Original research

Prediction of short-term atrial fibrillation risk using primary care electronic health records

Ramesh Nadarajah [1,2], Jianhua Wu [1,3], David Hogg,[4] Keerthenan Raveendra,[5] Yoko M Nakao [1,2], Kazuhiro Nakao,[1,2] Ronen Arbel [6,7], Moti Haim,[8,9] Doron Zahger,[9,10] John Parry,[11] Chris Bates [11], Campbel Cowan,[12] Chris P Gale[1,12]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/heartjnl-2022-322076).

For numbered affiliations see end of article.

Correspondence to
Dr Ramesh Nadarajah, Leeds Institute for Data Analytics, University of Leeds, Leeds LS2 9JT, UK;
r.nadarajah@leeds.ac.uk

RN and JW are joint first authors.

Received 1 November 2022
Accepted 26 January 2023
Published Online First
9 February 2023

Linked

► http://dx.doi.org/10.1136/heartjnl-2023-322395

To cite: Nadarajah R, Wu J, Hogg D, *et al*. *Heart* 2023;**109**:1072–1079.

ABSTRACT

Objective Atrial fibrillation (AF) screening by age achieves a low yield and misses younger individuals. We aimed to develop an algorithm in nationwide routinely collected primary care data to predict the risk of incident AF within 6 months (Future Innovations in Novel Detection of Atrial Fibrillation (FIND-AF)).

Methods We used primary care electronic health record data from individuals aged ≥30 years without known AF in the UK Clinical Practice Research Datalink-GOLD dataset between 2 January 1998 and 30 November 2018, randomly divided into training (80%) and testing (20%) datasets. We trained a random forest classifier using age, sex, ethnicity and comorbidities. Prediction performance was evaluated in the testing dataset with internal bootstrap validation with 200 samples, and compared against the CHA$_2$DS$_2$-VASc (Congestive heart failure, Hypertension, Age >75 (2 points), Stroke/transient ischaemic attack/thromboembolism (2 points), Vascular disease, Age 65–74, Sex category) and C$_2$HEST (Coronary artery disease/Chronic obstructive pulmonary disease (1 point each), Hypertension, Elderly (age ≥75, 2 points), Systolic heart failure, Thyroid disease (hyperthyroidism)) scores. Cox proportional hazard models with competing risk of death were fit for incident longer-term AF between higher and lower FIND-AF-predicted risk.

Results Of 2 081 139 individuals in the cohort, 7386 developed AF within 6 months. FIND-AF could be applied to all records. In the testing dataset (n=416 228), discrimination performance was strongest for FIND-AF (area under the receiver operating characteristic curve 0.824, 95% CI 0.814 to 0.834) compared with CHA$_2$DS$_2$-VASc (0.784, 0.773 to 0.794) and C$_2$HEST (0.757, 0.744 to 0.770), and robust by sex and ethnic group. The higher predicted risk cohort, compared with lower predicted risk, had a 20-fold higher 6-month incidence rate for AF and higher long-term hazard for AF (HR 8.75, 95% CI 8.44 to 9.06).

Conclusions FIND-AF, a machine learning algorithm applicable at scale in routinely collected primary care data, identifies people at higher risk of short-term AF.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ European Society of Cardiology Guidelines recommend opportunistic screening in individuals aged ≥65 years and systematic screening in individuals aged ≥75 years. However, this approach achieves low yields and misses the increasing number of people diagnosed with atrial fibrillation (AF) before the age of 65 years.

⇒ Several AF risk prediction algorithms have been tested using community-based electronic health records (EHRs). However, current models are limited by moderate discrimination performance, limited scalability and long prediction horizons, which are not relevant to the decision to investigate for AF in the short term.

WHAT THIS STUDY ADDS

⇒ In this nationwide primary care EHR study, we show that a random forest classifier (Future Innovations in Novel Detection of Atrial Fibrillation (FIND-AF)) can be used to accurately predict AF risk within 6 months, superior to the C$_2$HEST and CHA$_2$DS$_2$-VASc scores, and can be applied to all UK primary care EHRs.

⇒ One-fifth of incident AF cases in 6 months occurred in individuals younger than 65 years who would ordinarily be excluded from AF screening programmes. FIND-AF identified a cohort of higher-risk individuals younger than 65 years of age, and higher predicted AF risk was associated with elevated incident AF in the short and long term.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Leveraging FIND-AF, a scalable machine learning algorithm, in routinely collected EHRs may improve the efficiency of diagnostic pathways for AF.

⇒ External validation and evaluation of prospective clinical deployment of FIND-AF are in process, and a cost utility analysis and budget impact analysis will need to be conducted.

INTRODUCTION

Atrial fibrillation (AF) is a major public health issue. There are now more new cases of AF diagnosed each year in the English National Health Service (NHS) than the four most common causes of cancer combined.[1] Moreover, it is estimated that up to 35% of disease burden remains undiagnosed,[2] and 15% of strokes occur in the context of undiagnosed AF.[3]

Early detection of AF may permit the initiation of oral anticoagulation to reduce embolic stroke risk,[4] and early antiarrhythmic therapy to reduce the risk of death and stroke.[5] Accordingly, early AF detection is a key cardiovascular priority in the UK NHS Long Term Plan,[6] and the European Society of Cardiology recommends opportunistic screening by pulse palpation or ECG rhythm strip in persons aged ≥65 years and systematic ECG screening in those aged ≥75 years.[7] However, there is an increasing cohort of individuals aged younger than 65 years who are being diagnosed with AF and are eligible for anticoagulation.[1]

A large proportion of the population is registered in primary care with a routinely collected electronic health record (EHR).[8 9] An algorithm that uses routinely collected EHR data to calculate AF risk could give a scalable, efficient and fair approach to targeting AF detection. However, previous algorithms tested in community-based EHRs have a number of shortcomings (online supplemental tables 1 and 2). First, many algorithms developed using traditional regression techniques show only moderate discriminative performance.[10] Second, algorithm prediction horizons are often 5 or 10 years, making it difficult to judge the merits of investigating individuals in the short term.[9 11] Third, reports have infrequently investigated for variation in algorithm prediction performance by sex and ethnicity.[11] Fourth, algorithms often require variables frequently missing from routinely collected data such as height, weight and blood pressure thereby restricting the population to which they can be applied.[9 11]

Therefore, our objective was to train and test an algorithm (Future Innovations in Novel Detection of Atrial Fibrillation, FIND-AF) that predicts an individual's risk of AF in the next 6 months using routinely recorded data in primary care EHRs. We compared performance against other AF prediction algorithms and investigated for variation in performance by sex and ethnicity.

METHODS

Study design and population

In this population-based study, we used primary care EHRs from the UK Clinical Practice Research Datalink (CPRD)-GOLD dataset. CPRD is one of the largest databases of longitudinal medical records from primary care worldwide and contains anonymised patient data from approximately 7% of the UK population.[8] CPRD-GOLD represents the UK population in terms of age, sex and ethnicity,[8] and has been used to develop algorithms for predicting AF.[11] Data collection happens as part of routine clinical care in participating practices and patients are included in the primary care dataset from their first until their last contact with a participating practice.[8] Diagnostic coding for AF in CPRD has been shown to be consistent and valid, with a positive predictive value (PPV) of 98%.[12]

All individuals in the CPRD dataset were linked to Hospital Episode Statistics (HES) Admitted Patient Care (APC) records to obtain comprehensive coverage of AF cases diagnosed in secondary care. We included all adults registered at practices within CPRD who were ≥30 years of age at entry with no history of AF from either data source and at least 1-year follow-up between 2 January 1998 and 30 November 2018. Individuals were censored to a diagnosis of AF (or atrial flutter (AFl), since it has similar thromboembolic risk and anticoagulation guidelines),[7] withdrawal from CPRD or 6 months, whichever came first. Diagnoses of AF or AFl in primary care were identified using Read codes in CPRD and in secondary care with the 10th revision of the International Statistical Classification of Diseases and Related Health Problems codes in HES-APC (online supplemental table 3). Individuals were randomly split 4:1 to establish a training dataset (80%) and a testing dataset (20%) using the Mersenne twister pseudorandom number generator.

We followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis reporting guideline and the CODE-EHR best-practice framework for using structured electronic healthcare records in clinical research.[13 14]

FIND-AF algorithm development

A random forest (RF) classifier was trained to predict AF at 6 months. Our systematic review evidenced strong discriminative performance for AF prediction using RF across different EHR datasets.[10] RF is a machine learning method consisting of many individual decision trees that operate as an ensemble.[15] FIND-AF was trained using 10-fold cross-validation on the full training set (full details available in online supplemental methods).

To create an algorithm that could be implemented at scale in national primary care EHRs, we restricted candidate variables to age, sex, comorbidities (72 binary variables, indicating presence or absence of recorded diagnosis) and ethnicity (six categories; online supplemental table 6). Observations and laboratory results were not included. Ethnicity information is routinely collected in the UK NHS and so has increasingly high completeness,[16] and we included an 'ethnicity unrecorded' category where it was unavailable because missingness was considered to be informative.[17] Predictor variables were selected a priori from systematic review of variables included in previous AF risk prediction algorithms,[10] plus an updated literature review (online supplemental tables 4–6). Diagnostic code lists only included the primary care coding system (Read codes), ensuring that only information readily available within a primary care EHR could be incorporated within the algorithm. Concordantly, our entire analytical cohort had no missing data for any of the predictor variables and the algorithm could be applied to all records.

Statistical analyses

The baseline characteristics are summarised by incident AF status. Continuous variables were reported as mean±SD. Categorical variables were reported as frequencies with corresponding percentages.

The degree of variation of each feature in FIND-AF to classification was calculated using the mean decrease in the Gini coefficient, a measure of how each variable contributes to the homogeneity of nodes and leaves in the resulting RF.

Model performance of FIND-AF was determined using the full holdout test set with internal bootstrap validation with 200 samples and compared with a multivariable logistic regression (MLR) model developed with backward model selection with Akaike information criterion.[18] Performance was compared with the CHA$_2$DS$_2$-VASc (Congestive heart failure, Hypertension, Age >75 (2 points), Stroke/transient ischaemic attack/thromboembolism (2 points), Vascular disease, Age 65–74, Sex category) and C$_2$HEST (Coronary artery disease/Chronic obstructive pulmonary disease (1 point each), Hypertension, Elderly (age ≥75, 2 points), Systolic heart failure, Thyroid disease (hyperthyroidism)) scores. The CHA$_2$DS$_2$-VASc score was originally developed to predict stroke risk in individuals with AF, and the C$_2$HEST score for Asian people without structural heart disease.[10] These algorithms are robust to missing data in routinely collected primary care EHRs and have been tested for AF risk prediction in European cohorts (online supplemental table 2).[10] Other algorithms that can only be applied to a minority of European primary care

EHRs (Pfizer-AI, CHARGE-AF) were not considered.[9,19] The area under the receiver operating characteristic (AUROC) curve was used to evaluate predictive ability (concordance index) with 95% CIs calculated using the DeLong method. Youden Index was established for the outcome measure as a method of empirically identifying the optimal dichotomous cut-off to assess sensitivity, specificity, PPV and negative predictive value (NPV). Youden Index was calculated and optimised for each test set for each score to derive the optimal cut-off threshold. Calibration was assessed by plotting predicted AF risk against observed AF incidence and by the calibration slope. We calculated the Brier score, a measure of both discrimination and calibration, by taking the mean squared difference between predicted probabilities and the observed outcome. To assess the clinical impact of using FIND-AF as opposed to other risk prediction scores, we calculated the net reclassification index at 0.4% AF risk threshold (the average 6-month incidence rate in the cohort) and conducted a decision curve analysis.

