

# HCM-AF-Risk Model to Identify Cases and Predictors of Atrial Fibrillation in Hypertrophic Cardiomyopathy

Moumita Bhattacharya<sup>1</sup>, Dai-Yin Lu<sup>2</sup>, Gabriela Greenland<sup>3</sup>, Hulya Yalcin<sup>2</sup>, Joseph Marine<sup>4</sup>, Jeffrey Olgin<sup>5</sup>, Stefan Zimmerman<sup>2</sup>, Theodore Abraham<sup>3</sup>, Hagit Shatkay<sup>1</sup>, and M. Roselle Abraham<sup>3</sup>

<sup>1</sup>University of Delaware

<sup>2</sup>Johns Hopkins University

<sup>3</sup>University of California San Francisco

<sup>4</sup>Johns Hopkins Hospital

<sup>5</sup>UCSF

August 7, 2020

## Abstract

Background AF in HCM is associated with high stroke risk despite low CHA<sub>2</sub>DS<sub>2</sub>-VASc scores. Hence, there is need to understand AF pathophysiology and predict AF in HCM. We develop/apply a data-driven, machine learning-based method to identify AF cases and clinical features associated with AF in HCM, using electronic health record (EHR) data. Methods Patients with documented paroxysmal/ persistent/permanent AF (n=191) were considered AF cases, and the remaining patients in sinus rhythm (n=640) were tagged as No-AF. We evaluated 93 clinical variables; the most informative variables useful for distinguishing AF from No-AF cases were selected based on the 2-sample t-test and information gain criterion. Results We identified 18 highly informative clinical variables: 11 are positively associated (e.g. LA-diameter, LV-diastolic dysfunction, LV-LGE), and 7 are negatively correlated (e.g. several exercise parameters) with AF in HCM. Next, patient records were represented via these 18 variables. Data imbalance resulting from the relatively low number of AF cases was addressed via a combination of over- and under-sampling strategies. We trained and tested multiple classifiers under this sampling approach, showing effective classification. Specifically, an ensemble of logistic regression and naïve Bayes classifiers, trained based on the 18 variables and corrected for data imbalance, proved most effective for separating AF from No-AF cases (sensitivity=0.74, specificity=0.72, C-index=0.80). Conclusions Our model (HCM-AF-Risk Model), the first machine learning-based method for identification of AF cases in HCM, demonstrates good performance, and suggests that AF is associated with a more severe cardiac HCM-phenotype.

## INTRODUCTION

Hypertrophic Cardiomyopathy (HCM) is characterized by myocyte hypertrophy, myocyte disarray, interstitial/replacement fibrosis, and is associated with a high risk for atrial and ventricular arrhythmias. A large proportion (~25-30%) of HCM patients develop atrial fibrillation (AF) during their lifetime.[1] Notably, HCM patients with AF are at increased risk for stroke even in the setting of low CHA<sub>2</sub>DS<sub>2</sub>-VASc scores.[1-3] Furthermore, stroke can be the first manifestation of AF in HCM. Hence there is a need to understand AF pathophysiology and assess AF risk in HCM patients.[4]

Several small studies have reported an association between age,[5] NYHA class,[6] left atrial (LA) size/function,[5-7] EKG-P-wave dispersion,[7] N-terminal proB-type natriuretic peptide (NT-proBNP) levels,[7] fibrosis in the left ventricle (LV)[6] and AF in HCM patients. Clinical risk scores for AF prediction have been developed and validated using general populations from the Framingham Heart Study (FHS),[8]

Cardiovascular Heart Study (CHS),[9] Atherosclerosis Risk in Communities Study (ARIC),[10] Multi-Ethnic Study of Atherosclerosis Study (MESA)[11], Reykjavik Study (AGES)[12] and Rotterdam Study (RS).[13] But it is unknown whether these models are effective in assessing AF risk in HCM, given the differences in cardiac physiology between HCM patients and the general population.

We hypothesized that a machine learning-based model developed using HCM patients would perform better than existing models (that were developed using clinical data from the general population) to assess risk factors for AF in HCM. In order to achieve this goal, as a first step, we retrospectively identify AF cases and clinical features associated with higher/lower risk of AF, using electronic health record (EHR) data of HCM patients who underwent deep clinical phenotyping by multi-modality imaging. Similar to other machine learning methods and our work in ventricular arrhythmias, [14] our method has the advantage of allowing for thorough, unbiased scanning of the data along several dimensions, which permits derivation of a classification algorithm that stratifies HCM cases based on their respective likelihood to present with AF. Importantly, this method is flexible, can be easily updated as additional clinical data becomes available, and provides insights into the pathophysiology of AF in HCM patients. The next step is prospective testing of our model to predict AF development in HCM.

## METHODS

### 1. CLINICAL DATA AND OUTCOMES

**1.1 Patient Population:** Our HCM Registry (JHH-HCM Registry) is approved by the Institutional Review Boards of the Johns Hopkins Hospital (JHH) and the University of California San Francisco (UCSF). Patients were enrolled in the JHH-HCM Registry during their first visit to the Johns Hopkins HCM-Center of Excellence[15] if they met the standard diagnostic criteria for HCM, namely, maximal LV wall thickness  $\geq 15$  mm[16] in the absence of uncontrolled hypertension, valvular heart disease and HCM phenocopies (amyloidosis, storage disorders).[17]

We performed a retrospective study of all HCM patients from the JHH-HCM Registry, who were evaluated between January 1, 2003 to March 31, 2017. Clinical data including symptoms, comorbidities, medications, history of arrhythmias were ascertained by the examining physician (MRA, TPA) during the initial clinic visit, and during each follow up visit. Rest and treadmill exercise stress-echocardiography (ECHO) and cardiac magnetic resonance imaging (CMR) were performed at the first clinic visit as part of patients' clinical evaluation. Patients who were asymptomatic or had stable symptoms were followed yearly; symptomatic patients were followed more frequently (every 1-3 months) until symptom management was achieved. During yearly follow up visits, patients underwent treadmill exercise-echocardiography and 24h Holter monitoring or ICD interrogation. Patients with an ICD had remote device monitoring and device interrogation performed every 6 months, or more frequently if they were symptomatic or experienced ICD discharges. Patients (without ICD) who had palpitations without evidence of arrhythmias on Holter monitor or exercise-electrocardiogram (EKG), were provided event monitors to document cardiac rhythm during symptoms. A subset of HCM patients (n=145) were referred for perfusion- $^{13}\text{NH}_3$ -PET imaging after ruling out obstructive coronary artery disease by coronary angiography – these patients had angina, ventricular arrhythmias and/or exertional dyspnea despite optimal therapy.

