Title

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

Journal

JAMA Cardiology, 8(12)

ISSN

2380-6583

Authors

Yuan, Neal
Duffy, Grant
Dhruva, Sanket S
et al.

Publication Date

2023-12-01

DOI

10.1001/jamacardio.2023.3701

Deep learning of electrocardiograms in sinus rhythm from US Veterans to predict atrial fibrillation

Neal Yuan, M.D.[1], Grant Duffy[2,3], Sanket Dhruva, M.D., M.H.S. [1], Adam Oesterle M.D. [1], Cara Pellegrini M.D. [1], John Theurer[2,3], Marzieh Vali M.S.[4], Paul A. Heidenreich, M.D.[5], Salomeh Keyhani, M.D.[4], David Ouyang, M.D.[2,3]

1. Department of Medicine, University of California, San Francisco, CA; Division of Cardiology, San Francisco Veterans Affairs Medical Center, San Francisco, CA
2. Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA
3. Department of Medicine, Division of Artificial Intelligence in Medicine, Cedars-Sinai Medical Center, Los Angeles, CA
4. Department of Medicine, University of California, San Francisco, CA; Division of General Internal Medicine, San Francisco Veterans Affairs Medical Center, San Francisco, CA
5. Division of Cardiology, Palo Alto Veterans Affairs Medical Center, Palo Alto, CA; Division of Cardiovascular Medicine, Department of Medicine, Stanford University, Palo Alto, CA

Corresponding Author
Neal Yuan
4150 Clement St
San Francisco, CA 94121
(415) 221-4810
Neal.Yuan@ucsf.edu

Funding
Dr. David Ouyang is supported by grant NIH K99 HL157421.
Dr. Salomeh Keyhani and Marzieh Vali were supported by VA HSR&D IIR 18-231-2 and VA CSR&D I01 CX002417.
Funding sources were not involved in study design, data collection, or analysis.

Disclosures: The authors have no conflicts of interest to report.

Acknowledgments
We gratefully acknowledge Chris Freking from VISN 21 for assisting with access to ECG data and Nathan Wilairat M.S. for statistical consultation.

Data Sharing Statement
ECG data from US veterans is subject to strict privacy regulations but is available by reasonable request if appropriate approvals are obtained.

Keywords: Electrocardiogram, Atrial Fibrillation, Neural Network, Deep Learning, Stroke

Manuscript Word Count: 3190

Key Points

Question: Can a deep learning model using routinely acquired outpatient 12-lead ECGs predict the presence of atrial fibrillation within 31 days across diverse populations?

Findings: A model trained on data from two large Veterans Affairs (VA) hospital networks predicted atrial fibrillation with high accuracy in several separate patient populations (VA and non-VA) and across different demographic and comorbidity subgroups.

Meaning: Deep learning of ECGs holds promise for identifying patients at high risk of atrial fibrillation who could be considered for intensive monitoring programs to help prevent adverse cardiac events.

Abstract

Importance

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

Objective

To determine whether deep learning of outpatient ECGs in sinus rhythm can predict patients with AF in a large and diverse population.

Design

Retrospective cohort study from 1/1/1987 to 12/31/2022.

Setting

Multicenter study at 6 US Veterans Affairs (VA) hospital networks and 1 large non-VA academic medical center

Participants

All outpatient 12-lead ECGs in sinus rhythm

Methods and Outcomes

We trained a convolutional neural network using 12-lead ECGs from 2 US VA hospital networks 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

We used cohort of 908,341 ECGs. ECGs were from patients across 6 VA sites who had an average age of 62.4 years, were 6.4% female, 37.6% non-white, with an average CHA2DS2-VASc score of 1.9. At the non-VA academic medical center, the average age was 59.5 years, with 52.5% female, 25.2% non-white, and an average CHA2DS2-VASc score of 1.6. A deep learning model predicted the presence of atrial fibrillation within 31 days of a sinus ECG with AUCs of 0.86 (95% CI 0.85-0.86) and 0.93 (0.93-0.94), accuracies of 0.78 (0.77-0.78) and 0.87 (0.86-0.88), F1 scores of 0.30 (0.30-0.31) and 0.46 (0.44-0.48) on held-out test ECGs at VA and non-VA hospitals, respectively. 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 2.5 individuals at a testing sensitivity of 25% and 11.5 at 75%. Model performance was similar in patients who were black, female, younger than 65 years old, or had CHA2DS2-VASc score ≥ 2.

Conclusions

Deep learning of outpatient sinus rhythm ECGs predicted AF within 31 days in populations with diverse demographics and comorbidities.

Abbreviations

Atrial fibrillation (AF)
Electrocardiogram (ECG)
Veteran Affairs (VA)
International Classification of Diseases (ICD)
Current Procedural Terminology (CPT)
Receiver Operating Characteristic (ROC)
Area Under the Curve (AUC)
Stroke (CVA)
Transient ischemic attack (TIA)
Thromboembolism (TE)
Myocardial infarction (MI)
Standard Deviation (SD)
Confidence Interval (CI)

Background

Atrial fibrillation (AF) is the most common arrhythmia, affecting one quarter of patients older than 80 years old.[1] Patients with AF are five 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] AF 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 utilize information-dense data to draw associations that may be too complicated to be routinely identified by human clinicians. Deep learning of electrocardiograms (ECGs) has been used to successfully predict mortality, heart failure, cardiomyopathy, and valvular disease.[17–25] It 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 Veteran Affairs (VA) population, which includes almost 19 million individuals from a diversity of demographic backgrounds, many of whom are at higher risk for having cardiovascular disease including AF when compared to 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 of sinus rhythm ECGs in VA patients could predict the presence of concurrent AF.

Methods

ECG dataset selection

We extracted all 12-lead ECGs acquired at sites within the VA's Veterans Integrated Services Network Region (VISN) 21, which includes 6 separate VA medical center networks (San Francisco, Palo Alto, Fresno, Sacramento, Reno, and the Pacific Islands), each of which is composed of multiple clinics. ECGs were performed from 1/1/1987 to 12/31/2022. ECG tracings were linked to cardiologist ECG interpretations, patient demographic (age, sex, and race/ethnicity), and comorbidity information (atrial fibrillation, heart failure, hypertension, diabetes, prior stroke/transient ischemic attack/thromboembolism, prior myocardial infarction, peripheral vascular

disease, chronic kidney disease) from the VA Corporate Data Warehouse. Comorbidities were determined using International Classification of Diseases (ICD) and Current Procedural Terminology (CPT) codes.[32] Using comorbidity fields, we estimated CHA_2DS_2-VASc scores.