We investigated the performance of FIND-AF, CHA$_2$DS$_2$-VASc and C$_2$HEST within relevant subgroups defined by sex, ethnicity (white vs black vs Asian vs other non-white ethnic minorities) and age (≥65 years and ≥75 years). We plotted Kaplan-Meier plots for individuals identified as higher and lower FIND-AF-predicted risk of AF to assess the event rate for AF censored at 10 years, and calculated the HR for AF between higher and lower FIND-AF-predicted risk of AF using the Cox proportional hazard model with adjustment for the competing risk of death. We used R V.4.1.0 for all analyses.

Table 1 Baseline characteristics of analytical cohort with and without atrial fibrillation (AF)

	Incident AF	
	No AF n (%)	AF n (%)
	2 073 753	7386
Demographics		
Age, years	49.82 (15.37)	73.72 (12.62)
Sex (women)	1 051 942 (50.7)	3619 (49.0)
Comorbidities		
Diabetes mellitus	71 966 (3.5)	815 (11.0)
Stroke or TIA	37 773 (1.8)	892 (12.1)
Ischaemic heart disease	77 060 (3.7)	1542 (20.9)
Hypertension	247 436 (11.9)	2887 (39.1)
Heart failure	13 717 (0.7)	650 (8.8)
Dyslipidaemia	60 357 (2.9)	532 (7.2)
Hyperthyroidism	16 147 (0.8)	155 (2.1)
COPD	24 962 (1.2)	461 (6.2)
Chronic kidney disease	29 359 (1.4)	449 (6.1)
Anaemia	66 844 (3.2)	501 (6.8)
Cancer	72 621 (3.5)	887 (12.0)
Valvular heart disease	9 497 (0.5)	376 (5.1)
Mean CHA$_2$DS$_2$-VASc score (SD)	0.97 (1.03)	2.72 (1.42)

CHA$_2$DS$_2$-VASc, Congestive heart failure, Hypertension, Age >75 years (2 points), Stroke/transient ischaemic attack/thromboembolism (2 points), Vascular disease, Age 65–74 years, Sex category; COPD, chronic obstructive pulmonary disease; TIA, transient ischaemic attack.

Patient and public involvement

The Arrhythmia Alliance, an AF association, provided input on the FIND-AF scientific advisory board. The FIND-AF patient and public involvement group have given input to reporting and dissemination plans of the research.

RESULTS

Patient population

There were 2 081 139 individuals registered in our UK primary care cohort (1 664 911 in the training dataset, 416 228 in testing dataset), with average age 49.9 years (SD 15.4), 50.7% women and 86.7% white. Baseline characteristics and clinical outcomes were similar in the training and testing datasets (online supplemental table 7). Within 6 months, 7386 individuals (0.4%) were recorded as having AF. Those who developed AF were older and had a higher prevalence of baseline comorbidities than individuals who did not develop AF (table 1). Of new cases, 1546 (20.9%) were younger than 65 years old.

Prediction factors and model accuracy

According to mean decrease in the Gini coefficient, age contributed the most to the prediction, followed by ethnicity and history of heart failure (figure 1). AF discrimination and accuracy of predictions, by AUROC and Brier scores, were better using FIND-AF than the MLR, CHA$_2$DS$_2$-VASc and C$_2$HEST algorithms (table 2 and figure 2). Sensitivity was highest for the CHA$_2$DS$_2$-VASc algorithm, but specificity lowest.

According to the Youden Index, the optimal cut-off was 0.0032, leading to a sensitivity of 78% and a specificity of 73%, with a PPV of 2.5% and NPV of 99.8%. The low incidence of AF over 6 months led to similar values for PPV and NPV across the algorithms. Of the algorithms, FIND-AF was the best calibrated (calibration slope 0.782 (95% CI 0.743 to 0.824), table 2 and online supplemental figure 1), yet showed underestimation of risk in the mid-risk strata and overestimation in the highest risk strata.

Risk classification

Of the 416 228 individuals in the testing set, 82 942 (19.9%) were classified as higher risk using FIND-AF, 84 282 (20.2%) using the CHA$_2$DS$_2$-VASc score and 84 542 (20.3%) using the C$_2$HEST score, respectively. Net reclassification analyses at the 0.4% risk threshold demonstrated modestly favourable reclassification using FIND-AF as opposed to using CHA$_2$DS$_2$-VASc (net reclassification 0.032, 95% CI 0.029 to 0.051) and strong favourable reclassification using FIND-AF as opposed to using C$_2$HEST (net reclassification 0.113, 95% CI 0.098 to 0.135; online supplemental table 8). In a decision curve analysis, FIND-AF had a superior net benefit compared with the CHA$_2$DS$_2$-VASc and C$_2$HEST risk scores across all threshold probabilities (online supplemental figure 2).

Of the 82 942 individuals identified as higher risk by FIND-AF, 3483 were <65 years of age, of whom 3448 had a CHA$_2$DS$_2$-VASc score of at least 1. The incidence rate of AF in routine clinical practice at 6 months was 20-fold higher among individuals identified as a higher predicted risk of AF by FIND-AF compared with individuals identified as lower risk (2.0% vs 0.1%). In routine clinical practice, 1 in every 71 individuals aged ≥65 years were diagnosed with AF within 6 months, 1 in every 58 individuals aged ≥75 years and 1 in every 40 individuals identified at higher predicted AF risk.

Higher predicted AF risk was also associated with increased long-term AF occurrence. Within 5 and 10 years, respectively, 5.1% and 11.9% of the higher predicted risk cohort had been diagnosed with AF, with an 8.75-fold increased hazard (95% CI 8.44 to 9.06) relative to individuals at lower predicted risk (figure 3).

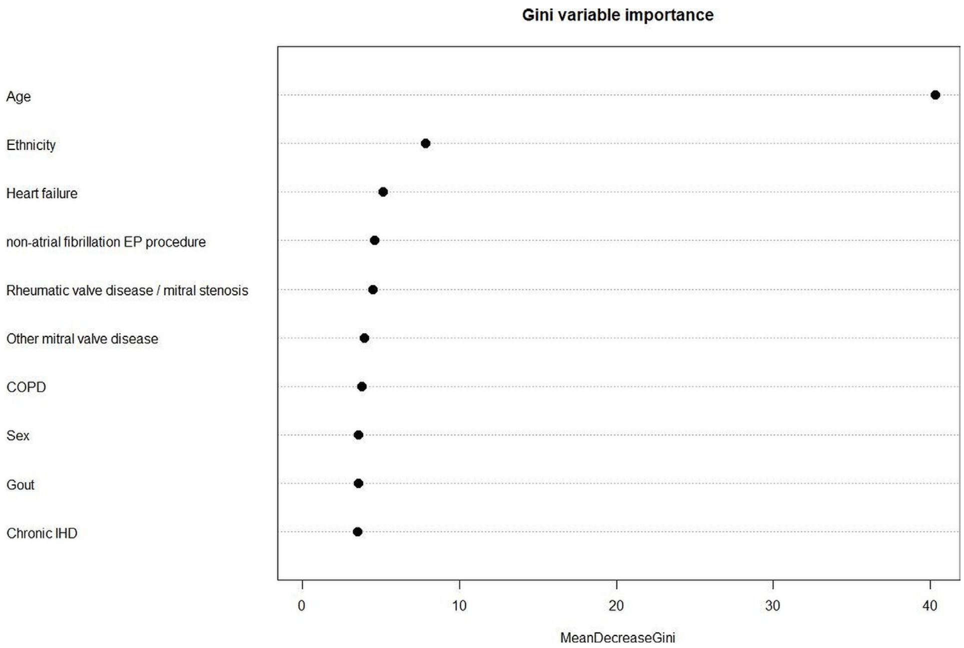


Figure 1 The top 10 most important variables for FIND-AF prediction in individuals aged ≥30 years quantified by mean decrease in Gini coefficient. COPD, chronic obstructive pulmonary disease; EP, electrophysiology; FIND-AF, Future Innovations in Novel Detection of Atrial Fibrillation; IHF, ischaemic heart disease.

Model performance in clinically relevant subgroups

FIND-AF discrimination performance remained strong in both sexes, whereas for the CHA$_2$DS$_2$-VASc and C$_2$HEST scores, performance was better in men than women (table 3). The scores performed differently across ethnic groups. In black individuals, AF discrimination was highest for CHA$_2$DS$_2$-VASc, and in white and Asian individuals, FIND-AF had the strongest discrimination performance.

DISCUSSION

In this population-based study, we trained a machine learning algorithm (FIND-AF) on more than 1.5 million individuals registered in UK primary care to predict the risk of incident AF within the next 6 months (figure 4). When tested in over 400 000 individuals, FIND-AF demonstrated good predictive accuracy, which was superior to other risk scores and robust in both sexes and across ethnic groups. FIND-AF identified a cohort of younger people at higher risk of AF and more efficiently identified individuals diagnosed with AF within 6 months compared with age-based risk stratification. Finally, short-term predicted AF risk also translated to long-term AF occurrence.

Current approaches to targeting investigation for undiagnosed AF are based on age.[7] Our analysis demonstrated that one-fifth of newly detected AF cases within 6 months occur in people aged ≤65 years, emphasising the opportunity lost when enhanced AF investigation is restricted to older populations. ECGs can be used to accurately predict AF risk,[20] but they are not widely available in the community, whereas 98% of the UK population are registered in primary care with an accompanying EHR.[8] Our meta-analysis of AF prediction algorithms using EHRs demonstrated that algorithms developed using traditional regression techniques provided only moderate discrimination performance.[10] In our study, a machine learning prediction algorithm (FIND-AF) outperformed the C$_2$HEST and CHA$_2$DS$_2$-VASc scores.

For a machine learning prediction algorithm to be useful in clinical practice, it must be implementable within the clinical workflow, provide prediction that meaningfully informs decision-making and engender confidence in how outputs were

Table 2 Performance for 6-month incident AF with optimal threshold determined by Youden Index

	Algorithm			
	FIND-AF	MLR	CHA$_2$DS$_2$-VASc	C$_2$HEST
AUROC (95% CI)	0.824 (0.814 to 0.834)	0.765 (0.755 to 0.769)	0.784 (0.773 to 0.794)	0.757 (0.744 to 0.770)
Sensitivity (95% CI)	0.781 (0.731 to 0.829)	0.760 (0.653 to 0.814)	0.847 (0.829 to 0.866)	0.642 (0.619 to 0.791)
Specificity (95% CI)	0.731 (0.693 to 0.771)	0.679 (0.635 to 0.776)	0.611 (0.608 to 0.612)	0.790 (0.622 to 0.792)
PPV (%(95% CI))	2.5% (2.3 to 2.7)	2.0% (1.8 to 2.6)	2.2% (2.1 to 2.3)	2.0% (1.5 to 2.2)
NPV (%(95% CI))	99.8% (99.8 to 99.8)	99.7% (99.6 to 99.7)	99.8% (99.8 to 99.8)	99.7% (99.7 to 99.8)
Calibration slope* (95% CI)	0.782 (0.743 to 0.824)	0.698 (0.654 to 0.735)	0.621 (0.589 to 0.652)	0.608 (0.576 to 0.648)
Brier score	0.069	0.097	0.093	0.102

*Calibration slope was derived from linear regression models by forcing the intercept through origin (0, 0).

