

Deep Learning of Electrocardiograms in Sinus Rhythm From US Veterans to Predict Atrial Fibrillation

Neal Yuan, MD; Grant Duffy, BS; Sanket S. Dhruva, MD, MHS; Adam Oesterle, MD; Cara N. Pellegrini, MD; John Theurer, BS; Marzieh Vali, MS; Paul A. Heidenreich, MD; Salomeh Keyhani, MD; David Ouyang, MD

IMPORTANCE Early detection of atrial fibrillation (AF) may help prevent adverse cardiovascular events such as stroke. Deep learning applied to electrocardiograms (ECGs) has been successfully used for early identification of several cardiovascular diseases.

OBJECTIVE To determine whether deep learning models applied to outpatient ECGs in sinus rhythm can predict AF in a large and diverse patient population.

DESIGN, SETTING, AND PARTICIPANTS This prognostic study was performed on ECGs acquired from January 1, 1987, to December 31, 2022, at 6 US Veterans Affairs (VA) hospital networks and 1 large non-VA academic medical center. Participants included all outpatients with 12-lead ECGs in sinus rhythm.

MAIN OUTCOMES AND MEASURES A convolutional neural network using 12-lead ECGs from 2 US VA hospital networks was trained to predict the presence of AF within 31 days of sinus rhythm ECGs. The model was tested on ECGs held out from training at the 2 VA networks as well as 4 additional VA networks and 1 large non-VA academic medical center.

RESULTS A total of 907 858 ECGs from patients across 6 VA sites were included in the analysis. These patients had a mean (SD) age of 62.4 (13.5) years, 6.4% were female, and 93.6% were male, with a mean (SD) CHA₂DS₂-VASc (congestive heart failure, hypertension, age, diabetes mellitus, prior stroke or transient ischemic attack or thromboembolism, vascular disease, age, sex category) score of 1.9 (1.6). A total of 0.2% were American Indian or Alaska Native, 2.7% were Asian, 10.7% were Black, 4.6% were Latinx, 0.7% were Native Hawaiian or Other Pacific Islander, 62.4% were White, 0.4% were of other race or ethnicity (which is not broken down into subcategories in the VA data set), and 18.4% were of unknown race or ethnicity. At the non-VA academic medical center (72 483 ECGs), the mean (SD) age was 59.5 (15.4) years and 52.5% were female, with a mean (SD) CHA₂DS₂-VASc score of 1.6 (1.4). A total of 0.1% were American Indian or Alaska Native, 7.9% were Asian, 9.4% were Black, 2.9% were Latinx, 0.03% were Native Hawaiian or Other Pacific Islander, 74.8% were White, 0.1% were of other race or ethnicity, and 4.7% were of unknown race or ethnicity. A deep learning model predicted the presence of AF within 31 days of a sinus rhythm ECG on held-out test ECGs at VA sites with an area under the receiver operating characteristic curve (AUROC) of 0.86 (95% CI, 0.85-0.86), accuracy of 0.78 (95% CI, 0.77-0.78), and F1 score of 0.30 (95% CI, 0.30-0.31). At the non-VA site, AUROC was 0.93 (95% CI, 0.93-0.94); accuracy, 0.87 (95% CI, 0.86-0.88); and F1 score, 0.46 (95% CI, 0.44-0.48). The model was well calibrated, with a Brier score of 0.02 across all sites. Among individuals deemed high risk by deep learning, the number needed to screen to detect a positive case of AF was 2.47 individuals for a testing sensitivity of 25% and 11.48 for 75%. Model performance was similar in patients who were Black, female, or younger than 65 years or who had CHA₂DS₂-VASc scores of 2 or greater.

CONCLUSIONS AND RELEVANCE Deep learning of outpatient sinus rhythm ECGs predicted AF within 31 days in populations with diverse demographics and comorbidities. Similar models could be used in future AF screening efforts to reduce adverse complications associated with this disease.

JAMA Cardiol. doi:10.1001/jamacardio.2023.3701
Published online October 18, 2023.

 [Invited Commentary](#)

 [Supplemental content](#)

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Neal Yuan, MD, Department of Medicine, University of California, San Francisco, 4150 Clement St, San Francisco, CA 94121 (neal.yuan@ucsf.edu).

Atrial fibrillation (AF) is the most common arrhythmia, affecting one-quarter of patients older than 80 years.¹ Patients with AF are 5 times more likely to experience a stroke and have up to a 25% risk of dying within 30 days of stroke.^{2,3} Many cases of AF go undetected, since at least one-third are asymptomatic.⁴⁻⁶ Among patients who experience an acute stroke of unknown origin, one-fifth will be found to have occult AF.⁷⁻¹⁰ Atrial fibrillation also causes long-term changes in cardiac structure, including atrial dilation and ventricular function deterioration, which can result in permanent AF, valvular regurgitation, and heart failure.^{11,12}

Effective clinical management can mitigate the complications of AF. Oral anticoagulation reduces the relative risk of stroke by two-thirds.¹³ Early use of antiarrhythmic medications or ablation may prevent more permanent AF and reduce symptoms and stroke risk.¹⁴⁻¹⁶ Earlier detection of AF therefore holds promise in preventing later adverse sequelae.

Deep learning, a subset of machine learning, can help diagnose early disease given its ability to use information-dense data to draw associations that may be too complicated to be routinely identified by human clinicians. Deep learning models of electrocardiograms (ECGs) have been used to successfully predict mortality, heart failure, cardiomyopathy, and valvular disease.¹⁷⁻²⁵ Deep learning has also been used recently to predict paroxysmal and incident AF, often in predominantly White, single-center patient populations.²⁶⁻²⁸

To date, few deep learning algorithms have been used for the US Veterans Affairs (VA) population, which includes almost 19 million individuals from diverse demographic backgrounds, many of whom are at higher risk for having cardiovascular disease, including AF, compared with the general adult population.²⁹⁻³¹ The VA patient population therefore represents a group in which deep learning-guided screening efforts may be most effective. We investigated whether deep learning applied to sinus rhythm ECGs in VA patients could predict the presence of concurrent AF.

Methods

This prognostic study was approved by the University of California, San Francisco and Cedars-Sinai, Los Angeles, California, Institutional Review Boards, which waived the need for informed consent for use of deidentified medical records. This study followed the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline.

ECG Data Set Selection

We extracted all 12-lead ECGs acquired at sites within the VA's Veterans Integrated Services Network Region 21, which includes 6 separate VA medical center networks (San Francisco, Palo Alto, Fresno, and Sacramento, California; Reno, Nevada; and the Pacific Islands), each of which is composed of multiple clinics. Electrocardiography was performed from January 1, 1987, to December 31, 2022. Electrocardiogram tracings were linked to cardiologist ECG interpretations, patient demographic characteristics (age, sex, and race and ethnicity), and

Key Points

Question Can a deep learning model using routinely acquired outpatient 12-lead electrocardiograms (ECGs) in sinus rhythm predict the presence of atrial fibrillation (AF) within 31 days across diverse populations?

