BMC Cardiovascular Disorders

RESEARCH

Open Access

Construction and validation of a predictive model for intracardiac thrombus risk in patients with dilated cardiomyopathy: a retrospective study

Xuetao Zhu[1], Jun Li[1], Yi Jiang[1], Tianqi Wang[1] and Zeping Hu[1*]

Abstract

Background Systemic embolic events due to exfoliation of intracardiac thrombus (ICT) are one of the catastrophic complications of dilated cardiomyopathy (DCM). This study intended to develop a prediction model to predict the risk of ICT in patients with DCM.

Methods Data from 632 patients with DCM from a hospital was collected. ICT was identified based on the results of transthoracic echocardiography. Basic information, vital signs, comorbidities, and biochemical data were measured and collected from each patient. The least absolute shrinkage and selection operator (LASSO) regression was used for the final model variable screening. Four classifiers including Logistic Regression, support vector machine (SVM), Random Forest, and eXtreme Gradient Boosting (XGBoost) were used for model construction respectively. The area under of the curve (AUC) with 95% confidence interval (CI), sensitivity, specificity, and accuracy of the models were calculated to assess the predictive ability of the models.

Results Of these 632 DCM patients, 88 (13.92%) had ICT and 544 (86.08%) did not. Eleven clinical variables were selected for the construction of predictive models. The AUC of the Logistic Regression model to predict ICT probability was 0.854 (95%CI: 0.811–0.896), the SVM model was 0.769 (95%CI: 0.715–0.824), the Random Forest model was 0.917 (95%CI: 0.887–0.947), and the XGBoost model was 0.947 (95%CI: 0.924–0.969). The Delong test demonstrated that the XGBoost model had the highest AUC for predicting the ICT probability compared to other models ($P < 0.05$). Moreover, D-dimer, age, and atrial fibrillation contributed the most to the XGBoost model among these 11 variables.

Conclusion The XGBoost model has a good predictive ability in predicting ICT risk in patients with DCM and may assist clinicians in identifying ICT risk.

Keywords Dilated cardiomyopathy, Intracardiac thrombus, Risk, Prediction model, XGBoost

*Correspondence:
Zeping Hu
huzepingdc@163.com
[1]Department of Cardiology, The First Affiliated Hospital of Anhui Medical University, No. 218 Jixi Road, Hefei, Anhui 230022, P.R. China

Background

Dilated cardiomyopathy (DCM) is a myocardial disease caused by left ventricular (LV) or biventricular dilatation and systolic dysfunction [1]. It has been reported that DCM occurred in 5–7 cases per 100,000 person-years [1, 2]. DCM is the most common indication for heart transplantation as it ultimately leads to heart failure (HF) and life-threatening arrhythmias [3, 4]. The high fatality rate and poor prognosis of DCM bring a great burden to the social economy and medical resources. Many factors affect the prognosis of DCM, including etiology, gender, age, and intracardiac thrombus (ICT), etc [1]. Systemic embolic events due to exfoliation of ICT are one of the common complications of DCM, which increase the mortality of DCM patients. Therefore, early identification of ICT in patients with DCM is important because it can improve clinical outcomes by timely and effective anticoagulant treatment.

When attempting to explore the risk factors of ICT in patients with DCM, both arterial and venous thrombus should be considered. Arterial thrombus is mostly formed at the location of endothelium damage, which triggers platelet aggregation through fibrin structure [5]. Venous thrombus commonly occurs at a low flow rate [5]. Decreased ventricular contraction and reduced myocardial motion in patients with DCM potentiates blood stasis [6]. Meanwhile, endothelium damage is favored by increased shear stress [6]. Previous studies have suggested that elevated D-dimer and brain natriuretic peptide (BNP) levels and reduced left ventricular ejection fraction (LVEF) are independently associated with left ventricular thrombus (LVT) [7]. However, previous studies have been limited to the LVT. To reflect the true clinical disease status, more clinical features such as atrial thrombus and right ventricular thrombus were also considered in this study. Furthermore, machine learning methods are widely used in the prediction of cardiovascular diseases and thrombosis [8, 9]. Therefore, this study intended to construct a prediction model for predicting the risk of ICT in patients with DCM based on their clinical features to assist clinicians in risk assessment. Moreover, there are differences in the predictive ability of prediction models constructed with different classifiers, and prediction models based on different classifiers were also constructed and compared in this study.

Methods

Study patients

The cohort of this study was collected from patients with DCM who were admitted to the First Affiliated Hospital of Anhui Medical University from April 2017 to December 2022. The diagnosis of DCM was in accordance with the Guidelines for the Diagnosis and Treatment of Dilated Cardiomyopathy in China: (1) left ventricular end-diastolic diameter (LVDd) > 5.0 cm (female) and LVDd > 5.5 cm (male); (2) LVEF < 45% (Simpsons Method), left ventricular fractional shortening (LVFS) < 25%. Patients diagnosed with DCM were included in this study. The exclusion criteria were: (1) patients with cardiac implantable electronic devices (CIEDs) or left ventricular assist devices; (2) patients with hypertension, valvular heart disease, congenital heart disease, ischemic heart disease, or malignancy; (3) patients with previous valve surgery or previous heart transplantation; (4) patients with a history of ICH; (5) patients with missing data on key laboratory tests (e.g., D-dimer). The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the First Affiliated Hospital of Anhui Medical University. Informed consent was obtained from all subjects involved in the study. Clinical trial number: not applicable.

The present study estimated the lower sample size of the development dataset based on the 10 events per variable (10EPV) rule of thumb, that is, when developing a predictive model for binary outcomes, an established rule of thumb for the required sample size is to ensure at least 10 events for each predictor parameter [10]. Finally, a total of 632 patients with DCM were enrolled in this study, including 544 patients without ICT and 88 patients with ICT. Therefore, the sample size was sufficient to build an effective predictive model.