AF, atrial fibrillation; AUROC, area under the receiver operating characteristic; CHA$_2$DS$_2$-VASc, Congestive heart failure, Hypertension, Age >75 (2 points), Stroke/transient ischaemic attack/thromboembolism (2 points), Vascular disease, Age 65–74, Sex category; C$_2$HEST, Coronary artery disease/Chronic obstructive pulmonary disease (1 point each), Hypertension, Elderly (age ≥75, 2 points), Systolic heart failure, Thyroid disease (hyperthyroidism); FIND-AF, Future Innovations in Novel Detection of Atrial Fibrillation; MLR, multivariable logistic regression; NPV, negative predictive value; PPV, positive predictive value.

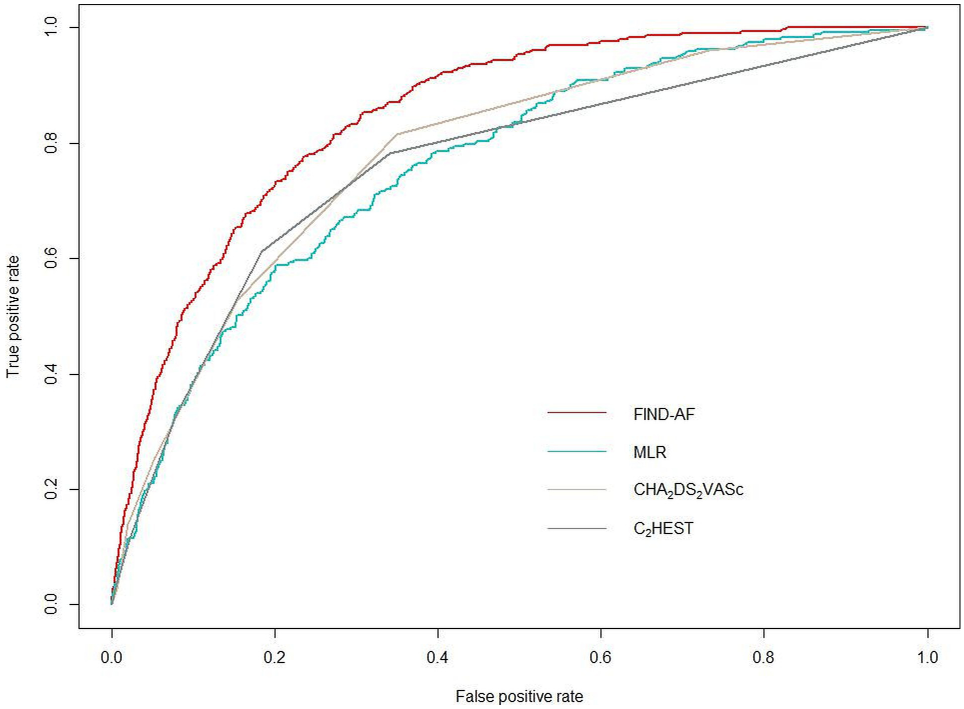


Figure 2 Receiver operating characteristic curves for FIND-AF, the multivariable logistic regression (MLR), CHA$_2$DS$_2$-VASc and C$_2$HEST algorithm. C$_2$HEST, Coronary artery disease/Chronic obstructive pulmonary disease (1 point each), Hypertension, Elderly (age ≥75, 2 points), Systolic heart failure, Thyroid disease (hyperthyroidism); CHA$_2$DS$_2$-VASc, Congestive heart failure, Hypertension, Age >75 (2 points), Stroke/transient ischaemic attack/ thromboembolism (2 points), Vascular disease, Age 65–74, Sex category.

arrived at.[21] FIND-AF has been designed to be implemented and displayed through EHR systems, so will be available in a platform that healthcare professionals are interacting with as part of routine care. By design, FIND-AF provides AF risk prediction over a short time frame and so could assist clinicians at point of care in identifying patients for targeted diagnostics such as ECG

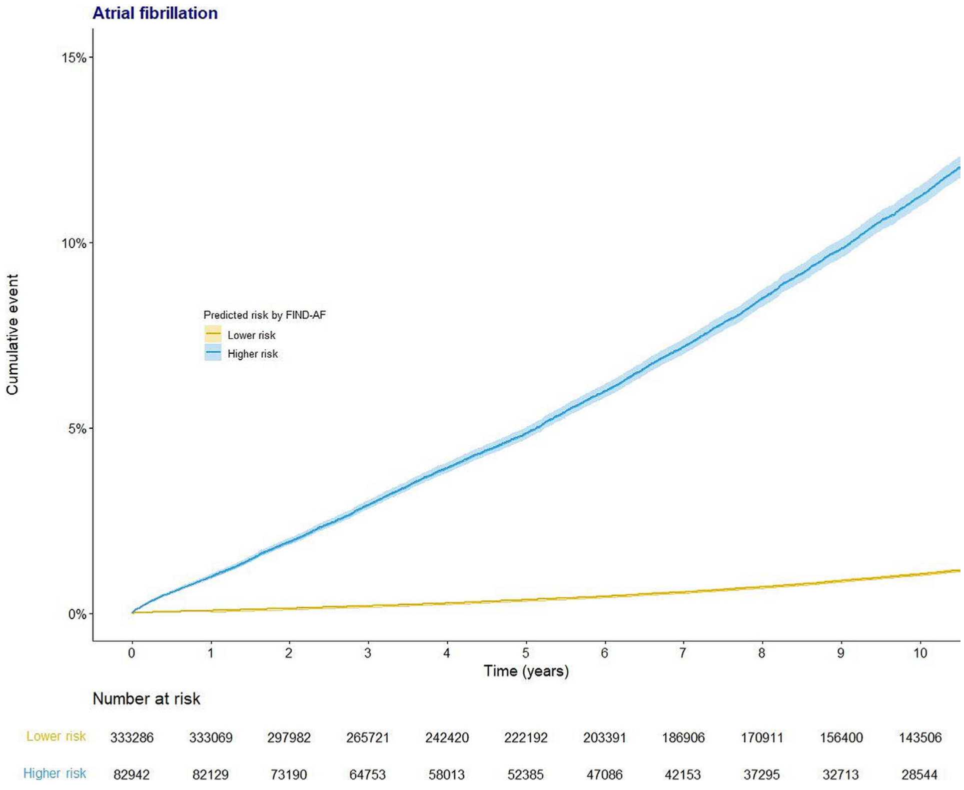


Figure 3 Kaplan-Meier plots for AF occurrence, by predicted risk from FIND-AF. AF, atrial fibrillation; FIND-AF, Future Innovations in Novel Detection of Atrial Fibrillation.

Table 3 Discrimination performance of FIND-AF, CHA_2DS_2-VASc and C_2HEST by sex, age and ethnicity

	FIND-AF	CHA_2DS_2-VASc	C_2HEST
	AUROC (95% CI)	AUROC (95% CI)	AUROC (95% CI)
Overall	0.824 (0.814 to 0.834)	0.784 (0.773 to 0.794)	0.757 (0.744 to 0.770)
Sex			
Men	0.819 (0.809 to 0.829)	0.807 (0.793 to 0.821)	0.793 (0.777 to 0.810)
Women	0.821 (0.810 to 0.831)	0.776 (0.760 to 0.793)	0.746 (0.727 to 0.765)
Age			
≥65 years	0.712 (0.698 to 0.727)	0.669 (0.654 to 0.684)	0.675 (0.661 to 0.690)
≥75 years	0.657 (0.638 to 0.675)	0.612 (0.593 to 0.632)	0.589 (0.570 to 0.608)
Ethnicity			
White	0.810 (0.799 to 0.821)	0.781 (0.769 to 0.792)	0.756 (0.743 to 0.770)
Asian	0.796 (0.693 to 0.899)	0.758 (0.639 to 0.876)	0.731 (0.611 to 0.850)
Black	0.801 (0.680 to 0.923)	0.843 (0.764 to 0.923)	0.707 (0.511 to 0.902)
Other non-white ethnic minority	0.805 (0.765 to 0.845)	0.768 (0.729 to 0.807)	0.805 (0.765 to 0.846)
Ethnicity unrecorded	0.823 (0.770 to 0.875)	0.838 (0.777 to 0.900)	0.788 (0.705 to 0.870)

The total number of individuals in each subgroup and number of incident AF cases are as follows: men (n=211 378, AF=720), women (n=204 850, AF=753), age ≥65 years (n=81 258, AF=1168), age ≥75 years (n=36 358, AF=796), white (n=279 027, AF=1301), Asian (n=8422, AF=16), black (n=6478, AF=11), other non-white ethnic minority (n=28 303, AF=96), ethnicity unrecorded (n=93 998, AF=49).

AF, atrial fibrillation; AUROC, area under the receiver operating characteristic; CHA_2DS_2-VASc, Congestive heart failure, Hypertension, Age >75 (2 points), Stroke/transient ischaemic attack/thromboembolism (2 points), Vascular disease, Age 65–74, Sex category; C_2HEST, Coronary artery disease/Chronic obstructive pulmonary disease (1 point each), Hypertension, Elderly (age ≥75, 2 points), Systolic heart failure, Thyroid disease (hyperthyroidism); FIND-AF, Future Innovations in Novel Detection of Atrial Fibrillation.

monitoring. Finally, the most important predictors in FIND-AF are already well-recognised risk factors for AF (for example, age, heart failure, valvular heart disease), which provide reassurance in the associations being made by the algorithm.[7]

Fairness is a critical characteristic when considering the impact of prediction algorithms in healthcare. The CHARGE-AF and PuLSE-AI algorithms have strong AF prediction performance,[9 11] yet incorporate variables that are frequently missing (height, weight and systolic and diastolic blood pressure).[10] Consequently, their applicability is limited to 17% and 35% of primary care EHRs, respectively.[9 11] Often, health data poverty disproportionately affects individuals from minority ethnicities and deprived backgrounds, so the application of these algorithms could reinforce health inequities.[22] Furthermore, whether their performance varies by sex and in minority ethnic groups in European populations is unknown. In our study, the C_2HEST and CHA_2DS_2-VASc scores were less accurate in women compared with men, and their performance varied substantially across different ethnic groups. FIND-AF's design enabled its application to every single patient record in a nationally representative

