pulmonary system to the vascular systems, thus causing vascular endothelial dysfunction and expediting the progression of atherosclerosis.[14] Another crucial mechanism in the progression of CLDs is oxidative stress, which is notably escalated in individuals with CLD.[15] Excessive accumulation of free radicals causes direct vascular endothelial injury and expedites atherosclerosis *via* the oxidation of low-density lipoprotein.[14] In addition, the dysregulation of the autonomic nervous system is also implicated in the progression of CLDs.[16] CLDs stimulate the activation of the sustained sympathetic nervous system, leading to elevated susceptibility to arrhythmias and hypertension (HTN).[17] Furthermore, the impacts of pulmonary HTN are a pivotal factor warranting clarification in the context of CLDs. CLDs frequently cause a sustained rise in pulmonary arterial pressure, which intensifies additional strain on the right side of the heart and eventually triggers right heart failure.[18] Hypoxemia, a prevalent symptom in the context of CLDs, may also stimulate myocardial hypoxia and trigger the risks of myocardial infarction and arrhythmias.[19] Intriguingly, CLDs and CVDs share risk factors (RFs), such as smoking status, sedentary lifestyle, and poor dietary choices.[7] The coexistence of the RFs facilitates the risk of both diseases, further fostering health burden among affected individuals.

The risk of CVD comorbidities in individuals with CLDs and the potentially severe outcomes make it urgent to develop strategies for prompt identification and prevention of CVDs. Conventional risk assessment approaches, such as Framingham risk scoring,[20] have been developed for the general population, but they may fail to accurately identify special risk profiles of individuals affected by CLDs.[21] Moreover, conventional models typically consider only a narrow range of RFs, thus constraining their ability to capture the complex pathophysiological state associated with CLDs.

The rapid evolution of artificial intelligence has led to an increasing application of machine learning (ML) in the medical field in recent years.[22–24] ML algorithms can process vast, complex data, recognize nonlinear patterns, and automatically extract the patterns and trends of data. The capability of ML to process complex data makes it a robust tool for risk prediction. In terms of risk prediction for CVDs, a diverse array of studies have documented the superiority of ML models over conventional risk assessment tools. One such study, based on UK Biobank data, has underscored that ML algorithms augmented the accuracy of risk prediction for 10-year CVDs by 4.9% relative to the conventional QRISK3 model.[25] However, current investigations into ML models on CVD-specific risk prediction in the context of CLDs are still relatively scarce. Existing studies have predominantly focused on individuals affected by COPD, with fewer explorations extending to other forms of CLDs. Furthermore, the algorithms developed thus far have largely been based on conventional clinical parameters and demographic data, whereas such pivotal variables as QoL and psychological state in CVD risk remain uncharacterized.

Given the aforementioned context, this research sought to generate a CVD-specific risk prediction model for CLD patients through ML algorithms. Our findings introduce several innovations to foster prediction performance. The set of RFs was expanded beyond conventional CVD risk indicators. Variables reflecting patients' QoL and psychological state were included in the study, such as activities of daily living (ADL), instrumental activities of daily living (IADL), and the 10-item Center for Epidemiologic Studies Depression Scale (CESD-10). The complex health conditions of patients with CLDs were captured in this context. In addition to the comprehensive inclusion of RFs, we examined the predictive performance of ML algorithms. 4 ML algorithms: logistic regression (LR),[26] extreme gradient boosting (XGBoost),[27] random forest,[28] and support vector machine (SVM)[29] to pinpoint the optimal algorithm in this specific population. A further innovation was the management of the challenge of data imbalance. The low prevalence of CVDs in individuals with CLDs was addressed by the Synthetic Minority Oversampling Technique (SMOTE), allowing for the improvement of model predictive capability for the underrepresented group of CVD cases.[30]

This study was conducted to design a promising CVD-specific risk prediction model to offer individualized decision support for CVD management in the context of CLDs. This model is expected to assist clinicians in the early recognition of high-risk populations and contribute to the establishment of targeted preventive strategies, thereby facilitating the long-term prognostic outcomes of CLD patients and efficient use of healthcare resources.

## 2. Methods

### 2.1. Study design and data source

The cross-sectional study design was adopted to process data from the first and third waves of the China Health and Retirement Longitudinal Study surveys. China Health and Retirement Longitudinal Study is a nationally representative project with extensive research data in China. The baseline survey (initiated in 2011) gathered responses from 17,708 individuals from 150 counties in 28 provinces of China.[31]

***2.1.1. Study participants.*** The study was initiated with a cohort of 2036 individuals with CLDs and complete blood test data. Participants were excluded if they had asthma, missing ADL or IADL data, missing CESD-10, or systolic/diastolic blood pressure data. Moreover, participants younger than 45 were also excluded. The final study population was reduced to 1357 individuals (Fig. 1).

