








ORIGINAL RESEARCH

Natural Language Processing to Improve Prediction of Incident Atrial Fibrillation Using Electronic Health Records

Jeffrey M. Ashburner , PhD, MPH; Yuchiao Chang, PhD; Xin Wang, MPH; Shaan Khurshid , MD, MPH; Christopher D. Anderson , MD, MMSC; Kumar Dahal, MS; Dana Weisenfeld, MS; Tianrun Cai, MD; Katherine P. Liao , MD, MPH; Kavishwar B. Waghlikar , MD, PhD; Shawn N. Murphy, MD, PhD; Steven J. Atlas, MD, MPH; Steven A. Lubitz , MD, MPH; Daniel E. Singer , MD

BACKGROUND: Models predicting atrial fibrillation (AF) risk, such as Cohorts for Heart and Aging Research in Genomic Epidemiology AF (CHARGE-AF), have not performed as well in electronic health records. Natural language processing (NLP) may improve models by using narrative electronic health record text.

METHODS AND RESULTS: From a primary care network, we included patients aged ≥ 65 years with visits between 2003 and 2013 in development ($n=32\,960$) and internal validation cohorts ($n=13\,992$). An external validation cohort from a separate network from 2015 to 2020 included 39\,051 patients. Model features were defined using electronic health record codified data and narrative data with NLP. We developed 2 models to predict 5-year AF incidence using (1) codified+NLP data and (2) codified data only and evaluated model performance. The analysis included 2839 incident AF cases in the development cohort and 1057 and 2226 cases in internal and external validation cohorts, respectively. The C-statistic was greater ($P<0.001$) in codified+NLP model (0.744 [95% CI, 0.735–0.753]) compared with codified-only (0.730 [95% CI, 0.720–0.739]) in the development cohort. In internal validation, the C-statistic of codified+NLP was modestly higher (0.735 [95% CI, 0.720–0.749]) compared with codified-only (0.729 [95% CI, 0.715–0.744]; $P=0.06$) and CHARGE-AF (0.717 [95% CI, 0.703–0.731]; $P=0.002$). Codified+NLP and codified-only were well calibrated, whereas CHARGE-AF underestimated AF risk. In external validation, the C-statistic of codified+NLP (0.750 [95% CI, 0.740–0.760]) remained higher ($P<0.001$) than codified-only (0.738 [95% CI, 0.727–0.748]) and CHARGE-AF (0.735 [95% CI, 0.725–0.746]).

CONCLUSIONS: Estimation of 5-year risk of AF can be modestly improved using NLP to incorporate narrative electronic health record data.

Key Words: atrial fibrillation ■ natural language processing ■ predicted risk

Atrial fibrillation (AF) is a common arrhythmia in aging populations,^{1,2} is a potent risk factor for ischemic stroke,^{3–5} and is often first identified at the time of stroke.^{6,7} Oral anticoagulation therapy is highly efficacious in preventing a large proportion of AF-related strokes.^{8–12} Screening for undiagnosed AF is a growing priority as it may enable earlier diagnosis of AF and implementation of oral anticoagulation therapy to prevent strokes.

Novel point-of-care technologies used in health care settings or at home, including wearable technology and mobile single-lead ECGs, make mass screening for undiagnosed AF feasible. Prior randomized studies assessing the impact of AF screening interventions in subjects aged ≥ 65 years have been mixed^{13–16}; however, the effectiveness of these studies may have been limited by screening solely based on age, which may

Correspondence to: Jeffrey M. Ashburner, PhD, MPH, Division of General Internal Medicine, Massachusetts General Hospital, 100 Cambridge St, 16th Floor, Boston, MA 02114. Email: jashburner@mgh.harvard.edu

Supplemental Material is available at <https://www.ahajournals.org/doi/suppl/10.1161/JAHA.122.026014>

For Sources of Funding and Disclosures, see page 14.

© 2022 The Authors. Published on behalf of the American Heart Association, Inc., by Wiley. This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVE

What Is New?

- We derived and validated 2 new models to predict the incidence of atrial fibrillation: one model that used only codified data and a second model that added information from narrative data using natural language processing.
- In internal and external validation, we found that estimation of the 5-year risk of atrial fibrillation in a primary care population can be modestly improved by using natural language processing to incorporate narrative electronic health record data.
- Both newly developed models (codified data only and codified plus natural language processing) demonstrated improved predictive utility compared with an established atrial fibrillation prediction model (Cohorts for Heart and Aging Research in Genomic Epidemiology Atrial Fibrillation).

What Are the Clinical Implications?

- The amount of data available in unstructured form within electronic health records is immense, and optimizing the use of natural language processing to meaningfully process this data offers an opportunity to improve prognostic models used within clinical care.

Nonstandard Abbreviations and Acronyms

CHARGE-AF	Cohorts for Heart and Aging Research in Genomic Epidemiology Atrial Fibrillation
NRI	net reclassification improvement

include many subjects with low short-term risk of developing AF.^{16,17} Using individual patient risk of AF can effectively allocate screening resources to those most likely to benefit.¹⁸

A widely used model to predict AF, the Cohorts for Heart and Aging Research in Genomic Epidemiology AF (CHARGE-AF) score,¹⁹ was developed and validated in research cohorts with rigorously collected and uniformly formatted data and has demonstrated weaker performance in health care–related data sets.^{20,21} AF risk stratification may be most often used in clinical settings that include subjects with more comorbidities than those included in research cohorts. Developing and validating an AF risk prediction model within a clinical setting using features available in electronic health records (EHRs) may improve model performance in these settings. Much of the data in EHRs exist as free-form typed narrative text within provider notes or reports. Natural language processing (NLP) represents a

range of computational techniques for processing text that can be used to extract medical concepts for analyses. Incorporating NLP in determination of risk factors within EHRs has several advantages and has resulted in improved risk prediction.^{22,23} NLP provides data that may be missing from codified data (ie, entered in a structured format, such as *International Classification of Diseases [ICD]*, Tenth Revision diagnosis codes), for example, information extracted from cardiology reports about left ventricular hypertrophy.^{24–26}

In this study, we developed and evaluated 2 models to identify patients at increased risk for AF in primary care patients at Massachusetts General Hospital. One model used codified and NLP data (codified+NLP), whereas one used codified data only (codified-only). It is not known if incorporating NLP data to ascertain risk factors within EHRs will result in an improved AF risk prediction model. As such, we compared performance of newly developed models with and without NLP and compared each to an existing and widely used AF risk model (CHARGE-AF) in both internal and external validation populations.²⁷

METHODS

Mass General Brigham data contain protected health information and cannot be shared publicly. The data processing scripts used to perform analyses will be made available to interested researchers on reasonable request to the corresponding author.

Study Sample

Model Development and Internal Validation

The study cohort for model development and internal validation consisted of patients from the Primary Care Practice-Based Research Network at Massachusetts General Hospital, identified using a validated attribution algorithm.^{28,29} All 18 practices in the network use EHRs and share the same data systems. Individuals included were aged ≥ 65 years, with primary care visits between 2003 and 2013. Individuals were excluded if they had diagnosed prevalent AF with no study follow-up before the AF diagnosis. Individuals aged < 65 years were excluded because AF prevalence is strongly associated with age, and these individuals are less likely to benefit from oral anticoagulation or be targeted in an AF screening program. The eligible cohort was randomly split into 2 subsets: two thirds for model development and one third for internal validation.

External Validation

The external validation cohort consisted of patients aged ≥ 65 years from Brigham and Women's Hospital with primary care visits between 2015 and 2020 and

without prevalent AF. Brigham and Women's 15 practices use EHRs, and data are accessed from the Mass General Brigham Research Patient Data Registry, a data warehouse containing data from 7 affiliated hospitals, including Massachusetts General Hospital.³⁰ This medical records-based study was approved with a waiver of informed consent by the local Mass General Brigham Institutional Review Board.

Ascertainment of Potential Features/Variables for the Model

Data extracted from the EHR included the following: (1) patient demographics, (2) diagnostic codes (*International Classification of Diseases, Ninth Revision [ICD-9]*, and *International Classification of Diseases, Tenth Revision [ICD-10]*), (3) procedure codes (Current Procedural Terminology), (4) medications, (5) cardiology test reports, (6) progress notes, (7) visit notes, (8) history and physical notes, (9) laboratory notes, (10) discharge summaries, and (11) vital status. We assembled a list of potential predictors of AF incidence based on a review of prior studies. The full list of potential model features is available in [Table 1](#), as well as the source of information (eg, codified EHR data versus data extracted using NLP). Although height, weight, and blood pressure are included in an existing AF prediction model, these were not considered as features for new models because of missing data.

Clinical Characteristics

Patient characteristics and comorbidities were ascertained using EHR data from the 3 years before each patient's first visit during the study period. Age, sex, and race and ethnicity were ascertained at the time of cohort entry. Height and weight recorded closest to cohort entry were obtained. Medication use was assessed on the basis of any medications listed in the EHR in the 3 years before cohort entry. Laboratory values were assessed within the 1 year before cohort entry. Current smoking status was assessed using the most recent smoking status update in the EHR before cohort entry. PR interval was extracted from ECGs and classified as shortened (<120 milliseconds), normal (120–200 milliseconds), or prolonged (>200 milliseconds).³¹ We used patient home zip code to link to the National Neighborhood Data Archive and ascertain the percentage of the population within a zip code tabulation area with household income of <\$50 000.³²

Features/Variables Defined Using Codified EHR Data

We used validated EHR algorithms to define the following variables: obesity, diabetes, hypertension, congestive heart failure, coronary artery disease, peripheral vascular

Table 1. List of Clinical Features Considered in Codified-Only and Codified+NLP models to Predict Incident AF

Variable	Codified	NLP
Age	X	
Sex	X	
Race and ethnicity	X	
Insurance	X	
Obesity	X	X
Current smoker	X	
Left ventricular hypertrophy		X
Left atrial enlargement		X
Mitral valve disease	X	X
Mitral valve prolapse	X	X
Mitral insufficiency	X	X
Mitral stenosis	X	X
Supraventricular tachycardia	X	X
Premature atrial contractions	X	X
Myocardial infarction	X	X
Chronic kidney disease	X	X
Chronic kidney disease: severe	X	X
Hyperlipidemia	X	X
Valvular disease	X	X
Prior stroke/transient ischemic attack	X	X
Systemic atherosclerosis	X	X
Cerebral atherosclerosis	X	X
Thyrotoxicosis	X	X
Hypothyroidism	X	X
Pulmonary disease	X	X
Chronic obstructive pulmonary disease	X	
Congenital heart disease	X	X
Cardiomegaly	X	X
Alcohol disorder	X	X
Pericarditis	X	X
Myocarditis	X	X
Sleep apnea	X	X
Prior cardiac surgery	X	
Hypertrophic cardiomyopathy	X	X
Other cardiomyopathy	X	X
Chronic liver disease	X	X
Cirrhosis	X	X
Liver complications	X	X
Diabetes	X	X
Hypertension	X	X
Congestive heart failure	X	X
Coronary artery disease	X	X
Peripheral vascular disease	X	X
Cerebrovascular disease	X	X
Long PR interval		X
Shortened PR interval		X
CRP	X	
NT-proBNP	X	

AF indicates atrial fibrillation; CRP, C-reactive protein; NLP, natural language processing; and NT-proBNP, N-terminal pro-B-type natriuretic peptide.

disease, and cerebrovascular disease (Table S1).^{33–36} Comorbidities without a validated algorithm were identified by a single ICD-9/ICD-10 code before and within 3 years of cohort entry. ICD-9/ICD-10 codes used are available in Table S2. We defined all components of the CHARGE-AF model using codified data (Table S3).

Features/Variables From the Narrative EHR Data Extracted Using NLP

All potential features were also ascertained using NLP, unless fully populated within structured fields (eg, demographics) or if the positive predictive value of the NLP-derived variable was low (Table 1). Health care provider progress notes, visit notes, history and physical notes, discharge summaries, laboratory notes, and cardiology test reports were processed to extract information from narrative data using a published approach²⁴ using Narrative Information Linear Extraction, an NLP package for EHR analysis (Figure 1).³⁷ Briefly, we created a dictionary of terms corresponding to potential model features (Table S4). The list of terms was mapped to concepts in the Unified Medical Language System.³⁸ For example, the terms “atrial fibrillation” and “auricular fibrillation” are different ways of expressing the same concept and are assigned a concept unique identifier (C0004238). We then processed free-text clinical notes using NLP to count the number of positive mentions of each concept unique identifier, while disregarding negative mentions, such as “no evidence of ...” Patients with at least 1 positive concept unique identifier match before cohort entry were considered to have the feature at baseline. Medical record review of 100 randomly selected patients was performed by author J.M.A. for NLP-defined variables with >10%

absolute increase compared with expected published prevalences (current smoking, myocardial infarction, mitral valve disease, mitral insufficiency, mitral valve prolapse, and mitral stenosis).^{39–43} Following this review, specific acronyms to be excluded and negation terms to be added were identified. We did not consider the NLP-defined variable for model inclusion if the positive predictive value on medical record review was <60%. For all features identified via NLP, we randomly selected 20 patients per feature with positive concept unique identifier mentions for medical record review performed by J.M.A. No additional systematic problems were identified following this review.

Outcomes

The primary outcome was incident AF within 5 years of entry into the study cohort. Incident AF status was ascertained using a previously validated EHR algorithm, which used problem list entries and inpatient or outpatient ICD-9/ICD-10 codes (positive predictive value, 96.3%).⁴⁴ Cases included in analyses occurred between 2003 and 2018 for the development and internal validation cohorts, and between 2015 and 2020 for the external validation cohort.

Model Derivation

In the development cohort, we developed 2 Cox proportional hazards models to predict AF incidence within 5 years using the following: (1) codified+NLP data (codified+NLP) and (2) codified data only (codified-only). Censoring occurred at the time of death, last primary care visit if leaving the primary care cohort, or after 5 years of follow-up. For variable selection in each model, we considered all potential model features in Table 1 and

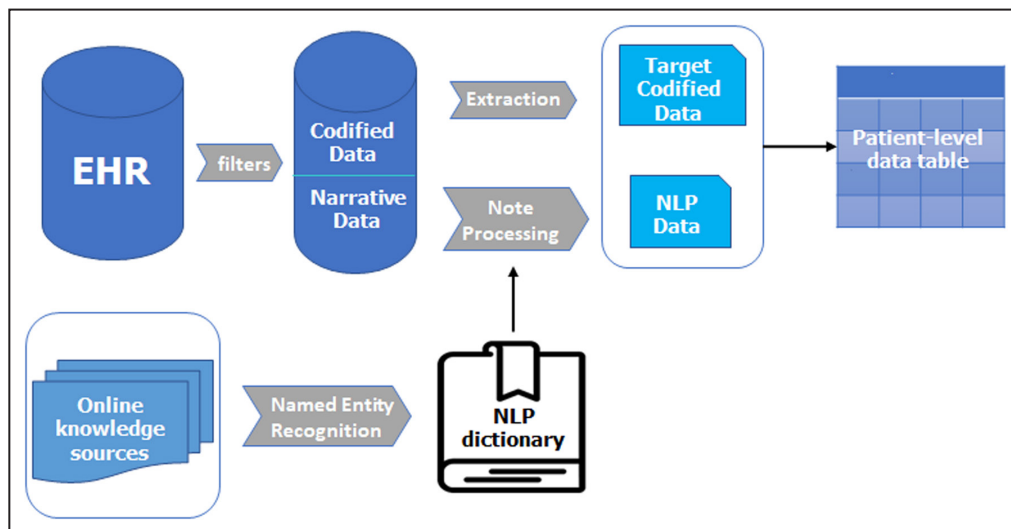


Figure 1. Overview of the process of extracting and processing codified and narrative electronic health record (EHR) data to determine patient-level predictors of atrial fibrillation. NLP indicates natural language processing.

ran 10-fold cross-validated Cox proportional hazards models using a least absolute shrinkage and selection operator penalty.⁴⁵ The largest tuning parameter (λ) was selected such that error was within 1 SD of the minimum cross-validated error.⁴⁶ As shown in Table 1, most features were defined by both codified and NLP data. For the codified+NLP model, the least absolute shrinkage and selection operator selection process was allowed to select the codified version of a feature, select the NLP version of a feature, or select both the codified and NLP version of the feature. If both the codified and NLP data version of a feature were selected, we recoded it into a single 4-level variable (0 indicates neither codified or NLP [feature not present]; 1, feature defined by codified data only; 2, feature defined by NLP data only; and 3, feature defined by both codified and NLP data). For example, both the codified and NLP versions of mitral valve disorder were selected in the codified+NLP model. In this model, there is a single variable representing both variables (0 indicates no mitral valve disorder by either codified version or NLP; 1, mitral valve disorder present in codified data, but not NLP; 2, mitral valve disorder present in NLP data, but not codified; and 3, mitral valve disorder present in both codified and NLP data). The proportional hazards assumption was verified graphically by examining the log of the minus log of the survival curve for each predictor.⁴⁷ Within each model, individual 5-year risk of AF was calculated.