**Atrial fibrillation :** Atrial fibrillation was diagnosed if atrial fibrillation/flutter of any duration was present on 12-lead EKG, telemetry, pacemaker/ICD interrogation, or Holter/event monitoring. Paroxysmal atrial fibrillation (PAF) was defined as AF that terminated spontaneously or with intervention in  $\leq 7$  days of AF onset[18]; persistent AF was defined as AF that lasted  $>7$  days and terminated spontaneously or with treatment; permanent AF was defined as AF that persisted despite treatment to restore sinus rhythm.[19] Atrial fibrillation was diagnosed by review of rest/stress EKGs, event recorder data, Holter monitor data and/or implantable cardioverter defibrillator (ICD) interrogation; a detailed chart review was performed to confirm this criterion. Patients with PAF had confirmed termination of AF within the 7-day window, either by Holter monitor or EKG. Review of medical records, Holter monitor/event recorder studies and ICD interrogations was performed in patients from the No-AF group to ensure no documented history of

AF prior to their first clinic visit and during follow up.

**1.2 Cardiac imaging:** Transthoracic echocardiography was performed using a GE Vivid 7 or E-9 ultrasound machine and a multi-frequency phased-array transducer. Left atrial diameter (anteroposterior) was measured in the parasternal long-axis view at the level of the aortic sinuses, perpendicular to the aortic root long axis, by using the leading-edge to leading-edge convention, just before mitral valve opening (LV end-systole).[20] Echocardiographic images for two dimensional speckle tracking strain analysis were acquired prospectively at frame rates of 50-90 Hz. Longitudinal strain/strain rate was analyzed from the apical two-, three-, and four-chamber views using EchoPAC 112 .[21] Cardiac magnetic resonance (CMR) imaging was performed using a 1.5T system at the first clinic visit. The contrast agent, gadopentetate dimeglumine was used at 0.2 mmol/kg.[22] Left ventricular mass and late gadolinium enhancement (LV-LGE) were quantified using QMass software (QMass 7.4; Medis, Leiden, The Netherlands). Cardiac  $^{13}\text{NH}_3$ -PET/CT imaging was performed using a GE Discovery VCT PET/CT System and a 1-day rest/stress protocol, as described previously.[23] Please refer to the *Supplementary file, Section A1* for detailed methods for rest/stress echocardiography, CMR and PET/CT imaging.

**1.3 Cardiovascular events during followup:** All analyses were blinded to AF outcome. Cardiovascular adverse events, including AF, stroke, heart failure, sustained ventricular tachycardia (VT), ventricular fibrillation (VF) and death were documented in the HCM Registry. All-cause mortality statistics for our study population were obtained by linking our database to the Social Security Death Index. A detailed description of methods is provided in the *Supplementary file, Section A2* .

**1.4 Statistics:** Descriptive statistics were performed on patient demographics, hemodynamics, echocardiographic and CMR parameters and cardiovascular events, stratified by the presence/absence of AF. Continuous variables are presented as mean  $\pm$  standard deviation and categorical variables as the total number and percentage. Comparisons between patients with/without AF was performed using the independent  $t$  test for continuous variables and the Fischer exact test for categorical variables. Statistical analyses were performed using STATA 14 (StataCorp LP, College Station, Texas).

## 2. COMPUTATIONAL METHODS

HCM patients with at least one episode of AF ( $n=191$ ), either prior to their first clinic visit ( $n=139$ ) or during follow up ( $n=52$ ), were considered AF cases, and the remaining patients who were in sinus rhythm ( $n=640$ ) were labeled as No-AF (Supplementary Figure 1).

Supplementary Figure 2 summarizes the computational framework (*HCM-AF-Risk Model*) that we introduce for identifying HCM patients with AF. It comprises 5 steps: 1) preprocessing to remove variables directly correlated with AF, and to address missing data; 2) feature selection, in which informative, predictive clinical variables that distinguish AF cases from No-AF are identified; 3) association analysis to quantify the degree of association between each predictor variable and the AF class; 4) supervised machine learning for building and training classifiers and performing classification; and 5) thorough quantitative and qualitative evaluation to assess the classifier's performance.

**2.1 Preprocessing:** We first removed variables that had no relevance to risk of AF (e.g. visit date, patient ID), as well as variables directly indicative of adverse outcomes (e.g. ventricular arrhythmia, heart failure, AF). The feature set remaining at the end of this step consisted of 93 clinical variables (Supplementary Table 1). As some of the records did not include values for all variables, data imputation was performed using a nearest neighbor approach (see *Supplementary file, Section B.1.1* for details).

**2.2 Feature selection:** When high-dimensional data is used for classifier training, the classifiers often exhibit low performance on the test set due to overfitting of the model to the specific training set. That is, certain features may show discriminating power within a limited dataset but not generalize beyond that small training set. Moreover, many of the features are not informative for distinguishing among the different classes (in this case *AF* vs No-AF records). We thus apply feature selection aiming to identify attributes most informative for AF, while reducing data dimensionality to avoid overfitting. We note that our dataset

comprises both nominal and continuous attributes, also referred to as features. While the classification method presented here is multivariate, the feature selection is performed by assessing individual variables one at a time. Selection of highly predictive nominal attributes relies on the well-known Information Gain criterion [24] which measures the information gained about the AF-class given the value assumed by the attribute. For continuous features, we used the two sample t-test under unequal variance [25, 26], testing whether the distribution of attribute values associated with AF cases is significantly different from that associated with No-AF cases. The resulting reduced feature set employed here contains only those continuous attributes for which the t-test indicated a highly statistically significant distributional difference ( $p \leq 0.01$ ), and those nominal attributes for which the information gain value was greater than 0.002. The threshold value was determined through an iterative process in which the least informative feature is removed and left out of the classification procedure at each iteration. Feature-removal was repeated in this way until the classification performance indicated deterioration, at which point all the remaining features were retained. This feature selection process resulted in 18 clinical variables deemed to be informative and predictive of AF in HCM patients (Table 1).

**2.3 Association Analysis:** Many of the attributes gathered per patient are nominal, as opposed to continuous-numerical (Supplementary Table 1). Nominal features include variables such as *HCM type* or *history of syncope*. Association among nominal variables cannot be calculated using the standard Pearson correlation. Thus, to express the degree and direction of association between the predictor variables and the outcome variable, we employ the polychoric correlation [27, 28], which takes on values in the range  $[-1, 1]$ , where a negative value indicates negative association and a positive value corresponds to positive association.

**2.4 Classification:** Our classifier operates by taking as input a vector of values representing a patient's record, and assigning a probability that indicates the patient's likelihood to belong to the AF vs No-AF class. Each patient in our dataset is represented using the 18 distinguishing features identified in the feature-selection step. Specifically, each of the 831 patients, denoted  $p^i$  ( $i \in [1, 831]$ ), is mapped to a 18-dimensional vector,  $V^i = \langle p_1^i, \dots, p_{18}^i \rangle$ , where each entry in the vector corresponds to the clinical value recorded for the respective variable. The classifier calculates for each 18 dimensional vector  $V^i$  representing the  $i^{th}$  patient, its probability to be an AF case,  $Pr(AF | V^i)$  vs its probability to be No-AF, ( $Pr(No-AF | V^i) = 1 - Pr(AF | V^i)$ ). The higher the value  $Pr(AF | V^i)$ , the more likely the patient is to have AF. As an illustration, a probability of 0.9 to be an AF case assigned to patient  $p$  indicates a high risk for AF, while a probability of 0.3 suggests that the risk for AF is much lower. We note that both the representation of the patients based on readily interpretable clinical values and the classification decision that corresponds to assigning a severity-probability are unique to this work. It stands in contrast to most recently published work in machine learning within the clinical domain [29-31] where a complex model architecture based on artificial neural networks is used, typically acting as a 'black-box' that provides the categorization of the patient without the ability to track down the justification or explanation.