Data collection

Patient information was collected from the electronic medical record system, including gender (female, male), body mass index (BMI), age, history of smoking (no, yes), history of drinking (no, yes), systolic blood pressure (SBP), diastolic blood pressure (DBP), heart rate, hypertension (no, yes), diabetes (no, yes), atrial fibrillation (AF) (no, yes), LVEF, left atrial diameter (LAD), LVDd, white blood cell, neutrophil, red blood cell (RBC), hemoglobin, hematocrit, RBC distribution width-variable coefficient (CV), RBC distribution width-standard deviation (SD), platelets, total protein, albumin, total bilirubin, direct bilirubin, indirect bilirubin, alanine aminotransferase (ALT), aspartate aminotransferase (AST), urea nitrogen, creatinine, uric acid, glucose, BNP, D-dimer, total cholesterol, triglyceride, high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), apolipoprotein A1, apolipoprotein B, spironolactone (no, yes), angiotensin receptor/neprilysin inhibitor (ARNI)/angiotensin-converting enzyme inhibitors (ACEIs)/angiotensin receptor blockers (ARBs) (no, yes), beta blockers (no, yes), and sodium-glucose cotransporter protein-2 (SGLT-2) inhibitors (no, yes). General information was collected within 8 h of hospitalization, including demographic information, the history of smoking and

drinking, the presence of comorbidities, etc. The blood test was completed within 24 h of hospitalization, and blood samples were sent to the laboratory for biochemical assays. Echocardiogram was completed within 48 h of hospitalization, and reports were reviewed by two senior physicians. AF was diagnosed by routine electrocardiogram or Holter electrocardiogram. These variables included the patient's routine clinical testing variables as well as those reported in previous studies [11]. Moreover, the CHA2DS2-VASc score was calculated [12] due to the predictive value of the CHA2DS2-VASc score for thrombosis [13].

Model construction and evaluation

A total of 632 DCM patients were included in the training set. A Bootstrap approach of put-back sampling was utilized to select 200 patients from these 632 patients as the testing set. The Bootstrapping method is recommended as the preferred method for internal validation of predictive models [14]. A series of variable screening methods were conducted to screen the variables that were ultimately included in the model. First, 45 variables of the initial 45 clinical variables remained after one-hot-code preprocessing. Then, highly correlated variables were excluded using the Pearson correlation coefficient, and variables with a Pearson correlation coefficient below 0.80 were retained (9 variables were excluded and 36 remained). To screen out variables that were closely related to the outcome, the Mann-Whitney U test was performed and variables with $P<0.05$ in the test were retained (16 variables were excluded and 20 remained). Finally, the remaining variables were screened using the Least absolute shrinkage and selection operator (LASSO) regression. LASSO regression compresses the coefficient of variables by constructing a penalty function and making some regression coefficients zero, and then selecting variables whose regression coefficients are not zero. Ten-fold cross-validation is utilized to ensure the stability and efficacy of the variables and to avoid overfitting. The number of variables entering the LASSO model decreases with the increase of λ, and the optimal regression coefficient can be obtained when the mean square error (MSE) is minimum. After screening by LASSO regression, 11 variables were used to construct the prediction model: age, beta blockers (yes), ACEIs/ARBs/ARNI (yes), RBC, DBP, RBC distribution width-CV, history of drinking, BNP, LVDd, D-dimer, and AF.

Four classifiers including Logistic Regression, support vector machine (SVM), Random Forest, and eXtreme Gradient Boosting (XGBoost) were used for prediction model construction. SVM is a classification algorithm that, according to the different input data can do different models, it can solve the nonlinear classification problem through the kernel function and the classification idea is simple. Random Forest algorithm is an extended variant of Bagging integration learning, with a decision tree as the base learner to build Bagging integration, and the model has strong generalization ability. XGBoost is an efficient gradient-boosting decision tree algorithm, which is improved on the basis of the original GBDT, so that the model effect is greatly improved, and the model is more accurate and flexible.

The predictive ability of these models was evaluated by the area under the receiver operating characteristic curve (AUC) with 95% confidence interval (CI), sensitivity, specificity, and accuracy. The AUC value is greater than 0.75, indicating that the model has good predictive ability. The Delong test was used to evaluate differences in AUC between models. The flowchart for the construction of the model was presented in Fig. 1. The parameters of these models were shown in Supplementary Table 1.

Statistical analysis

Continuous variables were reported as mean ± standard deviation (SD) or median and quartiles [M (Q1, Q3)], and comparisons between groups were performed using the t-test or Wilcoxon rank-sum test. Categorical variables were reported as numbers and percentages [n (%)], and comparisons between groups were conducted using the Fisher's exact test or Chi-square test. The outcome variable (ICT) was not missing, and other variables with a percentage of missing values greater than 20% were removed (Supplement Table 2), and less than 20%, the missing values were interpolated using random forest interpolation. Difference analysis was performed before and after missing data processing (Supplement Table 3). The prediction effect of the prediction model and CHA2DS2-VASc score on thrombosis was also compared. Moreover, the statistical power of the XGBoost calculated by the AUC was 0.889. Statistical analyses were completed using R software (Version 4.2.2) and Python software (version 3.9.12). A P-value < 0.05 was considered statistically significant.

Results

Patient characteristics

A total of 632 patients were included in this study (Supplement Fig. 1), of whom 441 were male and 191 were female. The mean age was 60.59 (± 14.07) years. Based on the results of echocardiogram, 88 (13.92%) patients had ICT and 544 (86.08%) did not. There were significant differences between the ICT and non-ICT groups in age, history of drinking, DBP, AF, LVEF, RBC, hemoglobin, hematocrit, RBC distribution width-CV, RBC distribution width-SD, total protein, albumin, total bilirubin, direct bilirubin, indirect bilirubin, ALT, AST, creatinine, uric acid, BNP, D-dimer, HDL-C, apolipoprotein A1, ACEIs/ARBs/ARNI, and beta blockers ($P<0.05$), but