Statistical Analysis

In validation cohorts, we used the coefficients from the models established in the development cohort to estimate predicted 5-year risks and used the product-limit method, which accounts for censoring, to calculate observed 5-year risks. In the external validation cohort, administrative censoring occurred at the end of follow-up in 2020. Because many participants in the external validation cohort were not followed up for 5 years, we also evaluated models over a 3-year period in supplementary analyses. For descriptive data, we calculated mean and SDs of numbers and percentages. In the validation data sets, we compared performance of the codified+NLP, codified-only, and CHARGE-AF models. To calculate the CHARGE-AF score, we used the published components and coefficients (Table S3).¹⁹ We compared hazard ratios (HRs) across groups, determined by the Cox method (based on the 16th, 50th, and 84th percentiles of the linear predictor values for each model).⁴⁸ In a normal distribution, the 84th percentile corresponds to the mean of the linear predictor +1 SD, whereas the 16th percentile corresponds to the mean of the linear predictor -1 SD.⁴⁸ We calculated the Harrell C-statistic to assess the model's ability to separate who developed AF from those who did not. We also plotted cumulative incidence curves for risk

groups. We compared the Harrell C-statistic between models with bias correction using 200 bootstrap samples. In addition to our main models, which included race as a predictor, we also assessed performance of models that included insurance and zip code-defined income instead of race in supplemental analyses. We assessed calibration by comparing the predicted and observed 5-year AF risks from each model with patients divided into risk groups based on deciles. To provide quantified summary measures of model calibration, we calculated the Integrated Calibration Index, which represents the weighted average absolute difference between observed and predicted probabilities.⁴⁹ For CHARGE-AF, we also assessed calibration after recalibration. To recalibrate, instead of using the published CHARGE-AF baseline survival when calculating predicted risk (published survival, 0.9718412736), we generated an updated baseline survival by calculating the average 5-year AF-free survival in the validation sample (updated baseline survival, 0.9331399).⁵⁰ To quantify how well each model reclassified subjects (codified+NLP compared with codified-only, and each compared with CHARGE-AF), we evaluated net reclassification improvement (NRI) for time-to-event data with censoring^{51,52} using percentile-based cut points, which allows for generalizability of the NRI when standard risk cut points are not available.⁵³ Percentile-based cut points were based on the 16th, 50th, and 84th percentiles for each model in the validation data.⁴⁸ Because the NRI may be sensitive to the cut points chosen, we also evaluated the NRI with 4 groups based on quartiles (25th, 50th, and 75th percentiles). We considered a 2-sided $P < 0.05$ to indicate statistical significance.

RESULTS

The development cohort included 32 960 patients aged ≥ 65 years without a diagnosis of AF at baseline. The mean age was 70.3 (SD, 6.9) years, 57.4% were women, and 85.3% were non-Hispanic White race and ethnicity. The internal validation cohort included 16 233 patients meeting eligibility criteria, with 2311 patients missing data to calculate CHARGE-AF, resulting in a population of 13 992 patients. The mean age was 69.5 (SD, 6.2) years, 57.5% were women, and 86.0% were non-Hispanic White race and ethnicity (Table 2). The external validation cohort included 42 234 patients meeting eligibility criteria, with 3183 patients missing data to calculate CHARGE-AF, resulting in a population of 39 051 patients (mean age, 71.5 years). Additional baseline patient characteristics for this cohort are included in Table S5.

Development Cohort

In the development cohort, there were 2839 incident AF diagnoses (5-year Kaplan-Meier cumulative

Table 2. Baseline Patient Characteristics for Development and Internal Validation Cohorts

Characteristic	Development (n=32960)	Internal validation (n=13992)	External validation (n=39051)*
Age, mean (SD), y	70.3 (6.9)	69.5 (6.2)	71.5 (6.8)
Aged 65–75	25233 (76.6)	11325 (80.9)	29153 (74.7)
Aged 75–85y	6200 (18.8)	2252 (16.1)	7680 (19.7)
Aged ≥85	1527 (4.6)	415 (3.0)	2218 (5.7)
Female sex	18910 (57.4)	8041 (47.5)	24391 (62.5)
Race and ethnicity			
Non-Hispanic White	28111 (85.3)	12031 (86.0)	30043 (76.9)
Black	1419 (4.3)	572 (4.1)	3280 (8.4)
Hispanic	1339 (4.1)	569 (4.1)	1911 (4.9)
Asian	1242 (3.8)	488 (3.5)	997 (2.6)
Unknown	447 (1.4)	177 (1.3)	1279 (3.3)
Other†	402 (1.2)	155 (1.1)	1541 (4.0)
Insurance			
Commercial	6980 (21.2)	3009 (21.5)	11557 (29.6)
Medicare	23566 (71.5)	9984 (71.4)	26383 (67.6)
Medicaid	993 (3.0)	401 (2.9)	908 (2.3)
Self-pay	1421 (4.3)	598 (4.3)	203 (0.5)
% With income <\$50000 by zip code, mean (SD)‡	27.2 (14.6)	27.0 (14.5)	21.3 (16.4)

Characteristic	Codified data	NLP data	Codified data	NLP data	Codified data	NLP data
Obesity	7707 (23.4)	8107 (24.6)	3532 (25.2)	3668 (26.2)	17398 (44.6)	10785 (27.6)
Current smoker	997 (3.0)	...	443 (3.2)
Mitral valve disorder	2988 (9.1)	6266 (19.0)	1290 (9.2)	2704 (19.3)	1137 (2.9)	7776 (19.9)
Mitral valve prolapse	2781 (8.4)	923 (2.8)	1196 (8.6)	398 (2.8)	758 (1.9)	...
Mitral valve insufficiency	326 (1.0)	5772 (17.5)	137 (1.0)	2504 (17.9)	328 (0.8)	7141 (18.3)
Mitral valve stenosis	73 (0.2)	95 (0.3)	25 (0.2)	44 (0.3)	32 (0.1)	265 (0.7)
Supraventricular tachycardia	297 (0.9)	1789 (5.4)	133 (1.0)	757 (5.4)	427 (1.1)	1276 (3.3)
Premature atrial contractions	106 (0.3)	5668 (17.2)	41 (0.3)	2357 (16.9)	91 (0.2)	6858 (17.6)
Left ventricular hypertrophy	...	3803 (11.5)	...	1623 (11.6)	...	6241 (16.0)
Left atrial enlargement	...	5543 (16.8)	...	2387 (17.1)
Myocardial infarction	1953 (5.9)	4683 (14.2)	782 (5.6)	2018 (14.4)	1565 (4.0)	4847 (12.4)
Chronic kidney disease	1033 (3.1)	915 (2.8)	475 (3.4)	385 (2.8)	3408 (8.7)	3036 (7.8)
Chronic kidney disease: severe	108 (0.3)	383 (1.2)	48 (0.3)	182 (1.3)	647 (1.7)	802 (2.1)
Hyperlipidemia	20243 (61.4)	13356 (40.5)	8605 (61.5)	5927 (42.4)
Valvular disease	941 (2.9)	491 (1.5)	391 (2.8)	213 (1.5)	1402 (3.6)	...
Prior stroke/transient ischemic attack	2182 (6.6)	7064 (21.4)	827 (5.9)	2980 (21.3)	2332 (6.0)	...
Systemic atherosclerosis	941 (2.9)	327 (1.0)	401 (2.9)	135 (1.0)	751 (1.9)	...
Cerebral atherosclerosis (codified)	1062 (3.2)	...	404 (2.9)
Thyrotoxicosis	1115 (3.4)	1190 (3.6)	484 (3.5)	508 (3.6)
Hypothyroidism	5219 (15.8)	4246 (12.9)	2197 (15.7)	1849 (13.2)
Pulmonary disease	6632 (20.1)	1926 (5.8)	2855 (20.4)	879 (6.3)
Chronic obstructive pulmonary disease	3211 (9.7)	...	1338 (9.6)	...	2328 (6.0)	...

(Continued)

Table 2. Continued

Characteristic	Codified data	NLP data	Codified data	NLP data	Codified data	NLP data
Congenital heart disease	468 (1.4)	32 (0.1)	176 (1.3)	20 (0.1)	316 (0.8)	...
Cardiomegaly	2353 (7.1)	788 (2.4)	955 (6.8)	332 (2.4)	300 (0.8)	...
Alcohol disorder	554 (1.7)	1944 (5.9)	260 (1.9)	819 (5.9)
Pericarditis	315 (1.0)	548 (1.7)	150 (1.1)	225 (1.6)
Myocarditis	37 (0.1)	125 (0.4)	11 (0.1)	57 (0.4)
Sleep apnea	1093 (3.3)	1278 (3.9)	491 (3.5)	584 (4.2)
Prior cardiac surgery	848 (2.6)	...	364 (2.6)
Hypertrophic cardiomyopathy	83 (0.3)	132 (0.4)	24 (0.2)	52 (0.4)
Other cardiomyopathy	1296 (3.9)	873 (2.7)	489 (3.5)	358 (2.6)	630 (1.6)	2186 (5.6)
Chronic liver disease	2521 (7.7)	176 (0.5)	1160 (8.3)	90 (0.6)
Cirrhosis	260 (0.8)	524 (1.6)	135 (1.0)	278 (2.0)
Liver complications	358 (1.1)	1294 (3.9)	183 (1.3)	670 (4.8)
Diabetes	4063 (12.3)	11 999 (36.4)	1741 (12.4)	5280 (37.7)
Hypertension	16295 (49.4)	21 339 (64.7)	6914 (49.4)	9093 (65.0)	...	26678 (68.3)
Congestive heart failure	663 (2.0)	2757 (8.4)	245 (1.8)	1134 (8.1)	3013 (7.7)	3344 (8.6)
Coronary artery disease	2973 (9.0)	8983 (27.3)	1163 (8.3)	3752 (26.8)	1196 (3.1)	13330 (34.1)
Peripheral vascular disease	777 (2.4)	1951 (5.9)	328 (2.3)	855 (6.1)	172 (0.4)	1576 (4.0)
Cerebrovascular disease	954 (2.9)	781 (2.4)	407 (2.9)	339 (2.4)
Long PR interval	...	156 (0.5)	...	68 (0.5)
Shortened PR interval	...	57 (0.2)	...	32 (0.2)

Data are given as number (percentage), unless otherwise indicated. NLP indicates natural language processing.

*Only final model parameters were ascertained in external validation cohort.

†“Other” Race represents American Indian/Alaskan Native, Indian, Middle Eastern, Multiracial, and those with “Other” listed in registration data.

‡Ascertained from National Neighborhood Data Archive.³² Missing income data in development population: n=1498; missing income data in internal validation population: n=507; missing income data in external validation population: n=54.

incidence, 9.5% [95% CI, 9.2%–9.9%]) and 8517 death events that occurred before an AF diagnosis or at the end of 5 years of follow-up (25.8%). The mean duration of follow-up among the entire development sample was 4.29 years; and among censored patients, it was 4.48 years. The estimated β coefficients and HRs for variables included in the codified+NLP model (22 features) and the codified-only model (23 features) are shown in Table 3. The codified+NLP model included 3 features defined from structured demographics fields, 5 features defined from codified data only, 7 features defined from NLP data only, and 7 features where both the codified and NLP versions were selected. For the 7 features where both the codified and NLP versions were selected, the prevalence identified by codified data only, NLP data only, and by both codified and NLP data is shown in Figure S1.

For the codified+NLP model, risk groups were defined on the basis of the \geq 84th percentile of 5-year predicted risk (\geq 14.6%), the 50th to <84th percentile (6.5%–<14.6%), the 16th to <50th percentile (3.6%–<6.5%), and the <16th percentile (<3.6%). For the codified-only model, the risk groups were defined as \geq 84th percentile (\geq 14.4%), the 50th to <84th percentile (6.8%–<14.4%), the 16th to <50th percentile (3.8%–<6.8%), and the <16th percentile (<3.8%). HRs

for incidence of AF within 5 years by risk group for each model are shown in Table 4. The C-statistic for the codified+NLP model was 0.744 (95% CI, 0.735–0.753), which was significantly greater ($P<0.001$) than the C-statistic for the codified-only model (0.730 [95% CI, 0.720–0.739]). Cumulative incidence plots stratified by groups of predicted risk for both models are shown in Figure 2. C-statistics were similar, although slightly reduced, in models excluding race and ethnicity and adding insurance and neighborhood-based income (Table S5). Calibration plots of both models are presented in Figure 3, with both models demonstrating good calibration (codified+NLP: Integrated Calibration Index=0.014 [95% CI, 0.003–0.024]; codified-only: Integrated Calibration Index=0.011 [95% CI, 0.001–0.021]).

Internal Validation Cohort

In the internal validation cohort, there were 1057 incident AF diagnoses (Kaplan-Meier cumulative incidence, 8.1% [95% CI, 7.7%–8.6%]) and 3201 death events (22.9%). The mean duration of follow-up among the entire internal validation sample was 4.44 years; and among censored patients, it was 4.61 years. Both the codified+NLP data model (mean, 8.4%; median, 5.8%) and codified-only model

Table 3. Estimated β Coefficients and HRs for Features Included in the Codified+NLP Data Model and the Codified-Only Data Model in the Development Cohort

Variable	Codified+NLP model		Codified-only model	
	Estimated β (SE)	HR (95% CI)	Estimated β (SE)	HR (95% CI)
Age (per 5 y)	0.067 (0.002)	1.40 (1.37–1.43)	0.067 (0.002)	1.40 (1.37–1.43)
Sex (female)	–0.504 (0.039)	0.60 (0.56–0.65)	–0.505 (0.040)	0.60 (0.56–0.65)
Race and ethnicity				
Black	–0.221 (0.167)	0.80 (0.58–1.11)	–0.089 (0.166)	0.92 (0.66–1.27)
Hispanic	–0.103 (0.172)	0.90 (0.64–1.26)	0.014 (0.171)	1.01 (0.73–1.42)
Other*	–0.516 (0.303)	0.60 (0.33–1.08)	–0.464 (0.303)	0.63 (0.35–1.14)
Unknown	0.454 (0.195)	1.58 (1.08–2.31)	0.469 (0.195)	1.60 (1.09–2.34)
Non-Hispanic White	0.273 (0.124)	1.31 (1.03–1.68)	0.347 (0.124)	1.41 (1.11–1.80)
Obesity				
Codified data	0.245 (0.044)	1.28 (1.17–1.39)
NLP data	0.222 (0.044)	1.25 (1.14–1.36)
Mitral valve disorder				
Codified data	–0.101 (0.199)	0.90 (0.61–1.33)	–0.110 (0.193)	0.90 (0.61–1.31)
NLP data	0.219 (0.165)	1.25 (0.90–1.72)
NLP+codified data	0.159 (0.246)	1.17 (0.72–1.90)
Mitral valve prolapse (codified)	0.312 (0.182)	1.37 (0.96–1.95)	0.455 (0.187)	1.58 (1.09–2.27)
Mitral valve insufficiency				
Codified data	0.208 (0.144)	1.23 (0.93–1.63)
NLP data	–0.152 (0.168)	0.86 (0.62–1.19)
Mitral valve stenosis				
Codified data	–0.085 (0.343)	0.92 (0.47–1.80)	0.641 (0.198)	1.90 (1.29–2.80)
NLP data	0.521 (0.262)	1.68 (1.01–2.82)
NLP+codified data	1.112 (0.233)	3.04 (1.93–4.79)
Supraventricular tachycardia				
Codified data	0.659 (0.212)	1.93 (1.28–2.93)	0.668 (0.121)	1.95 (1.54–2.47)
NLP data	0.405 (0.068)	1.50 (1.31–1.71)
NLP+codified data	0.634 (0.146)	1.89 (1.42–2.51)
Premature atrial contractions				
Codified data	0.543 (0.190)	1.72 (1.19–2.50)
NLP data	0.350 (0.046)	1.42 (1.30–1.55)
Left ventricular hypertrophy (NLP)	0.159 (0.053)	1.17 (1.06–1.30)
Myocardial infarction				
Codified data	0.136 (0.066)	1.15 (1.01–1.30)
NLP data	0.131 (0.052)	1.14 (1.03–1.26)
Chronic kidney disease				
Codified data	0.097 (0.073)	1.10 (0.96–1.27)
NLP data	0.041 (0.104)	1.04 (0.85–1.28)
Chronic kidney disease: severe				
Codified data	0.378 (0.138)	1.46 (1.11–1.91)
NLP data	0.367 (0.145)	1.44 (1.09–1.92)
Valvular disease (codified)	0.216 (0.087)	1.24 (1.05–1.47)	0.194 (0.097)	1.21 (1.01–1.47)
Stroke/transient ischemic attack (codified)	0.165 (0.061)	1.18 (1.05–1.33)
Systemic atherosclerosis (codified)	0.120 (0.081)	1.13 (0.96–1.32)	0.166 (0.081)	1.18 (1.01–1.38)

