ORIGINAL ARTICLE

MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

MAYO CLINIC

Beyond Atrial Fibrillation: Machine Learning Algorithm Predicts Stroke in Adult Patients With Congenital Heart Disease

Anca Chiriac, MD, PhD; Che Ngufor, PhD; Holly K. van Houten, BA; Raphael Mwangi, MS; Malini Madhavan, MBBS; Peter A. Noseworthy, MD; Samuel J. Asirvatham, MD; Sabrina D. Phillips, MD; and Christopher J. McLeod, MB ChB, PhD

Abstract

Objective: To develop and validate a robust risk prediction model for stroke and systemic embolism (SSE) in adult patients with congenital heart disease (ACHD), using artificial intelligence.

Patients and Methods: Deidentified insurance claims from the Optum Labs Data Warehouse, including enrollment records and medical and pharmacy claims for commercial and Medicare Advantage enrollees, were used to identify 49,276 patients with ACHD, followed between January 1, 2009, and December 31, 2014. The group was randomly divided into development (70%) and validation (30%) cohorts. The development cohort was used to train 2 machine learning (ML) algorithms, regularized Cox regression (RegCox), and extreme gradient boosting (XGBoost) to predict SSE at 1, 2, and 5 years. The Shapley additive explanations (SHAP) model was used to identify the variables particularly driving the SSE risk.

Results: Within this large and diverse cohort of patients with ACHD (mean age, 59 ± 19 years; 25,390 (51.5%) female, 35,766 [77.6%]) white), 1756 (3.6%) patients experienced SSE during follow-up. In the Validation cohort, CHA_2DS_2-VASC had an area under the receiver operating characteristics curve (AUC) of 0.66 for predicting SSE at 1-,2, and 5-years. RegCox had the best predictive performance, with AUCs of 0.82,.81, and.80 at 1-, 2, and 5-years. XGBoost had AUCs of 0.81, 0.80, and 0.79 respectively. Atrial septal defect (ASD) emerged as an important predictor for SSE uncovered by the unbiased ML algorithms. A new clinical risk score, the CHA_2DS_2-VASC-ASD_2 score, provides improved SSE prediction in ACHD. Yet, the ML models still outperformed this.

Conclusion: ML models significantly outperformed the clinical risk scores in patients with ACHD.

 ■ Mayo Clin Proc Digital Health 2024;2(1):92-103

From the Department of Cardiovascular Diseases (A.C., S.D.P., C.J.M.), Mayo Clinic, Jacksonville, FL; Department of Artificial Intelligence and Informatics (C.N.) and Robert D. and Patricia E. Kern Center for the Science of Healthcare Delivery (C.N., H.K.v.H.), Mayo Clinic, Rochester, MN; OptumLabs (H.K.v.H.), Minnetonka, MN; and Department of Quantitative Health Sciences (R.M.) and Department of Cardiovascular Medicine (M.M., P.A.N., S.J.A., C.J.M.) Mayo Clinic, Rochester, MN.

Patients with adult congenital heart disease (ACHD) are at increased risk of stroke from a young age and other mechanisms, beyond atrial fibrillation-driven cardioembolism, may play a role.[1–4] They are also at increased risk of atrial arrhythmias, often starting early in life.[5–7] For patients with documented atrial arrhythmias (atrial fibrillation, atrial tachycardia, or flutter), the clinician may refer to the CHA_2DS_2-VASC score to predict their risk of stroke and guide the initiation of systemic anticoagulation therapy. However, clinical prediction models such as the CHA_2DS_2-VASC score have not been validated in the ACHD population. Furthermore, atrial arrhythmias in ACHD often result from different pathogenic mechanisms compared with the anatomically normal heart, based on atriotomy scars, boundaries and substrates for reentry created by baffles and conduits, and progressive atrial myopathy caused by volume and pressure overload as a consequence of various anatomical defects.[8–13] This contrasts with the typically older age of the population and the different substrate and mechanisms of atrial fibrillation in the absence of congenital cardiac defects.

There is a paucity of data regarding the risk of stroke and systemic embolism (SSE) and the role of systemic anticoagulation in

patients with ACHD and atrial arrhythmias. In these cases, clinical practice is currently guided by expert opinion and consensus statements, and anticoagulation is recommended for the highest risk congenital defects.[14] Furthermore, SSE in this group may be driven by different mechanisms. There is an unmet need for a new risk prediction strategy to detect the real-world risk of SSE in the ACHD population.

We sought to harness the power of machine learning (ML) from a large and diverse data set of patients with ACHD to derive an unbiased algorithm able to predict the 5-year risk of SSE in this population based on their baseline characteristics.

METHODS

Data Source

A retrospective analysis of administrative claims data recorded over a 5-year period between January 1, 2009, and December 31, 2014, within the Optum Labs Data Warehouse was performed; this includes deidentified data for privately insured and Medicare Advantage enrollees, with longitudinal health information available on patients of diverse ages, ethnicities, and geographical regions across the United States. The plan provides comprehensive insurance coverage for physician visits, hospitals, and prescription drug services. This study involved analysis of preexisting and deidentified data, and institutional review board approval was exempted.

Study Population

A total of 49,276 adult patients with a coronary heart disease diagnosis (ICD-9 codes 745-747) were identified and followed up between January 2009 and December 2014. The index date of entrance to the study was defined as the earliest discharge date for qualifying inpatient claims or service date of the second qualifying office visit during the study timeframe. Continuous medical and pharmacy enrollment during the 12 months before the index event was required and used to obtain the baseline characteristics of the population.

Recorded patient characteristics included age (<65, 64−75, and 75+ years), sex, race (non-Hispanic White, Asian, Black, Hispanic, and unknown), census region (Midwest, Northeast, South, West, or unknown), diagnoses, medications, and length of follow-up. Comorbidities included the Charlson−Deyo comorbidity index (0, 1, 2, and 3+) and the 17 conditions that comprise the score. The CHA_2DS_2-VASC score was calculated for each patient at baseline, using diagnoses coded within the 12 month before the index diagnosis and categorized into groups (0−1, 2−3, and 4+). The specific congenital heart disease category was recorded and used as one of the input variables for the ML models (Supplemental Methods).

Primary Outcome

The primary outcome was defined as the 1-year, 2-year, and 5-year incidence of SSE after the index event which prompted entrance in the study. Patients were followed up until their first diagnosis of SSE, end of coverage, or until the end of the study period, whichever came first. SSE included ischemic stroke, transient ischemic attack, and peripheral arterial embolic events. Diagnosis codes for outcomes can be found in Supplemental Methods.