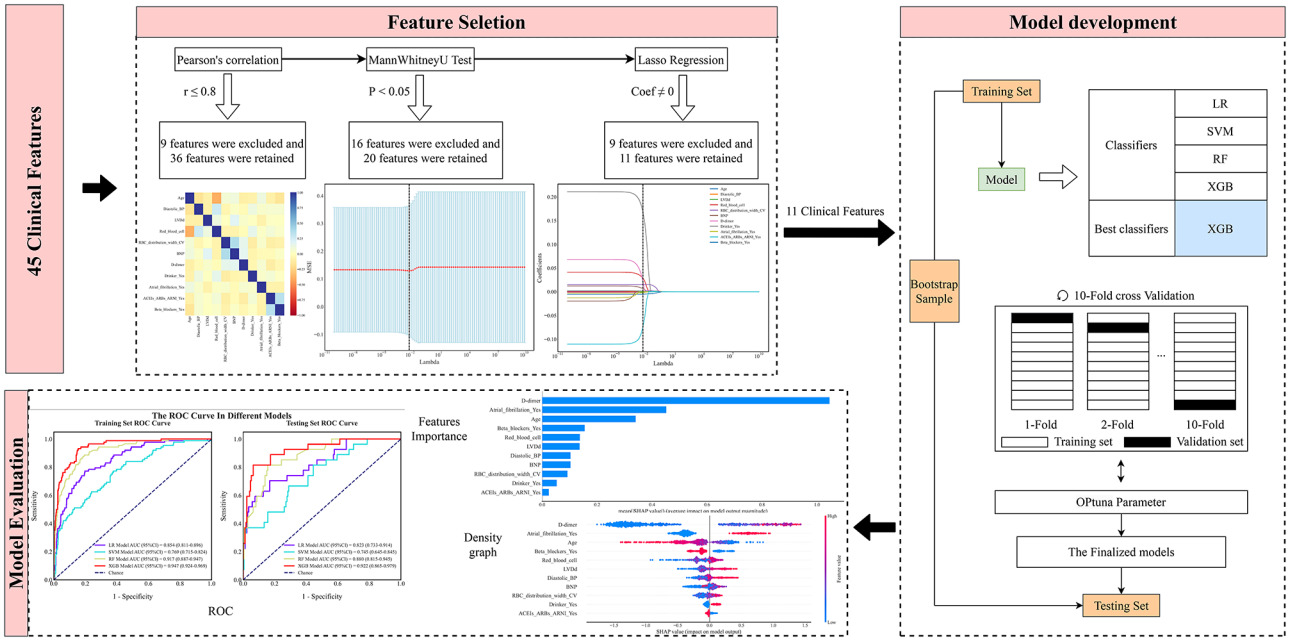

Fig. 1 The flowchart for the construction of the predictive model. The construction process of the prediction model mainly includes three parts: selection of model features, model development, and model evaluation

not in gender, BMI, history of smoking, SBP, heart rate, hypertension, diabetes, LAD, white blood cell, neutrophil, platelets, urea nitrogen, glucose, total cholesterol, triglyceride, LDL-C, apolipoprotein B, spironolactone, and SGLT-2 inhibitors ($P > 0.05$) (Table 1). The characteristics of patients in the training set and the testing set were shown in Supplement Table 4.

Model performance for predicting ICT probability in patients with DCM

The screening results of variables by LASSO regression were presented in Fig. 2. After LASSO regression screening, the λ-value (λ = 0.007565) with the smallest MSE value was determined (Fig. 2A), and 11 variables were selected to construct the model based on the λ value (Fig. 2B). These 11 variables were age, beta blockers-Yes, ACEIs/ARBs/ARNI-Yes, RBC, DBP, RBC distribution width-CV, history of drinking-Yes, BNP, LVDd, D-dimer, and AF-Yes. The LASSO regression coefficients for these 11 variables were listed in Supplement Table 5. The multicollinearity test indicated that there was no multicollinearity among these 11 variables (Supplement Table 6). In addition, the correlation heat map for these 11 variables were shown in Supplement Fig. 2.

The performance of Logistic Regression, SVM, Random Forest, and XGBoost models in predicting the ICT probability in patients with DCM was presented in Table 2. In the training set, the AUC of the Logistic Regression model to predict ICT probability was 0.854 (95%CI: 0.811–0.896), the SVM model was 0.769 (95%CI: 0.715–0.824), the Random Forest model was 0.917 (95%CI: 0.887–0.947), and the XGBoost model was 0.947 (95%CI: 0.924–0.969). In the testing set, the AUC was 0.823 (95%CI: 0.733–0.914) for the Logistic Regression model, 0.745 (95%CI: 0.645–0.845) for the SVM model, 0.880 (95%CI: 0.815–0.945) for the Random Forest model, and 0.922 (95%CI: 0.865–0.979) for the XGBoost model. The Delong test demonstrated that the XGBoost model had the highest AUC for predicting the ICT probability in patients with DCM compared to other models ($P < 0.05$) (Table 3). In addition, the XGBoost model also showed good sensitivity (0.932, 95%CI: 0.879–0.984), specificity (0.846, 95%CI: 0.815–0.876), and accuracy (0.858, 95%CI: 0.830–0.885) in predicting the probability of ICT (training set). The receiver operating characteristic (ROC) curves of these models in the training set and testing set were presented in Fig. 3. The SHAP summary plot of the variable importance analysis for the XGBoost model was shown in Fig. 4. The results showed that D-dimer, age, and AF contributed the most to the XGBoost prediction model. Furthermore, we compared the performance of the CHA2DS2-VASc score (a known thrombus prediction score) and our XGBoost model in predicting ICH risk in patients with DCM. The AUC of the CHA2DS2-VASc score to predict ICT risk in patients with DCM was 0.572 (95%CI: 0.511–0.633) in the training set and 0.591 (95%CI: 0.482-0.700) in the testing set, respectively (Supplement Table 7, Supplement Fig. 3). The Delong test indicated that the AUC was higher in the XGBoost model compared to the CHA2DS2-VASc score ($P < 0.001$).

Table 1 Characteristics of patients with dilated cardiomyopathy (DCM)

