SUPPLEMENTAL MATERIAL

Development of an Electronic Phenotyping Algorithm for Cardioembolic Stroke

Table of Contents

SUPPLEMENTAL METHODS

Mass General Brigham Biobank

For developing the feature extraction algorithms of features based on free text cardiology reports, we collected EHR data from a subset of 30,716 individuals in the Mass General Brigham Biobank as of December 2018 using the Research Patient Database Repository (RPDR). The subjects included in this analysis were a convenience sample comprising individuals with available genomic data, though genomic data were not utilized in this analysis.

Algorithm for Ascertaining Stroke Events

A combination of ICD9 and ICD10 codes were used to ascertain stroke events for patients missing stroke dates in the Massachusetts General Hospital and Brigham and Women's Hospital prospective ischemic stroke registry. The list of ICD9 stroke codes used were 433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.01, 434.11, and 434.91. The list of ICD10 stroke codes used were I63.00, I63.011, I63.012, I63.019, I63.02, I63.031, I63.032, I63.033, I63.039, I63.09, I63.10, I63.111, I63.112, I63.113, I63.119, I63.12, I63.131, I63.132, I63.133, I63.139, I63.19, I63.20, I63.211, I63.212, I63.219, I63.22, I63.231, I63.232, I63.233, I63.239, I63.29, I63.30, I63.311, I63.312, I63.319, I63.321, I63.322, I63.323, I63.329, I63.331, I63.332, I63.339, I63.341, I63.342, I63.343, I63.349, I63.39, I63.40, I63.411, I63.412, I63.413, I63.419, I63.421, I63.422, I63.423, I63.429, I63.431, I63.432, I63.433, I63.439, I63.441, I63.442, I63.443, I63.449, I63.49, I63.50, I63.511, I63.512, I63.513, I63.519, I63.521, I63.522, I63.523, I63.529, I63.531, I63.532, I63.539, I63.541, I63.542, I63.543, I63.549, I63.59, I63.6, I63.8, and I63.9.

Data Preparation for NLP Algorithm Development

We obtained free-form text notes from cardiology reports, including echocardiogram reports, in the EHR using RPDR. We pre-processed them to make them usable for natural language processing (NLP). First, we removed language abnormalities from the cardiology reports. Specifically, we lowercased the text, replaced single underscores with a space, removed two consecutive hyphens (to remove excessive hyphens while preserving negative signs), added a space after a colon, added a space after a semi-colon, added a space after a comma, replaced multiple consecutive question marks with one question mark, expanded contractions common in the English-language, and removed single apostrophes followed by an "s". Cardiology reports tend to have designated sections with several headings. Sometimes, headings were capitalized, so lowercasing the text could risk losing the possibility of finding specific search terms under a specified report section. Similarly, excessive whitespaces were not removed since multiple whitespaces could be used to delineate separate sections in cardiology reports. Stop words could have been removed, but removal substantially increased pre-processing run-time. In addition, stop words proved useful for manually reviewing reports, which included ambivalent language. Punctuation marks could have also been removed but were chosen to remain since sentence units had to be preserved for identifying specific search terms in individual sentences. After performing these pre-processing steps, we finally attained a clean corpus for data analysis.

NLP Algorithm Development

We developed NLP regular expressions algorithms for 11 cardioembolic stroke features (**Figure I**). Features included mitral stenosis, left atrial appendage thrombus, left ventricular thrombus, akinetic left ventricular segment, mitral valve prolapse, mitral annulus calcification, left atrial turbulence, delayed emptying velocity, atrial septal aneurysm, patent foramen ovale, and hypokinetic left ventricular segment. Algorithms and test characteristics for identification of each feature are summarized in **Table IV**.

Each NLP regular expressions algorithm (except for delayed emptying velocity and hypokinetic left ventricular segment) consisted of three types of rules: neutral rules, positive rules, and negation rules. First, strings matching the neutral rules were removed since their presence do not indicate whether the feature was present nor absent. For example, phrases containing the feature and "reason" were removed since physicians will often list the reasons behind the administration of a cardiology report at the heading of a cardiology report. Similarly, phrases containing the feature and one of the words: "eval" for evaluate, "exclude", "assess", "rule out", "r/o" for rule out, "icd-", and "diagnosis", were removed from the cardiology report. Second, cases of cardiology reports matching the positive rules were positively flagged for the feature. Third, cases of cardiology reports matching the negative rules will have their positive flags removed. After applying neutral, positive, and negation regular expression rules, we finally attained the covariates based on NLP regular expressions algorithms.

We developed a procedure for developing regular expressions rules. First, for each feature, a cardiology domain expert defined the search criteria for a feature: including search terms to search for and optionally, specific report locations to search in. Second, we searched for phrases which included those search terms in the cleaned corpus of cardiology reports using R packages 'quanteda' and 'corpus'. Third, after identifying the most common phrases including those search terms, regular expressions were written to match patterns in those common phrases and these regular expressions were used to define the positive rules and negation rules. Fourth, we reiteratively applied the regular expressions to cardiology reports and manually reviewed them for additional search terms and regular expressions to improve and develop additional neutral, positive, and negation rules. In this way, the NLP regular expressions algorithms were reiteratively revised to achieve a positive predictive value of ≥ 80% based on chart reviews for samples of 50 reports.

For the second step in the procedure for developing regular expressions rules, we developed a process for searching common phrases, which include specified search terms. First, R package quanteda tokens() function was used to tokenize the corpus into unigrams with the symbols removed and separators removed. Second, quanteda kwic() function was used to locate keywords in context by identifying an n-sized window of words around the search term. Third, data manipulation using R package dplyr was applied to filter for (or against) specific words before and/or after the search term to identify desirable phrases indicating the presence of the cardioembolic stroke feature. Fourth, R package corpus term_stats() function was applied to tabulate a frequency table for counting the occurrence of phrases of length n (n-grams) containing the search term, for phrases that occur at least 2 times. Since it was uncertain as to how long a phrase of words should be for capturing a feature, we tested the n-gram phrase length from 1 to 40 and counted the resulting number of phrases generated from each length. From this testing, a plot of the length of the phrases against the number of resulting phrases was generated. If the phrase length was 1, we would yield only a few (e.g. 1-2) results, which consist of a single word containing the search term. As the phrase length increased, the number of phrases including the search term increased. We searched for long phrases containing the search term since longer phrases were more informative for indicating the presence or absence of the feature. However, if the phrase length containing the search term was too high, the number of phrases including the search term started to decrease, because the phrase was too long and thus too unique. So, the frequency counts for these long phrases drop to 1 and do not meet the minimum count criteria of occurring at least 2 times. To allow for flexibility and add robustness, a range of n-gram length was chosen based on and around the peak of the graph in the plot (e.g. from 8 to 13 since peak was at 11). Fifth, R package corpus term_stats() function was applied again, now using the chosen range of n-gram length for tabulating the frequency of long phrases containing the search term. Sixth, since some phrases commonly occurred, usually with the same first 2 words, albeit with slight variation in their diction (e.g. "there is no evidence of left ventricular thrombus" vs. "there is no obvious evidence of left ventricular thrombus"), the length of phrases could be computed for each phrase and the phrases could be grouped by the concatenation of the first 2 words in each phrase. Then the longest phrase within each group was computed. Representative, long phrases were tabulated in descending count order to display the most common and informative phrases containing the search terms. These representative, long phrases were then analyzed for regular expressions that would capture the presence or absence of a cardioembolic stroke feature.

