SUPPLEMENTAL MATERIAL

EXPANDED METHODS

Atrial Fibrillation (AF) phenotype:

The following criteria were used for our AF phenotype definition –

Case 1

- Inclusion - Clinically reported finding of atrial fibrillation or atrial flutter from any 12-lead resting ECG.

Case 2

- Inclusion - Atrial fibrillation or atrial flutter diagnosis on the patients' problem list OR 2 or more encounter diagnoses.

Exclusion - Patients diagnosed with hyperthyroidism within 12 months prior to the ECG and patients with cardiac surgery within 30 days prior to the ECG.

Performance measures for the blinded chart review of 200 patients is shown in Supplemental Table I. In this review, patients were selected based on a structured phenotype definition for AF. Patients who met the structured phenotype definition for AF (n=100) along with patients who did not meet the definition (controls; n=100) were pulled from our electronic health record. The patients' medical record numbers and the associated AF index date for each were then provided to a cardiologist for chart review to confirm or negate a diagnosis of AF as of the index date. The cardiologist remained blinded to whether the patients they were reviewing met or did not meet our phenotype definition.

We included AF diagnoses by identifying relevant ICD 10 codes (I48.0, I48.1, I48.2, I48.3 and I48.91), ICD 9 codes (427.31) and 92 separate internal codes.

Input Data and Model architecture:

Input to the model architecture includes digital ECG traces and, for a second instance of the model, age and sex. The ECG input structure to the model includes "branch 1" comprising leads I, II, V1, and V5, acquired from time (t) = 0 (start of data acquisition) to t=5 seconds; "branch 2" comprising leads V1, V2, V3, II, and V5 from t=5 to t=7.5 seconds; and finally "branch 3" comprising leads V4, V5, V6, II,

and V1 from t=7.5 to t=10 seconds. This was designed to account for concurrent morphology changes throughout the standard clinical acquisition. All of the ECG traces were preprocessed to ensure that waveforms were centered around the zero baseline, while preserving variance and magnitude features. For the model including age and sex, sex was encoded as 1, 0 or 0.5 for male, female and unknown values respectively, and age was computed as days since patient's birth date from the ECG test date.

The deep neural network (DNN) model architecture comprises two major components: the convolutional component and the fully connected dense layer component. The convolutional component starts with an input for each branch followed by a convolutional block. Each convolutional block consists of a 1D convolutional layer, RELU activation function, and a batch norm layer, in series[45,46]. Next this convolutional block is followed by four inception blocks in series, where each inception block comprises three 1D convolutional blocks concatenated across the channel axis with decreasing filter window sizes[47]. Each of the four inception blocks are connected to a 1D maxpooling layer, where they are connected to another single convolutional block and a final global averaging pool layer[48]. The outputs for all three branches are concatenated and fully connected to the dense layer component. This dense layer component contains 4 dense layers of 256, 64, 8 and 1 unit(s) with a sigmoid function as the final layer. All layers in the architecture enforce kernel constraints and have no bias terms. Age and sex were input into a 64-unit hidden layer that was concatenated with the other branches. We used the AdaGrad optimizer[47] with a learning rate of $1e^{-2}$[45], a linear learning rate decay of 1/10 prior to early stopping[49] for efficient model convergence at patience of 3 epochs, and batch size of 2048. The patience for early stopping[49] was set to 9 epochs. The model was implemented using Keras (version: 2.2.4-tf) with a TensorFlow backend (version: 1.14.0) in python (version: 3.6.8) and default training parameters were used except where specified. All training was performed on NVIDIA DGX1 and DGX2 machines with eight and sixteen V100 GPUs and 32 GB of RAM per GPU, respectively.

Operating points for deployment model:

To account for potential variability in the clinical implementation of such a model (i.e., matching the performance to the scope of available resources and desired screening characteristics), we evaluated performance across a range of operating points (thresholds of the model risk that were used to classify

low or high risk for developing new onset AF). These thresholds were defined based on maxima of the F_β score (for β = 0.5, 1, and 2) and Youden's index[29] within the internal validation set (Supplemental Figure IV). F_β scores are functions of precision and recall. A β value of 1 is the harmonic mean of precision and recall (e.g. sensitivity), a value of 2 emphasizes recall, and value of 0.5 attenuates the influence of recall correspondingly. Youden's index combines sensitivity and specificity measures.