Variable Engineering

Unlike traditional statistical modeling, most ML algorithms require specific data formatting to in other for the algorithm to properly learn the underlying data distributions. Hence, before training, we conducted several data preprocessing steps: numerical variables were normalized to zero mean and unit variance, rare binary/categorical variables (<0.1%) were dropped, and rare levels of categorical variables were combined into a single category. Furthermore, categorical variables were converted into binary format using the 1-hot encoding, that is, a categorical predictor with k possible values were transformed into k binary predictors, with only 1 active predictor.

Risk Stratification Models

Two ML algorithms, regularized Cox proportional hazard (RegCOX) and extreme gradient boosting (XGBoost) were trained on a portion of the data set to predict the risk of SSE based on the baseline characteristics of the ACHD cohort (Table 1). These algorithms are commonly used for clinical risk stratification and have implementations for time-to-event

TABLE 1. Patient Characteristics

	No atrial fibrillation or flutter (n=39,725)	Atrial fibrillation or flutter (n=9551)	All patients (N=49,276)	*P*
Age (y), mean (SD)	55.8 (19.0)	71.5 (13.0)	58.9 (19.1)	<.0001
Age groups				<.0001
<65	24,810 (62.5)	2442 (25.6)	27,252 (55.3)	
65-75	6858 (17.3)	2303 (24.1)	9161 (18.6)	
75+	8057 (20.3)	4806 (50.3)	12,863 (26.1)	
Sex				<.0001
Female	21,043 (53.0)	4347 (45.5)	25,390 (51.5)	
Male	18,682 (47.0)	5204 (54.5)	23,886 (48.5)	
Region				<.0001
Unknown	20 (0.1)	13 (0.1)	33 (0.1)	
Midwest	9865 (24.8)	2781 (29.1)	12,646 (25.7)	
Northeast	7028 (17.7)	2139 (22.4)	9167 (18.6)	
South	17,537 (44.1)	3719 (38.9)	21,256 (43.1)	
West	5275 (13.3)	899 (9.4)	6174 (12.5)	
Race				<.0001
Unknown	1601 (4.0)	441 (4.6)	2042 (4.1)	
Asian	1241 (3.1)	196 (2.1)	1437 (2.9)	
Black	4898 (12.3)	975 (10.2)	5873 (11.9)	
Hispanic	3637 (9.2)	521 (5.5)	4158 (8.4)	
White	28,348 (71.4)	7418 (77.7)	35,766 (72.6)	
Year of index ACHD diagnosis				<.0001
2009	8292 (20.9)	1630 (17.1)	9922 (20.1)	
2010	6837 (17.2)	1426 (14.9)	8263 (16.8)	
2011	6304 (15.9)	1433 (15.0)	7737 (15.7)	
2012	6204 (15.6)	1555 (16.3)	7759 (15.7)	
2013	6490 (16.3)	1847 (19.3)	8337 (16.9)	
2014	5598 (14.1)	1660 (17.4)	7258 (14.7)	
Charlson index, mean (SD)	2.2 (2.6)	4.0 (3.1)	2.5 (2.8)	<.0001
Charlson index groups				<.0001
0	12,628 (31.8)	889 (9.3)	13,517 (27.4)	
1	8674 (21.8)	1246 (13.0)	9920 (20.1)	
2	5432 (13.7)	1357 (14.2)	6789 (13.8)	
3+	12,991 (32.7)	6059 (63.4)	19,050 (38.7)	
Charlson comorbidities				
Myocardial infarction	3985 (10.0)	1929 (20.2)	5914 (12.0)	<.0001
Congestive heart failure	7920 (19.9)	5604 (58.7)	13,524 (27.4)	<.0001
Peripheral vascular disease	8105 (20.4)	3408 (35.7)	11,513 (23.4)	<.0001
Cerebrovascular disease	10,632 (26.8)	3688 (38.6)	14,320 (29.1)	<.0001
Dementia	2610 (6.6)	1329 (13.9)	3939 (8.0)	<.0001
Chronic pulmonary disease	9976 (25.1)	4228 (44.3)	14,204 (28.8)	<.0001
Ulcer	682 (1.7)	327 (3.4)	1009 (2.0)	<.0001
Mild liver disease	2187 (5.5)	741 (7.8)	2928 (5.9)	<.0001
Diabetes without complications	9461 (23.8)	3505 (36.7)	12,966 (26.3)	<.0001
Diabetes with complications	3056 (7.7)	1378 (14.4)	4434 (9.0)	<.0001

Continued on next page

Mayo Clin Proc Digital Health ■ March 2024;2(1):92-103 ■ https://doi.org/10.1016/j.mcpdig.2023.12.002
www.mcpdigitalhealth.org

TABLE 1. Continued

	No atrial fibrillation or flutter (n=39,725)	Atrial fibrillation or flutter (n=9551)	All patients (N=49,276)	*P*
Charlson comorbidities, continued				
Hemiplegia or paraplegia	1283 (3.2)	381 (4.0)	1664 (3.4)	.0002
Moderate/severe renal disease	4171 (10.5)	2502 (26.2)	6673 (13.5)	<.0001
Moderate/severe liver disease	261 (0.7)	91 (1.0)	352 (0.7)	.0021
Metastatic solid tumor	624 (1.6)	252 (2.6)	876 (1.8)	<.0001
AIDS	135 (0.3)	18 (0.2)	153 (0.3)	.0170
Rheumatologic disease	1407 (3.5)	462 (4.8)	1869 (3.8)	<.0001
Other cancer	3564 (9.0)	1493 (15.6)	5057 (10.3)	<.0001
Other comorbidities				
Hypertension	22,676 (57.1)	8068 (84.5)	30,744 (62.4)	<.0001
Atrial fibrillation or flutter	0 (0.0)	9551 (100.0)	9551 (19.4)	<.0001
Pacemaker	613 (1.5)	656 (6.9)	1269 (2.6)	<.0001
CHA_2DS_2Vasc, mean (SD)	2.8 (2.1)	4.5 (2.1)	3.1 (2.2)	<.0001
CHADSVasc groups				<.0001
0-1	14,382 (36.2)	877 (9.2)	15,259 (31.0)	
2-3	11,603 (29.2)	2084 (21.8)	13,687 (27.8)	
4+	13,740 (34.6)	6590 (69.0)	20,330 (41.3)	
CHD groups				<.0001
ASD	12,461 (31.4)	2644 (27.7)	15,105 (30.7)	
Left-side CHD	12,019 (30.3)	3315 (34.7)	15,334 (31.1)	
Other CHD	12,996 (32.7)	3333 (34.9)	16,329 (33.1)	
Single ventricle	391 (1.0)	57 (0.6)	448 (0.9)	
Transposition	274 (0.7)	31 (0.3)	305 (0.6)	
VSD	1584 (4.0)	171 (1.8)	1755 (3.6)	

Values are n (%) unless specified.