(Continued)

Table 3. Continued

Variable	Codified+NLP model		Codified-only model	
	Estimated β (SE)	HR (95% CI)	Estimated β (SE)	HR (95% CI)
Chronic obstructive pulmonary disease (codified)	0.150 (0.053)	1.16 (1.05–1.29)	0.221 (0.053)	1.25 (1.13–1.38)
Congenital heart disease (codified)	0.236 (0.107)	1.27 (1.03–1.56)
Cardiomegaly (codified)	0.340 (0.061)	1.41 (1.25–1.58)	0.473 (0.059)	1.60 (1.43–1.80)
Other cardiomyopathy				
Codified data	0.162 (0.082)	1.18 (1.00–1.38)	0.176 (0.075)	1.19 (1.03–1.38)
NLP data	0.377 (0.121)	1.46 (1.15–1.85)
NLP+codified data	0.059 (0.117)	1.06 (0.84–1.34)
Hypertension (NLP)	0.212 (0.048)	1.24 (1.13–1.36)
Congestive heart failure				
Codified data	0.788 (0.203)	2.20 (1.48–3.27)	0.453 (0.087)	1.57 (1.33–1.87)
NLP data	0.137 (0.064)	1.15 (1.01–3.30)
NLP+codified data	0.397 (0.103)	1.49 (1.22–1.82)
Coronary artery disease				
Codified data	0.298 (0.207)	1.35 (0.90–2.02)	0.220 (0.057)	1.25 (1.12–1.39)
NLP data	0.030 (0.051)	1.03 (0.93–1.14)
NLP+codified data	0.024 (0.065)	1.02 (0.90–1.16)
Peripheral vascular disease				
Codified data	0.195 (0.145)	1.22 (0.91–1.61)	0.242 (0.087)	1.27 (1.07–1.51)
NLP data	0.184 (0.072)	1.20 (1.05–1.38)
NLP+codified data	0.242 (0.104)	1.27 (1.04–1.56)

All risk factors are classified at the start of follow-up. The presence of an estimated β and HR indicates the variable was included in the corresponding model. For the codified+NLP model, features may be defined by only codified data (only “codified data” row has a β), defined by only NLP (only “NLP data” row has a β), or defined by both codified and NLP (“codified data,” “NLP data,” and “NLP+codified data” rows have a β). An ellipses indicates the variable was not selected for inclusion in the model. HR indicates hazard ratio; and NLP, natural language processing.

*“Other” Race represents American Indian/Alaskan Native, Indian, Middle Eastern, Multiracial, and those with “Other” listed in registration data.

(mean, 9.0%; median, 6.4%) predicted higher AF risk than CHARGE-AF (mean, 5.6%; median, 5.6%). Distributions of risk predictions for all 3 models are displayed in Figure 4.

Table 4. HRs and 95% CIs for Incidence of AF by Risk Groups Defined by the 16th, 50th, and 84th Percentiles for Each Model in the Development Cohort

Variable	Codified+NLP HR (95% CI)	Codified-only HR (95% CI)
≥84th Percentile	15.66 (12.68–19.34)	12.73 (10.50–15.42)
50th–<84th Percentile	5.82 (4.71–7.20)	4.53 (3.73–5.49)
16th–<50th Percentile	2.19 (1.75–2.74)	2.03 (1.65–2.48)
<16th Percentile

The 5-year risk of AF was calculated for each model as $1 - s_0^{\exp(\sum \beta X - \sum \beta Y)}$ where s_0 is the average AF-free survival probability at 5 years in the sample, $\sum \beta X$ is an individual’s risk score calculated using the regression coefficients from the development model (β) and the level for each risk factor (X), and $\sum \beta Y$ is the average score of the sample. For the codified+NLP model, risk was calculated as $1 - 0.925769^{\exp(\sum \beta X - 5.1621047)}$; and for codified-only model, as $1 - 0.9234408^{\exp(\sum \beta X - 4.9497951)}$. AF indicates atrial fibrillation; HR, hazard ratio; and NLP, natural language processing.

In the internal validation cohort, the C-statistic was modestly higher in the codified+NLP data model (0.735 [95% CI, 0.720–0.749]) compared with the codified-only model (0.729 [95% CI, 0.715–0.744]; $P=0.06$) and the CHARGE-AF model (0.717 [95% CI, 0.703–0.731]; $P=0.002$). The C-statistic for the codified-only model was also significantly higher than for CHARGE-AF ($P=0.01$). HRs for AF by risk groups defined by the 16th, 50th, and 84th percentiles for each model are shown in Table 5. Like the development cohort, the C-statistic was similar, but slightly reduced, when excluding race and ethnicity and adding insurance and neighborhood-based income to the model (Table S5). Cumulative incidence plots stratified by risk groups of predicted risk are shown in Figure 5. Each model demonstrates separation in the cumulative incidence curves by risk group, with the codified+NLP data model (Kaplan-Meier estimate for ≥84th percentile, 21.9%; and 50th–<84th percentile, 9.1%) and codified-only model (Kaplan-Meier estimate for ≥84th percentile, 21.4%; and 50th–<84th percentile, 9.1%) having greater separation between the highest and next highest risk group compared

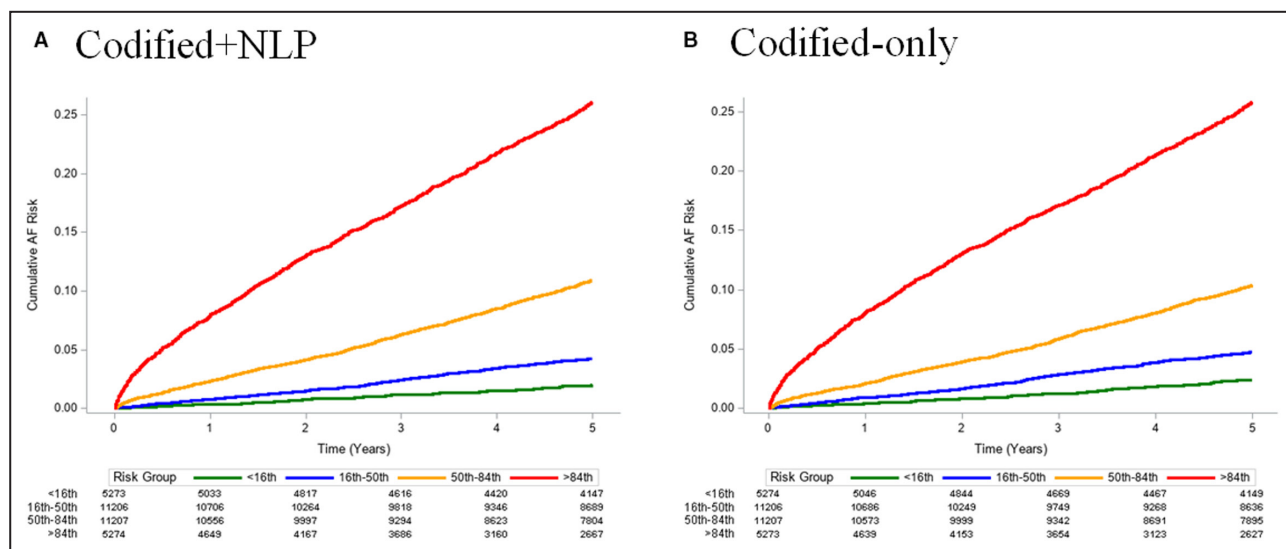


Figure 2. Cumulative incidence plots stratified by groups of predicted risk in development cohort. **A**, Depicts the cumulative risk of atrial fibrillation (AF) by groups defined by <16th percentile (green, <3.6%), 16th to <50th percentile (blue, 3.6%–<6.5%), 50th to <84th percentile (orange, 6.5%–<14.6%), and ≥84th percentile (red, ≥14.6%) of predicted AF risk for the codified+natural language processing (NLP) model. **B**, Depicts the cumulative risk of AF by groups defined by <16th percentile (green, <3.8%), 16th to <50th percentile (blue, 3.8%–<6.8%), 50th to <84th percentile (orange, 6.8%–<14.4%), and ≥84th percentile (red, ≥14.4%) of predicted AF risk for the codified-only model.

with CHARGE-AF (Kaplan-Meier estimate for ≥84th percentile, 19.3%; and 50th–<84th percentile, 9.9%). [Table 6](#) summarizes percentile-based NRI results. Codified+NLP and codified-only models demonstrate

significantly improved reclassification compared with CHARGE-AF according to the overall NRI, with both positive event and nonevent NRI. The overall NRI comparing codified+NLP and the codified-only model was

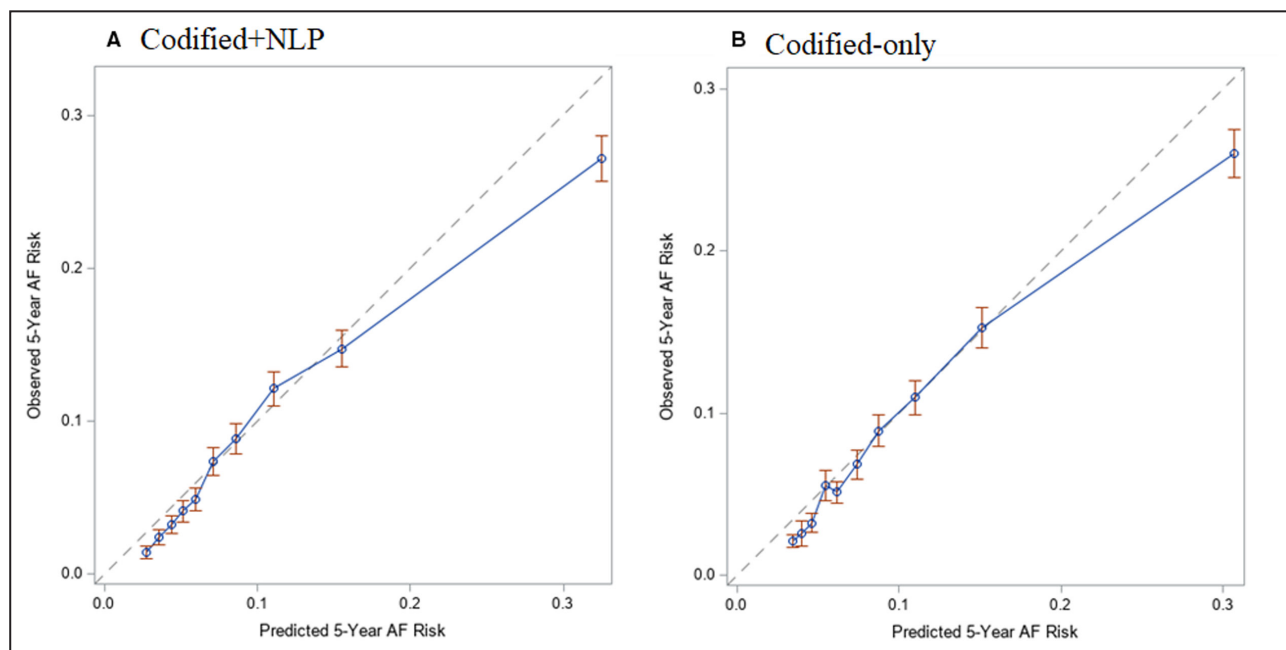


Figure 3. Calibration plots of observed 5-year atrial fibrillation (AF) risk vs predicted 5-year AF risk with patients divided into risk groups based on deciles in the development cohort. **A**, Depicts the plot of observed 5-year AF risk (y axis) vs predicted 5-year AF risk (x axis) for the codified+natural language processing (NLP) model in blue, whereas the optimal calibration is shown in gray. **B**, Depicts the plot of observed 5-year AF risk (y axis) vs predicted 5-year AF risk (x axis) for the codified-only model in blue, whereas the optimal calibration is shown in gray.

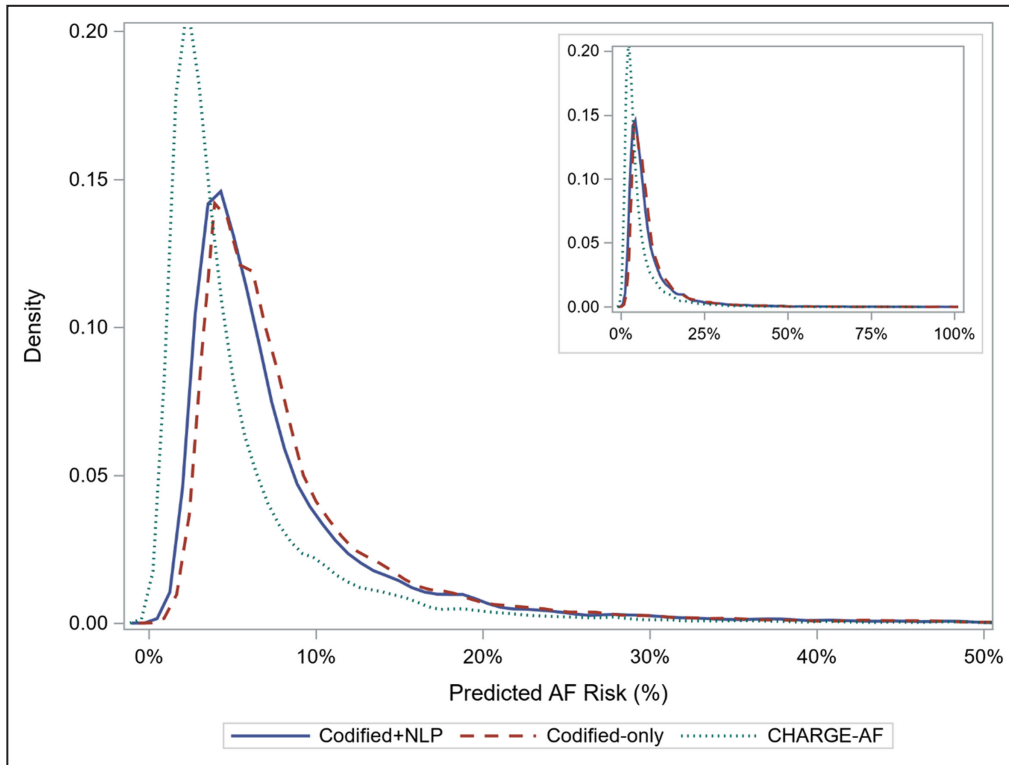


Figure 4. Distributions of predicted risk for codified+natural language processing (NLP), codified-only, and Cohorts for Heart and Aging Research in Genomic Epidemiology Atrial Fibrillation (CHARGE-AF) models in the internal validation cohort. AF indicates atrial fibrillation.

small, and the CI crosses 0. NRI results were similar when establishing groups based on quartiles.

The codified+NLP data model and the codified-only model both appeared well calibrated in the internal validation cohort (Figure 6). In contrast, calibration of CHARGE-AF was poor, with the plot of observed 5-year AF risk versus predicted 5-year AF risk demonstrating underestimation of AF risk (Figure 6). Calibration remained poor, even after recalibrating CHARGE-AF with the baseline survival of the internal validation sample (Figure S2). The Integrated Calibration Index estimate was smallest for the codified+NLP model (0.011 [95%

CI, 0.002–0.020]), compared with codified-only (0.016 [95% CI, 0.006–0.027]) and CHARGE-AF (0.023 [95% CI, 0.009–0.036]).

External Validation Cohort

In the external validation cohort, there were 2226 incident AF diagnoses (Kaplan-Meier cumulative incidence, 7.3% [95% CI, 7.0%–7.6%]). The mean duration of follow-up among the entire external validation sample was 3.64 years; and among censored patients, it was 3.75 years. The C-statistic was higher in the

Table 5. HRs and 95% CIs for Incidence of AF by Risk Groups Defined by the 16th, 50th, and 84th Percentiles for Each Model in Internal Validation Cohort

Variable	Codified+NLP HR (95% CI)	Codified-only HR (95% CI)	CHARGE-AF HR (95% CI)
≥84th Percentile	16.28 (11.31–23.43)	14.71 (10.47–20.66)	13.05 (9.20–18.53)
50th–<84th Percentile	6.19 (4.30–8.92)	5.67 (4.03–7.98)	6.35 (4.48–9.00)
16th–<50th Percentile	2.59 (1.77–3.80)	2.54 (1.77–3.63)	2.60 (1.81–3.75)
<16th Percentile

The 5-year risk of AF was calculated for each model as $1 - s_0^{\exp(\sum \beta X - \sum \beta Y)}$ where s_0 is the average AF-free survival probability at 5 years in the sample. For codified+NLP and codified-only, $\sum \beta X$ is an individual's risk score calculated using the regression coefficients from the development model (β) and the level for each risk factor (X), and $\sum \beta Y$ is the average score of the sample. For CHARGE-AF, $\sum \beta X$ is an individual's CHARGE-AF score calculated using the regression coefficients from the original CHARGE-AF publication (β) and the level for each risk factor (X), and $\sum \beta Y$ is a published constant (12.5815600).¹⁹ For the codified+NLP model, risk was calculated as $1 - 0.925769^{\exp(\sum \beta X - 5.1621047)}$; for codified-only, as $1 - 0.9234408^{\exp(\sum \beta X - 4.9497951)}$; and for CHARGE-AF, as $1 - 0.9718412736^{\exp(\sum \beta X - 12.5815600)}$. AF indicates atrial fibrillation; CHARGE-AF, Cohorts for Heart and Aging Research in Genomic Epidemiology AF; HR, hazard ratio; and NLP, natural language processing.