Chart review of strokes for anticoagulation medication:

All patients identified as having a potentially preventable AF-related stroke were chart reviewed by a cardiologist to determine if they were on anticoagulation at the time of the stroke. Anticoagulant medications considered were warfarin, dabigatran, apixaban, rivaroxaban, edoxaban, enoxaparin, tinzaparin, dalteparin, fondaparinux.

Random ECG selection for test set in DNN prediction proof-of-concept (POC):

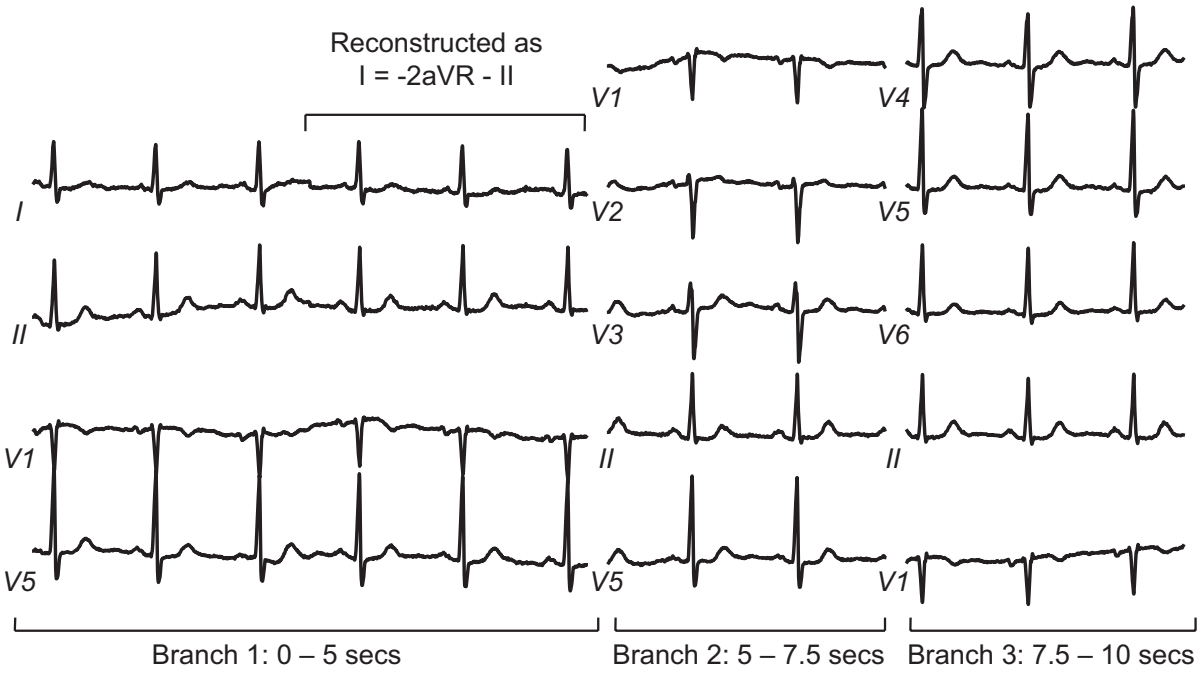
To demonstrate that there was no bias from selecting a single random ECG from each patient in the POC model we performed 100 random iterations of selections and found that performance of the M0 model was stable with mean and standard deviation of AUROCs and AUPRCs of 0.834 ± 0.002 and 0.209 +/- 0.004, respectively, for DNN-ECG (the model with input of ECG traces); and, 0.845 ± 0.002 and 0.220 ± 0.004 for DNN-ECG-AS (model with input of ECG traces with age and sex).