For comparison among the performance of different potential candidate machine learning classifiers, we tested four standard (yet probabilistic and explainable) machine learning classification methods, namely, Logistic Regression, Naive Bayes, Decision Tree and Random Forest, assessing how well they separate AF cases from No-AF (Supplementary Table 2). We used the Python scikit-learn package for training the baseline classifiers [32]. All four classifiers performed poorly when trained on our highly imbalanced dataset, failing to detect almost any AF records (Table 2). We have thus employed a method we have devised for addressing imbalance [14] by combining over- and under-sampling along with an ensemble classifier that integrates the most effective classifiers to separate AF records from No-AF records. The over- and under-sampling strategy is based on partitioning the training data, as shown in Supplementary Figure 3. For a more detailed description of the classification model and its testing see *Supplementary Data Section B1.2*.

**2.5 Model Evaluation:** We employed several standard measures [24] to assess the performance of our *HCM-AF-Risk Model*, namely, specificity, sensitivity (recall) and area under receiver operating characteristics (ROC) curve. The first two are defined as:  $Specificity = \frac{TN}{TN+FP}$ ,  $Sensitivity = \frac{TP}{TP+FN}$ , where TP (True Positives) denotes AF records that are correctly labeled as AF by the classifier; TN (true negatives) denotes

records that are not associated with AF and are not assigned to this class by the classifier; FP (false positives) denotes records not associated with AF that are misclassified by the classifier as AF; FN (false negatives) denotes AF records that were incorrectly labeled by the classifier as No-AF. The ROC curve plots the true positive rate (TPR), calculated as  $\frac{TP}{TP+FN}$ , as a function of false positive rate (FPR), calculated as  $\frac{FP}{FP+TN}$  (FP-false positive). The classifier performance is then reported based on the area under the ROC curve (C-index).

**2.6 Experiments:** We first employed univariate feature selection using information gain and the

two-sample t-test for unequal variance, to identify salient features that are associated with AF. We also employed polychoric correlation to determine the association between the pertinent features identified by the feature selection method and the outcome variable (AF/No-AF). We represented patients using the reduced set of features and trained simple classifiers as baseline, as described earlier. To address the data imbalance challenge, we applied the combination of under- and over-sampling to obtain a balanced training set, which was then used to train the four simple classifiers, and the ensemble classifier comprising logistic regression and naive Bayes. Each of the classifiers was trained and tested using five-fold cross-validation, in which the data is partitioned into five equal subsets – and five iterations of learning are performed, each of which uses 80% (4 of the subsets) for training while the fifth (leftout) subset is used for testing. After each training iteration, we evaluated the performance of the resulting classifier on the imbalanced test set in terms of specificity, sensitivity and the area under the ROC curve. We ran 10 complete sets of 5-fold cross validation experiments, for a total of 50 runs; the performance reported is averaged over these 50 runs.

To address data imbalance we first applied methods that were previously reported in the literature, such as simple oversampling, simple under-sampling, Adaptive Synthetic Sampling Approach (ADASYN)[33] and Meta-classification[34] and found that the performance of our combined under- and over-sampling using SMOTE was superior. Hence, we report only the results obtained using our method (*HCM-AF-Risk Model*) which combines under- and over-sampling, and compare this method against the baseline classifiers.

**2.7 Comparisons:** We also evaluated the performance attained by our model when trained on the dataset where it is represented based on three additional feature sets (Table 3, top three rows). Those feature sets correspond to sets of attributes that were reported as predictive in three seminal AF risk identification studies, namely the Framingham Heart Study,[8] ARIC[10] and CHARGE-AF Consortium[9].

## RESULTS

**Patient population :** We studied 831 patients with a clinical diagnosis of HCM. Atrial fibrillation was diagnosed in 22% of the HCM cohort: 139 patients were diagnosed with AF prior to/at the first clinic visit, and 52 patients were diagnosed with AF during follow up (Supplementary Figure 1). Prevalence of AF varied from 9-30% and increased with age; AF prevalence was highest in the age group of 61-80 years (Figure 1).

Demographic, clinical and imaging features of the HCM cohort at the time of their first clinic visit are presented in Table 4. Patients from the AF group were older, more likely to have obstructive HCM, higher NYHA class, and lower exercise capacity than the No-AF group. The AF group also had larger LA size (Figure 2), greater diastolic dysfunction, worse global longitudinal strain and greater amount of LV replacement fibrosis (reflected by LV-LGE), when compared to the No-AF group, suggesting greater degree of LV myopathy. No difference was observed in LV mass, maximum LV thickness or LVOT gradients between the AF and No-AF groups of HCM patients (Table 4).

Mean follow up was 3.1 years (median = 2.1, 25<sup>th</sup>-75<sup>th</sup> percentile = 1.0 - 4.8 years). HCM patients from the AF group had a higher incidence of heart failure and all-cause death, when compared to the No-AF group (Table 4).

**Machine learning-based identification of atrial fibrillation cases:** Our feature selection process identified 18 clinical variables whose values distinguish AF from No-AF cases within the HCM population. Table 1 provides a list of these predictive variables, along with the corresponding polychoric correlation and P values, indicating their degree of association (or lack thereof) with AF. We identified 7 variables that are

negatively correlated with AF, and 11 variables that are positively associated with AF. Left atrial diameter is highly correlated with AF. Several exercise-related parameters, including, lower exercise capacity (reflected by lower METs, exercise time, peak stress heart rate), abnormal BP response to exercise, lower diastolic BP at peak exercise and lower heart rate recovery post-exercise are associated with higher risk for AF. Other predictors of AF include replacement fibrosis in the LV (reflected by LV-LGE), greater diastolic dysfunction (reflected by higher E/A and E/e', lower LV-SR.E) and worse (more positive) global longitudinal systolic strain rate (LV-SR.S).

Notably, combining the ensemble classifier comprising logistic regression and naive Bayes with over- and under-sampling led to higher sensitivity and area under the receiver operating curve (AUC), compared to the four simple classifiers (naive Bayes, logistic regression, decision tree and random forest) alone (Table 2). Figure 3 illustrates the C-index (0.80) for our method (*HCM-AF-Risk Model*), which assigns an individualized probability to each patient to present with AF.

**Comparison of HCM-AF-Risk Model performance with previous AF models :** We compared the performance of the *HCM-AF-Risk Model* with that obtained by the Framingham Heart Study,[8] ARIC[10] and CHARGE-AF[9] risk models (Tables 3, 5). The *HCM-AF-Risk Model* demonstrates significantly higher performance ( $p < 0.001$ ) across all evaluation metrics including specificity, sensitivity and area under ROC curve (C-index) for HCM patients, when compared with published models [8-10] focused on AF prediction in the general population. The datasets used in these studies are not publicly available, which preclude their use for training/testing on our dataset and comparing their performance according to all the measures we have used. As such, we compared the performance level attained by our model with that reported by the other studies in terms of the C-index.