ACHD, adult congenital heart disease; CHA_2DS_2-VASC, clinically derived stroke prediction model with points for congestive heart failure, hypertension, age older than 65 or 75 years, diabetes, stroke, vascular disease, and female sex; CHD, congenital heart disease; VSD, ventricular septal defect.

outcomes.[15] See the Supplemental Methods for summary description of these algorithms.

Given that the CHA_2DS_2-VASC score can be computed for any patient at index (cohort entry), we also included the CHA_2DS_2-VASC score as a predictor variable in the ML models. The performance results of ML models without CHA_2DS_2-VASC score are presented in the Supplemental Material.

Training and Validation

We randomly divided the study data into a development cohort (n=40799, 70%) for training and validation of the ML models and a validation cohort (n=17485, 30%) for final evaluation. Because both RegCOX and XGBoost require choosing 1 or more tuning parameters for optimal performance, we set up a grid search to select the best (based on area under the receiver operating characteristic curve [AUC]) combination of hyperparameters through a 10-fold cross-validation procedure using the development cohort. In a 10-fold cross-validation, the training data were randomly partitioned into 10 mutually exclusive subsets (or folds); 9 subsets were used to train the model, and the other (holdout) subset was used to evaluate the performance of the model (Figure 1). This procedure ensures that each instance is included into the testing set once. Then, the best tuning parameters over the 10-fold cross-validation were used to train a final model based on the complete data set (all 10 folds combined).

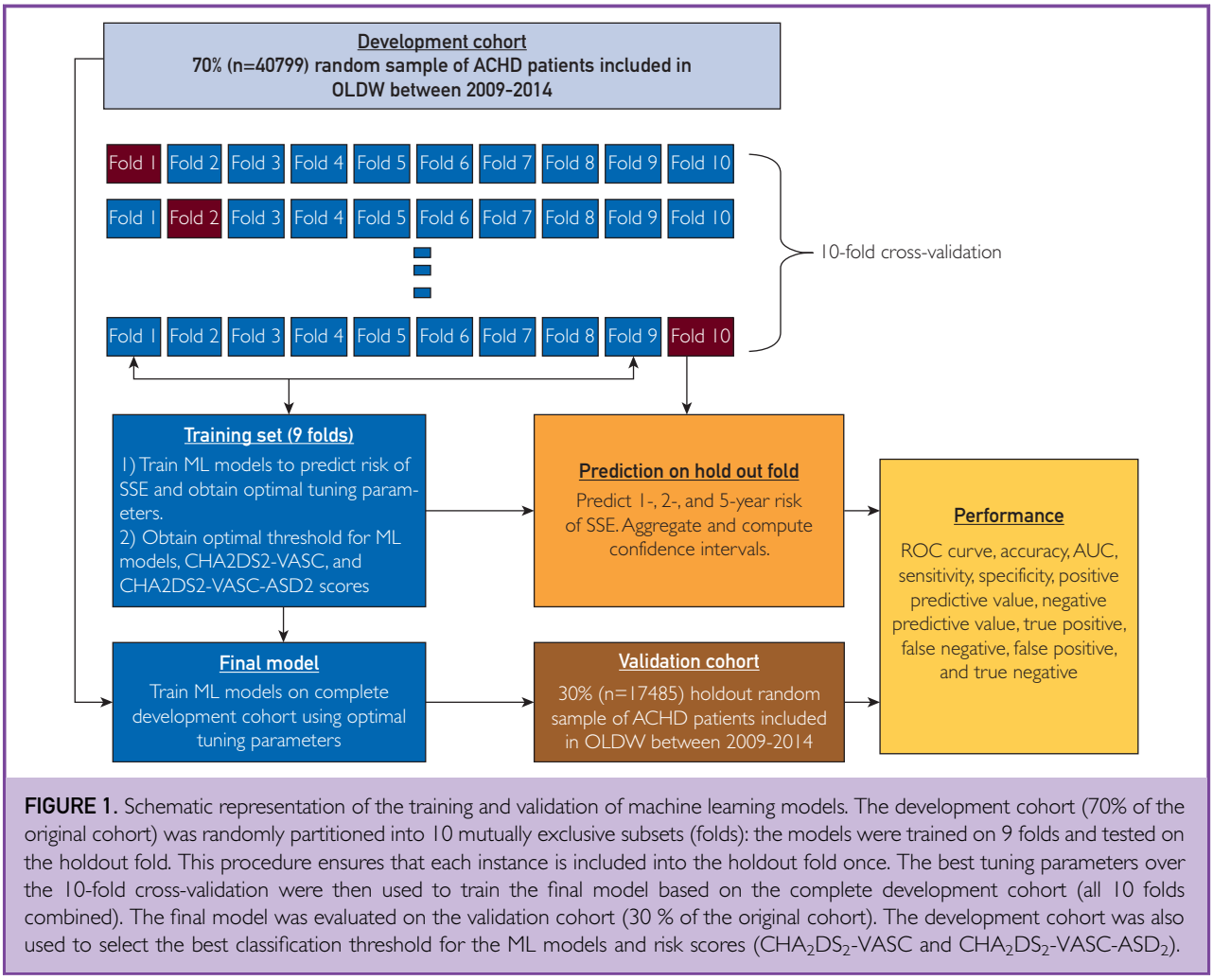

FIGURE 1. Schematic representation of the training and validation of machine learning models. The development cohort (70% of the original cohort) was randomly partitioned into 10 mutually exclusive subsets (folds): the models were trained on 9 folds and tested on the holdout fold. This procedure ensures that each instance is included into the holdout fold once. The best tuning parameters over the 10-fold cross-validation were then used to train the final model based on the complete development cohort (all 10 folds combined). The final model was evaluated on the validation cohort (30 % of the original cohort). The development cohort was also used to select the best classification threshold for the ML models and risk scores (CHA_2DS_2-VASC and CHA_2DS_2-VASC-ASD_2).

Performance Measures

To evaluate the performance of the models in predicting SSE at 1, 2, and 5 years, we computed the accuracy, receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), sensitivity, specificity, positive predictive value, number of true positives, false negatives, false positives, and true negatives at each time point.

Selecting the Optimal Classification Threshold

Using the predicted risk scores from each model, sensitivity, and specificity were computed across all possible threshold values that defined event assignments and used to construct the ROC curve. The optimal classification cutoff was defined as the value that minimizes the distance (d) between the point (0,1) and the ROC curve (Supplemental Figure 1, available online at https://www.mcpdigitalhealth.org/). The optimal cutoff value displaying the best predictive value for each of the models is reported in the results section, along with the corresponding sensitivity.