Additional validation, simulating external dataset:

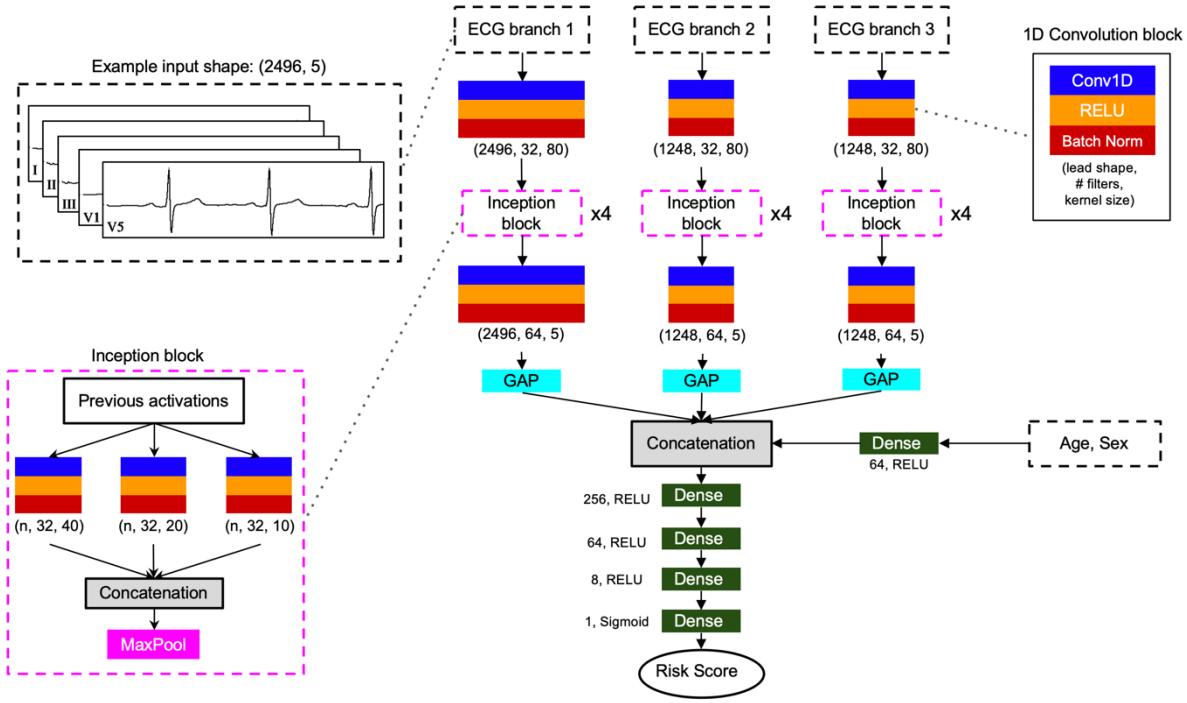
The ECGs were classified by the location where the ECG was taken as 'GMC' or 'non-GMC'. For GMC patients the ECGs were acquired at Geisinger Medical Center (GMC) in Danville, PA. Non-GMC patients had ECGs acquired at other facilities within the Geisinger system, comprising a mix of hospital and community clinic settings. All patients in the non-GMC group who were also in the GMC group were removed from non-GMC, such that there was no overlap of patients between the two groups. A model was trained with all ECGs relative to their events from the GMC group (380,433 ECGs from 131,472 patients) and tested on the non-GMC group (202,909 ECGs from 202,909 patients). A single random ECG for a patient was chosen in the test set for evaluation. The test set had 7,791 new onset AF events

in 202,909 ECGs with 912 new onset AF events among 86,878 normal ECGs and 6,879 events among 135,775 abnormal ECGs.

