Circulation

ORIGINAL RESEARCH ARTICLE

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

Editorial, see p 1299

BACKGROUND: Atrial fibrillation (AF) is associated with substantial morbidity, especially when it goes undetected. If new-onset AF could be predicted, targeted screening could be used to find it early. We hypothesized that a deep neural network could predict new-onset AF from the resting 12-lead ECG and that this prediction may help identify those at risk of AF-related stroke.

METHODS: We used 1.6 M resting 12-lead digital ECG traces from 430 000 patients collected from 1984 to 2019. Deep neural networks were trained to predict new-onset AF (within 1 year) in patients without a history of AF. Performance was evaluated using areas under the receiver operating characteristic curve and precision-recall curve. We performed an incidence-free survival analysis for a period of 30 years following the ECG stratified by model predictions. To simulate real-world deployment, we trained a separate model using all ECGs before 2010 and evaluated model performance on a test set of ECGs from 2010 through 2014 that were linked to our stroke registry. We identified the patients at risk for AF-related stroke among those predicted to be high risk for AF by the model at different prediction thresholds.

RESULTS: The area under the receiver operating characteristic curve and area under the precision-recall curve were 0.85 and 0.22, respectively, for predicting new-onset AF within 1 year of an ECG. The hazard ratio for the predicted high- versus low-risk groups over a 30-year span was 7.2 (95% CI, 6.9–7.6). In a simulated deployment scenario, the model predicted new-onset AF at 1 year with a sensitivity of 69% and specificity of 81%. The number needed to screen to find 1 new case of AF was 9. This model predicted patients at high risk for new-onset AF in 62% of all patients who experienced an AF-related stroke within 3 years of the index ECG.

CONCLUSIONS: Deep learning can predict new-onset AF from the 12-lead ECG in patients with no previous history of AF. This prediction may help identify patients at risk for AF-related strokes.

Sushravya Raghunath, PhD*

John M. Pfeifer, MD, MPH*

⋮

Brandon K. Fornwalt, MD, PhD†

Christopher M. Haggerty, PhD†

Clinical Perspective

What Is New?

- A deep learning model can identify patients at high risk for new-onset atrial fibrillation (AF).
- In patients with no history of AF who have an AF-related stroke, nearly two thirds would have been predicted to be high-risk for AF before the stroke by the deep learning model.

What Are the Clinical Implications?

- AF is a leading cause of stroke, and AF-related strokes can occur in patients with no known history of AF.
- A deep learning model capable of predicting future AF could be used in conjunction with a systematic monitoring strategy to find AF early and potentially prevent AF-related stroke.

Atrial fibrillation (AF) is a common cardiac rhythm disorder associated with several important adverse health outcomes including stroke and heart failure.[1–4] In patients with AF and risk factors for thromboembolism, early anticoagulation is effective at preventing strokes.[5–8] Unfortunately, AF is often unrecognized and untreated because it is frequently asymptomatic or minimally symptomatic.[9–11] Thus, methods to screen for and identify undetected AF are of significant interest[12–14] to ultimately prevent strokes.

Population-based screening for AF is challenging for 2 primary reasons. First, the yearly incidence of AF in the general population is low, with reported incidence rates of <10 per 1000 person-years younger than 70 years of age.[15–17] Second, AF is often paroxysmal, with many episodes lasting <24 hours.[18] At present, the most common screening strategy is opportunistic pulse palpation, sometimes in conjunction with a 12-lead ECG during routine medical visits. This has been shown to be cost-effective in certain populations and is recommended in some guidelines.[19–21] However, studies of implantable cardiac devices suggest that this strategy will miss many cases of AF.[10,11]

Many continuous monitoring devices are now available to detect paroxysmal and asymptomatic AF.[10,12,13] Patch monitors can be worn for up to 14 to 30 days, implantable loop recorders provide continuous monitoring for as long as 3 years, and wearable monitors such as the Apple Watch[13] can be worn indefinitely. Continuous monitoring devices overcome the problem of paroxysmal AF but must still contend with the overall low incidence of new-onset AF and cost and convenience limit their use for widespread population screening.

If future AF could be accurately predicted from a widely used and inexpensive test, this could identify a high-risk population that could then be screened with a continuous monitoring device. Machine learning, in particular deep neural networks (DNNs), can likely assist with this task. A recent study by Attia et al demonstrated the ability of a DNN to identify the electrocardiographic signature of paroxysmal AF from 12-lead ECGs showing sinus rhythm in a short time window.[22] A similar signature may be present in the ECG of patients without AF but who develop AF in the future. The prediction of truly future clinical outcomes from the ECG using machine learning methods is a new area of research with great potential. For example, recent work has demonstrated how a DNN can predict 1-year all-cause mortality directly from the 12-lead ECG with good performance, even in patients with ECGs clinically interpreted as normal.[23] In the present study, we trained a DNN to use ECGs to predict new-onset AF in patients with no history of AF. We then simulated a deployment scenario of this model retrospectively to demonstrate the high potential to identify patients who later have an AF-related stroke.

METHODS

Study data are available to researchers on reasonable request to the corresponding author. The methods can be reproduced based on details in the article; code will not be made available.

Data Selection and Phenotype Definitions

The Geisinger Institutional Review Board approved this retrospective study with a waiver of consent, in conjunction with our institutional patient privacy policies. We extracted 2.8 million standard 12-lead digital ECG traces from Geisinger's clinical MUSE (GE Healthcare, Milwaukee, WI) database, acquired between January 1984 and June 2019. Although 12-lead resting ECGs are acquired for 10-s, the ECG traces available for this study were in the standard clinical PDF format with 2.5-s traces for all 12 leads and 10-s rhythm strip traces for leads II, V1, and V5 (15 signal traces in total) at 500 Hz sampling frequency (42% of studies acquired at 250 Hz were resampled to 500 Hz by linear interpolation) and 1 μV resolution. We retained only ECGs (1) acquired in patients ≥18 years of age, and (2) with no significant artifacts as identified by the final ECG interpretation at the time of acquisition. This amounted to 1.6 million ECGs from 431 000 patients. The median (interquartile range) follow-up available after each ECG was 4.1 (1.5–8.5) years. Qualifying follow-up encounters were restricted to ECG, echocardiography, outpatient visit with internal medicine, family medicine or cardiology, any inpatient encounter, or any surgical procedure. An ECG was classified as normal if the findings text included strings that matched "normal ECG" or "within normal limits" and no other abnormalities were identified. All other ECGs were considered abnormal.

AF Outcome Definition

We excluded patients with preexisting or concurrent documentation of AF. The AF phenotype was defined as a clinically reported finding of AF from a 12-lead ECG or a diagnosis of AF applied to 2 or more inpatient or outpatient encounters or

AF listed on the patient problem list from our institutional electronic health record (August 1996 to January 2020). Any new diagnoses occurring within 30 days after cardiac surgery or within 1 year of a diagnosis of hyperthyroidism were excluded. Details on the applicable diagnostic codes and blinded chart review validation of the AF phenotype are provided in Methods in the Data Supplement and Table I in the Data Supplement. We chose to group atrial flutter with AF because the clinical consequences of the 2 rhythms are similar, including the risk of embolization, and because the 2 rhythms often coexist.

AF was considered to be new onset if it occurred at least 1 day after a baseline ECG that did not show AF in a patient with no known previous history of AF. This included patients with newly identified paroxysmal AF as well as incident AF. Electronic health record data were used to identify the most recent qualifying encounter date for censorship.

Model Development and Evaluation

We designed a deep convolutional neural network using only digital ECG traces as input in 3 temporally coherent branches. The data were restructured into 0- to 5-s signals for leads I, II, V1, and V5 in the first branch, 5- to 7.5-s signals for leads V1, V2, V3, II, and V5 in the second branch, and 7.5- to 10-s signals for leads II, V1, V4, V5, and V6 in the third branch (Figure I in the Data Supplement). The lead I signal between the 2.5- to 5-s interval was computed using the Goldberger equation[24] ($-aVR=[I+II]/2$) using signals from leads aVR and II.