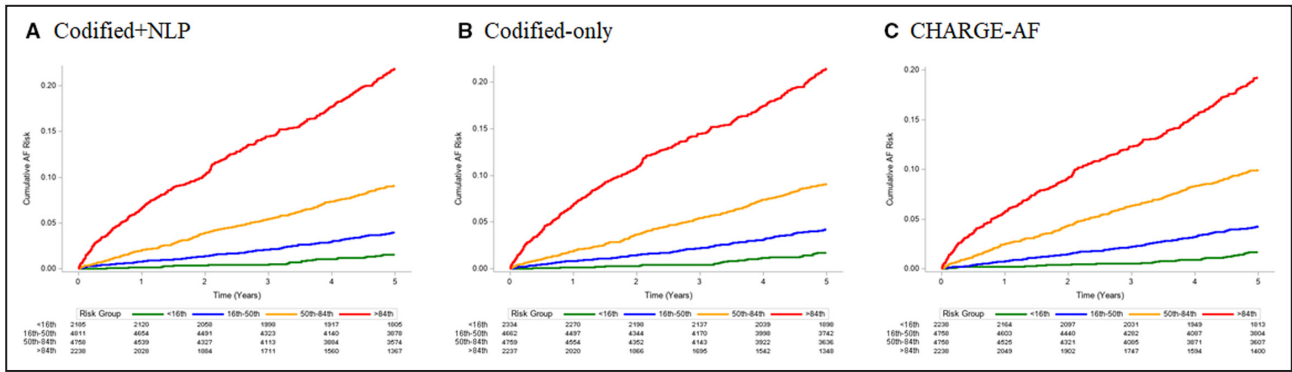


Figure 5. Cumulative incidence plots stratified by groups of predicted risk in the internal validation cohort. **A**, Depicts the cumulative risk of atrial fibrillation (AF) by groups defined by <16th percentile (green, <3.4%), 16th to <50th percentile (blue, 3.4%–<5.8%), 50th to <84th percentile (orange, 5.8%–<12.5%), and ≥84th percentile (red, ≥12.5%) of predicted AF risk for the codified+natural language processing (NLP) model. **B**, Depicts the cumulative risk of AF by groups defined by <16th percentile (green, <3.7%), 16th to <50th percentile (blue, 3.7%–<6.4%), 50th to <84th percentile (orange, 6.4%–<13.2%), and ≥84th percentile (red, ≥13.2%) of predicted AF risk for the codified-only model. **C**, Depicts the cumulative risk of AF by groups defined by <16th percentile (green, <1.8%), 16th to <50th percentile (blue, 1.8%–<3.5%), 50th to <84th percentile (orange, 3.5%–<8.8%), and ≥84th percentile (red, ≥8.8%) of predicted AF risk for the Cohorts for Heart and Aging Research in Genomic Epidemiology AF (CHARGE-AF) model.

codified+NLP data model (0.750 [95% CI, 0.740–0.760]) compared with the codified-only model (0.738 [95% CI, 0.727–0.748]; $P<0.001$) and the CHARGE-AF model (0.735 [95% CI, 0.725–0.746]; $P<0.001$). HRs for AF by risk groups defined by the 16th, 50th, and 84th percentiles for each model are shown in Table S6. Cumulative incidence plots stratified by risk groups of predicted risk are shown in Figure S3. Table S7 summarizes the NRI results. In external validation, the codified+NLP model demonstrated significantly improved reclassification compared with CHARGE-AF and codified-only by overall NRI, with most of the improvement attributable to correctly up-classifying events. The codified+NLP and codified-only models were not as well calibrated in the external validation cohort, with both models overestimating AF risk (Figure S4). Evaluation of model performance was similar when limiting to 3 years of follow-up (Table S8 and Figures S5 and S6).

DISCUSSION

In >86 000 older primary care patients, we derived and validated models to predict the incidence of AF. In one model, we used only codified data. In the second model, we added information from narrative data using

NLP to codified data. We observed that incorporating codified and NLP-derived data modestly improved model performance compared with using only codified data. Furthermore, both newly developed models were superior to an established AF prediction model, CHARGE-AF. In external validation, performance remained modestly greater for the model that incorporated NLP-derived data. Our findings suggest that incorporating narrative EHR data using NLP may improve identification of clinical predictors of AF and yield a better prediction model.

Our results demonstrate that clinical and demographic features routinely ascertained from the EHR can predict 5-year risk of AF, and that incorporating narrative data from the EHR using NLP can modestly improve model performance compared with using codified data only. We developed 2 models to predict incident AF in this study, 1 using only codified EHR data to identify clinical features and 1 that added narrative data using NLP. Prior studies have demonstrated the utility of NLP in increasing the sensitivity and positive predictive value of clinical features compared with using codified data only.^{22–26,54} NLP may reduce misclassification of clinical features by extracting information from narrative text that would

Table 6. Percentile-Based NRI With Groups Determined by 16th, 50th, and 84th Percentiles of Each Model in Internal Validation Cohort

Variable	Overall NRI	Event NRI	Nonevent NRI
Codified+NLP vs CHARGE-AF	0.070 (0.033–0.113)	0.052 (0.017–0.093)	0.018 (0.006–0.029)
Codified-only vs CHARGE-AF	0.054 (0.015–0.091)	0.036 (–0.002–0.069)	0.019 (0.007–0.030)
Codified+NLP vs codified-only	0.016 (–0.012–0.044)	0.020 (–0.007–0.046)	–0.004 (–0.013 to –0.005)

CHARGE-AF indicates Cohorts for Heart and Aging Research in Genomic Epidemiology Atrial Fibrillation; NLP, natural language processing; and NRI, net reclassification improvement.

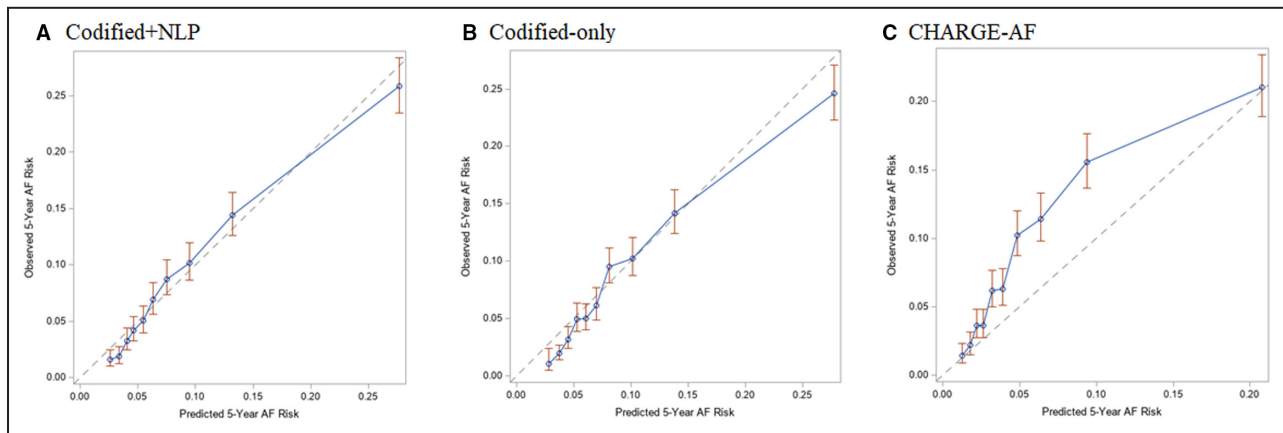


Figure 6. Calibration plots of observed 5-year atrial fibrillation (AF) risk vs predicted 5-year AF risk with patients divided into risk groups based on deciles in the internal validation cohort.

A, Depicts the plot of observed 5-year AF risk (y axis) vs predicted 5-year AF risk (x axis) for the codified+natural language processing (NLP) model in blue, whereas the optimal calibration is shown in gray. **B**, Depicts the plot of observed 5-year AF risk (y axis) vs predicted 5-year AF risk (x axis) for the codified-only model in blue, whereas the optimal calibration is shown in gray. **C**, Depicts the plot of observed 5-year AF risk (y axis) vs predicted 5-year AF risk (x axis) for the Cohorts for Heart and Aging Research in Genomic Epidemiology AF (CHARGE-AF) model in blue, whereas the optimal calibration is shown in gray.

otherwise not be considered for an algorithm. Using least absolute shrinkage and selection operator as a variable selection strategy, we found the NLP version of several features to be selected in the model over the codified version (eg, obesity). In other situations, both the NLP and codified versions were selected in the models (eg, supraventricular tachycardia). Among 22 model features included in the codified+NLP model, the NLP version was selected instead of the codified version for 6 features (eg, obesity), whereas both the NLP and codified versions were selected for 7 features (eg, supraventricular tachycardia) (Table 3). In the development and validation populations, the C-statistic and NRI were better in the model incorporating NLP compared with the model using codified data only. However, the magnitude of the improvement was modest. Accurately and efficiently assessing individual risk estimates for AF in clinical settings may enable targeted risk-based screening or prevention interventions and may be implemented within the EHR to guide clinical decision making. For example, assessing risk of AF could serve as a guide for use of longer-term cardiac monitor in survivors of acute stroke.^{17,35}

The CHARGE-AF risk model was developed in community-based research cohorts, which may represent subjects at lower underlying risk of developing AF and include more accurate assessment of covariates. External validation of CHARGE-AF in EHR-based cohorts has demonstrated poor calibration.^{20,21,35} Within both our internal and external validation samples, CHARGE-AF underestimated risk of developing AF, even after recalibrating to the baseline risk of each sample. In contrast, our newly developed models using codified data only or NLP and codified data

demonstrated modestly improved C-statistics compared with CHARGE-AF and were well calibrated in an internal validation population, although calibration was not as good in external validation. Our findings support prior work suggesting that prediction models developed within a clinical setting using EHR data perform better in real-world clinical settings than models derived from community-based research cohort data.²¹ This may be the result of differences in data quality and/or differences in populations between research cohorts and clinical settings.

NLP provides an opportunity to access the vast amount of data in narrative notes within EHRs. The use of NLP as part of clinical operations will be dependent on scaling the rapid processing of large amounts of data and optimizing the ability of NLP tools to efficiently and accurately identify clinical factors.^{55–57} Quality of narrative data in the EHR may differ over time and by provider and institution. Thus, porting our algorithm to another institution will require at minimum validation, and potential refitting before implementation if there are large differences in the population. In our samples, the addition of NLP provided a statistically significant improvement in predictive performance, but the magnitude of improvement was modest. Most components of the codified+NLP model, except for those extracted from cardiology reports, were available in a codified format. However, we did observe for most variables that either the NLP version was more predictive or the combination of codified and NLP was more predictive. If easily implemented, using NLP may be worthwhile to achieve the best possible model. The added benefit of NLP may differ at different institutions, depending on the quality of codified and narrative data. For

institutions unable to implement NLP, our model using only codified data performed well and was an improvement over CHARGE-AF.

This study has several potential limitations. Our models were developed within a single-center tertiary academic primary care practice network with patients who were largely of European ancestry, so generalizability may be limited. Like CHARGE-AF, we included race and ethnicity in our models.¹⁹ Although Black individuals have consistently had a lower prevalence of clinically detected AF compared with White individuals, this may represent differential detection rather than a biological mechanism.⁵⁸ Race and ethnicity may represent, and potentially poorly represent, a proxy for social determinants of health. As such, we presented models and evaluated the C-statistic adding insurance and insurance plus zip code–defined income as predictors instead of race and ethnicity and the performance of the models did not materially deteriorate. Ascertainment of clinical features and incidence of AF were based on retrospective assessment of EHR documentation, which may be associated with misclassification when classifying features using codified or NLP data. Data on clinical features and ascertainment of incident AF are limited to what is available within the Mass General Brigham EHR. We do not have the ability to fully ascertain information on clinical features or incident diagnoses of AF for patients seen outside of our network.

In conclusion, estimation of the 5-year risk of AF in a primary care population can be modestly improved by using NLP to incorporate narrative EHR data. The amount of data available in unstructured form within EHRs is immense. Optimizing the use of NLP tools to meaningfully process these data offers an opportunity to improve prognostic models used within clinical care.

ARTICLE INFORMATION

Received March 6, 2022; accepted June 29, 2022.

Affiliations

Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA (J.M.A., Y.C., S.J.A., D.E.S.); Harvard Medical School, Boston, MA (J.M.A., Y.C., T.C., K.P.L., K.B.W., S.N.M., S.J.A., D.E.S.); Cardiovascular Research Center (X.W., S.K., S.A.L.); and Division of Cardiology (S.K.), Massachusetts General Hospital, Boston, MA; Department of Neurology, Brigham and Women's Hospital, Boston, MA (C.D.A.); Department of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, Boston, MA (K.D., D.W., T.C., K.P.L.); Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA (K.B.W.); Research Information Science and Computing, Mass General Brigham, Somerville, MA (S.N.M.); and Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston, MA (S.A.L.).

Acknowledgments

We thank Meghan L. Rieu-Werden, BS, for help with data acquisition and cohort generation.

Sources of Funding

Drs Ashburner, Lubitz, and Anderson are supported by American Heart Association (AHA) 18SFRN34250007 and 18SFRN34150007. Dr Ashburner

is supported by National Institutes of Health (NIH) grant K01HL148506. Dr Lubitz is supported by NIH grants R01HL139731 and R01HL157635. Dr Anderson is supported by NIH grants R01NS103924 and U01NS069763, AHA-Bugher Foundation Centers for Excellence in Hemorrhagic Stroke, the Massachusetts General Hospital Center for Neuroscience, and Henry and Allison McCance Center for Brain Health. Dr Liao is supported by the NIH grant P30AR072577 (VERITY Bioinformatics Core).

Disclosures

Dr Lubitz receives sponsored research support from Bristol Myers Squibb, Pfizer, Boehringer Ingelheim, Fitbit, Medtronic, Premier, and IBM; and has consulted for Bristol Myers Squibb, Pfizer, Blackstone Life Sciences, and Invitae. Dr Anderson receives sponsored research support from Bayer AG; and has consulted for ApoPharma, Inc. Dr Singer receives research support from Bristol-Myers Squibb; and has consulted for Boehringer Ingelheim, Bristol-Myers Squibb, Fitbit, Johnson and Johnson, Merck, and Pfizer. Dr Atlas receives sponsored research support from Bristol Myers Squibb/Pfizer; and has consulted for Bristol Myers Squibb/Pfizer and Fitbit. Dr Murphy receives sponsored research support from AstraZeneca Pharmaceuticals LP, Analysis Group, Inc, Radius Health Inc, and Amgen, Inc. The remainder of the authors report no disclosures.

