ORIGINAL RESEARCH

Impact of Different Electronic Cohort Definitions to Identify Patients With Atrial Fibrillation From the Electronic Medical Record

Rashmee U. Shah, MD, MS; Rebeka Mukherjee, MS; Yue Zhang, PhD; Aubrey E. Jones, PharmD; Jennifer Springer, MD; Ian Hackett, MD; Benjamin A. Steinberg, MD, MHS; Donald M. Lloyd-Jones, MD, ScM; Wendy W. Chapman, PhD

***Background*—**Electronic medical records (EMRs) allow identification of disease-specific patient populations, but varying electronic cohort definitions could result in different populations. We compared the characteristics of an electronic medical record–derived atrial fibrillation (AF) patient population using 5 different electronic cohort definitions.

***Methods and Results*—**Adult patients with at least 1 AF billing code from January 1, 2010, to December 31, 2017, were included. Based on different electronic cohort definitions, we trained 5 different logistic regression models using a labeled training data set (n=786). Each model yielded a predicted probability; patients were classified as having AF if the probability was higher than a specified cut point. Test characteristics were calculated for each model. These models were then applied to the full cohort and resulting characteristics were compared. In the training set, the comprehensive model (including demographics, billing codes, and natural language processing results) performed best, with an area under the curve of 0.89, sensitivity of 0.90, and specificity of 0.87. Among a candidate population (n=22 000), the proportion of patients identified as having AF varied from 61% in the model using diagnosis or procedure *International Classification of Diseases* (*ICD*) billing codes to 83% in the model using natural language processing of clinical notes. Among identified AF patients, the proportion of patients with a CHA_2DS_2-VASc score ≥2 varied from 69% to 85%; oral anticoagulant treatment rates varied from 50% to 66% depending on the model.

***Conclusions*—**Different electronic cohort definitions result in substantially different AF study samples. This difference threatens the quality and reproducibility of electronic medical record–based research and quality initiatives. **(*J Am Heart Assoc*. 2020;9:e014527. DOI: 10.1161/JAHA.119.014527.)**

Key Words: atrial fibrillation • electronic health records • health services research • informatics • quality of care

Electronic medical records (EMR) are increasingly prevalent, resulting in an explosion of electronic health data available for research and quality initiatives. These data allow healthcare systems to capture large patient populations in order to study diagnoses, treatments, and outcomes. Specifically, atrial fibrillation (AF) is a common condition, and large patient cohort studies may allow health systems and researchers to monitor quality and outcomes. Using the EMR, a health system could monitor the number of AF patients treated with an oral anticoagulant (OAC) for quality improvement efforts, such as increasing appropriate treatment rates among eligible patients. However, no common method exists to identify patients for inclusion in EMR-based initiatives. Different approaches could result in different patient cohorts with respect to characteristics and apparent outcomes, and this would limit the potential of EMR-based initiatives.

Prior studies used varying methods to identify AF patients.[1–4] Medicare studies, for example, include patients who have at least 1 inpatient or 2 outpatient *International Classification of Diseases* (*ICD*) codes for AF.[2] Reports from Kaiser Permanente also include ECG results in the patient-selection process.[4] Other studies have included EMR data such as ablation codes or antiarrhythmic medication treatment, but these studies report only positive predictive value and include outcomes (eg, anticoagulation use) in the prediction model.[3] Although data collected for billing purposes, including *ICD* codes, follow a

From the Division of Cardiovascular Medicine, Department of Internal Medicine (R.U.S., R.M., J.S., B.A.S.), Division of Epidemiology, Department of Internal Medicine (Y.Z.), and Department of Population Health (A.E.J.), University of Utah School of Medicine, Salt Lake City, UT; Section of Cardiology, Department of Internal Medicine, University of Chicago Pritzker School of Medicine, Chicago, IL (I.H.); Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL (D.M.L.-J.); Centre for Digital Transformation of Health, Victorian Comprehensive Cancer Centre, Melbourne, Australia (W.W.C.).

Accompanying Tables S1 through S3 are available at https://www.ahajournals.org/doi/suppl/10.1161/JAHA.119.014527

Correspondence to: Rashmee U. Shah, MD, MS, 30 N. 1900 E, Room 4A100, Salt Lake City, UT 84132. E-mail: rashmee.shah@utah.edu

Received September 9, 2019; accepted January 7, 2020.

Clinical Perspective

What Is New?

- We evaluated 5 different electronic definitions for identifying atrial fibrillation patients from the medical record.
- The characteristics and outcomes of the population differ substantially between different definitions.
- Of 22 000 possible patients, the number of included atrial fibrillation patients varied by up 6690 patients, and the apparent oral anticoagulant treatment rate in patients with CHA_2DS_2-VASc ≥2 varied from half (49.6%) to two-thirds (66.3%) of the population, depending on the electronic cohort definition.

What Are the Clinical Implications?

- Quality improvement, learning healthcare systems, and real-world evidence require electronic cohort definitions to identify disease-specific patient populations.
- Different definitions will result in different populations and different apparent outcome rates.
- Validated, consistent electronic definitions are needed to ensure reproducibility and accuracy of studies that rely on electronic medical records.

controlled vocabulary, this approach may be inaccurate[5,6] and does not exploit other types of data, such as demographic information and non-AF diagnoses.

As an alternative to structured data, clinical notes are an untapped resource for detailed clinical information. Notes often include narrative references to patient conditions, such as “Patient was diagnosed with afib last year.” Text mining with natural language processing (NLP) leverages the unstructured narrative from routine care and is another option for identifying patient cohorts. An advantage of NLP is that the clinical narrative may be less prone to some types of variation seen with billing codes,[6,7] which could support more precise patient selection and portability between institutions. The overall goal of this study was (1) to develop and train various models using different electronic-cohort definitions to identify AF patients from the EMR, incorporating structured and unstructured data; (2) to compare the resulting patient samples and characteristics from each model; and (3) to compare apparent OAC treatment rates in each sample.

Methods

We developed and compared the performance of 5 cohort definitions to identify AF patients from the EMR:

1. Outpatient and inpatient *ICD* AF diagnosis billing codes (Medicare methodology)[2]: ≥1 inpatient billing code or ≥2 outpatient billing codes within 365 days
2. Outpatient AF diagnosis billing codes and ECG (Kaiser methodology)[4]: >1 outpatient diagnosis billing code, 1 outpatient diagnosis billing code and ECG consistent with AF
3. Demographics and *ICD* AF diagnosis billing codes: logistic regression model using patient demographics, presence of an inpatient AF diagnosis billing code, presence of an AF diagnosis billing code in the first position (primary), number of outpatient AF diagnoses billing codes, comorbid conditions and procedures from *ICD* codes, year-of-index-AF diagnosis billing code
4. NLP: at least 1 nonnegated mention of AF in the clinical text (negated AF mentions use phrases such as “patient denies AF,” whereas nonnegated references use phrases such as “Holter monitor showed AF”)
5. Comprehensive: comprehensive logistic regression model combining patient demographics, presence of an inpatient AF diagnosis billing code, presence of a primary AF diagnosis billing code, number of outpatient AF diagnoses billing codes, comorbid conditions and procedures from *ICD* codes, year of index AF diagnosis billing code, at least 1 nonnegated mention of AF in clinical text, ECG with reference to AF, Current Procedural Terminology (CPT) codes for ablation or cardioversion

ICD-9 codes were used through September, 2015 and *ICD-10* codes were used from October, 2015 onward.

Population and Reference Standard

We used data from the Enterprise Data Warehouse (EDW) from University of Utah Health for this study. Enterprise data warehouses are storage systems that integrate numerous data sources within an organization (eg, inpatient and outpatient facilities, radiology reporting, or laboratory result systems) into a central repository.[8] Our health system uses an internally developed EDW (as opposed to a third-party data warehousing solution). The candidate population included patients with at least 1 *ICD-9* or *ICD-10* code for AF between 2010 and 2017 (427.31, I48.0, I48.1, I48.2, I48.9, I48.91), and without an AF diagnosis from January 1, 2007, to December 31, 2009. For model development and training, 786 patients were randomly selected from the candidate population. Chart review by a team of 5 clinicians was used to classify each patient as *AF present* or *AF absent*, which served as the reference standard. This reference standard served as the outcome for all 5 models. Each patient was classified as having AF (1) if AF was referenced in a problem list or past medical history, (2) if AF was documented but appeared only as a transient event, as part of other acute conditions (eg, cardiac surgery or sepsis), or (3) if clinic notes described active AF management. Examples of active management include procedures or medications (eg, cardioversion, anticoagulation), outside records or procedures, or listing in the assessment and plan. Otherwise, the patient was classified as not having AF. At least 2 clinicians reviewed each patient. In case of disagreement, a third reviewer adjudicated the classification. If uncertainty was still present, the team discussed the case to arrive at consensus.