The DNN model was designed to analyze the ECG signals to yield a predicted risk score for new-onset AF within 1 year of the ECG. The model architecture is illustrated in Figure II in the Data Supplement (details in Methods in the Data Supplement). A second instance of the model also included age and sex as input features to the DNN.

For all experiments, data were divided into training, internal validation, and test sets. The composition of the training and test sets varied by experiment, as described in Study Design; however, the internal validation set in all cases was defined as a 20% subset of the training data to track validation area under the receiver operating characteristic curve (AUROC) during training to avoid overfitting (details in Methods in the Data Supplement).

The models were evaluated using the AUROC, which is a robust metric of model performance for binary classification. Higher AUROC suggests higher performance (with perfect discrimination represented by an AUROC of 1, and an AUROC of 0.5 equivalent to a random guess). We also computed a precision-recall curve, which summarizes the tradeoff between the true positive rate (sensitivity or recall) and the positive predictive value (precision) for the model at different thresholds. The area under the precision recall curve (AUPRC) was calculated as the average precision score by computing the weighted average of precisions achieved at each threshold by the increase in recall (with perfect discrimination represented by an AUPRC of 1 and random chance equivalent to the proportion of target class in the data—for example, 0.04 [Figure 1]—for the holdout set defined in Study Design).

Study Design

We performed 2 separate modeling experiments (Figure 1):

(1) Proof-of-concept model: Using all ECGs from January 1984 to June 2019, a holdout set (20%) was identified at the beginning of the study (Figure 1A). The model was trained with the remaining 80% of the data. There was no overlap of patients between the holdout set and the training set. All ECGs with known time-to-event or at least 1 year of follow-up were used during model training, and a single random ECG was selected for each patient in the holdout set for model evaluation (Figure IIIA in the Data Supplement), with results denoted as "M0". Two versions of the model architecture were compared: one with ECG traces alone as inputs (DNN-ECG), and a second with ECG traces, age, and sex (DNN-ECG-AS). For comparison, we implemented an extreme gradient boosting (XGBoost)[25] model using only age and sex as inputs. We also compared the DNN model with the published CHARGE-AF (Cohorts for Aging and Research in Genomic Epidemiology) 5-year risk prediction model[26] in a subset of patients who had all of the data necessary to calculate a CHARGE-AF score.

To establish model stability and generalizability, we performed 5-fold cross-validation within the M0 model training set to derive models M1 to M5 and evaluated each on the respective unique fold test set (cross-validation test sets). There was no overlap of patients between the training set and cross-validation test set in each fold. As earlier, all qualifying ECGs were used during model training, and a single random ECG for a patient was chosen from the cross-validation test sets so as not to overweight patients with multiple ECGs (Figure 1A).

We also performed Kaplan-Meier (KM) incidence-free survival analysis[27] with the available follow-up data in the holdout set stratified by the model prediction for all 3 of the models (age and sex only, DNN-ECG, and DNN-ECG-AS), using an optimal operating point to stratify the population into low- and high-risk groups for new-onset AF. The optimal operating point was defined as the point on the ROC curve on the highest iso-performance line (equal cost to misclassification of positives and negatives) in the internal validation set. Patients who did not develop AF were censored at the most recent encounter. We fit a Cox proportional hazard model[28] regressing time to development of AF on the model-predicted classification of low-risk groups and high-risk groups. The hazard ratios (HR; adjusted for age and sex) were reported for the DNN model predictions, as well as for subpopulations defined by age groups (<50, 50–65, and ≥65 years), sex (men and women), and ECG type (normal and abnormal) for the holdout set.

(2) Simulated deployment model: To simulate a real-world deployment scenario—evaluating model performance in patients who later had an AF-related stroke—we used a second modeling approach (Figure 1B). Because a standard digital ECG contains information on age and sex, we used the DNN model that included age and sex for the deployment scenario. All ECGs from 1984 through 2009 were used as a training set. Next, we identified all patients with an ECG between January 1, 2010, and December 31, 2014. For each patient, we chose the ECG with the highest model prediction risk score, and those ECGs comprised the deployment test set. The dates were chosen to align with our institutional stroke registry, which began tracking patients in 2009 as described later.
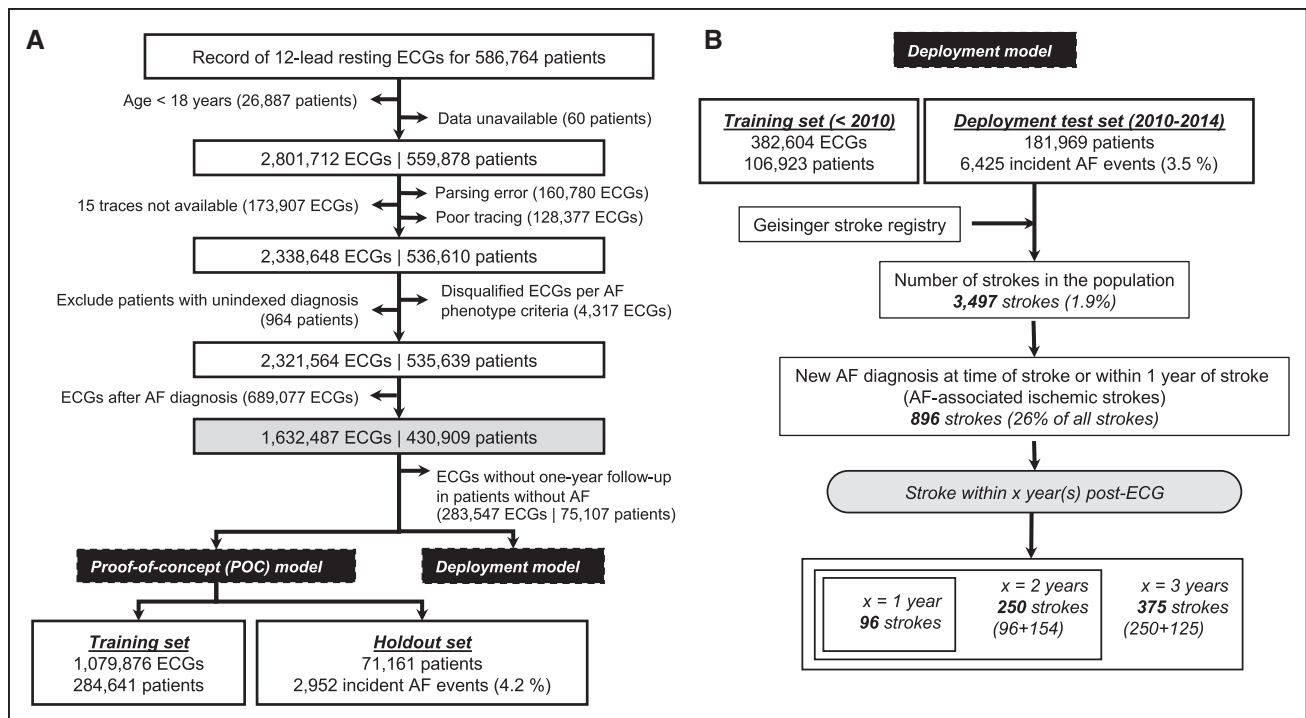
Figure 1. Flow chart illustrating the study design and data summary.
A, Data exclusions and data definition of proof-of-concept model. **B**, Data flow for deployment model. AF indicates atrial fibrillation.