Example: NLP Algorithm for Akinetic Left Ventricular Segment

The NLP regular expressions algorithm for akinetic left ventricular segment required identification of specific sections in the echocardiogram report. First, we removed phrases satisfying the neutral rules in the cleaned corpus. Second, we identified cases that contain one of the search terms. Third, we separated the report into sentences by applying R package tokenizers function tokenize_sentences() onto the original corpus text before pre-processing since the pre-processed corpus did not retain enough information (e.g. whitespaces, punctuation marks) for facile sentence tokenization. R package tokenizers function tokenize_sentences() was used since splitting text into sentences based on periods did not always work as periods were also used as decimal points. Fourth, we searched for sentences containing the desired cardiology report section heading (e.g. “left ventr”). Fifth, we searched for sentences containing the desired search term (e.g. “akine” and “dyskine”) and applied negation rules for the search term. Sixth, we found the location of all of the other report section headings, such as “venous”, “pulmonic valve”, “pericardium”, “interatrial septum”, “conclusions”, “dyssynchrony”, “pericardial disease”, “pulmonary valve”, “right ventr”, “interventricular septum”, “left atrium”, “right atrium”, “tricuspid valve”, “aortic valve”, “mitral valve”, “interatrial septum”, and “report_end”. This list of report section headings was constructed based on experience from reiteratively refining the algorithm. Finally, we identified positive cases of when the search term was between the desired section heading and another section heading. Additional effort was required to identify the search terms under specific report sections, because the search terms could also appear under other report sections.

Example: NLP Algorithm for Delayed Left Atrial Appendage Emptying Velocity

The NLP regular expressions algorithm for delayed emptying velocity was challenging since this feature could be reported in different formats, either as a number, a range of numbers, or as a qualitative description. From applying R code like that of **Figure II**, phrases containing “velocity” were extracted. Manual examination of the phrases showed that there were various formats used to express the positive cases of delayed emptying velocity (Table VI) and the negative cases of delayed emptying velocity (Table VII). These different formats guided the development of an NLP algorithm through a combination of if/else logic and regular expressions.

Example: NLP Algorithm for Left Ventricular Ejection Fraction

Extracted left ventricular ejection fraction quantities less than or equal to 40 were considered to qualify for the hypokinetic left ventricular segment feature. The transthoracic echocardiogram reports and transesophageal echocardiogram reports from MGH and BWH were analyzed for developing regular expression-based algorithms for extracting left ventricular ejection fraction (Figure **III**).

SUPPLEMENTAL TABLES

Table I. Modified cardioembolic stroke TOAST features

High-risk sources		**Data Sources**		**Time window of feature (relative to stroke)**
1	Mechanical prosthetic valve and Bioprosthetic cardiac valve	ICD codes, Procedure codes		Before or up to 90d after stroke
2	Mitral stenosis	ICD codes	Echo report	Before or up to 90d after stroke
3	Atrial fibrillation (including lone atrial fibrillation)	ICD codes		Before or up to 90d after stroke
4	Left atrial/atrial appendage thrombus		Echo report	90d before or after stroke
	Intracardiac thrombus*	ICD codes		90d before or after stroke
5	Sick sinus syndrome	ICD codes		Before or up to 90d after stroke
6	Recent myocardial infarction (<4 weeks)	ICD codes, CPT codes		<=4 weeks prior by definition
7	Left ventricular thrombus		Echo Report	Before or up to 90d after stroke
8	Dilated cardiomyopathy	ICD codes		Before or up to 90d after stroke
9	Akinetic left ventricular segment		Echo report	Before or up to 90d after stroke
10	Atrial myxoma	ICD codes		Before or up to 90d after stroke
11	Infective endocarditis	ICD codes		90d before or after stroke
Medium-risk sources				
12	Mitral valve prolapse	ICD codes	Echo report	Before or up to 90d after stroke
13	Mitral annulus calcification	ICD codes	Echo report	Before or up to 90d after stroke
14	Left atrial turbulence (smoke)		Echo report	90d before or after stroke
15	Atrial septal aneurysm	ICD codes	Echo report	Before or up to 90d after stroke

16	Patent foramen ovale	ICD codes	Echo report	Before or up to 90d after stroke
17	Atrial flutter	ICD codes		Before or up to 90d after stroke
18	Nonbacterial thrombotic endocarditis	ICD codes		90d before or after stroke
19	Congestive heart failure	ICD codes		Before or up to 90d after stroke
20	Hypokinetic left ventricular segment		Echo report	Before or up to 90d after stroke
21	Myocardial infarction (>4 weeks, <6 months)	ICD codes, Procedure codes		Between 6 mos to 4 weeks after by definition
	Delayed emptying velocity†		Echo report	90d before or after stroke

(1) **Mechanical prosthetic valve and Bioprosthetic cardiac valve.** Under TOAST, Mechanical prosthetic valve is a high-risk source and Bioprosthetic cardiac valve is a medium-risk source. Under this electronic phenotyping algorithm, we merge these two features into one feature since there is not enough resolution to discriminate between them. Many procedure codes could be used for either type of valve.

(2) **Mitral stenosis.** Under TOAST, Mitral stenosis with atrial fibrillation is a high-risk source and Mitral stenosis without atrial fibrillation is a medium-risk source. Under this electronic phenotyping algorithm, we treat mitral stenosis and atrial fibrillation as separate features.

(3) **Atrial fibrillation.** Under TOAST, atrial fibrillation and lone atrial fibrillation are treated as two separate features, where Atrial fibrillation (other than lone atrial fibrillation) is a high-risk source and lone atrial fibrillation is a medium-risk source. Under this electronic phenotyping algorithm, this feature has been subordinated into one "Atrial Fibrillation" feature.

(4) **Left atrial/atrial appendage thrombus**. Under this electronic phenotyping algorithm, an additional sub-feature, "Non-specific intracardiac thrombus", was created based on ICD codes since there are no ICD codes specific for left atrial appendage thrombus.

(16) **Patent foramen ovale**. Under this electronic phenotyping algorithm, patent foramen ovale feature consists of two sub-features: patent foramen ovale and atrial septal defect since they both render similar clinical effects.

*** Intracardiac thrombus**. An additional ICD-based feature was developed for intracardiac thrombus since ICD codes existed for intracardiac thrombus albeit not particularly for left atrial appendage thrombus.

† **Delayed emptying velocity**. An NLP feature was developed for delayed left atrial appendage emptying velocity since expert knowledge deemed it clinically relevant to cardioembolism.