Statistical Analyses

Analysis was performed using SAS 9.3 and the R programing environment for statistical computing, version 3.5.1. Baseline patient characteristics were described using means and standard deviation for continuous variables and percentages for categorical variables. The abovementioned performance measures,

Mayo Clin Proc Digital Health ■ March 2024;2(1):92-103 ■ https://doi.org/10.1016/j.mcpdig.2023.12.002
www.mcpdigitalhealth.org

and their CIs computed over the 10-fold cross-validation were used to assess the predictive power of the 2 ML models, the CHA_2DS_2-VASC score and a new CHA_2DS_2-VASC-ASD score in predicting SSE at 1, 2, and 5 years. Classification thresholds were defined as the value of the risk score that minimizes the distance to the ROC curve. Because the AUC is the most common and effective performance metric to define the discriminatory ability of a risk scoring system, the presentation of the results will be given in the AUC, using the following categories: excellent (0.9-1.0), good (0.80-0.89), fair (0.70-0.79), poor (0.60-0.69), or fail/no discriminatory capacity (0.50-0.59).

RESULTS

A cohort of 49,276 patients with insurance claims related to ACHD between 2009 and 2014 were included in the study (51.5% female; age, 59 ± 19 years). The most common congenital cardiac defect was atrial septal defect (ASD, 30.7%), followed by the group of left-sided defects (ie, congenital defects of the aortic and mitral valves, bicuspid aortic valve, Shone syndrome, coarctation of the aorta, and persistent ductus arteriosus, 31%). Ventricular septal defects (VSDs, 3.6%), single ventricle physiology (0.9%), and transposition of the great arteries (0.6%) were rare in this cohort, and this was consistent with real-life clinical practice. Baseline characteristics of the ACHD cohort are summarized in Table 1.

Atrial arrhythmias (atrial fibrillation or flutter) were present in one-fifth of the population (9551, 19%) at baseline. The mean CHA_2DS_2-VASC score was 3.1 ± 2.2 for the entire cohort, 4.5 ± 2.1 for patients with atrial arrhythmias, and 2.8 ± 2.1 for the rest. Therapeutic interventions included oral anticoagulation in 6344 (12.9%) patients and antiarrhythmic therapy in 2372 (4.8%) patients (Supplemental Table 2, available online at https://www.mcpdigitalhealth.org/).

The primary outcome, SSE, occurred in 1064 patients at 1 year, 1412 patients at 2 years, and 1717 patients at 5 years, at a rate of 25.77 per 1000 person-years. By the end of follow-up, SSE occurred at a rate of 38.62 per 1000 person-years in patients with known atrial arrhythmia, and in 23.00 per 1000 person-years in patients without known arrhythmia (Supplemental Table 3, available online at https://www.mcpdigitalhealth.org/).

Stroke Prediction Models

When applied to the entire ACHD population, with or without atrial arrhythmias, the CHA_2DS_2-VASC score achieved and maintained fair AUC performance of 0.76 (95% CI, 0.72-0.79) at each time point during follow-up (Table 2 and Figure 2). However, both ML models significantly outperformed the CHA_2DS_2-VASC score. In particular, RegCOX exhibited good AUC performance of 0.83 (95% CI, 0.80-0.85), 0.81 (95% CI, 0.80-0.83), and 0.80 (95% CI, 0.78-0.82) at 1-year, 2-year, and 5-year risk of SSE, respectively.

This powerful performance of the ML algorithms was also demonstrated on the independent validation cohort (RegCOX: 1-year AUC, 0.82; 2-year AUC, 0.81; and 5-year AUC, 0.80; XGBoost: 1-year AUC, 0.81; 2-year AUC, 0.80; and 5-year AUC, 0.79), thus both significantly outperforming the CHA_2DS_2-VASC score (1-year AUC, 0.75; 2-year AUC, 0.75; and 5-year AUC, 0.74). Similar superior performances of the ML models over the CHA_2DS_2-VASC score were observed for accuracy, sensitivity, specificity, true positives, false negatives, false positives, and true negatives.

ASD Emerged as an Important Predictor of Stroke in this ACHD Cohort

The Shapley additive explanations (SHAP) model was used to identify the variables particularly driving the SSE risk. SHAP plots were used to display the top 25 variables from the XGBoost model for patients in the validation cohort (Figure 3). Each variable in the model was assigned a weighting score that reflects the contribution of the variable toward the overall risk prediction. Positive SHAP values indicate the variable contributes toward high-risk prediction (SSE), whereas negative values indicate low risk (no SSE). In Figure 3, variables on the *y* axis are sorted by the sum of absolute SHAP values over all patients, and the *x* axis reports the raw SHAP values (impact of variables on the model). Each plotted point represents an individual patient, and the color represents the variable value: red indicates high and green

TABLE 2. Machine Learning Models, Clinical Risk Scores, and Stroke Prediction in the ACHD Population