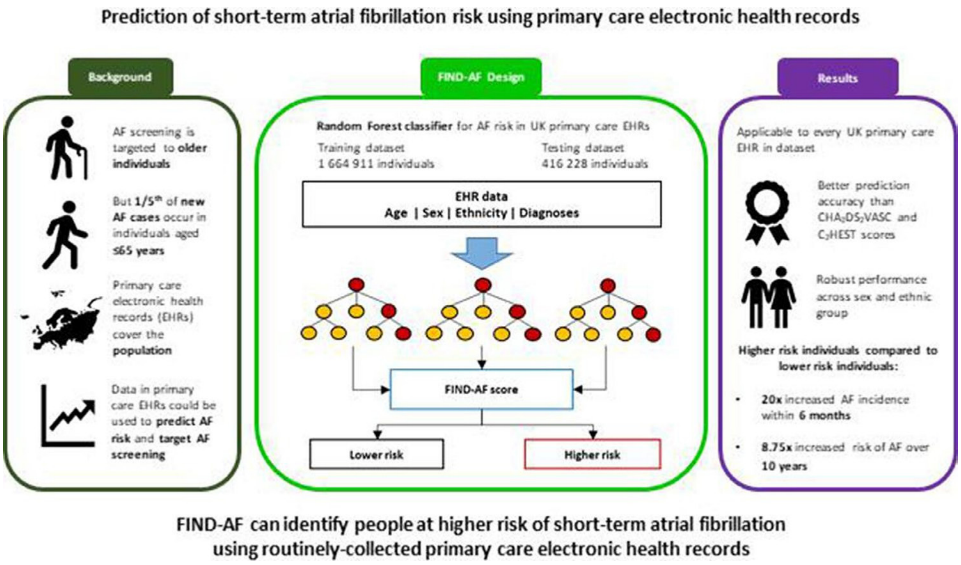

Prediction of short-term atrial fibrillation risk using primary care electronic health records
Background
AF screening is targeted to older individuals
But 1/5th of new AF cases occur in individuals aged ≤65 years
Primary care electronic health records (EHRs) cover the population
Data in primary care EHRs could be used to predict AF risk and target AF screening
FIND-AF Design
Random Forest classifier for AF risk in UK primary care EHRs
Training dataset 1 664 911 individuals
Testing dataset 416 228 individuals
EHR data
Age | Sex | Ethnicity | Diagnoses
FIND-AF score
Lower risk
Higher risk
Results
Applicable to every UK primary care EHR in dataset
Better prediction accuracy than CHA₂DS₂VASC and C₂HEST scores
Robust performance across sex and ethnic group
Higher risk individuals compared to lower risk individuals:
20x increased AF incidence within 6 months
8.75x increased risk of AF over 10 years
FIND-AF can identify people at higher risk of short-term atrial fibrillation using routinely-collected primary care electronic health records

Figure 4 Summary of the study and main findings. Hitherto implementation of screening for atrial fibrillation (AF) has been targeted to older persons in the general population, but this may miss one-fifth of new cases. A machine learning algorithm using routinely collected data in primary care electronic health records in the UK can accurately predict short-term risk of AF in persons aged ≥30 years. This may be a more efficient method for guiding AF screening. C_2HEST, Coronary artery disease/Chronic obstructive pulmonary disease (1 point each), Hypertension, Elderly (age ≥75, 2 points), Systolic heart failure, Thyroid disease (hyperthyroidism); CHA_2DS_2-VASc, Congestive heart failure, Hypertension, Age >75 (2 points), Stroke/transient ischaemic attack/thromboembolism (2 points), Vascular disease, Age 65–74, Sex category; FIND-AF, Future Innovations in Novel Detection of Atrial Fibrillation.

dataset of routinely collected primary care EHRs; and performance was robust in both sexes and across minority ethnic groups.

Three barriers need to be overcome for FIND-AF to be accepted into clinical practice. First, it requires external validation, which is currently underway using The Phoenix Partnership UK primary care EHR system (ResearchOne) and the Israeli Clalit Health Services. Second, prospective validation of FIND-AF is critical before implementation into clinical practice. We are launching a pilot implementation study across primary care sites where individuals identified at higher risk will be offered rhythm monitoring (The BHF Bristol Myers Squibb Cardiovascular Catalyst Award—CC/22/250026). Third, a cost utility analysis and budget impact analysis of the use of FIND-AF will need to be conducted.

Primary care EHRs in the UK are nationwide and held centrally, so FIND-AF could be activated at scale across geographically disparate sites to identify a subpopulation at elevated AF risk. The cohort identified as higher risk in this study included younger people who would currently be excluded from screening pathways, and higher predicted AF risk was associated with elevated AF occurrence both in the short and long term. Therefore, FIND-AF could facilitate efficient population-based AF screening or comprehensive programmes designed to improve risk factor profiles (including targeted weight loss and optimisation of blood pressure control).[23]

Screening for AF would adhere to many of the Wilson and Junger principles for a screening programme.[24] Opportunistic screening guided by age has not been demonstrated to increase AF detection rates,[25] but this may change in a more precisely defined higher-risk cohort. Systematic screening of older patients with intermittent or continuous (invasive or non-invasive) rhythm monitors is associated with increased AF detection rates, compared with routine care.[24] However, the yield of new cases is low (3% in the STROKESTOP trial)[26] and in our study, FIND-AF more efficiently identified a cohort with a higher rate of clinically detected AF than age-based approaches. Accurate risk assessment would be an integral component of a systematic screening process but ongoing research is needed to address the issues of the effectiveness and safety of treatment of screen-detected AF, and the costs of widespread use of ECG monitoring and prescription of oral anticoagulation, after the mixed results of the recently published LOOP and STROKESTOP trials.[26 27]

There are some limitations to our study. First, the CPRD database is routinely collected, retrospective primary care data. Underestimation of AF incidence is possible since there will have been individuals with unrecorded asymptomatic AF. Second, important predictor variables may have been 'missing by design'; nonetheless, we aimed to develop an algorithm that used routinely recorded data. Third, our choice of an RF classifier was based on a systematic review of AF prediction in EHRs,[10] and it is possible other machine learning methods may have performed differently in our study. Fourth, the algorithm will need to be updated as population characteristics change, data quality of EHRs improves and new or additional risk factors emerge. Fifth, electrophysiology procedures not specified as treating AF (including pacemaker implantations and percutaneous ablations) were a strong predictor of AF risk, and this may be a result of detection bias.

CONCLUSIONS

We trained and tested a novel machine learning algorithm (FIND-AF) that was applicable at scale within a nationwide routinely collected primary care EHR dataset. FIND-AF was able to accurately predict AF risk within 6 months and identify a cohort at elevated risk of AF in the longer term.

Author affiliations
[1]Leeds Institute for Data Analytics, University of Leeds, Leeds, UK
[2]Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, Leeds, UK
[3]Department of Dentistry, University of Leeds, Leeds, UK
[4]School of Computing, University of Leeds, Leeds, UK
[5]Faculty of Medicine and Health, University of Leeds, Leeds, UK
[6]Maximizing Health Outcomes Research Lab, Sapir College, Hof Ashkelon, Israel
[7]Community Medical Services Division, Clalit Health Services, Tel Aviv, Israel
[8]Department of Cardiology, Soroka University Medical Center, Beer Sheva, Israel
[9]Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel
[10]Cardiology, Soroka Medical Center, Beer Sheva, Israel
[11]The Phoenix Partnership, Leeds, UK
[12]Cardiology, Leeds General Infirmary, Leeds, UK

Twitter Ramesh Nadarajah @Dr_R_Nadarajah, Jianhua Wu @jianhuawu6, Keerthenan Raveendra @keerthraveendra, Yoko M Nakao @YokoMNakao and Chris P Gale @cpgale3

Acknowledgements We acknowledge the tremendous help of all members of the FIND-AF scientific advisory board and FIND-AF patient and public involvement group in directing this project to greatest clinical utility.

Contributors RN, JW and CPG conceived the idea of the study. JW undertook data extraction and statistical analysis. RN verified the underlying data. RN, JW and CPG interpreted the findings and RN drafted the manuscript. CB, JP, DH, CC, KR, YMN, KN, DZ, RA, MH, JW and CPG critically reviewed the manuscript and RN revised the manuscript for final submission. All authors had full access to all data in the study and accept responsibility to submit for publication. RN is the guarantor.

Funding RN is supported by the British Heart Foundation Clinical Research Training Fellowship (FS/20/12/34789).

Competing interests All authors have completed the ICMJE uniform disclosure form at www.icjme.org/col_disclosure.pdf and declare support from the British Heart Foundation for the submitted work. CPG has received grants for research from the British Heart Foundation, National Institute for Health Research, Horizon 2020, Abbott Diabetes, Bristol Myers Squibb and European Society of Cardiology; consulting fees from AstraZeneca, Bayer, Bristol Myers Squibb, Boehringer-Ingelheim, Boston Scientific, Chiesi, Daiichi Sankyo, Menarini and Organon; honoraria for speaking at meetings and educational events from AstraZeneca, Menarini, Novartis, Raisio Group, Wondr Medical and Zydus; support for attending meetings from Bayer and Bristol Myers Squibb; and has acted as an advisory board member for Amgen, AstraZeneca, Bayer, Bristol Myers Squibb, Boehringer-Ingelheim, Chiesi, Daiichi Sankyo, Menarini, Novartis and iRhythm. YMN has received advisory fees and a research grant from Bayer.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not required.

Ethics approval Ethical approval was granted by the Independent Scientific Advisory Committee (ISAC) of the Medicines and Healthcare Products Regulatory Agency (ref no: 19_076).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. Data used in this study can be accessed through CPRD subject to protocol approval. The algorithm can be shared with researchers who agree to use it only for research purposes with a data sharing agreement.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Ramesh Nadarajah http://orcid.org/0000-0001-9895-9356
Jianhua Wu http://orcid.org/0000-0001-6093-599X
Yoko M Nakao http://orcid.org/0000-0002-3627-5626
Ronen Arbel http://orcid.org/0000-0002-6058-8665
Chris Bates http://orcid.org/0000-0003-0113-2593

