Published in final edited form as:
Circ Arrhythm Electrophysiol. 2021 January ; 14(1): e008997. doi:10.1161/CIRCEP.120.008997.

Performance of Atrial Fibrillation Risk Prediction Models in Over Four Million Individuals

Shaan Khurshid, MD[1,2,*], Uri Kartoun, PhD[3,*], Jeffrey M. Ashburner, PhD[1,4], Ludovic Trinquart, PhD[5,6], Anthony Philippakis, MD, PhD[1], Amit V. Khera, MD MSc[1], Patrick T. Ellinor, MD, PhD[1,7], Kenney Ng, PhD[3], Steven A. Lubitz, MD, MPH[1,7]

[1]Cardiovascular Disease Initiative, Broad Institute of the Massachusetts Institute of Technology & Harvard University, Cambridge;

[2]Division of Cardiology, Massachusetts General Hospital, Boston;

[3]Center for Computational Health, IBM Research, Cambridge;

[4]Division of General Internal Medicine, Massachusetts General Hospital, Boston;

[5]Department of Biostatistics, Boston University School of Public Health, Boston;

[6]Boston University and National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA

[7]Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston;

Abstract

Background ––Atrial fibrillation (AF) is associated with increased risks of stroke and heart failure. Electronic health record (EHR) based AF risk prediction may facilitate efficient deployment of interventions to diagnose or prevent AF altogether.

Methods ––We externally validated an EHR atrial fibrillation (EHR-AF) score in IBM Explorys Life Sciences, a multi-institutional dataset containing statistically de-identified EHR data for over 21 million individuals ("Explorys Dataset"). We included individuals with complete AF risk data, ≥2 office visits within two years, and no prevalent AF. We compared EHR-AF to existing scores including CHARGE-AF, C_2HEST, and CHA_2DS_2-VASc. We assessed association between AF risk scores and 5-year incident AF, stroke, and heart failure using Cox proportional hazards modeling, 5-year AF discrimination using c-indices, and calibration of predicted AF risk to observed AF incidence.

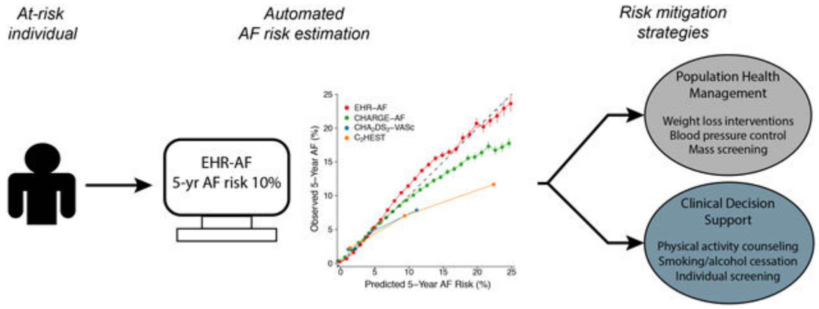
Results ––Of 21,825,853 individuals in the Explorys Dataset, 4,508,180 comprised the analysis (age 62.5, 56.3% female). AF risk scores were strongly associated with 5-year incident AF (hazard ratio [HR] per standard deviation [SD] increase 1.85 using CHA_2DS_2-VASc to 2.88 using EHR-AF), stroke (1.61 using C_2HEST to 1.92 using CHARGE-AF), and heart failure (1.91 using CHA_2DS_2-VASc to 2.58 using EHR-AF). EHR-AF (c-index 0.808 [95%CI 0.807–0.809]) demonstrated favorable AF discrimination compared to CHARGE-AF (0.806 [0.805–0.807]),

Correspondence: Steven A. Lubitz, MD, MPH, Cardiac Arrhythmia Service & Cardiovascular Research Center, Massachusetts General Hospital, 55 Fruit Street, GRB 109, Boston, MA 02114, Tel:617-643-7339, slubitz@mgh.harvard.edu.
[*]contributed equally

C_2HEST (0.683 [0.682–0.684]), and CHA_2DS_2-VASc (0.720 [0.719–0.722]). Of the scores, EHR-AF demonstrated the best calibration to incident AF (calibration slope 1.002 [0.997–1.007]). In subgroup analyses, AF discrimination using EHR-AF was lower in individuals with stroke (c-index 0.696 [0.692–0.700]) and heart failure (0.621 [0.617–0.625]).

Conclusions ––EHR-AF demonstrates predictive accuracy for incident AF using readily ascertained EHR data. AF risk is associated with incident stroke and heart failure. Use of such risk scores may facilitate decision-support and population health management efforts focused on minimizing AF-related morbidity.

Graphical Abstract


At-risk individual
Automated AF risk estimation
Risk mitigation strategies
EHR-AF
5-yr AF risk 10%
EHR–AF
CHARGE–AF
CHA2DS2–VASc
C2HEST
Observed 5–Year AF (%)
Predicted 5–Year AF Risk (%)
0
5
10
15
20
25
Population Health Management
Weight loss interventions
Blood pressure control
Mass screening
Clinical Decision Support
Physical activity counseling
Smoking/alcohol cessation
Individual screening

Keywords

atrial fibrillation; stroke; risk prediction; electronic health record

Journal Subject Terms:

Atrial Fibrillation; Epidemiology; Risk Factors; Quality and Outcomes

Introduction

Atrial fibrillation (AF) is a common arrhythmia associated with increased risks of ischemic stroke and heart failure.[1,2] Strokes related to AF are largely preventable with use of oral anticoagulation (OAC).[3,4] However, AF is frequently asymptomatic, with stroke commonly occurring as the first clinically-recognized manifestation.[5] At the same time, indiscriminate use of OAC in individuals without AF does not improve outcomes and leads to excess bleeding.[6] As a result, there has been substantial interest in identifying individuals at elevated AF risk, who may benefit from more aggressive diagnostic evaluation (e.g., AF screening) and targeted risk factor management (e.g., reduction in alcohol consumption, weight loss counseling) to mitigate risk of developing AF and related complications.[7,8]

AF risk can be estimated using clinical factors.[9,10] The Cohorts for Heart and Aging Research in Genomic Epidemiology AF (CHARGE-AF) score[10] has been validated in multiple community cohorts.[11,12] Existing AF risk schemes including CHARGE-AF, however, have peformed less favorably when deployed within electronic health records (EHRs).[13] In response, we recently developed an EHR-based AF score (EHR-AF),[9] which predicts incident AF using features readily available in most EHRs and demonstrates favorable performance in an EHR-based setting. By estimating AF risk accurately solely using EHR data, implementation of the EHR-AF score may facilitate improved population health management through identification of individuals most likely to benefit from targeted screening to detect occult AF or referral to intensive risk factor modification programs to prevent the development of AF altogether.