We included only ECGs in sinus rhythm. We excluded ECGs that had poor data quality, paced rhythms, or could not be paired with age and sex information (a sign that a patient was not followed consistently in VA health system or that the ECG patient data was entered incorrectly and not linkable) (**Figure 1**). We limited our dataset 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 be reflective of a general AF screening population.

ECGs from the San Francisco VA and Palo Alto VA were used for model training, validation, and testing. ECGs from the Fresno VA, Sacramento VA, Reno VA, and the Pacific Islands VA were used as separate held-out test datasets.

For an external test dataset, we used all 12-lead ECGs acquired at Cedars-Sinai Medical Center, a large urban tertiary care center, from 3/1/2005 to 12/31/2018. The same inclusion and exclusion criteria were applied as were used for the VA dataset.

This study was approved by the University of California, San Francisco IRB and the Cedars-Sinai IRB.

Definition of cases and controls

Cases of concurrent AF were defined as sinus rhythm ECGs that could be paired with at least one ECG in atrial fibrillation or flutter (based on the cardiologist ECG interpretation) within 31 days (**Figure 1**). Controls were defined as sinus ECGs in patients who did not have ECGs in atrial fibrillation or flutter or diagnoses of atrial fibrillation or flutter by ICD/CPT coding. A single patient could 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 to be sinus rhythm ECGs that were closest to and chronologically before the first diagnosis of AF for each patient. ECGs had to be within 1 year before AF diagnosis.

ECG processing and deep learning model training

ECG tracings were extracted from the VA's MUSE Cardiology Information System (GE HealthCare). ECG waveform data was acquired at 250 Hz and extracted as 10 second, 12 x 2500 matrices of amplitude values, stored as

base64 text. ECGs underwent baseline wander correction using median filtering at 200ms and 600ms intervals and z-score normalization.

We employed an atrous convolutional neural network based on a novel architecture previously used for predicting clinical phenotypes from ECGs (**Supplemental Figure 1**).[33] The model was trained using PyTorch. 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 1e-4. The training dataset, composed of ECGs from the San Francisco VA 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 datasets.

Assessing model performance

All performance analyses were from model prediction of held-out VA datasets and the external Cedars-Sinai dataset not involved in model training. We compared the deep learning model's performance to clinical prediction of AF for held-out testing data from the VA and Cedars-Sinai using the CHA_2DS_2-VASc score and a logistic regression model that incorporated all available demographic and comorbidity information (age, sex, history of heart failure, diabetes, CVA/TIA/TE, prior MI, peripheral vascular disease, chronic kidney disease). These patient characteristics approximate those used in AF clinical risk prediction models such as CHARGE-AF.[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 curve (AUC) for the ROC curve. We reported the sensitivity, specificity, and accuracy at Youden's index (defined as the maximum value of sensitivity+specificity-1) as well as the maximum F1 score (harmonic mean of the precision (positive predictive value) and recall (sensitivity)).[35] All metrics were reported with two-sided 95% confidence intervals (CI) from 1000 bootstrapped samples. ROC curve AUCs were compared using DeLong's 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/positive predictive value.

For model calibration, risk scores underwent Platt scaling using logistic regression trained on 80% of the test dataset and then applied to a held-out 20% of the test dataset.[37] We visualized a calibration plot for this held-out 20% test dataset by plotting the observed versus 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 representing perfect accuracy and 1 meaning perfect inaccuracy. Calibration was tested using Spiegelhalter's z test at a significance threshold of 0.05. The null hypothesis of Spiegelhalter's

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 VA hospitals and Cedars-Sinai.

Statistical analysis was performed in R and Python.

Results

There were 2,420,508 12-lead ECGs acquired within our network of VA hospitals. After excluding ECGs that had poor data quality, paced rhythms, incomplete clinical info and were non-sinus 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 1,736 cases of AF within 31 days. In the VA cohort, ECGs were from patients who were on average 62.4 (SD 13.5) years old, 6.4% female, 10.7% Black, with a high prevalence of comorbidities (11.2% heart failure, 32.4% diabetes, 8.8% prior stroke (CVA)/transient ischemic attack (TIA)/thromboembolism (TE), 11.1% prior myocardial infarction (MI)) and a mean CHA_2DS_2-VASc score of 1.9 (1.6) (**Table 1**). In the external test cohort, patients had an average of 59.5 years (SD 15.4) and were 52.5% female and 9.4% Black. Compared to the VA population, there was a lower prevalence of

comorbidities (8.4% heart failure, 8.5% diabetes, 4.6% prior CVA/TIA/TE, 1.8% prior MI) and mean CHA_2DS_2-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 to controls, patients with concurrent AF were on average older (70.4 vs. 61.9 years old), less often female (3.8% vs. 10.0%), 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/TE, 25.4% vs. 9.9% prior MI) and CHA_2DS_2-VASc score (3.1 (1.8) vs. 1.9 (1.6)) (**Supplemental Table 1**).

The deep learning model was trained on 359,886 ECGs from the San Francisco VA and Palo Alto VA. When tested on held-out training datasets at these two VA sites, the model had AUCs of 0.88 (95% CI 0.87-0.90), 0.89 (0.89-0.90) with accuracies of 0.81 (0.79-0.83), 0.82 (0.81-0.83) and F1 scores of 0.33 (0.29-0.37) and 0.49 (0.47-0.51), respectively (**Figure 2A**). The model was then applied to four other VA sites that were not included in model training and achieved AUCs of 0.86 (0.85-0.87) (Fresno VA), 0.84 (0.83-0.85) (Sacramento VA), 0.84 (0.83-0.85) (Reno VA), 0.83 (0.79-0.88) (Pacific Islands VA). When tested on an external test set at Cedars-Sinai Medical Center, the model achieved an AUC of 0.93 (0.93-0.94).

The deep learning model was also well-calibrated with Brier scores of 0.02, 0.02, and 0.02 across all sites, VA hospitals, and Cedars-Sinai Medical Center, respectively (a Brier score of 0 indicates perfect calibration, 1 indicates perfect miscalibration) (**Figure 2B**). Testing by Spigelhalter's z test also confirmed a failure to reject the null hypothesis of model calibration at a significance threshold of 0.05 ($p = 0.06, 0.07, 0.39$ across all sites, VA hospitals, Cedars-Sinai Medical Center).

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 CHA_2DS_2-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 AUC of 0.86 (0.86-0.87), the risk factor regression model had an AUC of 0.73 (0.73-0.74), and the CHA_2DS_2-VASc score had an AUC of 0.70 (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 by the regression model and 12.01 by CHA_2DS_2-VASc score (**Figure 3**).