***2.1.2. Variable definitions.*** Predictive variables incorporated in this study were demographic factors (age, gender, smoking status), clinical indicators (body mass index [BMI]), waist circumference, underlying diseases such as HTN, diabetes mellitus [DM], dyslipidemia [DLP], and chronic kidney disease, and laboratory markers (uric acid and C-reactive protein [CRP]), physical assessment metrics (peak expiratory flow), grip strength [GS], and QoL measures (ADL, IADL, and CESD-10 scores). The outcome variable was the presence or absence of comorbid heart diseases.

The definition of HTN included an average systolic blood pressure ≥ 140 mm Hg, an average diastolic blood pressure ≥ 90 mm Hg, or any patient-reported diagnosis of HTN confirmed by a physician. DLP was determined by total cholesterol ≥ 240 mg/dL, triglycerides ≥ 200 mg/dL, low-density lipoprotein cholesterol ≥ 160 mg/dL, or high-density lipoprotein

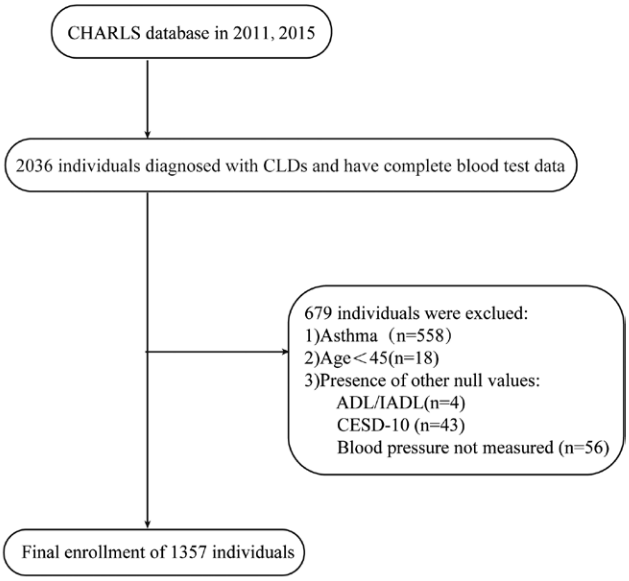
CHARLS database in 2011, 2015
2036 individuals diagnosed with CLDs and have complete blood test data
679 individuals were exclued:
1)Asthma (n=558)
2)Age<45(n=18)
3)Presence of other null values:
ADL/IADL(n=4)
CESD-10 (n=43)
Blood pressure not measured (n=56)
Final enrollment of 1357 individuals

**Figure 1.** Flowchart of participant inclusion.

cholesterol < 40 mg/dL, or a patient-reported diagnosis of DLP by a doctor. DM was defined by a fasting blood glucose ≥ 125 mg/dL (7.0 mmol/L), a random blood glucose ≥ 200 mg/dL (11.1 mmol/L), or a patient-reported diagnosis of DM. Chronic kidney disease was recorded through a question in the baseline survey and 3 subsequent follow-up waves, which inquired: "Have you been diagnosed with kidney diseases (excluding tumors or cancers) by a doctor?"

The CESD-10 is a 10-item scale to evaluate depressive symptoms, with total scores between 0 and 30. A score of 10 is regarded as the threshold for determining clinical depression. The ADL index was adopted to document functional limitations in basic activities, such as bathing, dressing, bed mobility, eating, toilet use, or bladder control, with scores (0–6) indicating ADL-related disabilities. Furthermore, the IADL index was employed to record functional impairments in more complex activities, such as housework, cooking, shopping, financial management, and medication adherence, with scores (0–6) suggesting IADL-related disabilities. The ADL–IADL index was obtained by combining these two measures (scores: 0–11) to reflect a more extensive overview of functional disabilities.

### 2.2. Statistical analysis

The data were processed using SPSS 26.0 software. Continuous variables were summarized using mean ± standard deviation or median (P25, P75), and categorical variables were summarized as n (%). Group comparisons between those with and without CVDs were made using *t*-test, Mann–Whitney *U* test, or chi-square test.