To link deployment model predictions with potentially preventable stroke events, we leveraged an internal registry of patients diagnosed with acute ischemic stroke after 2009 at any of the 3 main Geisinger hospitals. From January 1, 2010, to December 31, 2017, representing the time interval included in this analysis, there were 6569 patients in the registry who were treated for an ischemic stroke. We used this registry to identify patients within the deployment model test set with an ischemic stroke subsequent to the test set ECG. A stroke was considered AF-related and potentially preventable if the following criteria were met: (1) the ECG in the test set was before the stroke, and the model predicted risk score was above the given operating point (ie, high risk for new-onset AF); and (2) previously undiagnosed AF was identified at the time of the stroke or up to 365 days after the stroke (Figure IIIB in the Data Supplement). To allow for time lag on emergency department and hospital admission notes, we included AF that was identified up to 2 days before the date of the qualifying stroke encounter. To allow for adequate follow-up, we included strokes that occurred within 3 years of the ECG (Figure 1B, Figure IIIB in the Data Supplement). A total of 96, 250, and 375 potentially preventable AF-related strokes were identified within 1, 2, and 3 years after ECG, respectively. We performed a chart review to determine whether those patients were on anticoagulation at the time of the stroke (details in Methods in the Data Supplement). We explored F_β scores (for β = 0.5, 1, and 2) and the Youden index[29] for model operating points in the internal validation set (Figure IV in the Data Supplement). Figure V in the Data Supplement shows the relationship between the proportions of new-onset AF and stroke (within 3 years of ECG) as a function of age in the deployment test set.

Statistical Analysis

Multiple AUROCs were compared by bootstrapping 1000 instances (using random and variable sampling with replacement). Differences between models were considered statistically significant if the absolute difference in the 95% CI was >0.

The KM analysis and HR for proof-of-concept model were computed using the lifelines package (version 0.24.1) in Python (version 3.6.8) and R (version 4.0.0).

RESULTS

DNN Model Predicts AF at 1 Year

The AUROC and AUPRC of the proof-of-concept DNN models for the prediction of new-onset AF within 1 year in the holdout set (M0) were 0.83 (95% CI, 0.83–0.84) and 0.21 (95% CI, 0.20–0.22), respectively, for DNN-ECG; and 0.85 (95% CI, 0.84–0.85) and 0.22 (95% CI, 0.21–0.24), respectively, for DNN-ECG-AS (Figure 2). This performance represents a significant improvement compared with the XGBoost model using only age and sex (AUROC, 0.78 [95% CI, 0.77–0.79]; AUPRC, 0.13 [95% CI, 0.12–0.14]; *P*<0.05 for difference in 95% CI by bootstrapping for both DNN models). In the holdout set, we had sufficient data to calculate CHARGE-AF scores for 65% of the patients. In this subset, the DNN-ECG-AS showed superior performance (AUROC, 0.84 [95% CI, 0.83–0.85]; AUPRC, 0.20 [95% CI, 0.19–0.22])
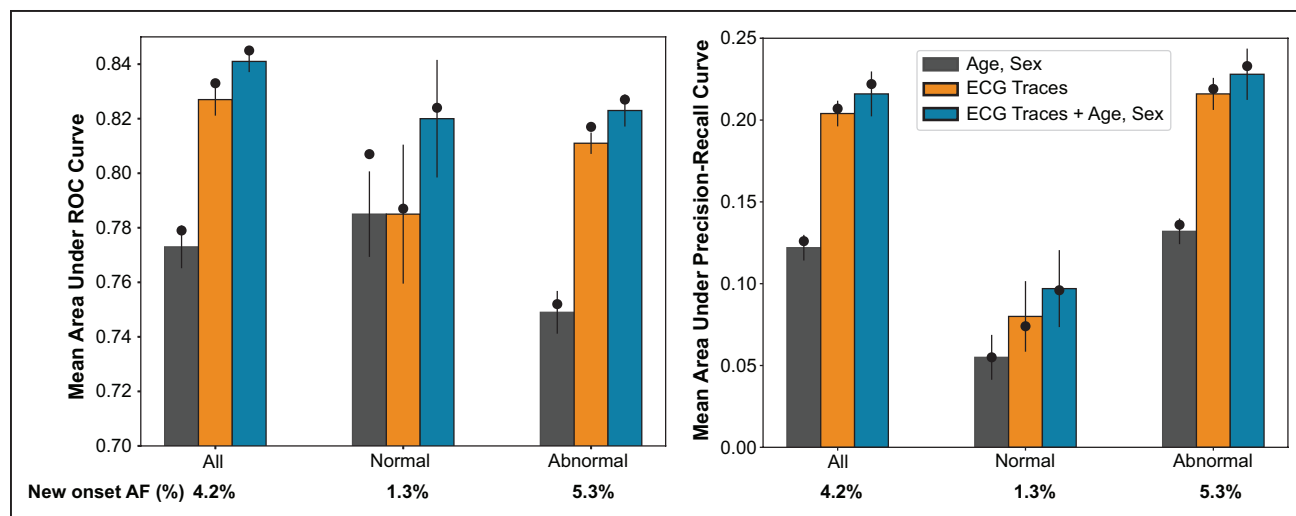
Figure 2. Illustration of model performance for all, normal, and abnormal subpopulations.
The model performance is represented for proof-of-concept model as area under the receiver operating characteristic (left) and precision-recall curves (right). The bars represent the mean performance across the 5-fold cross-validation with error bars showing 95% CIs. The black circle represents the M0 model performance on the holdout set. The 3 bars represent model performance for extreme gradient boosting (XGBoost) model with age and sex as inputs (gray); DNN model with digital ECG traces as input (DNN-ECG; orange); and DNN model with digital ECG traces, age, and sex as inputs (DNN-ECG-AS; blue). AF indicates atrial fibrillation; and ROC, receiver operating characteristic.

compared with the CHARGE-AF score (AUROC, 0.79 [95% CI, 0.78–0.80]; AUPRC, 0.12 [95% CI, 0.11–0.13]; Figure VI in the Data Supplement). The DNN models also maintained high performance within the subgroup of ECGs clinically reported as normal (Figure 2). These results were observed to be both generalizable and robust on the basis of the comparable performance of the M0 model on the holdout set and M1 to M5 (5-fold cross-validation models) on cross-validation test sets, as well as the stability of the M0 metrics with repeated iterations of random sampling within the holdout set (details in Methods in the Data Supplement). We simulated an external dataset by splitting the patient population by encounters at either Geisinger Medical Center or all other non–Geisinger Medical Center locations (Methods in the Data Supplement). The performance of the DNN-ECG-AS model trained on Geisinger Medical Center data and evaluated on a non–Geisinger Medical Center dataset was comparable with that obtained with the M0 model (AUROC, 0.85; and AUPRC, 0.21).

We also computed an AUROC of 0.87 (95% CI, 0.86–0.88; DNN-ECG model) for AF presenting exclusively between 1 to 31 days after the baseline ECG, consistent with the findings of Attia et al for the identification of paroxysmal AF from sinus rhythm.[22] We recognize that the DNN model both detects paroxysmal AF and predicts truly incident AF, and this is covered in detail in the Discussion.

DNN 1-Year AF Risk Prediction Is Associated With Long-Term AF Hazard

The KM curves and HR for the 3 AF-prediction models in Figure 2 are illustrated in Figure 3 with the operating points marked on the corresponding ROC curves (Figure 3A). The DNN models showed HRs of 6.7 (95% CI, 6.4–7.0) and 7.2 (95% CI, 6.9–7.6) in DNN-ECG and DNN-ECG-AS, respectively (Figure 3B). Adjusting for age (in increments of 10 years) and sex (interactions with sex and model were significant), the HR remained significant: 3.7 (95% CI, 3.6–4.1) and 3.1 (95% CI, 2.7–3.4) in women and men, respectively, for the DNN-ECG model and 3.8 (95% CI, 3.6–4.1) and 2.9 (95% CI, 2.5–3.4) in women and men, respectively, in the DNN-ECG-AS model (Figure 3C). For unadjusted comparisons, the DNN models had higher HR than the XGBoost model (age and sex) within all subsets defined by sex, age groups, and ECG type (normal or abnormal). Age alone is a powerful predictor of AF, so we further investigated the performance of the DNN-ECG-AS model by stratifying survival curves by age groups. Figure 4 (top row) shows the KM curves for age groups <50, 50 to 65, and ≥65 years in men and women. As expected, in both sexes, the survival curves are substantially different in each age group. However, Figure 4 (bottom row) shows that in each age group, the DNN model retains its ability to discriminate between high- and low-risk populations for the development of new-onset AF. The superiority of the DNN model over age and sex alone is most evident in younger age groups, and we note that no patient <58 years old was predicted as high-risk by the XGBoost model.