REFERENCES

1 Wu J, Nadarajah R, Nakao YM, *et al*. Temporal trends and patterns in atrial fibrillation incidence: a population-based study of 3·4 million individuals. *Lancet Reg Health Eur* 2022;17:100386.
2 Svennberg E, Engdahl J, Al-Khalili F, *et al*. Mass screening for untreated atrial fibrillation: the STROKESTOP study. *Circulation* 2015;131:2176–84.
3 Kamel H. Cryptogenic stroke and atrial fibrillation. *N Engl J Med* 2014;371:1261–2.
4 Ruff CT, Giugliano RP, Braunwald E, *et al*. Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: a meta-analysis of randomised trials. *Lancet* 2014;383:955–62.
5 Kirchhof P, Camm AJ, Goette A, *et al*. Early rhythm-control therapy in patients with atrial fibrillation. *N Engl J Med* 2020;383:1305–16.
6 NHS. Cardiovascular disease. 2019. Available: https://www.longtermplan.nhs.uk/areas-of-work/cardiovascular-disease/
7 Hindricks G, Potpara T, Dagres N, *et al*. 2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the european association for cardio-thoracic surgery (EACTS): the task force for the diagnosis and management of atrial fibrillation of the european society of cardiology (ESC) developed with the special contribution of the european heart rhythm association (EHRA) of the ESC. *Eur Heart J* 2021;42:373–498.
8 Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data resource profile: clinical practice research Datalink (CPRD). *Int J Epidemiol* 2015;44:827–36.
9 Himmelreich JCL, Lucassen WAM, Harskamp RE, *et al*. CHARGE-AF in a national routine primary care electronic health records database in the Netherlands: validation for 5-year risk of atrial fibrillation and implications for patient selection in atrial fibrillation screening. *Open Heart* 2021;8:e001459.
10 Nadarajah R, Alsaeed E, Hurdus B, *et al*. Prediction of incident atrial fibrillation in community-based electronic health records: a systematic review with meta-analysis. *Heart* 2022;108:1020–9.
11 Hill NR, Ayoubkhani D, McEwan P, *et al*. Predicting atrial fibrillation in primary care using machine learning. *PLoS One* 2019;14:e0224582.
12 Ruigómez A, Johansson S, Wallander MA, *et al*. Incidence of chronic atrial fibrillation in general practice and its treatment pattern. *J Clin Epidemiol* 2002;55:358–63.
13 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD group. *Circulation* 2015;131:211–9.
14 Kotecha D, Asselbergs FW, Achenbach S, *et al*. CODE-EHR best practice framework for the use of structured electronic healthcare records in clinical research. *BMJ* 2022;378:e069048.
15 Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
16 Routen A, Akbari A, Banerjee A, *et al*. Strategies to record and use ethnicity information in routine health data. *Nat Med* 2022;28:1338–42.
17 Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagn Progn Res* 2020;4:8:8.:.
18 Sakamoto Y, Ishiguro M, Kitagawa G. Akaike information criterion statistics. *Dordrecht Netherlands D Reidel* 1986;81:26853.
19 Szymanski T, Ashton R, Sekelj S, *et al*. Budget impact analysis of a machine learning algorithm to predict high risk of atrial fibrillation among primary care patients. *Europace* 2022;24:1240–7.
20 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;394:861–7.
21 van Smeden M, Heinze G, Van Calster B, *et al*. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. *Eur Heart J* 2022;43:2921–30.
22 Ibrahim H, Liu X, Zariffa N, *et al*. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health* 2021;3:e260–5.
23 Middeldorp ME, Pathak RK, Meredith M, *et al*. Prevention and regressive effect of weight-loss and risk factor modification on atrial fibrillation: the REVERSE-AF study. *Europace* 2018;20:1929–35.
24 Jones NR, Taylor CJ, Hobbs FDR, *et al*. Screening for atrial fibrillation: a call for evidence. *Eur Heart J* 2020;41:1075–85.
25 Uittenbogaart SB, Verbiest-van Gurp N, Lucassen WAM, *et al*. Opportunistic screening versus usual care for detection of atrial fibrillation in primary care: cluster randomised controlled trial. *BMJ* 2020;370:m3208.
26 Svennberg E, Friberg L, Frykman V, *et al*. Clinical outcomes in systematic screening for atrial fibrillation (STROKESTOP): a multicentre, parallel group, unmasked, randomised controlled trial. *Lancet* 2021;398:1498–506.
27 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* 2021;398:1507–16.

Supplementary Appendix

Prediction of short-term atrial fibrillation risk using primary care electronic health records

Ramesh Nadarajah*, Jianhua Wu*, David Hogg, Keerthenan Raveendra, Yoko M Nakao, Kazuhiro Nakao, John Parry, Chris Bates, Ronen Arbel, Moti Haim, Doron Zahger, Campbell Cowan, Chris P Gale

Supplementary Introduction

Supplementary Table S1. Algorithms that have been derived and/or validated in community-based EHR for predicting AF

Algorithm	**Study Aim**	**Study**	**EHR cohort (country)**	**Age eligibility (years)**	**Discrimination (c-statistic)**	**Follow-up**	**Variable frequently missing in routinely-collected primary care EHR**
Models originally derived for another purpose but tested for prediction of incident atrial fibrillation							
CHADS$_2$	EV	Chao 2013	NHIRD (TW)	≥18	0.713	10	N/A
	EV	Saliba 2016	ClalitHS (IL)	≥50	0.728	3	
	EV	Li 2019	YMID (CN)	≥18	0.632	11	
	EV	Li 2019	NHIS-HEALS (KR)	≥18	0.637	11	
	EV	Kim 2020	NHIS-NSC (KR)	≥18	0.652	5	
CHA$_2$DS$_2$-VASc	EV	Saliba 2016	ClalitHS (IL)	≥50	0.744	3	N/A
	EV	Li 2019	YMID (CN)	≥18	0.687	11	
	EV	Li 2019	NHIS-HEALS (KR)	≥18	0.637	11	
	EV	Himmelreich 2020	Nivel-PCD (NL)	≥40	0.669	5	
	EV	Kim 2020	NHIS-NSC (KR)	≥18	0.654	5	
HATCH	EV	Suenari 2017	NHIRD (TW)	≥20	0.716	9	N/A
	EV	Li 2019	YMID (CN)	≥18	0.633	11	
	EV	Li 2019	NHIS-HEALS (KR)	≥18	0.646	11	
	EV	Kim 2020	NHIS-NSC (KR)	≥18	0.669	5	
	EV	Hu-WS 2020	NHIRD (TW)	≥18	0.771	14	
Machine Learning models							
Pfizer-AI	D	Hill 2019	CPRD (UK)	≥30	0.827	11	Height, weight, BMI, SBP, DBP
	EV	Sekelj 2020	Discover (UK)	≥30	0.870	8	
NHIRD	D	Hu-WS 2019	NHIRD (TW)	≥18	0.948	14	Follow-up duration (years)
NHIS-NSC	D	Kim 2020	NHIS-NSC (KR)	≥18	0.845	5	BMI, SBP, Triglycerides, total cholesterol, HDL cholesterol, LDL cholesterol, eGFR, GGT, fasting blood glucose, Haemoglobin, AST, Socioeconomic status
Regression Models derived in electronic health records							

C_2HEST	D	Li 2019	YMID (CN)	≥18	0.750	11	N/A
	EV	Li 2019	NHIS-HEALS (KR)	≥18	0.654	11	
	EV	Hu-WS 2020	NHIRD (TW)	≥18	0.790	14	
	EV	Lip 2020	DCRS, DNPR, DPR (DK)	65	0.588	5	
				70	0.594		
				75	0.593		
MHS	D	Aronson 2018	MHS (IL)	≥50	0.743	10	BMI, SBP
Taiwan AF	D	Chao 2021	NHIRD (TW)	≥40	0.857	1	N/A
					0.825	5	
					0.797	10	
					0.756	16	
InGef	D	Schnabel 2022	InGef (G)	≥45	0.829	1	N/A
Regression model derived in a prospective cohort design							
CHARGE-AF	EV	Hill 2019	CPRD (UK)	≥30	0.725	11	Height, weight, SBP, DBP

Supplementary Table S2. Algorithms that have been derived and/or validated in European community-based EHRs for predicting AF

Algorithm	**Study Aim**	**Study**	**EHR cohort (country)**	**Age eligibility (years)**	**Discrimination (c-statistic)**	**Follow-up**	**Variable frequently missing in routinely-collected primary care EHR**
Models originally derived for another purpose but tested for prediction of incident atrial fibrillation							
CHA$_2$DS$_2$-VASc	EV	Himmelreich 2020	Nivel-PCD (NL)	≥40	0.669	5	N/A
Machine Learning models							
CPRD	D	Hill 2019	CPRD (UK)	≥30	0.827	11	Height, weight, BMI, SBP, DBP
	EV	Sekelj 2020	Discover (UK)	≥30	0.870	8	
Regression Models derived in electronic health records							
C$_2$HEST	EV	Lip 2020	DCRS, DNPR, DPR (DK)	65	0.588	5	N/A
				70	0.594		
				75	0.593		
InGef	D	Schnabel 2022	InGef (G)	≥45	0.829	1	N/A
Regression model derived in a prospective cohort design							
CHARGE-AF	EV	Hill 2019	CPRD (UK)	≥30	0.725	11	Height, weight, SBP, DBP

AF, Atrial Fibrillation; CHADS$_2$, Congestive heart failure, Hypertension, Age >75, Diabetes mellitus, prior Stroke or transient ischemic attack [2 points]; CHA$_2$DS$_2$-VASc, Congestive heart failure, Hypertension, Age >75 [2 points], Stroke/transient ischemic attack/thromboembolism [2 points], Vascular disease, Age 65-74, Sex Category; CHARGE-AF, Cohorts for Heart and Aging Research in Genomic Epidemiology; C$_2$HEST, Coronary artery disease / Chronic obstructive pulmonary disease [1 point each], Hypertension, Elderly (Age ≥75, 2 points), Systolic heart failure, Thyroid disease (hyperthyroidism); ClalitHS, Clalit Health Services; CPRD, Clinical Practice Research Datalink; D, derivation; DCRS, Danish Civil Registration system; DK, Denmark; DNPR, Danish National Patient Register; DPR, Danish Prescription Regster; EHR, electronic health record; EV, external validation; G, Germany; HATCH, Hypertension, Age, stroke or Transient ischemic attack, Chronic obstructive pulmonary disease, Heart failure; IL, Israel; KR, Republic of Korea; MHS, Maccabi Healthcare Services; NHIRD, National Health Insurance Research Database; NHIS-HEALS, National Health Insurance Service - Health screening Cohort; NHIS-NSC, National Health Insurance Service-based National Sample Cohort; Nivel-PCD, Netherlands Institute for Health Services Research Primary Care Database; NL, Netherlands; TW, Taiwan; UK, United Kingdom; YMID, Yunnan Medical Insurance Database.

Supplementary Methods

Supplementary Table S3. Read codes and ICD-10 codes used to define the outcomes of atrial fibrillation or atrial flutter

Code	**Description**
Readcodes	
G573200	Paroxysmal atrial fibrillation
G573400	Permanent atrial fibrillation
G573500	Persistent atrial fibrillation
3272	ECG: atrial fibrillation
G573000	Atrial fibrillation
G573300	Non-rheumatic atrial fibrillation
G573.00	Atrial fibrillation and flutter
G573z00	Atrial fibrillation and flutter NOS
3273	ECG: atrial flutter
G573100	Atrial flutter
ICD-10 codes	
I48	Atrial fibrillation and flutter

Training of the Random Forest classifier

Each decision tree used Gini impurity, commonly used in classification and regression tree (CART) algorithms, to measure the split quality.[1] The minimum impurity split threshold for each node, above which a node will split into two or more branches, was set to 10^{-7}. The minimum number of samples required to split a node was set to two. The minimum samples per leaf was set to one. All the algorithm's hyperparameters were tuned using the grid search method, in which all possible combinations were evaluated, resulting in 1000 trees, mtry = 8 (the number of random features to consider in each tree) and nodesize = 12 (number of patients classified at that node).