Missing data were processed *via* the Multiple Imputation by Chained Equations (MICE) approach. MICE was adopted by generating multiple sets of imputed values for missing data and producing multiple complete datasets for analysis. This tool enabled the identification of final imputed values by modeling the posterior distribution of the missing data through repeated estimations. The MICE method was implemented using the *miceforest* package in Python 3.11. The SMOTE was adopted to manage the data imbalance issue. The ratio of non-CVD to CVD cases was approximately 30:1. A more balanced dataset between both groups was generated by synthesizing 1932 additional data points using SMOTE and the *imbalanced-learn* package in Python 3.11.

For model construction and evaluation, the data was randomly allocated into the training set (80%) for model development and the validation set (20%) for performance assessment. Prior to model generation, the average contribution of each RF was calculated using the random forest algorithm, and the top 10 most influential RFs were highlighted. 4 ML models were then developed accordingly: LR, XGBoost, random forest, and SVM. During model training, 10-fold cross-validation was utilized to reduce overfitting risks. ML models were constructed using Python 3.11, with the *sklearn* and *xgboost* packages. Performance evaluation was conducted using a range of metrics, including area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, F1 score, and kappa value. Moreover, the importance analysis for RFs was carried out to examine the relative contribution of RFs to CVD risk estimation.

## 3. Results

### 3.1. Patient characteristics and CVD comorbidities

This study included 1357 individuals with CLDs, with 318 (23.4%) also presenting with CVDs. Baseline characteristics (Table 1) uncovered no marked differences in age between the CVD and non-CVD groups ($P = .052$). Nevertheless, pronounced differences were witnessed in BMI, gender, and waist circumference ($P < .001$), and smoking status ($P = .005$), with a higher proportion of nonsmokers in the CVD group (52.8% vs 45.3%).

Pulmonary function, measured by peak expiratory flow, was similar between the 2 groups. Nonetheless, the CVD-affected individuals exhibited lower GS ($P = .001$), implying an association between muscle strength and CVD risk. Furthermore, comorbid HTN, DLP, and chronic kidney disease were more frequent in CVD-affected individuals ($P < .001$), while no pronounced differences were witnessed in DM occurrence ($P = .127$). In addition, the CRP and uric acid levels also displayed no notable differences between the groups ($P > .05$), suggesting that conventional inflammatory and metabolic indicators may be insufficient for CVD risk prediction.

Furthermore, for QoL and psychological state, individuals with CVDs exhibited higher scores on ADL/IADL and CESD-10 than the non-CVD group ($P < .001$), indicating that patients with CVDs experience more restricted daily activities and more intense depressive symptoms.

### 3.2. Comparison of machine learning model performance

Among the 4 tested ML models, LR, XGBoost, random forest, and SVM, the XGBoost model outperformed the others in all evaluation criteria. It achieved an AUC of 0.788, an accuracy of 0.716, a sensitivity of 0.615, and a specificity above 0.803. The random forest model ranked closely behind, while the LR (AUC = 0.619) and SVM (AUC = 0.584) models displayed weaker performance. XGBoost also recorded a Kappa value of 0.427, indicating favorable predictive consistency (Table 2). Further ROC curve analysis reinforced the superiority of the XGBoost model. Its curve outperformed the other models at varying thresholds, offering a more favorable balance between sensitivity and specificity (Fig. 2).

### 3.3. Key RFs for CLD-specific CVDs identified by feature importance analysis

Feature importance analysis was carried out on the top-performing XGBoost and random forest models (Figs. 3 and 4) to pinpoint the key RFs for CLD-specific CVDs. The data implied that BMI and CRP were the most pronounced RFs in both models, highlighting the significance of obesity and systemic inflammation in the progression of CVDs. Age, a well-established RF for CVDs, similarly ranked as a major contributor in both models.

The significance of waist circumference and GS reflected a potential link between central obesity and muscle strength in relation to the risks for CVDs. Moreover, peak expiratory flow, IADL, and CESD-10 scores were recognized as key factors, suggesting the need to consider pulmonary functions, daily living capabilities, and mental health status for CVD risk assessment.

It should be noted that conventional CVD RFs, including HTN, DLP, and DM, did not rank among the top 10 RFs, suggesting that a broader range of indicators may be needed to predict CVD risk in CLD patients. The XGBoost model pinpointed BMI, CRP, age, uric acid, and GS as the predominant RFs. The importance rankings of RFs for the LR and SVM models are presented in Figure S1A–B, Supplemental Digital Content, http://links.lww.com/MD/O453.