Findings In this prognostic study, a model trained on data from 2 large Veterans Affairs (VA) hospital networks predicted AF with high accuracy in several separate patient populations (VA and non-VA) and across different demographic and comorbidity subgroups.

Meaning These findings suggest that deep learning applied to ECGs could help identify patients at high risk of AF who could be considered for intensive monitoring programs to help prevent adverse cardiac events.

comorbidity information (AF; heart failure; hypertension; diabetes; prior cerebrovascular accident [CVA], transient ischemic attack [TIA], or thromboembolism [TE]; prior myocardial infarction [MI]; peripheral vascular disease; and chronic kidney disease) from the VA Corporate Data Warehouse. Comorbidities were determined using *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* and *Current Procedural Terminology* codes.³² Using comorbidity fields, we estimated the CHA₂DS₂-VASC (congestive heart failure, hypertension, age, diabetes mellitus, prior stroke or transient ischemic attack or thromboembolism, vascular disease, age, sex category) scores.

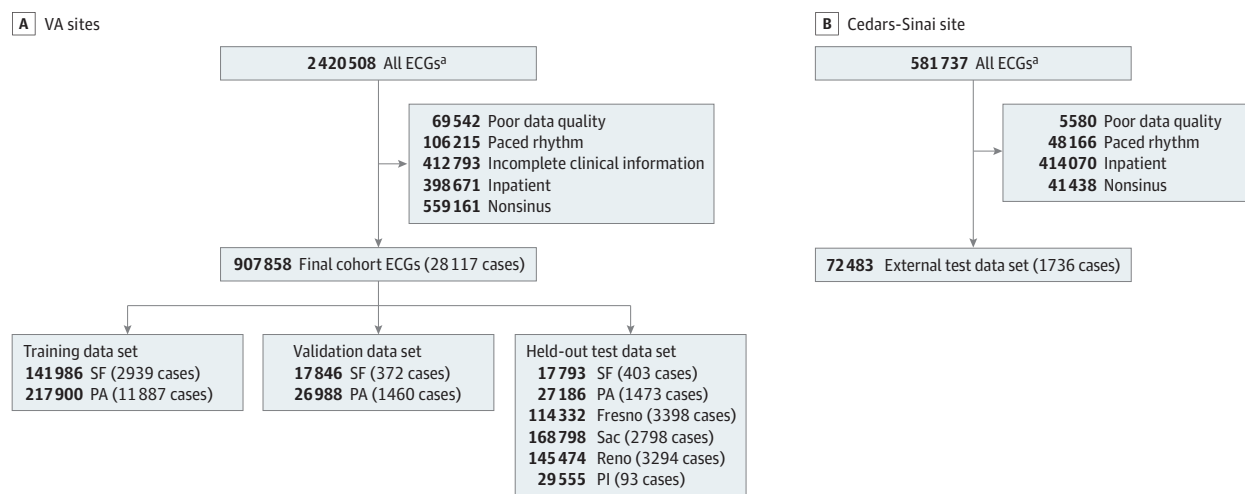
We included only ECGs in sinus rhythm. We excluded ECGs that had poor data quality or paced rhythms or that could not be paired with age and sex information (a sign that a patient was not followed up consistently in the VA health system or that the ECG patient data were entered incorrectly and were not linkable) (Figure 1). We limited our data set to outpatient ECGs, given that screening for AF would predominantly be implemented in an outpatient setting. Inpatient ECGs could introduce selection bias for sicker patients who may not reflect a general AF screening population.

We used ECGs from the San Francisco and Palo Alto VA sites for model training, validation, and testing. We used ECGs from the Fresno, Sacramento, Reno, and Pacific Islands VA sites as separate test data sets that were held out from model training. For an external test data set, we used all 12-lead ECGs acquired at Cedars-Sinai Medical Center, Los Angeles, California—a large urban tertiary care center—from March 1, 2005, to December 31, 2018. The same inclusion and exclusion criteria were applied as were used for the VA data set.

Definition of Cases and Controls

Cases of concurrent AF were defined as sinus rhythm ECGs that could be paired with at least 1 ECG in AF or flutter (based on the cardiologist ECG interpretation) within 31 days (eFigure 1 in Supplement 1). Controls were defined as sinus ECGs in patients who did not have ECGs in AF or flutter or with diagnoses of AF or flutter by *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* and *Current Procedural Terminology* coding. A single patient could

Figure 1. Study Flow Diagram



Inclusion and exclusion of 12-lead electrocardiograms (ECGs) at 6 Veterans Affairs (VA) sites and the Cedars-Sinai Medical Center. The model was trained and validated on ECGs from the San Francisco (SF) and Palo Alto (PA) VA sites. The model was then tested on held-out ECGs from the SF and PA VA sites in addition to ECGs from 4 other VA sites (Fresno, Sacramento [Sac], Reno, and

Pacific Islands [PI]) and the Cedars-Sinai site.

^a A single ECG could fall into multiple exclusion categories (eg, both a paced rhythm and nonsinus).

contribute multiple case and control ECGs, which has been shown to improve model performance.²⁶

In an additional exploratory analysis to simulate prospective prediction of a patient's first case of AF within a longer 1-year time frame, we defined cases as sinus rhythm ECGs that were closest to and chronologically before the first diagnosis of AF for each patient. Electrocardiograms had to be obtained within 1 year before AF diagnosis.

ECG Processing and Deep Learning Model Training

Electrocardiogram tracings were extracted from the VA's MUSE Cardiology Information System, version 9 (GE HealthCare). Electrocardiogram waveform data were acquired at 250 Hz and extracted as 10-second, 12×2500 matrices of amplitude values stored as base64 text. Electrocardiograms underwent baseline wander correction using median filtering at 200- and 600-millisecond intervals and z score normalization.

We used an atrous convolutional neural network based on a novel architecture previously used for predicting clinical phenotypes from ECGs (eFigure 1 in Supplement 1).³³ The model was trained using PyTorch software, version 2.0.1 (Linux Foundation). We initialized our model with random weights and trained using a binary cross-entropy loss function for 50 epochs with an Adam optimizer and an initial learning rate of 1×10^{-4} . The training data set, composed of ECGs from the San Francisco and Palo Alto VA sites, was split on a patient level in an 80:10:10 ratio to create training, validation, and held-out test data sets.