Prediction of New-Onset AF May Help Identify Patients at Risk of Future AF-Related Stroke

We observed 3497 patients out of 181969 (1.9%) with an ischemic stroke following an ECG within the
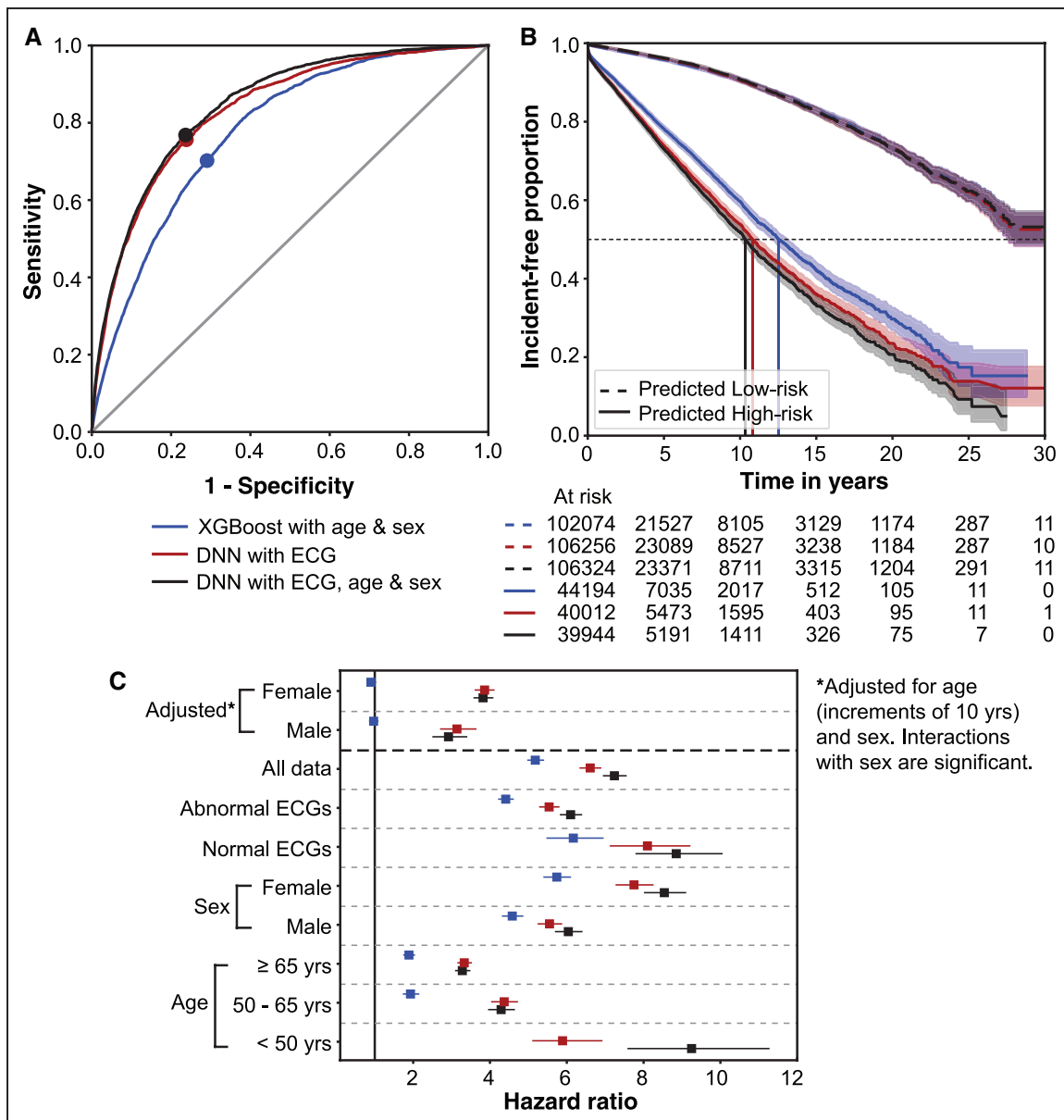
Figure 3. ROC curves, incidence-free KM survival curves, and HRs in subpopulations for the 3 models evaluated on the holdout set.
The 3 models are XGBoost model with age and sex only (blue); DNN model with ECG traces only (DNN-ECG; red); and DNN model with ECG traces, age, and sex (DNN-ECG-AS; black) for all ECGs in the holdout set. **A**, ROC curves with operating points marked for the 3 models. **B**, Incidence-free KM curves for the high- and low-risk groups for the operating point shown in **A** for a follow-up of 30 years. Note that curves corresponding to the low-risk group for all 3 models overlap. **C**, The plot of HR with 95% CIs for the 3 models in subpopulations defined by age groups, sex, and normal or abnormal ECG label. Note that there is no HR for age <50 years for the first model as there was no subject classified as high-risk for new-onset atrial fibrillation by the model for that subpopulation. DNN indicates deep neural network; HR, hazard ratio; KM, Kaplan-Meier; and ROC, receiver operating characteristic.

deployment test set (2010–2014). Of these, 96, 250, and 375 patients had a stroke within 1, 2, and 3 years, respectively, of an ECG and received a new diagnosis of AF within 365 days after the stroke. Of those 375 patients, 341 were not on an anticoagulant at the time of the stroke, 32 were on anticoagulant medications for reasons other than AF, and 2 patients had insufficient records to determine whether they were being treated with anticoagulants at the time of the stroke. Hence, these 375 represent a cohort at risk of AF-related strokes at the time of ECG. To reemphasize, we

hypothesized that the DNN would identify many of these ECGs as high-risk for AF.