Model	Time (d)	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	Threshold	Data
XGBoost	365	0.70 (0.54, 0.83)	0.81 (0.80, 0.84)	0.73 (0.45, 0.95)	0.70 (0.54, 0.84)	0.05 (0.03, 0.06)	0.99 (0.99, 1.00)	0.02 (0.02, 0.02)	Development
	730	0.68 (0.37, 0.87)	0.80 (0.78, 0.83)	0.74 (0.38, 0.98)	0.68 (0.36, 0.89)	0.06 (0.03, 0.08)	0.99 (0.98, 1.00)	0.03 (0.03, 0.03)	Development
	1825	0.70 (0.52, 0.82)	0.79 (0.77, 0.83)	0.74 (0.55, 0.90)	0.70 (0.51, 0.82)	0.07 (0.05, 0.09)	0.99 (0.98, 0.99)	0.05 (0.05, 0.05)	Development
	365	0.69	0.81	0.78	0.68	0.04	0.99	0.02	Validation
	730	0.76	0.80	0.70	0.76	0.07	0.99	0.03	Validation
	1825	0.61	0.79	0.83	0.60	0.06	0.99	0.05	Validation
RegCOX	365	0.78 (0.77, 0.79)	0.83 (0.80, 0.85)	0.72 (0.65, 0.80)	0.78 (0.77, 0.79)	0.05 (0.04, 0.07)	0.99 (0.99, 1.00)	0.03 (0.03, 0.03)	Development
	730	0.75 (0.74, 0.76)	0.81 (0.80, 0.83)	0.72 (0.66, 0.77)	0.75 (0.74, 0.76)	0.06 (0.05, 0.07)	0.99 (0.99, 0.99)	0.04 (0.04, 0.04)	Development
	1825	0.74 (0.73, 0.75)	0.80 (0.78, 0.82)	0.71 (0.66, 0.75)	0.74 (0.73, 0.76)	0.07 (0.06, 0.09)	0.99 (0.99, 0.99)	0.07 (0.07, 0.07)	Development
	365	0.78	0.82	0.73	0.78	0.06	0.99	0.03	Validation
	730	0.75	0.81	0.75	0.75	0.07	0.99	0.04	Validation
	1825	0.74	0.80	0.71	0.74	0.07	0.99	0.07	Validation
CHA_2DS_2VASC ASD_2	365	0.70 (0.69, 0.71)	0.79 (0.76, 0.82)	0.75 (0.70, 0.82)	0.70 (0.68, 0.71)	0.04 (0.03, 0.06)	0.99 (0.99, 1.00)	5	Development
	730	0.70 (0.69, 0.72)	0.78 (0.76, 0.81)	0.74 (0.71, 0.79)	0.70 (0.69, 0.72)	0.06 (0.04, 0.07)	0.99 (0.99, 0.99)	5	Development
	1825	0.70 (0.69, 0.72)	0.77 (0.74, 0.81)	0.72 (0.67, 0.78)	0.70 (0.69, 0.72)	0.07 (0.05, 0.08)	0.99 (0.99, 0.99)	5	Development
	365	0.70	0.77	0.74	0.70	0.04	0.99	5	Validation
	730	0.70	0.77	0.72	0.70	0.05	0.99	5	Validation
	1825	0.70	0.76	0.70	0.70	0.06	0.99	5	Validation
CHA_2DS_2VASC	365	0.66 (0.65, 0.67)	0.76 (0.72, 0.78)	0.72 (0.65, 0.77)	0.66 (0.65, 0.67)	0.04 (0.03, 0.05)	0.99 (0.99, 1.00)	4	Development
	730	0.66 (0.65, 0.68)	0.76 (0.73, 0.79)	0.74 (0.68, 0.79)	0.66 (0.65, 0.68)	0.05 (0.04, 0.06)	0.99 (0.99, 0.99)	4	Development
	1825	0.67 (0.65, 0.68)	0.76 (0.72, 0.79)	0.74 (0.69, 0.80)	0.66 (0.65, 0.68)	0.06 (0.05, 0.07)	0.99 (0.99, 0.99)	4	Development
	365	0.66	0.75	0.70	0.65	0.04	0.99	4	Validation
	730	0.66	0.75	0.71	0.66	0.05	0.99	4	Validation
	1825	0.66	0.74	0.70	0.66	0.06	0.99	4	Validation

Machine learning models versus the CHA_2DS_2-VASC score (RegCOX outperformed XGBoost, and both ML algorithms outperformed the traditional CHA_2DS_2-VASC score). The impact of ASD in the general population: improved stroke prediction compared of the CHA_2DS_2-VASC, but this was still outperformed by the ML models.

ACHD, adult congenital heart disease; ASD, atrial septal defect; CHA_2DS_2-VASC, clinically derived stroke prediction model with points for congestive heart failure, hypertension, age older than 65 or 75 years, diabetes, stroke, vascular disease, and female sex; ML, machine learning; RegCOX, regularized Cox proportional hazard; XGBoost, extreme gradient boosting.

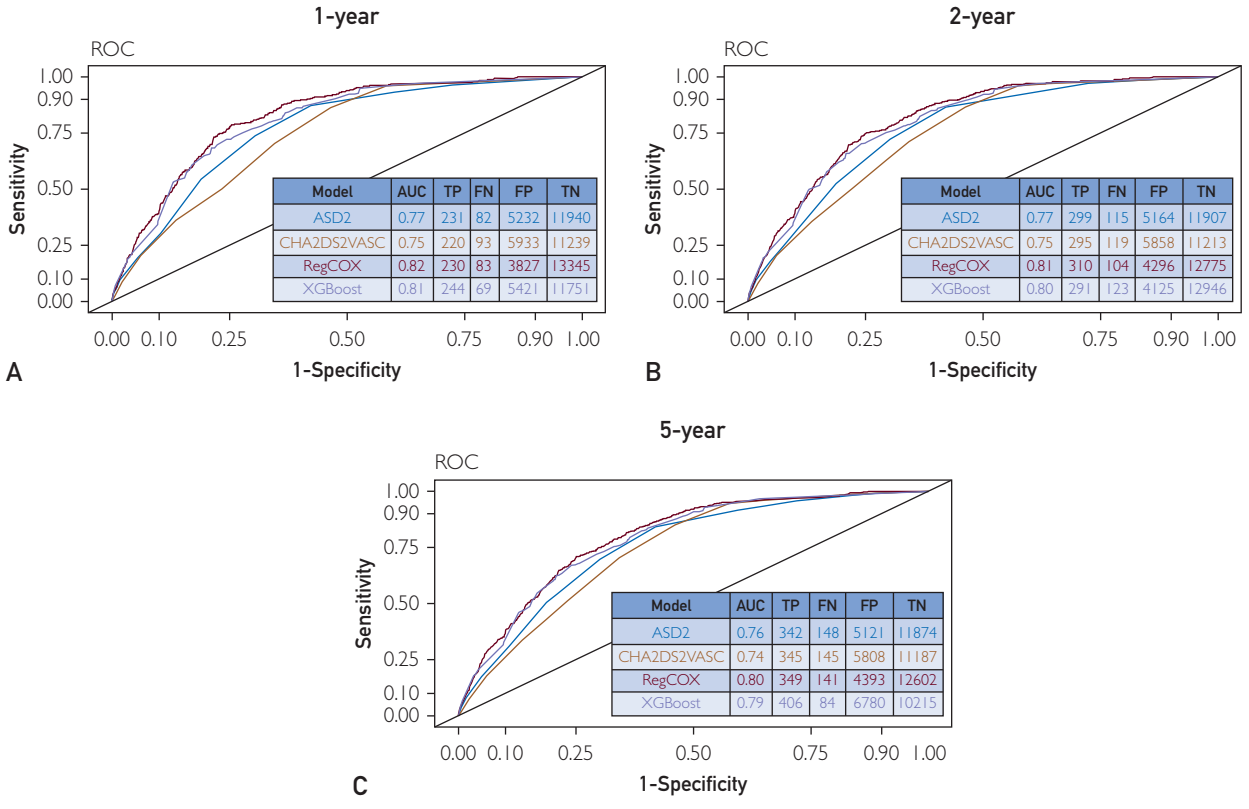

FIGURE 2. Model performance for 1-year, 2-year, and 5-year risk of SSE on the validation cohort. ASD2 represents the new CHA_2DS_2-VASC-ASD_2 score.

low. In this cohort of patients, ASD emerged as a top important variable. In particular, the top 3 drivers of SSE risk included high CHA_2DS_2-VASC scores, presence of cerebrovascular disease, and presence of ASD. However, the coloring of CHA_2DS_2-VASC score indicated that the abovementioned relationships do not necessarily apply to every patient. Although on average high CHA_2DS_2-VASC scores indicate an overall greater risk of SSE (the points are skewed toward positive SHAP values), for some patients (red points to the left), their high CHA_2DS_2-VASC scores were still associated with lower risk of SSE.