Variables	Total (*N* = 632)	Non-ICT (*N* = 544)	ICT (*N* = 88)	*P*
Gender, n (%)				0.600
Female	191 (30.22)	167 (30.70)	24 (27.27)	
Male	441 (69.78)	377 (69.30)	64 (72.73)	
BMI, Mean (± SD)	23.95 (± 4.19)	23.96 (± 4.25)	23.87 (± 3.77)	0.844
Age, Mean (± SD)	60.59 (± 14.07)	61.67 (± 13.38)	53.92 (± 16.30)	< 0.001
History of smoking, n (%)				0.171
No	485 (76.74)	423 (77.76)	62 (70.45)	
Yes	147 (23.26)	121 (22.24)	26 (29.55)	
History of drinking, n (%)				0.034
No	496 (78.48)	435 (79.96)	61 (69.32)	
Yes	136 (21.52)	109 (20.04)	27 (30.68)	
SBP, Mean (± SD)	120.38 (± 19.53)	120.49 (± 19.48)	119.72 (± 19.89)	0.731
DBP, Mean (± SD)	76.51 (± 14.44)	75.95 (± 14.09)	79.98 (± 16.05)	0.015
Heart rate, Mean (± SD)	86.12 (± 19.36)	85.81 (± 18.80)	88.07 (± 22.52)	0.374
Hypertension, n (%)				0.772
No	426 (67.41)	365 (67.10)	61 (69.32)	
Yes	206 (32.59)	179 (32.90)	27 (30.68)	
Diabetes, n (%)				1.000
No	543 (85.92)	467 (85.85)	76 (86.36)	
Yes	89 (14.08)	77 (14.15)	12 (13.64)	
Atrial fibrillation, n (%)				< 0.001
No	475 (75.16)	433 (79.60)	42 (47.73)	
Yes	157 (24.84)	111 (20.40)	46 (52.27)	
LVEF, Mean (± SD)	34.27 (± 5.90)	34.51 (± 5.86)	32.81 (± 6.00)	0.012
LAD, Mean (± SD)	5.06 (± 0.74)	5.04 (± 0.75)	5.15 (± 0.68)	0.224
LVDd, Mean (± SD)	6.87 (± 0.77)	6.84 (± 0.75)	7.11 (± 0.88)	0.008
White blood cell, M (Q_1, Q_3)	6.37 (5.24, 7.85)	6.34 (5.21, 7.77)	6.88 (5.50, 8.38)	0.057
Neutrophil, M (Q_1, Q_3)	4.15 (3.19, 5.38)	4.10 (3.18, 5.31)	4.42 (3.36, 5.85)	0.108
RBC, Mean (± SD)	4.53 (± 0.64)	4.50 (± 0.63)	4.72 (± 0.68)	0.003
Hemoglobin, Mean (± SD)	136.73 (± 18.97)	135.94 (± 18.95)	141.60 (± 18.48)	0.009
Hematocrit, Mean (± SD)	41.23 (± 5.39)	40.97 (± 5.38)	42.81 (± 5.21)	0.003
RBC distribution width-CV, Mean (± SD)	14.04 (± 1.80)	13.95 (± 1.74)	14.60 (± 2.02)	0.005
RBC distribution width-SD, Mean (± SD)	46.48 (± 5.67)	46.25 (± 5.50)	47.91 (± 6.46)	0.024
Platelets, Mean (± SD)	187.57 (± 69.49)	188.67 (± 68.90)	180.78 (± 73.09)	0.324
Total protein, Mean (± SD)	64.04 (± 6.89)	64.27 (± 6.80)	62.64 (± 7.28)	0.039
Albumin, Mean (± SD)	38.34 (± 4.60)	38.55 (± 4.51)	37.03 (± 4.99)	0.008
Total bilirubin, Mean (± SD)	23.82 (± 15.28)	22.79 (± 14.20)	30.20 (± 19.69)	0.001
Direct bilirubin, M (Q_1, Q_3)	4.50 (2.79, 7.06)	4.13 (2.65, 6.41)	7.42 (4.57, 11.90)	< 0.001
Indirect bilirubin, Mean (± SD)	17.69 (± 10.36)	17.23 (± 9.86)	20.53 (± 12.72)	0.022
ALT, M (Q_1, Q_3)	26.00 (17.00, 43.00)	26.00 (16.00, 41.00)	31.00 (20.50, 62.00)	0.009
AST, M (Q_1, Q_3)	28.00 (21.00, 38.25)	28.00 (21.00, 38.00)	34.00 (25.00, 55.50)	0.001
Urea nitrogen, M (Q_1, Q_3)	7.13 (5.78, 9.04)	7.12 (5.77, 8.90)	7.19 (6.14, 9.54)	0.207
Creatinine, M (Q_1, Q_3)	78.70 (64.70, 99.50)	77.85 (64.27, 98.90)	83.60 (73.88, 103.45)	0.015
Uric acid, Mean (± SD)	463.35 (± 157.06)	458.33 (± 154.99)	494.40 (± 166.87)	0.046
Glucose, M (Q_1, Q_3)	6.13 (5.26, 7.62)	6.13 (5.30, 7.62)	6.05 (4.86, 7.73)	0.188
BNP, Mean (± SD)	1510.36 (± 1448.45)	1432.98 (± 1422.14)	1988.67 (± 1524.68)	0.001
D-dimer, M (Q_1, Q_3)	0.68 (0.35, 1.55)	0.59 (0.33, 1.24)	2.15 (1.15, 3.50)	< 0.001
Total cholesterol, Mean (± SD)	4.02 (± 0.99)	4.03 (± 0.97)	3.91 (± 1.06)	0.279
Triglyceride, M (Q_1, Q_3)	1.14 (0.91, 1.50)	1.14 (0.90, 1.50)	1.15 (0.96, 1.49)	0.814
HDL-C, Mean (± SD)	1.00 (± 0.29)	1.02 (± 0.29)	0.90 (± 0.30)	0.001
LDL-C, Mean (± SD)	2.52 (± 0.75)	2.52 (± 0.74)	2.51 (± 0.81)	0.874
Apolipoprotein A1, Mean (± SD)	1.06 (± 0.27)	1.08 (± 0.25)	0.98 (± 0.33)	0.016

Table 1 (continued)

Variables	Total (*N* = 632)	Non-ICT (*N* = 544)	ICT (*N* = 88)	*P*
Apolipoprotein B, Mean (± SD)	0.81 (± 0.20)	0.80 (± 0.19)	0.83 (± 0.22)	0.231
Spironolactone, n (%)				0.088
No	53 (8.39)	41 (7.54)	12 (13.64)	
Yes	579 (91.61)	503 (92.46)	76 (86.36)	
ACEIs/ARBs/ARNI, n (%)				0.040
No	131 (20.73)	105 (19.30)	26 (29.55)	
Yes	501 (79.27)	439 (80.70)	62 (70.45)	
Beta blockers, n (%)				0.001
No	181 (28.64)	142 (26.10)	39 (44.32)	
Yes	451 (71.36)	402 (73.90)	49 (55.68)	
SGLT-2 inhibitors, n (%)				0.086
No	558 (88.29)	475 (87.32)	83 (94.32)	
Yes	74 (11.71)	69 (12.68)	5 (5.68)	
CHA2DS2-VASc score, Mean (± SD)	2.16 (± 1.13)	2.17 (± 1.14)	2.03 (± 1.04)	0.277
CHA2DS2-VASc score, n (%)				0.220
High	441 (69.78)	385 (70.77)	56 (63.64)	
Intermediate/Low	191 (30.22)	159 (29.23)	32 (36.36)	

Note: ICT, intracardiac thrombus; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; LVEF, left ventricular ejection fraction; LAD, left atrial diameter; LVDd, left ventricular end-diastolic diameter; RBC, red blood cell; CV, variable coefficient; SD, standard deviation; ALT, glutamic-pyruvic transaminase; AST, glutamic oxalacetic transaminase; BNP, brain natriuretic peptide; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipoprotein cholesterol; ACEIs, angiotensin-converting enzyme inhibitors; ARBs, angiotensin receptor blockers; ARNI, angiotensin receptor/neprilysin inhibitor; SGLT-2, sodium-glucose cotransporter protein-2