Applying the model (trained on data before 2010) to this deployment test set, we again observed good performance for the prediction of new-onset AF at 1 year (AUROC, 0.83; AUPRC, 0.17). Using an operating point determined by the F_2 score, the sensitivity was 69%, the specificity was 81%, and the number needed to screen (NNS) to find 1 case of new-onset AF at 1 year was 9 (Table). In addition, 62% (231 of 375) of patients who had an AF-related stroke within 3 years of an ECG
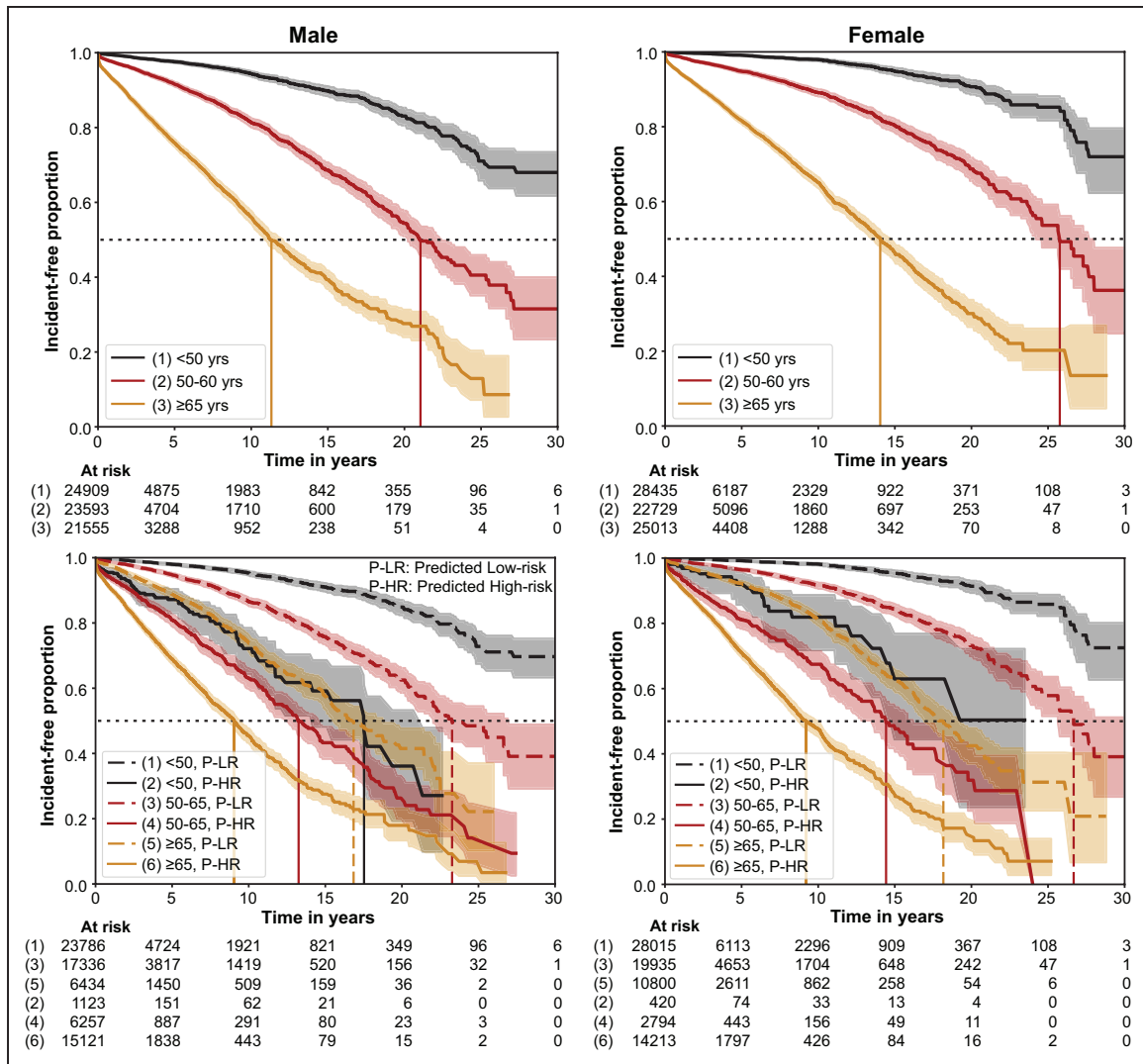
Figure 4. Incidence-free KM survival curves within the holdout set for subpopulations defined by sex, and age groups.
The top row shows the KM curves for subpopulations in age groups: <50 years, 50 to 65 years, and ≥65 years for men (left) and women (right). The bottom row shows the KM curves for the model-predicted (model M0 trained with ECG traces, age, and sex; DNN-ECG-AS) low-risk groups and high-risk groups for new-onset atrial fibrillation for each age group for men and women. It also reflects relative hazards between age groups. The horizontal dotted gray line represents incidence-free proportion of 50%, and vertical lines represent the median survival time for the respective curves. DNN indicates deep neural network; and KM, Kaplan-Meier.

were predicted to be high-risk for new-onset AF (Figure 5). The NNS to identify AF in 1 patient who developed an AF-related stroke within 3 years of a high-risk prediction was 162. The Table also shows favorable test characteristics in subgroups defined by age, sex, race, comorbidities, clinical setting, and CHA_2DS_2-VASc score.[30] The model performance and test characteristics at other operating points in Figure 5 are summarized in Table II in the Data Supplement.

DISCUSSION

We have shown that a DNN, trained on >1 million 12-lead resting ECGs, can predict new-onset AF within 1 year with good performance (AUROC, 0.85). We demonstrated that this DNN outperformed both a clinical model (CHARGE-AF) and an XGBoost model using age

and sex within the same dataset. We similarly note the superiority of our performance compared with the reported performances of other models in previous studies: CHARGE-AF (AUROC, 0.77), ARIC (Atherosclerosis Risk in Communities) (AUROC, 0.78), and Framingham heart study (AUROC, 0.78).[26,31,32] Moreover, the shorter prediction interval of our model (1 year compared with 5–10 years) allows for a more actionable prediction, and this prediction retains significant prognostic potential over the next 3 decades. We have shown that a large proportion of patients who had an AF-related stroke were predicted to be high risk for new-onset AF before stroke by the DNN model, demonstrating an important proof of concept for potentially using this model to prevent strokes through enhanced AF screening.

The DNN model is likely doing 2 different things: detecting paroxysmal AF and predicting incident AF. This is

Table. Performance Summary of the Deep Neural Network Model With Age and Sex for Predicting 1-Year New-Onset AF in a Deployment Scenario and the Potential to Identify Patients at Risk for AF-Related Stroke Within 3 Years of ECG

| Method/subgroup (operating threshold defined by F2 score) | Data | | New-onset AF within 1 year of ECG | | | | | Number predicted to be high risk for AF who developed an AF-related stroke within 3 y (number needed to screen) |
	Data (%)	AF incidence (%)	Proportion of ECGs flagged as high risk (%)	Number needed to screen to find 1 new-onset AF	Sensitivity (recall) (%)	Specificity (%)	
Full test set	100	3.5	21	9	69	81	231 (162)
Sex							
Men	45	4.1	25	9	70	77	109 (186)
Women	55	2.9	17	9	67	84	122 (141)
Race							
White	97	3.5	21	9	69	81	227 (162)
Black	2.3	1.7	11	13	49	90	3 (156)
Other	0.8	1.2	11	12	75	90	1 (179)
Comorbidities							
Coronary heart disease	9	7.8	52	8	84	50	66 (129)
Heart failure	1.3	18.8	77	4	92	27	17 (109)
Hypertension	47	4.6	28	9	70	74	162 (146)
Type 2 diabetes	14	5.3	33	8	74	69	63 (137)
None of the above	49	2.2	13	9	65	88	57 (202)
Patient setting							
Outpatient	49	2.1	13	13	51	87	63 (189)
Emergency	26	5.2	26	6	77	77	117 (105)
Inpatient	6	7.3	41	7	78	62	20 (232)
Unknown	18	3.4	27	11	73	75	31 (279)
Age group, y							
<50	32	0.5	2	15	23	98	2 (551)
50–65	33	2.2	12	12	47	89	23 (308)
Men, ≥65	15	8.4	54	8	81	48	91 (164)
Women, ≥65	19	6.7	42	8	76	61	115 (125)
CHA_2DS_2-VASc score[30]							
<2	53	1.4	7	12	43	93	18 (382)
≥2	47	5.8	36	8	76	66	213 (143)

Results are shown based on model predictions using the full test set, as well as specified population subsets with varying demographic, clinical setting, or comorbidity characteristics. AF indicates atrial fibrillation/flutter.

distinct from the study by Attia et al that focused solely on the identification of paroxysmal AF without claiming the ability to predict incident AF. As noted, the results indicate that our DNN model is doing both. Intuitively, the characteristics of the ECG that lead to a high-risk prediction by the DNN will be more prevalent in patients who already have AF but are currently in sinus rhythm. With this in mind, we expect a higher model performance for identification of paroxysmal AF compared with prediction of incident AF, and this is exactly what we see. We also expect a declining rate of new-onset AF over the course of 1 year. This is seen in Figure VII in the Data Supplement and is consistent with rapid identification of paroxysmal AF followed by a slower

identification of cases that represent incident AF. The largest piece of evidence supporting our assertion that the DNN model can predict incident AF is the continued separation of the KM incidence-free survival curves up to 30 years after the index ECG, as noted in Figures 3 and 4. In a retrospective analysis such as this, it is impossible to quantify how much of the new AF found within 1 year was detection of preexisting paroxysmal AF and how much was prediction of truly incident AF. However, from the perspective of preventing AF-related stroke, any finding of newly discovered AF is important, as it allows the opportunity to initiate anticoagulation.