Feature Specification

The feature specifications for the different models are provided in Table S1. Model features included demographics, comorbid conditions, procedures, ECG findings, and text-derived features. Each of the 5 models included some combination of these features, and all features were extracted from structured data fields in the EDW (except for the text-derived features). Table S1 specifies which features were included in which models. Briefly, demographic features included age, sex, race, Hispanic ethnicity, and primary insurance at the time of the index AF diagnosis. Comorbid conditions and procedures were identified based on the presence of an *ICD* diagnostic or procedure billing code any time during the study period. Codes were grouped into clinically meaningful groups according to the Clinical Classification Software (CCS) for the US Agency for Healthcare Research and Quality.[9] We used CPT codes to identify patients who had cardioversions (92960) or ablations (93651, 93655, 93656, 93657) at any time during the study period. For ECGs, we used the text interpretations and a simple regular-expression matching approach. If "atrial fibrillation," "afib," or "a fib" were present, the ECG was classified as positive.

Model Training

In the 786-patient training set, we trained 5 different logistic regression models using the definitions predict the presence of AF. The models yielded predicted probabilities of AF for each training case, and the optimal cut point for each model was identified using Liu's[10] method, which maximizes the product of the sensitivity and specificity. In other words, each model had its own cut point. If the predicted probability was higher than the cut point, the case was classified as AF present. Accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were calculated for each model, compared with the reference standard. Accuracy was defined as the number of correctly classified patients over the total number of patients. We generated and compared the area under the receiver operating characteristic curve for each model to the reference standard using Stata's "roccomp" command.

We used a rules-based NLP approach based on the pyConText algorithm, a freely available Python software package.[11–13] Using the training data and clinical expertise, we identified AF-specific target terms and relevant modifiers that allow classification of each AF mention as present or absent (Table S2).[14] For each patient, each note was analyzed for AF-specific mentions, and each mention was classified as AF present or absent based on the modifiers surrounding the AF target term. If 1 nonnegated mention was present in any note (eg, "Patient has had long standing AF for the past 10 years"), that patient was classified as AF present. In addition, we created a summary variable for each patient, counting the total number of times an AF-specific target term appeared in the notes, regardless of negation.

Application to Full Candidate Population

The candidate population included patients with at least 1 *ICD* billing diagnosis code for AF, seen between 2010 and 2017, excluding those with an AF diagnosis code going back to 2007. In other words, patients with an AF billing code in 2008 and again in 2011 would be excluded. Model training resulted in coefficients for each term in the models, which were then applied to the full candidate population. Patients with a predicted probability of AF higher than the cut point specified during training were classified as having AF. We evaluated the number of patients identified as AF per model and the apparent OAC treatment rates according to model. Patients were classified as treated with an OAC if they had an order for an OAC, including warfarin, dabigatran, apixaban, rivaroxaban, and edoxaban, in the EMR. The OAC treatment rate was calculated for patients with a CHA_2DS_2-VASc score ≥ 2.[15,16] The CHA_2DS_2-VASc score is automatically calculated in our EDW using all available prior diagnosis codes as well as ejection fraction from echocardiogram for classifying heart failure (R.U.S., unpublished data, 2019).

To compare patient characteristics between the different models, we created regression models with each characteristic as the outcome and each model as the predictor variables. In addition, we added a predictor variable that indicates whether all models agreed on whether a patient was included in the final cohort. In other words, the indicator variable equals 1 if all 5 models resulted in patients being included or excluded from the AF cohort. When the indicator variable was 1, all other predictors were reassigned to 0. Therefore, the indicator variable serves as a reference variable, or dummy variable. The likelihood ratio test was used to compare this model with a nested reduced model limited to the *y*-intercept. Thus, a significant *P* value with the likelihood ratio test indicates that at least 1 model differed in terms of inclusion of patients with a given outcome (eg, characteristic, in this case).

This study was approved by the institutional review board at the University of Utah, with a waiver of consent for patient

Table 1. Characteristics of the Model Training Population (n=786), According to the Presence or Absence of AF

Characteristic	AF Present (n=632)	AF Absent (n=154)	*P* Value
Age, y, mean (SD)	69.0 (14.2)	61.3 (17.9)	<0.01
Female sex	249 (39.4)	81 (52.6)	<0.01
White race	563 (89.1)	132 (85.7)	0.24
Medicare insured	411 (65.0)	81 (52.6)	<0.01
No. of outpatient AF diagnoses, mean (SD)	10.1 (21.6)	1.2 (1.7)	<0.01
Primary AF diagnosis†	404 (63.9)	99 (64.3)	0.93
Comorbid conditions‡			
Acute myocardial infarction	80 (12.7)	13 (8.4)	0.15
Coronary artery disease	302 (47.8)	53 (34.4)	<0.01
Valvular heart disease	216 (34.3)	47 (30.5)	0.39
Congestive heart failure	222 (35.1)	32 (20.8)	<0.01
Cerebrovascular disease	156 (24.7)	60 (39.0)	<0.01
Dementia	183 (29.0)	39 (25.3)	0.37
Liver disease	132 (20.9)	36 (23.4)	0.50
Diabetes mellitus	289 (45.7)	62 (40.3)	0.22
Acute renal failure	161 (25.5)	25 (16.2)	0.02
Chronic kidney disease	165 (26.1)	19 (12.3)	<0.01
Pulmonary heart disease	158 (25.0)	25 (16.2)	0.02
Hypertension	465 (73.6)	103 (66.9)	0.10
Thyroid disease	190 (30.1)	29 (18.8)	<0.01
Anemia	219 (34.7)	36 (23.4)	<0.01
Cancer	235 (37.2)	33 (21.4)	<0.01
Procedures, *ICD* codes			
Heart valve surgery	27 (4.3)	5 (3.3)	0.56
Coronary artery bypass grafting	21 (3.3)	4 (2.6)	0.65
Percutaneous coronary intervention	17 (2.7)	3 (1.8)	0.60
Angioplasty	56 (8.9)	8 (5.2)	0.14
Pacemaker/defibrillator	23 (3.6)	5 (3.3)	0.81
Cardioversion	65 (10.3)	2 (1.3)	<0.01
Procedures, CPT codes			
Ablation	19 (3.0)	2 (1.3)	0.24
Cardioversion	234 (37.0)	7 (4.6)	<0.01
Natural language processing			
At least 1 nonnegated mention	614 (97.2)	57 (37.0)	<0.01
No. of AF mentions§			
None	16 (2.5)	76 (49.3)	<0.01
First quartile	123 (19.5)	57 (37.0)	
Second quartile	159 (25.2)	15 (9.7)	
Third quartile	162 (25.6)	6 (3.9)	
Fourth quartile	172 (27.2)	0 (0)	
ECG with reference to AF	234 (37.0)	7 (4.6)	<0.01

Values are shown as n (%), unless otherwise specified. AF indicates atrial fibrillation; CPT, Current Procedural Terminology; *ICD*, *International Classification of Diseases*.

†Primary diagnosis refers to position 1 in the order of the billed codes.

‡Comorbid conditions were identified from *ICD* billing codes present in the patient medical record.

§Refers to the number of times a target term for AF was present in the clinical notes. The ranges are as follows: none, no mentions; first, 1–6; second, 7–19; third, 20–46; fourth, 48–670.

participation. Data set cleaning and analyses were completed using Stata v14.2, and the NLP was executed using Python. The Stata output for the model training is included in Table S3, along with the cut points and regression coefficients. The data that support the findings of this study are available from the corresponding author upon reasonable request. To protect patient information, sharing will be limited to Python scripts, in most cases.

Results

A total of 786 patients were included in the training set, with an AF prevalence of 80.4% per our reference standard. The mean age of the training population was 67.5 years (SD: 15.3), and 42.0% of participants were female. Comorbid conditions varied between patients with and without AF, including higher rates of coronary artery disease, congestive heart failure, thyroid disease, and cancer among AF patients. Aside from cardioversion, cardiac procedures did not differ significantly between patients with and without AF (Table 1).