## 4. Discussion

In the present study, ML-based models were constructed to predict CVD risks in patients with CLDs. The comparative analysis of 4 ML models uncovered that the XGBoost model provided optimal performance, as evidenced by high predictive accuracy and robust discriminative power. The findings offer a robust decision support tool for clinical settings and new insights into the complexity of CVD risk in individuals with CLDs.

### 4.1. Model performance and clinical significance

The XGBoost model, with an AUC of 0.788, displayed excellent performance in the present study and outperformed the other 3 ML algorithms. This result aligns with findings from recent studies in other fields. For instance, Abega et al have also pointed out the superior performance of the XGBoost model in CVD risks in patients with type 2 DM.[32] The XGBoost model may facilitate a comprehensive analysis of nonlinear relationships and high-dimensional feature interactions. Complex biological systems, including CVD risks in CLD patients, commonly involve nonlinear relationships among multiple variables. The XGBoost model potentiates prediction performance by generating decision trees and using boosting techniques, thereby effectively modeling these complex relationships.[27] Of note, our model displayed relatively high sensitivity (0.615) and specificity (0.803), indicating its capacity to identify high-risk patients and exclude low-risk patients. The equilibrium confers a crucial role in clinical decision-making by assisting physicians in identifying patients in need of active intervention,

**Table 1**

**Characteristics of study participants.**

| Variables | Total<br>(n = 1357) | Non-CVD group<br>(n = 1039) | CVD group<br>(n = 318) | *p*-value |
|---|---|---|---|---|
| Age (year) | 62.68 ± 9.40 | 62.40 ± 9.54 | 63.57 ± 8.85 | .052* |
| WC (cm) | 84 (76.5, 91.1) | 83.2 (76, 90.8) | 86.5 (78, 94.85) | <.001† |
| Gender | | | | |
| Male | 755 (55.6%) | 618 (59.5%) | 137 (43.1%) | |
| Female | 602 (44.4%) | 421 (40.5%) | 181 (56.9%) | <.001‡ |
| BMI | 23.13 ± 4.65 | 22.78 ± 4.19 | 24.26 ± 5.77 | <.001* |
| Smoking status | | | | |
| Nonsmoker | 639 (47.1%) | 471 (45.3%) | 168 (52.8%) | |
| Former smoker | 257 (18.9%) | 191 (18.4%) | 66 (20.8%) | |
| Current smoker | 461 (34%) | 377 (36.3%) | 84 (26.4%) | .005‡ |
| Hypertension | 605 (44.6%) | 422 (40.6%) | 183 (57.5%) | <.001‡ |
| Diabetes | 175 (12.9%) | 126 (12.1%) | 49 (15.4%) | .127‡ |
| Dyslipidemia | 565 (41.6%) | 401 (38.6%) | 164 (51.6%) | <.001‡ |
| CKD | 182 (13.4%) | 116 (11.2%) | 66 (20.8%) | <.001‡ |
| CRP (mg/L) | 1.4 (0.70, 2.80) | 1.4 (0.70, 2.90) | 1.5 (0.75, 2.70) | .670† |
| UA (mg/dL) | 4.70 (3.82, 5.61) | 4.72 (3.82, 5.68) | 4.60 (3.81, 5.50) | .502† |
| PEF (L/min) | 260 (170, 360) | 260 (170, 360) | 265 (170, 360) | .772† |
| Grip strength (kg) | 31 (24, 39) | 31 (24, 39) | 29 (22, 36.53) | .001† |
| ADL score | 0 (0, 1) | 0 (0, 0) | 0 (0, 1) | .001† |
| IADL score | 0 (0, 1) | 0 (0, 1) | 0 (0,1) | <.001† |
| CESD-10 score | 9 (5, 14) | 8 (5, 14) | 11 (6, 17) | <.001† |

Data are presented as mean ± standard deviation, median (interquartile range) or n (%).

ADL = activities of daily living, BMI = body mass index, CESD-10 = 10-item Center for Epidemiologic Studies Depression Scale, CKD = chronic kidney disease, CRP = C-reactive protein, CVD = cardiovascular disease, IADL = instrumental activities of daily living, PEF = peak expiratory flow, UA = uric acid, WC = waist circumference.

*Statistical tests include independent-sample *t*-test.

†Mann–Whitney *U* test.

‡Chi-squared test.