Statistical Analysis

All performance analyses were from model prediction of held-out VA data sets and the external Cedars-Sinai data set not involved in model training. We compared the deep learning mod-

el's performance with clinical prediction of AF for held-out testing data from the VA and Cedars-Sinai sites using the CHA₂DS₂-VAsC score and a logistic regression model that incorporated all available demographic and comorbidity information (age; sex; history of heart failure; diabetes; CVA, TIA, or TE; prior MI; peripheral vascular disease; and chronic kidney disease). These patient characteristics approximate those used in AF clinical risk prediction models such as the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) AF consortium risk score.³⁴ A CHARGE-AF risk score was not explicitly calculated because of the inability to reliably determine blood pressure and antihypertensive medication use at the time of the ECG.

Model discrimination was assessed by the area under the receiver operating characteristic curve (AUROC). We reported the sensitivity, specificity, and accuracy at the Youden index (defined as the maximum value of sensitivity plus specificity minus 1) as well as the maximum F1 score (harmonic mean of precision [positive predictive value] and recall [sensitivity]).³⁵ All metrics were reported with 2-sided 95% CIs from 1000 bootstrapped samples. The AUROCs were compared using the DeLong test.³⁶ We calculated the number needed to screen to detect a true-positive case of AF among patients deemed as high risk by the deep learning model as 1 divided by the positive predictive value.

For model calibration, risk scores underwent Platt scaling using logistic regression trained on 80% of the test data set and was then applied to a held-out 20% of the test data set.³⁷ We visualized a calibration plot for this held-out 20% test data set by plotting the observed vs predicted risk of AF for 50 equal-sized groups of increasing predicted risk. Calibration was quantified using the Brier score, which is the mean squared error between observed outcome and predicted risk, with 0 repre-

Table 1. ECG Patient Characteristics by Site^a

Characteristic	Study Site							
	All VA	San Francisco VA	Palo Alto VA	Fresno VA	Sacramento VA	Reno VA	Pacific Islands VA	Cedars-Sinai
No. of ECGs	907 858	177 625	272 074	114 332	168 798	145 474	29 555	72 483
ECGs per patient, mean (SD)	3.27 (4.14)	3.67 (4.82)	3.48 (4.55)	3.88 (4.65)	2.74 (3.09)	3.23 (3.81)	1.85 (1.49)	1.62 (2.78)
Age, mean (SD), y	62.4 (13.5)	62.4 (13.1)	61.6 (14.0)	64.1 (13.1)	62.2 (13.8)	62.7 (13.2)	61.8 (12.9)	59.5 (15.4)
Sex								
Women	58 158 (6.4)	10 820 (6.1)	18 548 (6.8)	5440 (4.8)	13 020 (7.7)	8796 (6.0)	1534 (5.2)	38 068 (52.5)
Men	849 700 (93.6)	166 805 (93.9)	253 526 (93.2)	108 892 (95.2)	155 778 (92.3)	136 678 (94.0)	28 021 (94.8)	34 415 (47.5)
Race and ethnicity								
American Indian or Alaska Native	1553 (0.2)	330 (0.2)	555 (0.2)	118 (0.1)	331 (0.2)	165 (0.1)	54 (0.2)	66 (0.1)
Asian	24 813 (2.7)	8257 (4.6)	9408 (3.5)	494 (0.4)	3562 (2.1)	196 (0.1)	2896 (9.8)	5743 (7.9)
Black	96 912 (10.7)	31 192 (17.6)	29 646 (10.9)	4731 (4.1)	27 930 (16.5)	2159 (1.5)	1254 (4.2)	6828 (9.4)
Latinx	41 446 (4.6)	5691 (3.2)	18 216 (6.7)	9617 (8.4)	6455 (3.8)	987 (0.7)	480 (1.6)	2119 (2.9)
Native Hawaiian or Other Pacific Islander	6193 (0.7)	502 (0.3)	1179 (0.4)	142 (0.1)	1000 (0.6)	59 (0.04)	3311 (11.2)	20 (0.03)
White	566 613 (62.4)	122 725 (69.1)	205 831 (75.7)	48 544 (42.5)	121 565 (72.0)	59 010 (40.6)	8938 (30.2)	54 245 (74.8)
Other ^b	3690 (0.4)	591 (0.3)	1192 (0.4)	142 (0.1)	916 (0.5)	56 (0.04)	793 (2.7)	69 (0.1)
Unknown	166 638 (18.4)	8337 (4.7)	6047 (2.2)	50 544 (44.2)	7039 (4.2)	82 842 (56.9)	11 829 (40.0)	3393 (4.7)
Comorbidities								
Heart failure	101 548 (11.2)	20 395 (11.5)	26 827 (9.9)	15 246 (13.3)	21 168 (12.5)	14 003 (9.6)	3909 (13.2)	6088 (8.4)
Hypertension	523 776 (57.7)	97 995 (55.2)	127 289 (46.8)	78 804 (68.9)	115 605 (68.5)	83 449 (57.4)	20 634 (69.8)	14 627 (20.2)
Diabetes	294 232 (32.4)	53 360 (30.0)	76 378 (28.1)	50 328 (44.0)	60 373 (35.8)	41 881 (28.8)	11 912 (40.3)	6170 (8.5)
CVA, TIA, or TE	80 006 (8.8)	15 402 (8.7)	19 365 (7.1)	11 844 (10.4)	17 689 (10.5)	13 783 (9.5)	1923 (6.5)	3309 (4.6)
MI	100 788 (11.1)	20 954 (11.8)	29 135 (10.7)	15 305 (13.4)	18 899 (11.2)	14 123 (9.7)	2372 (8.0)	1339 (1.8)
PVD	37 596 (4.1)	8339 (4.7)	7837 (2.9)	4316 (3.8)	7938 (4.7)	7829 (5.4)	1337 (4.5)	3740 (5.2)
CKD	92 461 (10.2)	18 226 (10.3)	21 900 (8.0)	14 347 (12.5)	21 025 (12.5)	13 490 (9.3)	3473 (11.8)	5943 (8.2)
CHA ₂ DS ₂ -VASc score, mean (SD)	1.9 (1.6)	1.9 (1.6)	1.7 (1.6)	2.3 (1.6)	2.1 (1.6)	1.9 (1.6)	2.1 (1.5)	1.6 (1.4)
Concurrent AF	28 117 (3.1)	3714 (2.1)	14 820 (5.4)	3398 (3.0)	2798 (1.7)	3294 (2.3)	93 (0.3)	1736 (2.4)

Abbreviations: AF, atrial fibrillation; CHA₂DS₂-VASc, congestive heart failure, hypertension, age, diabetes mellitus, prior stroke or transient ischemic attack or thromboembolism, vascular disease, age, sex category; CKD, chronic kidney disease; CVA, cerebrovascular accident; ECG, electrocardiogram; MI, myocardial infarction; PVD, peripheral vascular disease; TE, thromboembolism; TIA,

transient ischemic attack; VA, Veterans Affairs.