The test characteristics for the training models are seen in Figures 1 and 2. Compared with the reference standard, accuracy, sensitivity, and negative predictive value were highest using the NLP model, whereas specificity and positive predictive value were highest using the comprehensive model. Figure 1 shows the receiver operating characteristic curves for each model compared with the reference standard. The areas under the receiver operating characteristic curve were highest for the comprehensive and NLP models, at 0.887 and 0.801, respectively ($P<0.01$). The *ICD* and NLP models did not differ significantly regarding discrimination (area under the receiver operating characteristic curve: 0.801 versus 0.798, respectively; $P=0.91$); the *ICD* model had higher specificity at the cost of lower sensitivity. The previously published models using AF-specific diagnosis codes and ECGs resulted in high false-negative rates (Figure 2).

The full candidate population included 22 000 patients, with a mean age of 67.1 years (SD: 15.1); 42.3% were female. The number of patients, patient characteristics, and OAC treatment rates varied substantially when the models were applied to the candidate population (Table 2, Figure 3). The number of patients who could be included in an AF sample varied by up to 6690 patients. The model using outpatient AF codes and ECG resulted in the smallest AF sample, including 11 512 patients, or 52.3% of the candidate population. Comparatively, the NLP model resulted in the largest AF sample, including 18 202 patients, or 82.7% of the candidate population. The mean age of patients identified as AF was

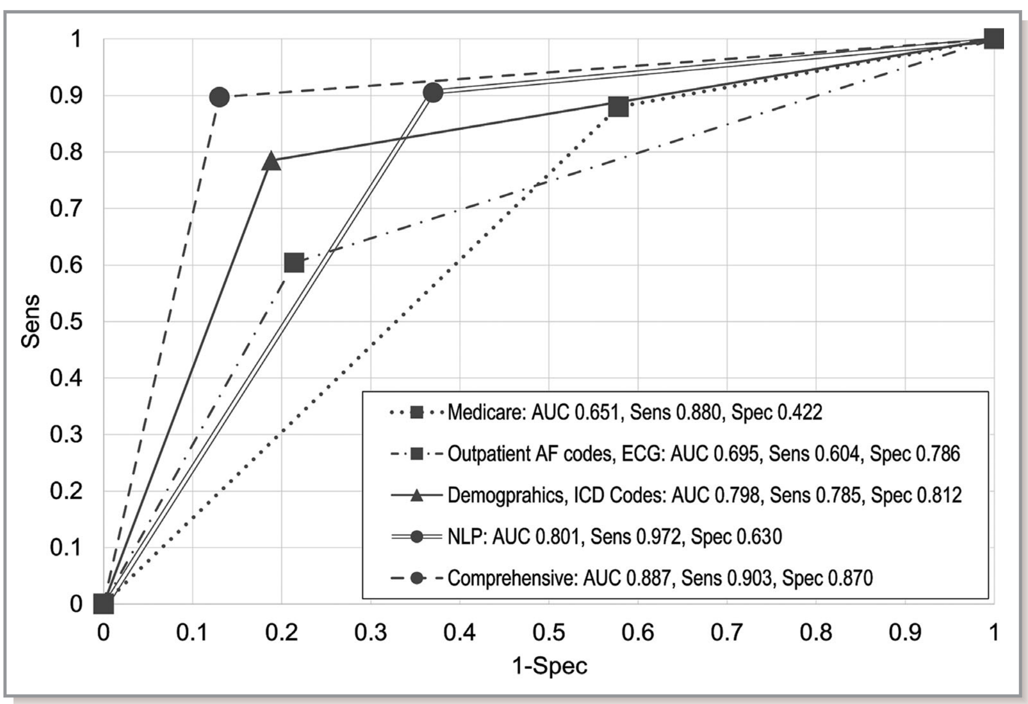

Figure 1. Receiver operating characteristic curves for different models to identify atrial fibrillation patients using the electronic medical record. In the training set (n=786), the AUC was highest for the comprehensive model and lowest for the Medicare model. AUC indicates area under the receiver operating characteristic curve; *ICD, International Classification of Diseases*; NLP, natural language processing; Sens, sensitivity; Spec, specificity.

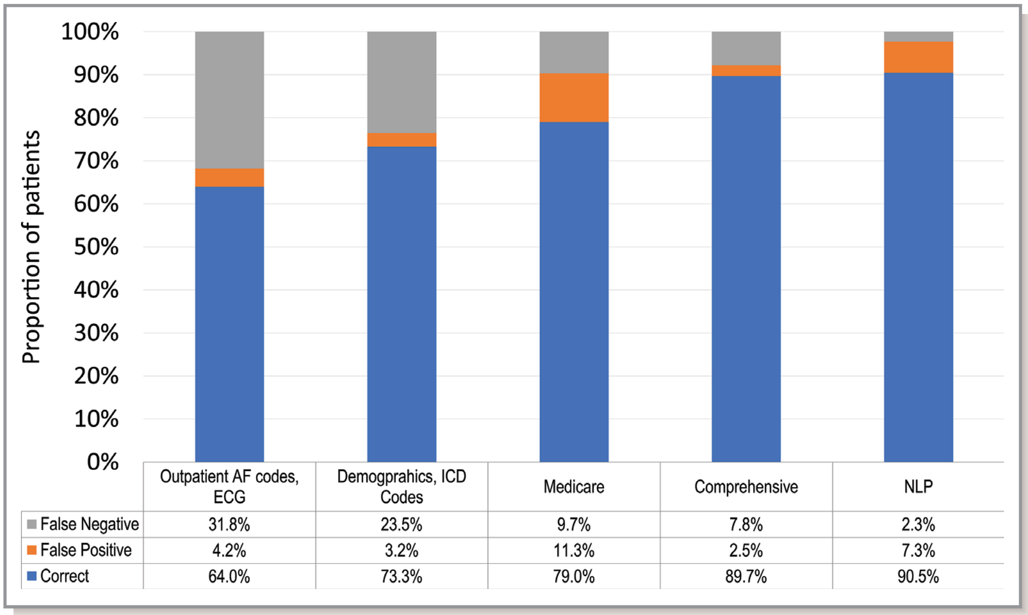

Figure 2. Proportion of correct, false-positive, and false-negative classifications for each model in the training set. In the training set (n=786), the NLP model resulted in the highest number of correctly classified patients, at the expense of a high false-positive rate. The outpatient billing codes and ECG method had the lowest number of correctly classified patients and the highest number of false negatives. AF indicates atrial fibrillation; *ICD, International Classification of Diseases*; NLP, natural language processing.

lowest using the outpatient AF/ECG model (68.1 years) and highest using the demographics/*ICD* codes model (70.1 years), with additional variation in comorbid conditions (Table 2). Finally, the proportion of AF patients with a CHA_2DS_2-VASc score ≥2 ranged from 78.5% to 85.3%. The OAC treatment rates for patients with a score ≥2 also varied, from 49.6% in the NLP model to 66.3% in the outpatient AF/ECG model.

Discussion

Accurate identification of patient populations is critical for effective quality-improvement efforts. The EMR provides important opportunities to identify patient populations, but standard, electronic cohort definitions do not exist. We found that the number of AF patients included in a cohort varies by an absolute range of up to 30%, depending on which electronic cohort definition is used. In large health systems, this translates into cohorts that differ by thousands of patients. In addition, quality measures such as OAC treatment rates varied by 16.7%, between 49.6% and 66.3%, depending on the cohort definition. These findings have important implications for quality-improvement initiatives, research endeavors, and case-mix analyses.

From a quality perspective, health systems use EMR-based tools to characterize patient populations and find opportunities for improvement.[17] In addition, OAC treatment rate in AF is a quality measure in the federal government's Merit-Based Incentive Payment System (MIPS)[18]; accurate estimation of the denominator—the number of patients with AF, in this case—is critical to its success. The AF quality measure for MIPS relies on AF billing codes and outpatient CPT codes for evaluation and management.[18] Based on our findings, this type of cohort selection could result in underestimating the true AF population; some electronic cohort definitions are biased and can omit a substantial number of patients, affecting the impact of the quality measure. Less sensitive AF cohort definitions would omit a large proportion of patients from any assessment of the quality of their care or related interventions. Conversely, false positives are also problematic for quality reporting because patients who do not truly have AF are unlikely to receive (or benefit from) guideline-recommended treatment.