We tested the model's performance in specific patient cohort subsets (**Table 2, Supplemental Table 2**). Across the different sites, there were

substantial differences in the proportion of patients that were female (ranging from 4.8%-52.5%), Black (1.5%-17.2%), younger than 65 years old (48.1%-59.9%), and with a CHA_2DS_2-VASc score ≥ 2 (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 old and those with a CHA_2DS_2-VASc score ≥ 2. However, these differences were not observed consistently across all sites and performance was largely unchanged across the different subgroups.

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 AUCs ranging from 0.80 (0.79-0.81) to 0.85 (0.84-0.86) and accuracies from 0.73 (0.72-0.75) to 0.77 (0.76-0.78) at VA sites (**Supplemental Table 3, Supplemental Figure 2**). When tested on Cedars-Sinai ECGs, the AUC was 0.79 (0.78-0.79) with an accuracy of 0.72 (0.71-0.72).

Discussion

In this multi-site retrospective 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 atrial

fibrillation 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 could potentially help predict new onset atrial fibrillation within a longer 1-year time horizon.

Early detection of AF holds particular 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 studies 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.[40–46] However, most of these screening interventions are resource-intensive, sometimes invasive, and have not been adopted as part of routine clinical practice. One recent large randomized controlled trial of an AF screening program for all individuals 75-76 years old in two regions of Sweden 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 endpoint of stroke, bleeding, and mortality.[47]

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 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 2.5 patients for a test sensitivity of 25% and up to 11.5 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 CHA_2DS_2-VASc score, which had a number needed to screen of 11.5 and 12, respectively, for a test sensitivity of 25% and 25.4 and 20.8 for a test sensitivity of 75%. Our work builds upon previous research which has also used deep learning to identify AF from sinus ECGs with simulated and real pilot deployments in different patient populations.[48,27]

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 pre-test probability of disease, which could help limit the rate of false positives, as well as the higher average CHA_2DS_2-VASc score, which could increase the net benefit of starting anticoagulation.[29–31] The same characteristics that make the Veteran population particularly apt for AF screening, however, also make it different from other well-studied patient populations. These differences can pose challenges for the generalizability into and out from the VA for deep learning models, which remain limited in their interpretability and at risk for overfitting and confounding.[49] A recent study showed that a deep learning algorithm designed to recognize acute kidney injury did not perform equally well across VA and non-VA populations possibly due to differences in demographics (i.e. a significantly lower proportion of VA patients being female).[50]

We found that despite there being 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, there were small decreases in performance in patients who were older and had higher CHA_2DS_2-VASc scores. This could be because these patients had more comorbidities that introduced competing changes to the ECG and made predicting AF more difficult. Female patients in this cohort were also overall younger (69.7% < 65 years old 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 Cedars-Sinai. This may be because this cohort was relatively enriched for patients who were female, younger, and with a lower CHA_2DS_2-VASc score.

Limitations

Several limitations warrant consideration. As this 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 average 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 to using the model when screening a broader population of patients. Still, 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, it remains likely that there were patients in the control group who 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 stroke and thromboembolism.

Conclusion

A convolutional neural network trained using outpatient 12-lead ECGs in sinus rhythm from US Veterans successfully predicted the presence of AF within 31 days in populations of Veterans and non-Veterans 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.

Figure Legends

Figure 1. Cohort flow diagram

Inclusion and exclusion of 12-lead ECGs at 6 VA sites and Cedars-Sinai. All available ECGs were initially included and then excluded if they had poor data quality, paced rhythm, incomplete clinical information, were acquired during inpatient stays, or were non-sinus rhythm. The model was trained and validated on ECGs from the San Francisco and Palo Alto VA sites. The model was then tested on held-out ECGs from San Francisco and Palo Alto VA sites in addition to ECGs from 4 other VA sites and Cedars-Sinai.

Abbreviations: ECG = electrocardiogram, SF = San Francisco, PA = Palo Alto, Sac = Sacramento, PI = Pacific Islands

*A single ECG could fall into multiple exclusion categories (E.g. both a paced rhythm and non-sinus)

Figure 2. Model performance

A. Model discrimination performance characteristics for deep learning model trained on data from San Francisco and Palo Alto VA sites and tested on held out ECGs from these two sites as well as additional VA sites and Cedars-Sinai.

B. Model calibration performance characteristics. Observed versus predicted risk of AF for equal-sized groups of increasing predicted risk for all sites, VA hospitals only, and Cedars-Sinai only.

Abbreviations: AUC = area under the curve of the receiver operating characteristic curve

Figure 3. Deep learning model performance compared to clinical risk factor models.

Performance of deep learning model on all ECGs held out from model training compared to predicting AF using a clinical risk factors model (age, sex, history of heart failure, diabetes, stroke/transient ischemic attack/thromboembolism, prior myocardial infarction, peripheral vascular disease, chronic kidney disease) or CHA_2DS_2-VASc score.

Abbreviations: PPV = positive predictive value, NNS = number needed to screen to detect one true positive case of AF

Supplemental Figure 1. Study design schematic.

Outpatient 12-lead ECGS in sinus rhythm from the San Francisco and Palo Alto VA centers were used for model training. Cases of concurrent AF were defined as sinus ECGs with an AF ECG within 31 days. Controls were sinus ECGs with no AF by ECG or by diagnoses available in the electronic health records system. An atrous convolutional neural network was trained to predict cases and was then tested on held-out ECGs from San Francisco and Palo Alto VA sites in addition to ECGs from 4 other VA sites and Cedars-Sinai. The model was also tested in specific patient subgroup. Both prediction discrimination and calibration performance characteristics were reported.

Supplemental Figure 2. Model performance for exploratory analysis to simulate prediction of first case of AF within 1 year.

The model was used to predict the first case of AF within 1 year of a sinus rhythm ECG.

References

1. 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

2. 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

3. 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

4. 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

5. 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

6. 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

7. 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

8. 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

9. 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

10. 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

11. 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

12. 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

13. 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

14. Goette A, Borof K, Breithardt G, et al. 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

15. Kirchhof P, Camm AJ, Goette A, et al. Early Rhythm-Control Therapy in Patients with Atrial Fibrillation. *N Engl J Med*. 2020;383(14):1305-1316. doi:10.1056/NEJMoa2019422

16. Andrade JG, Wells GA, Deyell MW, et al. Cryoablation or Drug Therapy for Initial Treatment of Atrial Fibrillation. *N Engl J Med*. 2021;384(4):305-315. doi:10.1056/NEJMoa2029980

17. 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

18. 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

19. Adedinsewo 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

20. 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

21. 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

22. 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

23. 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 J, Lee SY, Jeon K, et al. Deep Learning–Based Algorithm for Detecting Aortic Stenosis Using Electrocardiography. *Journal of the American Heart Association*. 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-cncr2820030106>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.