Supplementary Table S4. Baseline demographic and comorbidity variables used in algorithms tested for predicting incident AF in community-based electronic health records

Algorithm	**Demographics**	**Comorbidities**
$CHADS_2$	Age	Hypertension, CHF, diabetes mellitus, CVA
CHA_2DS_2-VASc	Age, sex	Hypertension, CHF, stroke/TIA/thromboembolism, vascular disease
CHARGE-AF	Age, race, smoking status	Anti-hypertensive medication, MI, CHF, DM
C_2HEST	Age	Hypertension, ischaemic heart disease, CHF, COPD, thyroid disease
HATCH	Age	Hypertension, CHF, stroke/TIA, COPD
InGef	Age, sex	Anti-hypertension medication, heart failure medication, chronic kidney disease, disorderd of lipoprotein metabolism and other lipidaemias, pulmonary heart diseases cardiac arrhythmias, other cerebrovascular disease, diverticular disease of intestine, dorsalgia, breathing abnormalities
MHS	Age, sex	Anti-hypertensive medication, MI, CHF, peripheral vascular disease, inflammatory disease in a female, COPD
NHIRD	Age (years), age group, sex	Hypertension, CHF, COPD, rheumatological disease, dyslipidaemia, DM, CVA or TIA, sleep disorder, cancer, hyperthyroidism, vascular disease, gout, CKD or ESRD, anaemia
NHIS-NSC*	Age, sex, smoking (pack-year), alcohol	Hypertension, CHF, MI, vascular disease, stroke/TIA, COPD
Pfizer-AI	Age, sex, race, smoking status	Hypertension, anti-hypertensive medication, CHF, congenital heart disease, MI, LVH, type 1 DM, type 2 DM
Taiwan AF	Age, sex, alcohol excess	Hypertension, CHF, IHD, ESRD

AF, Atrial Fibrillation; $CHADS_2$, Congestive heart failure, Hypertension, Age >75, Diabetes mellitus, prior Stroke or transient ischemic attack [2 points]; CHA_2DS_2-VASc, Congestive heart failure, Hypertension, Age >75 [2 points], Stroke/transient ischemic attack/thromboembolism [2 points]; CHARGE-AF, Cohorts for Heart and Aging Research in Genomic Epidemiology; C_2HEST, Coronary artery disease / Chronic obstructive pulmonary disease [1 point each], Hypertension, Elderly (Age ≥75, 2 points), Systolic heart failure, Thyroid disease (hyperthyroidism); CHF, chronic heart failure; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; CPRD, Clinical Practice Research Datalink; CVA, cerebrovascular accident; DM, diabetes mellitus; ESRD, end-stage renal disease; HATCH, Hypertension, Age, stroke or Transient ischemic attack, Chronic obstructive pulmonary disease, Heart failure; IHD, ischaemic heart disease; LVH, left ventricular hypertrophy; MHS, Maccabi Healthcare Services; MI, myocardial infarction; NHIRD, National Health Insurance Research Database; NHIS-HEALS, National Health Insurance Service - Health screening Cohort; NHIS-NSC, National Health Insurance Service-based National Sample Cohort; TIA, transient ischaemic attack.

* In Kim 2020 prediction model development using machine learning was completed both with and without the predictor $PM_{2.5}$ - which is fine particular matter air pollution. In this analysis we have only included the model without $PM_{2.5}$ as it is judged not to be a predictor that would be routinely available in primary care or population EHR.

Supplementary Table S5. Candidate variables added after literature search with accompanying reference demonstrating association

Comorbidity associated with / predictive of atrial fibrillation	**Categorisation**	**Reference demonstrating association with AF and rationale for categorisation**
Cardiac surgery	Valvular,	Greenberg JW, Lancaster TS, Schuessler RB, et al. Postoperative atrial fibrillation following cardiac surgery: a persistent complication. Eur J Cardiothorac Surg 2017;52(4):665-72. Within overall cardiac surgical procedures incidence of post-operative AF is 35%, isolated CABG has an incidence of 20—30% and isolated valve surgeries have an incidence of 35-40
	Non-valvular	
Deep venous thrombosis	-	Lutsey P, Norby F, Alonso A, et al. Atrial fibrillation and venous thromboembolism: evidence of bidirectionality in the Atherosclerosis Risk in Communities Study. J Thromb Haemost 2018;16(4):670-79.
Infective Endocarditis	-	Ferrera C, Vilacosta I, Fernández C, et al. Usefulness of new-onset atrial fibrillation, as a strong predictor of heart failure and death in patients with native left-sided infective endocarditis. The American journal of cardiology 2016;117(3):427-33.
Electrophysiology procedure affecting the atria	-	Strickberger SA, Man KC, Daoud EG, et al. Adenosine-induced atrial arrhythmia: a prospective analysis. Ann Intern Med 1997;127(6):417-22. Khachab, H., and B. Brembilla-Perrot. "Prevalence of atrial fibrillation in patients with history of paroxysmal supraventricular tachycardia." International journal of cardiology 166.1 (2013): 221-224.
Hypertrophic cardiomyopathy	-	Siontis KC, Geske JB, Ong K, et al. Atrial fibrillation in hypertrophic cardiomyopathy: prevalence, clinical correlations, and mortality in a large high-risk population. Journal of the American Heart Association 2014;3(3):e001002.
Inflammatory bowel disease	-	Boos CJ. Infection and atrial fibrillation: inflammation begets AF. Eur Heart J 2020
Intensive care unit admission	-	Klein Klouwenberg PM, Frencken JF, Kuipers S, et al. Incidence, predictors, and outcomes of new-onset atrial fibrillation in critically ill patients with sepsis. A cohort study. Am J Respir Crit Care Med 2017;195(2):205-11.
Infection	Gastrointestinal	Gundlund A, Olesen JB, Butt JH, et al. One-year outcomes in atrial fibrillation presenting during infections: a nationwide registry-based study. Eur Heart J 2020;41(10):1112-19. Chang T-Y, Chao T-F, Liu C-J, et al. The association between influenza infection, vaccination, and atrial fibrillation: A nationwide case-control study. Heart Rhythm 2016;13(6):1189-94. Klein Klouwenberg PM, Frencken JF, Kuipers S, et al. Incidence, predictors, and outcomes of new-onset atrial fibrillation in critically ill patients with sepsis. A cohort study. Am J Respir Crit Care Med
	Influenza	
	Respiratory	
	Sepsis	

		2017;195(2):205-11. In a cohort study among infections precipitating AF the order of risk is as follows: Pneumonia > sepsis > urinary tract infection > gastrointestinal infection
	Urinary	
Myocarditis	-	Wang Z, Wang Y, Lin H, et al. Early characteristics of fulminant myocarditis vs non-fulminant myocarditis: a meta-analysis. Medicine 2019;98(8)
Pulmonary embolus	-	Ptaszynska-Kopczynska K, Kiluk I, Sobkowicz B. Atrial fibrillation in patients with acute pulmonary embolism: clinical significance and impact on prognosis. BioMed research international 2019;2019
Pericarditis	-	Imazio M, Lazaros G, Picardi E, et al. Incidence and prognostic significance of new onset atrial fibrillation/flutter in acute pericarditis. Heart 2015;101(18):1463-67.
Pulmonary hypertension	-	Olsson KM, Nickel NP, Tongers J, et al. Atrial flutter and fibrillation in patients with pulmonary hypertension. Int J Cardiol 2013;167(5):2300-05.
Surgery (non-cardiac)	Colorectal	Siu CW, Tung HM, Chu KW, et al. Prevalence and predictors of new-onset atrial fibrillation after elective surgery for colorectal cancer. Pacing Clin Electrophysiol 2005;28:S120-S23. Onaitis M, D'Amico T, Zhao Y, et al. Risk factors for atrial fibrillation after lung cancer surgery: analysis of the Society of Thoracic Surgeons general thoracic surgery database. The Annals of thoracic surgery 2010;90(2):368-74. Philip I, Berroëta C, Leblanc I. Perioperative challenges of atrial fibrillation. Current Opinion in Anesthesiology 2014;27(3):344-52. Thoracic surgery is associated with the greatest risk of post-operative AF amongst non-cardiac surgeries followed by colorectal then vascular surgery
	Thoracic	
	Vascular	
Valvular heart disease	Mitral stenosis / rheumatic valvular disease	Iung B, Leenhardt A, Extramiana F. Management of atrial fibrillation in patients with rheumatic mitral stenosis. Heart 2018;104(13):1062-68. Levy S. Factors predisposing to the development of atrial fibrillation. Pacing Clin Electrophysiol 1997;20(10):2670-74. Grigioni F, Avierinos J-F, Ling LH, et al. Atrial fibrillation complicating the course of degenerative mitral regurgitation: determinants and long-term outcome. J Am Coll Cardiol 2002;40(1):84-92. The association of mitral stenosis and rheumatic valve disease with AF is greater than mitral regurgitation followed by diseases of other valves
	Non-mitral valve / other valves	
	Mitral regurgitation	
Vascular dementia	-	Ott A, Breteler MM, De Bruyne MC, et al. Atrial fibrillation and dementia in a population-based study: the Rotterdam Study. Stroke 1997;28(2):316-21.
Weight	Obese	Lavie CJ, Pandey A, Lau DH, et al. Obesity and atrial fibrillation prevalence, pathogenesis, and prognosis:
	Overweight	

	Under-weight	effects of weight loss and exercise. J Am Coll Cardiol 2017;70(16):2022-35. Frost L, Hune LJ, Vestergaard P. Overweight and obesity as risk factors for atrial fibrillation or flutter: the Danish Diet, Cancer, and Health Study. The American journal of medicine 2005;118(5):489-95. Lee S-R, Choi E-K, Park CS, et al. Direct oral anticoagulants in patients with nonvalvular atrial fibrillation and low body weight. J Am Coll Cardiol 2019;73(8):919-31. Obesity is associated with a greater risk of AF than being overweight. Low body weight is associated with a higher risk of AF than normal weight.