Just as clinical trials and disease registries have specific inclusion and exclusion criteria, the same is needed for EMR-based research and initiatives. Trials and registries use "human-readable" definitions; for example, persistent AF is defined as "sustained for ≥7 days."[19] These definitions should have corresponding "machine-readable" definitions to increase uniformity and reproducibility in EMR-based initiatives. The challenge is creating machine-readable definitions that are portable across institutions. In this study, for example, the outpatient AF/ECG model had suboptimal

Table 2. Population Characteristics Based on the Patient-Selection Model

Selected Characteristics	Medicare	Outpatient AF Codes, ECG	Demographics, *ICD* Codes	NLP	Comprehensive	*P* Value
Proportion identified as AF, %	18 030 (82.0)	11 512 (52.3)	13 427 (61.0)	18 202 (82.7)	15 962 (72.6)	<0.01
Age, y, mean (SD)	67.8 (14.3)	68.1 (14.1)	70.8 (12.4)	68.7 (13.8)	69.8 (13.1)	<0.01
Female sex	7434 (41.2)	4846 (42.1)	5113 (38.1)	7538 (41.4)	6528 (40.9)	<0.01
White race	15 707 (87.2)	10 143 (88.1)	11 980 (89.2)	15 957 (87.7)	14 110 (88.4)	<0.01
Medicare	11 092 (61.5)	7116 (61.8)	8874 (66.1)	11 481 (63.1)	10 389 (65.1)	<0.01
CHA_2DS_2-VASc ≥2	14 920 (82.8)	9156 (79.5)	11 450 (85.3)	15 110 (83.0)	13 286 (83.2)	<0.01
OAC prescribed[†]	7838 (52.5)	6074 (66.3)	6572 (57.4)	7502 (49.6)	8127 (61.2)	<0.01
Comorbid conditions						
Acute myocardial infarction	2690 (14.9)	1493 (13.0)	2198 (16.4)	2567 (14.2)	2356 (14.8)	<0.01
Coronary artery disease	8463 (46.9)	5365 (46.6)	6809 (50.7)	8431 (46.3)	7496 (47.0)	<0.01
Valvular heart disease	6801 (37.7)	4001 (34.7)	5024 (37.4)	6604 (36.3)	5665 (35.5)	<0.01
Congestive heart failure	6859 (38.0)	4352 (37.8)	5766 (42.9)	3828 (37.5)	6173 (38.7)	<0.01
Cerebrovascular disease	5914 (32.8)	3077 (26.7)	3132 (23.3)	5506 (30.3)	4265 (27.7)	<0.01
Dementia	2488 (13.8)	1340 (11.6)	1776 (13.2)	2386 (13.1)	2092 (13.1)	<0.01
Diabetes mellitus	8283 (45.9)	4779 (41.5)	6219 (46.3)	8106 (44.5)	7080 (44.4)	<0.01
Chronic kidney disease	4487 (24.9)	2610 (26.7)	4082 (30.4)	4504 (24.7)	4306 (27.0)	<0.01
Hypertension	14 109 (78.3)	8729 (75.8)	10 797 (80.4)	14 068 (77.3)	12 261 (76.8)	<0.01
Cancer	6116 (33.9)	3886 (33.8)	5387 (40.1)	6257 (34.4)	5631 (35.3)	<0.01
Procedures						
Heart valve surgery	867 (4.8)	502 (4.4)	627 (4.7)	844 (4.6)	672 (4.2)	<0.01
Coronary artery bypass grafting	644 (3.6)	325 (2.8)	457 (3.4)	615 (3.4)	583 (3.7)	<0.01
Percutaneous coronary intervention	608 (3.4)	333 (2.9)	553 (4.1)	558 (3.1)	481 (3.0)	<0.01
Pacemaker/defibrillator	812 (4.5)	563 (4.9)	582 (4.3)	783 (4.3)	706 (4.4)	<0.01

Values shown as n (%), unless otherwise specified. AF indicates atrial fibrillation; *ICD*, *International Classification of Diseases*; NLP, natural language processing; OAC, oral anticoagulant.
[†]Including only patients with CHA_2DS_2-VASc ≥2.

performance, whereas it may perform well in the system for which it was designed. We showed it cannot easily be applied to a system in which patients receive fragmented care from different institutions using different EMRs.

Furthermore, many administrative data sets and definitions are used to calculate observed-to-expected event ratios in efforts to understand quality of care (and, on occasion, payment, scoring, etc). Underlying these calculations is a case mix, to account for severity of illness, on which to base expected outcomes. Once again, such case-mix analyses could vary dramatically with the definition of the underlying disease-based cohort, leading to wide variability in expected outcomes, observed outcomes, and downstream effects. With more precise, portable, cohort-definition methods, precision and utility of such analyses could improve dramatically.

NLP may have some advantages over billing data models because the clinical narrative may be less prone to certain types of variation, given that there are relatively few ways that clinicians state that a patient has AF. However, we were limited by low specificity with our rules-based approach. Machine learning, as opposed to rules-based approaches, can also be used for NLP but often requires large sets of labeled training data. Our future efforts will focus on improving NLP specificity by using the comprehensive model to automatically label patients and create a large training set for a machine learning approach. Still, whatever methods result in the ideal performing model (text or structured data, eg, *ICD* codes), we will need to ensure that the model is calibrated and portable. Efforts are underway using common data models such as the Observational Medical Outcomes Partnership (OMOP) common data model,[20] but standardization of the data that go into the common data model must also be a part of the process. For example, 2 different NLP systems can extract AF patients and map the concept to OMOP, but the systems are different; the common data model alone does not solve the portability issue. In the

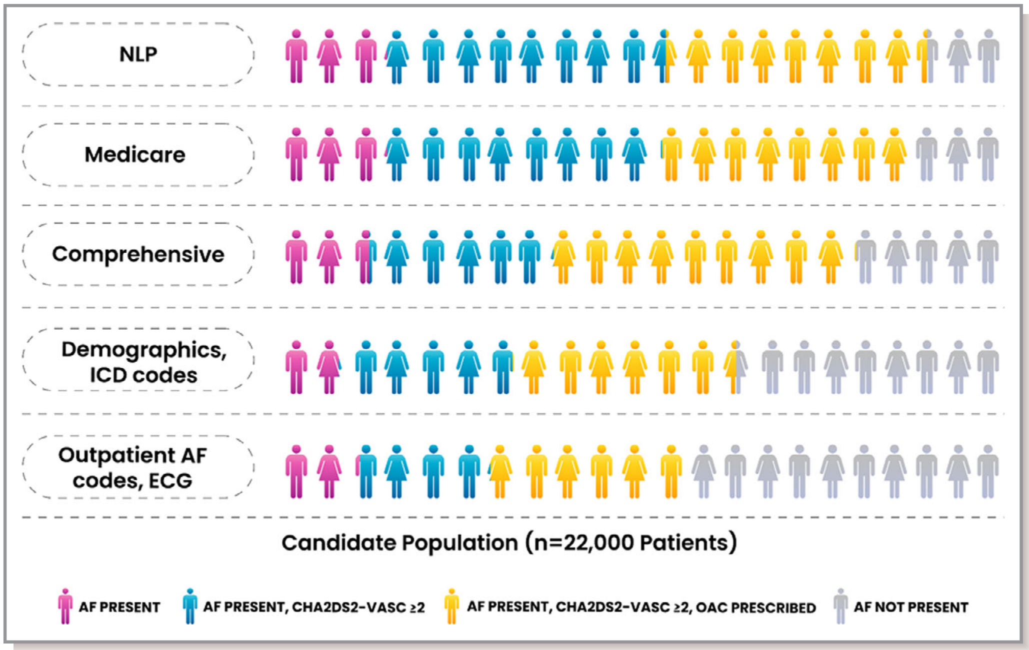

Figure 3. Proportion of patients included with CHA_2DS_2-VASc score ≥2 and treated with an OAC for each model. When applied to the candidate population, different patient-selection models resulted in populations with different sizes, stroke risks, and OAC treatment rates. The corresponding values are found in Table 2. "Outpatient AF codes, ECG" refers to the method used in prior publications from Kaiser Permanente. AF indicates atrial fibrillation; *ICD, International Classification of Diseases*; NLP, natural language processing; OAC, oral anticoagulant.

future, if EMRs become more similar, an option is for guidelines and regulations to include validated algorithms along with recommendations.

Limitations

We used diverse approaches to patient electronic cohort definitions in this study, and this is only 1 factor that can skew outcome results. CHA_2DS_2-VASc score calculations and OAC treatment classification methods can also vary and yield different apparent treatment rates. In addition, we used only a small fraction of the variables available in the EMR. We chose features based on widespread availability (eg, demographics) and controlled vocabularies (eg, *ICD* and CPT) for this demonstration project. Additional features, such as ejection fraction, have varying capture and format across institutions; adding features and increasing model complexity could decrease bias but would probably increase model overfitting and result in site-specific, nonportable models. Our candidate population, including patients with at least 1 AF billing code, was enriched with a high prevalence of "AF present" patients. Generation of models that accurately identify low-prevalence conditions, such as AF patients in an entire health system, is limited by challenges in creating a reference standard; manual chart review to identify 1% of the population is cumbersome, if not impossible. Our reference standard definition of AF was broad, and the results would differ with narrower definitions. From this larger group, health systems could apply criteria to select patient subsets, such as patients who have at least 2 outpatient encounters, a designated primary care physician within the health system, or a first AF encounter during admission for cardiac surgery. Finally, both billing codes and text-based terms vary between institutions. We did not include internal and external validation populations for each model because the purpose of this study was not to identify the optimal model to select AF patients but rather to compare population characteristics and outcomes from different approaches.