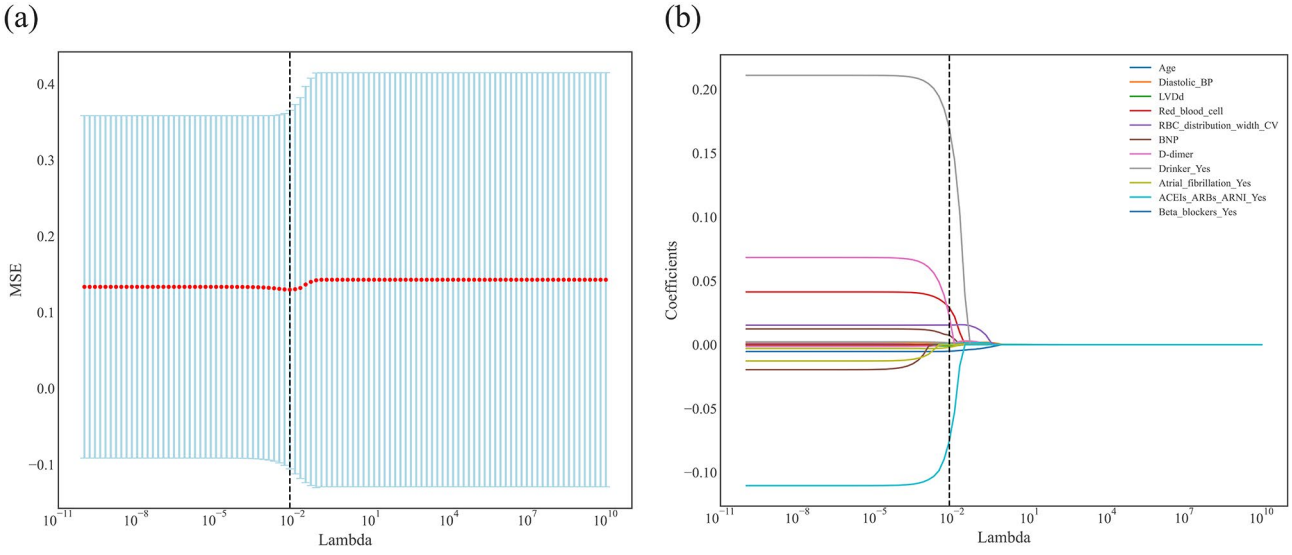

Fig. 2 The least absolute shrinkage and selection operator (LASSO) regression for variable screening. (**A**) changes in mean squared error (MSE) during LASSO regression screening; (**B**) changes in the coefficient profiles during LASSO regression screening

Discussion

In this study, we constructed a prediction model to predict the risk of ICT in patients with DCM and compared the predictive ability of different classifiers. A total of 11 clinical features were utilized to construct the predictive model. The XGBoost model showed better predictive ability than Logistic Regression, SVM, and Random Forest models, with an AUC of 0.947 (training set) and 0.922 (testing set). In addition, D-dimer, age, and AF contributed the most to the XGBoost model among the 11 variables included in the model.

Causes of ICT include arrhythmias, valvular disease, CIEDs-associated thrombus, ischemic and non-ischemic cardiomyopathy, and several systemic diseases-associated thrombus [15]. ICT may also develop elsewhere in the body, and then be transported to the heart. While numerous studies regarding the risk factors of ICT among patients with AF, valvular disease, CIEDs and

Table 2 The performance of different models in predicting the ICT probability in patients with DCM

Models	Datasets	Sensitivity (95%CI)	Specificity (95%CI)	AUC (95%CI)	Accuracy (95%CI)
Logistic Regression	Training set	0.773 (0.685–0.860)	0.803 (0.770–0.837)	0.854 (0.811–0.896)	0.799 (0.768–0.830)
	Testing set	0.704 (0.531–0.876)	0.827 (0.770–0.883)	0.823 (0.733–0.914)	0.810 (0.756–0.864)
SVM	Training set	0.773 (0.685–0.860)	0.623 (0.582–0.664)	0.769 (0.715–0.824)	0.644 (0.607–0.681)
	Testing set	0.667 (0.489–0.844)	0.653 (0.582–0.724)	0.745 (0.645–0.845)	0.655 (0.589–0.721)
Random Forest	Training set	0.886 (0.820–0.953)	0.807 (0.774–0.840)	0.917 (0.887–0.947)	0.818 (0.788–0.848)
	Testing set	0.815 (0.668–0.961)	0.838 (0.783–0.893)	0.880 (0.815–0.945)	0.835 (0.784–0.886)
XGBoost	Training set	0.932 (0.879–0.984)	0.846 (0.815–0.876)	0.947 (0.924–0.969)	0.858 (0.830–0.885)
	Testing set	0.815 (0.668–0.961)	0.879 (0.830–0.927)	0.922 (0.865–0.979)	0.870 (0.823–0.917)

Note: ICT, intracardiac thrombus; DCM, dilated cardiomyopathy; SVM, support vector machine; XGBoost, eXtreme Gradient Boosting; AUC, the area under the receiver operating characteristic curve; CI, confidence interval

ischemic cardiomyopathy have been published, only a few studies have been published regarding non-ischemic cardiomyopathy. Within this subgroup, DCM represents the most common cardiomyopathy. The present study aimed to screen the risk factors of ICT in the context of DCM and develop a predictive model. After excluding the patients with CIEDs, this study enrolled a total of 632 patients with DCM, 88 of whom had ICT. Although the prevalence of ICT in patients with DCM has not been reported, the prevalence was approximately 13.92% in this study. Our study constructed a prediction model to predict the risk of ICT in patients with DCM. In the comparison of the prediction performance of the four different classifiers, the XGBoost model showed the best prediction effect, with an AUC of 0.947 and 0.922 in the training set and testing set, respectively. A recent study used a logistic regression model to construct a nomogram predicting ICT risk in DCM patients, with an AUC of 0.833 [11]. In our study, we compared the predictions of four different classifiers and showed that the XGBoost model was significantly better than the logistic regression model (0.947 vs. 0.854). The XGBoost model also had good sensitivity (0.932) and specificity (0.846) for the prediction of ICT risk. In this study, the sensitivity and specificity of the model represent the model's ability to identify patients with and without ICT risk, respectively. Moreover, of the 11 variables included in the XGBoost model D-dimer, AF, age, beta blockers, RBC, and LVDd all showed significance to the model.

D-dimer is a soluble fibrin degradation product produced by the orderly breakdown of thrombi by the fibrinolytic system and is an important indicator of intravascular thrombosis [16]. Circulating D-dimer levels are low in healthy individuals and elevated in conditions associated with thrombosis [17]. A cohort study demonstrated that higher basal plasma D-dimer concentrations were associated with an increased risk of ischemic stroke [18]. Huang et al. also showed that D-dimer is an important predictor of ICT risk in patients with DCM [11]. For AF, the disorganized electrical activity in the atria leads to ineffective irregular contraction of the atria, which results in incomplete injection of blood into the ventricles and blood stasis [19]. Previous studies have reported that the left atrial appendage (LAA) is the main site of ICT in patients with AF, and more than 90% of the thrombus found in patients with nonvalvular AF and 57% found in patients with valvular AF is formed in the LAA [20, 21]. A meta-analysis showed that patients with DCM had the highest incidence of AF compared to other cardiomyopathy, at about 24%, which was generally consistent with this study's findings (24.8%) [22]. Catheter ablation therapy can significantly improve LVEF and quality of life in patients with AF [23, 24].