We also repeated our experiments, using only LA diameter to represent our dataset, based on results from several previous studies [2, 35, 36] that identified LA enlargement as most predictive of AF in HCM. We observed reduction in our model's performance when LA diameter alone is included in the feature set: specifically, area under ROC curve decreased by 20% (C-index: 0.66 from 0.80), sensitivity decreased by 37% (0.54 from 0.74) and specificity decreased by 14% (0.63 from 0.72).

## DISCUSSION

The *HCM-AF-Risk Model* is the first machine learning-based method for the identification of AF cases and clinical features associated with higher/lower risk of AF in HCM, using electronic health record data. In our model, individual patient data is represented as an N-dimensional vector, and the model output is a probability score for AF (AF risk) in HCM. We identified 18 clinical variables that are highly associated (positively/negatively) with AF in HCM patients. In addition to age, NYHA class, LA size and LV fibrosis that have been previously associated with AF in HCM, we found additional clinical features such as LV diastolic dysfunction/lower LV-systolic strain that are positively associated with AF, and greater exercise capacity that is negatively associated with AF in HCM.

**HCM-AF-Risk Model :** Employing a statistical machine learning method is advantageous as it allows automatic quantification of the likelihood of an event (AF, in this case) based on the combination of feature values - as obtained from the patient's electronic health records - and their level of association with the event. Moreover, unlike traditional rule based models, machine learning methods are robust in the face of new data, as these methods can tune and update their parameters, which govern the classification algorithm based on the added data. Thus, machine learning methods are well-suited for use in the clinical setting where additional patients' data is frequently accumulated.

In contrast to the majority of current 'black-box' machine learning methods that are based on artificial neural networks [29, 37, 38] and whose output decision typically cannot be explained, our method is based on modeling a clear probabilistic decision process, that can be tracked back and used to justify the decision - we believe that this is a critical aspect when using machine learning for supporting clinical decisions. Our *HCM-AF-Risk Model* addresses data imbalance, and utilizes a set of 18 clinical variables to identify AF cases, and clinical features associated with higher/lower risk for AF in HCM patients.

We note that heart failure, along with VT/VF and stroke, were not included in the list of clinical variables considered by our method. This is because our goal is to identify demographic, clinical, and imaging features that predict adverse outcomes (AF in this case) in HCM patients, and using such adverse outcomes as predictors defeats this purpose. However, as several previous studies including the Framingham Heart Study,[8] ARIC [10] and CHARGE-AF [9] have shown heart failure to be a predictor of risk for AF, we assessed the performance of our model while including heart failure. Inclusion of heart failure did not increase AUC or sensitivity, but led to a slight increase in specificity of our model, from 0.72 to 0.73.

**Clinical predictors of AF in HCM using the HCM-AF-Risk Model:** Left atrial diameter is the strongest predictor of AF in our study. The association between LA size and AF has been extensively documented in the general population[39-43] and HCM patients.[2, 44-48] The association between LA enlargement and AF has been attributed to stretch-induced LA structural and electrophysiologic remodeling.[49] In the case of HCM, since most causal HCM mutations are expressed in both atrial and ventricular myocytes, atrial myopathy and LV diastolic dysfunction could underlie the high prevalence of AF in HCM.

Our *HCM-AF-Risk Model* indicates an association between diastolic dysfunction and AF in HCM. We found that higher values for E/A, E/e'[50] and lower (worse) global diastolic strain rate reflecting greater degree of diastolic dysfunction are associated with higher risk for AF in HCM. Similar results have been reported in studies conducted in non-HCM patients.[39, 51, 52] The mechanism whereby diastolic dysfunction has been proposed to predispose to AF is by increasing LA preload (stretch), afterload and wall stress (dilation), which lead to ion channel remodeling, fibrosis and increase susceptibility for reentrant arrhythmias such as atrial fibrillation/flutter. [51]

Left ventricular fibrosis (LV-LGE) and worse LV global longitudinal peak systolic strain rate, which reflect greater degree of LV myopathy are associated with AF in our model. Several previous studies have detected an association between LV fibrosis and AF in HCM.[53-55] A recent CMR study in HCM patients reported greater amounts of LA fibrosis in HCM patients with PAF, as well as a positive association between atrial and ventricular fibrosis (LGE).[56] Since fibrosis slows conduction and predisposes to reentry, LA fibrosis would be expected to increase risk for AF.

Lower exercise capacity, lower chronotropic response/heart rate recovery, abnormal BP response to exercise and lower diastolic BP at peak exercise are associated with higher risk for AF in our study. Similar results of exercise intolerance in HCM patients with PAF have been reported in a previous study of 265 HCM patients during sinus rhythm[57] – here, the authors did not observe an association between lower exercise capacity and diastolic dysfunction or LA volume. Additionally, ECHO[58] and CMR[5, 56] studies in HCM patients have revealed greater impairment of LA function and greater degree of LA fibrosis in HCM patients with PAF, suggesting that PAF is a marker of LA myopathy.

One mechanism underlying reduced exercise capacity in HCM patients (with PAF), even during sinus rhythm[57] could be impairment of LV hemodynamics in the setting of LA myopathy, since the LA modulates LV performance by its reservoir function during ventricular systole, conduit function during early ventricular diastole and booster pump function during late ventricular diastole. A second possibility is higher pulmonary capillary wedge pressure (PCWP) in HCM patients with AF, based on results of a study in 123 patients who underwent simultaneous left and right heart catheterization, where PCWP was higher than LV end-diastolic pressure (LVEDP) among AF patients and lower than LVEDP among patients in sinus rhythm.[59] Other contributors to lower exercise capacity in HCM patients with AF include sympathovagal imbalance[60] leading to systemic vasodilation, chronotropic incompetence induced by atrial remodeling/medications, and greater degree of LV myopathy.

**Comparison of predictors for atrial fibrillation and ventricular arrhythmias identified by the HCM-AF-Risk and HCM-VAR-Risk Models:** In an earlier study[14] we developed a machine-learning based model for predicting lethal ventricular arrhythmias (VT/VF) in HCM patients. We identified 5 predictors (exercise time, METs, E/e' ratio, LV global longitudinal peak systolic strain rate and LV global longitudinal early diastolic strain rate) that are common in the two models and 13 variables associated with

AF, but not VT/VF (Supplementary Table 3).