Supplementary Table S6. Variable categorisations with rationale

Comorbidity associated with / predictive of atrial fibrillation	**Categorisation**	**References and Rationale for categorisation**
Demographics		
Age	-	Hindricks G, Potpara T, Dagres N, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association of Cardio-Thoracic Surgery (EACTS). Eur Heart J 2020 Incidence of AF increases with age (therefore included as a continuous variable)
Sex	Men	Hindricks G, Potpara T, Dagres N, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association of Cardio-Thoracic Surgery (EACTS). Eur Heart J 2020 AF is more common in men
	Women	
Ethnicity	Asian	Shen AY-J, Contreras R, Sobnosky S, et al. Racial/ethnic differences in the prevalence of atrial fibrillation among older adults—a cross-sectional study. J Natl Med Assoc 2010;102(10):906-14. Chiang C-E, Zhang S, Tse HF, et al. Atrial fibrillation management in Asia: from the Asian expert forum on atrial fibrillation. Int J Cardiol 2013;164(1):21-32. White, Asian, pacific Asian, and black ethnicities have different odds ratios of development of AF
	Black	
	Mixed	
	Other	
	Pacific Asian	
	White	
Alcohol use	Ex-	Samokhvalov AV, Irving HM, Rehm J. Alcohol consumption as a risk factor for atrial fibrillation: a systematic review and meta-analysis. European Journal of Preventive Cardiology 2010;17(6):706-12. There is a monotonic dose-response relationship between alcohol consumption and AF incidence
	Light,	
	Moderate	
	Excess	
	Unspecified	
Smoking	Current	Heeringa J, Kors JA, Hofman A, et al. Cigarette smoking and risk of atrial fibrillation: the Rotterdam Study. Am Heart J 2008;156(6):1163-69. Watanabe I. Smoking and risk of atrial fibrillation: Elsevier, 2018. Current and ex-smokers are at increased risk of AF, with a higher risk in current smokers.
	Ex	
Weight	Obese	See table S4
	Overweight	
	Under-weight	
Comorbidities		
Adult congenital heart disease	-	-
Anaemia	-	-
Cancer	Leukaemia	Thompson PA, Lévy V, Tam CS, et al. Atrial fibrillation in CLL patients treated with ibrutinib. An international retrospective study. Br J Haematol 2016;175(3):462-66.
	Lymphoma	
	Metastasis	

	Skin cancers other than melanoma	Sorigue M, Gual-Capllonch F, Garcia O, et al. Incidence, predictive factors, management, and survival impact of atrial fibrillation in non-Hodgkin lymphoma. Ann Hematol 2018;97(9):1633-40. Han H, Chen L, Lin Z, et al. Prevalence, trends, and outcomes of atrial fibrillation in hospitalized patients with metastatic cancer: findings from a national sample. Cancer medicine 2021;10(16):5661-70. AF risk is higher in patients with leukaemia and lymphoma, especially treated with iritunib. Solid organ cancers (such as lung and colorectal cancer) are more likely to undergo surgery. Metastatic disease is associated with higher risk of AF compared to non-metastatic disease. Skin cancers other than melanoma have a lower risk of metastasis and hence AF.
	Solid organ	
Cardiac surgery	Valvular,	See table S4
	Non-valvular	
Chronic kidney disease	Stage 1-2	Alonso A, Lopez FL, Matsushita K, et al. Chronic kidney disease is associated with the incidence of atrial fibrillation: the Atherosclerosis Risk in Communities (ARIC) study. Circulation 2011;123(25):2946-53. Risk of AF increases as CKD stage worsens and if there is proteinuria
	Stage 3	
	Stage 4	
	Stage 5	
	Unspecified	
	Other	
COPD	-	-
Cerebro-vascular accident	Intracerebral haemorrhage	Hindricks G, Potpara T, Dagres N, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association of Cardio-Thoracic Surgery (EACTS). Eur Heart J 2020 Association with AF is higher for ischaemic strokes than haemorrhagic strokes
	Subarachnoid haemorrhage	
	Unspecified	
Diabetes Mellitus	Good control	Dublin S, Glazer NL, Smith NL, et al. Diabetes mellitus, glycemic control, and risk of atrial fibrillation. J Gen Intern Med 2010;25(8):853-58. Poorer glycaemic control is associated with a higher risk of AF compared to better glycaemic control or no diabetes
	Poor control	
	Unspecified / secondary	
Deep venous thrombosis	-	-
Dyslipidaemia	-	-
Infective Endocarditis	-	-
Electrophysiology procedure affecting the atria	-	-
Gout	-	-
Hypertrophic cardiomyopathy	-	-
Heart failure	-	-
Hypertension	Poor control	Dzeshka MS, Shantsila A, Shantsila E, et al. Atrial fibrillation and hypertension. Hypertension 2017;70(5):854-61. Poorer control of hypertension and end organ damage is
	Unspecified / secondary	

		associated with a higher risk of developing AF
Hyperthyroidism	-	-
Inflammatory bowel disease	-	-
Intensive care unit admission	-	-
Ischaemic heart disease	Chronic	Huxley RR, Lopez FL, Folsom AR, et al. Absolute and attributable risks of atrial fibrillation in relation to optimal and borderline risk factors: the Atherosclerosis Risk in Communities (ARIC) study. Circulation 2011;123(14):1501-08. Pizzetti F, Turazza F, Franzosi M, et al. Incidence and prognostic significance of atrial fibrillation in acute myocardial infarction: the GISSI-3 data. Heart 2001;86(5):527-32. There is a high risk of AF in the acute setting of myocardial infarction as well as evidence in the context of underlying chronic coronary syndromes.
	Myocardial infarction	
	Percutaneous coronary intervention	
Infection	Gastrointestinal	See table S4
	Influenza	
	Respiratory	
	Sepsis	
	Urinary	
Left ventricular hypertrophy	-	-
Myocarditis	-	-
Obstructive sleep apnoea	-	-
Pulmonary embolus	-	-
Pericarditis	-	-
Pulmonary hypertension	-	-
Peripheral vascular disease	-	-
Rheumatological condition	Autoimmune connective tissue diseases	Lee E, Choi E-K, Jung J-H, et al. Increased risk of atrial fibrillation in patients with Behçet's disease: a nationwide population-based study. Int J Cardiol 2019;292:106-11. Moon I, Choi E-K, Jung J-H, et al. Ankylosing spondylitis: a novel risk factor for atrial fibrillation—a nationwide population-based study. Int J Cardiol 2019;275:77-82. Melduni RM, Cooper LT, Gersh BJ, et al. Association of Autoimmune Vasculitis and Incident Atrial Fibrillation: A Population-Based Case-Control Study. Journal of the American Heart Association 2020;9(18):e015977. Naaraayan A, Meredith A, Nimkar A, et al. Arrhythmia prevalence among patients with Polymyositis-Dermatomyositis in the United States: an observational study. Heart Rhythm 2021 Songnan W, Shengma C. GW24-e2483 Catheter ablation of atrial fibrillation in patients with autoimmune rheumatic diseases. Heart 2013;99(Suppl 3):A197-A97. Giallafos I, Triposkiadis F, Oikonomou E, et al. Incident
	Rheumatoid arthritis	
	Spondyloarthropathies	
	Vasculitides	

		atrial fibrillation in systemic sclerosis: the predictive role of B-type natriuretic peptide. Hellenic J Cardiol 2014;55:313-21. Pugnet G, Gouya H, Puéchal X, et al. Cardiac involvement in granulomatosis with polyangiitis: a magnetic resonance imaging study of 31 consecutive patients. Rheumatology 2017;56(6):947-56. Lindhardsen J, Ahlehoff O, Gislason GH, et al. Risk of atrial fibrillation and stroke in rheumatoid arthritis: Danish nationwide cohort study. BMJ 2012;344 Each of the subtypes of rheumatological disease are associated with differing risks of development of AF. Here they have been categorised in clinical sub-type.
Smoking	Current	See table S4
	Ex	
Surgery (non-cardiac)	Colorectal	See table S4
	Thoracic	
	Vascular	
Systemic Embolism	-	-
Valvular heart disease	Mitral stenosis / rheumatic valvular disease	See table S4
	Non-mitral valve / other valves	
	Mitral regurgitation	
Vascular dementia	-	-

Supplementary Results

Supplement Table S7. Baseline characteristics of training and testing datasets

	Training set **n (%)**	**Testing set** **n (%)**
	1 664 911	416 228
Demographics		
Age, years	49.90 (15.43)	49.90 (15.42)
Sex (women)	844 083 (50.7)	211 478 (50.8)
Comorbidities		
Diabetes mellitus	58 513 (3.5)	14 268 (3.4)
Stroke or TIA	30 871 (1.9)	7 794 (1.9)
Ischaemic heart disease	62 980 (3.8)	15 622 (3.8)
Hypertension	200 217 (12.0)	50 106 (12.0)
Heart failure	11 577 (0.7)	2 790 (0.7)
Dyslipidaemia	48 719 (2.9)	12 170 (2.9)
Hyperthyroidism	13 069 (0.8)	3 233 (0.8)
COPD	20 294 (1.2)	5 129 (1.2)
Chronic kidney disease	23 794 (1.4)	6 014 (1.4)
Anaemia	53 962 (3.2)	13 383 (3.2)
Cancer	58 725 (3.5)	14 783 (3.6)
Valvular heart disease	7 946 (0.5)	1 927 (0.5)
Mean CHA_2DS_2-VASc score	0.98 (1.04)	0.98 (1.04)

Supplementary Table S8. Net reclassification using FIND-AF

AF cases

	FIND-AF			**FIND-AF**	
CHA$_2$DS$_2$-VASc	≥0.4%	<0.4%	**C$_2$HEST**	≥0.4%	<0.4%
≥0.4%	1 121	37	≥0.4%	893	10
<0.4%	82	191	<0.4%	310	218

🟩 Appropriate upclassification

🟥 Inappropriate downclassification

Non-AF cases

	FIND-AF			**FIND-AF**	
CHA$_2$DS$_2$-VASc	≥0.4%	<0.4%	**C$_2$HEST**	≥0.4%	<0.4%
≥0.4%	65 322	17 511	≥0.4%	38 640	3 053
<0.4%	16 417	315 547	<0.4%	43 099	330 005

🟩 Appropriate downclassification

🟥 Inappropriate upclassification

Net reclassification indices

Index	**CHA$_2$DS$_2$-VASc**	**C$_2$HEST**
Case reclassification (NRI+ [95% CI])	0.031 (0.026-0.048)	0.021 (0.19-0.23)
Non-case reclassification (NRI- [95% CI])	0.0026 (0.0015-0.0032)	-0.096 (-0.098 - -0.095)
Net reclassification (NRI [95% CI])	0.032 (0.029-0.051)	0.113 (0.098-0.135)

Supplementary Table S9. Baseline characteristics of testing set, stratified by incident AF and predicted AF risk