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. Steinhubl 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. 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. Incidence of Previously Undiagnosed Atrial Fibrillation Using Insertable Cardiac Monitors in a High-Risk Population: The REVEAL AF Study. *JAMA Cardiology*. 2017;2(10):1120-1127. doi:10.1001/jamacardio.2017.3180

46. Healey JS, Alings M, Ha A, et al. Subclinical Atrial Fibrillation in Older Patients. *Circulation*. 2017;136(14):1276-1283. doi:10.1161/CIRCULATIONAHA.117.028845

47. Svennberg E, Friberg L, Frykman V, Al-Khalili F, Engdahl J, Rosenqvist M. Clinical outcomes in systematic screening for atrial fibrillation (STROKESTOP): a multicentre, parallel group, unmasked, randomised controlled trial. *Lancet*. 2021;398(10310):1498-1506. doi:10.1016/S0140-6736(21)01637-8

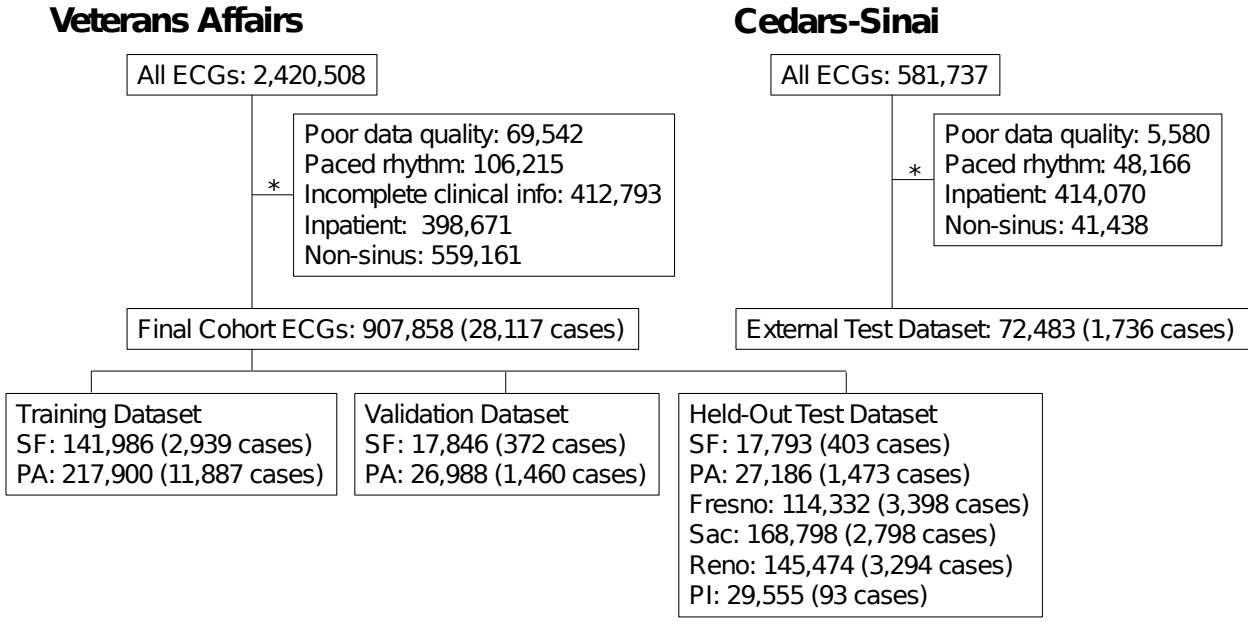
48. Noseworthy PA, Attia ZI, Behnken EM, et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. *Lancet*. 2022;400(10359):1206-1212. doi:10.1016/S0140-6736(22)01637-3

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*. Published online December 1, 2022:1-9. doi:10.1038/s42256-022-00563-8

699
700

Figure 1. Cohort flow diagram



Abbreviations: ECG = electrocardiogram, SF = San Francisco, PA = Palo Alto, Sac = Sacramento, PI = Pacific Islands

*A single ECG could fall into multiple exclusion categories (E.g. both a paced rhythm and non-sinus)

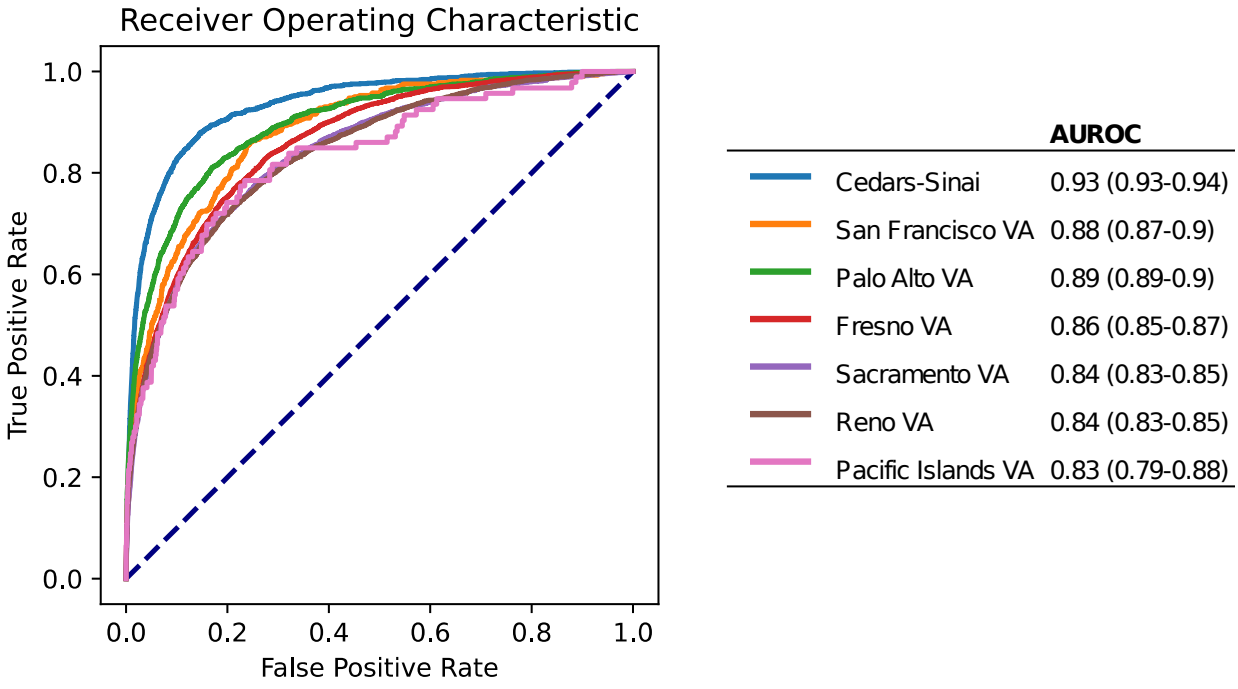
Table 1. ECG patient characteristics by site