Higher age is associated with increased risk for AF, but lower risk for VT/VF, which has been confirmed by other studies.[2, 61] HCM type (non-obstructive), family history of HCM or sudden cardiac death and non-sustained VT are associated with VT/VF but not AF, which may reflect differences in arrhythmic substrate in the LV and LA in HCM. Notably, LV hypertrophy (max IVS thickness, IVS/PW ratio) is associated with VT/VF but not AF – higher risk for VT/VF but not AF could be attributed to greater degree of myocardial ischemia,[62] interstitial fibrosis[63] and myocyte disarray [64] in the hypertrophied LV.[65] The association of replacement fibrosis (LV-LGE) with AF but not VT/VF could reflect the impact of greater degree of diastolic dysfunction induced by LV fibrosis resulting in LA dilatation/remodeling and AF. Taken together, our results suggest distinct pathophysiologic mechanisms underlying atrial and ventricular arrhythmias in HCM (Supplementary Table 3).

## LIMITATIONS

This is a single center, retrospective study. Although our approach is especially effective when the size of the dataset and the number of examples in the under-represented class (AF in our case) are limited, our approach has limitations. When working with a larger imbalanced dataset, the under-sampling step involved in creating a balanced training set eliminates a sizeable portion of the over-represented class, while the over-sampling step applied via the SMOTE process to the under-represented class generates a large number of synthesized samples that were not in the original dataset. Both of these lead to a potential loss of useful information and alter the distribution of characteristic feature values across both the minority and the majority classes. In the current study, this issue was mitigated by thorough experimentation to determine the effective rates of under- and over-sampling.

We grouped all AF cases (paroxysmal, persistent, permanent), as well as prevalent and incident AF into one set (because of low event number) and only included ECHO/CMR imaging features obtained at the patients' first clinic visit in the model. Left atrial volume,[46] LA strain,[58] LA fibrosis,[56] EKG/blood biomarkers,[66] genotype,[67-69] or sleep apnea were not included in our model, because this data is not available for a large proportion of our cohort. Lastly, we were unable to assess the generalizability of our approach by applying our developed model to additional HCM patients – beyond the cross-validation study, due to the unavailability of data from other HCM cohorts reported in other studies. We expect to address the latter issue in a future prospective study.

## CONCLUSION

The *HCM-AF-Risk Model* effectively identifies HCM patients with AF, in addition to predictors of AF in HCM. Our approach attains good performance (0.74 sensitivity, 0.72 specificity, C-index = 0.80) improving upon current AF risk scores from community cohorts, while addressing the imbalance between high-risk and low-risk cases that is inherent in most clinical data. The set of clinical attributes identified by our method as indicative of AF – and serving to justify the severity level assigned by the classifier, includes several hitherto un-identified markers of AF in HCM, and suggests that HCM patients with AF have a more severe cardiac HCM phenotype.

**Funding:** This work was funded in part by NSF IIS EAGER grant#1650851 (to HS), an award from the John Taylor Babbitt (JTB) foundation (Chatham, New Jersey) and startup funds from the UCSF Division of Cardiology (to MRA).

## REFERENCES

- [1] Camm CF, Camm AJ. Atrial Fibrillation and Anticoagulation in Hypertrophic Cardiomyopathy. Arrhythmia & electrophysiology review. 2017;6:63-8.
- [2] Olivetto I, Cecchi F, Casey SA, Dolara A, Traverse JH, Maron BJ. Impact of atrial fibrillation on the clinical course of hypertrophic cardiomyopathy. Circulation. 2001;104:2517-24.

- [3] Tsuda T, Hayashi K, Fujino N, Konno T, Tada H, Nomura A, et al. Effect of hypertrophic cardiomyopathy on the prediction of thromboembolism in patients with nonvalvular atrial fibrillation. *Heart Rhythm*. 2019;16:829-37.
- [4] Maron BJ, Olivetto I, Bellone P, Conte MR, Cecchi F, Flygenring BP, et al. Clinical profile of stroke in 900 patients with hypertrophic cardiomyopathy. *J Am Coll Cardiol*. 2002;39:301-7.
- [5] Maron BJ, Haas TS, Maron MS, Lesser JR, Browning JA, Chan RH, et al. Left atrial remodeling in hypertrophic cardiomyopathy and susceptibility markers for atrial fibrillation identified by cardiovascular magnetic resonance. *Am J Cardiol*. 2014;113:1394-400.
- [6] Cochet H, Morlon L, Verge MP, Salel M, Camaioni C, Reynaud A, et al. Predictors of future onset of atrial fibrillation in hypertrophic cardiomyopathy. *Arch Cardiovasc Dis*. 2018;111:591-600.
- [7] Tuluze K, Yakar Tuluze S, Kahya Eren N, Kocabas U, Akyildiz Akcay F, Gunduz R, et al. Predictors of Future Atrial Fibrillation Development in Patients with Hypertrophic Cardiomyopathy: A Prospective Follow-Up Study. *Echocardiography*. 2016;33:379-85.
- [8] Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, Sr., et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet*. 2009;373:739-45.
- [9] Alonso A, Krijthe BP, Aspelund T, Stepas KA, Pencina MJ, Moser CB, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *J Am Heart Assoc*. 2013;2:e000102.
- [10] Chamberlain AM, Agarwal SK, Folsom AR, Soliman EZ, Chambless LE, Crow R, et al. A clinical risk score for atrial fibrillation in a biracial prospective cohort (from the Atherosclerosis Risk in Communities [ARIC] study). *Am J Cardiol*. 2011;107:85-91.
- [11] Bundy JD, Heckbert SR, Chen LY, Lloyd-Jones DM, Greenland P. Evaluation of Risk Prediction Models of Atrial Fibrillation (from the Multi-Ethnic Study of Atherosclerosis [MESA]). *Am J Cardiol*. 2019.
- [12] Harris TB, Launer LJ, Eiriksdottir G, Kjartansson O, Jonsson PV, Sigurdsson G, et al. Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. *Am J Epidemiol*. 2007;165:1076-87.
- [13] Ikram MA, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, et al. The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol*. 2017;32:807-50.
- [14] Bhattacharya M, Lu DY, Kudchadkar SM, Greenland GV, Lingamaneni P, Corona-Villalobos CP, et al. Identifying Ventricular Arrhythmias and Their Predictors by Applying Machine Learning Methods to Electronic Health Records in Patients With Hypertrophic Cardiomyopathy (HCM-VAr-Risk Model). *Am J Cardiol*. 2019;123:1681-9.
- [15] Lu DY, Ventoulis I, Liu H, Kudchadkar SM, Greenland GV, Yalcin H, et al. Sex-specific cardiac phenotype and clinical outcomes in patients with hypertrophic cardiomyopathy. *Am Heart J*. 2019;219:58-69.
- [16] Gersh BJ, Maron BJ, Bonow RO, Dearani JA, Fifer MA, Link MS, et al. 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2011;124:e783-831.
- [17] Sankaranarayanan R, E JF, C JG. Mimics of Hypertrophic Cardiomyopathy - Diagnostic Clues to Aid Early Identification of Phenocopies. *Arrhythm Electrophysiol Rev*. 2013;2:36-40.
- [18] January CT, Wann LS, Alpert JS, Calkins H, Cigarroa JE, Cleveland JC, Jr., et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: executive summary: a

report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. *Circulation*. 2014;130:2071-104.