**Table 2**

**Comparison of the predictive performance of ML models.**

| Model | F1 Score | AUC | Kappa | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| LR | 0.531 | 0.619 | 0.159 | 0.584 | 0.508 | 0.649 |
| XGBoost | 0.667 | 0.788 | 0.427 | 0.716 | 0.615 | 0.803 |
| random forest | 0.559 | 0.709 | 0.256 | 0.633 | 0.503 | 0.745 |
| SVM | 0.1380 | 0.584 | 0.062 | 0.548 | 0.078 | 0.952 |

AUC = area under the receiver operating characteristic curve, LR = logistic regression, ML = machine learning, SVM = support vector machine, XGBoost = extreme gradient boosting.

while preventing unnecessary overtreatment for individuals with a lower risk profile.

### 4.2. Implications of RF importance

Our study found that conventional cardiovascular RFs (such as age, BMI), inflammatory markers (CRP), physical characteristics (waist circumference, GS), and indicators reflecting QoL and mental health (IADL, CESD-10) all imparted critical effects in predicting CVD risk in patients with CLDs. This finding highlights the need for a comprehensive approach to CVD risk assessment in these patients.

In the current study, BMI and CRP levels ranked in the top two in XGBoost and random forest models, which is consistent with prior studies.[33,34] Obesity is an independent RF for CVDs, and it might increase CVD risks by augmenting chronic inflammatory reactions.[33] The identification of CRP levels as an indicator of inflammation pointed to chronic inflammation as a potential driving mechanism in the development of CVDs in the context of CLDs. This result reinforced the hypothesis that early intervention with anti-inflammatory therapies might aid in diminishing CVD risks in this population. Notably, the significance of GS as a key RF highlighted the potential role of muscle strength in CVD risk assessment. This observation aligns with recent research on the association between sarcopenia and cardiovascular health.[35] The research of Yang et al has clarified that individuals with stable normal muscle strength displayed the lowest CVD risks, and the CVD risks progressively escalated in the Low-Normal (HR: 1.20), Normal-Low (HR: 1.35), and Low-Low (HR: 1.76) groups.[36] Our study further substantiated the importance of muscle strength evaluation in patients with CLDs, which might offer insights for the development of targeted exercise intervention strategies. The data of IADL and CESD-10 scores as important RFs highlighted the role of QoL and mental health in CVD risk assessment. This result is consistent with previous findings on the relationships between depressive symptoms and increased CVD risks. A prior pooled analysis of individual-participant data from 22 prospective cohorts uncovered a pronounced association between depressive symptoms and the risk of CVDs, even at CESD scores below the usual threshold for depressive disorders (CESD ≥ 16), with each 1-SD increase in the log-transformed CES-D score corresponding to an HR of 1.06.[37] Our study further validates the importance of mental health in cardiovascular health management for patients with CLDs.

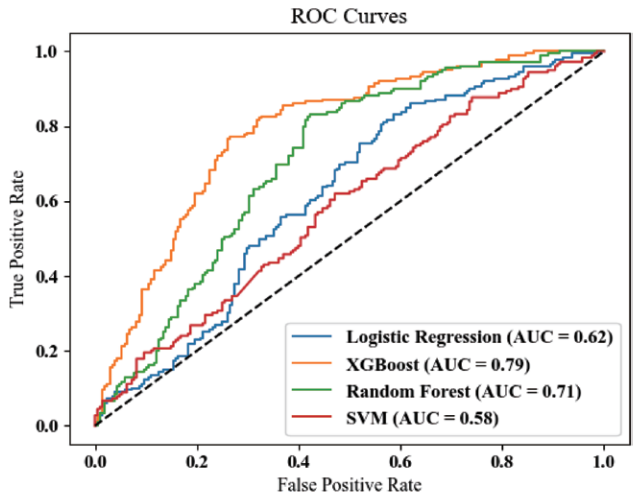

**Figure 2.** Comparison of ROC curves for 4 ML models on the testing dataset. ML = machine learning.

However, conventional CVD RFs such as HTN, DLP, and DM were relatively less important in the XGBoost model. This finding may reflect the uniqueness of CLD-specific CVD risk, suggesting that conventional risk assessment tools may be insufficient to accurately predict CVD risks in patients with CLDs. The findings highlight the necessity of specific CVD risk assessment tools for patients with CLDs.