Prior to widespread implementation of the EHR-AF score, however, external validation in a large, independent EHR dataset is needed. We leveraged the EHR data of over four million individuals in the Explorys Dataset to assess the performance of EHR-AF, along with other established AF risk prediction schemes, for predicting AF, stroke, and heart failure.

Methods

Data Availability

The institutional review boards of Partners HealthCare and IBM approved this study and all its methods, including the EHR cohort assembly using the Explorys Dataset, data extraction, and analyses. Partners HealthCare data contain potentially identifying information and may not be shared publicly. Explorys data can be made available via a commercial license (for details see: https://www.ibm.com/downloads/cas/4P0QB9JN).

Study Population

The Explorys Dataset is comprised of the healthcare data of over 21 million individuals, pooled from multiple different healthcare systems with distinct EHRs which has been utilized previously for medical research.[14] Data were statistically de-identified,[15] standardized and normalized using common ontologies, and made searchable after upload to a Health Insurance Portability and Accountability Act-enabled platform. Data utilized in the current study included EHR entries of all patients who were seen in multiple healthcare systems between January 1st, 1999 and December 31st, 2019. We limited the scope of the current analysis to EHR data given our intent to assess the performance of EHR-AF as an AF prediction tool with potential EHR-based point-of-care applications.

To parallel the selection process of the original EHR-AF derivation study, we included individuals with at least two outpatient visits ≥2 years apart (Supplemental Figure I).[9] Follow-up was initiated at the first point in time after the second qualifying visit where complete data for AF risk estimation was available. All data available before the second qualifying visit were used to ascertain the presence of baseline conditions. The primary analysis cohort was referred to as the “Explorys Subset.”

Baseline Characteristics

Baseline age, sex, race, height, weight, and blood pressure values were obtained from the EHR, where the values most closely preceding the baseline visit were used to define baseline height, weight, and blood pressure. Tobacco use was classified as present or absent. Race was classified as white or non-white, in accordance with CHARGE-AF and EHR-AF definitions.[9,10] The presence of baseline conditions was ascertained using a combination of EHR data and diagnostic codes (International Classification of Diseases-9th revision [ICD-9] and −10th revision [ICD-10]). Clinical factor definitions can be found in Supplemental Table I.

Outcomes

The primary outcome of incident AF was defined using a modified version of a previously validated EHR-based AF ascertainment algorithm (positive predictive value 88%), in which electrocardiographic criteria were not utilized given the absence of electrocardiogram reports in the Explorys Dataset.[16] Secondary outcomes included incident stroke and incident heart failure, each defined using diagnostic codes unrestricted by encounter type (i.e., inpatient or outpatient). In additional analyses, we assessed incident AF and heart failure considering only events linked to inpatient encounters. Outcome definitions can be found in Supplemental Table I.

Determination of Predicted AF Risk

We estimated clinical AF risk for each participant by calculating the linear predictor of the EHR-AF score, along with three previously-validated risk scores: CHARGE-AF, C_2HEST, and CHA_2DS_2-VASc. Although originally derived and validated to predict risk of stroke in patients with AF, we included the CHA_2DS_2-VASc score as a comparator since it is commonly used in clinical settings and has demonstrated some ability to estimate AF risk.[12,17] The covariates and weights comprising each score have been described previously[3,9,10,18] (Supplemental Tables II–V). We converted the scores to a 5-year probability of AF utilizing the formula $1 - s_0^{\exp\left(\sum \beta X - \sum \beta Y\right)}$ where s_0 is the average AF-free survival probability at five years in the sample, $\sum \beta X$ is the individual's score, and $\sum \beta Y$ is the average score of the sample. We used the sample-level survival probability and average score in the primary analysis both to facilitate equitable comparison among the scores and to allow estimation of predicted 5-year AF risk using models without an explicit method of obtaining a 5-year probability in the original model publication (i.e., C_2HEST and CHA_2DS_2-VASc).[3,18]

Statistical Analysis

To facilitate comparison among AF risk prediction schemes, we utilized a complete-case approach in all analyses. The cumulative incidence of outcome events was calculated using the Kaplan-Meier method and incidence rates were calculated by dividing the number of events by total person-time. We assessed the association between each AF risk score and 5-year incident AF using Cox proportional hazards regression with 5-year AF risk as the outcome of interest and the AF risk score as the sole predictor in each model. We assessed the validity of the proportional hazards assumption by inspecting plots of scaled Schoenfeld residuals against time and by testing the correlation between Schoenfeld residuals and time (Supplemental Appendix). Scores were centered upon the mean and scaled by standard deviation to facilitate comparison. All models were censored at the earliest of death, end of follow-up (defined as last office visit or hospital encounter), or five years. We further assessed the strength of association with incident AF by comparing hazard ratios (HRs) and model fit using the Wald χ^2, the Nagelkerke R^2 and the Akaike Information Criterion (AIC). We compared discrimination of each score for incident AF using c-indices derived using the inverse probability of censoring method.[19,20] We compared model calibration by plotting predicted AF risk against observed AF incidence, both with and without bias-correction using 200-iteration bootstrapping. We also compared calibration slopes, each defined as the beta coefficient of a univariable Cox proportional hazards model with incident AF at five years as the outcome and the linear predictor of the repsective risk score as the sole covariate, where an optimally calibrated score has a value of one.[21] For models with a published 5-year AF risk equation (i.e., EHR-AF and CHARGE-AF), we additionally assessed calibration by plotting observed AF incidence across increasing deciles of predicted AF risk obtained using the original published equations.[9,22] Since calibration slope is not a sufficient indicator of calibration in models not recalibrated to the baseline hazard of the sample,[23] we tested calibration of the original models using the Greenwood-Nam-D'Agostino test.[24] We also assessed the hazard ratio for AF incidence and the 5-year cumulative incidence of AF as a function of categories of increasing predicted AF risk (i.e., <2.5%, 2.5–5%, and ≥5%).[9,10] To assess the clinical impact of utilizing EHR-AF as opposed to other scores for identifying individuals at elevated AF risk, we calculated net reclassification indices at the 5% AF risk threshold.[9,10,25]

We repeated the association and discrimination analyses outlined above with incident stroke at five years and incident heart failure at five years as the outcomes of interest. We performed secondary analyses in which we assessed incident AF and heart failure considering only diagnoses linked to a hospital encounter.