^a Unless otherwise indicated, data are expressed as No. (%) of ECGs.

^b Category is not broken down into subcategories in the VA data set.

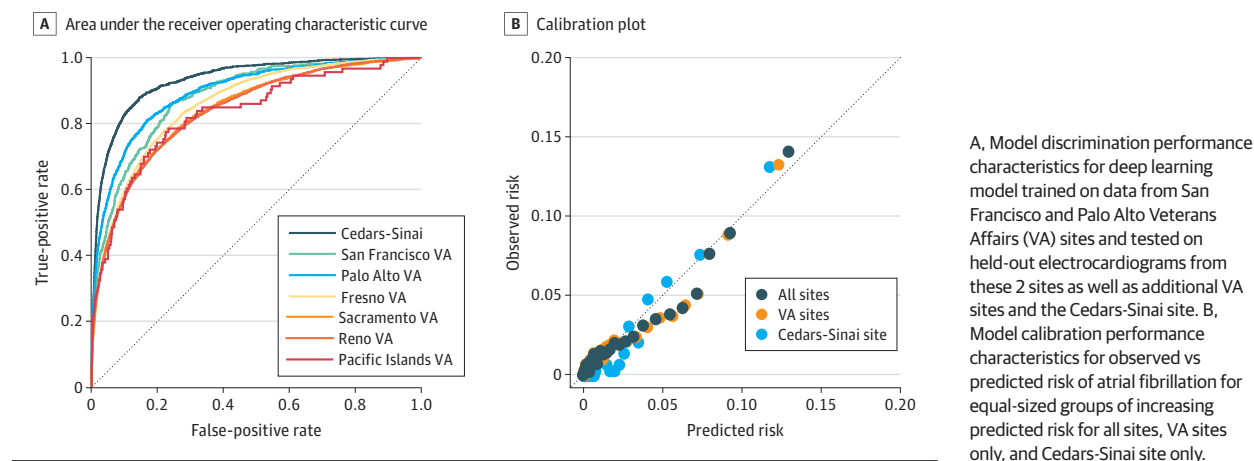
senting perfect accuracy and 1, perfect inaccuracy. Calibration was tested using the Spiegelhalter z test at a significance threshold of 2-sided $P < .05$. The null hypothesis of the Spiegelhalter z test is that the model is well calibrated; a statistically significant score indicates poor calibration. Calibration was visualized and tested across all sites and separately across the VA and Cedars-Sinai sites. Statistical analysis was performed in R, version 4.1.0 (R Project for Statistical Computing) and Python, version 3.10.0 (Python Software Foundation).

Results

A total of 2 420 508 twelve-lead ECGs were acquired within our network of VA hospitals. After excluding ECGs with poor data

quality, paced rhythms, or incomplete clinical information and that were nonsinus or acquired in inpatient settings (62.5% of all ECGs), the final VA cohort included 907 858 outpatient ECGs in sinus rhythm from 277 528 patients, with 28 117 ECGs having a documented case of AF within 31 days (Figure 1). The Cedars-Sinai external testing cohort included 72 483 outpatient ECGs in sinus rhythm from 44 754 patients, with 1736 cases of AF within 31 days. In the VA cohort, ECGs were from patients with a mean (SD) age of 62.4 (13.5) years; 6.4% were female and 93.6% were male. Female patients were on average younger than male patients (69.7% younger than 65 years compared with 50.6% of male patients). In terms of race and ethnicity, 0.2% were American Indian or Alaska Native, 2.7% were Asian, 10.7% were Black, 4.6% were Latinx, 0.7% were Native Hawaiian or Other Pacific Islander, 62.4% were White, 0.4% were of other race or ethnicity (which is not broken down

Figure 2. Model Performance



into subcategories in the VA data set), and 18.4% were of unknown race or ethnicity. The VA cohort had a high prevalence of comorbidities (11.2% heart failure; 32.4% diabetes; 8.8% prior CVA, TIA, or TE; 11.1% prior MI), with a mean (SD) CHA₂DS₂-VASc score of 1.9 (1.6) (Table 1). In the external test cohort (n = 72 483), patients had a mean (SD) age of 59.5 (15.4) years, 52.5% were women, and 47.5% were men. In terms of race and ethnicity, 0.1% were American Indian or Alaska Native, 7.9% were Asian, 9.4% were Black, 2.9% were Latinx, 0.03% were Native Hawaiian or Other Pacific Islander, 74.8% were White, 0.1% were of other race or ethnicity, and 4.7% were of unknown race or ethnicity. Compared with the VA population, there was a lower prevalence of comorbidities (8.4% heart failure; 8.5% diabetes; 4.6% prior CVA, TIA, or TE; 1.8% prior MI), with a mean (SD) CHA₂DS₂-VASc score of 1.6 (1.4).

The prevalence of sinus ECGs with AF detected within 31 days on ECG was 3.1%. When comparing cases with controls, patients with concurrent AF were older (mean [SD] age, 70.4 [10.5] vs 61.9 [13.7] years), less often female (3.8% vs 10.0%), and more often White (78.3% vs 62.9%) with a higher incidence of comorbidities (37.3% vs 10.2% heart failure; 45.0% vs 30.2% diabetes; 16.2% vs 8.3% prior CVA, TIA, or TE; 25.4% vs 9.9% prior MI) and a higher CHA₂DS₂-VASc score (mean [SD], 3.1 [1.8] vs 1.9 [1.6]) (eTable 1 in Supplement 1).

The deep learning model was trained on 359 886 ECGs from the San Francisco and Palo Alto VA sites. When tested on held-out training data sets, the model had AUROCs of 0.88 (95% CI, 0.87-0.90) at the San Francisco VA site and 0.89 (95% CI, 0.89-0.90) at the Palo Alto VA site; accuracy of 0.81 (95% CI, 0.79-0.83) at the San Francisco VA site and 0.82 (95% CI, 0.81-0.83) at the Palo Alto VA site; and F1 scores of 0.33 (95% CI, 0.29-0.37) at the San Francisco VA site and 0.49 (95% CI, 0.47-0.51) at the Palo Alto VA site (Figure 2A and eTable 2 in Supplement 1). The model was then applied to 4 other VA sites that were not included in model training and achieved AUROCs of 0.86 (95% CI, 0.85-0.87) (Fresno VA), 0.84 (95% CI, 0.83-0.85) (Sacramento VA), 0.84 (95% CI, 0.83-0.85) (Reno VA), and 0.83 (95% CI, 0.79-0.88) (Pacific Islands VA). In summary, across held-out test ECGs at all VA sites, the model had an overall AUROC of 0.86 (95% CI, 0.85-0.86), accuracy of 0.78 (95% CI, 0.77-0.78), and F1 score of 0.30 (95% CI, 0.30-0.31).