Defining a New CHA_2DS_2-VASC + ASD_2 Score

Sensitivity analyzes were performed by successively adding between 0 and 10 points for the presence of an ASD to the CHA_2DS_2-VASC score and calculating the resulting predictive performance for each model (Supplemental Table 5, available online at https://www.mcpdigitalhealth.org/). The optimal performance was achieved by adding 2 points for the presence of an ASD. The new CHA_2DS_2-VASC-ASD_2 score outperformed the CHA_2DS_2-VASC score in its ability to predict the 1-year (AUC, 0.79 [95% CI, 0.76-0.82], 2-year (AUC, 0.78 [95% CI, 0.76-0.81]), and 5-year (AUC, 0.77 [95% CI, 0.74-0.81]) risk of SSE on the development cohort. This predictive advantage was maintained for the validation cohort. However, the improved CHA_2DS_2-VASC-ASD_2 score still could not match the performance of the ML models (Table 2).

ML: Performance of Models in Patients With Known Atrial Arrhythmia

The predictive ability of the ML models was greater for patients without a known atrial

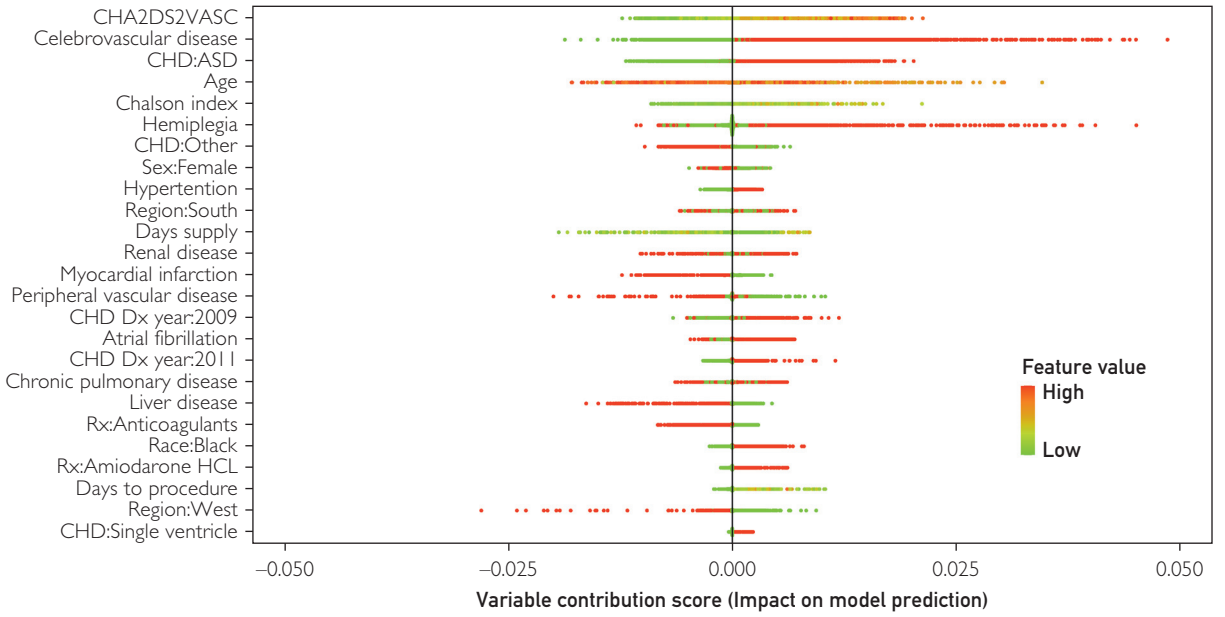

FIGURE 3. SHAP summary plot for the top 25 variables of the XGBoost model contributing to predict risk of SSE. Positive SHAP values (*x* axis) indicate the variable contribute to push the model to make high SSE risk prediction, whereas negative values indicate the variable contribute to push the model to make low SSE risk prediction. Variables on the *y* axis are sorted by the sum of the absolute SHAP values over all patients. Each plotted point represents a patient, and the color represents the variable value: red indicates high and green low. See Supplemental Figure 2 (available online at https://www.mcpdigitalhealth.org/) for SHAP summary plot for the RegCOX model.

arrhythmia diagnosis at baseline, and it was paradoxically weakest in patients with known atrial arrhythmia. Supplemental Table 4 (available online at https://www.mcpdigitalhealth.org/) presents the performance of the models in the validation cohort for patients with and without an atrial arrhythmia diagnosis. The final ML model and the optimal classification threshold were used for the risk models. The best performing model, RegCOX, predicted the 1-year, 2-year, and 5-year risk of SSE in arrhythmia-free patients with good AUCs of 0.85, 0.84, and 0.84 respectively. Incorporating the presence of an ASD into the CHA_2DS_2-VASC score improved its predictive ability for arrhythmia-free patients. However, the presence of an ASD did not make any difference in stroke prediction in patients with arrhythmia. Indeed, the performance of all models was weaker for patients with known atrial arrhythmia.

DISCUSSION

In this study, 2 well-established ML models (XGBoost and RegCOX) were trained and validated in a large population of patients with ACHD of diverse clinical, racial, and geographical backgrounds in the United States to recognize a unique pathologic signature and forecast the 1-year, 2-year, and 5-year risk of SSE with greater precision and accuracy than the CHA_2DS_2-VASC score.

ML has the ability to scan large data sets and identify complex patient signatures, which together may predict the real-life risk of SSE better than clinical risk scores. In this study, both ML models outperformed the CHA_2DS_2-VASC score, and RegCOX had the best predictive power, AUC 0.80 (good performance), at 5 years. XGBoost had a similar performance (AUC 0.79). Comparatively, the CHA_2DS_2-VASC score had a lower predictive power, AUC 0.74 (fair performance).

Mayo Clin Proc Digital Health ■ March 2024;2(1):92-103 ■ https://doi.org/10.1016/j.mcpdig.2023.12.002
www.mcpdigitalhealth.org

Therefore, these ML algorithms provide superior risk stratification for SSE in the ACHD population.