[19] January CT, Wann LS, Alpert JS, Calkins H, Cigarroa JE, Cleveland JC, Jr., et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. *Circulation*. 2014;130:e199-267.

[20] Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *J Am Soc Echocardiogr*. 2015;28:1-39 e14.

[21] Liu H, Pozios I, Haileselassie B, Nowbar A, Sorensen LL, Phillip S, et al. Role of Global Longitudinal Strain in Predicting Outcomes in Hypertrophic Cardiomyopathy. *Am J Cardiol*. 2017;120:670-5.

[22] Bravo PE, Zimmerman SL, Luo HC, Pozios I, Rajaram M, Pinheiro A, et al. Relationship of delayed enhancement by magnetic resonance to myocardial perfusion by positron emission tomography in hypertrophic cardiomyopathy. *Circ Cardiovasc Imaging*. 2013;6:210-7.

[23] Yalcin H, Valenta I, Yalcin F, Corona-Villalobos C, Vasquez N, Ra J, et al. Effect of Diffuse Subendocardial Hypoperfusion on Left Ventricular Cavity Size by (13)N-Ammonia Perfusion PET in Patients With Hypertrophic Cardiomyopathy. *Am J Cardiol*. 2016;118:1908-15.

[24] Murphy KP. Machine learning : a probabilistic perspective. Cambridge, Mass.: MIT Press; 2012.

[25] Sokal RR, Rohlf FJ. Introduction to biostatistics. 2nd ed. New York: Freeman; 1987.

[26] Welch BL. The Significance of the Difference Between Two Means when the Population Variances are Unequal. *Biometrika*. 1938;29:350-62.

[27] Maron BJ, Wolfson JK, Epstein SE, Roberts WC. Intramural ("small vessel") coronary artery disease in hypertrophic cardiomyopathy. *J Am Coll Cardiol*. 1986;8:545-57.

[28] Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*. 1979;44:443-60.

[29] Fernandez-Ruiz I. Artificial intelligence to improve the diagnosis of cardiovascular diseases. *Nat Rev Cardiol*. 2019;16:133.

[30] Zou L, Yu S, Meng T, Zhang Z, Liang X, Xie Y. A Technical Review of Convolutional Neural Network-Based Mammographic Breast Cancer Diagnosis. *Comput Math Methods Med*. 2019;2019:6509357.

[31] Daoud M, Mayo M. A survey of neural network-based cancer prediction models from microarray data. *Artif Intell Med*. 2019;97:204-14.

[32] Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform*. 2014;8:14.

[33] Tan AC, Gilbert D, Deville Y. Multi-class protein fold classification using a new ensemble machine learning approach. *Genome Inform*. 2003;14:206-17.

[34] Bhattacharya M, Jurkowitz C, Shatkay H. Assessing chronic kidney disease from office visit records using hierarchical meta-classification of an imbalanced dataset. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)2017. p. 663-70.

[35] Debonnaire P, Joyce E, Hiemstra Y, Mertens BJ, Atsma DE, Schalij MJ, et al. Left Atrial Size and Function in Hypertrophic Cardiomyopathy Patients and Risk of New-Onset Atrial Fibrillation. *Circ Arrhythm Electrophysiol*. 2017;10.

- [36] Guttman OP, Rahman MS, O'Mahony C, Anastasakis A, Elliott PM. Atrial fibrillation and thromboembolism in patients with hypertrophic cardiomyopathy: systematic review. *Heart*. 2014;100:465-72.
- [37] Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J*. 2019;40:1975-86.
- [38] Vallee A, Cinaud A, Blachier V, Lelong H, Safar ME, Blacher J. Coronary heart disease diagnosis by artificial neural networks including aortic pulse wave velocity index and clinical parameters. *J Hypertens*. 2019;37:1682-8.
- [39] Tsang TS, Gersh BJ, Appleton CP, Tajik AJ, Barnes ME, Bailey KR, et al. Left ventricular diastolic dysfunction as a predictor of the first diagnosed nonvalvular atrial fibrillation in 840 elderly men and women. *J Am Coll Cardiol*. 2002;40:1636-44.
- [40] Patton KK, Ellinor PT, Heckbert SR, Christenson RH, DeFilippi C, Gottdiener JS, et al. N-terminal pro-B-type natriuretic peptide is a major predictor of the development of atrial fibrillation: the Cardiovascular Health Study. *Circulation*. 2009;120:1768-74.
- [41] Schotten U, Neuberger HR, Allessie MA. The role of atrial dilatation in the domestication of atrial fibrillation. *Prog Biophys Mol Biol*. 2003;82:151-62.
- [42] Tsang TS, Barnes ME, Gersh BJ, Bailey KR, Seward JB. Risks for atrial fibrillation and congestive heart failure in patients  $\geq 65$  years of age with abnormal left ventricular diastolic relaxation. *Am J Cardiol*. 2004;93:54-8.
- [43] Rosenberg MA, Gottdiener JS, Heckbert SR, Mukamal KJ. Echocardiographic diastolic parameters and risk of atrial fibrillation: the Cardiovascular Health Study. *Eur Heart J*. 2012;33:904-12.
- [44] Tani T, Tanabe K, Ono M, Yamaguchi K, Okada M, Sumida T, et al. Left atrial volume and the risk of paroxysmal atrial fibrillation in patients with hypertrophic cardiomyopathy. *J Am Soc Echocardiogr*. 2004;17:644-8.
- [45] Spirito P, Autore C, Formisano F, Assenza GE, Biagini E, Haas TS, et al. Risk of sudden death and outcome in patients with hypertrophic cardiomyopathy with benign presentation and without risk factors. *Am J Cardiol*. 2014;113:1550-5.
- [46] Yang WI, Shim CY, Kim YJ, Kim SA, Rhee SJ, Choi EY, et al. Left atrial volume index: a predictor of adverse outcome in patients with hypertrophic cardiomyopathy. *J Am Soc Echocardiogr*. 2009;22:1338-43.
- [47] Tian T, Wang Y, Sun K, Wang J, Zou Y, Zhang W, et al. Clinical profile and prognostic significance of atrial fibrillation in hypertrophic cardiomyopathy. *Cardiology*. 2013;126:258-64.
- [48] Losi MA, Betocchi S, Aversa M, Lombardi R, Miranda M, D'Alessandro G, et al. Determinants of atrial fibrillation development in patients with hypertrophic cardiomyopathy. *Am J Cardiol*. 2004;94:895-900.
- [49] Kotecha D, Piccini JP. Atrial fibrillation in heart failure: what should we do? *Eur Heart J*. 2015;36:3250-7.
- [50] Nagueh SF, Lakkis NM, Middleton KJ, Spencer WH, 3rd, Zoghbi WA, Quinones MA. Doppler estimation of left ventricular filling pressures in patients with hypertrophic cardiomyopathy. *Circulation*. 1999;99:254-61.
- [51] Rosenberg MA, Manning WJ. Diastolic dysfunction and risk of atrial fibrillation: a mechanistic appraisal. *Circulation*. 2012;126:2353-62.
- [52] Jons C, Joergensen RM, Hassager C, Gang UJ, Diken U, Johannesen A, et al. Diastolic dysfunction predicts new-onset atrial fibrillation and cardiovascular events in patients with acute myocardial infarction and depressed left ventricular systolic function: a CARISMA substudy. *Eur J Echocardiogr*. 2010;11:602-7.