	All VA Sites	San Francisco VA	Palo Alto VA	Fresno VA	Sacramento VA	Reno VA	Pacific Islands VA	Cedars-Sinai
n	907858	177625	272074	114332	168798	145474	29555	72483
ECGs/Patient (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 (SD)	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)
Female	58158 (6.4)	10820 (6.1)	18548 (6.8)	5440 (4.8)	13020 (7.7)	8796 (6.0)	1534 (5.2)	38068 (52.5)
Race (%)								
American Indian	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	24813 (2.7)	8257 (4.6)	9408 (3.5)	494 (0.4)	3562 (2.1)	196 (0.1)	2896 (9.8)	5743 (7.9)
Black	96912 (10.7)	31192 (17.6)	29646 (10.9)	4731 (4.1)	27930 (16.5)	2159 (1.5)	1254 (4.2)	6828 (9.4)
Latinx	41446 (4.6)	5691 (3.2)	18216 (6.7)	9617 (8.4)	6455 (3.8)	987 (0.7)	480 (1.6)	2119 (2.9)
Pacific Islander	6193 (0.7)	502 (0.3)	1179 (0.4)	142 (0.1)	1000 (0.6)	59 (0.0)	3311 (11.2)	20 (0.0)
White	566613 (62.4)	122725 (69.1)	205831 (75.7)	48544 (42.5)	121565 (72.0)	59010 (40.6)	8938 (30.2)	54245 (74.8)
Other	3690 (0.4)	591 (0.3)	1192 (0.4)	142 (0.1)	916 (0.5)	56 (0.0)	793 (2.7)	69 (0.1)
Unknown	166638 (18.4)	8337 (4.7)	6047 (2.2)	50544 (44.2)	7039 (4.2)	82842 (56.9)	11829 (40.0)	3393 (4.7)
HF	101548 (11.2)	20395 (11.5)	26827 (9.9)	15246 (13.3)	21168 (12.5)	14003 (9.6)	3909 (13.2)	6088 (8.4)
HTN	523776 (57.7)	97995 (55.2)	127289 (46.8)	78804 (68.9)	115605 (68.5)	83449 (57.4)	20634 (69.8)	14627 (20.2)
DM	294232 (32.4)	53360 (30.0)	76378 (28.1)	50328 (44.0)	60373 (35.8)	41881 (28.8)	11912 (40.3)	6170 (8.5)
CVA/TIA/TE	80006 (8.8)	15402 (8.7)	19365 (7.1)	11844 (10.4)	17689 (10.5)	13783 (9.5)	1923 (6.5)	3309 (4.6)
MI	100788 (11.1)	20954 (11.8)	29135 (10.7)	15305 (13.4)	18899 (11.2)	14123 (9.7)	2372 (8.0)	1339 (1.8)
PVD	37596 (4.1)	8339 (4.7)	7837 (2.9)	4316 (3.8)	7938 (4.7)	7829 (5.4)	1337 (4.5)	3740 (5.2)
CKD	92461 (10.2)	18226 (10.3)	21900 (8.0)	14347 (12.5)	21025 (12.5)	13490 (9.3)	3473 (11.8)	5943 (8.2)
CHADSVASc (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	28117 (3.1)	3714 (2.1)	14820 (5.4)	3398 (3.0)	2798 (1.7)	3294 (2.3)	93 (0.3)	1736 (2.4)

Abbreviations: SD = standard deviation, HF = heart failure, HTN = hypertension, DM = diabetes mellitus, CVA = cerebrovascular accident, TIA = transient ischemic attack, TE = thromboembolism, MI = myocardial infarction, PVD = peripheral vascular disease, CKD = chronic kidney disease, AF = atrial fibrillation

Figure 2. Model performance by test site

A. Model discrimination. Performance characteristics for deep learning model trained on data from San Francisco and Palo Alto VA sites and tested on held out ECGs from these two sites as well as additional VA sites and Cedars-Sinai



B. Model calibration. Observed versus predicted risk of AF for equal-sized groups of increasing predicted risk for all sites, VA hospitals only, and Cedars-Sinai only.

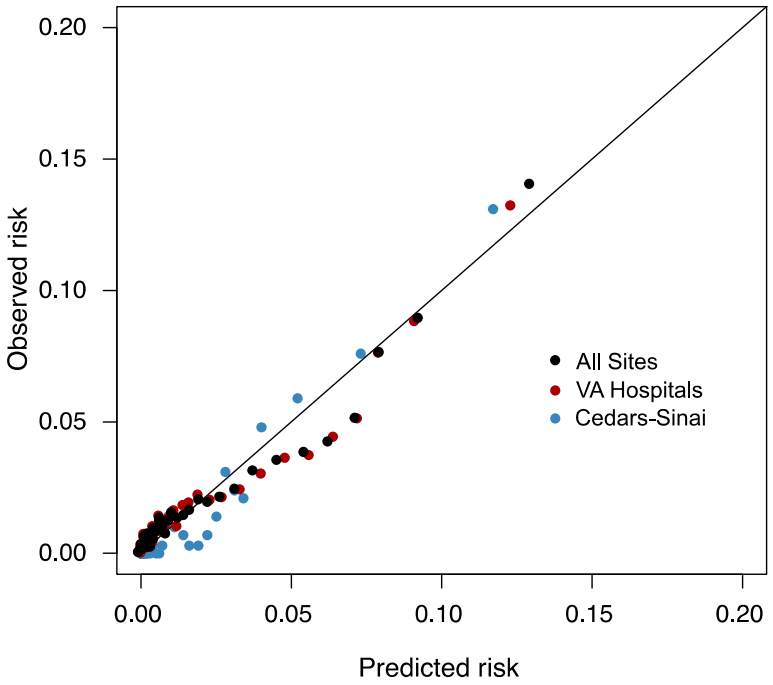


Figure 3. Deep learning model performance compared to clinical risk factor models. Performance of deep learning model on all ECGs held out from model training compared to predicting AF using a clinical risk factors model (age, sex, history of heart failure, diabetes, stroke/transient ischemic attack/thromboembolism, prior myocardial infarction, peripheral vascular disease, chronic kidney disease) or CHA_2DS_2-VASc score.