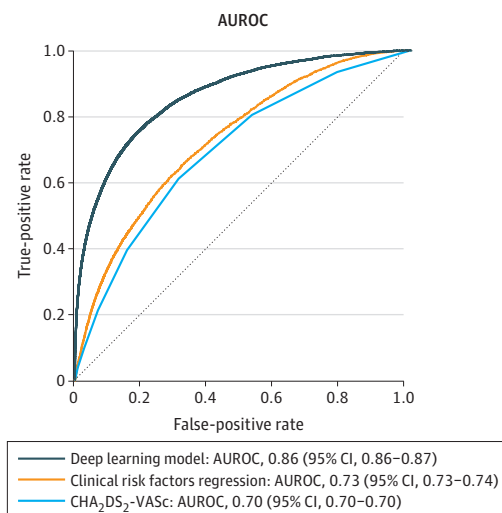
When tested on an external data set at Cedars-Sinai Medical Center, the model achieved an AUROC of 0.93 (95% CI, 0.93-0.94), accuracy of 0.87 (95% CI, 0.86-0.88), and F1 score of 0.46 (95% CI, 0.44-0.48).

The deep learning model was also well calibrated, with Brier scores of 0.02 across all sites, 0.02 across all VA sites, and 0.02 at the Cedars-Sinai site (a Brier score of 0 indicates perfect calibration; 1, perfect miscalibration) (Figure 2B). Testing by Spiegelhalter z test also confirmed a failure to reject the null hypothesis of model calibration at a significance threshold of .05 ($P = .06$ across all sites, $P = .07$ across VA sites, and $P = .39$ at the Cedars-Sinai site).

To establish the deep learning model's performance relative to conventional clinical prediction tools, we compared the deep learning model's predictions to AF predictions made by using the CHA₂DS₂-VASc score as well as regression using all available demographic and clinical risk factor information. When applied to test patients not involved in model training across all VA and Cedars-Sinai sites, the deep learning model had an AUROC of 0.86 (95% CI, 0.86-0.87), the risk factor regression model had an AUROC of 0.73 (95% CI, 0.73-0.74), and the CHA₂DS₂-VASc score had an AUROC of 0.70 (95% CI, 0.70-0.70) (Figure 3). Choosing a screening threshold to fix testing sensitivity at 25% resulted in the number needed to screen to find a true-positive case of AF being 2.47 individuals using the deep learning model vs 11.48 using the regression model and 12.01 using CHA₂DS₂-VASc score (eTable 3 in Supplement 1).

We tested the model's performance in specific patient cohort subsets (Table 2 and eTable 4 in Supplement 1). Across the different sites, there were substantial differences in the proportion of patients who were women (range, 4.8%-52.5%), Black (1.5%-17.2%), younger than 65 years (48.1%-59.9%), and with a CHA₂DS₂-VASc score of 2 or greater (41.4%-64.4%). At some sites, the model showed small significant increases in performance in female patients and small decreases in performance in patients older than 65 years and those with a CHA₂DS₂-VASc score of 2 or greater. However, these differences were not observed consistently across all sites, and performance was largely unchanged across the different subgroups.

Figure 3. Deep Learning Model Performance Compared With Clinical Risk Factor Models



Performance of deep learning model on all electrocardiograms held out from the model training compared with predicting atrial fibrillation using a clinical risk factors model (age; sex; history of heart failure; diabetes; cerebrovascular accident, transient ischemic attack, or thromboembolism; prior myocardial infarction; peripheral vascular disease; and chronic kidney disease) or the CHA₂DS₂-VASc (congestive heart failure, hypertension, age, diabetes mellitus, prior stroke or transient ischemic attack or thromboembolism, vascular disease, age, sex category) score. AUROC indicates area under the receiver operating characteristic curve.

We conducted an additional exploratory analysis to simulate the prediction of new undiagnosed AF within a longer 1-year time frame by redefining cases as sinus rhythm ECGs closest to and chronologically before the first known diagnosis of AF for each patient (limited to ECGs within 1 year before AF diagnosis). In this analysis, the model had AUROCs ranging from 0.80 (95% CI, 0.79-0.81) to 0.85 (95% CI, 0.84-0.86) and accuracies from 0.73 (95% CI, 0.72-0.75) to 0.79 (95% CI, 0.73-0.85) at VA sites (eTable 5 and eFigure 2 in Supplement 1). When tested on Cedars-Sinai ECGs, the AUROC was 0.79 (95% CI, 0.78-0.79) with an accuracy of 0.72 (95% CI, 0.71-0.72).

Discussion

In this multisite prognostic study of a large and diverse population, we found that a deep learning model using convolutional neural networks predicted with high discrimination and calibration the occurrence of AF within 31 days from 12-lead ECGs in sinus rhythm. Prediction performance was robust across 6 different VA hospital networks as well as a separate non-VA large urban academic medical center. Predictions were better than those using conventional clinical risk factors and were largely preserved across multiple patient subgroups, including women and Black patients. We additionally showed that this model may help to predict new-onset AF within a longer 1-year time horizon.

Early detection of AF holds promise because it can inform management decisions that change the natural progression and complication profile of this disease. Anticoagulation reduces the risk of stroke by two-thirds.¹³ Antiarrhythmic medications and ablation can prevent the development of permanent AF and may also reduce the rate of stroke and cardiovascular death.¹⁴⁻¹⁶ While guidelines support opportunistic screening for AF, the ideal population and best method for screening remain unclear.^{38,39} Multiple studies⁴⁰⁻⁴⁶ have proven that more intensive monitoring, whether by structured 12-lead ECG screening programs, remote monitoring, or implanted devices, results in more detection of occult AF. However, most of these screening interventions are resource intensive and sometimes invasive, and they have not been adopted as part of routine clinical practice. One recent large randomized clinical trial of an AF screening program for all individuals aged 75 to 76 years in 2 regions of Sweden⁴⁷ revealed that one of the major barriers in screening was convincing patients to participate in the program, even though those who did participate had a significantly lower composite end point of stroke, bleeding, and mortality.