- [53] Pujadas S, Vidal-Perez R, Hidalgo A, Leta R, Carreras F, Barros A, et al. Correlation between myocardial fibrosis and the occurrence of atrial fibrillation in hypertrophic cardiomyopathy: a cardiac magnetic resonance imaging study. *Eur J Radiol.* 2010;75:e88-91.
- [54] Yamaji K, Fujimoto S, Yutani C, Ikeda Y, Mizuno R, Hashimoto T, et al. Does the progression of myocardial fibrosis lead to atrial fibrillation in patients with hypertrophic cardiomyopathy? *Cardiovasc Pathol.* 2001;10:297-303.
- [55] Papavassiliu T, Germans T, Fluchter S, Doesch C, Suriyakamar A, Haghi D, et al. CMR findings in patients with hypertrophic cardiomyopathy and atrial fibrillation. *J Cardiovasc Magn Reson.* 2009;11:34.
- [56] Sivalokanathan S, Zghaib T, Greenland GV, Vasquez N, Kudchadkar SM, Kontari E, et al. Hypertrophic Cardiomyopathy Patients With Paroxysmal Atrial Fibrillation Have a High Burden of Left Atrial Fibrosis by Cardiac Magnetic Resonance Imaging. *JACC Clin Electrophysiol.* 2019;5:364-75.
- [57] Azarbal F, Singh M, Finocchiaro G, Le VV, Schnittger I, Wang P, et al. Exercise capacity and paroxysmal atrial fibrillation in patients with hypertrophic cardiomyopathy. *Heart.* 2014;100:624-30.
- [58] Vasquez N, Ostrander BT, Lu DY, Ventoulis I, Haileselassie B, Goyal S, et al. Low Left Atrial Strain Is Associated With Adverse Outcomes in Hypertrophic Cardiomyopathy Patients. *J Am Soc Echocardiogr.* 2019;32:593-603 e1.
- [59] Dickinson MG, Lam CS, Rienstra M, Vonck TE, Hummel YM, Voors AA, et al. Atrial fibrillation modifies the association between pulmonary artery wedge pressure and left ventricular end-diastolic pressure. *Eur J Heart Fail.* 2017;19:1483-90.
- [60] Luo HC, Dimaano VL, Kembro JM, Hilser A, Hurtado-de-Mendoza D, Pozios I, et al. Exercise heart rates in patients with hypertrophic cardiomyopathy. *Am J Cardiol.* 2015;115:1144-50.
- [61] Maron BJ, Rowin EJ, Casey SA, Haas TS, Chan RH, Udelson JE, et al. Risk stratification and outcome of patients with hypertrophic cardiomyopathy  $\geq 60$  years of age. *Circulation.* 2013;127:585-93.
- [62] Bravo PE, Pinheiro A, Higuchi T, Rischpler C, Merrill J, Santaularia-Tomas M, et al. PET/CT assessment of symptomatic individuals with obstructive and nonobstructive hypertrophic cardiomyopathy. *J Nucl Med.* 2012;53:407-14.
- [63] Hurtado-de-Mendoza D, Corona-Villalobos CP, Pozios I, Gonzales J, Soleimanifard Y, Sivalokanathan S, et al. Diffuse interstitial fibrosis assessed by cardiac magnetic resonance is associated with dispersion of ventricular repolarization in patients with hypertrophic cardiomyopathy. *J Arrhythm.* 2017;33:201-7.
- [64] Kobayashi T, Popovic Z, Bhonsale A, Smedira NG, Tan C, Rodriguez ER, et al. Association between septal strain rate and histopathology in symptomatic hypertrophic cardiomyopathy patients undergoing septal myectomy. *Am Heart J.* 2013;166:503-11.
- [65] Spirito P, Watson RM, Maron BJ. Relation between extent of left ventricular hypertrophy and occurrence of ventricular tachycardia in hypertrophic cardiomyopathy. *Am J Cardiol.* 1987;60:1137-42.
- [66] D'Amato R, Tomberli B, Castelli G, Spoladore R, Girolami F, Fornaro A, et al. Prognostic value of N-terminal pro-brain natriuretic Peptide in outpatients with hypertrophic cardiomyopathy. *Am J Cardiol.* 2013;112:1190-6.
- [67] Gruver EJ, Fatkin D, Dodds GA, Kisslo J, Maron BJ, Seidman JG, et al. Familial hypertrophic cardiomyopathy and atrial fibrillation caused by Arg663His beta-cardiac myosin heavy chain mutation. *Am J Cardiol.* 1999;83:13H-8H.
- [68] Ogimoto A, Hamada M, Nakura J, Miki T, Hiwada K. Relation between angiotensin-converting enzyme II genotype and atrial fibrillation in Japanese patients with hypertrophic cardiomyopathy. *J Hum Genet.* 2002;47:184-9.

[69] Bongini C, Ferrantini C, Girolami F, Coppini R, Arretini A, Targetti M, et al. Impact of Genotype on the Occurrence of Atrial Fibrillation in Patients With Hypertrophic Cardiomyopathy. *Am J Cardiol.* 2016;117:1151-9.

## FIGURES AND FIGURE LEGENDS

**Figure 1. Age distribution of HCM patients with AF in HCM cohort .** AF prevalence increases with age in HCM.

**Figure 2. Distribution of LA size in HCM patients with/without AF .** Significant overlap exists between LA diameter values in the AF and No-AF groups of HCM patients, but mean values for LA diameter are significantly higher in the AF group, when compared to the No-AF group ( $p < 0.001$ ). Each dot represents a patient; mean  $\pm$  1.96 SD is presented.

**Figure 3. Receiver operating characteristic (ROC) curve for HCM-AF-Risk Model :** The ROC curve depicting the performance of the *HCM-AF-Risk Model* that combines under- and over-sampling approach. False-positive rate is shown on the x-axis and the true-positive rate is indicated on the y-axis.