Supplemental Material

Tables S1–S8
Figures S1–S6

REFERENCES

- Go AS, Hylek EM, Phillips KA, Chang Y, Henault LE, Selby JV, Singer DE. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. *JAMA*. 2001;285:2370–2375. doi: 10.1001/jama.285.18.2370
- January CT, Wann LS, Alpert JS, Calkins H, Cigarroa JE, Cleveland JC, Conti JB, Ellinor PT, Ezekowitz MD, Field ME, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. *Circulation*. 2014;130:e199–e267. doi: 10.1161/CIR.0000000000000041
- Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke*. 1991;22:983–988. doi: 10.1161/01.str.22.8.983
- 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.
- Colilla S, Crow A, Petkun W, Singer DE, Simon T, Liu X. Estimates of current and future incidence and prevalence of atrial fibrillation in the U.S. adult population. *Am J Cardiol*. 2013;112:1142–1147. doi: 10.1016/j.amjcard.2013.05.063
- Borowsky LH, Regan S, Chang Y, Ayres A, Greenberg SM, Singer DE. First diagnosis of atrial fibrillation at the time of stroke. *Cerebrovasc Dis*. 2017;43:192–199. doi: 10.1159/000457809
- Lubitz SA, Yin X, McManus DD, Weng LC, 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. doi: 10.1161/STROKEAHA.116.015071
- Bjorck S, Palaszewski B, Friberg L, Bergfeldt L. Atrial fibrillation, stroke risk, and warfarin therapy revisited: a population-based study. *Stroke*. 2013;44:3103–3108. doi: 10.1161/STROKEAHA.113.002329
- Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, Pogue J, Reilly PA, Themeles E, Varrone J, et al. Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med*. 2009;361:1139–1151. doi: 10.1056/NEJMoa0905561
- Granger CB, Alexander JH, McMurray JJ, Lopes RD, Hylek EM, Hanna M, Al-Khalidi HR, Ansell J, Atar D, Avezum A, et al. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med*. 2011;365:981–992. doi: 10.1056/NEJMoa1107039
- 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

12. Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, Breithardt G, Halperin JL, Hankey GJ, Piccini JP, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med*. 2011;365:883–891. doi: [10.1056/NEJMoa1009638](https://doi.org/10.1056/NEJMoa1009638)
13. Fitzmaurice DA, Hobbs FD, Jowett S, Mant J, Murray ET, Holder R, Raftery JP, Bryan S, Davies M, Lip GY, et al. Screening versus routine practice in detection of atrial fibrillation in patients aged 65 or over: cluster randomised controlled trial. *BMJ*. 2007;335:383. doi: [10.1136/bmj.39280.660567.55](https://doi.org/10.1136/bmj.39280.660567.55)
14. Kaasenbrood F, Hollander M, de Bruijn SH, Dolmans CP, Tieleman RG, Hoes AW, Rutten FH. Opportunistic screening versus usual care for diagnosing atrial fibrillation in general practice: a cluster randomised controlled trial. *Br J Gen Pract*. 2020;70:e427–e433. doi: [10.3399/bjgp20X708161](https://doi.org/10.3399/bjgp20X708161)
15. Uittenbogaart SB, Verbiest-van Gurp N, Lucassen WAM, Winkens B, Nielen M, Erkens PMG, Knottnerus JA, van Weert HCPM, Stoffers HEJH. Opportunistic screening versus usual care for detection of atrial fibrillation in primary care: cluster randomised controlled trial. *BMJ*. 2020;370:m3208. doi: [10.1136/bmj.m3208](https://doi.org/10.1136/bmj.m3208)
16. Lubitz SA, Atlas SJ, Ashburner JM, Lipsanopoulos ATT, Borowsky LH, Guan W, Khurshid S, Ellinor PT, Chang Y, McManus DD, et al. Screening for atrial fibrillation in older adults at primary care visits: VITAL-AF randomized controlled trial. *Circulation*. 2022;145:946–954. doi: [10.1161/CIRCULATIONAHA.121.057014](https://doi.org/10.1161/CIRCULATIONAHA.121.057014)
17. Ashburner JM, Khurshid S, Atlas SJ, Singer DE, Lubitz SA. Point-of-care screening for atrial fibrillation: where are we, and where do we go next? *Cardiovasc Digit Health J*. 2021;2:294–297. doi: [10.1016/j.cvdhj.2021.10.001](https://doi.org/10.1016/j.cvdhj.2021.10.001)
18. Khurshid S, Mars N, Haggerty CM, Huang Q, Weng L-C, Hartzel DN, Center RG, Lunetta KL, Ashburner JM, Anderson CD, et al. Predictive accuracy of a clinical and genetic risk model for atrial fibrillation. *Circ Genomic Precis Med*. 2021;14:e003355. doi: [10.1161/CIRCGEN.121.003355](https://doi.org/10.1161/CIRCGEN.121.003355)
19. 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](https://doi.org/10.1161/JAHA.112.000102)
20. 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 Cardiol*. 2016;1:1007–1013. doi: [10.1001/jamacardio.2016.3366](https://doi.org/10.1001/jamacardio.2016.3366)
21. 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. doi: [10.1016/j.jacep.2019.07.016](https://doi.org/10.1016/j.jacep.2019.07.016)
22. Himes BE, Dai Y, Kohane IS, Weiss ST, Ramoni MF. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *J Am Med Inform Assoc*. 2009;16:371–379. doi: [10.1197/jamia.M2846](https://doi.org/10.1197/jamia.M2846)
23. Wasfy JH, Singal G, O'Brien C, Blumenthal DM, Kennedy KF, Strom JB, Spertus JA, Mauri L, Normand S-LT, Yeh RW. Enhancing the prediction of 30-day readmission after percutaneous coronary intervention using data extracted by querying of the electronic health record. *Circ Cardiovasc Qual Outcomes*. 2015;8:477–485. doi: [10.1161/CIRCOUTCOMES.115.001855](https://doi.org/10.1161/CIRCOUTCOMES.115.001855)
24. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthkrishnan AN, Gainer VS, Shaw SY, Xia Z, Szolovits P, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*. 2015;350:h1885. doi: [10.1136/bmj.h1885](https://doi.org/10.1136/bmj.h1885)
25. Liao KP, Ananthkrishnan AN, Kumar V, Xia Z, Cagan A, Gainer VS, Goryachev S, Chen P, Savova GK, Agniel D, et al. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PLoS One*. 2015;10:e0136651. doi: [10.1371/journal.pone.0136651](https://doi.org/10.1371/journal.pone.0136651)
26. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. *AMIA Annu Symp Proc*. 2008;2008:404–408.
27. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–247. doi: [10.1016/j.jclinepi.2015.04.005](https://doi.org/10.1016/j.jclinepi.2015.04.005)
28. Atlas SJ, Chang Y, Lasko TA, Chueh HC, Grant RW, Barry MJ. Is this “my” patient? Development and validation of a predictive model to link patients to primary care providers. *J Gen Intern Med*. 2006;21:973–978. doi: [10.1111/j.1525-1497.2006.00509.x](https://doi.org/10.1111/j.1525-1497.2006.00509.x)
29. Atlas SJ, Grant RW, Ferris TG, Chang Y, Barry MJ. Patient-physician connectedness and quality of primary care. *Ann Intern Med*. 2009;150:325–335. doi: [10.7326/0003-4819-150-5-200903030-00008](https://doi.org/10.7326/0003-4819-150-5-200903030-00008)
30. Murphy SN, Morgan MM, Barnett GO, Chueh HC. Optimizing health-care research data warehouse design through past COSTAR query analysis. *Proc AMIA Symp*. 1999;892–896.
31. Mason JW, Ramseth DJ, Chanter DO, Moon TE, Goodman DB, Mendzelevski B. Electrocardiographic reference ranges derived from 79743 ambulatory subjects. *J Electrocardiol*. 2007;40:228–234. doi: [10.1016/j.jelectrocard.2006.09.003](https://doi.org/10.1016/j.jelectrocard.2006.09.003)
32. Robert Melendez U of MI for SR, Philippa Clarke U of MI for SR, Anam Khan U of MI for SR, Iris Gomez-Lopez U of MI for SR, Mao Li U of MI for SR, Megan Chenoweth U of MI for SR. National Neighborhood Data Archive (NaNDA): Socioeconomic Status and Demographic Characteristics of ZIP Code Tabulation Areas, United States, 2008–2017 [Internet]. 2020. Available at: <https://www.openicpsr.org/openicpsr/project/120462/version/V1/view>. Accessed Jan 25, 2022.
33. Grant RW, Cagliero E, Sullivan CM, Dubey AK, Estey GA, Weil EM, Gesmundo J, Nathan DM, Singer DE, Chueh HC, et al. A controlled trial of population management: diabetes mellitus: putting evidence into practice (DM-PEP). *Diabetes Care*. 2004;27:2299–2305. doi: [10.2337/diacare.27.10.2299](https://doi.org/10.2337/diacare.27.10.2299)
34. Leong A, Berkowitz SA, Triant VA, Porneala B, He W, Atlas SJ, Wexler DJ, Meigs JB. Hypoglycemia in diabetes mellitus as a coronary artery disease risk factor in patients at elevated vascular risk. *J Clin Endocrinol Metab*. 2016;101:659–668. doi: [10.1210/jc.2015-3169](https://doi.org/10.1210/jc.2015-3169)
35. Ashburner JM, Wang X, Li X, Khurshid S, Ko D, Trisini Lipsanopoulos A, Lee PR, Carmichael T, Turner AC, Jackson C, et al. Re-CHARGE-AF: recalibration of the CHARGE-AF model for atrial fibrillation risk prediction in patients with acute stroke. *J Am Heart Assoc*. 2021;10:e022363. doi: [10.1161/JAHA.121.022363](https://doi.org/10.1161/JAHA.121.022363)
36. Berkowitz SA, Atlas SJ, Grant RW, Wexler DJ. Individualizing HbA1c targets for patients with diabetes: impact of an automated algorithm within a primary care network. *Diabet Med J Br Diabet Assoc*. 2014;31:839–846.
37. Yu S, Cai T. NILE: Fast Natural Language Processing for Electronic Health Records. ArXiv13116063 Cs [Internet]. 2019; Available at: <http://arxiv.org/abs/1311.6063>. Accessed Feb 7, 2022.
38. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32:D267–D270.
39. Cornelius ME, Loretan CG, Wang TW, Jamal A, Homa DM. Tobacco product use among adults - United States, 2020. *MMWR Morb Mortal Wkly Rep*. 2022;71:397–405. doi: [10.15585/mmwr.mm7111a1](https://doi.org/10.15585/mmwr.mm7111a1)
40. Yu B, Akushevich I, Yashkin AP, Kravchenko J. Epidemiology of geographic disparities of myocardial infarction among older adults in the United States: analysis of 2000–2017 medicare data. *Front Cardiovasc Med*. 2021;8:707102. doi: [10.3389/fcvm.2021.707102](https://doi.org/10.3389/fcvm.2021.707102)
41. Delling FN, Vasan RS. Epidemiology and pathophysiology of mitral valve prolapse. *Circulation*. 2014;129:2158–2170. doi: [10.1161/CIRCULATIONAHA.113.006702](https://doi.org/10.1161/CIRCULATIONAHA.113.006702)
42. Tsao CW, Aday AW, Almarazgoq ZI, Alonso A, Beaton AZ, Bittencourt MS, Boehme AK, Buxton AE, Carson AP, Commodore-Mensah Y, et al. Heart disease and stroke statistics-2022 update: a report from the American Heart Association. *Circulation*. 2022;145:e153–e639. doi: [10.1161/CIR.0000000000001052](https://doi.org/10.1161/CIR.0000000000001052)
43. Vahanian A, Beyersdorf F, Praz F, Milojevic M, Baldus S, Bauersachs J, Capodanno D, Conradi L, De Bonis M, De Paulis R, et al. 2021 ESC/EACTS Guidelines for the management of valvular heart disease. *Eur Heart J*. 2022;43:561–632. doi: [10.1093/eurheartj/ehab395](https://doi.org/10.1093/eurheartj/ehab395)
44. Ashburner JM, Singer DE, Lubitz SA, Borowsky LH, Atlas SJ. Changes in use of anticoagulation in patients with atrial fibrillation within a primary care network associated with the introduction of direct oral anticoagulants. *Am J Cardiol*. 2017;120:786–791. doi: [10.1016/j.amjcard.2017.05.055](https://doi.org/10.1016/j.amjcard.2017.05.055)
45. Tibshirani R. The lasso method for variable selection in the Cox Model. *Stat Med*. 1997;16:385–395. doi: [10.1002/\(sici\)1097-0258\(19970228\)16:4](https://doi.org/10.1002/(sici)1097-0258(19970228)16:4)
46. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. 2nd ed. Berlin/Heidelberg, Germany: Springer; 2009.
47. Kuitunen I, Ponkilainen VT, Uimonen MM, Eskelinen A, Reito A. Testing the proportional hazards assumption in cox regression and dealing

- with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review. *BMC Musculoskelet Disord*. 2021;22:489. doi: [10.1186/s12891-021-04379-2](https://doi.org/10.1186/s12891-021-04379-2)
48. Cox D. Note on grouping. *J Am Stat Assoc*. 1957;52:543–547.
 49. Austin PC, Harrell FE, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Stat Med*. 2020;39:2714–2742. doi: [10.1002/sim.8570](https://doi.org/10.1002/sim.8570)
 50. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. doi: [10.1186/s12916-019-1466-7](https://doi.org/10.1186/s12916-019-1466-7)
 51. 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. *Stat Med*. 2008;27:157–172. discussion 207–212
 52. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30:11–21.
 53. McKearnan SB, Wolfson J, Vock DM, Vazquez-Benitez G, O'Connor PJ. Performance of the net reclassification improvement for nonnested models and a novel percentile-based alternative. *Am J Epidemiol*. 2018;187:1327–1335. doi: [10.1093/aje/kwx374](https://doi.org/10.1093/aje/kwx374)
 54. Zhao SS, Hong C, Cai T, Xu C, Huang J, Ermann J, Goodson NJ, Solomon DH, Cai T, Liao KP. Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records. *Rheumatol Oxf Engl*. 2020;59:1059–1065. doi: [10.1093/rheumatology/kez375](https://doi.org/10.1093/rheumatology/kez375)
 55. Maddox TM, Matheny MA. Natural language processing and the promise of big data: small step forward, but many miles to go. *Circ Cardiovasc Qual Outcomes*. 2015;8:463–465. doi: [10.1161/CIRCOUTCOMES.115.002125](https://doi.org/10.1161/CIRCOUTCOMES.115.002125)
 56. Liao KP, Sun J, Cai TA, Link N, Hong C, Huang J, Huffman JE, Grönsbell J, Zhang Y, Ho Y-L, et al. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. *J Am Med Inform Assoc*. 2019;26:1255–1262. doi: [10.1093/jamia/ocz066](https://doi.org/10.1093/jamia/ocz066)
 57. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Cai T. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc*. 2015;22:993–1000. doi: [10.1093/jamia/ocv034](https://doi.org/10.1093/jamia/ocv034)
 58. Heckbert SR, Austin TR, Jensen PN, Chen LY, Post WS, Floyd JS, Soliman EZ, Kronmal RA, Psaty BM. Differences by race/ethnicity in the prevalence of clinically detected and monitor-detected atrial fibrillation. *Circ Arrhythm Electrophysiol*. 2020;13:e007698. doi: [10.1161/CIRCEP.119.007698](https://doi.org/10.1161/CIRCEP.119.007698)

SUPPLEMENTAL MATERIAL

Table S1. Validated algorithm to ascertain obesity from electronic health record

<p>Eligibility: Inclusion Criteria</p>	<p>Adult patients (18 and older) with</p> <ol style="list-style-type: none"> 1) Height and weight measurement at a frequency based upon patient age from structured data field in physical examination section of electronic health record 2) Assessment of body mass index (BMI = Weight / [Height]²) to categorize individuals as follows <ol style="list-style-type: none"> 1. BMI ≥ 30 2. Most recent BMI ≥29 w/ any BMI ≥30 in the past year 3. No BMI data available 3) If no BMI data available, obesity define based upon <ol style="list-style-type: none"> 1. ICD-9/10 diagnosis code or 2. Problem list term
<p>Eligibility: Exclusion Criteria</p>	<ol style="list-style-type: none"> 1) Most recent BMI < 30 2) Most recent BMI ≤ 29 with no BMI ≥ 30 in the past year
<p>Frequency of weight/height measurements</p>	<ol style="list-style-type: none"> 1) Weight measurement at a frequency based upon age <ol style="list-style-type: none"> a) Patients age ≥65, most recent weight in past 2 years b) Patients age 40-64, most recent weight in past 3 years c) Patients age 23-39, most recent weight in past 5 years d) Patients aged 22, most recent weight in past 4 years e) Patients aged 21, most recent weight in past 3 years f) Patients aged 18-20, most recent weight in past 2 years 2) Height measurement at a frequency based upon age <ol style="list-style-type: none"> a) Patients aged ≥28, most recent height in past 10 years b) Patients aged 27, most recent height in past 9 years c) Patients aged 26, most recent height in past 8 years

	<ul style="list-style-type: none"> d) Patients aged 25, most recent height in past 7 years e) Patients aged 24, most recent height in past 6 years f) Patients aged 23, most recent height in past 5 years g) Patients aged 22, most recent height in past 4 years h) Patients aged 21, most recent height in past 3 years i) Patients aged 18-20, most recent height in past 2 years
Data Sources	<ul style="list-style-type: none"> 1) Problem list terms (any prior) 2) ICD 9/10 diagnosis codes in prior 3 years <ul style="list-style-type: none"> a) Hospitalization or any outpatient visit b) Any primary or secondary codes 3) Physical exam / flow sheets
Problem List Terms	Obese; Obesity; Morbid Obesity; Simple Obesity; Body Mass Index 30 + -Obesity
ICD-9 codes	278.00: Obesity, unspecified 278.01: Morbid obesity 278.03: Obesity Hypoventilation syndrome
ICD-10 Codes	E66.01: Morbid (severe) obesity due to excess calories E66.09: Other obesity due to excess calories E66.1: Drug-induced obesity E66.2: Morbid (severe) obesity with alveolar hypoventilation E66.3: Overweight E66.8: Other obesity E66.9: Obesity, unspecified Z68.30: Body mass index (BMI) 30.0-30.9, adult Z68.31: Body mass index (BMI) 31.0-31.9, adult Z68.32: Body mass index (BMI) 32.0-32.9, adult Z68.33: Body mass index (BMI) 33.0-33.9, adult Z68.34: Body mass index (BMI) 34.0-34.9, adult Z68.35: Body mass index (BMI) 35.0-35.9, adult