In this study, we show that deep learning of 12-lead ECGs acquired as a part of routine clinical practice may be a relatively easy method for identifying patients who are at highest risk for having unidentified AF. This could be incorporated into existing workflows without necessarily requiring significant additional patient participation or clinical resources. High-risk patients could then be funneled into a more intensive AF identification program using additional monitoring. Among patients determined to be at high risk by the deep learning model, the number needed to screen to detect a true positive case of AF is tunable based on the desired test sensitivity and could be as low as 2.47 patients for a test sensitivity of 25% and up to 11.48 patients for a sensitivity of 75%. This is substantially lower than the number needed to screen using risk assessment based on clinical risk factor regression or the CHA₂DS₂-VASc score, which had numbers needed to screen of 11.5 and 12.0, respectively, for a test sensitivity of 25% and 25.4 and 20.8, respectively, for a test sensitivity of 75%. Our work builds on previous research that has also used deep learning to identify AF from sinus ECGs with simulated and real pilot deployments in different patient populations.^{27,48}

Our findings are unique in applying deep learning to multicenter cardiovascular data from US veterans with additional external site validation. Implementation of a screening program in this large population may be particularly effective given the high pretest probability of disease, which could help limit the rate of false-positive results, as well as the higher mean CHA₂DS₂-VASc score, which could increase the net benefit of starting anticoagulation therapy.²⁹⁻³¹ The same characteristics that make the veteran population appropriate for AF screening, however, also make it different from other well-studied patient populations. These differences can pose challenges for the generalizability of models developed outside of the VA when applied to VA populations and vice versa; deep learning models remain limited in their interpretability and at risk for overfitting and confounding.⁴⁹ A recent study⁵⁰ showed that a deep learning algorithm designed to recognize acute kid-

Table 2. Model Performance in Patient Subgroups^a

Subgroup	Study site						
	San Francisco VA (n = 17 793)	Palo Alto VA (n = 27 186)	Fresno VA (n = 114 332)	Sacramento VA (n = 168 798)	Reno VA (n = 145 174)	Pacific Islands VA (n = 29 555)	Cedars-Sinai (n = 72 483)
All patients, AUROC (95% CI)	0.88 (0.87-0.90)	0.89 (0.89-0.90)	0.86 (0.85-0.87)	0.84 (0.83-0.85)	0.84 (0.83-0.85)	0.83 (0.79-0.88)	0.93 (0.93-0.94)
Women							
No. (%)	1048 (5.9)	1785 (6.6)	5440 (4.8)	13 020 (7.7)	8796 (6.1)	1534 (5.2)	38 068 (52.5)
AUROC (95% CI)	0.92 (0.88-0.97) ^b	0.88 (0.79-0.97)	0.88 (0.84-0.92)	0.87 (0.82-0.92)	0.87 (0.84-0.91)	0.96 (0.91-1.00) ^b	0.95 (0.94-0.96) ^b
Black patients							
No. (%)	3058 (17.2)	2827 (10.4)	4731 (4.1)	27 930 (16.5)	2159 (1.5)	1254 (4.2)	6828 (9.4)
AUROC (95% CI)	0.90 (0.85-0.94)	0.88 (0.84-0.92)	0.84 (0.81-0.88)	0.86 (0.84-0.89)	0.80 (0.71-0.89)	0.86 (0.73-0.99)	0.92 (0.88-0.95)
Aged <65 y							
No. (%)	9834 (55.3)	14 884 (54.7)	55 035 (48.1)	90 427 (53.6)	75 162 (51.8)	15 549 (52.6)	43 431 (59.9)
AUROC (95% CI)	0.88 (0.85-0.92)	0.90 (0.88-0.91) ^b	0.86 (0.85-0.88)	0.84 (0.83-0.86)	0.85 (0.83-0.86)	0.80 (0.72-0.88)	0.94 (0.93-0.95) ^b
Aged ≥65 y							
No. (%)	7959 (44.7)	12 302 (45.3)	59 297 (51.9)	78 371 (46.4)	70 312 (48.4)	14 006 (47.4)	29 052 (40.1)
AUROC (95% CI)	0.85 (0.83-0.88)	0.87 (0.86-0.89)	0.83 (0.82-0.84) ^b	0.81 (0.80-0.82) ^b	0.81 (0.80-0.82) ^b	0.85 (0.80-0.89)	0.92 (0.91-0.93) ^b
CHA ₂ DS ₂ -VAsC score ≥2							
No. (%)	9340 (52.5)	12 872 (47.3)	73 633 (64.4)	101 830 (60.3)	78 041 (53.60)	17 938 (60.7)	29 990 (41.4)
AUROC (95% CI)	0.86 (0.84-0.88)	0.87 (0.86-0.88)	0.84 (0.83-0.84) ^b	0.82 (0.81-0.83) ^b	0.82 (0.81-0.83) ^b	0.84 (0.78-0.90)	0.92 (0.91-0.93) ^b

Abbreviations: AUROC, area under the receiver operating characteristic curve; CHA₂DS₂-VAsC, congestive heart failure, hypertension, age, diabetes mellitus, prior stroke or transient ischemic attack or thromboembolism, vascular disease, age, sex category; VA, Veterans Affairs.

^a Subgroups for women and Black patients are presented because these groups

have substantial differences in representation across our study sites as well as compared with prior research study cohorts.

^b $P < .01$ when comparing with AUROC for all patients at site.

ney injury did not perform equally well across VA and non-VA populations, possibly due to differences in demographic characteristics (ie, a significantly lower proportion of VA patients being female).

We found that despite substantial differences in patient makeup across different VA cohorts and our external non-VA test site, the predictive performance of our deep learning model for concurrent AF was largely preserved. At some sites, small decreases in performance occurred in patients who were older and had higher CHA₂DS₂-VAsC scores. These patients may have had more comorbidities that introduced competing changes to the ECG and made predicting AF more difficult. Female patients in this cohort were overall younger (69.7% younger than 65 years compared to 50.6% of male patients), which could explain the improved performance in this subgroup. Overall, these differences were not seen across all sites and, given the small magnitude of difference, may not be clinically meaningful. Similarly, our model displayed small improvements in discriminatory abilities when applied to the external test cohort from the Cedars-Sinai site. This may be because this cohort was relatively enriched for patients who were women, were younger, and had a lower CHA₂DS₂-VAsC score.

Limitations

Several limitations of this study warrant consideration. Because it was a retrospective study, the population with

12-lead ECGs may be different from a prospective AF screening population. While ECGs in our VA system are routinely obtained during clinic visits, there was site-to-site variability in the mean number of ECGs per patient, and we might expect that this study's patient population with ECGs has a higher prevalence of cardiovascular disease and AF. This selection bias could increase the positive predictive value of the model and decrease the number needed to screen compared with using the model when screening a broader population of patients. Prospective model performance could be similar if a higher-risk population is chosen for prospective screening. While we used all data from the ECG database and electronic health records to identify cases of AF, some patients in the control group likely had undiagnosed AF. This would bias our results to the null and cause underestimation of our model's performance. Some patients predicted to be cases could have in fact been correctly predicted but unknown at the time or had AF identified at an outside health system. Future prospective studies using continuous monitoring of high-risk patients by our model could confirm AF prediction and clarify whether this method improves downstream outcomes such as CVA and TE.