Cardiac blood flow characteristics are also different at different ages. ICT is closely related to specific blood flow

Table 3 Delong test for AUC comparison of different models

Datasets	Model	AUC (95%CI)	Statistic	*P*
Training set	Logistic Regression	0.854 (0.811–0.896)	-5.57	< 0.001
	SVM	0.769 (0.715–0.824)	-6.89	< 0.001
	Random Forest	0.917 (0.887–0.947)	-3.97	< 0.001
	XGBoost	0.947 (0.924–0.969)	Ref	
Testing set	Logistic Regression	0.823 (0.733–0.914)	-2.59	0.01
	SVM	0.745 (0.645–0.845)	-3.39	0.001
	Random Forest	0.880 (0.815–0.945)	-2.44	0.015
	XGBoost	0.922 (0.865–0.979)	Ref	

Note: SVM, support vector machine; XGBoost, eXtreme Gradient Boosting; AUC, the area under the receiver operating characteristic curve; CI, confidence interval

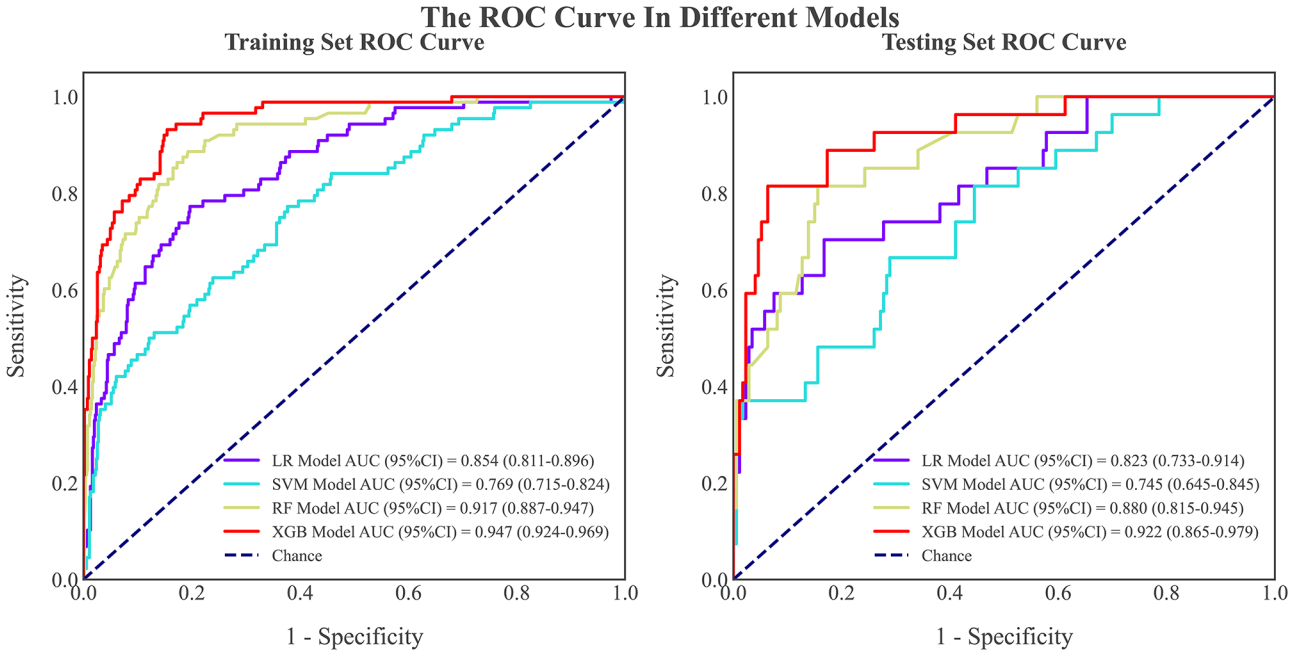

Fig. 3 The receiver operating characteristic (ROC) curves of the predictive models in the training set and testing set

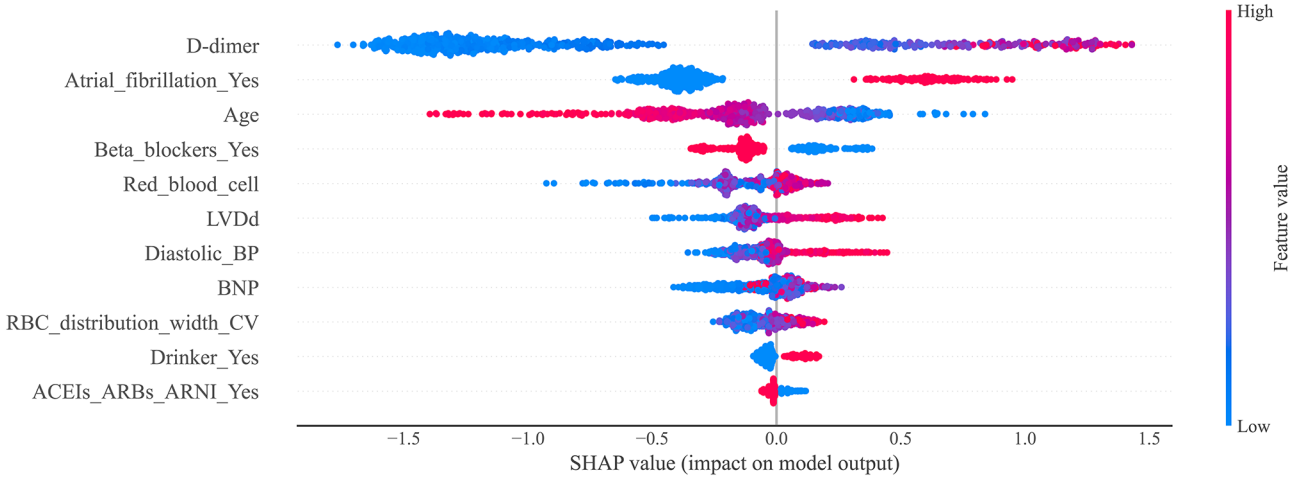

Fig. 4 The SHAP summary plot of the variable importance analysis for the XGBoost model. Each point on the plot specifically refers to a distinct feature of a particular patient. Red represents a high feature value, whereas blue represents a low feature value. The relative importance of each variable is depicted by its y position on the plot. The concentration of red dots for each variable (horizontal axis: >0, <0) represents the direction of the variable's influence on ICH risk (> 0 represents a positive association)