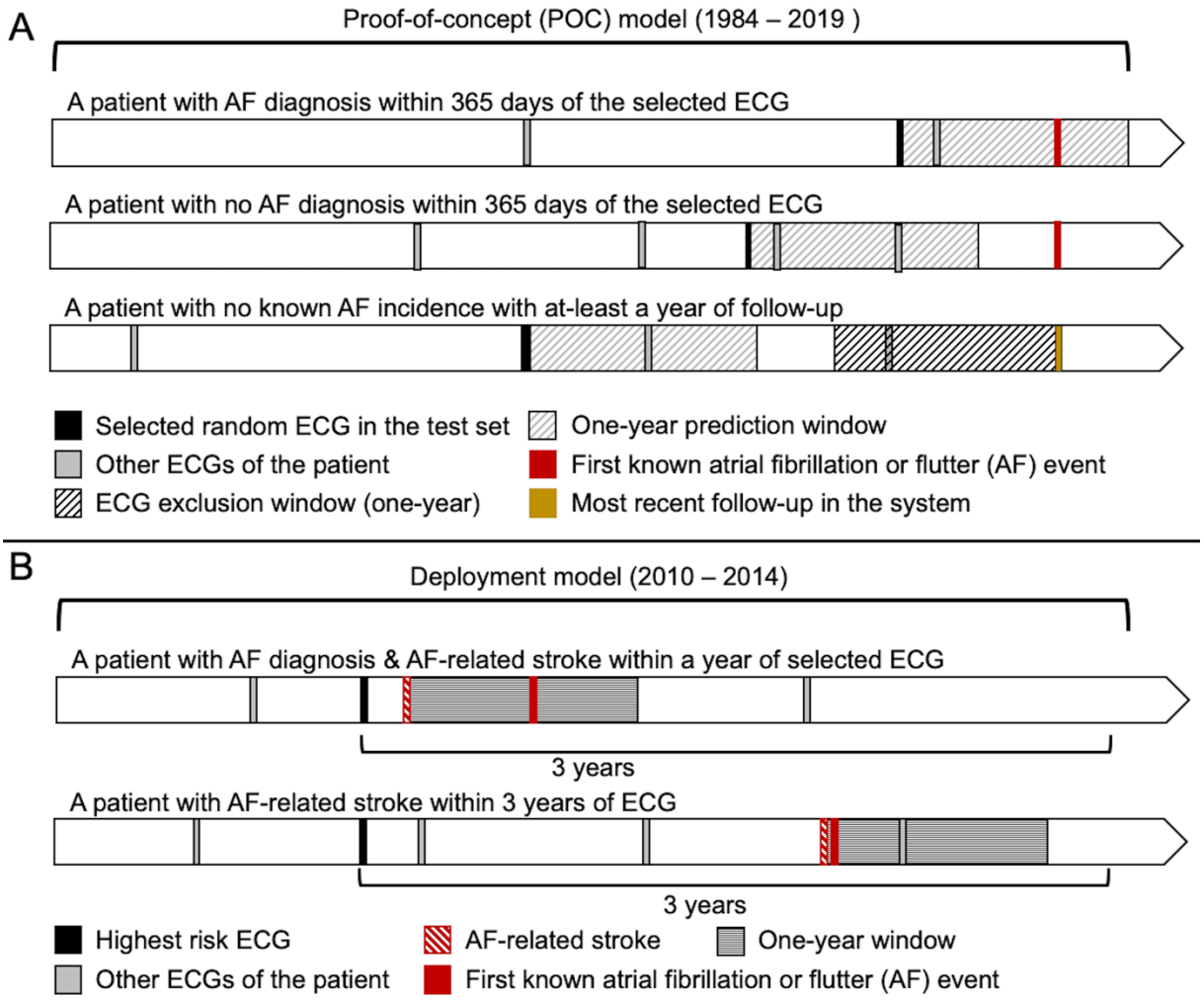
SUPPLEMENTAL FIGURES



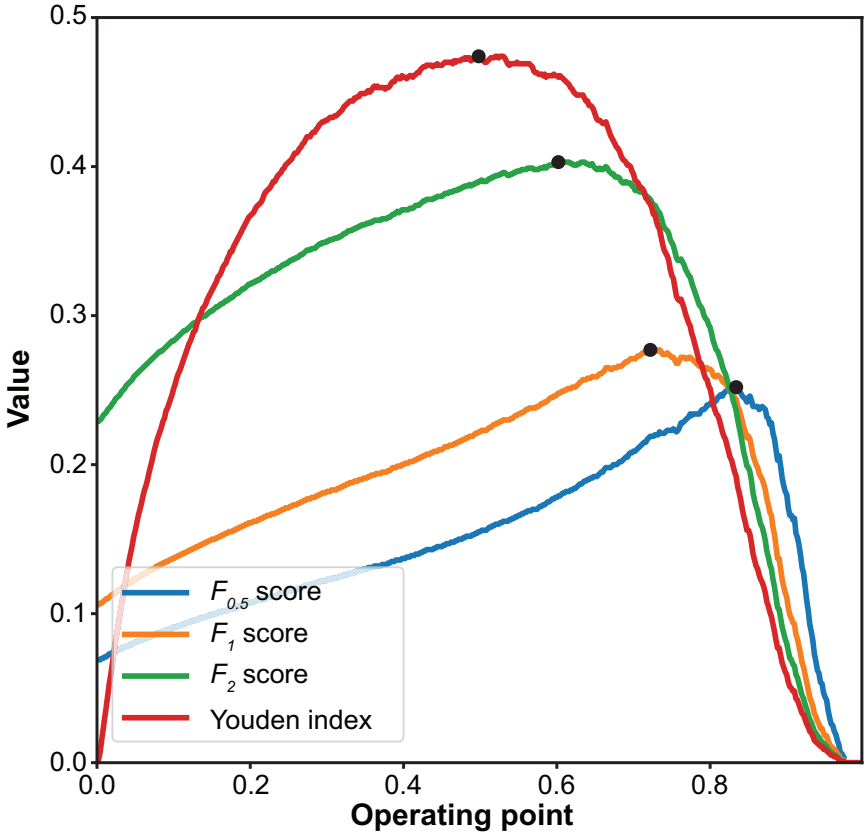
Supplementary Figure I: Illustration of the rearrangement of the signal traces from a standard 12-lead ECG which has 12 signal traces of 2.5 seconds and 3 rhythm strips of 10 seconds (for