Given our interest in assessing the broad applicability of the EHR-AF score, we performed a secondary analysis in which we assessed the performance of EHR-AF within clinically relevant subgroups defined by sex, race, the presence of stroke, and the presence of heart failure. We then validated the findings of this analysis in the original EHR-AF validation dataset, which has been described previously.[15]

We considered two-sided p-values <0.05 to indicate statistical significance. All analyses were performed using R version 3.5 including the 'survival', 'rms', 'data.table', and 'prodlim' packages.[26]

Results

Of 21,825,853 million individuals in the Explorys Dataset, 4,791,963 met criteria for inclusion in the Explorys Subset. Of these 4,791,963, 283,783 had AF at baseline, leaving 4,508,180 individuals in the primary analysis (Supplemental Figure I). Of these 4,508,180 individuals, the mean age was 62.5±10.9 and 56.3% were women (other characteristics are listed in Table 1; baseline characteristics of the original EHR-AF derivation and validation sets[9] are shown for comparison). The distribution of the EHR-AF score both in Explorys and in the original EHR-AF validation set are depicted in Supplemental Figure II.

Over five years, 153,151 individuals developed incident AF over median follow-up 3.1 years (Q1: 1.1, Q3: 5.0). The 5-year cumulative incidence of AF was 5.5% (95% CI 5.5–5.5) and the yearly incidence rate was 11.7 per 1,000 person-years (95% CI 11.7–11.8). Each risk score was strongly associated with 5-year incident AF (HR 1.85 [95% CI 1.84–1.86] per 1 standard deviation [SD] increase in CHA$_2$DS$_2$-VASc; 1.88 [95% CI 1.88–1.89] per 1-SD increase in C$_2$HEST; 2.87 [95% CI 2.85–2.89] per 1-SD increase in CHARGE-AF; 2.87 [95% CI 2.86–2.89] per 1-SD increase in EHR-AF). AF discrimination was highest using the EHR-AF score (c-index 0.808, 95% CI 0.807–0.809), although it was also good using CHARGE-AF (0.806, 95% CI 0.805–0.807), but less favorable using C$_2$HEST (0.683, 95% CI 0.682–0.684) or CHA$_2$DS$_2$-VASc (0.720, 95% CI 0.719–0.722). Detailed comparisons of model fit using each score are listed in Table 2. Distributions of predicted AF risk by score are displayed in Supplemental Figure III. Of the scores, EHR-AF was best calibrated to observed AF risk (calibration slope 1.002, 95% CI 0.997–1.007). Correlation between predicted AF risk and observed AF incidence using each score is depicted in Figure 1. A summary of calibration results for each score is depicted in Figure 2. Associations with 5-year incident AF as well as the cumulative incidence of 5-year AF in categories of predicted AF risk are shown in Supplemental Table VI and Supplemental Figure IV. Net reclassification analyses at the 5% risk threshold demonstrated modest favorable reclassification using EHR-AF as opposed to CHARGE-AF (case reclassification −0.2% [95% CI −0.3% to −0.008%], non-case reclassification 1% [95% CI 0.9%−1%], net reclassification 0.08 [95% CI 0.06–0.01]), and strong favorable reclassification using EHR-AF as opposed to C$_2$HEST (case reclassification 12% [95% CI 12%−13%], non-case reclassification −2.8% [95% CI −2.8% to −2.8%], net reclassification 0.095 [95% CI 0.094–0.098]) and CHA$_2$DS$_2$-VASc (case reclassification 15% [95% CI 15%−15%], non-case reclassification −3.5% [95% CI −3.5% to −3.4%], net reclassification 0.12 [95% CI 0.12–0.13] (Supplemental Table VII).

Of 4,508,180 individuals in the primary analysis, 225,416 (5.0%) had prevalent stroke, resulting in 4,282,764 for incident stroke analyses. Over five years, 190,049 developed incident stroke over a median follow-up 3.1 years (Q1: 1.1, Q3: 5.0). The 5-year cumulative incidence of stroke was 7.2% (95% CI 7.2–7.3) and the yearly incidence rate was 15.3 per 1,000 person-years (95% CI 15.2–15.4). Each risk score was associated with 5-year incident stroke (HR 1.76 [95% CI 1.76–1.77] per 1-SD increase in CHA$_2$DS$_2$-VASc; 1.61 [95% CI 1.60–1.61] per 1-SD increase in C$_2$HEST; 1.92 [95% CI 1.91–1.93] per 1-SD increase in CHARGE-AF; 1.84 [95% CI 1.83–1.85] per 1-SD increase in EHR-AF). Stroke discrimination was greater with EHR-AF (0.700, 95% CI 0.699–0.702) and CHARGE-AF (0.710, 95% CI 0.709–0.711), than C$_2$HEST (0.670, 95% CI 0.669–0.672) or CHA$_2$DS$_2$-VASc (0.686, 95% CI 0.684–0.687). Detailed comparisons of model fit for incident stroke using each score are listed in Table 2.

Of 4,508,180 individuals in the primary analysis, 165,872 (3.7%) had prevalent HF, resulting in 4,342,308 for incident HF analyses. Over five years, 230,187 developed incident HF over median follow-up of 3.1 years (Q1: 1.1, Q3: 5.0). The 5-year cumulative incidence of heart failure was 8.4% (95% CI 8.4–8.5) and the yearly incidence rate was 18.0 per 1,000 person-years (95% CI 18.0–18.1). Each risk score was associated with 5-year incident HF (HR 1.91 [95% CI 1.91–1.92] per 1-SD increase in CHA$_2$DS$_2$-VASc; 1.94 [95% CI 1.93–1.944] per 1-SD increase in C$_2$HEST; 2.51 [95% CI 2.50–2.52] per 1-SD increase in CHARGE-AF; 2.58 [95% CI 2.57–2.59] per 1-SD increase in EHR-AF). HF discrimination was highest using EHR-AF (0.775, 95% CI 0.774–0.776) as compared to CHARGE-AF (0.770, 95% CI 0.769–0.771), CHA$_2$DS$_2$-VASc (0.730, 95% CI 0.729–0.732), and C$_2$HEST (0.733, 95% CI

0.732–0.734). Detailed comparisons of model fit for incident heart failure using each score are listed in Table 2.

Results were similar in analyses including only incident AF and heart failure events linked to an inpatient encounter (Supplemental Table VIII). When evaluated using the baseline risk estimates provided from their original derivation sets, both EHR-AF and CHARGE-AF showed evidence of miscalibration, with EHR-AF tending to overestimate AF risk at the high extreme of the AF risk distribution, and CHARGE-AF tending to underestimate AF risk throughout most of the AF risk distribution (Greenwood-Nam-D'Agostino $p<0.01$ for both, Supplemental Figure V).

