

Is machine learning the future for atrial fibrillation screening?



Pavidra Sivanandarajah, BSc, MBBS, MRCP,^{*†} Huiyi Wu, PhD,^{*} Nikesh Bajaj, PhD,^{*} Sadia Khan, BM, BCh, MD, FRCP,[†] Fu Siong Ng, BSc, MBBS, PhD, FRCP, FHRS^{*†}

From the ^{*}National Heart and Lung Institute, Imperial College London, London, United Kingdom, and [†]Chelsea and Westminster NHS Foundation Trust, London, United Kingdom.

Atrial fibrillation (AF) is the most common arrhythmia and causes significant morbidity and mortality. Early identification of AF may lead to early treatment of AF and may thus prevent AF-related strokes and complications. However, there is no current formal, cost-effective strategy for population screening for AF. In this review, we give a brief overview of targeted screening for AF, AF risk score models used for screening and describe the different screening tools. We then go on to extensively discuss the potential applications of machine learning in AF screening.

KEYWORDS Atrial fibrillation; Machine learning; Screening; Artificial intelligence; Electronic health records

(Cardiovascular Digital Health Journal 2022;3:136–145) © 2022 Heart Rhythm Society. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Introduction

Atrial fibrillation (AF) is the most common arrhythmia and causes significant morbidity and mortality, including strokes, heart failure, cognitive decline, depression, impaired quality of life, and hospitalizations.¹ AF can be difficult to identify, as it can be paroxysmal and patients may be asymptomatic. Individuals with AF have a 5-fold increased risk of developing strokes, which can be reduced significantly (by up to two-thirds) with appropriate anticoagulation.² However, currently, only opportunistic screening for AF is recommended by the European Society of Cardiology (ESC)¹ and UK National Institute for Health and Care Excellence (NICE)³ and relies on patients presenting to general practices themselves, while in the United States, the US Preventive Services Task Force does not recommend AF screening.⁴ In this review, we will discuss how AF screening has previously been carried out and the current application of machine learning (ML) in AF screening.

Recent studies on clinical outcomes of AF screening

Two recent studies reported the long-term clinical outcomes of AF screening. The STROKESTOP⁵ study performed in Sweden recruited 75- to 76-year-old participants between 2012 and 2014. They were randomized to a screening group or a control group (usual care). Individuals in the screening group were screened for AF over a 2-week period twice daily,

using a single-lead electrocardiogram (ECG) recorder, Zenicor. There was a small net benefit of screening with regard to the primary endpoint, which was a composite of ischemic or hemorrhagic stroke, systemic embolism, bleeding requiring hospitalization, and all-cause death.

The LOOP study,⁶ carried out in neighboring Denmark, also looked at long-term clinical outcomes of AF screening. In this study, participants aged 70–90 years with at least 1 additional stroke risk factor were recruited. They were randomized in a 1:3 ratio to an implantable loop recorder monitoring group or control group (usual care). If a minimum of 6 minutes of AF was detected, anticoagulation was recommended. However, in this study, there was no reduction in stroke or systemic arterial embolism with screening.

The STROKESTOP study showed benefit for systematic AF screening, while the LOOP study did not. It has been postulated that a reason for this was that the background AF detection rate in the control group in LOOP was significantly higher than would be expected, thus reducing the benefit of AF screening. The duration of 6 minutes of AF requiring anticoagulation in LOOP may have also been too short a duration and perhaps anticoagulation should only be considered for longer periods of AF, although the minimum duration of AF that should require anticoagulation is still unknown.

Targeted AF screening groups

Although there may be some clinical benefit in screening for AF, it is not cost-effective to conduct systematic screening, as only 1 case is identified after screening 145 individuals.⁷ Different screening studies have chosen specific groups of

Address reprint requests and correspondence: Dr Fu Siong Ng, Imperial College London, 72 Du Cane Rd, 4th Floor, ICTEM Building, London W12 0NN, UK. E-mail address: f.ng@imperial.ac.uk.

KEY FINDINGS

- There may be long-term clinical benefits of atrial fibrillation (AF) screening suggested by the recent STROKE-STOP study, but it is currently not cost-effective to conduct systematic AF screening.
- Machine learning can be applied to AF screening to help with AF risk prediction or to improve the automated diagnosis of AF using different rhythm monitoring modalities.
- Machine learning AF risk prediction can be used to predict incident, paroxysmal, and future AF. Prediction models have been developed using 12-lead electrocardiograms (ECGs), ambulatory ECGs, and electronic health records.
- Machine learning automated AF diagnosis has been achieved using a variety of rhythm modalities including 12-lead ECGs, ambulatory ECGs, and photoplethysmography.

individuals for targeted screening to make systematic screening more cost-effective. The criteria for inclusion into these studies often have similarities, such as age, a certain CHA₂DS₂-VASc score or prior history of stroke.

It is known that the incidence of AF increases with age. Many screening studies choose age 65 as the minimum age to enter their screening study for AF. However, 1 meta-analysis of single-time point AF screening studies calculated the number needed to screen (NNS) to identify 1 case of AF in a population greater than 65 years old was still as high as 69, with the number rising to 83 to identify 1 treatable new AF case. This number increased the lower the age of the population chosen to screen.⁸

Another approach that has been used is to screen for AF based on the CHA₂DS₂-VASc score, which will be discussed in the next section, while others have used a prior history of stroke as an inclusion criterion for AF screening, as AF increases one's risk of a stroke by 5-fold. However, despite targeting screening to the above groups, a cost-effective screening program with sufficiently low NNS has not been developed. Therefore, more work needs to be done to identify appropriate individuals for systematic targeted screening.