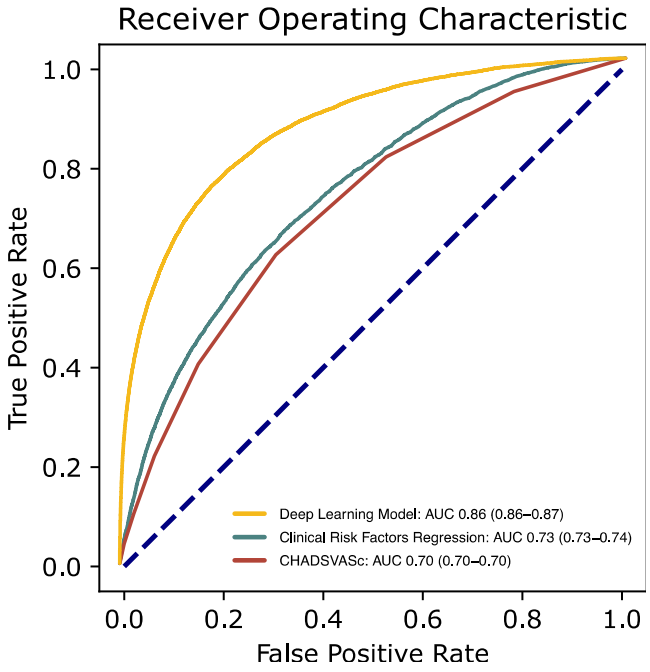


Table 2. Model performance in patient subgroups

	San Francisco VA	Palo Alto VA	Fresno VA	Sacramento VA	Reno VA	Pacific Islands VA	Cedars-Sinai
All Test Patients	17793	27186	114332	168798	145474	29555	72483
AUC All Patients	0.88 (0.87-0.9)	0.89 (0.89-0.9)	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)
Female (%)	1048 (5.9)	1785 (6.6)	5440 (4.8)	13020 (7.7)	8796 (6.0)	1534 (5.2)	38068 (52.5)
AUC in Female Patients	0.92* (0.88-0.97)	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)	0.95* (0.94-0.96)
Black (%)	3058 (17.2)	2827(10.4)	4731 (4.1)	27930 (16.5)	2159 (1.5)	1254 (4.2)	6828 (9.4)
AUC in Black Patients	0.9 (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)
Age < 65 y.o. (%)	9834 (55.3)	14884 (54.7)	55035 (48.1)	90427 (53.6)	75162 (51.7)	15549 (52.6)	43431 (59.9)
AUC in < 65 y.o.	0.88 (0.85-0.92)	0.90* (0.88-0.91)	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)
Age ≥ 65 y.o. (%)	7959 (44.7)	12302 (45.3)	59297 (51.9)	78371 (46.4)	70312 (48.3)	14006 (47.4)	29052 (40.1)
AUC in ≥ 65 y.o.	0.85 (0.83-0.88)	0.87 (0.86-0.89)	0.83* (0.82-0.84)	0.81* (0.8-0.82)	0.81* (0.8-0.82)	0.85 (0.8-0.89)	0.92* (0.91-0.93)
CHA_2DS_2-VASc ≥ 2 (%)	9340 (52.5)	12872 (47.3)	73633 (64.4)	101830 (60.3)	78041 (53.6)	17938 (60.7)	29990 (41.4)
AUC in CHA_2DS_2-VASc ≥ 2	0.86 (0.84-0.88)	0.87 (0.86-0.88)	0.84* (0.83-0.84)	0.82* (0.81-0.83)	0.82* (0.81-0.83)	0.84 (0.78-0.9)	0.92* (0.91-0.93)

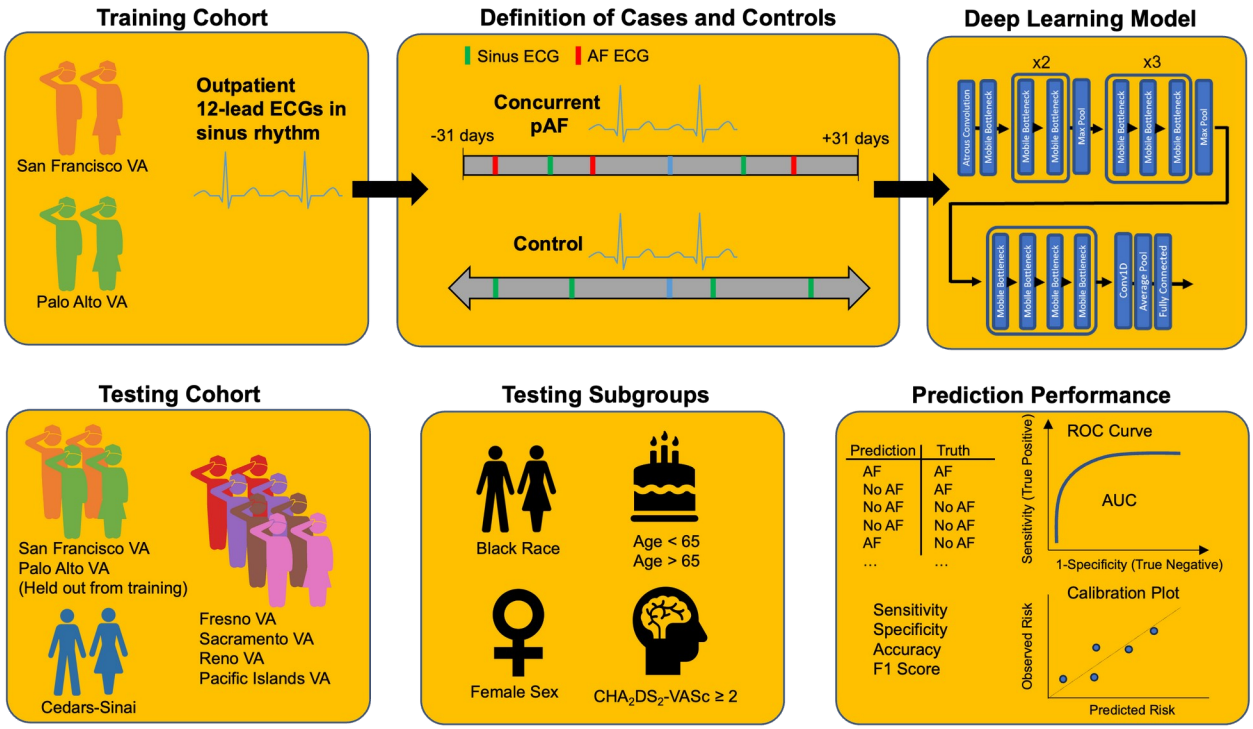
* = statistically significant, $p < 0.01$ when comparing to AUC for all patients at site