Table II. ICD codes and procedure codes for algorithm dictionary

Code Type	Code	Code Description
Mitral stenosis		
ICD9	394.0	Mitral stenosis
ICD9	394.2	Mitral stenosis with insufficiency
ICD9	396.0	Mitral valve stenosis and aortic valve stenosis
ICD9	396.1	Mitral valve stenosis and aortic valve insufficiency
ICD9	746.5	Congenital mitral stenosis
ICD10	I05.0	Rheumatic mitral stenosis
ICD10	I05.2	Rheumatic mitral stenosis with insufficiency
ICD10	I34.2	Nonrheumatic mitral (valve) stenosis
ICD10	Q23.2	Congenital mitral stenosis
Atrial fibrillation (including lone atrial fibrillation)		
ICD9	427.3	Atrial fibrillation and flutter
ICD9	427.31	Atrial fibrillation
ICD10	I48	Atrial fibrillation and flutter
ICD10	I48.0	Paroxysmal atrial fibrillation
ICD10	I48.1	Persistent atrial fibrillation
ICD10	I48.2	Chronic atrial fibrillation
ICD10	I48.9	Unspecified atrial fibrillation and atrial flutter
ICD10	I48.91	Unspecified atrial fibrillation
Intracardiac thrombus		
ICD10	I23.6	Thrombosis of atrium, auricular appendage, and ventricle as current complications following acute myocardial infarction
ICD10	I51.3	Intracardiac thrombosis, not elsewhere classified
Sick sinus syndrome		
ICD9	427.81	Sinoatrial dysfunction
ICD10	I49.5	Sick sinus syndrome
Atrial myxoma		
ICD9	212.7	Benign neoplasm of heart
ICD10	D15.1	Benign neoplasm of heart
Recent myocardial infarction (<4weeks)		
ICD9	410	Acute myocardial infarction
ICD9	410.0	Acute myocardial infarction, of anterolateral wall
ICD9	410.01	Acute myocardial infarction, of anterolateral wall, initial episode of care
ICD9	410.1	Acute myocardial infarction, of other anterior wall
ICD9	410.11	Acute myocardial infarction, of other anterior wall, initial episode of care
ICD9	410.21	Acute myocardial infarction, of inferolateral wall, initial episode of care
ICD9	410.3	Acute myocardial infarction, of inferoposterior wall
ICD9	410.31	Acute myocardial infarction, of inferoposterior wall, initial episode of care

ICD9	410.4	Acute myocardial infarction, of other inferior wall
ICD9	410.41	Acute myocardial infarction, of other inferior wall, initial episode of care
ICD9	410.51	Acute myocardial infarction, of other lateral wall, initial episode of care
ICD9	410.6	Acute myocardial infarction, true posterior wall infarction
ICD9	410.61	Acute myocardial infarction, true posterior wall infarction, initial episode of care
ICD9	410.7	Acute myocardial infarction, subendocardial infarction
ICD9	410.71	Acute myocardial infarction, subendocardial infarction, initial episode of care
ICD9	410.8	Acute myocardial infarction, of other specified sites
ICD9	410.81	Acute myocardial infarction, of other specified sites, initial episode of care
ICD9	410.9	Acute myocardial infarction, unspecified site
ICD9	410.91	Acute myocardial infarction, unspecified site, initial episode of care
ICD10	I21	St elevation (stemi) and non-st elevation (nstemi) myocardial infarction
ICD10	I21.0	St elevation (stemi) myocardial infarction of anterior wall
ICD10	I21.01	St elevation (stemi) myocardial infarction involving left main coronary artery
ICD10	I21.02	St elevation (stemi) myocardial infarction involving left anterior descending coronary artery
ICD10	I21.09	St elevation (stemi) myocardial infarction involving other coronary artery of anterior wall
ICD10	I21.1	St elevation (stemi) myocardial infarction of inferior wall
ICD10	I21.11	St elevation (stemi) myocardial infarction involving right coronary artery
ICD10	I21.19	St elevation (stemi) myocardial infarction involving other coronary artery of inferior wall
ICD10	I21.2	St elevation (stemi) myocardial infarction of other sites
ICD10	I21.21	St elevation (stemi) myocardial infarction involving left circumflex coronary artery
ICD10	I21.29	St elevation (stemi) myocardial infarction involving other sites
ICD10	I21.3	St elevation (stemi) myocardial infarction of unspecified site
ICD10	I21.4	Non-st elevation (nstemi) myocardial infarction
Dilated cardiomyopathy		
ICD9	425.4	Other primary cardiomyopathies
ICD10	I42.0	Dilated cardiomyopathy
Infective Endocarditis		
ICD9	421	Acute and subacute endocarditis
ICD9	421.9	Acute endocarditis, unspecified
ICD9	424.9	Endocarditis, valve unspecified
ICD9	424.90	Endocarditis, valve unspecified, unspecified cause
ICD9	424.91	Endocarditis in diseases classified elsewhere
ICD9	424.99	Other endocarditis, valve unspecified
ICD10	I33	Acute and subacute endocarditis
ICD10	I33.0	Acute and subacute infective endocarditis
ICD10	I33.9	Acute and subacute endocarditis, unspecified

ICD10	I38	Endocarditis, valve unspecified
ICD10	I39	Endocarditis and heart valve disorders in diseases classified elsewhere
Mitral valve prolapse		
ICD10	I34.1	Nonrheumatic mitral (valve) prolapse
Mitral annulus calcification		
ICD10	I34.8	Other nonrheumatic mitral valve disorders
Patent foramen ovale (including atrial septal defect)		
ICD10	I23.1	Atrial septal defect as current complication following acute myocardial infarction
ICD10	Q21.1	Atrial septal defect
ICD9 Procedure Code	35.41	Enlargement Of Existing Atrial Septal Defect
Atrial flutter		
ICD9	427.32	Atrial flutter
ICD10	I48.3	Typical atrial flutter
ICD10	I48.4	Atypical atrial flutter
ICD10	I48.92	Unspecified atrial flutter
Nonbacterial thrombotic endocarditis		
ICD9	391.1	Acute rheumatic endocarditis
ICD10	I01.1	Acute rheumatic endocarditis
ICD10	M32.11	Endocarditis in systemic lupus erythematosus
Congestive heart failure		
ICD9	402.01	Malignant hypertensive heart disease with congestive heart failure
ICD9	402.11	Benign hypertensive heart disease with congestive heart failure
ICD9	402.91	Unspecified hypertensive heart disease with congestive heart failure
ICD9	404.01	Hypertensive heart and renal disease, malignant, with congestive heart failure
ICD9	404.03	Hypertensive heart and renal disease, malignant, with congestive heart failure and renal failure
ICD9	404.11	Hypertensive heart and renal disease, benign, with congestive heart failure
ICD9	404.13	Hypertensive heart and renal disease, benign, with congestive heart failure and renal failure
ICD9	404.91	Hypertensive heart and renal disease, unspecified, with congestive heart failure
ICD9	404.93	Hypertensive heart and renal disease, unspecified, with congestive heart failure and renal failure
ICD9	428.0	Congestive heart failure
ICD10	I11.0	Hypertensive Heart Disease With Heart Failure
ICD10	I13.0	Hypertensive Heart And Chronic Kidney Disease With Heart Failure And Stage 1 Through Stage 4 Chronic Kidney Disease, Or Unspecified Chronic Kidney Disease
ICD10	I13.2	Hypertensive Heart And Chronic Kidney Disease With Heart Failure And With Stage 5 Chronic Kidney Disease, Or End Stage Renal Disease
ICD10	I50.20	Unspecified Systolic (Congestive) Heart Failure