More than 25% of all strokes are deemed a result of AF, and ≈20% of strokes caused by AF occur in
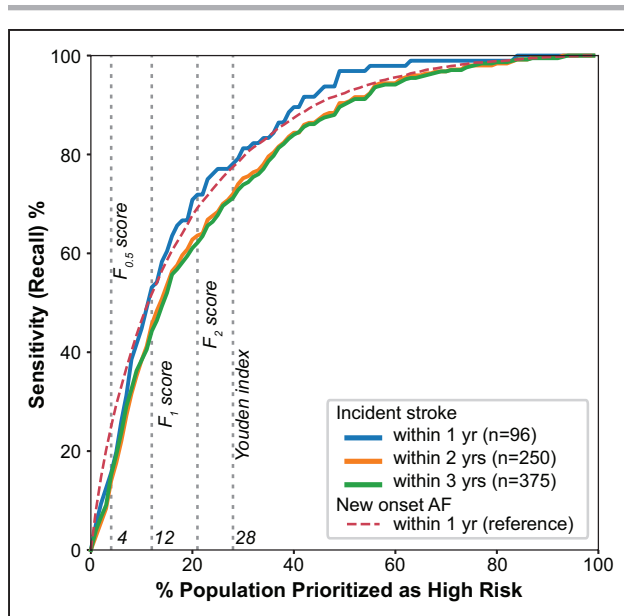
Figure 5. Illustration of model sensitivity to detect patients at risk of AF-related strokes as a function of the proportion of the population flagged as high risk to develop new-onset AF.
Colored curves denote patients with strokes occurring within 1 (blue), 2 (orange), and 3 (green) years after ECG in the deployment test set. Gray dotted lines represent the corresponding optimal operating thresholds from Table II in the Data Supplement. AF indicates atrial fibrillation.

individuals not previously diagnosed with AF.[33–35] Once AF is detected, anticoagulation is effective at preventing stroke, but screening for AF is difficult because of the paroxysmal nature of AF and the fact that it is often asymptomatic. Screening strategies involving patch monitors, wearables, and other devices can be used to detect AF but are most effective in populations with a high prevalence of AF. The underlying goal for developing this prediction model is to identify a high-risk population that can then be selected for additional monitoring with the goal of finding AF before a stroke.

We simulated such a real-world scenario by applying our model to all ECGs acquired within our large regional health system (Geisinger) over a 5-year period by cross-referencing predicted high-risk ECGs with future ischemic stroke incidences that were deemed potentially preventable (concurrent/subsequent identification of AF). We found that a high proportion (62%) of patients who had an AF-related stroke were correctly predicted as high-risk for AF. The NNS to identify AF in 1 patient who later had an AF-related stroke was 162. This compares favorably with other well-accepted screening tests, including mammography (NNS 476 to prevent 1 breast cancer death ages 60–69 years),[36] prostate specific antigen (NNS 1410 to prevent 1 death from prostate cancer),[37] and cholesterol (NNS 418 to prevent 1 death from cardiovascular disease).[38] Not all patients with AF are at high risk for stroke, and scoring systems such as CHA_2DS_2-VASc[30] are commonly used to determine the need for anticoagulation. A CHA_2DS_2-VASc score of 2

or greater is the cut point most commonly used to start an anticoagulant, and the Table shows that the model performs well within that subgroup, with an NNS of 8 to find 1 new case of AF. The Table also shows that 92% of patients predicted to be high-risk for AF who later had an AF-related stroke had a CHA_2DS_2-VASc score of 2 or greater and were potentially eligible for anticoagulation.

Three points are important to note in evaluating these findings. First, we have counted strokes occurring only at 3 Geisinger hospitals based on the exclusive use of an internal registry. Despite Geisinger's predominantly rural clinical population with low outmigration, some patients in the deployment test set likely had an incident stroke at another facility and were not captured in the registry. This leads to an underestimate of the number of patients at risk for stroke. Second, there was no systematic monitoring strategy to identify AF in the patients in our test set. Identification of new AF undoubtedly occurred in multiple ways, including fortuitous capture of asymptomatic AF as well as ECGs obtained in symptomatic patients. A systematic monitoring strategy implemented as part of the predictive model will capture more AF, as has been borne out in studies of continuous monitors. For example, in the mSTOPS trial (mHealth Screening to Prevent Stroke), monitoring with a patch monitor for up to 4 weeks identified new AF with an incidence of 6.7 per 100 person-years compared with 2.6 per 100 person-years without monitoring.[12] Third, our population of AF-related strokes was purposefully restricted by our definition that AF developed at the time of stroke or within 1 year after the stroke. We expect that some patients with an AF-related stroke would not have had their AF discovered in the 1 year after the stroke. For all of these reasons, we posit that the numbers we report for NNS with respect to both AF and stroke ascertainment represent worst-case scenarios of what would be prospectively realized. A prospective clinical trial is needed to confirm this speculation.

Once the ability to prevent strokes given this AF-prediction paradigm is demonstrated, this screening could be initiated in many different settings and performed through many different methods. With regard to setting, a promising opportunity—particularly for integrated care delivery systems—is the systematic screening of all ECGs in a health system. Specifically, the DNN could be incorporated into the existing workflow, such that every ECG is evaluated, and high-risk studies could be flagged for follow-up and surveillance. Such increased surveillance could take many different forms, including systematic pulse palpation, systematic ECG screening, continuous patch monitors worn once or multiple times, intermittent home screening with a device such as the Kardia mobile, or wearable monitors such as an Apple Watch.[12,13,39] Although these methods could be used in isolation to screen for AF, and many clinical trials are currently underway to that end,[40,41] combination with a DNN predictive model could help to overcome the challenges associated

with the overall low incidence of AF in the general population, especially in younger age groups. Age is generally the predominant risk factor in guiding AF screening strategies, yet in our study, 38% of all new AF (within 1 year of ECG) and 36% of all potentially preventable strokes (within 3 years of ECG) occurred among those younger than 70 years of age (Figure V in the Data Supplement). Our model can be used in all patients older than 18 years of age and outperformed a machine learning–based model that used age and sex alone.

Our focus in this article has been on the potential to prevent AF-related stroke by early identification of new-onset AF, but there are other ways in which a model that predicts future AF could be useful. AF is a frequent cause of arrhythmia-induced cardiomyopathy, and a hospital presentation with decompensated heart failure can be the first clinical manifestation of new-onset AF.[42] Enhanced surveillance in those predicted to be high-risk for future AF may therefore lead to a reduction in arrhythmia-induced cardiomyopathy. In addition to allowing early treatment for new-onset AF, a clinical risk prediction tool such as this could be used for the prevention of AF. A high-risk prediction of future AF could bring increased attention to modifiable risk factors such as obesity and obstructive sleep apnea, with the goal of avoiding AF altogether.