characteristics. Maintaining intracardiac flow homeostasis can effectively avoid blood stasis and ICT. Benito et al. showed an inverted U-shaped relationship between blood residence time in the left ventricle and age [25]. It rose sharply from 1.0 cycles in newborns (62–192 days) to 1.8 cycles in adolescents (10–17 years), and then became shorter, reaching 1.5 cycles by the age of 65. To a certain extent, this explained that middle-aged patients with DCM were also prone to ICT. Our study found that the age of patients in the ICT group was significantly younger than that in the non-ICT group (54 vs. 63 years), which was also consistent with the curve characteristics of Benito et al. [25]. Huang et al. showed that age was negatively associated with the development of ICT in patients with DCM [26], which may be related to the diverse etiology of DCM. Myocarditis is a major cause of DCM, and myocardial cells in patients with myocarditis tend to be severely damaged and more prone to thrombus formation, while younger patients are more likely to develop myocarditis-induced DCM. Yao et al. demonstrated that the mean age of patients in the LVT group was significantly lower than that of the non-LVT

group [27], which further confirms that younger patients are more susceptible to the development of ICT. For the effect of medication, beta-blockers may be related to the formation of ICT. Beta-blockers can inhibit and reduce platelet aggregation, and non-selective beta-blockers are superior to selective beta-blockers [28]. Moreover, beta-blockers are associated with enhanced fibrinolysis and fibrinogen reduction [29].

Epidemiologic studies have shown an association between quantitative and qualitative abnormalities in RBCs (e.g., altered hematocrit) and thrombosis [30, 31]. Our study found that RBC and RBC distribution width-CV were important for predicting ICH in the XGBoost model. RBC distribution width-CV is a parameter reflecting the size of the RBC volume, and its increase indicates an abnormally high degree of RBC size heterogeneity. RBC may contribute to hemostasis and thrombosis through a variety of mechanisms, such as platelet margination leading to an increase in near-wall platelet concentration, blood viscosity, thrombin generation, and platelet activation [32, 33]. Substantial evidence suggested that left ventricular dysfunction was the strongest independent predictor of ICT [34]. Ventricular dilatation can lead to raised wall stress and depressed systolic function [35]. LVDd was the most intuitive indicator of left ventricular dilatation, and this study found that LVDd was an independent risk factor for ICT. A previous study showed that 40% of DCM patients with LVEF less than 40% and LVDd greater than 60 mm had ICT [36]. Wu et al. found that the LVEF was lower and LVDd was larger in the LVT group compared to the no LVT group in patients with DCM [7].

We present the performance of the XGBoost model for predicting ICH risk in DCM patients, and that the XGBoost model based on these 11 features can assist clinicians in identifying patients at high risk for ICH. Inputting these 11 features of any patient with DCM, the XGBoost model can produce a predictive probability of the patient's ICH risk, and patients with a high predictive probability (e.g., >80%) should be promptly treated with anticoagulant therapy or imaging to clarify the presence of a thrombus and its location. This study provides a reference for the selection of clinical features of DCM patients and the application of machine learning in ICH. Although the present study developed a model to predict the risk of ICT in patients with DCM by considering various factors, there were still some limitations of this study. First, compared with transthoracic echocardiography, transesophageal echocardiography or cardiac magnetic resonance examination has higher sensitivity and specificity for ICT, so the incidence of ICT by transthoracic echocardiography may be underestimated. However, transthoracic echocardiography is a non-invasive examination with low medical costs and is still the most commonly used diagnostic method. Second, this was a retrospective study and selection bias cannot be avoided, but this study set strict inclusion and exclusion criteria in order to accurately reflect the actual circumstances of the event. Third, all patients were from the same hospital, and while the predictive model was validated internally in this study, more data from other multicenter cohorts are needed to test it. The current dataset is from the eastern region of China (Anhui Province), and datasets from other regions of China (e.g., central, southern, northern, and western regions) are needed to further validate the broad applicability of the model.

Conclusions

This study developed a prediction model to predict the risk of ICT in patients with DCM. For four different models, the XGBoost model showed better predictive ability than Logistic Regression, SVM, and Random Forest models, with an AUC of 0.947. Of the 11 variables included in the prediction model D-dimer, AF, age, beta blockers, RBC, and LVDd all showed significance to the model. Data from other centers may be needed in the future to further validate the model.

Abbreviations

DCM	Dilated cardiomyopathy
LV	Left ventricular
HF	Heart failure
ICT	Intracardiac thrombus
LVEF	Left ventricular ejection fraction
LVT	Left ventricular thrombus
LVDd	Left ventricular end-diastolic diameter
CIEDs	Cardiac implantable electronic devices
10EPV	10 events per variable
BMI	Body mass index
SBP	Systolic blood pressure
DBP	Diastolic blood pressure
AF	Atrial fibrillation
LAD	Left atrial diameter
RBC	Red blood cell
SD	Standard deviation
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
HDL-C	High-density lipoprotein cholesterol
LDL-C	Low-density lipoprotein cholesterol
ARNI	Angiotensin receptor/neprilysin inhibitor
ACEIs	Angiotensin-converting enzyme inhibitors
ARBs	Angiotensin receptor blockers
SGLT-2	Sodium-glucose cotransporter protein-2

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12872-025-04581-3.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions
Xuetao Zhu and Zeping Hu designed the study. Xuetao Zhu wrote the manuscript. Jun Li, Yi Jiang, and Tianqi Wang collected, analyzed, and interpreted the data. Zeping Hu critically reviewed, edited, and approved the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the Major Project of Natural Science Research of Colleges and Universities in Anhui Province (grant number: KJ2019ZD65) and Anhui Provincial Natural Science Foundation (grant number: 2208085MH200).