### 4.3. Strengths and limitations of the study

The main strength of this study is the inclusion of a wide range of RFs. This multi-dimensional data collection facilitates

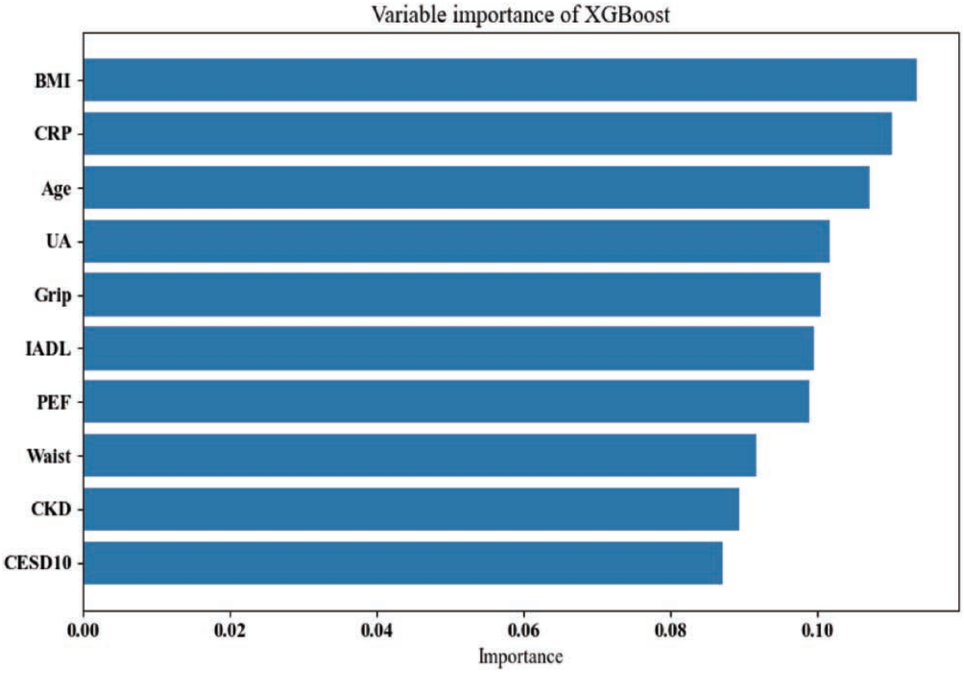

**Figure 3.** Top 10 key risk factors for CVDs in CLD patients in the XGBoost model. CVDs = cardiovascular disease, CLD = chronic lung diseases.

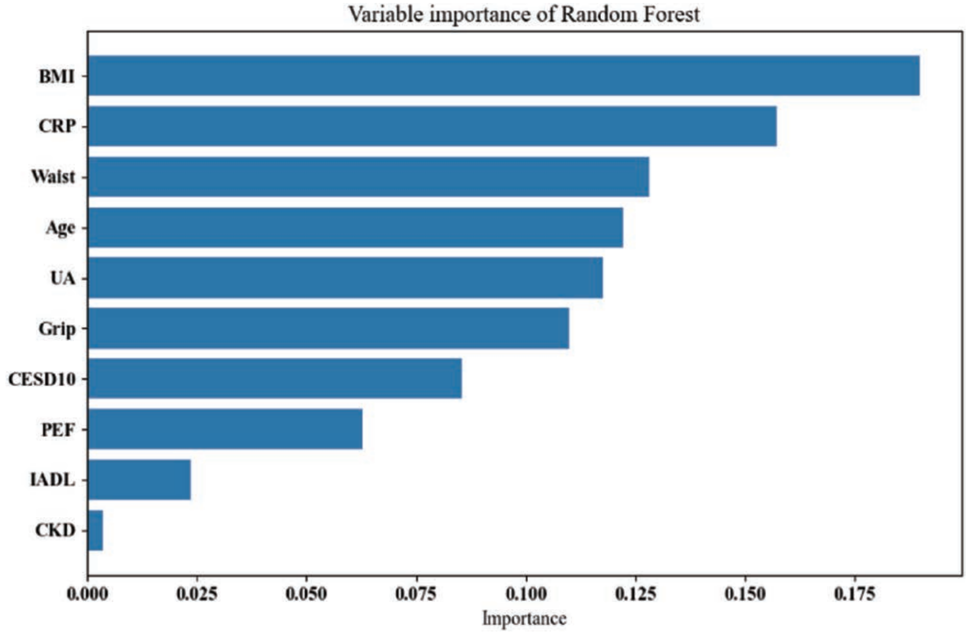

**Figure 4.** Top 10 key risk factors for CVDs in CLD patients in the random forest model. CVDs = cardiovascular disease, CLD = chronic lung diseases.

comprehensive analysis of the health status of patients with CLDs. The XGBoost model with the best performance was identified among the 4 ML algorithms tested. The advantages of ML models enable the XGBoost model to process complex nonlinear relationships. The importance analysis of RFs facilitated the interpretability of the model and offered promising insights for clinical practice.