We also assessed for potentially important variation in EHR-AF risk score performance in subgroups of interest (Figure 3). Both association with incident AF and discrimination were moderately higher in females (HR 3.18 [95% CI 3.15–3.20] per 1-SD increase; c-index 0.24 [95% CI 0.823–0.826]) than males (2.61 [95% CI 2.59–2.62] per 1-SD increase; 0.781 [95% CI 0.780–0.783]), but similar in whites (2.91 [95% CI 2.89–2.93] per 1-SD increase; 0.810 [95% CI 0.809–0.811]) compared to non-whites (2.75 [95% CI 2.70–2.79] per 1-SD increase; 0.796 [95% CI 0.791–0.800]). Association with incident AF and discrimination of AF were lower in individuals with a history of stroke (2.14 [95% CI 2.10–2.18] per 1-SD increase; 0.744 [95% CI 0.740–0.749]) and in individuals with a history of heart failure (1.57 [95% CI 1.55–1.60] per 1-SD increase; 0.683 [95% CI 0.678–0.687]) as compared to the overall sample (2.87 [95% CI 2.86–2.89] per 1-SD increase; 0.808 [95% CI 0.807–0.809]). Replication of these subgroup analyses in the independent EHR-AF validation sample showed very consistent results (Supplemental Table IX and Supplemental Figure VI).

Discussion

In over four million individuals in the Explorys Subset, the EHR-AF score demonstrated good predictive accuracy for 5-year incident AF. As previously demonstrated,[9] discrimination of incident AF was greater using EHR-AF relative to other AF risk schemes including CHARGE-AF, C_2HEST, and CHA_2DS_2-VASc. Moreover, calibration of estimated AF risk to observed AF incidence was also favorable using EHR-AF, suggesting the predicted probabilities of AF are more accurate using EHR-AF. Consistent with identification of individuals at high risk for both AF and AF-related complications, higher EHR-AF scores also predicted incident stroke and heart failure. In analyses focused on clinically relevant subgroups, we found that estimated AF risk was less accurate among individuals with a history of stroke and heart failure, indicating that improved methods of prediction may be necessary in these individuals.

Our findings provide further demonstration that AF prediction can be achieved with reasonable accuracy using clinical factors. The CHARGE-AF score has been shown to have good predictive performance in multiple populations.[9,11,12] The recently-developed C_2HEST score showed moderate predictive accuracy in an independent dataset comprised primarily of Asian individuals.[18] Our study adds to these findings by demonstrating that EHR-AF, a score derived and internally validated solely using EHR-based features, identifies individuals at higher risk for AF, stroke, and heart failure in a very large, independent EHR-based dataset in spite of differences in available data types (e.g., no electrocardiograms in the Explorys Dataset).

By demonstrating favorable discrimination and calibration for incident AF using EHR-AF within the independent Explorys Subset, our results support the broad use of EHR-AF for AF risk prediction using EHR data.[9] Although the improvement in discrimination we observed over CHARGE-AF was small, even numerically modest improvements in discrimination may have substantial impact within a population health management context. Consistent with past findings,[13] CHARGE-AF tended to overestimate observed AF risk in an EHR setting, particularly among individuals at higher predicted AF risk. In contrast, EHR-AF remained fairly well-calibrated across the spectrum of predicted AF risk represented by a sizeable EHR-based sample. Furthermore, reclassification analyses demonstrated modest appropriate reclassification using EHR-AF as opposed to CHARGE-AF and strong reclassification versus C_2HEST and CHA_2DS_2-VASc. In the current study, improved model performance is likely related to derivation of EHR-AF within a similarly-ascertained population (i.e., a large EHR sample). Our findings therefore suggest that risk model performance may be optimized when models are derived within populations representative of the specific clinical context in which risk stratification is intended.[9] Future prospective studies may clarify whether the gains in predictive performance we observed translate to improved clinical outcomes in patients at risk for AF and AF-related events.

More broadly, our results support the concept that disease risk stratification schemes based on EHR features have potential for deployment across EHR platforms. Our observation of good calibration to observed risk using EHR-AF in an external dataset suggests that application of EHR-AF should generally enable accurate individual-level absolute AF risk estimation. At the point of care, deployment of EHR-AF as a decision support aide may assist clinicians in identifying patients for targeted diagnostics such as ECG screening to detect occult AF, or more aggressive testing and treatment of AF risk factors (e.g., hypertension, sleep apnea).[27,28] Since healthcare systems or payers can readily leverage population-level EHR data to identify subpopulations of individuals at elevated AF risk, tools like EHR-AF may facilitate efficient population-based AF screening, or comprehensive programs designed to improve risk factor profiles (e.g., targeted weight loss, alcohol cessation).[7,8] Dedicated implementation studies are needed to determine whether deployment of EHR-based risk estimators, either in the context of decision support or population health management, leads to improved outcomes.

Although the EHR-AF score performed well in the Explorys Subset, dedicated analyses suggested the presence of performance heterogeneity within subpopulations in whom AF risk prediction may be particularly relevant. Specifically, EHR-AF was associated with roughly 6% greater AF discrimination among women versus men. Previous studies have suggested that AF scores may perform better in women, but the reasons are unclear and more detailed investigation is needed.[29] In contrast, EHR-AF did not discriminate AF as effectively in individuals with stroke and heart failure. Since both are known AF risk factors, it is possible that a greater baseline risk may decrease the relative importance of comorbidities captured by AF risk scores. Conversely, it is possible that features contributing to AF risk differ in individuals with pre-existing heart or cerebrovascular disease. In the latter case, it may be reasonable to derive dedicated risk schemes within subpopulations of interest. Since incident AF is a key determinant of outcome in individuals with both stroke and heart failure,[30,31] future studies are needed to determine whether any improvement in prediction accuracy achieved through development of dedicated risk stratification tools within subgroups of interest justifies the resources required to develop and implement those tools.

Our findings must be interpreted in the context of design. First, although we submit that our findings demonstrate that EHR-AF is externally valid in a large, independent EHR dataset, whether EHR-AF retains predictive performance outside of EHR-based populations (e.g., community-based cohorts) requires further study. Second, the Explorys Dataset is largely of European ancestry, and therefore our findings cannot establish the accuracy of EHR-AF in more racially diverse populations. Furthermore, as in the original EHR-AF derivation study, white race was associated with decreased AF risk whereas prospective cohort studies have generally suggested higher AF risk among whites.[10,32] Further work is needed to assess whether the observed association is due to selection bias or residual confounding, or instead may be reflective of disparities associated with real-world healthcare delivery. Third, although we utilized a validated AF detection algorithm to establish incident AF, ascertainment of other clinical factors relied on EHR-based features and diagnosis codes, which we cannot manually validate in the Explorys Dataset and may contribute to misclassification. Fourth, although EHR-AF demonstrated favorable discrimination, calibration, and reclassification, the degree of discrimination improvement over CHARGE-AF was small. Nevertheless, the improvement in calibration was more substantial, and even modest improvements in discrimination may be impactful at population scale. Fifth, since this is an observational study we cannot establish causal relations or eliminate residual confounding.