Risk score models to target AF screening

Over the last 15 years, several clinical risk score models have been developed to predict the likelihood of the development of AF. They can provide a 5-year or 10-year prediction risk of AF, and they are listed in [Table 1](#). Most of these were developed from data from particular population cohorts. The earlier risk scores were developed in the United States (US) from population cohorts, including FRAMINGHAM,⁹ ARIC,¹⁰ and CHARGE-AF,¹¹ which have C-statistic values of 0.78, 0.78, and 0.765, respectively.

Following the development of a clinical risk score model from a particular population cohort, the model is usually evaluated on a different population cohort where the data are readily available. However, it has been recurrently found that the performance of a model is not as accurate as when performed on the original population cohort. The C-statistic value is often greater than 0.75 in the original population cohort and this drops below 0.75 when external validation of the model is performed. AGES, the Rotterdam study, and the EPIC Norfolk cohort are examples of population cohorts used for external validation. Recalibration or modifications of the model are required to maintain adequate performance.^{12,13} There are many differences in clinical variables between the cohorts that could affect the performance of the model, including median age, proportions of different ethnicities, and prevailing cardiovascular risk factors within the population cohort.

The CHA₂DS₂-VASc risk score is a score designed for stroke risk prediction in individuals with AF. Although CHA₂DS₂-VASc risk score is not formally validated for AF risk prediction, it has been used to identify individuals to screen for AF. It has been tested on a retrospective Israeli database to give an area under the receiver operating characteristic curve (AUROC) of 0.744 for prediction of new-onset AF.¹⁴ There are also a few clinical risk score models developed from Asian population cohorts, which include a Japanese simple risk score,¹⁵ Suita score,¹⁶ C2HEST,¹⁷ and the Taiwan AF score,¹⁸ which were developed more recently. There was a need for these, as previous risk scores have been developed from populations where Asian groups were underrepresented.

As listed in [Table 1](#), there are now many clinical risk score models that have been developed for AF risk prediction. However, they are not applicable worldwide, as risk factors for AF will vary from population to population. One study compared 9 clinical risk score models on the ARIC cohort and found that only 5 risk scores were superior to using age alone. The best risk model, CHARGE-AF, was only able to identify 82% of the AF patients in the ARIC group.¹⁹ Many risk score models have attempted to incorporate examination findings, blood tests, ECG data, and echo data without improving the performance of the risk score model significantly. It is also difficult to obtain these types of data for individuals, as they are often missing and therefore affect the performance of the risk score model. These factors have limited the ability of these clinical risk score models to be used effectively for systematic AF screening.

Screening tools to diagnose AF

There has been a recent growth in the number of technologies available for heart rhythm monitoring, which provides more opportunities for ambulatory monitoring, therefore increasing the likelihood of detecting AF. There are currently over 400 wearable devices available for monitoring and detection of AF.¹ Clinically validated mobile health technologies include handheld or smartphone single-lead ECG

Table 1 Risk score models for atrial fibrillation detection

Risk score model	Factors	Internal validation	External validation	Prediction
FRAMINGHAM ⁹	Age, blood pressure, hypertension, body mass index, PR interval, murmur, age of heart failure	C-statistic 0.78	C-statistic 0.65–0.73 Montefiore Medical Center (Bronx County): 0.734/0.724 ¹² ARIC 0.68 ¹⁰ AGES 0.652 ¹¹ Rotterdam 0.686 ¹¹ CHARGE-AF 0.734 ¹¹	10 -year risk of AF
ARIC ¹⁰	Age, race, height, systolic blood pressure, hypertension medication, smoking, murmur, left atrial enlargement, LVH, diabetes mellitus, heart failure, age of coronary heart disease	C-statistic 0.78	-	10-year risk of AF
CHARGE-AF ¹¹	Age, race, height, weight, systolic blood pressure, diastolic blood pressure, murmur, hypertension medication, diabetes mellitus, heart failure, myocardial infarction, LVH on ECG, PR interval	C-statistic 0.765	C-statistic 0.66–0.81 Montefiore Medical Center (Bronx County): 0.707/0.691 ¹² ARIC: 0.775 FHS: 0.75 EPIC Norfolk: 0.81/0.8 (2-fold overestimation of AF) ¹³ MESA: 0.779 ²⁷ Electronic health record only: 0.747/0.753 ³⁸ FHS: 0.71	5-year risk of AF
CHA ₂ DS ₂ -VASc	Congestive heart failure, hypertension, age, diabetes mellitus, vascular disease, stroke, sex	-	ARIC: 0.66 Israeli database: 0.7440 ¹⁴ Chinese cohort: 0.687 ¹⁷ Korean cohort: 0.637 ¹⁷ Electronic health record only: 0.701/0.702 ³⁸ Taiwan cohort: 0.7711 ⁴⁰	Stroke risk
HATCH ³⁹	Heart failure, age, previous transient ischemic attack, chronic pulmonary obstructive disease, hypertension	-	-	Progression of paroxysmal AF to persistent AF
Suita ¹⁶	Age, systolic hypertension, weight, alcohol intake, smoking, other arrhythmia, coronary artery disease, cardiac murmur	C-statistic 0.749	-	10-year risk of AF
Japanese simple risk score ¹⁵	Age, sex, waist circumference, diastolic blood pressure, alcohol intake	C-statistic 0.77 Added model C-statistic 0.79	-	7-year risk of AF
C2HEST ¹⁷	Added model: ECG variable Coronary artery disease, chronic obstructive pulmonary disease, hypertension, elderly, systolic heart failure, thyroid disease	C-statistic 0.75	Korean cohort: 0.65 Taiwan cohort: 0.7895 ⁴⁰ Electronic health record only: 0.735/0.754 ³⁸	AF risk
Taiwan AF score ¹⁸	Age, sex, hypertension, heart failure, coronary artery disease, end-stage renal disease, alcoholism	AUROC 0.862 at 1 year, 0.755 at 16 years	-	1- to 16-year risk of AF