Conclusion

In this prognostic study, a convolutional neural network trained using outpatient 12-lead ECGs in sinus rhythm from US vet-

erans successfully predicted the presence of AF within 31 days in VA and non-VA populations with a diversity of demographic characteristics and comorbidities. Such a model holds

promise for AF screening and could be used in future efforts to reduce adverse complications associated with this disease.

ARTICLE INFORMATION

Accepted for Publication: July 31, 2023.

Published Online: October 18, 2023.

doi:10.1001/jamacardio.2023.3701

Author Affiliations: Department of Medicine, University of California, San Francisco (Yuan, Dhruva, Oesterle, Pellegrini); Division of Cardiology, San Francisco Veterans Affairs Medical Center, San Francisco, California (Yuan, Dhruva, Oesterle, Pellegrini, Ouyang); Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California (Duffy, Theurer, Ouyang); Division of Artificial Intelligence in Medicine, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, California (Duffy, Theurer); Department of Medicine, University of California, San Francisco (Vali, Keyhani); Division of General Internal Medicine, San Francisco Veterans Affairs Medical Center, San Francisco, California (Vali, Keyhani); Division of Cardiology, Palo Alto Veterans Affairs Medical Center, Palo Alto, California (Heidenreich); Division of Cardiovascular Medicine, Department of Medicine, Stanford University, Palo Alto, California (Heidenreich).

Author Contributions: Dr Yuan had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Yuan, Duffy, Theurer, Vali, Keyhani, Ouyang.

Acquisition, analysis, or interpretation of data: Yuan, Duffy, Dhruva, Oesterle, Pellegrini, Heidenreich, Keyhani, Ouyang.

Drafting of the manuscript: Yuan.

Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: Yuan, Pellegrini, Theurer, Ouyang.

Obtained funding: Ouyang.

Administrative, technical, or material support: Duffy, Oesterle, Pellegrini, Vali, Keyhani, Ouyang.

Supervision: Ouyang.

Conflict of Interest Disclosures: Dr Dhruva reported receiving research support from Arnold Ventures LLC and the Department of Veterans Affairs unrelated to the present research and serving on the Medicare Evidence Development and Coverage Advisory Committee and Institute for Clinical and Economic Review California Technology Assessment Forum. Dr Pellegrini reported receiving personal fees from Abbott Laboratories, Cook Medical, and Biosense Webster Inc outside the submitted work. No other disclosures were reported.

Funding/Support: This study was supported by grant K99 HL157421 from the National Institutes of Health (Dr Ouyang) and by grants HSR&D IIR 18-231-2 and CSR&D IO1 CX002417 from the Department of Veterans Affairs (Dr Kayhani and Ms Vali).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or

approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 2.

Additional Contributions: Chris Freking, Veterans Integrated Services Network Region 21, assisted with access to electrocardiographic data and Nathan Wilairat, MS, provided statistical consultation. They were not compensated for this work.

REFERENCES

- Go AS, Hylek EM, Phillips KA, et al. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the Anticoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. *JAMA*. 2001;285(18):2370-2375. doi:10.1001/jama.285.18.2370
- Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke*. 1991;22(8):983-988. doi:10.1161/01.STR.22.8.983
- Fang MC, Go AS, Chang Y, et al. Long-term survival after ischemic stroke in patients with atrial fibrillation. *Neurology*. 2014;82(12):1033-1037. doi:10.1212/WNL.0000000000000248
- Hindricks G, Piorkowski C, Tanner H, et al. Perception of atrial fibrillation before and after radiofrequency catheter ablation: relevance of asymptomatic arrhythmia recurrence. *Circulation*. 2005;112(3):307-313. doi:10.1161/CIRCULATIONAHA.104.518837
- Quirino G, Giammaria M, Corbucci G, et al. Diagnosis of paroxysmal atrial fibrillation in patients with implanted pacemakers: relationship to symptoms and other variables. *Pacing Clin Electrophysiol*. 2009;32(1):91-98. doi:10.1111/j.1540-8159.2009.02181.x
- Silberbauer J, Veasey RA, Cheek E, Maddekar N, Sulke N. Electrophysiological characteristics associated with symptoms in pacemaker patients with paroxysmal atrial fibrillation. *J Interv Card Electrophysiol*. 2009;26(1):31-40. doi:10.1007/s10840-009-9411-x
- Rizos T, Rasch C, Jenetzky E, et al. Detection of paroxysmal atrial fibrillation in acute stroke patients. *Cerebrovasc Dis*. 2010;30(4):410-417. doi:10.1159/000316885
- Seet RCS, Friedman PA, Rabinstein AA. Prolonged rhythm monitoring for the detection of occult paroxysmal atrial fibrillation in ischemic stroke of unknown cause. *Circulation*. 2011;124(4):477-486. doi:10.1161/CIRCULATIONAHA.111.029801
- Kalman JM, Sanders P, Rosso R, Calkins H. Should we perform catheter ablation for asymptomatic atrial fibrillation? *Circulation*. 2017;136(5):490-499. doi:10.1161/CIRCULATIONAHA.116.024926
- Sgreccia D, Manicardi M, Malavasi VL, et al. Comparing outcomes in asymptomatic and symptomatic atrial fibrillation: a systematic review and meta-analysis of 81 462 patients. *J Clin Med*. 2021;10(17):3979. doi:10.3390/jcm10173979
- Farhan S, Silbiger JJ, Halperin JL, et al. Pathophysiology, echocardiographic diagnosis, and treatment of atrial functional mitral regurgitation: JACC state-of-the-art review. *J Am Coll Cardiol*. 2022;80(24):2314-2330. doi:10.1016/j.jacc.2022.09.046
- Santhanakrishnan R, Wang N, Larson MG, et al. Atrial fibrillation begets heart failure and vice versa: temporal associations and differences in preserved versus reduced ejection fraction. *Circulation*. 2016;133(5):484-492. doi:10.1161/CIRCULATIONAHA.115.018614
- 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(12):857-867. doi:10.7326/0003-4819-146-12-200706190-00007
- Goette A, Borof K, Breithardt G, et al; EAST-AFNET 4 Investigators. Presenting pattern of atrial fibrillation and outcomes of early rhythm control therapy. *J Am Coll Cardiol*. 2022;80(4):283-295. doi:10.1016/j.jacc.2022.04.058
- Kirchhoff P, Camm AJ, Goette A, et al; EAST-AFNET 4 Trial Investigators. Early rhythm-control therapy in patients with atrial fibrillation. *N Engl J Med*. 2020;383(14):1305-1316. doi:10.1056/NEJMoa2019422
- Andrade JG, Wells GA, Deyell MW, et al; EARLY-AF Investigators. Cryoablation or drug therapy for initial treatment of atrial fibrillation. *N Engl J Med*. 2021;384(4):305-315. doi:10.1056/NEJMoa2029980
- Raghunath S, Ulloa Cerna AE, Jing L, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med*. 2020;26(6):886-891. doi:10.1038/s41591-020-0870-z
- Akbilgic O, Butler L, Karabayir I, et al. ECG-AI: electrocardiographic artificial intelligence model for prediction of heart failure. *Eur Heart J Digit Health*. 2021;2(4):626-634. doi:10.1093/ehjdh/ztab080
- Adedinsawo D, Carter RE, Attia Z, et al. Artificial intelligence-enabled ECG algorithm to identify patients with left ventricular systolic dysfunction presenting to the emergency department with dyspnea. *Circ Arrhythm Electrophysiol*. 2020;13(8):e008437. doi:10.1161/CIRCEP.120.008437
- Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25(1):70-74. doi:10.1038/s41591-018-0240-2
- Attia ZI, Kapa S, Yao X, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. *J Cardiovasc Electrophysiol*. 2019;30(5):668-674. doi:10.1111/jce.13889
- Ko WY, Siontis KC, Attia ZI, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *J Am Coll Cardiol*. 2020;75(7):722-733. doi:10.1016/j.jacc.2019.12.030
- Elias P, Poterucha TJ, Rajaram V, et al. Deep learning electrocardiographic analysis for detection