In summary, within an EHR-based sample including over 4.5 million individuals, we demonstrate that the EHR-AF score predicts incident AF risk accurately, comparing favorably to other established risk scores. Importantly, EHR-AF was well-calibrated to observed AF risk, suggesting the ability to provide accurate AF risk estimation either individually at the point of care, or at the population level using readily ascertainable EHR data. Future work is warranted to assess whether routine deployment of EHR-based AF risk estimation to guide interventions to diagnose or prevent AF leads to improved outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Sources of Funding:

This work was supported by the Broad/IBM Cardiovascular Disease Research Project, NIH grants 1K08HG010155 (Khera), R01HL139731 (Lubitz), 2R01HL092577 (Ellinor), 1R01HL128914 (Ellinor), R01HL104156 and K24HL105780 (Ellinor); T32HL007208 (Khurshid); American Heart Association (Dallas, Texas) 18SFRN34250007 (Lubitz); a Doris Duke Charitable Foundation Clinical Scientist Development Award 2014105 (Lubitz); and by the Fondation Leducq 14CVD01 (Ellinor).

Disclosures: Dr. Lubitz receives sponsored research support from Bristol-Myers Squibb / Pfizer, Bayer AG, Biotronik, and Boehringer Ingelheim, and has consulted for Bristol-Myers Squibb and Bayer AG. Dr. Ellinor is the principal investigator on a grant at the Broad Institute from Bayer AG.

Nonstandard Abbreviations and Acronyms

AF	Atrial fibrillation
CCS	Clinical Classification Software
CHARGE-AF	Cohorts for Heart and Aging Research in Genomic Epidemiology atrial fibrillation
ECG	Electrocardiogram
EHR	Electronic health record
EHR-AF	Electronic health record atrial fibrillation
ICD	International Classification of Diseases
OAC	Oral anticoagulant

References:

1. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation: a major contributor to stroke in the elderly. The Framingham Study. Arch Intern Med. 1987;147:1561–1564. [PubMed: 3632164]
2. Corley SD, Epstein AE, DiMarco JP, Domanski MJ, Geller N, Greene HL, Josephson RA, Kellen JC, Klein RC, Krahn AD, et al. Relationships between sinus rhythm, treatment, and survival in the Atrial Fibrillation Follow-Up Investigation of Rhythm Management (AFFIRM) Study. Circulation. 2004;109:1509–1513. [PubMed: 15007003]
3. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. 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. [PubMed: 19762550]
4. Stroke Prevention in Atrial Fibrillation Study. Final results. Circulation. 1991;84:527–539. [PubMed: 1860198]
5. Lubitz SA, Yin X, McManus DD, Weng L-C, Aparicio HJ, Walkey AJ, Rafael Romero J, Kase CS, Ellinor PT, Wolf PA, et al. Stroke as the Initial Manifestation of Atrial Fibrillation: The Framingham Heart Study. Stroke. 2017;48:490–492. [PubMed: 28082669]
6. Hart RG, Sharma M, Mundl H, Kasner SE, Bangdiwala SI, Berkowitz SD, Swaminathan B, Lavados P, Wang Y, Wang Y, et al. Rivaroxaban for Stroke Prevention after Embolic Stroke of Undetermined Source. N Engl J Med. 2018;378:2191–2201. [PubMed: 29766772]
7. Voskoboinik A, Kalman JM, De Silva A, Nicholls T, Costello B, Nanayakkara S, Prabhu S, Stub D, Azzopardi S, Vizi D, et al. Alcohol Abstinence in Drinkers with Atrial Fibrillation. N Engl J Med. 2020;382:20–28. [PubMed: 31893513]
8. Middeldorp ME, Pathak RK, Meredith M, Mehta AB, Elliott AD, Mahajan R, Twomey D, Gallagher C, Hendriks JML, Linz D, et al. PREVEntion and regReSsive Effect of weight-loss and risk factor modification on Atrial Fibrillation: the REVERSE-AF study. Europace. 2018;20:1929–1935. [PubMed: 29912366]
9. Hulme OL, Khurshid S, Weng L-C, Anderson CD, Wang EY, Ashburner JM, Ko D, McManus DD, Benjamin EJ, Ellinor PT, et al. Development and Validation of a Prediction Model for Atrial Fibrillation Using Electronic Health Records. JACC Clin Electrophysiol. 2019;5:1331–1341. [PubMed: 31753441]

10. Alonso A, Roetker NS, Soliman EZ, Chen LY, Greenland P, Heckbert SR. Prediction of Atrial Fibrillation in a Racially Diverse Cohort: The Multi-Ethnic Study of Atherosclerosis (MESA). J Am Heart Assoc. 2016;5:e003077. [PubMed: 26908413]
11. Shulman E, Kargoli F, Aagaard P, Hoch E, Di Biase L, Fisher J, Gross J, Kim S, Krumerman A, Ferrick KJ. Validation of the Framingham Heart Study and CHARGE-AF Risk Scores for Atrial Fibrillation in Hispanics, African-Americans, and Non-Hispanic Whites. Am J Cardiol. 2016;117:76–83. [PubMed: 26589820]
12. Christophersen IE, Yin X, Larson MG, Lubitz SA, Magnani JW, McManus DD, Ellinor PT, Benjamin EJ. A comparison of the CHARGE-AF and the CHA2DS2-VASc risk scores for prediction of atrial fibrillation in the Framingham Heart Study. Am Heart J. 2016;178:45–54. [PubMed: 27502851]
13. Kolek MJ, Graves AJ, Xu M, Bian A, Teixeira PL, Shoemaker MB, Parvez B, Xu H, Heckbert SR, Ellinor PT, et al. Evaluation of a Prediction Model for the Development of Atrial Fibrillation in a Repository of Electronic Medical Records. JAMA Cardiology. 2016;1:1007. [PubMed: 27732699]
14. Kartoun U, Corey KE, Simon TG, Zheng H, Aggarwal R, Ng K, Shaw SY. The MELD-Plus: A generalizable prediction risk score in cirrhosis. PLoS ONE. 2017;12:e0186301. [PubMed: 29069090]
15. Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington (DC): National Academies Press (US); 2015 Apr 20. Appendix B, Concepts and Methods for De-identifying Clinical Trial Data Available from: https://www.ncbi.nlm.nih.gov/books/NBK285994/.
16. Khurshid S, Keaney J, Ellinor PT, Lubitz SA. A Simple and Portable Algorithm for Identifying Atrial Fibrillation in the Electronic Medical Record. Am J Cardiol. 2016;117:221–225. [PubMed: 26684516]
17. Saliba W, Gronich N, Barnett-Griness O, Rennert G. Usefulness of CHADS2 and CHA2DS2-VASc Scores in the Prediction of New-Onset Atrial Fibrillation: A Population-Based Study. Am J Med. 2016;129:843–849. [PubMed: 27012854]
18. Li Y-G, Pastori D, Farcomeni A, Yang P-S, Jang E, Joung B, Wang Y-T, Guo Y-T, Lip GYH. A Simple Clinical Risk Score (C2HEST) for Predicting Incident Atrial Fibrillation in Asian Subjects: Derivation in 471,446 Chinese Subjects, With Internal Validation and External Application in 451,199 Korean Subjects. Chest. 2018;
19. Hung H, Chiang C-T. Estimation methods for time-dependent AUC models with survival data. Can J Statistics. 2009;n/a–n/a.
20. Uno H, Cai T, Lu T, Wei LJ. Evaluating Prediction Rules for t-Year Survivors with Censored Regression Models. J Am Stat Assoc. 2007;102:527–537.
21. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958;45:562–565.
22. Alonso A, Krijthe BP, Aspelund T, Stepas KA, Pencina MJ, Moser CB, Sinner MF, Sotoodehnia N, Fontes JD, Janssens ACJW, 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. [PubMed: 23537808]
23. Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the “calibration slope” really measure? J Clin Epidemiol. 2020;118:93–99. [PubMed: 31605731]
24. Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness-of-fit in the survival setting. Stat Med. 2015;34:1659–1680. [PubMed: 25684707]
25. Pencina MJ, D’ Agostino RB, D’ Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statist Med. 2008;27:157–172.
26. R Core Team (2015). R: A language and environment for statistical computing R Foundation for Statistical Computing Vienna, Austria URL https://www.R-project.org/.
27. Linz D, McEvoy RD, Cowie MR, Somers VK, Nattel S, Lévy P, Kalman JM, Sanders P. Associations of Obstructive Sleep Apnea With Atrial Fibrillation and Continuous Positive Airway Pressure Treatment: A Review. JAMA Cardiol. 2018;3:532–540. [PubMed: 29541763]