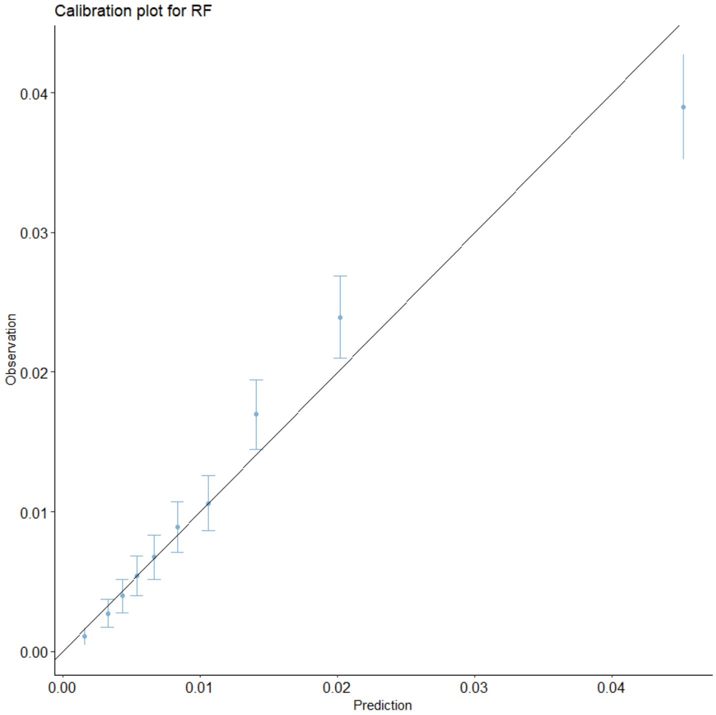
	Incident atrial fibrillation		**FIND-AF predicted risk**	
	no AF **n (%)**	**AF** **n (%)**	**Lower risk** **n (%)**	**Higher risk** **n (%)**
	414 676	1 552	333 286	82 942
Demographics				
Age, years	49.82 (15.38)	73.87 (12.47)	44.11 (10.40)	73.24 (8.75)
Sex (women)	210 646 (50.8)	755 (48.6)	170 568 (51.2)	41 210 (49.7)
Ethnicity				
Asian	8 258 (2.0)	21 (1.5)	7 385 (2.2)	894 (1.1)
Black	6 390 (1.5)	9 (0.6)	5 786 (1.7)	613 (0.7)
Other	27 805 (6.7)	106 (7.4)	22 033 (6.6)	5 878 (7.1)
Unknown	93 630 (22.6)	36 (2.5)	91 505 (27.5)	2 161 (2.6)
White	278 714 (67.2)	1 259 (88.0)	206 577 (62.0)	73 396 (88.5)
Comorbidities				
Diabetes mellitus	14 649 (3.5)	171 (11.0)	6 328 (1.9)	8 072 (9.7)
Stroke or TIA	7 467 (1.8)	189 (12.2)	1 376 (0.4)	6 375 (7.7)
Ischaemic heart disease	15 483 (3.7)	314 (20.2)	3 299 (1.0)	12 486 (15.1)
Hypertension	49 494 (11.9)	621 (40.0)	20 139 (6.0)	29 594 (35.7)
Heart failure	2 745 (0.7)	132 (8.5)	163 (0.0)	2 748 (3.3)
Dyslipidaemia	12 122 (2.9)	121 (7.8)	6 095 (1.8)	5 984 (7.2)
Hyperthyroidism	3 203 (0.8)	44 (2.8)	1 883 (0.6)	1 370 (1.7)
COPD	4 987 (1.2)	106 (6.8)	1 111 (0.3)	4 019 (4.8)
Chronic kidney disease	5 839 (1.4)	99 (6.4)	2 938 (0.9)	2 990 (3.6)

Anaemia	13 165 (3.2)	106 (6.8)	9118 (2.7)	4251 (5.1)
Cancer	14 710 (3.5)	186 (12.0)	6120 (1.8)	8303 (10.0)
Valvular heart disease	1 881 (0.5)	84 (5.4)	562 (0.2)	1414 (1.7)
Mean CHA_2DS_2-VASc score (SD)	0.97 (1.03)	2.74 (1.40)	0.62 (0.62)	2.42 (1.14)

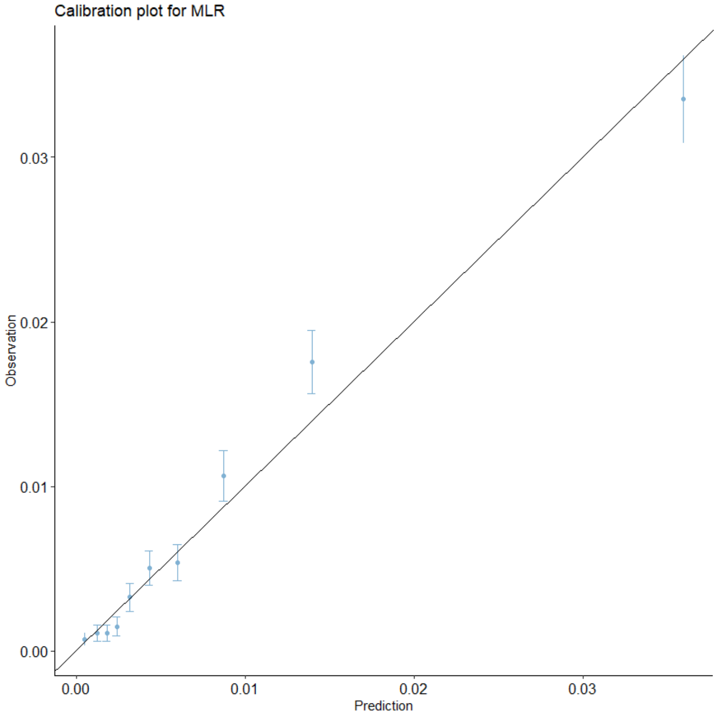
AF, atrial fibrillation; CHA_2DS_2-VASc, Congestive heart failure, Hypertension, Age >75 years [2 points], Stroke/transient ischemic attack/thromboembolism [2 points], Vascular disease, Age 65-74 years, Sex Category; COPD, chronic obstructive pulmonary disease; TIA, transient ischaemic attack

Supplement Figure S1. Calibration plots

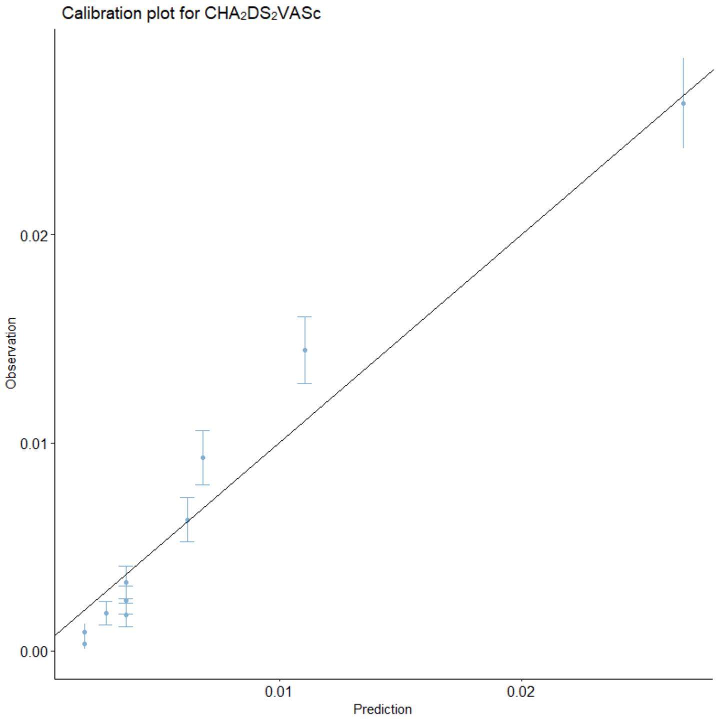
FIND-AF



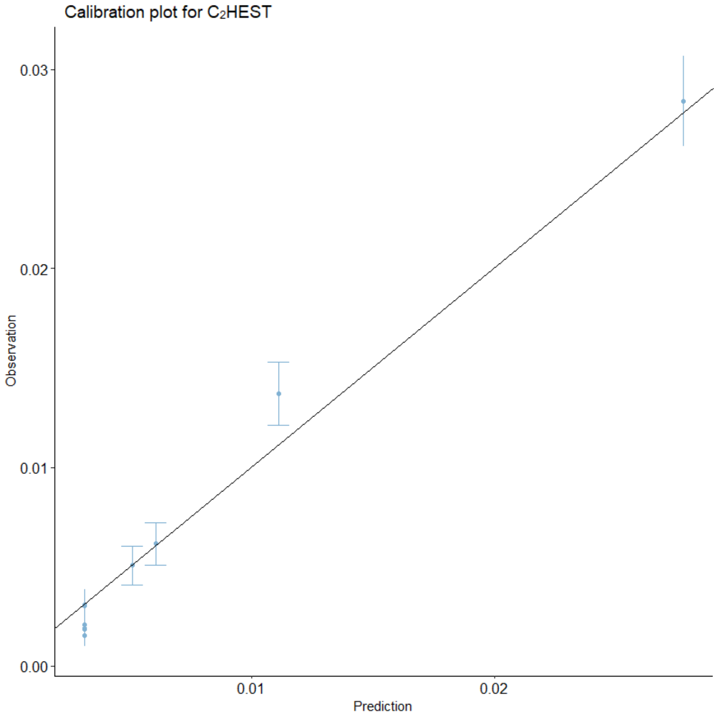
Multivariable logistic regression



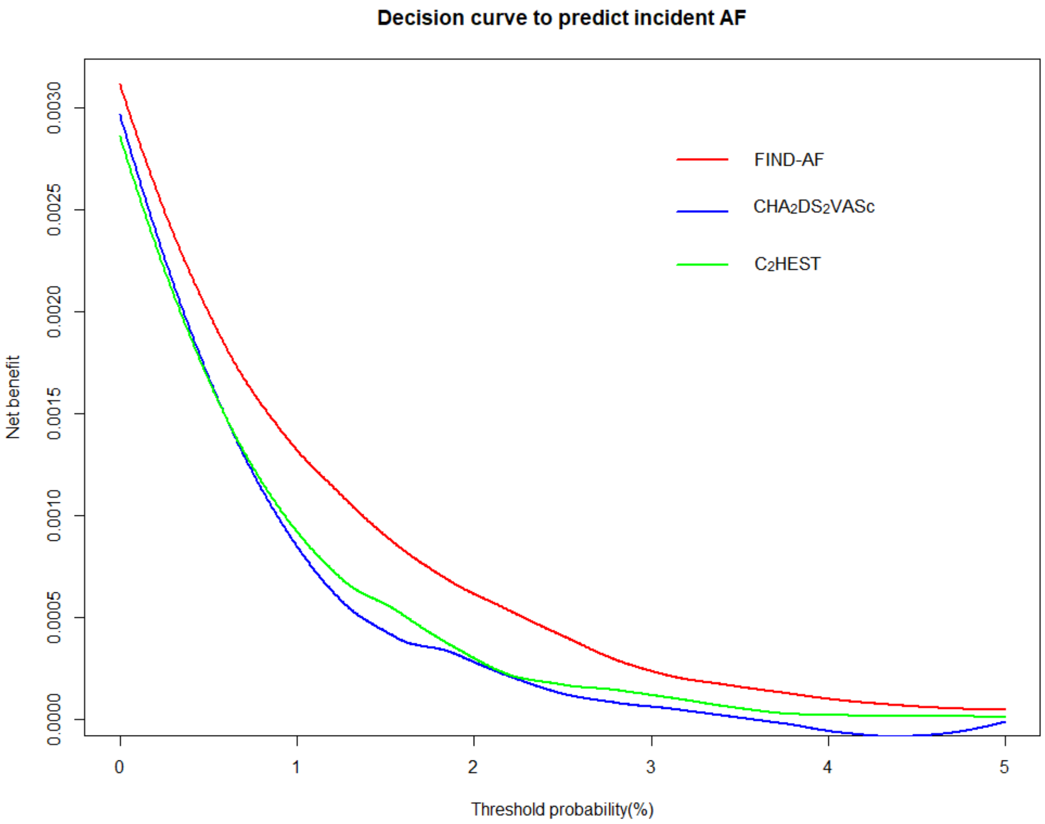
CHA_2DS_2VASc



C_2HEST



Supplementary Figure S2. Decision curve analysis for FIND-AF versus CHA_2DS_2-VASc and C_2HEST



References

1. Raileanu LE, Stoffel K. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence* 2004;41(1):77-93.