ICD10	I50.21	Acute Systolic (Congestive) Heart Failure
ICD10	I50.22	Chronic Systolic (Congestive) Heart Failure
ICD10	I50.23	Acute On Chronic Systolic (Congestive) Heart Failure
ICD10	I50.30	Unspecified Diastolic (Congestive) Heart Failure
ICD10	I50.31	Acute Diastolic (Congestive) Heart Failure
ICD10	I50.32	Chronic Diastolic (Congestive) Heart Failure
ICD10	I50.33	Acute On Chronic Diastolic (Congestive) Heart Failure
ICD10	I50.40	Unspecified Combined Systolic (Congestive) And Diastolic (Congestive) Heart Failure
ICD10	I50.41	Acute Combined Systolic (Congestive) And Diastolic (Congestive) Heart Failure
ICD10	I50.42	Chronic Combined Systolic (Congestive) And Diastolic (Congestive) Heart Failure
ICD10	I50.43	Acute On Chronic Combined Systolic (Congestive) And Diastolic (Congestive) Heart Failure
ICD10	I50.814	Right heart failure due to left heart failure
ICD10	I50.9	Heart failure, unspecified
Myocardial infarction (>4 weeks, <6 months)		
ICD9	410	Acute myocardial infarction
ICD9	410.0	Acute myocardial infarction, of anterolateral wall
ICD9	410.00	Acute myocardial infarction, of anterolateral wall, episode of care unspecified
ICD9	410.01	Acute myocardial infarction, of anterolateral wall, initial episode of care
ICD9	410.02	Acute myocardial infarction, of anterolateral wall, subsequent episode of care
ICD9	410.1	Acute myocardial infarction, of other anterior wall
ICD9	410.10	Acute myocardial infarction, of other anterior wall, episode of care unspecified
ICD9	410.11	Acute myocardial infarction, of other anterior wall, initial episode of care
ICD9	410.12	Acute myocardial infarction, of other anterior wall, subsequent episode of care
ICD9	410.2	Acute myocardial infarction, of inferolateral wall
ICD9	410.20	Acute myocardial infarction, of inferolateral wall, episode of care unspecified
ICD9	410.21	Acute myocardial infarction, of inferolateral wall, initial episode of care
ICD9	410.22	Acute myocardial infarction, of inferolateral wall, subsequent episode of care
ICD9	410.3	Acute myocardial infarction, of inferoposterior wall
ICD9	410.30	Acute myocardial infarction, of inferoposterior wall, episode of care unspecified
ICD9	410.31	Acute myocardial infarction, of inferoposterior wall, initial episode of care
ICD9	410.32	Acute myocardial infarction, of inferoposterior wall, subsequent episode of care
ICD9	410.4	Acute myocardial infarction, of other inferior wall
ICD9	410.40	Acute myocardial infarction, of other inferior wall, episode of care unspecified

ICD9	410.41	Acute myocardial infarction, of other inferior wall, initial episode of care
ICD9	410.42	Acute myocardial infarction, of other inferior wall, subsequent episode of care
ICD9	410.50	Acute myocardial infarction, of other lateral wall, episode of care unspecified
ICD9	410.51	Acute myocardial infarction, of other lateral wall, initial episode of care
ICD9	410.52	Acute myocardial infarction, of other lateral wall, subsequent episode of care
ICD9	410.6	Acute myocardial infarction, true posterior wall infarction
ICD9	410.60	Acute myocardial infarction, true posterior wall infarction, episode of care unspecified
ICD9	410.61	Acute myocardial infarction, true posterior wall infarction, initial episode of care
ICD9	410.62	Acute myocardial infarction, true posterior wall infarction, subsequent episode of care
ICD9	410.7	Acute myocardial infarction, subendocardial infarction
ICD9	410.70	Acute myocardial infarction, subendocardial infarction, episode of care unspecified
ICD9	410.71	Acute myocardial infarction, subendocardial infarction, initial episode of care
ICD9	410.72	Acute myocardial infarction, subendocardial infarction, subsequent episode of care
ICD9	410.8	Acute myocardial infarction, of other specified sites
ICD9	410.80	Acute myocardial infarction, of other specified sites, episode of care unspecified
ICD9	410.81	Acute myocardial infarction, of other specified sites, initial episode of care
ICD9	410.82	Acute myocardial infarction, of other specified sites, subsequent episode of care
ICD9	410.9	Acute myocardial infarction, unspecified site
ICD9	410.90	Acute myocardial infarction, unspecified site, episode of care unspecified
ICD9	410.91	Acute myocardial infarction, unspecified site, initial episode of care
ICD9	410.92	Acute myocardial infarction, unspecified site, subsequent episode of care
ICD9	412	Old myocardial infarction
ICD10	I21	St elevation (stemi) and non-st elevation (nstemi) myocardial infarction
ICD10	I21.0	St elevation (stemi) myocardial infarction of anterior wall
ICD10	I21.01	St elevation (stemi) myocardial infarction involving left main coronary artery
ICD10	I21.02	St elevation (stemi) myocardial infarction involving left anterior descending coronary artery
ICD10	I21.09	St elevation (stemi) myocardial infarction involving other coronary artery of anterior wall
ICD10	I21.1	St elevation (stemi) myocardial infarction of inferior wall
ICD10	I21.11	St elevation (stemi) myocardial infarction involving right coronary artery
ICD10	I21.19	St elevation (stemi) myocardial infarction involving other coronary artery of inferior wall
ICD10	I21.2	St elevation (stemi) myocardial infarction of other sites

ICD10	I21.21	St elevation (stemi) myocardial infarction involving left circumflex coronary artery
ICD10	I21.29	St elevation (stemi) myocardial infarction involving other sites
ICD10	I21.3	St elevation (stemi) myocardial infarction of unspecified site
ICD10	I21.4	Non-st elevation (nstemi) myocardial infarction
ICD10	I22	Subsequent st elevation (stemi) and non-st elevation (nstemi) myocardial infarction
ICD10	I22.0	Subsequent st elevation (stemi) myocardial infarction of anterior wall
ICD10	I22.1	Subsequent st elevation (stemi) myocardial infarction of inferior wall
ICD10	I22.2	Subsequent non-st elevation (nstemi) myocardial infarction
ICD10	I22.8	Subsequent st elevation (stemi) myocardial infarction of other sites
ICD10	I22.9	Subsequent st elevation (stemi) myocardial infarction of unspecified site
ICD10	I25.2	Old myocardial infarction
Mechanical prosthetic valve or bioprosthetic mechanical valve		
CPT	33361	Transcatheter aortic valve replacement (TAVR/TAVI) with prosthetic valve; percutaneous femoral artery approach
CPT	33362	Transcatheter aortic valve replacement (TAVR/TAVI) with prosthetic valve; open femoral artery approach
CPT	33363	Transcatheter aortic valve replacement (TAVR/TAVI) with prosthetic valve; open axillary artery approach
CPT	33364	Transcatheter aortic valve replacement (TAVR/TAVI) with prosthetic valve; open iliac artery approach
CPT	33365	Transcatheter aortic valve replacement (TAVR/TAVI) with prosthetic valve; transaortic approach (eg, median sternotomy, mediastinotomy)
CPT	33366	Transcatheter aortic valve replacement (TAVR/TAVI) with prosthetic valve; transapical exposure (eg, left thoracotomy)
CPT	33367	Transcatheter aortic valve replacement (TAVR/TAVI) with prosthetic valve; cardiopulmonary bypass support with percutaneous peripheral arterial and venous cannulation (eg, femoral vessels) (List separately in addition to code for primary procedure)
CPT	33368	Transcatheter aortic valve replacement (TAVR/TAVI) with prosthetic valve; cardiopulmonary bypass support with open peripheral arterial and venous cannulation (eg, femoral, iliac, axillary vessels) (List separately in addition to code for primary procedure)
CPT	33369	Transcatheter aortic valve replacement (TAVR/TAVI) with prosthetic valve; cardiopulmonary bypass support with central arterial and venous cannulation (eg, aorta, right atrium, pulmonary artery) (List separately in addition to code for primary procedure)
CPT	33405	Replacement, aortic valve, with cardiopulmonary bypass; with prosthetic valve other than homograft or stentless valve
CPT	33406	Replacement, aortic valve, open, with cardiopulmonary bypass; with allograft valve (freehand)
CPT	33410	Replacement, aortic valve, open, with cardiopulmonary bypass; with stentless tissue valve
CPT	33411	Replacement, aortic valve; with aortic annulus enlargement, noncoronary cusp
CPT	33412	Replacement, aortic valve; with transventricular aortic annulus enlargement (Konno procedure)