Conclusions

EMRs provide an opportunity to identify large patient cohorts for research and quality initiatives. Cohort selection is a critical step to realizing the potential of EMRs for quality improvement and research and a prerequisite to developing learning healthcare systems. Cohort definitions should be based on validated portable definitions to maximize comparability. In the case of AF, number of patients, characteristics, and outcomes vary depending on the patient-selection method. To optimize the impact of EMR-driven research and quality improvement, we need an unbiased, portable approach to identify patient populations. Combining multiple types of data from EMRs may serve this goal. Nevertheless, regardless of the data sources—structured data like *ICD* codes or unstructured data like text—we will ultimately require a common AF definition for use in research and quality improvement.

Sources of Funding

This work and R.U.S. are supported by a grant from the National Heart, Lung, and Blood Institute of the National Institutes of Health (K08HL136850). B.A.S. is supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (K23HL143156).

Disclosures

B.A.S. receives research support from Boston Scientific and Janssen; provides consulting to Janssen, Bayer, and Merit Medical; and speaks for North American Center for Continuing Medical Education (funded by Sanofi). The remaining authors have no disclosures to report.

References

1. Jensen PN, Johnson K, Floyd J, Heckbert SR, Carnahan R, Dublin S. A systematic review of validated methods for identifying atrial fibrillation using administrative data. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 1):141–147.
2. Piccini JP, Hammill BG, Sinner MF, Jensen PN, Hernandez AF, Heckbert SR, Benjamin EJ, Curtis LH. Incidence and prevalence of atrial fibrillation and associated mortality among Medicare beneficiaries, 1993–2007. *Circ Cardiovasc Qual Outcomes*. 2012;5:85–93.
3. Khurshid S, Keaney J, Ellinor PT, Lubitz SA. A simple and portable algorithm for identifying atrial fibrillation in the electronic medical record. *Am J Cardiol*. 2016;117:221–225.
4. Go AS, Hylek EM, Borowsky LH, Phillips KA, Selby JV, Singer DE. Warfarin use among ambulatory patients with nonvalvular atrial fibrillation: the anticoagulation and risk factors in atrial fibrillation (ATRIA) study. *Ann Intern Med*. 1999;131:927.
5. Molnar AO, van Walraven C, McArthur E, Fergusson D, Garg AX, Knoll G. Validation of administrative database codes for acute kidney injury in kidney transplant recipients. *Can J Kidney Health Dis*. 2016;3:1–10.
6. McCarthy C, Murphy S, Cohen JA, Rehman S, Jones-O'Connor M, Olshan DS, Singh A, Vaduganathan M, Januzzi JL Jr, Wasfy JH. Misclassification of myocardial injury as myocardial infarction: implications for assessing outcomes in value-based programs. *JAMA Cardiol*. 2019;4:460–464.
7. Chang TE, Lichtman JH, Goldstein LB, George MG. Accuracy of ICD-9-CM codes by hospital characteristics and stroke severity: paul coverdell national acute stroke program. *J Am Heart Assoc*. 2016;5:e003056. DOI: 10.1161/JAHA.115.003056.
8. Evans RS, Lloyd JF, Pierce LA. Clinical use of an enterprise data warehouse. *AMIA Annu Symp Proc*. 2012;2012:189–198.
9. HCUP-US Tools & Software Page [Internet]. Available at: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. November 5, 2018.
10. Liu X. Classification accuracy and cut point selection. *Stat Med*. 2012;31:2676–2686.
11. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform*. 2009;42:839–851.
12. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *J Biomed Inform*. 2011;44:728–737.
13. pyConTextNLP [Internet]. PyPI. Available at: https://pypi.org/project/pyConTextNLP/. 2019 Apr 10 April, 2019.
14. Taggart M, Chapman WW, Steinberg BA, Ruckel S, Pregenzer-Wenzler A, Du Y, Ferraro J, Bucher BT, Lloyd-Jones DM, Rondina MT, Shah RU. Comparison of 2 natural language processing methods for identification of bleeding among critically Ill patients. *JAMA Netw Open*. 2018;1:e183451.
15. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*. 2010;137:263–272.
16. January CT, Wann LS, Alpert JS, Calkins H, Cigarroa JE, Cleveland JC, Conti JB, Ellinor PT, Ezekowitz MD, Field ME, Murray KT, Sacco RL, Stevenson WG, Tchou PJ, Tracy CM, Yancy CW. 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.
17. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, Specchia ML. The impact of electronic health records on healthcare quality a systematic review and meta-analysis. *Eur J Public Health*. 2016;26:60–64.
18. Quality ID #326 (NQF 1525): atrial fibrillation and atrial flutter: chronic anticoagulation therapy—national quality strategy domain: effective clinical care – meaningful measure area: management of chronic condition. Available from: https://qpp.cms.gov/docs/QPP_quality_measure_specifications/Claims-Registry-Measures/2019_Measure_326_MedicarePartBClaims.pdf. Accessed August 16, 2019.
19. Packer DL, Mark DB, Robb RA, Monahan KH, Bahnson TD, Moretz K, Poole JE, Mascette A, Rosenberg Y, Jeffries N, Al-Khalidi HR, Lee KL; CABANA Investigators. Catheter ablation versus antiarrhythmic drug therapy for atrial fibrillation (CABANA) trial: study rationale and design. *Am Heart J*. 2018;199:192–199.
20. OMOP Common Data Model—OHDSI [Internet]. Available at: https://www.ohdsi.org/data-standardization/the-common-data-model/. Accessed 7 October, 2019.

SUPPLEMENTAL MATERIAL

Table S1. Model specifications.

Feature	**Definition**	**Models**
Kaiser	>1 outpatient ICD AF diagnosis billing code OR 1 outpatient ICD AF diagnosis billing code and ECG consistent with AF	Model 2
Medicare	≥1 inpatient ICD AF diagnosis billing code OR ≥2 outpatient ICD AF diagnoses billing codes within 365 days	Model 1
Demographics: Based on status at the time of the index AF encounter		
Age	Ordinal; roughly according to decade (18 to 29, 30 to 39, 40 to 49, 50 to 59, 60 to 69, 70 to 79, 80 and older)	Model 3 Model 5
Sex	Binary; female (reference) or male	Model 3 Model 5
Race	Categorical; white (reference), black, Asian, other/missing	Model 3 Model 5
Hispanic	Categorical; Hispanic, not Hispanic, or missing	Model 3 Model 5
Primary payer	Categorical; Medicare (reference), Medicaid, private, self, other/missing	Model 3 Model 5
Inpatient AF diagnosis	Binary; presence of an inpatient billing code at any time during the study period	Model 1 Model 3 Model 5
Number of outpatient AF diagnoses	Numeric; the total number of outpatient AF billing codes during the study period	Model 3 Model 5
Primary AF diagnosis	Binary; presence of an AF billing code in the first billing position any time during the study period	Model 3 Model 5
Acute myocardial infarction	Binary; presence of an ICD diagnosis billing code for condition any time during the study period	Model 3 Model 5
Comorbid Conditions: Based on ICD billing codes		
Coronary artery disease Valvular heart disease Congestive heart failure	Binary; presence of an ICD diagnosis billing code for condition any time during the study period	Model 3 Model 5

Peripheral vascular disease Cerebrovascular disease Dementia Pulmonary heart disease Rheumatologic disease Gastrointestinal ulcer Liver disease Diabetes mellitus Acute renal disease Chronic renal disease Lymphoma Hypertension Coagulopathy Electrolyte disorder Anemia Cancer Dialysis		
Procedures: Based on ICD billing codes		
Heart valve surgery Coronary artery bypass grafting Percutaneous coronary intervention Angiogram Pacemaker/defibrillator Cardioversion (ICD based)	Binary; presence of an ICD procedure billing code for procedure any time during the study period	Model 3 Model 5
Procedures: Based on CPT billing codes		
Cardioversion (CPT based) Ablation	Binary; presence of a CPT billing code for procedure any time during the study period	Model 5
Electrocardiograms and text		
Electrocardiogram	Binary; presence of an ECG interpretation that includes an AF-specific term any time during the study period	Model 2 Model 5
Number of AF mentions in the text	Categorical, split into zero and quartiles for values >0; the total number of AF mentions, as extracted by NLP, in the available text any time during the study period	Model 5
Non-negated AF mention in the text	Binary; the presence of a non-negated reference to AF in the available text any time during the study period	Model 4