	Z68.36: Body mass index (BMI) 35.0-35.9, adult Z68.37: Body mass index (BMI) 37.0-37.9, adult Z68.38: Body mass index (BMI) 38.0-38.9, adult Z68.39: Body mass index (BMI) 39.0-39.9, adult Z68.41: Body mass index (BMI) 40.0-44.9, adult Z68.42: Body mass index (BMI) 45.0-49.9, adult Z68.43: Body mass index (BMI) 50-59.9 , adult Z68.44: Body mass index (BMI) 60.0-69.9, adult Z68.45: Body mass index (BMI) 70 or greater, adult
Cut off values for height and weight	1) Max height: 84 inches 2) Min height: 48 inches 3) Max weight: 515 pounds 4) Min weight: 60 pounds
Chart Review Selection Criteria	1) Blinded list of 630 patients (25 positive and 10 negative patients per primary care practice in the network) 2) Review performed by a Research Nurse
Chart Review Results	Sensitivity: 98% Specificity: 97% Positive predictive value: 97% Negative predictive value: 96%

BMI: body mass index

Table S2. ICD-9/10 codes used to define codified predictors not ascertained using validated algorithms

Variable	ICD-9/10 Codes
Myocardial Infarction	410.00, 410.01, 410.02, 410.10, 410.11, 410.12, 410.20, 410.21, 410.22, 410.30, 410.31, 410.32, 410.40, 410.41, 410.42, 410.50, 410.51, 410.52, 410.60, 410.61, 410.62, 410.70, 410.71, 410.72, 410.80, 410.81, 410.82, 410.90, 410.91, 410.92, 412, 429.79, I21.01, I21.02, I21.09, I21.11, I21.19, I21.21, I21.29, I21.3, I21.4, I21.9, I21.A1, I21.A9, I22.0, I22.1, I22.2, I22.8, I22.9, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.7, I23.8, I24.1, I25.2,
Chronic Kidney Disease	250.4, 250.41, 250.42, 250.43, 403, 403.01, 403.1, 403.11, 403.9, 403.91, 404, 404.01, 404.02, 404.03, 404.1, 404.11, 404.12, 404.13, 404.9, 404.91, 404.92, 404.93, 582, 582.1, 582.2, 582.4, 582.81, 582.89, 582.9, 583, 583.1, 583.2, 583.4, 583.6, 583.7, 583.81, 583.89, 583.9, 584.5, 584.6, 584.7, 584.8, 584.9, 585.1, 585.2, 585.3, 586, 587, 588, 588.81, 588.89, 588.9, 753, 753.12, 753.13, 753.14, 753.15, 753.16, 753.17, 753.19, 788.5, 792.5, V42.0, V45.11, V45.12, V56.0, V56.1, V56.2, V56.31, V56.32, V56.8, E08.22, E09.22, E10.22, E11.22, E13.22, I12.0, I12.9, I13.0, I13.1, I13.10, I13.11, I13.2, N18.1, N18.2, N18.3, N18.4, N18.5, N18.9, N19, N99.0, R34, Z49.01, Z49.02, Z49.31, Z49.32, Z99.2
Chronic Kidney Disease - Severe	403.01, 403.11, 403.91, 404.02, 404.03, 404.12, 404.13, 404.92, 404.93, 585.4, 585.5, 585.6, 788.5, 792.5, V42.0, V45.11, V45.12, V56.0, V56.1, V56.2, V56.31, V56.32, V56.8, I12.0, I13.11, I13.2, N18.4, N18.5, N18.6, R34, Z99.2
Hyperlipidemia	272, 272.1, 272.2, 272.3, 272.4, 272.5, 272.6, 272.7, 272.8, 272.9, 759.9, E71.30, E75.21, E75.22, E75.5, E75.6, E77.0, E77.1, E78.0, E78.1, E78.2, E78.3, E78.4, E78.5, E78.6, E78.7, E78.70, E78.79, E78.81, E78.89, E78.9, E88.1, E88.89
Valvular Disease	35.05, 35.12, 35.10, 35.11, 35.14, 35.20, 35.21, 35.06, 35.13, 35.22, 35.24, 35.25, 35.26, 35.27, 35.28, 35.96, 35.23, 394.1, 394.2, 396.3, 396.2, 396.9, 394.9, 396.0, 396.1, 396.8, 394.0, V42.3, V43.3, I05.0, I05.1, I05.2, I05.8, I05.9, I06.8, I06.9, I07.8, I07.9, I08.0, I08.1, I08.3, I08.8, I08.9, I09.1, I34.0, I34.1, I34.2, I34.8, I34.9, I35.0, I35.1, I35.2, I35.8, I35.9, I36.0, I36.1, I36.2, I36.8, I36.9, I37.0, I37.1, I37.2, I37.8, I37.9, I38
Prior Stroke / Transient Ischemic Attack	362.31, 362.32, 362.33, 362.34, 388.02, 430, 431, 432.9, 433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.00, 434.01, 434.10, 434.11, 434.91, 435.0, 435.1, 435.2,

	435.3, 435.8, 435.9, 437.1, 437.7, 437.9, 438.10, 438.11, 438.12, 438.13, 438.14, 438.20, 438.21, 438.22, 438.81, 438.82, 438.83, 438.89, 438.9, 997.02, V12.54, G45.0, G45.1, G45.2, G45.3, G45.4, G45.8, G46.3, G46.4, H34.00, H34.01, H34.02, H34.03, H34.10, H34.11, H34.12, H34.13, H34.211, H34.212, H34.213, H34.219, H34.231, H34.232, H34.233, H34.239, H93.099, I60.9, I61.9, I62.9, I63.00, I63.011, I63.012, I63.019, I63.111, I63.112, I63.119, I63.12, I63.131, I63.132, I63.139, I63.19, I63.20, I63.211, I63.212, I63.219, I63.22, I63.231, I63.232, I63.239, I63.29, I63.30, I63.311, I63.312, I63.319, I63.321, I63.322, I63.329, I63.331, I63.332, I63.339, I63.341, I63.342, I63.349, I63.40, I63.411, I63.412, I63.419, I63.421, I63.422, I63.429, I63.431, I63.432, I63.439, I63.49, I63.50, I63.511, I63.512, I63.519, I63.521, I63.522, I63.529, I63.531, I63.532, I63.539, I63.541, I63.542, I63.549, I63.59, I63.6, I63.8, I63.9, I66.01, I66.02, I66.03, I66.09, I66.11, I66.12, I66.13, I66.19, I66.21, I66.22, I66.23, I66.29, I66.3, I66.8, I66.9, I67.81, I67.82, I67.841, I67.848, I67.89, I67.9, I69.80, I69.81, I69.820, I69.821, I69.822, I69.823, I69.828, I69.831, I69.832, I69.833, I69.834, I69.839, I69.841, I69.842, I69.843, I69.844, I69.849, I69.851, I69.852, I69.853, I69.854, I69.859, I69.861, I69.862, I69.863, I69.864, I69.865, I69.869, I69.890, I69.891, I69.892, I69.893, I69.898, I69.90, I69.91, I69.920, I69.921, I69.922, I69.923, I69.928, I69.931, I69.932, I69.933, I69.934, I69.939, I69.941, I69.942, I69.943, I69.944, I69.949, I69.951, I69.952, I69.953, I69.954, I69.959, I69.961, I69.962, I69.963, I69.964, I69.965, I69.969, I69.990, I69.991, I69.992, I69.993, I69.998, I97.810, I97.811, I97.820, I97.821, Z86.73, G45.9
Systemic Atherosclerosis	441, 441.01, 441.02, 441.03, 441.1, 441.2, 441.3, 441.4, 441.5, 441.6, 441.7, 441.9, I70.0, I71.00, I71.01, I71.02, I71.03, I71.1, I71.2, I71.3, I71.4, I71.5, I71.6, I71.8, I71.9
Cerebral Atherosclerosis	433, 433.01, 433.1, 433.11, 433.2, 433.21, 433.3, 433.31, 433.8, 433.81, 433.91, 434.9, 434.91, 435, 435.1, 435.2, 435.3, 437, 437.1, 438.13, 438.14, G45.0, G45.8, I63.00, I63.011, I63.012, I63.019, I63.111, I63.112, I63.119, I63.12, I63.131, I63.132, I63.139, I63.19, I63.20, I63.211, I63.212, I63.219, I63.22, I63.231, I63.232, I63.239, I63.29, I63.30, I63.311, I63.312, I63.319, I63.321, I63.322, I63.329, I63.331, I63.332, I63.339, I63.341, I63.342, I63.349, I63.40, I63.411, I63.412,

	I63.419, I63.421, I63.422, I63.429, I63.431, I63.432, I63.439, I63.49, I63.50, I63.511, I63.512, I63.519, I63.521, I63.522, I63.529, I63.531, I63.532, I63.539, I63.541, I63.542, I63.549, I63.59, I65.01, I65.02, I65.03, I65.09, I65.1, I65.21, I65.22, I65.23, I65.29, I65.8, I65.9, I66.9, I67.2, I67.81, I67.82, I67.89
Thyrotoxicosis	242.00, 242.01, 242.10, 242.11, 242.20, 242.21, 242.30, 242.31, 242.40, 242.41, 242.80, 242.81, 242.90, 242.91, E05.00, E05.01, E05.10, E05.20, E05.21, E05.80, E05.41, E05.90, E05.30, E05.31, E05.40, E05.81, E05.11, E05.91
Hypothyroidism	243, 244, 244.1, 244.8, 244.9, 245, 245.1, 245.2, 245.9, E03.1, E03.8, E03.9, E06.1, E06.3, E06.5, E06.9, E89.0
Pulmonary Disease	490, 491.0, 491.1, 491.20, 491.21, 491.22, 491.8, 491.9, 492.0, 492.8, 493.00, 493.01, 493.02, 493.10, 493.11, 493.12, 493.20, 493.21, 493.22, 493.81, 493.82, 493.90, 493.91, 493.92, 494.0, 494.1, 495.0, 495.1, 495.2, 495.3, 495.4, 495.5, 495.6, 495.7, 495.8, 495.9, 496, 500, 501, 502, 503, 504, 505, 506.0, 506.1, 506.2, 506.3, 506.4, 506.9, A15.0, A52.72, B38.1, B39.1, B40.1, D86.0, D86.2, E84.0, J40, J41.0, J41.1, J41.8, J42, J43.0, J43.1, J43.2, J43.8, J43.9, J44, J44.0, J44.1, J44.9, J45.20, J45.21, J45.22, J45.30, J45.31, J45.32, J45.40, J45.41, J45.42, J45.50, J45.51, J45.52, J45.901, J45.902, J45.909, J45.990, J45.991, J45.998, J47, J47.0, J47.1, J47.9, J60, J61, J62.0, J62.8, J63.0, J63.1, J63.2, J63.3, J63.4, J63.5, J63.6, J64, J65, J66.0, J66.1, J66.2, J66.8, J67.0, J67.2, J67.4, J67.5, J67.6, J67.7, J67.8, J67.9, J671, J673, J68.0, J68.1, J68.2, J68.3, J68.4, J68.9, J70.1, J70.3, J70.4, J81.8, J82, J84.02, J84.03, J84.10, J84.112, J84.115, J84.17, J84.82, J84.842, J84.89, J84.9, J95.3, J98.2, J98.3, M30.1, M32.13, M34.81, M35.02
Chronic Obstructive Pulmonary Disease	491, 491.0, 491.1, 491.2, 491.20, 491.21, 491.22, 491.8, 491.9, 492, 492.0, 492.8, 493.2, 493.20, 493.21, 493.22, 496, J44, J44.0, J44.1, J44.9, J41, J41.0, J41.1, J41.8, J42, J43, J43.1, J43.2, J43.8, J43.9, J45.5, J45.50, J45.51, J45.52
Congenital Heart Disease	745.0, 745.1, 745.2, 745.4, 745.5, 746.1, 746.2, 746.3, 746.4, 746.5, 746.6, 746.7, 746.81, 746.82, 746.83, 746.85, 746.86, 747.1, 747.11, 747.29, 747.31, 747.49, 745.7, 745.11, 745.3, 745.12, 745.8, 746.9, 745.69, 745.9, 746.01, 746.02, 746.09, 746.00, 746.1, 746.89, 747.6, 747.0, 747.9, 748.5, Q20.0, Q20.3, Q21.3, Q21.0, Q21.1, Q22.4, Q22.5, Q23.0, Q23.1, Q23.2, Q23.3, Q23.4, Q24.4, Q24.2, Q24.3, Q24.5, Q24.6, Q25.1, Q25.2, Q25.3, Q25.4, Q25.8, Q25.9, Q25.5, Q25.6,

	Q25.7, Q26, Q20.8, Q20.1, Q20.2, Q20.4, Q20.5, Q20.6, Q20.8, Q20.9, Q21.2, Q21.4, Q21.8, Q21.9, Q22.0, Q22.1, Q22.2, Q22.3, Q22.6, Q22.8, Q22.9, Q23.8, Q23.9, Q27.4, Q24.8, Q24.9, Q25.0, Q28.9, Q33.2
Cardiomegaly	429.3, I51.7
Mitral Valve Disorder	394, 394.0, 394.1, 394.2, 394.9, 424.0, I34, I34.0, I34.1, I34.2, I34.8, I34.9, I05, I05.0, I05.1, I05.2, I05.8, I05.9
Mitral Stenosis	394.0, 394.2, I05.0, I05.2, I34.2
Mitral Insufficiency	394.1, 394.2, I05.1, I34.0, I05.2
Mitral Valve Prolapse	I34.1, 424.0
Supraventricular Tachycardia	427.0, I47.1
Premature Atrial Contractions	427.61, I49.1
Alcohol Use – Heavy	303, 303.0, 303.00, 303.01, 303.02, 303.9, 303.90, 303.91, 303.92, 305, 305.00, 305.01, 305.02, F10.1, F10.10, F10.12, F10.120, F10.121, F10.129, F10.14, F10.15, F10.150, F10.151, F10.159, F10.18, F10.180, F10.181, F10.182, F10.188, F10.19, F10.2, F10.20, F10.22, F10.220, F10.221, F10.229, F10.23, F10.230, F10.231, F10.232, F10.239, F10.24, F10.25, F10.250, F10.251, F10.259, F10.26, F10.27, F10.28, F10.280, F10.281, F10.282, F10.288, F10.9, F10.92, F10.920, F10.921, F10.929, F10.94, F10.95, F10.950, F10.951, F10.959, F10.96, F10.97, F10.98, F10.980, F10.981, F10.982, F10.988, F10.99
Pericarditis	036.41, 074.21, 093.81, 420, 098.83, 420.9, 420.99, 423.1, 423.2, 393, 420.91, 420.9, 115.93, 391.0, 115.03, 115.13, 115.93, 423.0, 423.3, 423.8, 423.9, 420.0, A39.53, B33.23, I01.0, I09.2, I30.0, I30.1, I30.8, I30.9, I31.0, I31.1, I31.2, I31.3, I31.4, I31.8, I31.9, I32
Myocarditis	42.9, 032.82, 036.43, 074.23, 093.82, 13.03, 391.2, 398.0, 422.0, 422.90, 422.91, 422.92, 422.93, 422.99, I40.0, I40.1, I40.8, I40.9, I41, I51.4, J10.82, J11.82, A38.1, A39.52, B26.82, B33.22, B58.81, D86.85, I01.2, I09.0
Obstructive Sleep Apnea	780.57, 327.23, G47.30, G47.33, G47.39
Prior Cardiac Surgery	35.00, 35.01, 35.02, 35.03, 35.04, 35.11, 35.22, 35.31, 35.32, 35.33, 35.34, 35.35, 35.39, 35.41, 35.42, 35.5, 35.51, 35.52, 35.53, 35.54, 35.55, 35.6, 35.61, 35.62, 35.63, 35.7, 35.71, 35.72, 35.73, 35.8, 36.03, 36.04, 36.07, 36.09, 36.10, 36.1, 36.11, 36.12, 36.13, 36.14, 36.15, 36.16, 36.17, 36.19, 36.2, 37.0, 37.1, 37.11, 37.12, 37.2, 37.21, 37.22, 37.23, 37.24, 37.25, 37.26, 37.27, 37.28, 37.29, 37.3, 37.31, 37.32, 37.33, 37.34, 37.35, 37.36, 37.37, 37.4, 37.41, 37.49, 37.7, 37.8, 37.9, I97.190, Z98.61, I25.700, I25.701, I25.708, I25.709, I25.710, I25.711, I25.718, I25.719, I25.720, I25.721,