Table 1 (Continued)

Risk score model	Factors	Internal validation	External validation	Prediction
SAAFE ¹⁹	Age, height, weight, congestive heart failure, coronary artery disease, chronic obstructive pulmonary disease, cardiac arrest, coronary artery stenting, stroke, diabetes, kidney transplant	C-statistic 0.785/0.804	ARIC 0.766	Incident AF risk
Electronic health record risk score ³⁸	Sex, age, race, smoking, height, weight, blood pressure, cardiovascular and cardiometabolic factors	C-statistic 0.77	C-index 0.808 ⁴¹	5-year risk of AF

AF = atrial fibrillation; AUROC = area under the receiver operating characteristic curve; ECG = electrocardiogram; LVH = left ventricular hypertrophy.

recorders, smartwatch single-lead ECG recorders, photoplethysmography (PPG) smartwatches or smartphones, ECG patch monitors, blood pressure monitors, and external Holter monitors. Several studies have been performed for each modality showing that they are feasible to use for AF detection. The summaries of their results are listed in Table 2. A meta-analysis suggests that the screening modality does not influence the yield of screening²⁰; therefore, with the appropriate targeted screening group, the use of any of these mobile health technologies could be very promising.

Machine learning

ML is one of the approaches of artificial intelligence, which itself is a broader concept of automation and prediction. Artificial intelligence has been proven beneficial in a wide range of applications, from trivial tasks for humans such as recognizing faces in an image or video to understanding large data to make decisions. There are various models for ML, which are shown in Table 3.

ML algorithms can be applied to a wide range of health-care data, which include and are not limited to mobile health technology data, electronic health records, biomedical imaging data, implanted electronic device data, biochemical data, and genomic data. This makes ML invaluable for automated detection of pathology or conditions, for risk prediction of future conditions, and for prediction of those who would benefit from specific treatment strategies. Some examples of the uses of ML for AF include prediction of recurrence of AF following AF ablation,²¹ improving cardiac activation mapping for AF,²² prediction of successful electrical cardioversion of AF,²³ and AF screening. Thus far, supervised and unsupervised ML have been used for AF screening. ML can be applied to AF screening to help with 2 different tasks: (1) to identify the individuals with the highest risk for AF to target the screening to minimize the number needed to screen, and (2) to improve the automated diagnosis of AF using data from different rhythm monitoring modalities. We discuss both these applications below. Figure 1 shows a summary of ML in AF screening.

Machine learning for AF risk prediction to target AF screening

Several studies have looked at the use of ML to predict future AF using 12-lead ECGs, electronic health records, or sinus rhythm ambulatory ECGs. These approaches can be used to help target any screening efforts to maximize the efficiency of screening and the number of new AF cases detected.

Machine learning for AF risk prediction using 12-lead ECGs

ML has been used to predict future or paroxysmal AF from sinus rhythm ECGs in 3 studies. The first one of these is from the Mayo Clinic.²⁴ This study used 649,931 normal sinus rhythm ECGs from 180,922 patients collected over a 24-year period. Convolutional neural networks were used to create an algorithm that predicted from a sinus rhythm ECG the likelihood of paroxysmal AF. The AUROC was 0.87 and the sensitivity was 0.79. The algorithm was further validated on the Mayo Clinic Olmsted Study of Aging cohort.

Another study²⁵ looked at predicting current AF and future AF from a sinus rhythm ECG with 2 separate models. The study used 12,863 sinus rhythm ECGs without a history of structural heart disease from the Shinken Database 2010–2017. The predictive models were based on a random forest algorithm. This study also explored ECG parameters that could affect predictive capability using logistic regression and Spearman coefficient of correlation. It identified that the P wave, QRS, and ST-T segment were similar in their parameter importance to the model. The C-statistic for prediction of current paroxysmal AF was 0.913 and the C-statistic for prediction of future AF was 0.991.

The third study that predicts the future risk of AF from the ECG is from Geisinger.²⁶ This study used 1.6 million resting 12-lead ECG traces from 430,000 patients collected over a 35-year period for training, and then using new-onset AF at 1 year for testing. This model achieved sensitivity of 0.69, specificity of 0.81, and AUROC of 0.85. In addition, this model has been tested on patients who experienced an AF-related stroke within 3 years of the indexed ECG, and it

Table 2 Devices to detect and screen for atrial fibrillation

Rhythm modality	Reference	Results
Handheld/smartphone single-lead ECG recorder	William et al 2018, ⁴²	Algorithm sensitivity 0.966
	Maputo et al 2020 ⁴³	Algorithm specificity 0.94
		Physician review sensitivity 1 Physician review specificity 0.987
Handheld/smartphone single-lead ECG recorder	Vaes et al 2014 ⁴⁴	Sensitivity 0.94 (0.87–0.98) Specificity 0.93 (0.85–0.97)
Smartwatch single-lead ECG recorder	Wasserlauf et al 2019 ⁴⁵	Sensitivity 0.975 when compared to implantable cardiac monitor
Smartphone PPG	Brasier et al 2019 ⁴⁶	Sensitivity 0.915 Specificity 0.996
Smartphone PPG	Proesmans et al 2019 ⁴⁷	Sensitivity 0.96 Specificity 0.97
Smartphone PPG	Poh et al 2018 ⁴⁸	Sensitivity 1 Specificity 0.996
Smartphone PPG (iPhone 4S)	McManus et al 2013 ⁴⁹	Sensitivity 0.962 Specificity 0.975
Smartphone PPG (iPhone 4S)	Lee et al 2013 ⁵⁰	Sensitivity 0.7461–0.9763 Specificity 0.9961–1
Smartwatch PPG	Bashar et al 2019 ⁵¹	Sensitivity 0.982 Specificity 0.974
Smartwatch PPG	Dörr et al 2019 ⁵²	Sensitivity 0.937 (0.898–0.964) Specificity 0.982 (0.958–0.994)
Smartwatch PPG ⁵¹	Bumgarner et al 2018 ⁵³	Sensitivity 0.93 (0.86–0.99) Specificity 0.84 (0.73–0.95)
ECG patch monitor	Okubo et al 2021 ⁵⁴	Sensitivity 0.98 Specificity 0.982
Blood pressure monitor	Wiesel et al 2009 ⁵⁵	Sensitivity 0.95 Specificity 0.86