leads V1, II and V5). Leads aVL, aVF, and III were not used since they are combinations of other leads. Lead I was reconstructed from Goldberger's equation: -aVR = (I + II) / 2.



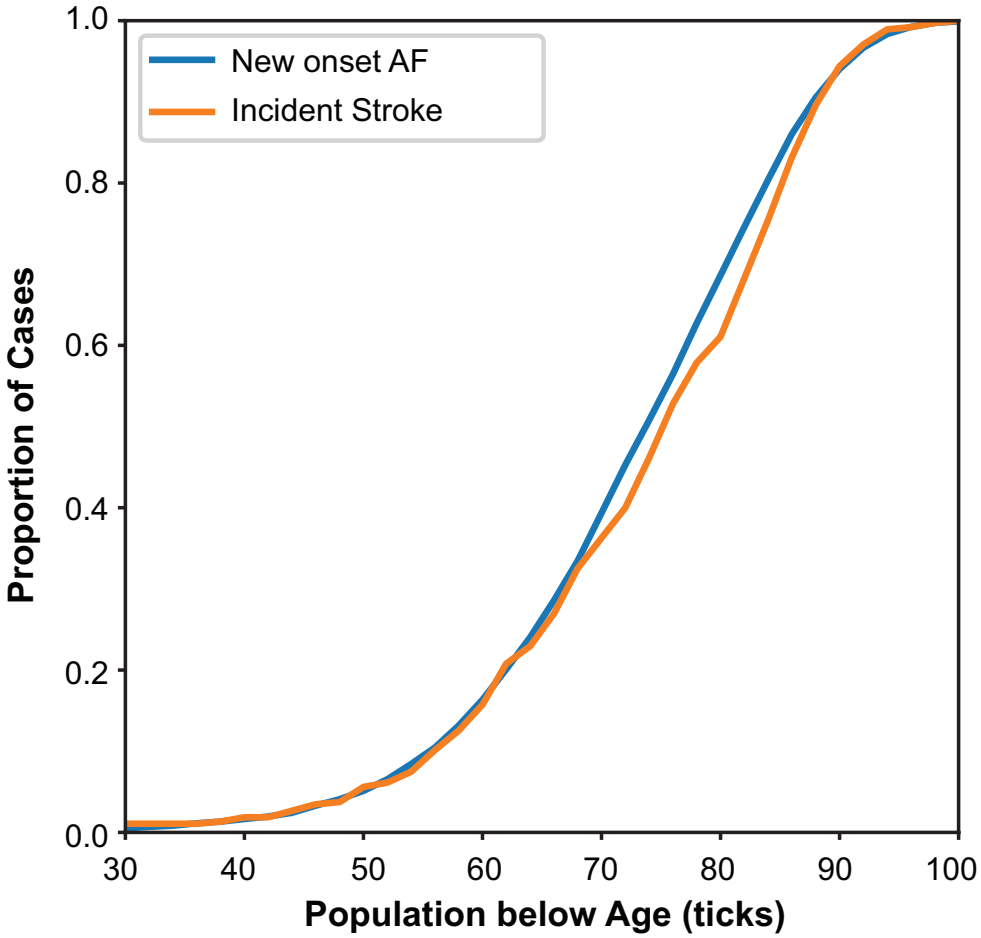
Supplementary Figure II: Deep neural network model architecture.