However, this study also has some limitations. The study design makes it impossible to determine the causal relationships between RFs and CVDs. Furthermore, the data were retrieved from China Health and Retirement Longitudinal Study in China, which may limit the generalizability of the model. Although multiple RFs were included in the study, potential variables, such as genetic factors and environmental stimuli, may be omitted.

### *4.4. Future prospects*

Based on our findings, future prospective cohort studies should be conducted to substantiate the predictive performance of the XGBoost model and further examine the causal relationships between RFs and CVD development. Moreover, external validation should be conducted in different populations to assess the generalizability of the model. Additionally, more potential RFs, such as genetic factors and environmental exposure, should be considered to further improve predictive accuracy. Meanwhile, the combined effects of this model with clinical data support systems should be explored to evaluate its application in clinical settings and its impacts on patient prognosis. Finally, based on the key RFs identified in this study, targeted intervention strategies, such as exercise interventions and mental health management, should be developed and evaluated to reduce CVD risk in patients with CLDs.

## 5. Conclusion

This study successfully developed an ML-based model for predicting CVD risk in patients with CLDs. The XGBoost algorithm showed superior predictive performance, providing a potentially powerful tool for clinical practice. Our findings emphasize the need to consider the combined effects of traditional cardiovascular RFs, inflammatory markers, physical characteristics, and QoL and mental health indicators for CVD risk assessment in patients with CLDs. This multi-dimensional risk assessment approach provides new insights for the development of individualized CVD prevention strategies.

## Acknowledgments

We would like to extend our sincere thanks to the Institute of Social Science Survey at Peking University for orchestrating the China Health and Retirement Longitudinal Study (CHARLS). Our appreciation also goes to all the participants, researchers, and staff who contributed to the success of CHARLS.

## Author contributions

**Conceptualization:** Huiming Xi, Qingxin Kang.
**Formal analysis:** Qingxin Kang.
**Investigation:** Xunsheng Jiang.
**Methodology:** Huiming Xi, Qingxin Kang, Xunsheng Jiang.
**Project administration:** Xunsheng Jiang.
**Software:** Huiming Xi.
**Visualization:** Huiming Xi, Qingxin Kang, Xunsheng Jiang.
**Writing – original draft:** Huiming Xi, Qingxin Kang.
**Writing – review & editing:** Xunsheng Jiang.

## References

[1] GBD 2021 Diseases and Injuries Collaborators. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 2024;403:2133–61.

[2] GBD 2019 Chronic Respiratory Diseases Collaborators. Global burden of chronic respiratory diseases and risk factors, 1990-2019: an update from the Global Burden of Disease Study 2019. EClinicalMedicine. 2023;59:101936.

[3] GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Respir Med. 2020;8:585–96.

[4] Chen S, Kuhn M, Prettner K, et al. The global economic burden of chronic obstructive pulmonary disease for 204 countries and territories

in 2020-50: a health-augmented macroeconomic modelling study. Lancet Glob Health. 2023;11:e1183–93.

[5] Wang C, Xu J, Yang L, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study. Lancet. 2018;391:1706–17.

[6] Fabbri LM, Celli BR, Agustí A, et al. COPD and multimorbidity: recognising and addressing a syndemic occurrence. Lancet Respir Med. 2023;11:1020–34.

[7] Rabe KF, Hurst JR, Suissa S. Cardiovascular disease and COPD: dangerous liaisons? Eur Respir Rev. 2018;27:180057.

[8] Kunisaki KM, Dransfield MT, Anderson JA, et al. Exacerbations of chronic obstructive pulmonary disease and cardiac events. a post hoc cohort analysis from the SUMMIT randomized clinical trial. Am J Respir Crit Care Med. 2018;198:51–7.

[9] Graul EL, Nordon C, Rhodes K, et al. Temporal risk of nonfatal cardiovascular events after chronic obstructive pulmonary disease exacerbation: a population-based study. Am J Respir Crit Care Med. 2024;209:960–72.

[10] Yang HM, Ryu MH, Carey VJ, et al. Chronic obstructive pulmonary disease exacerbations increase the risk of subsequent cardiovascular events: a longitudinal analysis of the COPDGene study. J Am Heart Assoc. 2024;13:e033882.

[11] Morgan AD, Zakeri R, Quint JK. Defining the relationship between COPD and CVD: what are the implications for clinical practice? Ther Adv Respir Dis. 2018;12:1753465817750524.

[12] Dalal AA, Shah M, Lunacsek O, Hanania NA. Clinical and economic burden of patients diagnosed with COPD with comorbid cardiovascular disease. Respir Med. 2011;105:1516–22.