CPT	33418	Transcatheter mitral valve repair, percutaneous approach, including transseptal puncture when performed; initial prosthesis
CPT	33419	Transcatheter mitral valve repair, percutaneous approach, including transseptal puncture when performed; additional prosthesis(es) during same session (List separately in addition to code for primary procedure)
CPT	33430	Replacement, mitral valve, with cardiopulmonary bypass
ICD9 Procedure Code	35.05	Endovascular replacement of aortic valve
ICD9 Procedure Code	35.06	Transapical replacement of aortic valve
ICD9 Procedure Code	35.07	Endovascular Replacement Of Pulmonary Valve
ICD9 Procedure Code	35.08	Transapical Replacement Of Pulmonary Valve
ICD9 Procedure Code	35.09	Endovascular Replacement Of Unspecified Heart Valve
ICD9 Procedure Code	35.20	Open and other replacement of unspecified heart valve
ICD9 Procedure Code	35.21	Open and other replacement of aortic valve with tissue graft
ICD9 Procedure Code	35.22	Other replacement of aortic valve
ICD9 Procedure Code	35.23	Replacement of mitral valve with tissue graft
ICD9 Procedure Code	35.24	Other replacement of mitral valve
ICD9 Procedure Code	35.25	Open And Other Replacement Of Pulmonary Valve With Tissue Graft
ICD9 Procedure Code	35.26	Open And Other Replacement Of Pulmonary Valve
ICD9 Procedure Code	35.27	Open And Other Replacement Of Tricuspid Valve With Tissue Graft
ICD9 Procedure Code	35.28	Open And Other Replacement Of Tricuspid Valve

ICD9 Procedure Code	35.97	Percutaneous mitral valve repair with implant
Hypokinetic left ventricular segment		
ICD9	414.1	Aneurysm of heart
ICD9	414.10	Aneurysm of heart (wall)
ICD9	414.19	Other aneurysm of heart
ICD10	I23.1	Atrial septal defect as current complication following acute myocardial infarction
ICD10	I25.3	Aneurysm of heart

Table III. Initial and final search criteria for each echo report feature.

<table>
<tr><th>Feature</th><th>Cardiology Report Type</th><th>Search Location</th><th>Initial Search Terms for Search Criteria</th><th>Final Search Terms for Search Criteria</th></tr>
<tr><td>Mitral stenosis (with atrial fibrillation, without atrial fibrillation)</td><td></td><td></td><td>mitral stenosis</td><td>mitral stenosis</td></tr>
<tr><td>Left atrial/atrial appendage thrombus</td><td>Transesophageal echo report (not transthoracic echo report),</td><td>Under [left atrium field]</td><td>thrombus, clot</td><td>left atrial appendage < thrombus

thrombus < left atrial appendage

left atrial appendage < clot

clot < left atrial appendage</td></tr>
<tr><td>Left ventricular thrombus</td><td></td><td></td><td>left ventricular thrombus</td><td>lv thrombus
lv clot
left ventr (thromb | clot)
(thromb | clot) left ventr
thromb | clot lv

left ventr (apex|apical) (thromb | clot)</td></tr>
<tr><td>Akinetic left ventricular segment</td><td></td><td>Under [left ventricular function field]</td><td>akinesis, dyskinesis</td><td>“akine” under [Left Ventricle] field
“dyskine” under [Left Ventricle] field

[See Example: NLP Algorithm for Search Terms in Specified Report Location (Akinetic Left Ventricular Segment)]</td></tr>
<tr><td>Mitral valve prolapse</td><td></td><td>Under [mitral valve field]</td><td>prolapse,
flail leaflet</td><td>mitral prolapse</td></tr>
<tr><td>Mitral annulus calcification</td><td></td><td>Under [mitral valve field]</td><td>calcification</td><td>mitral annul* calcif*</td></tr>
<tr><td>Left atrial turbulence (smoke)</td><td>Transesophageal echo report (not transthoracic echo report)</td><td></td><td>smoke</td><td>smoke,
then in sentences with smoke, find “left atr”, “laa”, “la appendage”, “left atrial appendage”, “appendage”

left atrium < spontaneous echo contrast</td></tr>
</table>

<table>
<tr><td></td><td></td><td></td><td></td><td>spontaneous echo contrast < (laa | la appendage | left atr | left atrial appendage | appendage)</td></tr>
<tr><td>Delayed emptying velocity</td><td></td><td></td><td></td><td>[See Example: NLP Algorithm for Features Reported in Different Formats (Delayed Emptying Velocity)]</td></tr>
<tr><td>Atrial septal aneurysm</td><td></td><td></td><td>atrial septal aneurysm,
interatrial septal aneurysm</td><td>atrial sept < aneurysm
aneurysm < atrial sept

interatrial sept < aneurysm
aneurysm < interatrial sept</td></tr>
<tr><td>Patent foramen ovale</td><td></td><td></td><td>patent foramen ovale,
shunting,
atrial septal defect</td><td>patent foramen ovale
pfo

atrial septal defect

interatrial (shunt | communication)
intracards shunt
residual shunt</td></tr>
<tr><td>Hypokinetic left ventricular segment</td><td></td><td>[See derived left ventricular ejection fraction field]</td><td>Use LVEF $\leq$ 40%</td><td>LV EF
ejection fraction
left ventricular function</td></tr>
</table>

Note that “Akinetic left ventricular segment” feature includes “akinesis” and “dyskinesis” of the left ventricle and is considered separate from “Hypokinetic left ventricular segment”. In addition, “<” denotes “before”; for example, the final search term “left atrial appendage < thrombus” denotes that phrases wherein “left atrial appendage” appears before “thrombus” were identified for extraction.

Table IV. Descriptive summary of algorithms of NLP-based TOAST-based features

Cardioembolic Feature based on NLP	PPV	Number of Charts Reviewed	Number of Cases Detected	Number of Cases Without Feature (Total = 440985)
Mitral Stenosis	94%	100	919	440066
Left atrial appendage thrombus	90%	100	142	440843
Left ventricular thrombus	94%	50	104	440881
Akinetic left ventricular segment	92%	100	2411	438574
Mitral valve prolapse	88%	100	1248	439737
Mitral annulus calcification	100%	50	10021	430964
Left atrial turbulence	100%	100	360	440625
Delayed emptying velocity	96%	50	330	440655
Atrial septal aneurysm	98%	100	277	440708
Patent foramen ovale (including atrial septal defect)	98%	100	2070	438915
Hypokinetic left ventricular segment	94%	50	8703	54257 (378025 NA's)

The Mass General Brigham Biobank was used to test and develop the algorithms for NLP-based features. The algorithms were applied to the entire dataset to detect the prevalence of the TOAST-based features and adjudication was performed to assess their PPVs.