Data availability
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The study was conducted in accordance with the Declaration of Helsinki, and was approved by the Ethics Committee of the First Affiliated Hospital of Anhui Medical University. Informed consent was obtained from all subjects involved in the study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 November 2024 / Accepted: 17 February 2025
Published online: 27 March 2025

References

1. Heymans S, Lakdawala NK, Tschöpe C, Klingel K. Dilated cardiomyopathy: causes, mechanisms, and current and future treatment approaches. Lancet (London England). 2023;402:998–1011.
2. Weintraub RG, Semsarian C, Macdonald P. Dilated cardiomyopathy. Lancet (London England). 2017;390:400–14.
3. Merlo M, Cannatà A, Gobbo M, Stolfo D, Elliott PM, Sinagra G. Evolving concepts in dilated cardiomyopathy. Eur J Heart Fail. 2018;20:228–39.
4. Lakdawala NK, Winterfield JR, Funke BH. Dilated cardiomyopathy. Circulation Arrhythmia Electrophysiol. 2013;6:228–37.
5. Patel M, Wei X, Weigel K, Gertz ZM, Kron J, Robinson AA, et al. Diagnosis and treatment of intracardiac Thrombus. J Cardiovasc Pharmacol. 2021;78:361–71.
6. Lemaître AI, Picard F, Maurin V, Faure M, Dos Santos P, Girerd N. Clinical profile and midterm prognosis of left ventricular thrombus in heart failure. ESC Heart Fail. 2021;8:1333–41.
7. Wu HS, Dong JZ, Du X, Hu R, Jia CQ, Li X, et al. Risk factors for left ventricular Thrombus formation in patients with dilated cardiomyopathy. Semin Thromb Hemost. 2023;49:673–8.
8. Yilmaz A, Hayıroğlu Mİ, Salturk S, Pay L, Demircali AA, Coşkun C, et al. Machine learning approach on high risk treadmill exercise test to predict obstructive coronary artery disease by using P, QRS, and T waves' features. Curr Probl Cardiol. 2023;48:101482.
9. Cicek V, Orhan AL, Saylik F, Sharma V, Tur Y, Erdem A, et al. Predicting Short-Term mortality in patients with acute pulmonary embolism with deep learning. Circ J. 2024. https://doi.org/10.1253/circj.CJ-24-0630.
10. Riley RD, Ensor J, Snell KIE, Harrell FE Jr., Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ (Clinical Res ed). 2020;368:m441.
11. Huang Y, Li LC, Li YX, Gui C, Yang LH. Development and validation of a risk model for intracardiac thrombosis in patients with dilated cardiomyopathy: a retrospective study. Sci Rep. 2024;14:1431.
12. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the Euro heart survey on atrial fibrillation. Chest. 2010;137:263–72.
13. Çınar T, Hayıroğlu MI, Tanık VO, Aruğaslan E, Keskin M, Uluganyan M, et al. The predictive value of the CHA2DS2-VASc score in patients with mechanical mitral valve thrombosis. J Thromb Thrombolysis. 2018;45:571–7.
14. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245–7.
15. Egolum UO, Stover DG, Anthony R, Wasserman AM, Lenihan D, Damp JB. Intracardiac thrombus: diagnosis, complications and management. Am J Med Sci. 2013;345:391–5.
16. Johnson ED, Schell JC, Rodgers GM. The D-dimer assay. Am J Hematol. 2019;94:833–9.
17. Weitz JI, Fredenburgh JC, Eikelboom JW. A test in context: D-Dimer. J Am Coll Cardiol. 2017;70:2411–20.
18. Folsom AR, Gottesman RF, Appiah D, Shahar E, Mosley TH. Plasma d-Dimer and incident ischemic stroke and coronary heart disease: the atherosclerosis risk in communities study. Stroke. 2016;47:18–23.
19. Safavi-Naeini P, Rasekh A. Thromboembolism in atrial fibrillation: role of the left atrial appendage. Cardiac Electrophysiol Clin. 2020;12:13–20.
20. Blackshear JL, Odell JA. Appendage obliteration to reduce stroke in cardiac surgical patients with atrial fibrillation. Ann Thorac Surg. 1996;61:755–9.
21. Kapłon-Cieślicka A, Budnik M, Gawałko M, Peller M, Gorczyca I, Michalska A, et al. Atrial fibrillation type and renal dysfunction as important predictors of left atrial thrombus. Heart. 2019;105:1310–5.
22. Noubiap JJ, Bigna JJ, Agbor VN, Mbanga C, Ndoadoumgue AL, Nkeck JR, et al. Meta-analysis of atrial fibrillation in patients with various cardiomyopathies. Am J Cardiol. 2019;124:262–9.
23. Şaylık F, Çınar T, Akbulut T, Hayıroğlu Mİ. Comparison of catheter ablation and medical therapy for atrial fibrillation in heart failure patients: A meta-analysis of randomized controlled trials. Heart Lung. 2023;57:69–74.
24. Nesapiragasan V, Hayıroğlu Mİ, Sciacca V, Sommer P, Sohns C, Fink T. Catheter ablation approaches for the treatment of arrhythmia recurrence in patients with a durable pulmonary vein isolation. Balkan Med J. 2023;40:386–94.
25. Benito Y, Martinez-Legazpi P, Rossini L, Del Pérez C, Yotti R, Martín Peinador Y, et al. Age-Dependence of flow homeostasis in the left ventricle. Front Physiol. 2019;10:485.
26. Huang Y, Zhou WW, Li YX, Chen XZ, Gui C. The use of D-dimer in the diagnosis and risk assessment of intracardiac thrombus among patients with dilated cardiomyopathy. Sci Rep. 2023;13:18075.
27. Yao H, Chen QF, Katsouras CS, Lu Y, Zhou XD. Clinical characteristics of left ventricular thrombus and the use of anticoagulants in patients with dilated cardiomyopathy and sinus rhythm. Eur J Intern Med. 2024;119:146–8.
28. Bonten TN, Plaizier CE, Snoep JJ, Stijnen T, Dekkers OM, van der Bom JG. Effect of β-blockers on platelet aggregation: a systematic review and meta-analysis. Br J Clin Pharmacol. 2014;78:940–9.
29. Teger-Nilsson AC, Larsson PT, Hjemdahl P, Olsson G. Fibrinogen and plasminogen activator inhibitor-1 levels in hypertension and coronary heart disease. Potential effects of beta-blockade. Circulation. 1991;84:Vi72–7.
30. Byrnes JR, Wolberg AS. Red blood cells in thrombosis. Blood. 2017;130:1795–9.
31. Braekkan SK, Mathiesen EB, Njølstad I, Wilsgaard T, Hansen JB. Hematocrit and risk of venous thromboembolism in a general population. The Tromso study. Haematologica. 2010;95:270–5.
32. Weisel JW, Litvinov RI. Red blood cells: the forgotten player in hemostasis and thrombosis. J Thromb Haemost. 2019;17:271–82.
33. Mackman N. The red blood cell death receptor and thrombosis. J Clin Invest. 2018;128:3747–9.
34. Massussi M, Scotti A, Lip GYH, Proietti R. Left ventricular thrombosis: new perspectives on an old problem. Eur Heart J Cardiovasc Pharmacotherapy. 2021;7:158–67.
35. Jefferies JL, Towbin JA. Dilated cardiomyopathy. Lancet (London England). 2010;375:752–62.
36. Blondheim DS, Jacobs LE, Kotler MN, Costacurta GA, Parry WR. Dilated cardiomyopathy with mitral regurgitation: decreased survival despite a low frequency of left ventricular thrombus. Am Heart J. 1991;122:763–71.

Publisher's note