In clinical practice, the physician will often refer to the CHA_2DS_2-VASC score after an atrial arrhythmia diagnosis is firmly established; however, this risk score was previously demonstrated to be predictive of stroke/SSE even in the absence of atrial arrhythmia because it is a compilation of important cardiovascular risk factors for stroke.[16] It has been suggested that perhaps the individual components of the CHA_2DS_2-VASC score may be more important than the atrial rhythm per se. These factors collectively contribute to a proinflammatory and prothrombotic milieu and can set in motion a cascade of events that ultimately culminates in both thrombotic (in situ) and cardioembolic events.[16]

In this study, the CHA_2DS_2-VASC score paradoxically performed better in predicting the risk of SSE in patients with ACHD without a known atrial arrhythmia diagnosis at baseline (which was not the population it was designed for). Furthermore, both ML models also performed better in these patients (XGBoost AUC was 0.78-0.80 for all patients, 0.8-0.83 for arrhythmia-free patients, and 0.65-0.67 for atrial arrhythmia patients). From a clinical standpoint, one may hypothesize that once an atrial arrhythmia diagnosis is established, the clinician may take steps to mitigate the risk by implementing systemic anticoagulation therapy and, perhaps, arrhythmia-targeted interventions. Owing to the nature of this study, it was not possible to analyze the individual effects of specific interventions, whether a rhythm control strategy was pursued and whether it was effective in maintaining sinus rhythm, or whether patients were compliant with anticoagulation therapy when prescribed. However, the large number of patients in this cohort, and the fact that the majority of patients (81%) did not have a known atrial arrhythmia diagnosis at baseline, and were not treated with antiarrhythmics or systemic anticoagulation, only increases the strength of the ML model for SSE prediction in a naïve real-world population. However, ML mirrored the behavior of the CHA_2DS_2-VASC score in patients with or without known atrial arrhythmias (with better performance), suggesting that SSE prediction in ACHD is more complex than the simple dichotomic presence or absence of atrial arrhythmia, and other mechanisms of stroke may be at play.

Although clinical risk prediction models are developed as the sum of a finite number of clinical variables already known to be important (ie, age, hypertension, heart failure, diabetes, vascular disease, and previous stroke), ML models are free of this bias and avoid the "streetlamp paradox" of only considering variables that are already expected to be relevant. Furthermore, ML can incorporate a large number of clinical variables in building the model—limited only by the input—and the algorithm can learn and improve the model with each successive iteration.

In this study, ML also underscored the importance of several key variables, found to be prioritized by the unbiased algorithms. Among the different types of congenital cardiac defects, ASD emerged as a particularly important predictor of stroke in patients with ACHD, seemingly independent of atrial fibrillation. In fact, the top 4 high-risk variables identified by the ML models were the CHA_2DS_2-VASC score, a history of cerebrovascular disease, ASD, and the Charlson comorbidity index. Although the enhanced CHA_2DS_2-VASC-ASD2 risk score demonstrated improved predictive power in arrhythmia-free patients compared with the CHA_2DS_2-VASC score, it performed similarly in patients with atrial arrhythmia. Paradoxical embolization may be an independent mechanism for stroke in these patients, although this requires further study.

The mechanisms of stroke in ACHD are heterogeneous and likely different from the general population.[1–4,16] Pressure and volume overload can lead to atrial fibrosis, stasis and thrombosis, intracardiac shunts and the presence of pacemaker leads can lead to paradoxical embolization, and the hyperviscosity of cyanotic congenital heart disease can lead to in situ thrombosis.[2–4] At the same time, patients with cyanotic ACHD are at increased risk of bleeding, making empirical initiation of systemic anticoagulation without supporting data problematic. In ACHD, there is a paucity of data regarding primary prevention anticoagulation, and current recommendations are based on expert opinion and consensus statements.[14,17] In these

documents, anticoagulation is reserved for the more complex congenital defects and for high burden of atrial arrhythmias only.

This study harnesses the power of advanced ML from a large data set of patients with ACHD within the United States to identify a unique signature that predicts SSE better than the CHA_2DS_2-VASC score and even in the absence of atrial arrhythmia.

We followed a robust model development process through cross-validation to prevent model bias and overfitting and validation on a holdout independent cohort to assess model generalizability outside the development cohort, thus mimicking how the model will be used in clinical practice. We demonstrated that the ML models were consistent, reliable, and interpretable and, thus, can be translated into clinical practice to help with the selection of patients with ACHD at high risk for SSE.

Furthermore, ASD—identified by the ML algorithm as a particular risk factor for stroke in this population—is known to be associated with significant late morbidity and mortality after the fourth decade of life.[18] Studies have reported normal survival in patients who undergo early ASD closure, during childhood, adolescence or early adulthood, and more recently improved survival was also demonstrated in patients who undergo closure later in life.[3,19,20] Although ASD is a known risk factor for stroke due to paradoxical embolization, and ASD closure is an established therapy for stroke prevention,[21,22] multiple studies have also suggested an increased risk of atrial arrhythmias after ASD closure.[23,24] These iatrogenic atrial arrhythmias hypothetically confer a higher risk of stroke in certain cohorts; however, this has not yet been explored. Owing to the nature of this study, using deidentified data from a large cohort of patients with ACHD selected based insurance claims using ICD codes, it was not possible to know the types and severities of various ASDs and whether they had any specific catheter-based or surgical interventions.

In this study, we randomly divided the cohort of patients with ACHD into a development group (70%), used to train the ML algorithms, and a validation group (30%). The ML models demonstrated good performance on the development cohort, and this was also confirmed on the independent validation cohort. Further studies are needed to validate these ML models on different adult congenital populations. Potential implementation in the form of a smartphone or personal computer application integrated with the electronic medical record would expand the use in clinical practice. The updated and simple CHA_2DS_2-VASC-ASD_2 score has a better prediction power than the CHA_2DS_2-VASC score in ACHD and would be readily available for clinical implementation.

Study Limitations

This study was performed using ML from a large data set of diverse patients with ACHD from across the United States; however, 1 requirement for enrollment was to have and maintain a specific insurance coverage throughout the duration of the study. Therefore, this data set may not be representative of the worldwide ACHD population, with different medical insurance coverage or treated outside of the United States or across different periods (as clinical practice may change). However, having a large cohort of patients with diverse clinical, ethnic, and demographic characteristics tried to mitigate this limitation. Future studies are needed to further validate the ML models on different ACHD populations before full clinical application.

CONCLUSIONS

This study highlights that ML can be leveraged in large data sets to reveal unique patient signatures in complex and rare diseases such as ACHD. Being able to provide insight into the prediction of real-life risk of SSE at 1, 2, and 5 years, beyond standard clinical methods, provides the physician with the armamentarium to prevent such events in this population. Both ML algorithms (XGBoost and RegCOX) significantly outperformed the CHA_2DS_2-VASC score, and a new clinical risk score was derived, the CHA_2DS_2-VASC-ASD_2 score, providing improved SSE prediction in ACHD.

POTENTIAL COMPETING INTERESTS

C.N. reports that, in the past 36 months, he has received research support through the Department of Defense (DOD) Prostate Cancer Research Program (PCRP), Health Equity Research and Outcomes

Mayo Clin Proc Digital Health ■ March 2024;2(1):92-103 ■ https://doi.org/10.1016/j.mcpdig.2023.12.002
www.mcpdigitalhealth.org

Improvement Consortium (HEROICA) Award W81XWH2210968. The other authors report no competing interests.