ECG = electrocardiogram; PPG = photoplethysmography.

predicted new-onset AF with accuracy of 62% in this subgroup of patients. This suggests that the model could also help predict future AF-related strokes. However, all these models have yet to be used prospectively.

Machine learning for AF risk prediction using ambulatory ECGs

ML has been applied to sinus rhythm segments of ambulatory ECGs to predict the presence of paroxysmal AF. [Supplemental Table 1](#) summarizes these studies. The preceding sinus rhythm segments can be used to predict the probability of an incoming episode of paroxysmal AF. There are 2 approaches for this, which are using heart rate variability features or premature atrial complexes (PACs) features. The initial approach was using PACs, as it is found that there is an increase in PACs preceding a paroxysmal AF episode. However, recent approaches have involved using heart rate

variability features, which include time domains and frequency domains. The accuracy of models developed in the last year for predicting paroxysmal AF have been above 0.9.

Machine learning for AF risk prediction using electronic health records

It is not possible for everyone to have an ECG or a physiological signal recording to predict the risk of AF. Another approach is to use healthcare data from electronic healthcare records to identify possible patients that may have AF. This would reduce the number of individuals needed to screen. The use of ML in combination with clinical risk scoring shows potential for reducing the NNS and improving the effectiveness of AF screening.

Random survival forests have been used in the Multi-Ethnic Study of Atherosclerosis (MESA) cohort²⁷ to identify highly predictive variables for the prediction of AF. These variables included coronary calcium score, carotid measurements, blood test results, ECG measurements, and echocardiographic measurements. The variables were combined to the CHARGE-AF enriched score (CHARGE-AF score with NT-proBNP) to create a new score, which gave a C-statistic of only 0.806. The C-statistic for the CHARGE-AF enriched score was 0.804. In this case, the use of ML was not beneficial in improving AF risk prediction.

There have been 4 further AF screening studies that have applied an ML screening tool previously using data from healthcare databases. They all compared different ML models to find the most optimum model for AF screening and are summarized in [Supplemental Table 2](#).

A team at the University of Colorado developed their own ML model for 6-month risk prediction of AF using electronic health records.²⁸ They compared the use of different ML models, including Naïve Bayes, logistic regression, random forest, shallow and deep neural networks, and gradient-boosted machine, and found that single-layer neural networks and random oversampling was the best model. The sensitivity was 75.2% and the specificity was 84.9%. The AUROC was 0.8. However, they found that a simple logistic regression model using known clinical risk factors performed nearly as well as their ML model. This study, however, did not take into account the time-varying effects of the variables.

Another group,²⁹ using secondary care data from Japanese databases, also developed their own ML methods for the identification of nonvalvular AF patients. The ML models used included deep learning with lasso regression. Stacking models of multiple single classifiers were also used, including lasso regression, ridge regression, support vector machine, random forest, deep learning, AdaBoost, and gradient boosting. The AUROC was greater than 0.8 for each individual model and, for stacking models, was as high as 0.9. The sensitivity varied between 0.09 and 0.84. The specificity varied between 0.79 and 0.99 for each model, while for stacking models they were 0.8 or above. The best model was a stacking model that included deep learning (neural networks) with lasso regression.

Table 3 Examples of machine learning and deep learning models

Machine learning model	Principle
<u>Traditional models</u>	
Logistic regression	Uses a linear combination of variable with logistic function for the classification
Decision trees / random forest	Uses the binary tree-based approach
Naïve Bayes	Uses probabilistic approach-based Bayes theorem and assumes that all the input variables are independent
Linear discriminant analysis	A dimensionality reduction method for classification. It looks for linear combinations of variables that best explain the data, and separates 2 or more classes of objects
Support vector machines	Uses a kernel-based hyperplane that separates the data points with maximum margin
K-nearest neighbor	Uses the distance (eg, Euclidian) of a given data point to K nearest data points using known data points (training set) to predict the class/value. It is based on assumption that data points in near proximity have similar values
Multilayer perceptron	Fully connected neural networks
<u>Deep learning model</u>	
Deep neural network	Originated from multilayer perceptron, uses different architectures and methods of weight sharing
Convolutional neural networks	ResNet U-Net Uses shared weights to exploit translation invariance in convolution approach. ResNet uses skip connections, U-Net architecture is used to label pixels of an image, widely used for image segmentation
Recurrent neural networks	GRU LSTM Transformers Uses temporal dynamic behavior of data with recursive operations. It is efficient to carry the useful information over a long time using GRU and LSTM mechanisms. Transformers also process the temporal input data using attention mechanism, which allows it to handle very large temporal data.

(Continued)

Table 3 (Continued)

Machine learning model	Principle
Generative models	Auto-encoders, GANs Generative models are based on approach to learn the distribution of given dataset, which can be used to reduce the dimension of datasets (eg, auto-encoders) or to generate new dataset (GANs)

GRU = gated recurrent units; LSTM = long short-term memory; GANs = generative adversarial networks.