28. Dzeshka MS, Shantsila A, Shantsila E, Lip GYH. Atrial Fibrillation and Hypertension. Hypertension. 2017;70:854–861. [PubMed: 28893897]

29. Aronson D, Shalev V, Katz R, Chodick G, Mutlak D. Risk Score for Prediction of 10-Year Atrial Fibrillation: A Community-Based Study. Thromb Haemost. 2018;118:1556–1563. [PubMed: 30103243]

30. Marrouche NF, Brachmann J, Andresen D, Siebels J, Boersma L, Jordaens L, Merkely B, Pokushalov E, Sanders P, Proff J, et al. Catheter Ablation for Atrial Fibrillation with Heart Failure. N Engl J Med. 2018;378:417–427. [PubMed: 29385358]

31. Penado S, Cano M, Acha O, Hernández JL, Riancho JA. Atrial fibrillation as a risk factor for stroke recurrence. Am J Med. 2003;114:206–210. [PubMed: 12637135]

32. Rodriguez CJ, Soliman EZ, Alonso A, Swett K, Okin PM, Goff DC, Heckbert SR. Atrial fibrillation incidence and risk factors in relation to race-ethnicity and the population attributable fraction of atrial fibrillation risk factors: the Multi-Ethnic Study of Atherosclerosis. Ann Epidemiol. 2015;25:71–76, 76.e1. [PubMed: 25523897]

What Is Known?

- The development of future atrial fibrillation (AF) can be predicted with reasonable accuracy using clinical factors.
- Electronic health records are ubiquitous and comprise routinely collected clinical risk factors.
- AF risk estimation using electronic health records may be feasible and accurate but requires validation.

What the Study Adds?

- When compared to existing AF risk scores, an EHR-AF score demonstrates favorable AF discrimination and is well-calibrated to AF risk in an independent EHR comprising over 4 million individuals.
- AF risk scores were associated with higher risks of incident AF, stroke, and heart failure.
- An EHR-AF score may provide accurate and automated AF risk estimation to enable both clinical decision support and population health management interventions designed to reduce AF risk and downstream consequences.

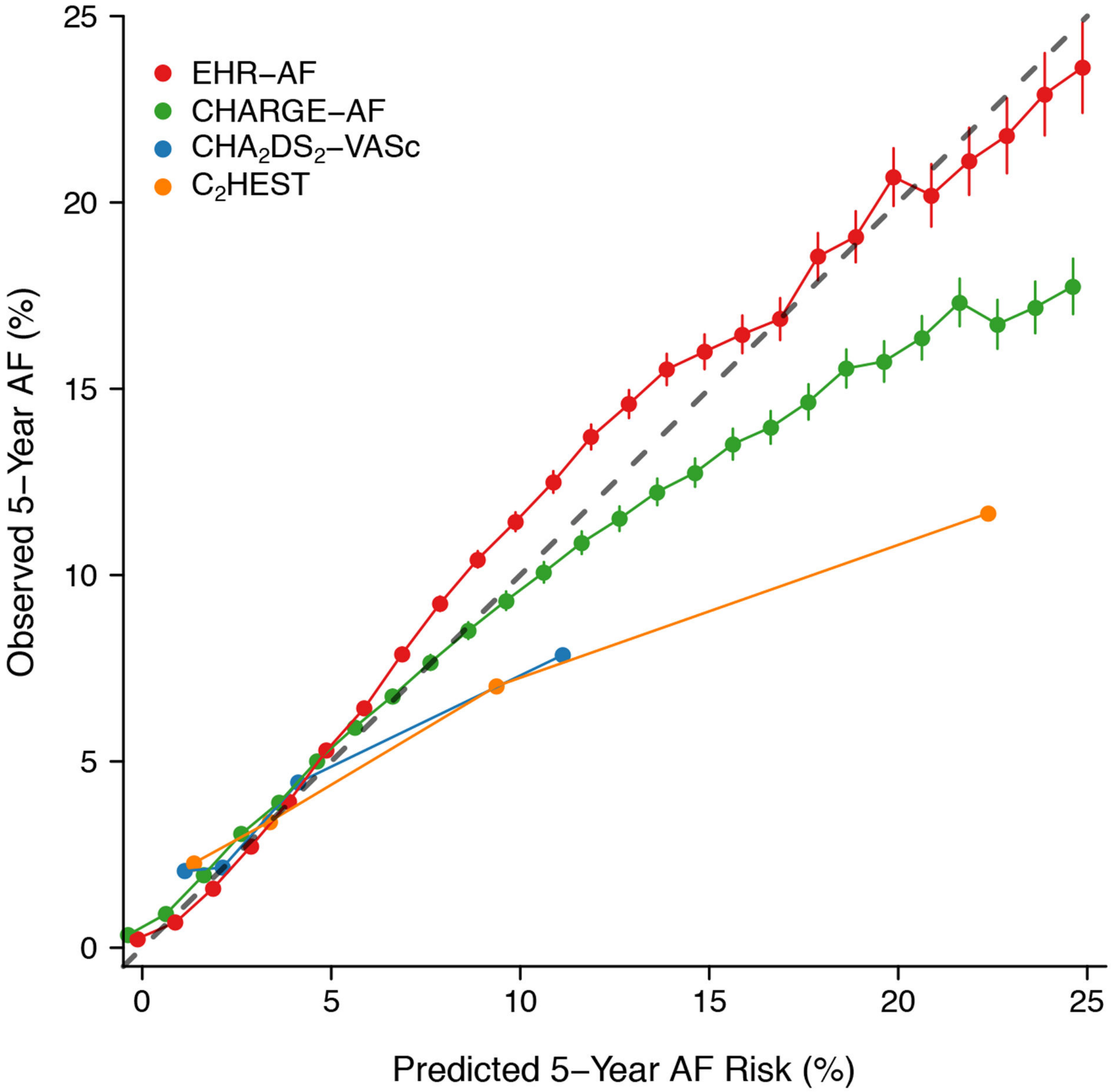


Figure 1.