ACKNOWLEDGMENTS

Raphael Mwangi and Holly K. Van Houten had full access to all data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **ACHD**, adult congenital heart disease; **ASD**, atrial septal defect; **CHA_2DS_2-VASC**, clinically derived stroke prediction model with points for congestive heart failure, hypertension, age older than 65 or 75 years, diabetes, stroke, vascular disease, and female sex; **ML**, machine learning; **RegCOX**, regularized Cox proportional hazard; **SSE**, stroke and systemic embolism; **XGBoost**, extreme gradient boosting

Grant Support: This study was supported by the Mayo Clinic Robert D and Patricia E Kern Center for the Science of Health Care Delivery and a grant from the Mayo Clinic Cardiovascular Department.

Correspondence: Address to Anca Chiriac, MD, PhD, Department of Cardiovascular Diseases, Cardiac Electrophysiology, Mayo Clinic, 4500 San Pablo Rd S, Jacksonville, FL 32224 (chiriac.anca@mayo.edu).

ORCID

Anca Chiriac: https://orcid.org/0000-0002-6634-5293

REFERENCES

1. Hoffmann A, Chockalingam P, Balint OH, et al. Cerebrovascular accidents in adult patients with congenital heart disease. *Heart*. 2010;96(15):1223-1226.
2. Mandalenakis Z, Rosengren A, Lappas G, et al. Ischemic stroke in children and young adults with congenital heart disease. *J Am Heart Assoc*. 2016;5(2):e003071.
3. Goldstein SA, Goldstein LB. Neurologic complications of congenital heart disease in adults. *Handb Clin Neurol*. 2021;177:15-22.
4. Lu N, Smith JMC, Andrade JG, Flexman AM, Field TS. Considerations in adult congenital heart disease and stroke: a case report. *Stroke*. 2020;51(8):e148-e150.
5. Triedman JK. Arrhythmias in adults with congenital heart disease. *Heart*. 2002;87(4):383-389.
6. Walsh EP, Cecchin F. Arrhythmias in adult patients with congenital heart disease. *Circulation*. 2007;115(4):534-545.
7. Bouchardy J, Therrien J, Pilote L, et al. Atrial arrhythmias in adults with congenital heart disease. *Circulation*. 2009;120(17):1679-1686.
8. Khairy P, Van Hare GF. Catheter ablation in transposition of the great arteries with Mustard or Senning baffles. *Heart Rhythm*. 2009;6(2):283-289.
9. Frankish K, Daly M, Greenslade J, et al. Electrophysiology assessment and radiofrequency ablation of arrhythmias in adult patients with congenital heart defects: the Christchurch experience. *N Z Med J*. 2014;127(1402):88-96.
10. Houck CA, Teuwen CP, Bogers AJ, de Groot NM. Atrial tachyarrhythmias after atrial switch operation for transposition of the great arteries: treating old surgery with new catheters. *Heart Rhythm*. 2016;13(8):1731-1738.
11. Baysa SJ, Olen M, Kanter RJ. Arrhythmias following the mustard and senning operations for dextro-transposition of the great arteries: clinical aspects and catheter ablation. *Card Electrophysiol Clin*. 2017;9(2):255-271.
12. Houck CA, Chandler SF, Bogers A, et al. Arrhythmia mechanisms and outcomes of ablation in pediatric patients with congenital heart disease. *Circ Arrhythm Electrophysiol*. 2019;12(11):e007663.
13. Chiriac A, Cheema KP, Giardi D, et al. Atrial arrhythmia ablation in patients with D-transposition of the great arteries and atrial switch. *Circ Arrhythm Electrophysiol*. 2022;15(7):e010546.
14. Khairy P, Van Hare GF, Balaji S, et al. PACES/HRS expert consensus statement on the recognition and management of arrhythmias in adult congenital heart disease: developed in partnership between the Pediatric and Congenital Electrophysiology Society (PACES) and the Heart Rhythm Society (HRS). Endorsed by the governing bodies of PACES, HRS, the American College of Cardiology (ACC), the American Heart Association (AHA), the European Heart Rhythm Association (EHRA), the Canadian Heart Rhythm Society (CHRS), and the International Society for Adult Congenital Heart Disease (ISACHD). *Heart Rhythm*. 2014;11(10):e102-e165.
15. Herrin J, Abraham NS, Yao X, et al. Comparative effectiveness of machine learning approaches for predicting gastrointestinal bleeds in patients receiving antithrombotic treatment. *JAMA Netw Open*. 2021;4(5):e2110703.
16. Killu AM, Granger CB, Gersh BJ. Risk stratification for stroke in atrial fibrillation: a critique. *Eur Heart J*. 2019;40(16):1294-1302.
17. Hernandez-Madrid A, Paul T, Abrams D, et al. Arrhythmias in congenital heart disease: a position paper of the European Heart Rhythm Association (EHRA), Association for European Paediatric and Congenital Cardiology (AEPC), and the European Society of Cardiology (ESC) Working Group on Grown-up Congenital heart disease, endorsed by HRS, PACES, APHRS, and SOLAECE. *Europace*. 2018;20(11):1719-1753.
18. Webb G, Gatzoulis MA. Atrial septal defects in the adult: recent progress and overview. *Circulation*. 2006;114(15):1645-1653.
19. Brida M, Diller GP, Kempny A, et al. Atrial septal defect closure in adulthood is associated with normal survival in the mid to longer term. *Heart*. 2019;105(13):1014-1019.
20. Abrahamyan L, Dharma C, Alnasser S, et al. Long-term outcomes after atrial septal defect transcatheter closure by age and against population controls. *JACC Cardiovasc Interv*. 2021;14(5):566-575.
21. Khositseth ACA, Sweeney JP, Fortuin FD, et al. Transcatheter Amplatzer device closure of atrial septal defect and patent foramen ovale in patients with presumed paradoxical embolism. *Mayo Clin Proc*. 2004;79(1):35-41.
22. Bartz PJ, Cetta F, Cabalka AK, et al. Paradoxical emboli in children and young adults: role of atrial septal defect and patent foramen ovale device closure. *Mayo Clin Proc*. 2006;81(5):615-618.
23. Himelfarb JD, Shulman H, Olesovsky CJ, et al. Atrial fibrillation following transcatheter atrial septal defect closure: a systematic review and meta-analysis. *Heart*. 2022;108(15):1216-1224.
24. Evertz R, Reinders M, Houck C, et al. Atrial fibrillation in patients with an atrial septal defect in a single centre cohort during a long clinical follow-up: its association with closure and outcome of therapy. *Open Heart*. 2020;7(2):e001298.