	I25.728, I25.729, I25.730, I25.731, I25.738, I25.739, I25.790, I25.791, I25.798, I25.799, I25.810, I25.811, I25.812, I97.0, I97.110, I97.120, I97.130, V45.82, Z45.018, Z45.02, Z45.09, Z48.21, Z48.280, Z95.1, Z95.2, Z95.811, Z95.812
Hypertrophic Cardiomyopathy	425.1, 425.11, 425.18, I42.1, I42.2
Other Cardiomyopathy	425, 425.0, 425.2, 425.3, 425.4, 425.5, 425.7, 425.8, 425.9, I42, I42.0, I42.3, I42.4, I42.5, I42.6, I42.7, I42.8, I42.9
Chronic Liver Disease	070.0, 070.20, 070.21, 070.22, 070.23, 070.30, 070.31, 070.32, 070.33, 070.41, 070.42, 070.43, 070.49, 070.51, 070.52, 070.53, 070.54, 070.59, 070.6, 070.70, 070.71, 070.9, 273.4, 275.01, 275.1, 453.0, 570, 571.1, 571.3, 571.40, 571.41, 571.49, 571.8, 571.9, 573.0, 573.1, 573.2, 573.3, 573.4, 573.8, 573.9, 576.1, B15.0, B15.9, B16.0, B16.1, B16.2, B16.9, B17.0, B17.10, B17.11, B17.2, B17.8, B17.9, B18.0, B18.1, B18.2, B18.8, B18.9, B19.0, B19.10, B19.11, B19.20, B19.21, B19.9, B94.2, K70.0, K70.10, K70.11, K70.2, K70.30, K70.31, K70.9, K71.0, K71.10, K71.11, K71.2, K71.3, K71.4, K71.50, K71.51, K71.6, K71.7, K73.0, K73.1, K73.2, K73.8, K73.9, K74.0, K74.1, K75.3, K75.4, K75.8, K75.89, K75.9, K76.89, Z22.50, Z22.51, Z22.52, Z22.59, K72.00, K72.01, K72.1, K72.10, K72.22, K72.90, K72.91
Cirrhosis	571.2, 571.5, 571.6, K70.40, K70.41, K71.9, K74.2, K74.3, K74.5, K74.60, K74.69, K744, K7460, K75.0, K75.2
Liver Complications	456.0, 456.20, 456.1, 456.21, 572.2, 572.3, 572.4, 572.8, 573.5, 567.0, 567.23, 567.21, 567.29, 567.1, 567.89, 567.9, 789.5, 789.59, K76.0, K76.1, K76.2, K76.3, I85.01, I85.11, I85, K76.6, K76.7, K76.81, K76.80, K76.8, K65.0, K65.2, K65.8, K65.9, R18, R18.0, R18.8

ICD: International Classification of Disease

Table S3. CHARGE-AF score components and weights*

Covariate	Estimated β (SE)
Age (per 5-year increase)	0.508 (0.022)
Race (white)	0.465 (0.093)
Height (per 10 cm increase), cm	0.248 (0.036)
Weight (per 15 kg increase), kg	0.115 (0.033)
Systolic blood pressure (per 20 mmHg increase), mmHg	0.197 (0.033)
Diastolic blood pressure (per 10 mmHg increase), mmHg	-0.101 (0.032)
Current smoker	0.359 (0.091)
Anti-hypertensive medication use	0.349 (0.063)
Diabetes	0.237 (0.073)
Heart failure	0.701 (0.106)
Myocardial infarction	0.496 (0.089)

CHARGE-AF: Cohorts for Heart and Aging Research in Genomic Epidemiology Atrial Fibrillation; SE: standard error

* Alonso A, Krijthe BP, Aspelund T, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *J Am Heart Assoc.* 2013;2(2):e000102.

Table S4. Concept unique identifiers (CUIs) mapped to potential predictors from linkage to the Unified Medical Language System (UMLS)*

Variable	CUI
Alcohol Abuse	C0085762
Alcohol Abuse	C0001973
Alcohol Abuse	C0560219
Cardiomegaly	C0018800
Cerebral Atherosclerosis	C4024924
Cerebrovascular disease	C0007775
Cerebrovascular disease	C0007820
Chronic Kidney Disease - Severe	C2316810
Chronic Kidney Disease	C1561643
Chronic Liver Disease	C0341439
Chronic Liver Disease	C0085605
Chronic Obstructive Pulmonary Disease	C0024117
Chronic Obstructive Pulmonary Disease	C0034067
Chronic Obstructive Pulmonary Disease	C0008677
Cirrhosis	C0023890
Congenital Heart Disease	C0152021
Congestive heart failure	C0018802
Coronary artery disease	C0010054
Coronary artery disease	C0010068
Diabetes	C0011849
Diabetes	C0011860
Diabetes	C0011854
Hyperlipidemia	C0020473
Hypertension	C0020538
Hypertrophic Cardiomyopathy	C0007194
Hypertrophic Cardiomyopathy	C4551472
Hypothyroidism	C0020676
Left atrial enlargement	C0232309
Left atrial enlargement	C0232310
Left atrial enlargement	C0238705
Left Ventricular Hypertrophy	C0149721
Left Ventricular Hypertrophy	C0232306
Left Ventricular Hypertrophy	C0344398
Liver Complications	C0015695
Liver Complications	C0267821
Liver Complications	C0014867
Liver Complications	C0019151
Liver Complications	C0020541
Liver Complications	C0019212
Liver Complications	C0600452

Mitral Insufficiency	C0026266
Mitral Stenosis	C0026269
Mitral Valve Disease	C0026265
Mitral Valve Prolapse	C0026267
Myocardial Infarction	C0027051
Myocarditis	C0027059
Obesity	C0028754
Obesity	C0028756
Obstructive Sleep Apnea	C0520679
Other Cardiomyopathy	C0878544
Other Cardiomyopathy	C0007193
Other Cardiomyopathy	C0007192
Other Cardiomyopathy	C0264834
Pericarditis	C0031046
Peripheral vascular disease	C0085096
Premature atrial contractions	C0033036
Prolonged PR Interval	C0600125
Pulmonary Disease	C0024115
Shortened PR Interval	C0520878
Stroke	C0038454
Supraventricular Tachycardia	C0039240
Supraventricular Tachycardia	C1963244
Supraventricular Tachycardia	C3815188
Supraventricular Tachycardia	C0030590
Systemic Atherosclerosis	C0155733
Thyrotoxicosis	C0040156
Thyrotoxicosis	C0020550
Transient Ischemic Attack	C0007787
Valvular Disease	C3258293
Valvular Disease	C0018824

CUI: concept unique identifier; UMLS: Unified Medical Language System

* Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32:D267-270.

Table S5. C-statistics and hazard ratios in prediction models that exclude race and include social determinants of health

	C-statistic (95% CI)	Hazard Ratio (95% CI)
Development (Add insurance)		
Codified+NLP	0.742 (0.733-0.751)	
≥ 84 th percentile		15.91 (12.88-19.64)
50-<84 th percentile		5.74 (4.65-7.10)
16-<50 th percentile		2.32 (1.86-2.90)
< 16 th percentile		-
Codified-only	0.728 (0.719-0.737)	
≥ 84 th percentile		12.80 (10.55-15.52)
50-<84 th percentile		4.64 (3.82-5.63)
16-<50 th percentile		2.17 (1.77-2.66)
< 16 th percentile		-
Development (Add insurance + Income)*		
Codified+NLP	0.741 (0.732-0.750)	
≥ 84 th percentile		15.44 (12.52-19.04)
50-<84 th percentile		5.81 (4.71-7.17)
16-<50 th percentile		2.29 (1.83-2.86)
< 16 th percentile		-
Codified-only	0.727 (0.717-0.736)	
≥ 84 th percentile		12.59 (10.33-15.36)
50-<84 th percentile		4.68 (3.84-5.72)
16-<50 th percentile		2.13 (1.73-2.63)
< 16 th percentile		-
Internal Validation (Add insurance)		
Codified+NLP	0.731 (0.716-0.746)	
≥ 84 th percentile		12.91 (9.42-17.71)
50-<84 th percentile		4.38 (3.18-6.02)
16-<50 th percentile		2.00 (1.43-2.80)
< 16 th percentile		-
Codified-only	0.722 (0.707-0.737)	
≥ 84 th percentile		12.02 (8.76-16.50)
50-<84 th percentile		4.74 (3.45-6.52)
16-<50 th percentile		1.96 (1.40-2.74)
< 16 th percentile		-
Internal Validation (Add insurance + Income)*		
Codified+NLP	0.731 (0.716-0.746)	
≥ 84 th percentile		12.72 (9.27-17.46)
50-<84 th percentile		4.26 (3.09-5.86)

16-<50 th percentile		1.95 (1.39-2.73)
< 16 th percentile		-
Codified-only	0.722 (0.707-0.738)	
≥ 84 th percentile		11.36 (8.27-15.60)
50-<84 th percentile		4.55 (3.31-6.25)
16-<50 th percentile		1.99 (1.42-2.78)
< 16 th percentile		-

NLP: natural language processing; CI: confidence interval

* Income represents proportion of population by zip code with income < \$50,000 from 2008-2012 ascertained from Melendez, Robert, Clarke, Philippa, Khan, Anam, Gomez-Lopez, Iris, Li, Mao, and Chenoweth, Megan. National Neighborhood Data Archive (NaNDA): Socioeconomic Status and Demographic Characteristics of ZIP Code Tabulation Areas, United States, 2008-2017. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-07-30. <https://doi.org/10.3886/E120462V1>. Missing income data in development population: n=1,498; Missing income data in internal validation population: n=507

Table S6. Hazard ratios and 95% confidence intervals for incidence of AF by risk groups defined by the 16th, 50th, and 84th percentiles for each model in the external validation cohort

	Codified+NLP	Codified-only	CHARGE-AF
	HR (95% CI)	HR (95% CI)	HR (95% CI)
≥ 84 th percentile	17.74 (13.59-23.16)	13.26 (10.47-16.78)	14.60 (11.37-18.77)
50-<84 th percentile	5.67 (4.34-7.43)	4.26 (3.36-5.41)	5.24 (4.07-6.74)
16-<50 th percentile	2.41 (1.82-3.20)	2.01 (1.56-2.58)	2.23 (1.71-2.90)
< 16 th percentile	-	-	-

AF: atrial fibrillation; NLP: natural language processing; CI: confidence interval

Table S7. Percentile-based net reclassification improvement (NRI) with groups determined by 16th, 50th, 84th percentile of each model in external validation cohort

	Overall NRI (95% CI)	Event NRI (95% CI)	Non-event NRI (95% CI)
Codified+NLP vs. CHARGE-AF	0.044 (0.018 - 0.069)	0.040 (0.016 – 0.064)	0.003 (-0.003 – 0.010)
Codified-only vs. CHARGE-AF	0.001 (-0.025 - 0.026)	-0.003 (-0.026 – 0.021)	0.004 (-0.002 – 0.011)
Codified+NLP vs. Codified-only	0.051 (0.027 - 0.072)	0.055 (0.033 – 0.075)	-0.004 (-0.010 – 0.002)

NRI: net reclassification improvement; NLP: natural language processing; CI: confidence interval; CHARGE-AF: Cohorts for Heart and Aging Research in Genomic Epidemiology Atrial Fibrillation

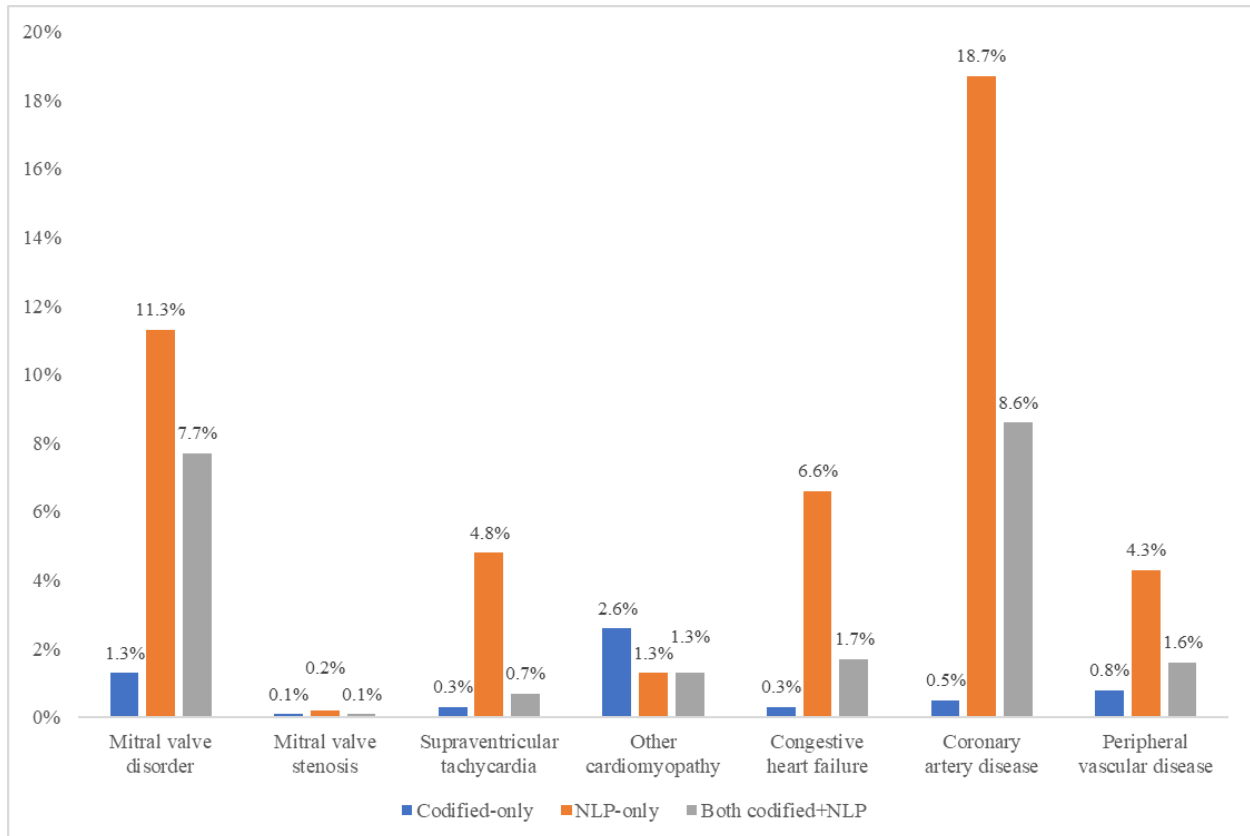
Table S8. C-statistics and hazard ratios in external validation cohort over 3-years of follow-up

	C-statistic (95% CI)	Hazard Ratio (95% CI)
Codified+NLP	0.758 (0.746-0.769)*	
≥ 84 th percentile		19.36 (14.29-26.21)
50-<84 th percentile		5.93 (4.37)
16-<50 th percentile		2.39 (1.73-3.30)
< 16 th percentile		-
Codified-only	0.745 (0.733-0.757)	
≥ 84 th percentile		13.22 (10.15-17.23)
50-<84 th percentile		4.35 (3.33-5.69)
16-<50 th percentile		1.79 (1.34-2.38)
< 16 th percentile		-
CHARGE-AF	0.741 (0.730-0.753)	
≥ 84 th percentile		16.89 (12.51-22.82)
50-<84 th percentile		6.00 (4.43-8.12)
16-<50 th percentile		2.44 (1.78-3.36)
< 16 th percentile		-

CI: confidence interval, NLP: natural language processing; CHARGE-AF: Cohorts for Heart and Aging Research in Genomic Epidemiology Atrial Fibrillation