We acknowledge some limitations to our study. Although 10-s digital ECG traces are acquired during a resting 12-lead ECG, we had access to only 2.5-s for 9 of the leads and 10-s for the remaining 3 leads. A model using 10-s for all the leads could be considered in the future to maximize model training capabilities. Our analysis was limited to a single health system with a predominantly White population, so the generalizability to other organizations—particularly with a racially diverse population—must be established. We refer to the strokes in this study as potentially preventable, but in reality, identification of AF alone will not prevent all AF-related strokes. Some patients will either have a contraindication to or are not eligible for anticoagulation, and some who are treated with anticoagulation will still have a stroke. A chart review of the patients identified as having a potentially preventable AF-related stroke revealed that 9% of them were already on anticoagulation for reasons other than AF at the time of the stroke. It is unknown whether a diagnosis of new-onset AF would have affected the treatment plan or outcome in this small subset of patients. A prospective clinical trial is needed to confirm how many strokes can be prevented using a screening strategy on the basis of enhanced monitoring as a result of an AF risk prediction. This DNN approach represents a black-box model such that we do not know the specific features forming the basis of model predictions. Structural changes occur in the atria of patients with AF, and it is possible the DNN is using ECG manifestations of this atrial myopathy to guide the prediction.[43] Although previous work has shown some initial results for model interpretability specific to ECG-based DNN models for mortality predictions, these methods are challenging to generalize on a population level.[23] Acceptance of this limitation is warranted at the present time as more interpretable machine learning methods are not designed to directly leverage the digital ECG data as the DNN does, and there are currently no robust methods available to provide this insight into DNNs, although it remains an active area of investigation.

CONCLUSIONS

We have shown that a DNN can automatically analyze data from a resting 12-lead digital ECG to predict the risk of new-onset AF within 1 year with good performance. The model can both detect paroxysmal AF and predict incident AF. This predictive performance surpasses that of currently available clinical models, persists even within ECGs interpreted as normal, and is associated with significant hazard for AF development over the next 30 years. Preliminary data simulating a real-world deployment scenario demonstrate that using this tool identifies a high-risk population for new-onset AF that can be targeted for increased screening and may prove useful for helping to prevent AF-related strokes.

Authors

Sushravya Raghunath, PhD; John M. Pfeifer, MD, MPH; Alvaro E. Ulloa-Cerna, PhD; Arun Nemani, PhD; Tanner Carbonati, BS; Linyuan Jing, PhD; David P. vanMaanen, MS; Dustin N. Hartzel, BS; Jeffery A. Ruhl, MS; Braxton F. Lagerman, BCS; Daniel B. Rocha, MM; Nathan J. Stoudt, BS; Gargi Schneider, MD; Kipp W. Johnson, PhD; Noah Zimmerman, PhD; Joseph B. Leader, BA; H. Lester Kirchner, PhD; Christoph J. Griessenauer, MD; Ashraf Hafez, BS; Christopher W. Good, DO; Brandon K. Fornwalt, MD, PhD; Christopher M. Haggerty, PhD

Correspondence

Christopher M. Haggerty, PhD, 100 North Academy Avenue, Danville, PA 17822. Email chris.m.haggerty@gmail.com

Affiliations

Department of Translational Data Science and Informatics (S.R., A.E.U.-C., L.J., D.P.v.M, J.A.R., N.J.S., G.S., B.K.F., C.M.H.), Phenomic Analytics and Clinical Data Core (D.N.H., B.F.L., D.B.R., J.B.L.), Department of Population Health Sciences (H.L.K.), Department of Vascular and Endovascular Neurosurgery (C.J.G.), Heart Institute (C.W.G., B.K.F., C.M.H.), Department of Radiology (B.K.F.), Geisinger, Danville, PA. Heart and Vascular Center, Evangelical Hospital, Lewisburg, PA (J.M.P.). Tempus Labs Inc, Chicago, IL (A.N., T.C., K.W.J.,

N.Z., A.H.). Research Institute of Neurointervention, Paracelsus Medical University, Salzburg, Austria (C.J.G.). Heart and Vascular Institute at University of Pittsburgh Medical Center Hamot, Erie, PA (C.W.G.).

REFERENCES

1. Britton M, Gustafsson C. Non-rheumatic atrial fibrillation as a risk factor for stroke. *Stroke.* 1985;16:178–181. doi: 10.1161/01.STR.16.2.182
2. Wolf PA, Dawber TR, Thomas HE Jr, Kannel WB. Epidemiologic assessment of chronic atrial fibrillation and risk of stroke: the Framingham study. *Neurology.* 1978;28:973–977. doi: 10.1212/wnl.28.10.973
3. Stewart S, Hart CL, Hole DJ, McMurray JJ. A population-based study of the long-term risks associated with atrial fibrillation: 20-year follow-up of the Renfrew/Paisley study. *Am J Med.* 2002;113:359–364. doi: 10.1016/s0002-9343(02)01236-6
4. Gopinathannair R, Etheridge SP, Marchlinski FE, Spinale FG, Lakkireddy D, Olshansky B. Arrhythmia-induced cardiomyopathies: mechanisms, recognition, and management. *J Am Coll Cardiol.* 2015;66:1714–1728. doi: 10.1016/j.jacc.2015.08.038
5. Hart RG, Pearce LA, Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Ann Intern Med.* 2007;146:857–867. doi: 10.7326/0003-4819-146-12-200706190-00007
6. Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, Pogue J, Reilly PA, Themeles E, Varrone J, et al; RE-LY Steering Committee and Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med.* 2009;361:1139–1151. doi: 10.1056/NEJMoa0905561
7. Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, Breithardt G, Halperin JL, Hankey GJ, Piccini JP, et al; ROCKET AF Investigators. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med.* 2011;365:883–891. doi: 10.1056/NEJMoa1009638
8. Granger CB, Alexander JH, McMurray JJ, Lopes RD, Hylek EM, Hanna M, Al-Khalidi HR, Ansell J, Atar D, Avezum A, et al; ARISTOTLE Committees and Investigators. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med.* 2011;365:981–992. doi: 10.1056/NEJMoa1107039
9. Page RL, Wilkinson WE, Clair WK, McCarthy EA, Pritchett EL. Asymptomatic arrhythmias in patients with symptomatic paroxysmal atrial fibrillation and paroxysmal supraventricular tachycardia. *Circulation.* 1994;89:224–227. doi: 10.1161/01.cir.89.1.224
10. Reiffel JA, Verma A, Kowey PR, Halperin JL, Gersh BJ, Wachter R, Pouliot E, Ziegler PD; 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;2:1120–1127. doi: 10.1001/jamacardio.2017.3180
11. Healey JS, Connolly SJ, Gold MR, Israel CW, Van Gelder IC, Capucci A, Lau CP, Fain E, Yang S, Bailleul C, et al; ASSERT Investigators. Subclinical atrial fibrillation and the risk of stroke. *N Engl J Med.* 2012;366:120–129. doi: 10.1056/NEJMoa1105575
12. Steinhubl SR, Waalen J, Edwards AM, Ariniello LM, Mehta RR, Ebner GS, Carter C, Baca-Motes K, Felicione E, Sarich T, 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;320:146–155. doi: 10.1001/jama.2018.8102
13. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Balasubramanian V, Russo AM, Rajmane A, Cheung L, et al; Apple Heart Study Investigators. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med.* 2019;381:1909–1917. doi: 10.1056/NEJMoa1901183
14. Guo Y, Wang H, Zhang H, Liu T, Liang Z, Xia Y, Yan L, Xing Y, Shi H, Li S, et al; MAFA II Investigators. Mobile photoplethysmographic technology to detect atrial fibrillation. *J Am Coll Cardiol.* 2019;74:2365–2375. doi: 10.1016/j.jacc.2019.08.019
15. Heeringa J, van der Kuip DA, Hofman A, Kors JA, van Herpen G, Stricker BH, Stijnen T, Lip GY, Witteman JC. Prevalence, incidence and lifetime risk of atrial fibrillation: the Rotterdam study. *Eur Heart J.* 2006;27:949–953. doi: 10.1093/eurheartj/ehi825
16. Krahn AD, Manfreda J, Tate RB, Mathewson FA, Cuddy TE. The natural history of atrial fibrillation: incidence, risk factors, and prognosis in the Manitoba Follow-Up Study. *Am J Med.* 1995;98:476–484. doi: 10.1016/S0002-9343(99)80348-9
17. Lloyd-Jones DM, Wang TJ, Leip EP, Larson MG, Levy D, Vasan RS, D'Agostino RB, Massaro JM, Beiser A, Wolf PA, et al. Lifetime risk for development of atrial fibrillation: the Framingham Heart Study. *Circulation.* 2004;110:1042–1046. doi: 10.1161/01.CIR.0000140263.20897.42
18. Turakhia MP, Ziegler PD, Schmitt SK, Chang Y, Fan J, Than CT, Keung EK, Singer DE. Atrial fibrillation burden and short-term risk of stroke: case-crossover analysis of continuously recorded heart rhythm from cardiac electronic implanted devices. *Circ Arrhythm Electrophysiol.* 2015;8:1040–1047. doi: 10.1161/CIRCEP.114.003057
19. Meschia JF, Bushnell C, Boden-Albala B, Braun LT, Bravata DM, Chaturvedi S, Creager MA, Eckel RH, Elkind MS, Fornage M, et al; American Heart Association Stroke Council; Council on Cardiovascular and Stroke Nursing; Council on Clinical Cardiology; Council on Functional Genomics and Translational Biology; Council on Hypertension. Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke.* 2014;45:3754–3832. doi: 10.1161/STR.0000000000000046
20. Hobbs FD, Fitzmaurice DA, Mant J, Murray E, Jowett S, Bryan S, Raftery J, Davies M, Lip G. A randomised controlled trial and cost-effectiveness study of systematic screening (targeted and total population screening) versus routine practice for the detection of atrial fibrillation in people aged 65 and over. The SAFE study. *Health Technol Assess.* 2005;9:iii–iv, ix. doi: 10.3310/hta9400
21. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, Castella M, Diener HC, Heidbuchel H, Hendriks J, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur Heart J.* 2016;37:2893–2962. doi: 10.1093/eurheartj/ehw210
22. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet.* 2019;394:861–867. doi: 10.1016/S0140-6736(19)31721-0
23. Raghunath S, Ulloa Cerna AE, Jing L, vanMaanen DP, Stough J, Hartzel DN, Leader JB, Kirchner HL, Stumpe MC, Hafez A, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med.* 2020;26:886–891. doi: 10.1038/s41591-020-0870-z