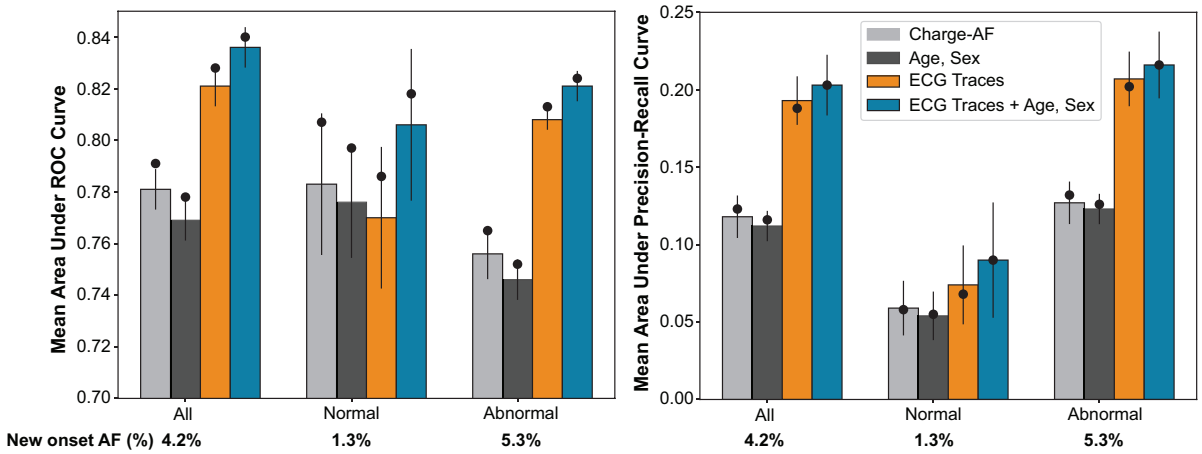
Supplementary Figure III: The timeline illustrating the composition of the test sets of proof-of-concept (POC) model (A) and deployment model (B), where a single ECG is selected for patients with multiple ECGs in the timeframe.