[13] Barnes PJ. Inflammatory mechanisms in patients with chronic obstructive pulmonary disease. J Allergy Clin Immunol. 2016;138:16–27.

[14] Khedoe PP, Rensen PC, Berbée JF, Hiemstra PS. Murine models of cardiovascular comorbidity in chronic obstructive pulmonary disease. Am J Physiol Lung Cell Mol Physiol. 2016;310:L1011–27.

[15] Barnes PJ. Oxidative stress-based therapeutics in COPD. Redox Biol. 2020;33:101544.

[16] Spiesshoefer J, Regmi B, Ottaviani MM, et al. Sympathetic and vagal nerve activity in COPD: pathophysiology, presumed determinants and underappreciated therapeutic potential. Front Physiol. 2022;13:919422.

[17] van Gestel AJ, Kohler M, Clarenbach CF. Sympathetic overactivity and cardiovascular disease in patients with chronic obstructive pulmonary disease (COPD). Discov Med. 2012;14:359–68.

[18] Olsson KM, Corte TJ, Kamp JC, et al. Pulmonary hypertension associated with lung disease: new insights into pathomechanisms, diagnosis, and management. Lancet Respir Med. 2023;11:820–35.

[19] Nathan SD, Barbera JA, Gaine SP, et al. Pulmonary hypertension in chronic lung disease and hypoxia. Eur Respir J. 2019;53:1801914.

[20] Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. Lancet. 2014;383:999–1008.

[21] Cimmino G, Natale F, Alfieri R, et al. Non-conventional risk factors: “fact” or “fake” in cardiovascular disease prevention? Biomedicines. 2023;11:2353.

[22] Kim HY, Lampertico P, Nam JY, et al. An artificial intelligence model to predict hepatocellular carcinoma risk in Korean and Caucasian patients with chronic hepatitis B. J Hepatol. 2022;76:311–8.

[23] Singal AG, Mukherjee A, Elmunzer BJ, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol. 2013;108:1723–30.

[24] Wei C, Wang J, Yu P, et al. Comparison of different machine learning classification models for predicting deep vein thrombosis in lower extremity fractures. Sci Rep. 2024;14:6901.

[25] Devani RN, Kirubakaran A, Molokhia M. Digital health RCT interventions for cardiovascular disease risk reduction: a systematic review and meta-analysis. Health Technol (Berl). 2022;12:687–700.

[26] Liu Y, Hannig J. Generalized fiducial inference for logistic graded response models. Psychometrika. 2017;82:1097–125.

[27] Mitchell R, Frank E. Accelerating the XGBoost algorithm using GPU computing. PeerJ Comput Sci. 2017;3:e127.

[28] Amaratunga D, Cabrera J, Lee YS. Enriched random forests. Bioinformatics. 2008;24:2010–4.

[29] Wang H, Shao Y, Zhou S, Zhang C, Xiu N. Support vector machine classifier via l(0/1) soft-margin loss. IEEE Trans Pattern Anal Mach Intell. 2022;44:7253–65.

[30] Nguyen T, Mengersen K, Sous D, Liquet B. SMOTE-CD: SMOTE for compositional data. PLoS One. 2023;18:e0287705.

[31] 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.

[32] Abegaz TM, Baljoon A, Kilanko O, Sherbeny F, Ali AA. Machine learning algorithms to predict major adverse cardiovascular events in patients with diabetes. Comput Biol Med. 2023;164:107289.

[33] Koliaki C, Liatis S, Kokkinos A. Obesity and cardiovascular disease: revisiting an old relationship. Metabolism. 2019;92:98–107.

[34] Leuzzi G, Galeone C, Taverna F, Suatoni P, Morelli D, Pastorino U. C-reactive protein level predicts mortality in COPD: a systematic review and meta-analysis. Eur Respir Rev. 2017;26:160070.

[35] Gao K, Cao LF, Ma WZ, et al. Association between sarcopenia and cardiovascular disease among middle-aged and older adults: findings from the China health and retirement longitudinal study. EClinicalMedicine. 2022;44:101264.

[36] Yang Z, Wei J, Liu H, et al. Changes in muscle strength and risk of cardiovascular disease among middle-aged and older adults in China: Evidence from a prospective cohort study. Chin Med J (Engl). 2024;137:1343–50.

[37] Harshfield EL, Pennells L, Schwartz JE, et al. Association between depressive symptoms and incident cardiovascular diseases. JAMA. 2020;324:2396–405.