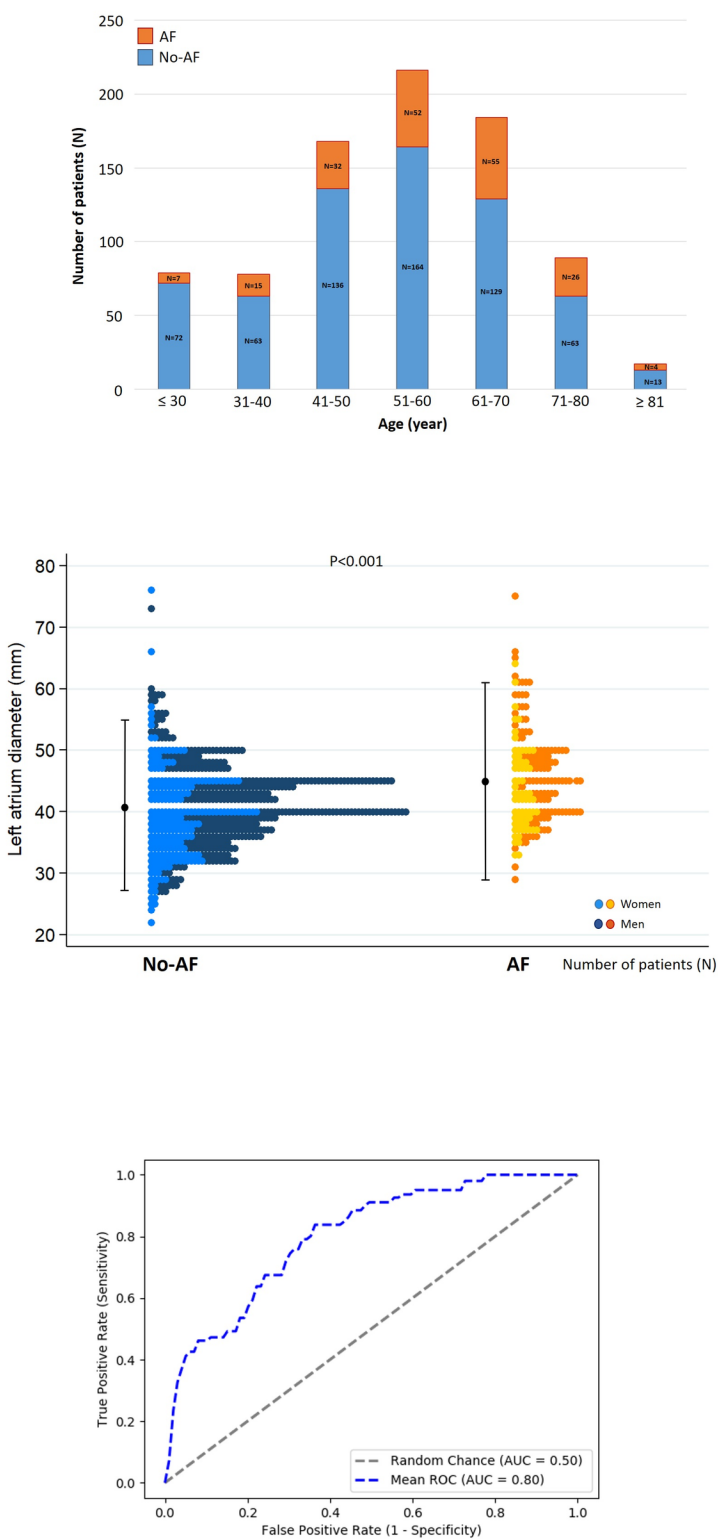
**Supplementary Figure 1. Flow chart illustrating identification of AF and No-AF cases in HCM cohort**

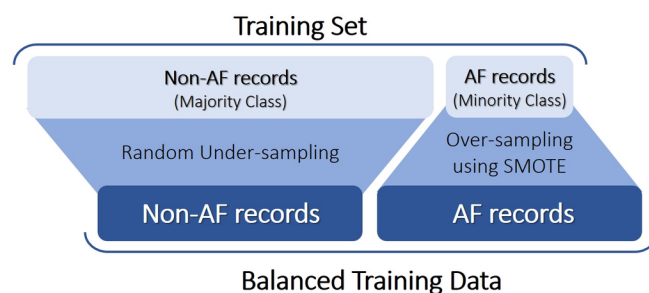
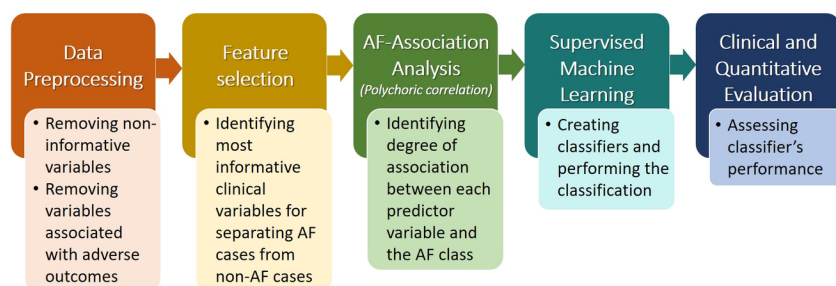
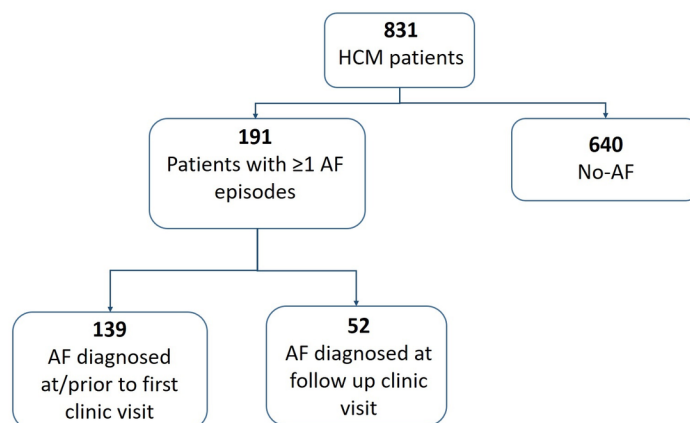
**Supplementary Figure 2. HCM-AF-Risk Model Schematic :** The overall framework for identification of atrial fibrillation cases using clinical attributes within electronic health records of HCM patients (*HCM-AF-Risk Model* ). In the data preprocessing step, variables known to be non-informative with respect to AF, and variables associated with adverse outcomes (e.g. heart failure, ventricular arrhythmia, stroke) are removed. The feature selection step identifies the most informative clinical variables for separating AF cases from No-AF cases. Next, the degree of association between each predictor variable and the AF class is identified via association analysis. Supervised machine learning is then used to build classifiers and perform classification. Last, a thorough evaluation, both qualitative and quantitative was performed to assess the classifier's performance.

**Supplementary Figure 3. Methods for addressing data imbalance:** An illustration of our classification scheme for combining over- and under-sampling. The topmost layer represents the entire training set, which comprises a majority of No-AF records (shown on the left) and the minority of AF records (shown on the right). The majority class in the training set (No-AF) is randomly under-sampled such that the No-AF to AF record ratio is 2:1. The minority class (AF) is over-sampled using *SMOTE* to generate synthetic new AF-like records, doubling the number of AF records. The resulting set forms a balanced training set, containing the same number of AF and No-AF records.

## Hosted file

Tables\_August\_5\_Final.docx available at <https://authorea.com/users/349335/articles/474387-hcm-af-risk-model-to-identify-cases-and-predictors-of-atrial-fibrillation-in-hypertrophic-cardiomyopathy>





## Hosted file

Supplemental\_tables\_August\_5.docx available at <https://authorea.com/users/349335/articles/474387-hcm-af-risk-model-to-identify-cases-and-predictors-of-atrial-fibrillation-in-hypertrophic-cardiomyopathy>