Table V. Summary of echocardiograms among patients in MGH Stroke Registry

		MGH Stroke Registry
Total number of patients		1598
Stroke patients		1598 (100%)
Stroke patients without either TEE/TTE/echo not clearly labeled as TEE/TTE		107 (6.7%)
Stroke patients with either TEE/TTE/echo not clearly labeled as TEE/TTE		1491 (93.3%)
Stroke patients with TEE		218 (13.6%)
	Stroke patients with TEE within 4 weeks before or after stroke event	85 (5.3%)
	Stroke patients with TEE within 90 weeks before or after stroke event	113 (7.1%)
	Stroke patients with TEE within 4 weeks to 6 months stroke after event	49 (3.1%)
Stroke patients with TTE		792 (49.6%)
	Stroke patients with TTE within 4 weeks before or after stroke event	585 (36.6%)
	Stroke patients with TTE within 90 weeks before or after stroke event	629 (39.4%)
	Stroke patients with TTE within 4 weeks to 6 months stroke after event	79 (4.9%)
Stroke patients with echocardiogram (not clearly labeled as TEE/TTE)		1044 (65.3%)
	Stroke patients with echoes not clearly labeled as TEE/TTE within 4 weeks before or after stroke event	888 (55.6%)
	Stroke patients with echoes not clearly labeled as TEE/TTE within 90 days before or after stroke event	959 (60.0%)

	Stroke patients with echoes not clearly labeled as TEE/TTE within 4 weeks to 6 months stroke after event	95 (5.9%)

TEE=transesophageal echocardiogram; TTE=Transthoracic echocardiogram

Table VI. Different formats for reporting positive cases of delayed emptying velocity

Secondary Keywords	Positive Keywords	Sentence Structure	Modifier	Number	Unit
"emptying"	"left atrial appendage" or "la appendage" or "laa"	In same sentence	Check if "less 0.40 m/s"	Check if less than 0.40 m/s	Check if m/s (change units)
"diastolic"	"velocit.*"		Check if "low", "reduced"	Check if <1 → correct scale	
"ejection"			Check if "<0.40" m/s	Range: Get first number in range	
		"Filling and emptying velocit*" OR "Inflow and outflow velocit*" OR "left atrial appendage velocity systolic filling velocity 0.5 m/s and diastolic emptying velocity 0.6 m/s."		1 number → Get 1st number 2 numbers → Get 1st number	
"left atrial appendage" or "la appendage" or "laa" In prior sentence	"velocit" In target sentence	Difference sentences	Check if "less 0.40 m/s"	Check if <0.40 m/s	Check if m/s (change units)
	"emptying"		Check if "low", "reduced"	Check if <1 → correct scale	
	"diastolic"			Range: Get first number in range	
	"ejection"				

Table VII. Different formats for reporting negative cases of delayed emptying velocity

Secondary Keywords	Negative Keywords	Sentence Structure	Modifier	Number	Unit
"aortic valve", "aorta", "transaortic"	"velocit.*"	Same sentence	>0.4 m/s	Check if >0.4 m/s	Check if m/s (change units)
"mitral valve"			Greater		
"regurgitant"			"normal"		
"tricuspid"					
"pulmonary", "pv"					
"vein", "veins", "venous"					
"lvad"					
"Systolic emptying velocity"					
"lvot"					
"transgastric"					
"transvalvular"					

Table VIII. Multivariable logistic regression model applied to MGH Stroke Registry

Variables	Multivariable-adjusted OR (95% CI)	P-value
Age	0.0301 (0.00699 - 0.119)	0.11
Gender (Male)	1.0164 (0.997 - 1.037)	0.02
Atrial fibrillation	0.5417 (0.322 - 0.909)	2×10^{-16}
Atrial flutter	20.1595 (12 - 34.891)	0.98
Akinetic left ventricular segment	0.9851 (0.363 - 2.768)	0.79
Atrial myxoma	1.1381 (0.432 - 2.827)	0.99
Atrial septal aneurysm	NA	0.02
Congestive heart failure	0.033 (0.001 - 0.449)	0.42
Dilated cardiomyopathy	0.7716 (0.406 - 1.443)	0.59
Delayed Emptying Velocity	1.2045 (0.606 - 2.375)	0.99
Hypokinetic left ventricular segment	2321740.4733 (1.06E-55 - NA)	0.0000331
Infective endocarditis	5.1265 (2.38 - 11.189)	0.01
Intracardiac Thrombus	9.9354 (1.68 - 56.921)	NA
Left atrial appendage thrombus	NA	0.63
Left atrial turbulence	3.1467 (0.0549 - 221.118)	0.78
Left ventricular thrombus	1.7904 (0.0355 - 71.746)	0.94
Mitral annulus calcification	1.1277 (0.0546 - 40.397)	0.19
Mechanical and bioprosthetic cardiac valve	0.6686 (0.364 - 1.205)	0.12
Myocardial infarction (>4 weeks, <6 months)	6.7377 (0.682 - 83.105)	0.11
Recent myocardial infarction (<4 weeks)	0.2688 (0.0474 - 1.252)	0.15
Mitral stenosis	1.6784 (0.823 - 3.415)	0.64
Mitral valve prolapse	1.3786 (0.365 - 5.4)	0.76
Nonbacterial thrombotic endocarditis	1.4969 (0.107 - 20.61)	NA
Patent foramen ovale	NA	0.5
Sick sinus syndome	0.7891 (0.389 - 1.551)	0.0021

Some cardioembolic features are similar to one another, such as Atrial fibrillation and Atrial flutter; Akinetic left ventricular segment and Hypokinetic left ventricular segment; Delayed emptying velocity and Left atrial turbulence; and Congestive heart failure and Dilated Cardiomyopathy. These similarities are confirmed by moderate levels of positive correlation in Figure V as well as by similarly signed coefficients in Table VIII. Multivariable logistic regression model applied to MGH Stroke Registry

.