- of left-sided valvular heart disease. *J Am Coll Cardiol*. 2022;80(6):613-626. doi:10.1016/j.jacc.2022.05.029
24. Kwon JM, Lee SY, Jeon KH, et al. Deep learning-based algorithm for detecting aortic stenosis using electrocardiography. *J Am Heart Assoc*. 2020;9(7):e014717. doi:10.1161/JAHA.119.014717
25. Cohen-Shelly M, Attia ZI, Friedman PA, et al. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *Eur Heart J*. 2021;42(30):2885-2896. doi:10.1093/eurheartj/ehab153
26. 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(10201):861-867. doi:10.1016/S0140-6736(19)31721-0
27. 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(13):1287-1298. doi:10.1161/CIRCULATIONAHA.120.047829
28. Khurshid S, Friedman S, Reeder C, et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation*. 2022;145(2):122-133. doi:10.1161/CIRCULATIONAHA.121.057480
29. Farmer CM, Hosek SD, Adamson DM. Balancing demand and supply for veterans' health care: a summary of three RAND assessments conducted under the Veterans Choice Act. RAND Corporation. 2016. Accessed January 17, 2023. https://www.rand.org/pubs/research_reports/RR1165z4.html
30. Agha Z, Lofgren RP, VanRuiswyk JV, Layde PM. Are patients at Veterans Affairs medical centers sicker? a comparative analysis of health status and medical resource use. *Arch Intern Med*. 2000;160(21):3252-3257. doi:10.1001/archinte.160.21.3252
31. Assari S. Veterans and risk of heart disease in the United States: a cohort with 20 years of follow up. *Int J Prev Med*. 2014;5(6):703-709.
32. Keyhani S, Cohen BE, Vali M, et al. The Heart and Cannabis (THC) Cohort: differences in baseline health and behaviors by cannabis use. *J Gen Intern Med*. 2022;37(14):3535-3544. doi:10.1007/s11606-021-07302-6
33. Holmstrom L, Christensen M, Yuan N, et al. Deep learning-based electrocardiographic screening for chronic kidney disease. *Commun Med (Lond)*. 2023;3(1):73. doi:10.1038/s43856-023-00278-w
34. 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(2):e000102. doi:10.1161/JAHA.112.000102
35. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR280030106>3.0.CO;2-3
36. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845. doi:10.2307/2531595
37. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc*. 2020;27(4):621-633. doi:10.1093/jamia/ocz228
38. Kirchhof P, Benussi S, Kotecha D, et al. 2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur J Cardiothorac Surg*. 2016;50(5):e1-e88. doi:10.1093/ejcts/ezw313
39. January CT, Wann LS, Calkins H, et al. 2019 AHA/ACC/HRS focused update of the 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. *J Am Coll Cardiol*. 2019;74(1):104-132. doi:10.1016/j.jacc.2019.01.011
40. Engdahl J, Andersson L, Mirskaya M, Rosenqvist M. Stepwise screening of atrial fibrillation in a 75-year-old population: implications for stroke prevention. *Circulation*. 2013;127(8):930-937. doi:10.1161/CIRCULATIONAHA.112.126656
41. Fitzmaurice DA, Hobbs FDR, Jowett S, et al. Screening versus routine practice in detection of atrial fibrillation in patients aged 65 or over: cluster randomised controlled trial. *BMJ*. 2007;335(7616):383. doi:10.1136/bmj.39280.660567.55
42. Steinhilb SR, Waalen J, Edwards AM, et al. Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed atrial fibrillation: the mSToPS randomized clinical trial. *JAMA*. 2018;320(2):146-155. doi:10.1001/jama.2018.8102
43. Lubitz SA, Atlas SJ, Ashburner JM, et al. Screening for atrial fibrillation in older adults at primary care visits: VITAL-AF randomized controlled trial. *Circulation*. 2022;145(13):946-954. doi:10.1161/CIRCULATIONAHA.121.057014
44. Perez MV, Mahaffey KW, Hedlin H, et al; Apple Heart Study Investigators. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med*. 2019;381(20):1909-1917. doi:10.1056/NEJMoa1901183
45. Reiffel JA, Verma A, Kowey PR, et al; REVEAL AF Investigators. Incidence of previously undiagnosed atrial fibrillation using insertable cardiac monitors in a high-risk population: the REVEAL AF study. *JAMA Cardiol*. 2017;2(10):1120-1127. doi:10.1001/jamacardio.2017.3180
46. Healey JS, Alings M, Ha A, et al; ASSERT-II Investigators. Subclinical atrial fibrillation in older patients. *Circulation*. 2017;136(14):1276-1283. doi:10.1161/CIRCULATIONAHA.117.028845
47. Sennaroglu 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*. 2021;398(10310):1498-1506. doi:10.1016/S0140-6736(21)01637-8
48. Noseworthy PA, Attia ZI, Behnken EM, et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. *Lancet*. 2022;400(10359):1206-1212. doi:10.1016/S0140-6736(22)01637-3
49. Duffy G, Clarke SL, Christensen M, et al. Confounders mediate AI prediction of demographics in medical imaging. *NPJ Digit Med*. 2022;5(1):188. doi:10.1038/s41746-022-00720-8
50. Cao J, Zhang X, Shahinian V, et al. Generalizability of an acute kidney injury prediction model across health systems. *Nat Mach Intell*. 2022;4:1121-1129. doi:10.1038/s42256-022-00563-8