The most recent study³⁰ to develop an ML model to apply to electronic healthcare records compared only different logistic regression models. The healthcare records used were derived from the Indiana Network for Patient Care. The best model, the UNAFIED model, used 10 variables and had a C-statistic of 0.8061 following validation phase. The model was implemented prospectively as part of a 6-week pilot study on the Eskenazi Health Information Systems. However, the prospective patients identified did not undergo further monitoring, screening, or intervention to confirm if they had AF.

Our group recently helped to develop and optimize another ML algorithm to identify individuals at highest risk of AF. The algorithm was trained by using a dataset from a large United Kingdom primary care database called UK Clinical Practice Research Datalink.³¹ It was created using time-varying neural networks and the published AF risk score models, Framingham, ARIC, and CHARGE-AF. Baseline ML models were compared initially, which included neural networks, lasso regression, random forest, and support vector machine. Time-varying covariates were then considered and added. The AUROC was 0.827 and the NNS was 9. We then evaluated and validated this ML algorithm on the separate DISCOVER database retrospectively.³² The DISCOVER database contains primary care data from over 2.5 million people living in northwest London. The AUROC increased to 0.87 and the NNS remained at 9. We also evaluated and validated the ML algorithm on our secondary care database retrospectively.³³ The sensitivity was 0.5 and the specificity was 0.95 with an NNS of 5. It showed an additional 5444 individuals who had AF that were not identified by algorithm based on their primary care data alone. However, as with the other ML algorithms, this has not yet been tested prospectively.

Machine learning to improve automated AF diagnosis

There are a wide range of rhythm monitoring modalities including ECGs used for AF detection. Another application of ML in AF screening is to improve the accuracy of automated AF diagnosis using different rhythm monitoring

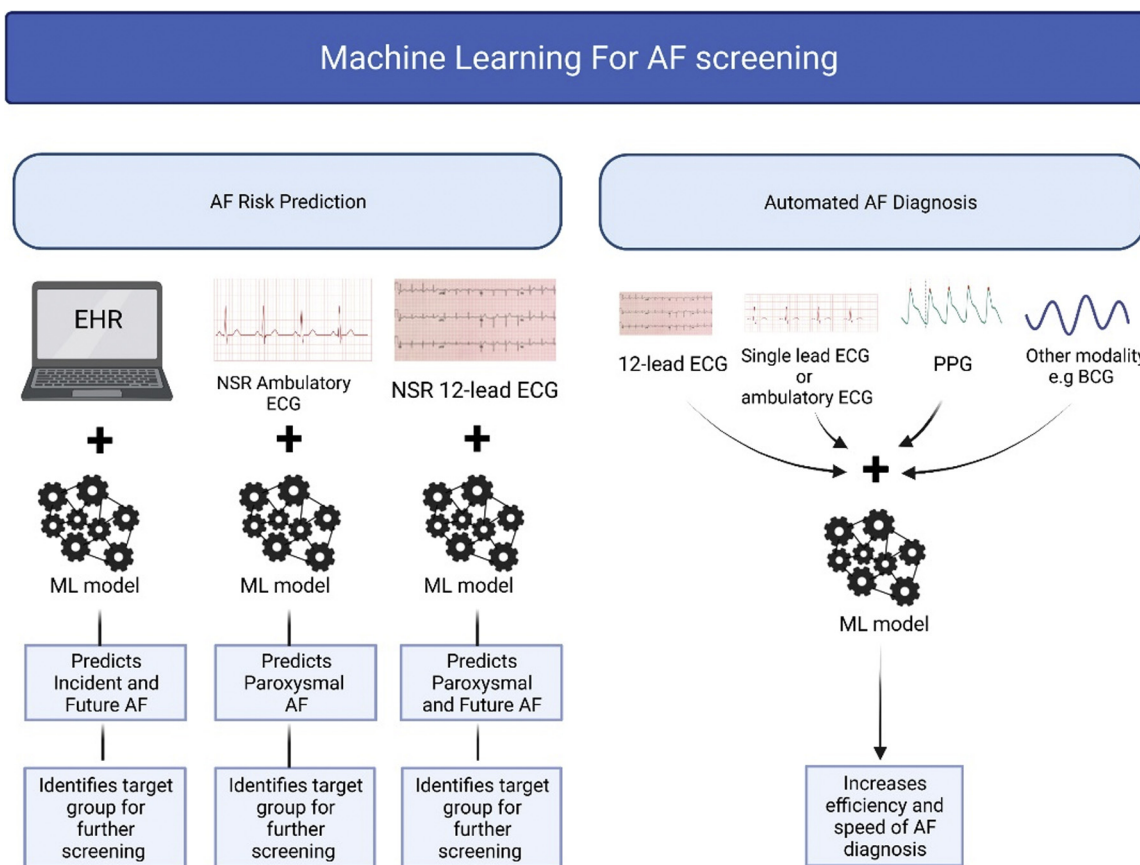


Figure 1 Summary of machine learning (ML) for atrial fibrillation (AF) screening. There are 2 main categories: AF risk prediction and automated AF diagnosis. AF risk prediction involves the application of ML to electronic health records, normal sinus rhythm ambulatory electrocardiograms (ECGs), and normal sinus rhythm 12-lead ECGs to identify the target group for further screening. For automated AF diagnosis, ML is applied to different rhythm monitoring modalities to increase efficiency and the speed of AF diagnosis. BCG = ballistocardiography; EHR = electronic health records; NSR = normal sinus rhythm; PPG = photoplethysmography.

modalities, which could have the potential to increase the numbers of individuals screened and allow quicker population or mass screening. It could mean that a physician is no longer required to interpret the result of the rhythm monitoring modality. The use of ML with different rhythm monitoring modalities is discussed below and is summarized in [Supplemental Tables 3–6](#).