Model Definitions (also see Methods section of manuscript):

Model #1: Outpatient and inpatient AF billing codes (Medicare methodology)[2]: ≥1 inpatient diagnosis or ≥2 outpatient diagnoses within 365 days

Model #2: Outpatient AF codes and electrocardiogram (Kaiser methodology)[4]: >1 outpatient ICD code, 1 outpatient ICD code and ECG consistent with AF

Model #3: Demographics and International Classification of Diseases (ICD) billing AF codes: A logistic regression model using patient demographics, presence of an inpatient AF diagnosis, presence of a primary AF diagnosis, number of outpatient AF diagnoses, comorbid conditions and procedures from ICD codes, year of index AF diagnosis

Model #4: Natural language processing: At least one non-negated mention of AF in the clinical text

Model #5: Comprehensive: A comprehensive logistic regression model combining patient demographics, presence of an inpatient AF diagnosis, presence of a primary AF diagnosis, number of outpatient AF diagnoses, comorbid conditions and procedures from ICD codes, year of index AF diagnosis, at least one non-negated mention of AF in clinical text, ECG with reference to AF, CPT codes for ablation or cardioversion

AF=atrial fibrillation; CPT=current procedural terminology; ICD=International Classification of Diseases

Table S2. Target terms used in natural language processing task to identify atrial fibrillation patients from clinical notes.

<table>
<tr><th>Target term</th><th>Regular Expression</th></tr>
<tr><td>afib</td><td>\bafib\b|\batrial\sfib|a-fib|a\.\sfib|a\.fib|\ba\sfib\b</td></tr>
<tr><td></td><td></td></tr>
<tr><td>Modifier terms</td><td></td></tr>
<tr><td>no</td><td>\bno(?!\sfurther)\b</td></tr>
<tr><td>not</td><td>\bnot\b</td></tr>
<tr><td>none</td><td>\bnone\b</td></tr>
<tr><td>negative</td><td>\bnegative\b</td></tr>
<tr><td>denies</td><td>denies|denied|denying</td></tr>
<tr><td>family</td><td>\bmother\b|\bfather\b|\bsister\b|\bbrother\b|\bdaughter\b|\bson\b|\baunt\b|\buncle\b|\bgranddaughter\b|\bgrandson\b</td></tr>
<tr><td>rule out</td><td>r/o|r\\o|\brule\s+out\b|\brules\s+out\b|\bruled\s+out\b</td></tr>
<tr><td>unlikely</td><td>\bunlikely\b</td></tr>
<tr><td>investigate</td><td>\binvestigate\b|\binvestigating\b</td></tr>
<tr><td>look for</td><td>\blook\s+for\b\b</td></tr>
<tr><td>differential</td><td>\bdifferential\b\b|ddx</td></tr>
<tr><td>possible</td><td>\bpossible\b</td></tr>
<tr><td>holter</td><td>\b(holter|event)\s+(monitor(ing)?\s+)?ordered\s+for\b\b|ddx</td></tr>
<tr><td>etc</td><td>\betc\b</td></tr>
<tr><td>screen for</td><td>\bscreen\s+for\b</td></tr>
<tr><td>risk of</td><td>\brisk\s+(of|for)\b</td></tr>
<tr><td>suspicious</td><td>\bsuspicious\b</td></tr>
<tr><td>question of</td><td>\bquestion\s+of\b</td></tr>
</table>

Table S3. Training model regression results.

KAISER MODEL

```
Logistic regression                              Number of obs     =        786
                                                 LR chi2(1)        =      78.77
                                                 Prob > chi2       =     0.0000
Log likelihood = -349.45524                      Pseudo R2         =     0.1013

-----------------------------------------------------------------------------------
binary_adj_goldstd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+---------------------------------------------------------------
            kaiser |   1.723243   .2125683     8.11   0.000     1.306616    2.139869
             _cons |   .7256704    .110745     6.55   0.000     .5086141    .9427267
-----------------------------------------------------------------------------------


Logistic model for binary_adj_goldstd
number of observations =      786
area under ROC curve   =   0.6951


Empirical cutpoint estimation
Method:                               Liu
Reference variable:                   binary_adj_goldstd (0=neg, 1=pos)
Classification variable:              kaiser_lr
Empirical optimal cutpoint:           .7971682
Sensitivity at cutpoint:              0.60
Specificity at cutpoint:              0.79
Area under ROC curve at cutpoint:     0.70
(415 real changes made)


Detailed report of sensitivity and specificity
-----------------------------------------------------------------------------
                                          Correctly
Cutpoint      Sensitivity   Specificity   Classified          LR+          LR-
-----------------------------------------------------------------------------
```

```
( >= 1 )          60.44%        78.57%       63.99%       2.8207       0.5035
------------------------------------------------------------------------------
                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
     ------------------------------------------------------------
           786     0.6951       0.0192        0.65738     0.73276

binary_adj |     kaiser_class
  _goldstd |      Pos.       Neg. |     Total
-----------+----------------------+----------
  Abnormal |       382        250 |       632
    Normal |        33        121 |       154
-----------+----------------------+----------
     Total |       415        371 |       786

True abnormal diagnosis defined as binary_adj_goldstd = 1
                                                  [95% Confidence Interval]
---------------------------------------------------------------------------
Prevalence                         Pr(A)       80%       77%      83.1%
---------------------------------------------------------------------------
Sensitivity                      Pr(+|A)     60.4%     56.5%     64.3%
Specificity                      Pr(-|N)     78.6%     71.2%     84.8%
ROC area               (Sens. + Spec.)/2      .695      .657      .733
---------------------------------------------------------------------------
Likelihood ratio (+)     Pr(+|A)/Pr(+|N)      2.82      2.07      3.84
Likelihood ratio (-)     Pr(-|A)/Pr(-|N)      .503      .443      .572
Odds ratio                   LR(+)/LR(-)       5.6       3.7      8.48
Positive predictive value        Pr(A|+)       92%       89%     94.5%
Negative predictive value        Pr(N|-)     32.6%     27.9%     37.6%
---------------------------------------------------------------------------
```

MEDICARE MODEL

```
Logistic regression                               Number of obs     =        786
                                                  LR chi2(1)        =      65.41
                                                  Prob > chi2       =     0.0000
```

```
Log likelihood = -356.13488                     Pseudo R2         =     0.0841

------------------------------------------------------------------------------------
binary_adj_goldstd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
         simpleicd |   1.675786   .2039045     8.22   0.000      1.27614    2.075431
             _cons |   .1563462   .1689453     0.93   0.355    -.1747804    .4874729
------------------------------------------------------------------------------------


Logistic model for binary_adj_goldstd
number of observations =      786
area under ROC curve   =   0.6509

Empirical cutpoint estimation
Method:                                Liu
Reference variable:                    binary_adj_goldstd (0=neg, 1=pos)
Classification variable:               medicare_lr
Empirical optimal cutpoint:            .70051131
Sensitivity at cutpoint:               0.88
Specificity at cutpoint:               0.42
Area under ROC curve at cutpoint:      0.65
(645 real changes made)

Detailed report of sensitivity and specificity
------------------------------------------------------------------------------
                                          Correctly
Cutpoint      Sensitivity   Specificity   Classified          LR+          LR-
------------------------------------------------------------------------------
( >= 1 )           87.97%        42.21%       79.01%       1.5223       0.2849
------------------------------------------------------------------------------
                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
     ------------------------------------------------------------
           786     0.6509       0.0210        0.60978     0.69205
```

```
binary_adj |    medicare_class
  _goldstd |      Pos.       Neg. |     Total
-----------+----------------------+----------
  Abnormal |       556         76 |       632
    Normal |        89         65 |       154
-----------+----------------------+----------
     Total |       645        141 |       786

True abnormal diagnosis defined as binary_adj_goldstd = 1
                                                  [95% Confidence Interval]
---------------------------------------------------------------------------
Prevalence                         Pr(A)       80%       77%      83.1%
---------------------------------------------------------------------------
Sensitivity                      Pr(+|A)       88%     85.2%     90.4%
Specificity                      Pr(-|N)     42.2%     34.3%     50.4%
ROC area               (Sens. + Spec.)/2      .651       .61      .692
---------------------------------------------------------------------------
Likelihood ratio (+)     Pr(+|A)/Pr(+|N)      1.52      1.33      1.75
Likelihood ratio (-)     Pr(-|A)/Pr(-|N)      .285      .215      .377
Odds ratio                   LR(+)/LR(-)      5.34      3.59      7.96
Positive predictive value        Pr(A|+)     86.2%     83.3%     88.8%
Negative predictive value        Pr(N|-)     46.1%     37.7%     54.7%
---------------------------------------------------------------------------
```