Predicted versus observed AF rates by score. Depicted is the observed 5-year AF rate (y-axis) versus increasing predicted 5-year AF risk (x-axis) determined using each of four AF risk schemes: EHR-AF, CHARGE-AF, C_2HEST, and CHA_2DS_2-VASc. Each point represents estimated AF risk rounded to the nearest whole number. The gray line demonstrates perfect correspondence between predicted and observed risk. Points above the gray bar represent underestimation of true AF risk, while points below the gray bar represent overestimation of true AF risk. Fewer points are present for CHA_2DS_2-VASc and C_2HEST given the discrete nature of these scores.

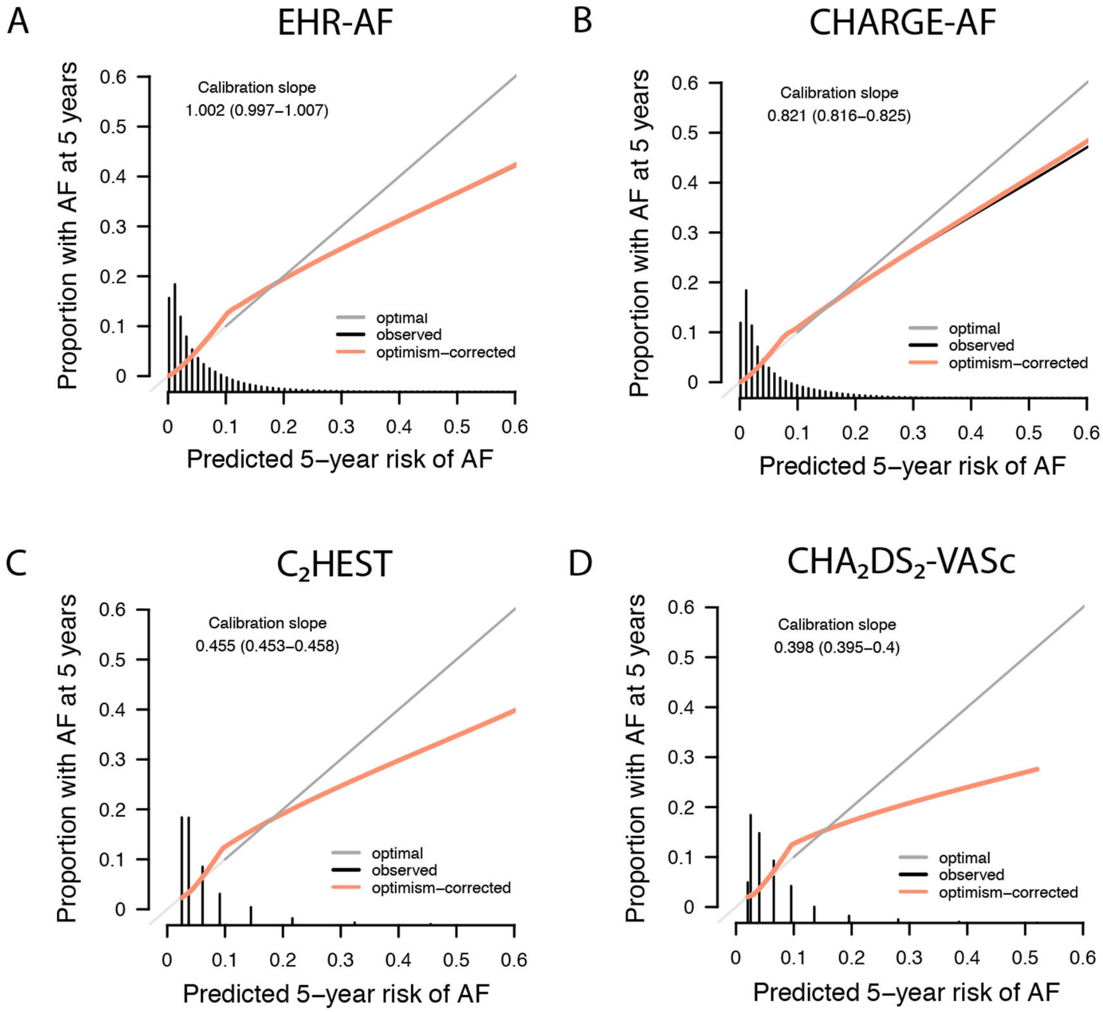


Figure 2.

Calibration of incident AF risk prediction models. Depicted is the calibration of the four AF risk prediction models evaluated: EHR-AF, CHARGE-AF, C_2HEST, and CHA_2DS_2-VASc. In each plot, the y-axis depicts the observed 5-year AF incidence and the x-axis depicts the predicted 5-year AF risk. Perfect calibration is represented by the dark gray diagonal line. The black line corresponds to a smoothed fit of the observed calibration, and the orange line to a smoothed fit of the calibration after correction for model optimism using 200-iteration bootstrapping. Smoothed fits of the observed calibration are present only for the continuous scores EHR-AF and CHARGE-AF given inadequate data points to support fits for the discrete scores CHA_2DS_2-VASc and C_2HEST, although optimism-corrected fits are present for each score. Calibration slopes (and 95% confidence intervals) are shown on the graphs, where optimal calibration slope is equal to one. The distribution of predicted 5-year AF risk is depicted by histograms along the x-axis of each plot.