Machine learning to improve automated AF diagnosis using 12-lead ECGs

[Supplemental Table 3](#) shows different studies that have developed ML models to detect AF from a 12-lead ECG. The models typically used are neural networks, as these allow unknown features of the ECG to be used in the decision process to determine if the ECG demonstrates AF. There was only 1 study that used a vector machine instead. Neural networks do not necessarily require the ECG data to have undergone feature extraction and selection.

There were high accuracy and AUC (>0.98) values for models differentiating between AF and sinus rhythm. However, the performance of the models for AF detection reduced when the models were also developed to differentiate other rhythms too.

Machine learning to improve automated AF diagnosis using photoplethysmography

There have been multiple studies that have tested ML algorithms to detect AF from PPG signals. PPG involves reading pulse pressure signals resulting from the propagation of blood pressure pulses along arterial blood vessels.³⁴ It can be measured peripherally. Detection of irregularities in pulse pressure signals aids in the diagnosis of AF. The ML studies using PPG for AF diagnosis are shown in [Supplemental Table 4](#). The ML models used were neural networks, logistic regression, random forests, and support vector machines. When the models were compared, neural networks were found to be superior. The accuracy, specificity, and sensitivity of these models were above 0.90. The features that ML models review include pulse interval features and pulse amplitude features, which can make distinguishing AF from ectopic beats difficult, although 1 study has used deep learning classifiers to overcome this. PPG is more susceptible to noise and artifact than a 12-lead ECG; however, DeepBeat (another deep learning model) has also overcome this. PPG showing AF still needs a confirmatory ECG to confirm the AF diagnosis. However, the use of ambulatory monitoring PPG could

reduce the number of individuals requiring a confirmatory ECG to diagnose AF.

Machine learning to improve automated AF diagnosis using single-lead or ambulatory ECGs

Supplemental Table 5 lists a large number of studies involving ML algorithms to identify AF from single-lead or ambulatory ECGs. PhysioNet Challenge 2017, AF prediction, MIT-BIH AF, and MIT-BIH Arrhythmia databases are readily available to the public and were used by several studies for training and testing their models. The easy accessibility of a large database of single-lead ECGs has made it possible for many ML models to be developed compared with 12-lead ECGs. A wide range of models were used. When measured, accuracy, sensitivity, and specificity tended to be above 0.85. However, when F1 scores were measured, these would vary from 0.782 to 0.92. In the last year, feature importance interpretability and real-time detection have been researched. Their aims have been to explore explainable artificial intelligence and improve AF diagnosis time.

Machine learning to improve automated AF diagnosis using other diagnostic modalities

Supplemental Table 6 shows 3 studies showing ML algorithms applied to other forms of monitoring for AF. One method is ballistocardiography, which provides waves based on ejection of blood from the heart to the great vessels. Other methods were using arterial pulse waveforms and inertia measurements using a gyroscope and accelerometer on a smartphone. Different models were used for the different methods. Again, accuracy, sensitivity, and specificity for these models tended to be above 0.9.

Limitations of studies applying machine learning for AF screening

One general limitation of the ML studies is the imbalance in data between the size of the AF cohort and the non-AF cohort, as the non-AF cohort is usually considerably larger. However, this was not the case for the Indiana Network owing to the population containing a higher proportion of older patients. To deal with the imbalances in data, resampling methods such as oversampling or undersampling have been used.

A limitation regarding the development of these ML models from electronic patient records is that the non-AF cohort may include patients who have undiagnosed AF and therefore are not labeled as having AF when training the ML algorithm. It is also not clear if, in the AF cohort, there may be patients who are also misdiagnosed with AF. A prospective, interventional study actively diagnosing AF patients could address this issue.

With regard to automated rhythm diagnosis, in the presence of multiple ectopic beats (premature atrial complexes or ventricular complexes), there can be a high risk of false-positive AF findings owing to rhythm irregularity.³⁵ This can consequently impact accuracy of ML for

automated AF diagnosis. Memory-stored templates of ectopic beats can help to overcome misinterpretation of ectopic beats for AF. For PPG, tachograms, Lorenz plots,³⁶ or Poincare plots³⁷ can be used to aid in improving accurate AF diagnosis.

The impact of missing data on the efficacy of an ML algorithm for AF screening is also unclear. There is limited transparency about how the processes of ML lead to their outputs. Although there are inherent limitations to the accuracy of the ML algorithms, in theory, they will most likely outperform current practice of opportunistic screening for AF.

The ideal ML algorithm for AF risk prediction would use electronic health records in combination with 12-lead ECGs. However, the availability of 12-lead ECG data from electronic health records and in an appropriate form limits our ability to develop such an algorithm.

Conclusion

There have been many developments in the use of ML to improve AF screening. These ML algorithms have either been used to identify the individuals at highest risk of AF to better target the screening or to facilitate more accurate automated AF diagnosis during the screening process. Most ML algorithms have been developed and tested on retrospective data, but there needs to be more work to test them prospectively and then to assess their impact on real-world data.

Funding Sources

PS and SK are supported by Chelsea and Westminster Hospitals NHS Foundation Trust, United Kingdom. FSN is supported by the British Heart Foundation, United Kingdom (RG/16/3/32175) and the National Institute for Health Research Imperial Biomedical Research Centre, United Kingdom.

Disclosures

PS, SK, and FSN are currently conducting a study funded by Bristol-Myers Squibb Pharmaceuticals Limited and Chelsea & Westminster Charity+ to test a machine learning AF risk prediction algorithm across primary care networks. The other authors have no conflicts of interest to declare.