ICD MODEL

```
Logistic regression                               Number of obs     =        786
                                                  LR chi2(43)       =     260.73
                                                  Prob > chi2       =     0.0000
Log likelihood = -258.47662                       Pseudo R2         =     0.3353

------------------------------------------------------------------------------------
   binary_adj_goldstd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------------+-------------------------------------------------------------
              agegrp1 |   .0446484   .0099897     4.47   0.000      .025069    .0642279
                  sex |    .358715   .2435289     1.47   0.141    -.1185929    .8360229
```

race_categ						
Black	-1.942464	1.308129	-1.48	0.138	-4.506349	.6214212
Asian	-1.375892	1.244579	-1.11	0.269	-3.815221	1.063437
Other/missing	-.4826733	.4840383	-1.00	0.319	-1.431371	.4660243
hispanic	.0927108	.1684888	0.55	0.582	-.2375211	.4229427
index_pay1_categ						
2	.8464419	.6084834	1.39	0.164	-.3461637	2.039047
3	.0162844	.3308754	0.05	0.961	-.6322194	.6647882
4	.4823887	.6985767	0.69	0.490	-.8867964	1.851574
5	2.056091	1.095191	1.88	0.060	-.0904435	4.202626
inpatientdx	1.640814	.3087453	5.31	0.000	1.035684	2.245943
countoutpatient_afib	.4895864	.0782739	6.25	0.000	.3361723	.6430004
afibicd_primary	-.9130516	.2387842	-3.82	0.000	-1.38106	-.4450431
index_year	.0646928	.0550775	1.17	0.240	-.0432571	.1726428
amidiag_all	.3809207	.4520047	0.84	0.399	-.5049923	1.266834
caddiag_all	.0267004	.2832404	0.09	0.925	-.5284405	.5818414
valvediag_all	-.121885	.2692074	-0.45	0.651	-.6495218	.4057518
chfdiag_all	.2366036	.3143618	0.75	0.452	-.3795342	.8527414
pvddiag_all	-.6911086	.3083174	-2.24	0.025	-1.2954	-.0868175
cvddiag_all	-1.278336	.2852031	-4.48	0.000	-1.837324	-.7193479
dementiadiag_all	-.0838705	.3846242	-0.22	0.827	-.8377201	.669979
pulmdzdiag_all	-.2523725	.2855902	-0.88	0.377	-.812119	.3073739
rheumdiag_all	.7370864	.5824089	1.27	0.206	-.404414	1.878587
ulcerdiag_all	.0170389	.5962036	0.03	0.977	-1.151499	1.185576
liverdiag_all	-.4056906	.3076358	-1.32	0.187	-1.008646	.1972644
dmdiag_all	-.0513154	.2568892	-0.20	0.842	-.554809	.4521781
renaldiag_all	.0839594	.4110108	0.20	0.838	-.7216069	.8895257
ckddiag_all	.7758547	.4285092	1.81	0.070	-.0640079	1.615717
lymphdiag_all	.7620735	.8795399	0.87	0.386	-.961793	2.48594
pulmhtndiag_all	.1058361	.3386697	0.31	0.755	-.5579443	.7696164
htndiag_all	.0271604	.2786255	0.10	0.922	-.5189355	.5732563

```
      thyroiddiag_all |   .2414638   .2949052     0.82   0.413    -.3365399    .8194674
         coagdiag_all |   .3509488   .3714666     0.94   0.345    -.3771123     1.07901
         elecdiag_all |  -.1480849   .3050795    -0.49   0.627    -.7460297      .44986
       anemiadiag_all |   .1875522   .3232971     0.58   0.562    -.4460984    .8212028
       cancerdiag_all |   .3973601   .2820185     1.41   0.159     -.155386    .9501062
 dialysis_icdproc_all |  -.3362597   .7277529    -0.46   0.644    -1.762629     1.09011
    valve_icdproc_all |  -.1382759   .6785085    -0.20   0.839    -1.468128    1.191576
     cabg_icdproc_all |  -.0303663   .8015717    -0.04   0.970    -1.601418    1.540685
      pci_icdproc_all |   .2712517   .9442262     0.29   0.774    -1.579398    2.121901
    angio_icdproc_all |   .0102707   .6437773     0.02   0.987     -1.25151    1.272051
ppm_defib_icdproc_all |  -.8168417   .6898823    -1.18   0.236    -2.168986    .5353027
     dccv_icdproc_all |   1.196939   .8358149     1.43   0.152    -.4412278    2.835106
                _cons |   -132.932   110.9735    -1.20   0.231     -350.436    84.57198
--------------------------------------------------------------------------------------
Note: 0 failures and 44 successes completely determined.

Logistic model for binary_adj_goldstd
number of observations =      786
area under ROC curve   =   0.8738

Empirical cutpoint estimation
Method:                                Liu
Reference variable:                    binary_adj_goldstd (0=neg, 1=pos)
Classification variable:               icd_lr
Empirical optimal cutpoint:            .77046734
Sensitivity at cutpoint:               0.78
Specificity at cutpoint:               0.81
Area under ROC curve at cutpoint:      0.80
(525 real changes made)

Detailed report of sensitivity and specificity
------------------------------------------------------------------------------
                                          Correctly
Cutpoint      Sensitivity   Specificity   Classified          LR+          LR-
------------------------------------------------------------------------------
```

```
( >= 1 )           78.48%        81.17%       79.01%       4.1676       0.2651
------------------------------------------------------------------------------


                     ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
     ------------------------------------------------------------
           786     0.7982       0.0178        0.76337     0.83313


binary_adj |       icd_class
  _goldstd |      Pos.       Neg. |     Total
-----------+----------------------+----------
  Abnormal |       496        136 |       632
    Normal |        29        125 |       154
-----------+----------------------+----------
     Total |       525        261 |       786


True abnormal diagnosis defined as binary_adj_goldstd = 1


                                                  [95% Confidence Interval]
---------------------------------------------------------------------------
Prevalence                         Pr(A)       80%       77%      83.1%
---------------------------------------------------------------------------
Sensitivity                      Pr(+|A)     78.5%     75.1%     81.6%
Specificity                      Pr(-|N)     81.2%     74.1%       87%
ROC area               (Sens. + Spec.)/2      .798      .763      .833
---------------------------------------------------------------------------
Likelihood ratio (+)     Pr(+|A)/Pr(+|N)      4.17      2.99       5.8
Likelihood ratio (-)     Pr(-|A)/Pr(-|N)      .265      .224      .313
Odds ratio                   LR(+)/LR(-)      15.7      10.1      24.5
Positive predictive value        Pr(A|+)     94.5%     92.2%     96.3%
Negative predictive value        Pr(N|-)     47.9%     41.7%     54.1%
---------------------------------------------------------------------------
```