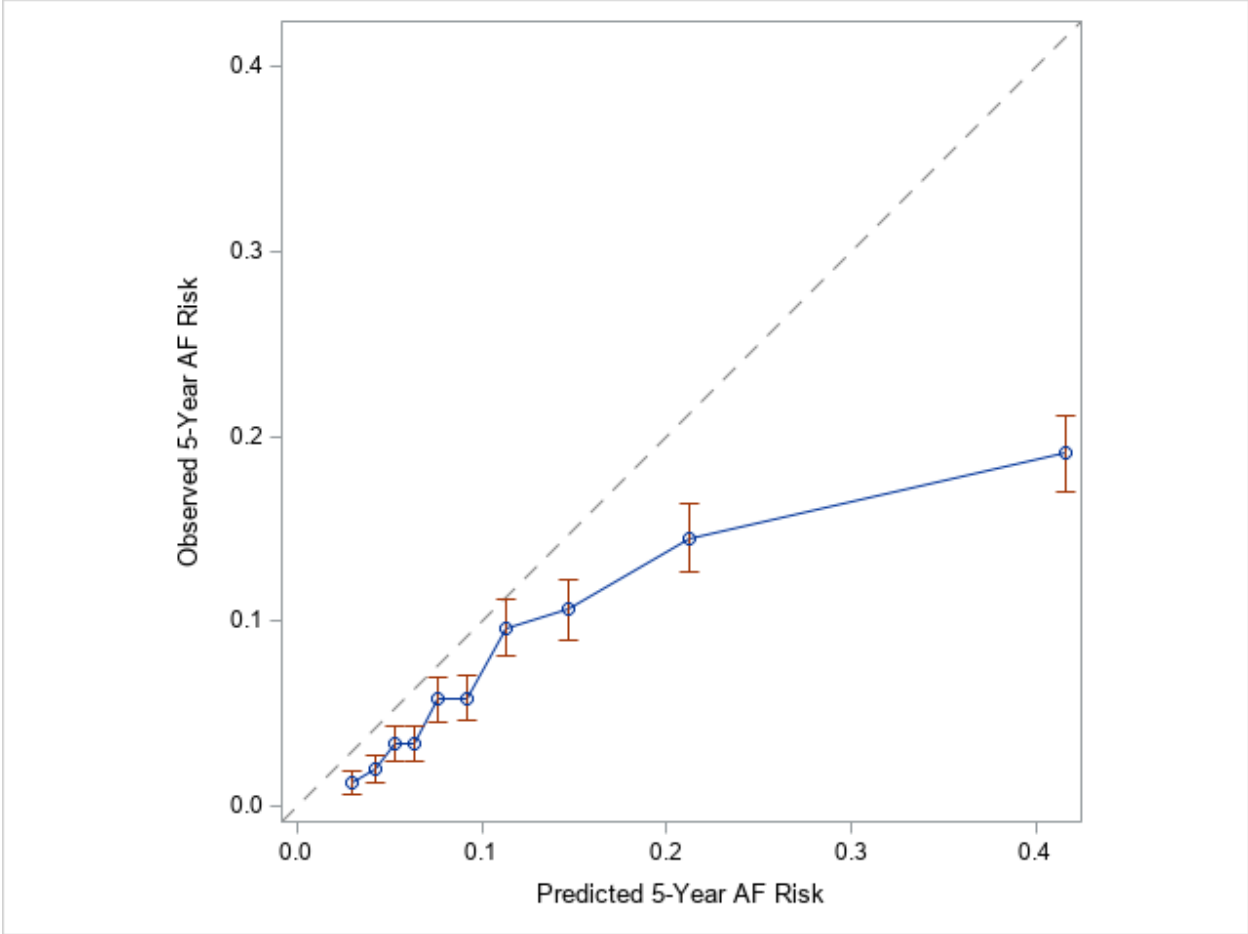
* p<0.001 comparing C-statistic in Codified+NLP compared to codified-only and CHARGE-AF

Figure S1. Prevalence of features identified by codified data only, NLP data only, and by both codified and NLP data among those where both the codified and NLP version were selected for inclusion in the codified+NLP model in the development cohort



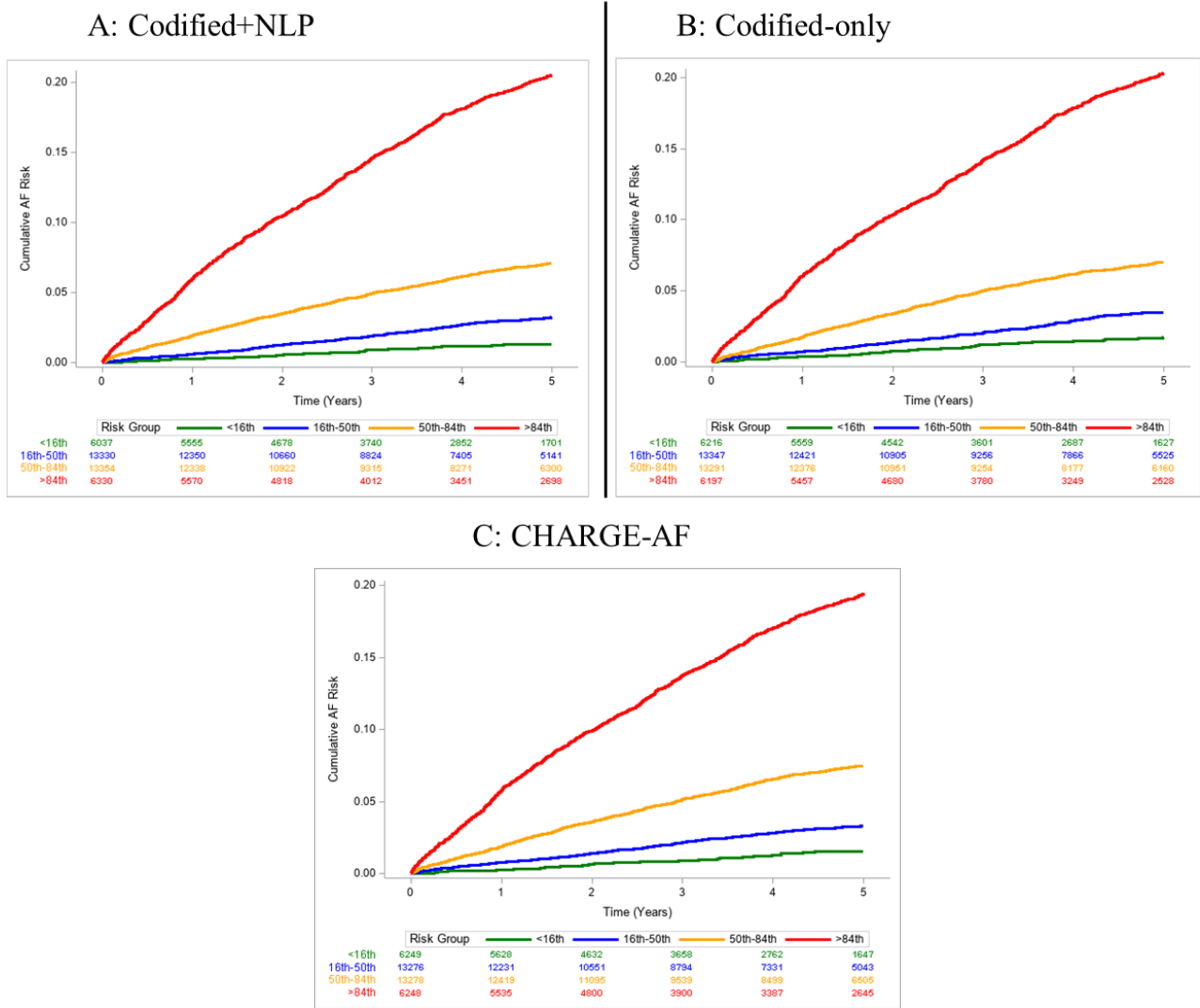
NLP: natural language processing

Figure S2. Plot of observed 5-year AF risk versus predicted 5-year AF risk for recalibrated CHARGE-AF with patients divided into risk groups based on deciles in the internal validation cohort



AF: atrial fibrillation; CHARGE-AF: Cohorts for Heart and Aging Research in Genomic Epidemiology Atrial Fibrillation

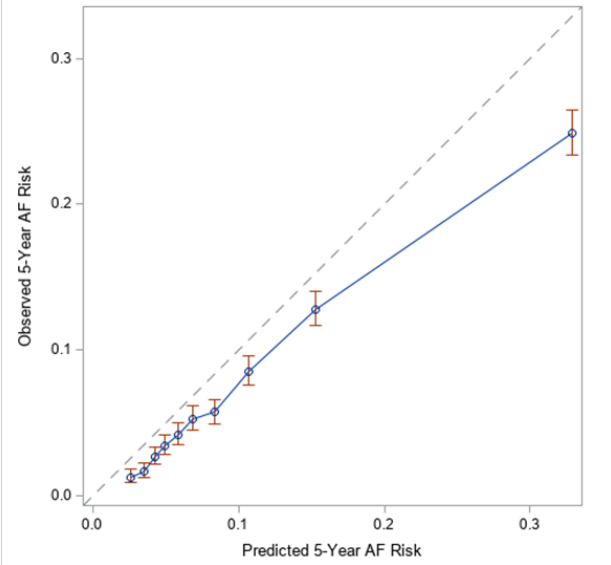
Figure S3. Cumulative incidence plots stratified by groups of predicted risk for codified + NLP, codified-only, and CHARGE-AF models in external validation cohort



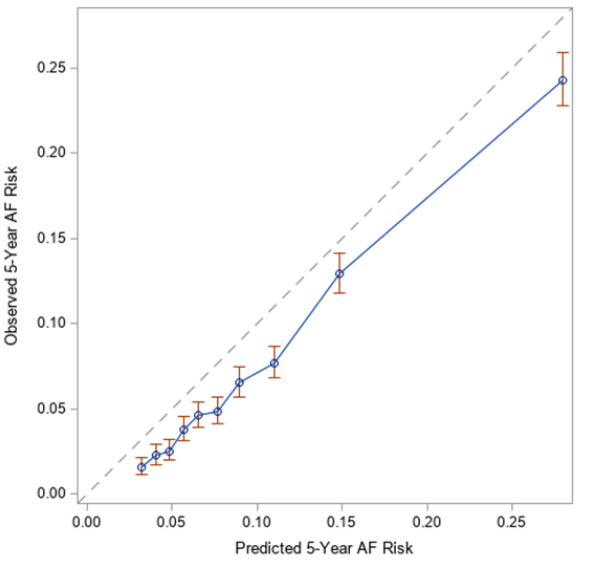
NLP: natural language processing; AF: atrial fibrillation; CHARGE-AF: Cohorts for Heart and Aging Research in Genomic Epidemiology Atrial Fibrillation

Figure S4. Plots of observed 5-year AF risk versus predicted 5-year AF risk with patients divided into risk groups based on deciles in external validation cohort

A: Codified+NLP

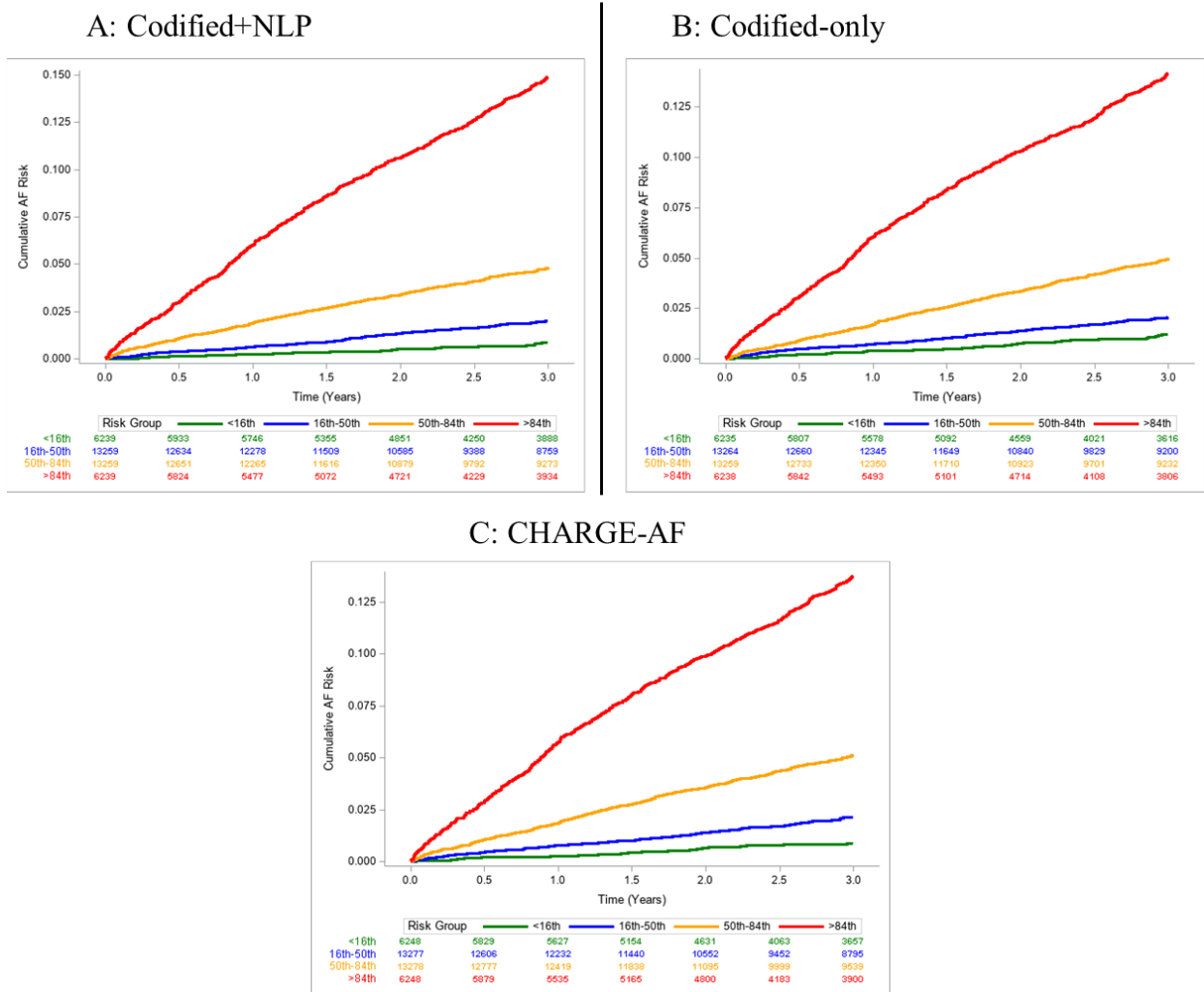


B: Codified-only



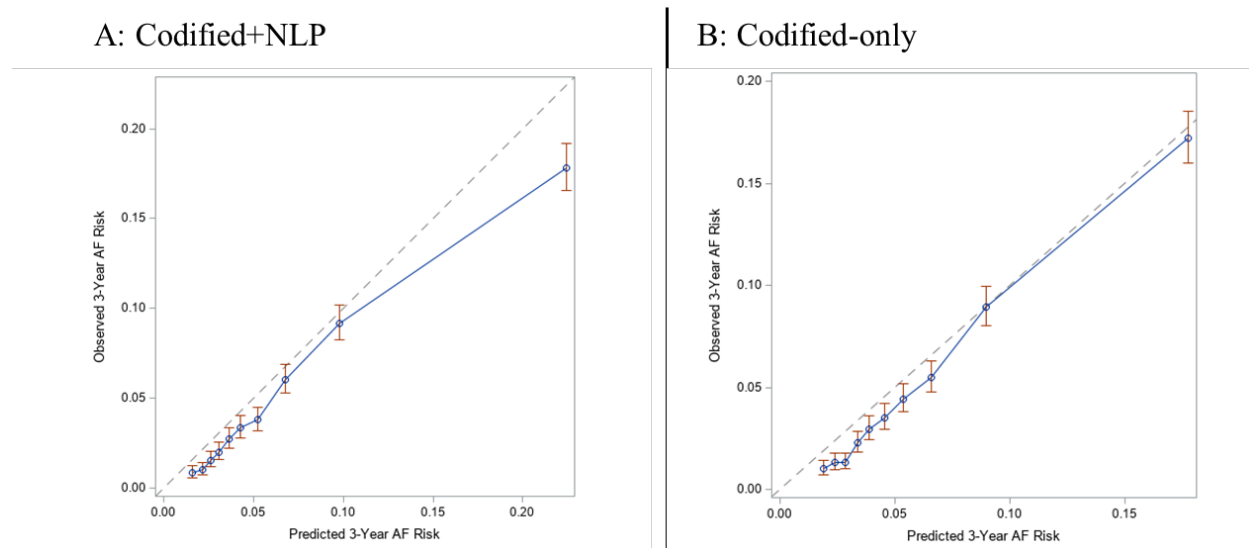
AF: atrial fibrillation; NLP: natural language processing

Figure S5. Cumulative incidence plots stratified by groups of predicted risk for codified + NLP, codified-only, and CHARGE-AF models in external validation cohort over 3-years of follow-up



NLP: natural language processing; CHARGE-AF: Cohorts for Heart and Aging Research in Genomic Epidemiology Atrial Fibrillation; AF: atrial fibrillation

Figure S6. Plots of observed 3-year AF risk versus predicted 3-year AF risk with patients divided into risk groups based on deciles in external validation cohort



AF: atrial fibrillation; NLP: natural language processing