Authorship

All authors attest they meet the current ICMJE criteria for authorship.

Appendix Supplementary data

Supplementary data associated with this article can be found in the online version at <https://doi.org/10.1016/j.cvdhj.2022.04.001>.

References

- Hindricks G, Potpara T, Dagres N, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with EACTS. *Eur Heart J* 2021;42:373–498.
- Hart RG, Pearce LA, Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Ann Intern Med* 2007;146:857–867.
- Surveillance decision | Evidence | Atrial fibrillation: management | Guidance | NICE. <https://www.nice.org.uk/guidance/cg180/resources/surveillance-report-2017-atrial-fibrillation-management-2014-nice-guideline-cg180-4597399263/chapter/Surveillance-decision#reason-for-the-decision>. Accessed November 5, 2020
- US Preventive Services Task Force. Screening for atrial fibrillation with electrocardiography: US Preventive Services Task Force Recommendation Statement. *JAMA* 2018;320:478–484.
- Svennberg E, Friberg L, Frykman V, Al-Khalili F, Engdahl J, Rosenqvist M. Clinical outcomes in systematic screening for atrial fibrillation (STROKESTOP): a multicentre, parallel group, unmasked, randomised controlled trial. *Lancet*. [https://doi.org/10.1016/S0140-6736\(21\)01637-8](https://doi.org/10.1016/S0140-6736(21)01637-8)
- Svendsen JH, Diederichsen SZ, Højberg S, et al. Implantable loop recorder detection of atrial fibrillation to prevent stroke (The LOOP Study): a randomised controlled trial. *Lancet*. [https://doi.org/10.1016/S0140-6736\(21\)01698-6](https://doi.org/10.1016/S0140-6736(21)01698-6)
- Chan NY. Systematic screening for atrial fibrillation in the community: evidence and obstacles. *Arrhythm Electrophysiol Rev* 2018;7:39–42.
- Lowres N, Olivier J, Chao TF, et al. Estimated stroke risk, yield, and number needed to screen for atrial fibrillation detected through single time screening: a multicountry patient-level meta-analysis of 141,220 screened individuals. *PLOS Med* 2019;16:e1002903.
- Schnabel RB, Sullivan LM, Levy D, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet* 2009;373:739–745.
- Chamberlain AM, Agarwal SK, Folsom AR, 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. <https://doi.org/10.1016/j.amjcard.2010.08.049>
- Alonso A, Krijthe BP, Aspelund T, 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–e000102.
- Shulman E, Kargoli F, Aagaard P, et al. Validation of the Framingham Heart Study and CHARGE-AF risk scores for atrial fibrillation in Hispanics, African-Americans, and non-Hispanic whites. *Am J Cardiol* 2016;117:76–83.
- Pfister R, Brägelmann J, Michels G, Wareham NJ, Luben R, Khaw KT. Performance of the CHARGE-AF risk model for incident atrial fibrillation in the EPIC Norfolk cohort. *Eur J Prev Cardiol* 2014;22:932–939.
- Saliba W, Gronich N, Barnett-Griness O, Rennert G. Usefulness of CHADS2 and CHA2DS2-VASc scores in the prediction of new-onset atrial fibrillation: a population-based study. *Am J Med* 2016;129:843–849.
- Hamada R, Muto S. Simple risk model and score for predicting of incident atrial fibrillation in Japanese. *J Cardiol* 2019;73:65–72.
- Kokubo Y, Watanabe M, Higashiyama A, Nakao YM, Kusano K, Miyamoto Y. Development of a basic risk score for incident atrial fibrillation in a Japanese general population—the Suita Study. *Circ J* 2017;81:1580–1588.
- Li YG, Bisson A, Bodin A, et al. C2 HEST score and prediction of incident atrial fibrillation in poststroke patients: a French nationwide study. *J Am Heart Assoc* 2019;8:e012546.
- Chao T, Chiang C, Chen T, Liao J, Tuan T, Chen S. Clinical risk score for the prediction of incident atrial fibrillation: derivation in 7 220 654 Taiwan patients with 438 930 incident atrial fibrillations during a 16-year follow-up. *J Am Heart Assoc* 2021;10:e020194.
- Linker DT, Murphy TB, Mokdad AH. Selective screening for atrial fibrillation using multivariable risk models. *Heart* 2018;104:1492–1499.
- Petryszyn P, Niewinski P, Staniak A, et al. Effectiveness of screening for atrial fibrillation and its determinants. A meta-analysis. *PLoS One* 2019;14:e0213198–e0213198.
- Shade JK, Ali RL, Basile D, et al. Preprocedure application of machine learning and mechanistic simulations predicts likelihood of paroxysmal atrial fibrillation recurrence following pulmonary vein isolation. *Circ Arrhythm Electrophysiol* 2020;13:e008213.
- Alhuseini MI, Abuzaid F, Rogers AJ, et al. Machine learning to classify intracardiac electrical patterns during atrial fibrillation. *Circ Arrhythm Electrophysiol* 2020;13:e008160.
- Vinter N, Frederiksen AS, Albertsen AE, et al. Role for machine learning in sex-specific prediction of successful electrical cardioversion in atrial fibrillation? *Open Heart* 2020;7:e001297.
- Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019;394:861–867.
- Hirota N, Suzuki S, Arita T, et al. Prediction of current and new development of atrial fibrillation on electrocardiogram with sinus rhythm in patients without structural heart disease. *Int J Cardiol* 2021;327:93–99.
- Raghunath S, Pfeifer JM, Ulloa-Cerna AE, et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation-related stroke. *Circulation* 2021;143:1287–1298.
- 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* 2020;125:55–62.
- Tiwari P, Colborn KL, Smith DE, Xing F, Ghosh D, Rosenberg MA. Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation. *JAMA Netw Open* 2020;3:e1919396–e1919396.
- Suzuki R, Katada J, Ramagopalan S, McDonald L. Potential of machine learning methods to identify patients with nonvalvular atrial fibrillation. *Future Cardiol* 2019;16:43–52.
- Grout RW, Hui SL, Imler TD, et al. Development, validation, and proof-of-concept implementation of a two-year risk prediction model for undiagnosed atrial fibrillation using common electronic health data (UNAFIED). *BMC Med Inform Decis Mak* 2021;21:112.
- Hill NR, Ayoubkhani D, McEwan P, et al. Predicting atrial fibrillation in primary care using machine learning. *PLoS One* 2019;14:e0224582.
- Sekelj S, Sandler B, Johnston E, et al. Detecting undiagnosed atrial fibrillation in UK primary care: validation of a machine learning prediction algorithm in a retrospective cohort study. *Eur J Prev Cardiol*. Published online August 13, 2020. <https://doi.org/10.1177/2047487320942338>
- Pollock KG, Sekelj S, Johnston E, et al. Application of a machine learning algorithm for detection of atrial fibrillation in secondary care. *IJC Heart Vasc* 2020;31:100674.
- Pereira T, Tran N, Gadhomi K, et al. Photoplethysmography based atrial fibrillation detection: a review. *NPJ Digit Med* 2020;3:3.
- Han D, Bashar SK, Mohagheghian F, et al. Premature atrial and ventricular contraction detection using photoplethysmographic data from a smartwatch. *Sensors (Basel)* 2020;20:5683.
- Lown M, Brown M, Brown C, et al. Machine learning detection of atrial fibrillation using wearable technology. *PLoS One* 2020;15:e0227401.
- Gruwez H, Evens S, Proesmans T, et al. Accuracy of physicians interpreting photoplethysmography and electrocardiography tracings to detect atrial fibrillation: INTERPRET-AF. *Front Cardiovasc Med* 2021;8:734737.
- Hulme OL, Khurshid S, Weng LC, et al. Development and validation of a prediction model for atrial fibrillation using electronic health records. *JACC Clin Electrophysiol* 2019;5:1331–1341.
- de Vos CB, Pisters R, Nieuwlaar R, et al. Progression from paroxysmal to persistent atrial fibrillation: clinical correlates and prognosis. *J Am Coll Cardiol* 2010;55:725–731.
- Hu WS, Lin CL. Prediction of new-onset atrial fibrillation for general population in Asia: a comparison of C2HEST and HATCH scores. *Int J Cardiol* 2020;313:60–63.
- Khurshid S, Kartoun U, Ashburner JM, et al. Performance of atrial fibrillation risk prediction models in over 4 million individuals. *Circ Arrhythm Electrophysiol* 2021;14:e008997.
- William AD, Kanbour M, Callahan T, et al. Assessing the accuracy of an automated atrial fibrillation detection algorithm using smartphone technology: the iREAD Study. *Heart Rhythm* 2018;15:1561–1565.
- Maputo T, Zaprutko J, Baszko A, et al. Feasibility of atrial fibrillation screening with mobile health technologies at pharmacies. *J Cardiovasc Pharmacol Ther* 2020;25:142–151.
- Vaes B, Stalpaert S, Tavernier K, et al. The diagnostic accuracy of the MyDiagnostick to detect atrial fibrillation in primary care. *BMC Fam Pract* 2014;15:113.
- Wasserlauf J, You C, Patel R, Valys A, Albert D, Passman R. Smartwatch performance for the detection and quantification of atrial fibrillation. *Circ Arrhythm Electrophysiol* 2019;12:e006834.
- Brasier N, Raichle CJ, Dörr M, et al. Detection of atrial fibrillation with a smartphone camera: first prospective, international, two-centre, clinical validation study (DETECT AF PRO). *Europace* 2019;21:41–47.
- Proesmans T, Mortelmans C, Van Haelst R, Verbrugge F, Vandervoort P, Vaes B. Mobile phone-based use of the photoplethysmography technique to detect atrial fibrillation in primary care: diagnostic accuracy study of the FibrCheck app. *JMIR Mhealth Uhealth* 2019;7:e12284.