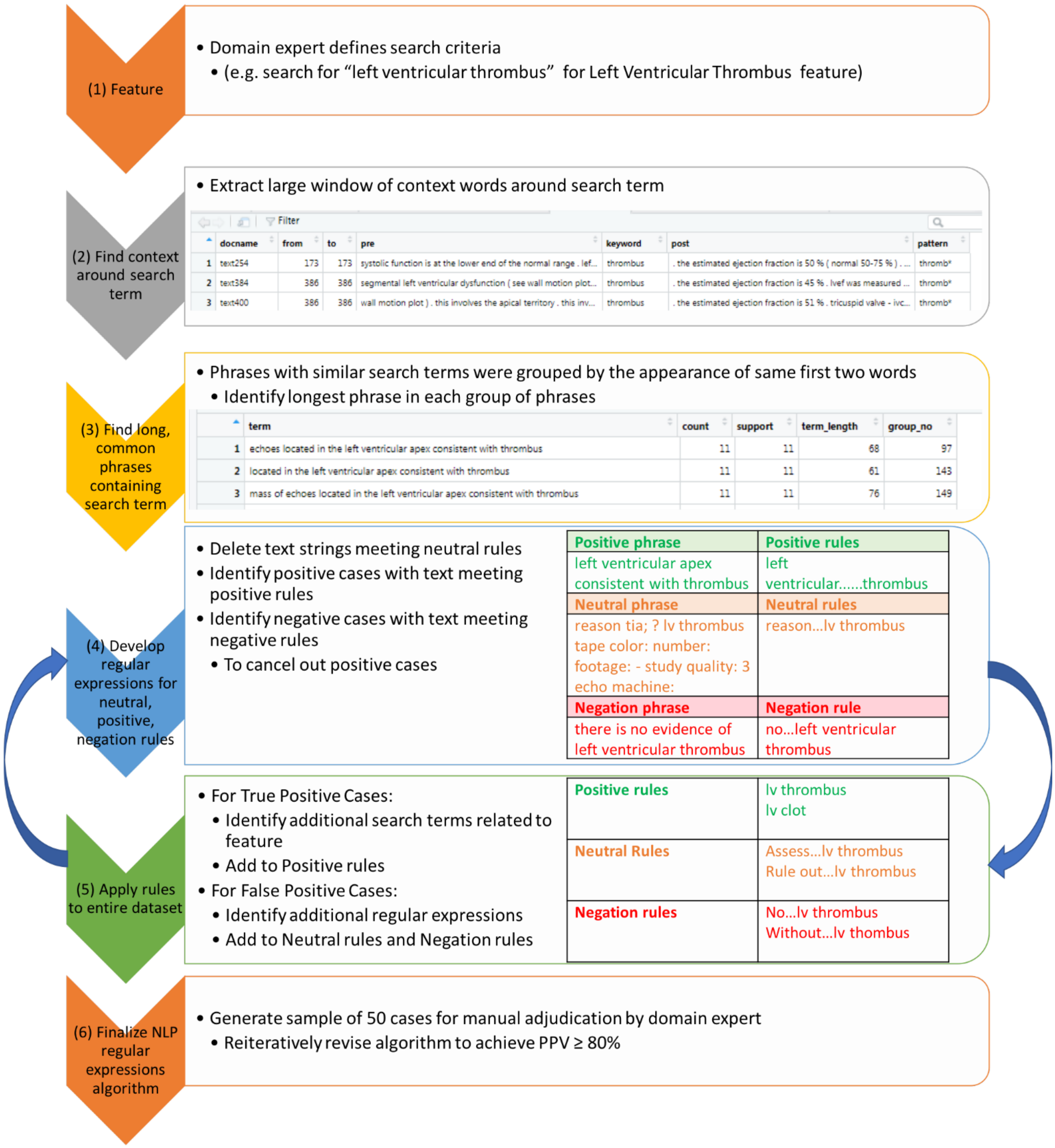
Table IX. Random forest model performance under inclusion of PFO compared to exclusion of PFO

	Accuracy	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	F1 Score	AUC
Random Forest without PFO	0.92 (0.89 to 0.95)	0.94 (0.90 - 0.97)	0.88 (0.82 - 0.94)	0.94 (0.91 - 0.97)	0.88 (0.80 - 0.93)	0.94 (0.92 - 0.96)	0.91 (0.89 - 0.94)
Random Forest with PFO	0.92 (0.89 to 0.94)	0.94 (0.91 - 0.96)	0.88 (0.80 - 0.93)	0.94 (0.91 - 0.97)	0.88 (0.82 - 0.93)	0.94 (0.92 - 0.96)	0.91 (0.87 - 0.94)

A comparison of model performance shows that the removal of PFO had minimal effect on model performance.

SUPPLEMENTAL FIGURES

Figure I. Process for feature extraction by identifying rules and regular expressions.



Domain expert clinicians defined the search criteria and search terms for a TOAST cardioembolic stroke feature. For each search term, a window of words was extracted to show the context around a certain keyword. The results were filtered to identify the presence of other keywords in the search term. Then similar phrases containing the search term were group together. Out of each group of phrases, the longest phrase was chosen to represent each group of phrases since longer phrases tend to contain more information. Then regular expressions were devised based on these phrases. Rules for identifying positive, neutral, and negatives usages of the search term were developed, such that neutral usages could be removed, positive usages were flagged, and then cancelled out in the presence of negative usages. NLP algorithms based on rules and regular expressions were iteratively improved to include more rules and search terms. Our NLP algorithms were manually adjudicated by a domain expert and iteratively revised until PPV $\geq$ 80% was achieved.

Figure II. Example R script for finding common long phrases containing search term.

```
library(quanteda) # An R package for the quantitative analysis of textual
data
library(corpus) # An R package for term_statistics
library(dplyr) # An R package for data manipulation

### set working directory
setwd(...)

#### load transformed text file
load(file="docs.transf.RData")

#### quanteda - find phrases containing the keywords from entire corpus

docs.transf.quanteda <- quanteda::corpus(docs.transf)

doc.tokens.1 <- quanteda::tokens(docs.transf.quanteda, remove_numbers = F,
remove_punct = F,
         remove_symbols = T, remove_separators = T, ngrams = 1, skip = 0L,
         concatenator = " ")
#### find keywords in context around thrombus/thrombi
#### window of 30 was chosen to help capture long 1-2 sentences, rather than
paragraphs, around search term
contexts <- as.data.frame(quanteda::kwic(doc.tokens.1, pattern = "thromb*",
window=30, valuetype = "glob"))
#### search for phrases with other keywords in search term
contexts1 <- contexts %>% filter(grepl("left ventr",pre) | grepl("left
ventr",post) |
         grepl("left ventr",pre) | grepl("left ventr",post))
View(contexts1)

#### data manipulation for finding phrases containing search term in a
sentence
contexts2 <- contexts1 %>%
 mutate(new_pre = gsub("\\s*.*\\.","",pre)) %>%
 mutate(new_post = gsub("\\s*\\..*","",post)) %>%
 filter(!grepl("left atr",new_pre), !grepl("left atr",new_post),
   !grepl("laa",new_pre), !grepl("laa",new_post),
   !grepl("appendage",new_pre), !grepl("appendage",new_post)) %>%
 select(new_pre, keyword, new_post) %>%
 mutate(text=paste(new_pre,keyword,new_post)) %>%
 select(text) %>%
 arrange(text)

#### estimate the best length(s) of phrases containing search terms
a1 <- c()
for(i in 1:40) {
 a <- corpus::term_stats(contexts2, ngrams = i, subset = (grepl("thromb",
term)), min_count = 2)
 a1 <- c(a1,nrow(a))
}
plot(a1,
```

```
  main="Length of Phrases Containing Keyword vs. Number of Phrases Occurring
>=2",
  xlab = "Length of Phrases Containing Keyword",
  ylab = "Number of Phrases Occurring >=2")
which.max(a1)

#### from plot a1, we estimate the best length(s) of phrases are around 8 to
13
## a0 is a frequency table of the most common phrases of length 8 to 13
a0 <- corpus::term_stats(contexts2, ngrams = 8:13, subset = (grepl("thromb",
term)), min_count = 2, types = F)
View(a0)
## a1 is a frequency table of the most common phrases of length 8 to 13,
including columns for each word in the phrase
a <- corpus::term_stats(contexts2, ngrams = 8:13, subset = (grepl("thromb",
term)), min_count = 2, types = T)

#### compute phrase length
#### compute group number to group phrases by first two words
a2 <- a %>% mutate(term_length = nchar(str_squish(a$term))) %>%
 mutate(group_no = as.integer(factor(paste(type1,type2))))
View(a2)

#### group phrases by first two words (can change two),
#### find longest phrase in each group,
#### generate frequency table for long phrases to find most common long
phrases
a3 <- a2 %>% arrange(group_no,desc(term_length)) %>%
 group_by(group_no) %>%
 filter(row_number()==1) %>%
 select(-starts_with("type")) %>%
 arrange(desc(count))

View(a3)

#### visual inspection of contexts2 and a3
#### informs us to search for the following phrases
#### for positive rules:
# left ventricular thrombus
# "left ventr", "thrombus" in same sentence
# thrombus is present within the left ventricular apex
# left ventricular apex consistent with thrombus
# echoes within the ventricular apex c / w thrombus
# lv thrombus
# thrombus in the left ventricle

#### for negation rules:
# left ventricular thrombus can not be excluded
# no obvious evidence of left ventricular thrombus
# no obvious lv thrombus
```