Supplementary Online Content

Yuan N, Duffy G, Dhruva SS, et al. Deep Learning in Electrocardiograms in Sinus Rhythm From US Veterans to Predict Atrial Fibrillation. *JAMA Cardiol*. Published online October 18, 2023. doi:10.1001/jamacardio.2023.3701

This supplementary material has been provided by the authors to give readers additional information about their work.

eFigure 1. Study Design Schematic
Outpatient 12-lead ECGS in sinus rhythm from the San Francisco and Palo Alto VA centers were used for model training. Cases of concurrent AF were defined as sinus ECGs with an AF ECG within 31 days. Controls were sinus ECGs with no AF by ECG or by diagnoses available in the electronic health records system. An atrous convolutional neural network was trained to predict cases and was then tested on held-out ECGs from San Francisco and Palo Alto VA sites in addition to ECGs from 4 other VA sites and Cedars-Sinai. The model was also tested in specific patient subgroup. Both prediction discrimination and calibration performance characteristics were reported.



eTable 1. ECG Patient Characteristics by Case or Control

	Concurrent AF	No Concurrent AF
n	29853	950488
ECGs/Patient	3.4 (4.9)	3.1 (4.0)
Age (SD)	70.4 (10.5)	61.9 (13.7)
Female	1147 (3.8)	95079 (10.0)
Race (%)		
American Indian	39 (0.1)	1580 (0.2)
Asian	674 (2.3)	29882 (3.1)
Black	1797 (6.0)	101943 (10.7)
Latinx	1042 (3.5)	42523 (4.5)
Pacific Islander	161 (0.5)	6052 (0.6)
White	23373 (78.3)	597485 (62.9)
Other	71 (0.2)	3688 (0.4)
Unknown	2696 (9.0)	167335 (17.6)
HF	11130 (37.3)	96506 (10.2)
HTN	22665 (75.9)	515738 (54.3)
DM	13443 (45.0)	286959 (30.2)
CVA/TIA/TE	4850 (16.2)	78465 (8.3)
MI	7573 (25.4)	94554 (9.9)
PVD	2655 (8.9)	38681 (4.1)
CKD	6469 (21.7)	91935 (9.7)
CHA_2DS_2-VASc (SD)	3.1 (1.8)	1.9 (1.6)

Abbreviations: SD = standard deviation, HF = heart failure, HTN = hypertension, DM = diabetes mellitus, CVA = cerebrovascular accident, TIA = transient ischemic attack, TE = thromboembolism, MI = myocardial infarction, PVD = peripheral vascular disease, CKD = chronic kidney disease, AF = atrial fibrillation

eTable 2. Model Discrimination Performance by Test Site.

Performance characteristics for deep learning model trained on data from San Francisco and Palo Alto VA sites and tested on held out ECGs from these two sites as well as additional VA sites and Cedars-Sinai

Site	AUROC	Sensitivity	Specificity	Accuracy	F1
Cedars-Sinai	0.93 (0.93-0.94)	0.87 (0.83-0.9)	0.87 (0.83-0.9)	0.87 (0.86-0.88)	0.46 (0.44-0.48)
San Francisco VA	0.88 (0.87-0.9)	0.86 (0.82-0.9)	0.76 (0.74-0.79)	0.81 (0.79-0.83)	0.33 (0.29-0.37)
Palo Alto VA	0.89 (0.89-0.9)	0.74 (0.72-0.76)	0.83 (0.81-0.85)	0.82 (0.81-0.83)	0.49 (0.47-0.51)
Fresno VA	0.86 (0.85-0.87)	0.78 (0.72-0.84)	0.78 (0.71-0.84)	0.78 (0.77-0.79)	0.32 (0.30-0.33)
Sacramento VA	0.84 (0.83-0.85)	0.75 (0.67-0.82)	0.78 (0.71-0.85)	0.76 (0.76-0.77)	0.24 (0.23-0.26)
Reno VA	0.84 (0.83-0.85)	0.73 (0.7-0.76)	0.79 (0.76-0.82)	0.76 (0.75-0.77)	0.28 (0.27-0.30)
Pacific Islands VA	0.83 (0.79-0.88)	0.77 (0.66-0.89)	0.80 (0.71-0.88)	0.78 (0.74-0.83)	0.18 (0.12-0.25)

eTable 3. Number Needed to Screen (NNS) Across Different Atrial Fibrillation Detection Sensitivities to Identify One True Case of Atrial Fibrillation

	Deep Learning Model		Risk Factors Regression		CHA_2DS_2VASc	
Sensitivity	**PPV**	**NNS**	**PPV**	**NNS**	**PPV**	**NNS**
0.10	0.61	1.65	0.10	9.68	0.10	10.03
0.25	0.40	2.47	0.09	11.48	0.08	12.01
0.50	0.19	5.40	0.06	17.59	0.06	15.63
0.75	0.09	11.53	0.04	25.39	0.05	20.76
0.90	0.05	20.75	0.03	31.58	0.04	26.25

eTable 4. Number Needed to Screen Across Patients Subgroups

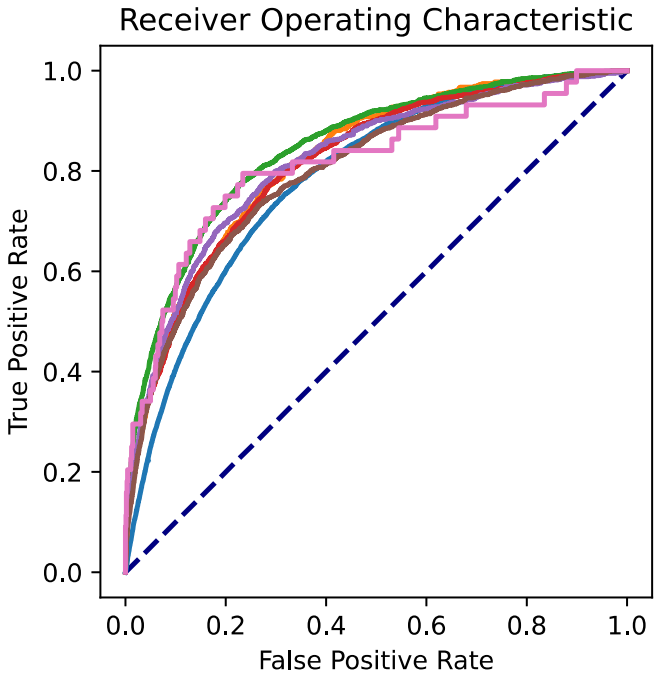
Number needed to screen to detect one true positive case of AF in patient subgroups across different sensitivities when deep learning model is applied to held out test data.