48. Poh MZ, Poh YC, Chan PH, et al. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. *Heart* 2018;104:1921.
49. McManus DD, Lee J, Maitas O, et al. A novel application for the detection of an irregular pulse using an iPhone 4S in patients with atrial fibrillation. *Heart Rhythm* 2013;10:315–319.
50. Lee J, Reyes BA, McManus DD, Maitas O, Chon KH. Atrial fibrillation detection using an iPhone 4S. *IEEE Trans Biomed Eng* 2013;60:203–206.
51. Bashar SK, Han D, Hajeb-Mohammadalipour S, et al. Atrial fibrillation detection from wrist photoplethysmography signals using smartwatches. *Sci Rep* 2019; 9:15054.
52. Dörr M, Nohturfft V, Brasier N, et al. The WATCH AF trial: SmartWATCHes for Detection of Atrial Fibrillation. *JACC Clin Electrophysiol* 2019;5:199–208.
53. Bumgarner JM, Lambert CT, Hussein AA, et al. Smartwatch algorithm for automated detection of atrial fibrillation. *J Am Coll Cardiol* 2018;71:2381–2388.
54. Okubo Y, Tokuyama T, Okamura S, et al. Evaluation of the feasibility and efficacy of a novel device for screening silent atrial fibrillation (MYBEAT trial). *Circ J* 2022;86:182–188.
55. Wiesel J, Fitzig L, Herschman Y, Messineo FC. Detection of atrial fibrillation using a modified microlife blood pressure monitor. *Am J Hypertens* 2009;22:848–852.