ORIGINAL RESEARCH
ARTICLE

24. Goldberger E. The aVl, aVr, and aVf leads. A simplification of standard lead electrocardiography. *Am Heart J.* 1942;24:378–396. doi: 10.1016/S0002-8703(42)90821-4

25. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, NY: Association for Computing Machinery. 2016:785–794.

26. Alonso A, Krijthe BP, Aspelund T, Stepas KA, Pencina MJ, Moser CB, Sinner MF, Sotoodehnia N, Fontes JD, Janssens AC, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *J Am Heart Assoc.* 2013;2:e000102. doi: 10.1161/JAHA.112.000102

27. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Am Stat Assoc.* 1958;53:457–481. doi: 10.2307/2281868

28. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B.* 1972;34:187–202. doi: 10.1111/j.2517-6161.1972.tb00899.x

29. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3:32–35. doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3

30. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM, Andresen D, Camm AJ, Davies W, Capucci A, Olsson B, et al. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the Euro Heart Survey on atrial fibrillation. *Chest.* 2010;137:263–272. doi: 10.1378/chest.09-1584

31. Chamberlain AM, Agarwal SK, Folsom AR, Soliman EZ, Chambless LE, Crow R, Ambrose M, Alonso A. A clinical risk score for atrial fibrillation in a biracial prospective cohort (from the Atherosclerosis Risk in Communities [ARIC] study). *Am J Cardiol.* 2011;107:85–91. doi: 10.1016/j.amjcard.2010.08.049

32. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB Sr, Newton-Cheh C, Yamamoto JF, Magnani JW, Tadros TM, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet.* 2009;373:739–745. doi: 10.1016/S0140-6736(09)60443-8

33. Hannon N, Sheehan O, Kelly L, Marnane M, Merwick A, Moore A, Kyne L, Duggan J, Moroney J, McCormack PM, et al. Stroke associated with atrial fibrillation–incidence and early outcomes in the north Dublin population stroke study. *Cerebrovasc Dis.* 2010;29:43–49. doi: 10.1159/000255973

34. Asplund K. High prevalence of atrial fibrillation among patients with ischemic stroke. *Stroke.* 2014;45:2599–2605. doi: 10.1161/STROKEAHA.114.006070

35. Lin HJ, Wolf PA, Benjamin EJ, Belanger AJ, D'Agostino RB. Newly diagnosed atrial fibrillation and acute stroke. The Framingham Study. *Stroke.* 1995;26:1527–1530. doi: 10.1161/01.str.26.9.1527

36. Siu AL; U.S. Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med.* 2016;164:279–296. doi: 10.7326/M15-2886

37. Schröder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, Kwiatkowski M, Lujan M, Lilja H, Zappa M, et al; ERSPC Investigators. Screening and prostate-cancer mortality in a randomized European study. *N Engl J Med.* 2009;360:1320–1328. doi: 10.1056/NEJMoa0810084

38. Rembold CM. Number needed to screen: development of a statistic for disease screening. *BMJ.* 1998;317:307–312. doi: 10.1136/bmj.317.7154.307

39. Koltowski L, Balsam P, Głowczynska R, Rokicki JK, Peller M, Maksym J, Blicharz L, Maciejewski K, Niedziela M, Opolski G, Grabowski M. Kardia Mobile applicability in clinical practice: a comparison of Kardia Mobile and standard 12-lead electrocardiogram records in 100 consecutive patients of a tertiary cardiovascular care center [published online January 15, 2019]. *Cardiol J.* doi: 10.5603/cj.a2019.0001. https://journals.viamedica.pl/cardiology_journal/article/view/58839

40. A Study to Determine If Identification of Undiagnosed Atrial Fibrillation in People at Least 70 Years of Age Reduces the Risk of Stroke (GUARD-AF). clinicaltrials.gov. 2019. Accessed January 4, 2021. https://clinicaltrials.gov/ct2/show/NCT04126486.

41. Aliot E, Brandes A, Eckardt L, Elvan A, Gulizia M, Heidbuchel H, Kautzner J, Mont L, Morgan J, Ng A, et al. The EAST study: redefining the role of rhythmcontrol therapy in atrial fibrillation: EAST, the Early treatment of Atrial fibrillation for Stroke prevention Trial. *Eur Heart J.* 2015;36:255–256. doi: 10.1093/eurheartj/ehu476

42. Huizar JF, Ellenbogen KA, Tan AY, Kaszala K. Arrhythmia-induced cardiomyopathy: JACC State-of-the-Art Review. *J Am Coll Cardiol.* 2019;73:2328–2344. doi: 10.1016/j.jacc.2019.02.045

43. Kottkamp H. Human atrial fibrillation substrate: towards a specific fibrotic atrial cardiomyopathy. *Eur Heart J.* 2013;34:2731–2738. doi: 10.1093/eurheartj/eht194

44. Morandi A, Limousin M, Sayers J, Golwala SR, Czakon NG, Pierpaoli E, Jullo E, Richard J, Ameglio S. X-ray, lensing and Sunyaev-Zel'dovich triaxial analysis of Abell 1835 out to R 200. *Mon Not R Astron Soc.* 2012;425:2069–2082.

45. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015 Accessed January 4, 2021. http://arxiv.org/abs/1502.03167.

46. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. *J Mach Learn Res.* 2011;12:2121–2159.

47. Lin M, Chen Q, Yan S. Network in network. In: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings. 2014. Accessed January 4, 2021. https://arxiv.org/abs/1312.4400.

48. Prechelt L. Early stopping-but when? In: *Neural Networks: Tricks of the Trade.* Springer; 1998:55–69.