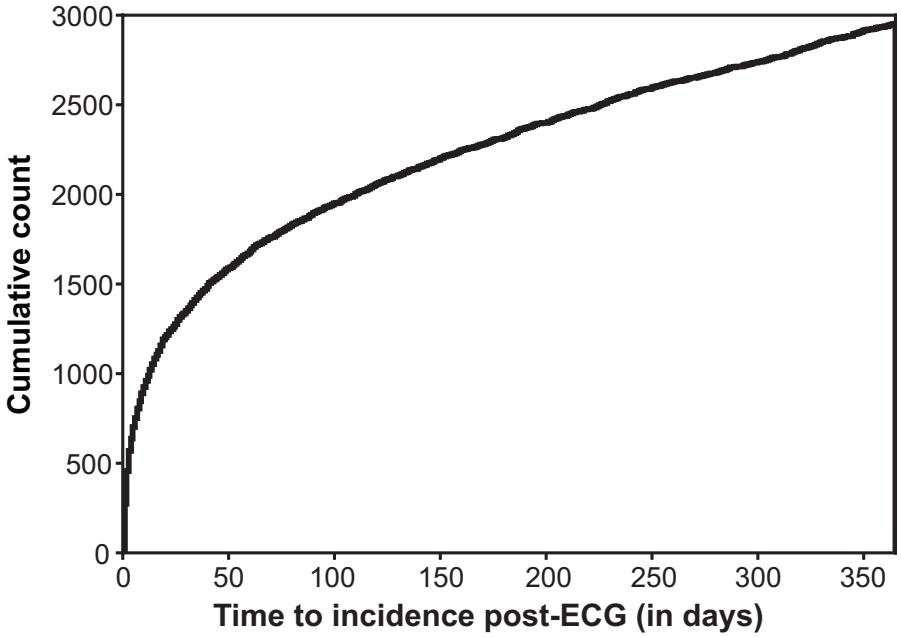
Supplementary Figure IV: Selection of the operating point for the model using the internal validation set in the simulated deployment model for the F_β scores ($\beta = 0.5$, 1 and 2) and Youden's index.



Supplementary Figure V: Proportion of all new onset AF (within one year post-ECG) and strokes (within 3 years post-ECG) in the population as a function of patients below the given age threshold.



Supplementary Figure VI: Illustration of model performance (proof-of-concept model) as area under the receiver operating characteristic (AUROC, left) and precision-recall curves (AUPRC, right) for the population with sufficient data for computation of the CHARGE-AF score. The bars represent the mean performance across the 5-fold cross-validation with error bars showing 95% confidence intervals. The black circle represents the M0 model performance on the holdout set. The three bars represent model performance for (i) Extreme gradient boosting (XGB) model with age and sex as inputs (gray); (ii) DNN model with digital ECG traces as input (DNN-ECG; orange) and (iii) DNN model with digital ECG traces, age and sex as inputs (DNN-ECG-AS; blue).



Supplementary Figure VII: The cumulative distribution of time to AF incidence after ECG in the holdout set of the proof-of-concept model.

SUPPLEMENTAL TABLES

Supplementary Table I: The performance metrics of blinded chart review (N = 200) of definition of atrial fibrillation or flutter (AF) phenotype.

Blinded chart review validation (AF phenotype)	
Positive Predictive Value	94%
Negative Predictive Value	98%
Sensitivity	98%
Specificity	94%
True Positive	94
True Negative	98
False Positive	6
False Negative	2

Supplementary Table II: Summary of the performance of the deep learning model trained with ECGs, age and sex to predict one-year new onset atrial fibrillation (AF) in the deployment scenario for four different operating points (defined in the independent internal validation set) and the potential to identify patients at risk for AF-associated stroke within 1, 2 and 3 years after ECG.

	Model predicted risk for new onset AF within 1 year of ECG					Number of patients predicted high risk for AF who developed an AF-related stroke within x years of ECG (Number needed to screen)		
Operating Point	# of ECGs flagged high risk	% of all ECGs flagged high risk	NNS to find 1 new onset AF	Sensitivity (Recall) (%)	Specificity (%)	*x* = 1	*x* = 2	*x* = 3
$F_{0.5}$ score	7958	4.4	5	26.9	96.4	17 (468)	41 (194)	65 (122)
F_1 score	21831	12.1	7	52.0	89.3	51 (428)	115 (190)	167 (131)
F_2 score	37428	20.7	9	68.7	81.0	69 (542)	158 (237)	231 (162)
Youden index	50995	28.3	11	77.8	73.5	75 (680)	182 (280)	269 (190)