Sensitivity	All Patients	Female	Black	Age < 65	Age ≥ 65	CHA_2DS_2-VASc ≥ 2
0.1	1.65	1.83	1.94	1.82	1.58	1.56
0.25	2.47	2.17	3.41	2.94	2.31	2.33
0.5	5.4	3.31	7.65	8.1	4.8	4.89
0.75	11.53	9.89	15.77	20.1	9.61	9.96
0.9	20.75	26.3	30.69	39.19	15.82	16.56

eTable 5. ECG Patient Characteristics for Exploratory Analysis to Simulate Prediction of First Case of AF Within 1 Year

	All VA Sites	San Francisco VA	Palo Alto VA	Fresno VA	Sacramento VA	Reno VA	Pacific Islands VA	Cedars-Sinai
n	760976	126698	181893	112448	166939	143494	29504	306789
Age (SD)	62.58 (13.25)	62.97 (12.47)	62.02 (13.34)	63.96 (13.08)	62.09 (13.82)	62.56 (13.21)	61.82 (12.90)	62.15 (17.18)
Female	45707 (6.0)	6736 (5.3)	10314 (5.7)	5392 (4.8)	12970 (7.8)	8763 (6.1)	1532 (5.2)	145394 (47.4)
Race (%)								
American Indian	1306 (0.2)	249 (0.2)	394 (0.2)	118 (0.1)	327 (0.2)	164 (0.1)	54 (0.2)	379 (0.1)
Asian	18715 (2.5)	5602 (4.4)	6008 (3.3)	481 (0.4)	3537 (2.1)	196 (0.1)	2891 (9.8)	19020 (6.2)
Black	81236 (10.7)	24272 (19.2)	21161 (11.6)	4675 (4.2)	27737 (16.6)	2138 (1.5)	1253 (4.2)	47933 (15.6)
Latinx	34660 (4.6)	4410 (3.5)	12880 (7.1)	9535 (8.5)	6387 (3.8)	968 (0.7)	480 (1.6)	8668 (2.8)
Pacific Islander	5518 (0.7)	306 (0.2)	726 (0.4)	141 (0.1)	989 (0.6)	55 (0.0)	3301 (11.2)	241 (0.1)
White	457460 (60.1)	86666 (68.4)	136723 (75.2)	47250 (42.0)	120040 (71.9)	57870 (40.3)	8911 (30.2)	211646 (69.0)
Other	3038 (0.4)	392 (0.3)	750 (0.4)	142 (0.1)	909 (0.5)	56 (0.0)	789 (2.7)	693 (0.2)
Unknown	159043 (20.9)	4801 (3.8)	3251 (1.8)	50106 (44.6)	7013 (4.2)	82047 (57.2)	11825 (40.1)	18209 (5.9)
HF	87807 (11.5)	16873 (13.3)	19164 (10.5)	14341 (12.8)	20378 (12.2)	13147 (9.2)	3904 (13.2)	44935 (14.6)
HTN	460657 (60.5)	76735 (60.6)	90400 (49.7)	77152 (68.6)	113974 (68.3)	81802 (57.0)	20594 (69.8)	74555 (24.3)
DM	260372 (34.2)	42960 (33.9)	55739 (30.6)	49272 (43.8)	59502 (35.6)	41013 (28.6)	11886 (40.3)	36016 (11.7)
CVA/TIA/TE	71193 (9.4)	12727 (10.0)	14358 (7.9)	11435 (10.2)	17357 (10.4)	13397 (9.3)	1919 (6.5)	36737 (12.0)
MI	89655 (11.8)	18108 (14.3)	22422 (12.3)	14697 (13.1)	18407 (11.0)	13653 (9.5)	2368 (8.0)	20741 (6.8)
PVD	33795 (4.4)	7160 (5.7)	5787 (3.2)	4173 (3.7)	7760 (4.6)	7582 (5.3)	1333 (4.5)	33163 (10.8)
CKD	82415 (10.8)	15144 (12.0)	16442 (9.0)	13780 (12.3)	20532 (12.3)	13054 (9.1)	3463 (11.7)	34340 (11.2)
CHADSVASc (SD)	2.00 (1.62)	2.02 (1.66)	1.78 (1.61)	2.25 (1.63)	2.12 (1.60)	1.90 (1.60)	2.07 (1.48)	2.05 (1.76)
AF in 1 year	5628 (0.7)	578 (0.5)	1726 (0.9)	1151 (1.0)	916 (0.5)	1213 (0.8)	44 (0.1)	7170 (2.3)

eFigure 2. Model Performance for Exploratory Analysis to Simulate Prediction of First Case of AF Within 1 Year
The model was used to predict the first case of AF within 1 year of a sinus rhythm ECG.



Site	AUC	Sensitivity	Specificity	Accuracy	F1
Cedars-Sinai	0.79 (0.78-0.79)	0.74 (0.71-0.77)	0.70 (0.67-0.72)	0.72 (0.71-0.72)	0.33 (0.33-0.34)
San Francisco VA	0.82 (0.81-0.84)	0.78 (0.72-0.84)	0.72 (0.66-0.78)	0.75 (0.73-0.77)	0.10 (0.08-0.12)
Palo Alto VA	0.85 (0.84-0.86)	0.77 (0.74-0.80)	0.78 (0.75-0.80)	0.77 (0.76-0.78)	0.19 (0.17-0.2)
Fresno VA	0.82 (0.80-0.83)	0.77 (0.73-0.82)	0.71 (0.68-0.75)	0.74 (0.73-0.75)	0.15 (0.13-0.17)
Sacramento VA	0.82 (0.80-0.83)	0.75 (0.65-0.84)	0.76 (0.66-0.86)	0.75 (0.74-0.77)	0.11 (0.09-0.13)
Reno VA	0.80 (0.79-0.81)	0.72 (0.65-0.78)	0.75 (0.69-0.81)	0.73 (0.72-0.75)	0.12 (0.11-0.13)
Pacific Islands VA	0.81 (0.73-0.89)	0.77 (0.62-0.92)	0.82 (0.73-0.90)	0.79 (0.73-0.85)	0.17 (0.06-0.28)