NLP MODEL

```
Logistic regression                              Number of obs     =        786
```

```
                                                LR chi2(1)        =     287.78
                                                Prob > chi2       =     0.0000
Log likelihood = -244.94682                     Pseudo R2         =     0.3701

------------------------------------------------------------------------------------
binary_adj_goldstd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
 afnlp_mrn_predict |   4.061283   .2916117    13.93   0.000     3.489735    4.632831
             _cons |  -1.684339   .2566414    -6.56   0.000    -2.187347   -1.181331
------------------------------------------------------------------------------------
Logistic model for binary_adj_goldstd
number of observations =      786
area under ROC curve   =   0.8007

Empirical cutpoint estimation
Method:                                Liu
Reference variable:                    binary_adj_goldstd (0=neg, 1=pos)
Classification variable:               afnlp_lr
Empirical optimal cutpoint:            .53578696
Sensitivity at cutpoint:               0.97
Specificity at cutpoint:               0.63
Area under ROC curve at cutpoint:      0.80
(671 real changes made)

Detailed report of sensitivity and specificity
------------------------------------------------------------------------------
                                          Correctly
Cutpoint      Sensitivity   Specificity   Classified          LR+          LR-
------------------------------------------------------------------------------
( >= 1 )           97.15%        62.99%       90.46%       2.6248       0.0452
------------------------------------------------------------------------------


                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
     ------------------------------------------------------------
```

```
         786     0.8007       0.0198        0.76189     0.83949


binary_adj |      afnlp_class
  _goldstd |      Pos.       Neg. |     Total
-----------+----------------------+----------
  Abnormal |       614         18 |       632
    Normal |        57         97 |       154
-----------+----------------------+----------
     Total |       671        115 |       786


True abnormal diagnosis defined as binary_adj_goldstd = 1


                                                  [95% Confidence Interval]
---------------------------------------------------------------------------
Prevalence                         Pr(A)       80%       77%      83.1%
---------------------------------------------------------------------------
Sensitivity                      Pr(+|A)     97.2%     95.5%     98.3%
Specificity                      Pr(-|N)       63%     54.8%     70.6%
ROC area               (Sens. + Spec.)/2      .801      .762      .839
---------------------------------------------------------------------------
Likelihood ratio (+)     Pr(+|A)/Pr(+|N)      2.62      2.14      3.23
Likelihood ratio (-)     Pr(-|A)/Pr(-|N)     .0452     .0282     .0724
Odds ratio                   LR(+)/LR(-)        58      32.9       102
Positive predictive value        Pr(A|+)     91.5%     89.1%     93.5%
Negative predictive value        Pr(N|-)     84.3%     76.4%     90.5%
---------------------------------------------------------------------------
```

COMPREHENSIVE MODEL

```
Logistic regression                               Number of obs     =        786
                                                  LR chi2(47)       =     459.16
                                                  Prob > chi2       =     0.0000
Log likelihood = -159.26033                       Pseudo R2         =     0.5904


------------------------------------------------------------------------------------
   binary_adj_goldstd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

afnlp_mrn_predict	2.2063	.5049817	4.37	0.000	1.216554	3.196046
agegrp1	.0328202	.0130682	2.51	0.012	.0072069	.0584335
sex	.006181	.3288827	0.02	0.985	-.6384173	.6507793
race_categ						
Black	-2.187707	1.514649	-1.44	0.149	-5.156364	.7809499
Asian	.2546778	3.239757	0.08	0.937	-6.095129	6.604485
Other/missing	-.656952	.6036018	-1.09	0.276	-1.83999	.5260858
hispanic	.1670156	.2225115	0.75	0.453	-.269099	.6031302
index_pay1_categ						
2	.5163838	.7883102	0.66	0.512	-1.028676	2.061443
3	-.2421059	.4610076	-0.53	0.599	-1.145664	.6614523
4	.9414504	.8906946	1.06	0.291	-.804279	2.68718
5	2.136887	1.231747	1.73	0.083	-.2772917	4.551066
inpatientdx	-.1369238	.4172727	-0.33	0.743	-.9547633	.6809157
countoutpatient_afib	.1403936	.0740849	1.90	0.058	-.0048101	.2855974
afibicd_primary	-1.262697	.3451367	-3.66	0.000	-1.939153	-.5862419
index_year	-.0405696	.075878	-0.53	0.593	-.1892877	.1081485
amidiag_all	.561467	.6321116	0.89	0.374	-.6774491	1.800383
caddiag_all	-.0842321	.3797421	-0.22	0.824	-.8285129	.6600487
valvediag_all	-.0823799	.3639308	-0.23	0.821	-.7956711	.6309113
chfdiag_all	-.113876	.4480115	-0.25	0.799	-.9919625	.7642104
pvddiag_all	-.6556172	.4073686	-1.61	0.108	-1.454045	.1428106
cvddiag_all	-1.12224	.3907799	-2.87	0.004	-1.888155	-.3563258
dementiadiag_all	.0200449	.5326683	0.04	0.970	-1.023966	1.064056
pulmdzdiag_all	-.3854925	.3913458	-0.99	0.325	-1.152516	.3815311
rheumdiag_all	.5394939	.8438259	0.64	0.523	-1.114374	2.193362
ulcerdiag_all	.2380188	.8312635	0.29	0.775	-1.391228	1.867265
liverdiag_all	-.4841053	.4199376	-1.15	0.249	-1.307168	.3389572
dmdiag_all	-.0962111	.3521609	-0.27	0.785	-.7864338	.5940115
renaldiag_all	-.2154304	.5936916	-0.36	0.717	-1.379045	.9481838

```
          ckddiag_all |   1.626106   .6297206     2.58   0.010      .391876    2.860335
        lymphdiag_all |   .5713589   1.101511     0.52   0.604    -1.587562     2.73028
      pulmhtndiag_all |   .4727443   .4709309     1.00   0.315    -.4502633    1.395752
          htndiag_all |  -.5871352   .3913627    -1.50   0.134    -1.354192    .1799215
      thyroiddiag_all |    .012911   .4038323     0.03   0.974    -.7785859    .8044078
         coagdiag_all |   .4569073   .5011965     0.91   0.362    -.5254199    1.439234
         elecdiag_all |   -.613357   .4176761    -1.47   0.142    -1.431987    .2052732
       anemiadiag_all |   .2884396   .4747701     0.61   0.543    -.6420927    1.218972
       cancerdiag_all |   .1483555     .38792     0.38   0.702    -.6119538    .9086648
 dialysis_icdproc_all |  -.9619088   1.086018    -0.89   0.376    -3.090466    1.166648
    valve_icdproc_all |  -1.106743   .9657247    -1.15   0.252    -2.999529    .7860425
     cabg_icdproc_all |   1.961944   1.187286     1.65   0.098    -.3650924    4.288981
      pci_icdproc_all |   -.623917    1.42916    -0.44   0.662    -3.425019    2.177185
    angio_icdproc_all |  -.3559314   .8904829    -0.40   0.689    -2.101246    1.389383
ppm_defib_icdproc_all |   .0118631    1.02624     0.01   0.991     -1.99953    2.023257
   dccv_icdcpt_binary |  -.0867684   1.027261    -0.08   0.933    -2.100162    1.926625
    ablate_cpt_binary |  -1.658924   1.220243    -1.36   0.174    -4.050557    .7327092
             ecg_afib |   2.881581   .6455071     4.46   0.000     1.616411    4.146752
        q_afnlp_total |   1.182809   .2549187     4.64   0.000      .683178    1.682441
                _cons |   78.62454   152.8337     0.51   0.607    -220.9241    378.1731
------------------------------------------------------------------------------------
Note: 0 failures and 14 successes completely determined.


Logistic model for binary_adj_goldstd
number of observations =      786
area under ROC curve   =   0.9504


Empirical cutpoint estimation
Method:                               Liu
Reference variable:                   binary_adj_goldstd (0=neg, 1=pos)
Classification variable:              comprehensive_lr
Empirical optimal cutpoint:           .7892637
Sensitivity at cutpoint:              0.90
Specificity at cutpoint:              0.87
Area under ROC curve at cutpoint:     0.89
```

```
(591 real changes made)


Detailed report of sensitivity and specificity
------------------------------------------------------------------------------
                                           Correctly
Cutpoint      Sensitivity   Specificity   Classified          LR+          LR-
------------------------------------------------------------------------------
( >= 1 )           90.35%        87.01%       89.69%       6.9568       0.1109
------------------------------------------------------------------------------
                      ROC                   -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
     ------------------------------------------------------------
           786     0.8868       0.0148        0.85779     0.91582


binary_adj |  comprehensive_class
  _goldstd |      Pos.       Neg. |     Total
-----------+----------------------+----------
  Abnormal |       571         61 |       632
    Normal |        20        134 |       154
-----------+----------------------+----------
     Total |       591        195 |       786


True abnormal diagnosis defined as binary_adj_goldstd = 1


                                                  [95% Confidence Interval]
---------------------------------------------------------------------------
Prevalence                         Pr(A)       80%       77%      83.1%
---------------------------------------------------------------------------
Sensitivity                      Pr(+|A)     90.3%     87.8%     92.5%
Specificity                      Pr(-|N)       87%     80.7%     91.9%
ROC area               (Sens. + Spec.)/2      .887      .858      .916
---------------------------------------------------------------------------
Likelihood ratio (+)     Pr(+|A)/Pr(+|N)      6.96      4.62      10.5
Likelihood ratio (-)     Pr(-|A)/Pr(-|N)      .111     .0867      .142
Odds ratio                   LR(+)/LR(-)      62.7      36.7       107
```

```
Positive predictive value        Pr(A|+)     96.6%     94.8%     97.9%

Negative predictive value        Pr(N|-)     68.7%     61.7%     75.2%

---------------------------------------------------------------------------
```