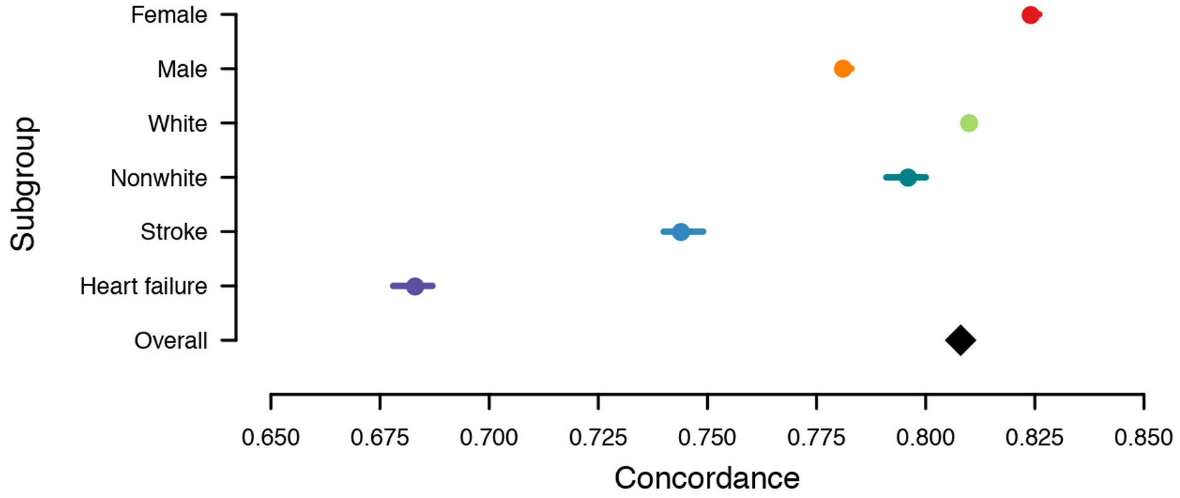


Figure 3.

EHR-AF score prediction performance for 5-year incident AF in subgroups of interest. Depicted is the c-index with 95% confidence interval (panel A), and hazard ratio per one standard deviation increase) with 95% confidence interval (panel B) for 5-year incident atrial fibrillation using the EHR-AF score within subgroups of clinical interest. The effect estimate in the overall sample is depicted with the black diamond. The total number of individuals in each subgroup and number of incident AF events are as follows: female (N=2,537,855, AF=72,195), male (N=1,970,325, AF=80,956), white (N=3,791,377, AF=137,852), non-white (N=716,803, AF=15,299), stroke (N=225,416, AF=14,997), heart failure (N=165,872, AF=20,058), overall (N=4,508,180, AF=153,151).

Table 1.

Baseline characteristics

	Original EHR-AF Sample*		Explorys Subset
	Derivation (n=206,042)	Validation (n= 206,043)	(n=4,508,180)
	% or Mean (SD)	% or Mean (SD)	% or Mean (SD)
Female	58.3	58.1	56.3
Age	60.7 (10.6)	60.6 (10.6)	62.5 (10.9)
White race	85.5	85.3	84.1
Smoking	9.7	9.8	18.1
Height, cm	167 (10)	167 (10)	169 (11)
Weight, kg	79 (19)	79 (19)	87 (22)
Systolic blood pressure, mmHg	129 (17)	129 (17)	131 (18)
Diastolic blood pressure, mmHg	76 (10)	76 (10)	77 (11)
Hypertension	28.5	28.8	52.7
Diabetes	9.4	9.4	21.7
Hyperlipidemia	30.2	30.2	58.0
Heart failure	3.1	3.2	3.7
Coronary heart disease	9.3	9.4	12.1
Valvular disease	1.4	1.5	3.0
Stroke/TIA	3.8	3.8	5.0
Myocardial infarction†	4.1	4.0	3.3
Peripheral artery disease†	3.9	3.8	5.3
Systemic atherosclerosis†	1.4	1.4	1.6
Cerebral atherosclerosis†	3.5	3.5	3.9
Chronic kidney disease	3.4	3.4	6.7
Thyrotoxicosis	1.5	1.5	1.2
Hypothyroidism	8.7	8.6	12.8

*Characteristics of original EHR-AF derivation and validation cohorts listed for comparison to Explorys Subset

†Component of the vascular disease category of the CHA_2DS_2-VASc score

Table 2.

AF risk score performance for 5-year incident AF, stroke, and heart failure.

	Score			
	CHA_2DS_2-VASc	**C_2HEST**	**CHARGE-AF**	**EHR-AF**
AF[*]				
Hazard ratio per SD increase (95% CI)	1.85 (1.84–1.86)	1.88 (1.88–1.89)	2.87 (2.85–2.89)	2.87 (2.86–2.89)
AIC[†]	4,473,906	4,458,257	4,394,587	4,390,568
Wald χ^2	88,782, p<0.01	113,049, p<0.01	142,925, p<0.01	152,686, p<0.01
R^2 (95% CI)	0.039 (0.038–0.039)	0.039 (0.039–0.040)	0.065 (0.064–0.065)	0.065 (0.064–0.065)
c-index (95% CI)	0.720 (0.719–0.722)	0.683 (0.682–0.684)	0.806 (0.805–0.807)	0.808 (0.807–0.809)
Calibration slope (95% CI)	0.398 (0.295–0.400)	0.455 (0.453–0.458)	0.821 (0.816–0.825)	1.002 (0.997–1.007)
Stroke[‡]				
Hazard ratio per SD increase (95% CI)	1.76 (1.76–1.77)	1.61 (1.60–1.61)	1.92 (1.91–1.93)	1.84 (1.83–1.85)
AIC[†]	5,550,757	5,562,546	5,541,413	5,551,439
Wald χ^2	73,374, p<0.01	68,772, p<0.01	77,083, p<0.01	68,101, p<0.01
R^2 (95% CI)	0.022 (0.022–0.022)	0.018 (0.018–0.019)	0.025 (0.025–0.025)	0.022 (0.021–0.022)
c-index (95% CI)	0.686 (0.684–0.687)	0.670 (0.669–0.672)	0.710 (0.709–0.711)	0.700 (0.699–0.702)
Heart failure[§]				
Hazard ratio per SD increase (95% CI)	1.91 (1.91–1.92)	1.94 (1.93–1.94)	2.51 (2.50–2.52)	2.58 (2.57–2.59)
AIC[†]	6,687,160	6,678,595	6,638,481	6,628,574
Wald χ^2	148,962, p<0.01	156,206, p<0.01	161,868, p<0.01	173,488, p<0.01
R^2 (95% CI)	0.037 (0.036–0.037)	0.039 (0.039–0.040)	0.050 (0.050–0.051)	0.053 (0.053–0.054)
c-index (95% CI)	0.730 (0.729–0.732)	0.733 (0.732–0.734)	0.770 (0.769–0.771)	0.775 (0.774–0.776)

[*] N events=153,151, N total=4,508,180, median follow-up, years (Q1,Q3): 3.1 (1.1,5.0)

[†] Akaike Information Criterion (AIC), a penalized likelihood metric. Lower values suggest better model fit while accounting for model complexity.

[‡] N events=190,049, N total=4,282,764, median follow-up, years (Q1,Q3): 3.1 (1.1,5.0)

[§] N events=230,187, N total=4,342,308, median follow-up, years (Q1,Q3): 3.1 (1.1,5.0)