`docs.transf` was a character vector wherein each element contains a cardiac report. Quanteda's kwic() function extracted the window of words around a keyword in a search term.

Data manipulation was applied to find the entire search term from the extraction. Then we graphed the number of unique phrases containing the keyword over the length of the phrase (n) containing the keyword, where the phrase occurs at least twice. The shortest phrase containing the keyword was the keyword itself, so the number of unique phrases, occurring at least twice, containing the keyword would be minimal (close to 1). The longest phrase containing the keyword would be a very long sentence with extraneous information; however, they do not occur more than once, so they would be omitted and the total number of unique phrases would be close to minimal. The option "min_count" in the Corpus's term_stats() function excluded phrases with extraneous information such that such phrases would not occur more than once.

Between 1 and large n, there was an optimal length of phrase containing the keyword. At this length, there would be greater and more various information within each phrase, so that the number of unique phrases would be high. Each feature we developed had its own the optimal length of phrase. The graph was used to find the optimal length of phrase, for which there would be a peak of the number of unique phrases containing the keyword. Based on the peak in the graph, a range of lengths around the optimal length, say 8 to 13 in the above example, was used to compute the frequency table of the most common phrases of the range lengths (8 to 13).

While the most common phrases were found, many of the common phrases were very alike, since they may differ by a word or two. Our approach to this problem was to group the most common phrases by their first two words and then find the longest phrase within each grouping. Finally, we obtained a frequency table for long and popular phrases containing the search term. Based on this table, regular expressions were devised. Iteratively testing the algorithms helped identify additional positive, neutral, and negative rules.

Figure III. Left ventricular ejection fraction (LVEF) regular expressions algorithm.

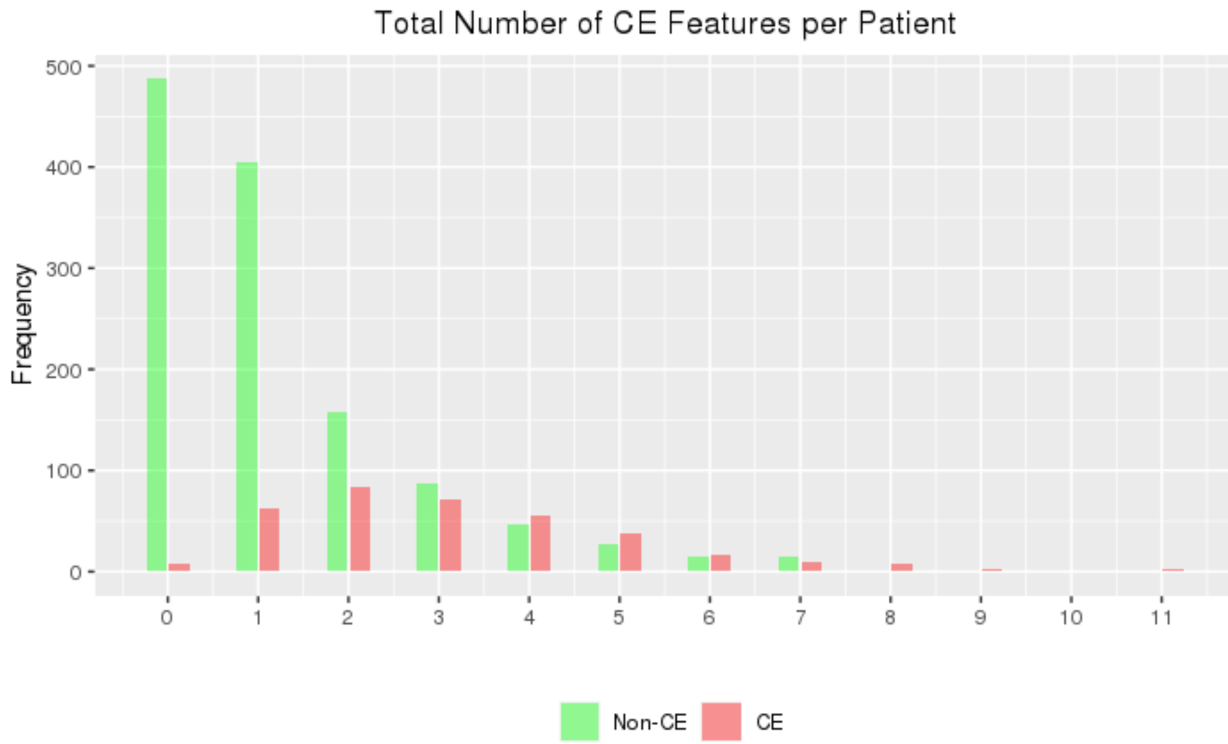
```
library(stringr)

#### LVEF code
pat <- "LV\\s?(\\w+\\s){0,4}EF\\s(\\w+\\s){0,4}\\d+|LV\\s?EF (\\w+\\s){0,4}is
(\\w+\\s){0,4}\\d+\\s?(-|to)\\s?\\d+|LV\\s?EF at
\\d+|((L?V?EF:*|Ejection\\s+fraction\\s+is|ejection\\s+fraction\\s+of|ejectio
n\\s+fraction\\s+is|ejection\\s+fraction\\s+is \\D* at|The calculated LVEF
\\D*is\\D*|LVEF is|LVEF estimated at|left\\s+ventricular function
is|ejection\\s+fraction\\s+is \\D* of)\\s+(\\S+)\\s*\\%)"
b <- stringr::str_extract(corpus.txt, pat) # find phrases of "LVEF"
p <- gsub("'", "", b, fixed=F)
d <- gsub(":", "", p, fixed=F)
e <- gsub(",", "", d, fixed=F)
f <- gsub("[a-z]+", "", e, fixed=F, ignore.case=T)
g <- gsub("%", "", f, fixed=F)
n = 5
h = substr(g,(nchar(g)+1)-n,nchar(g))
h <- trimws(g, "both")
i <- strsplit(h, "-")
j <- lapply(i, as.numeric)
k <- lapply(j, mean, na.rm=T) #take mean when \\1 group measurement or range
(i.e., "45-50%")
z <- as.numeric(unlist(k))
car$LVEF <- z
car$LVEF[which(car$LVEF == 0)] <- NA
sum(is.na(car$LVEF))

# if LVEF <= 40%, then hypokinetic_left_ventricular_segment present
ret <- as.numeric(z <= 40)
```

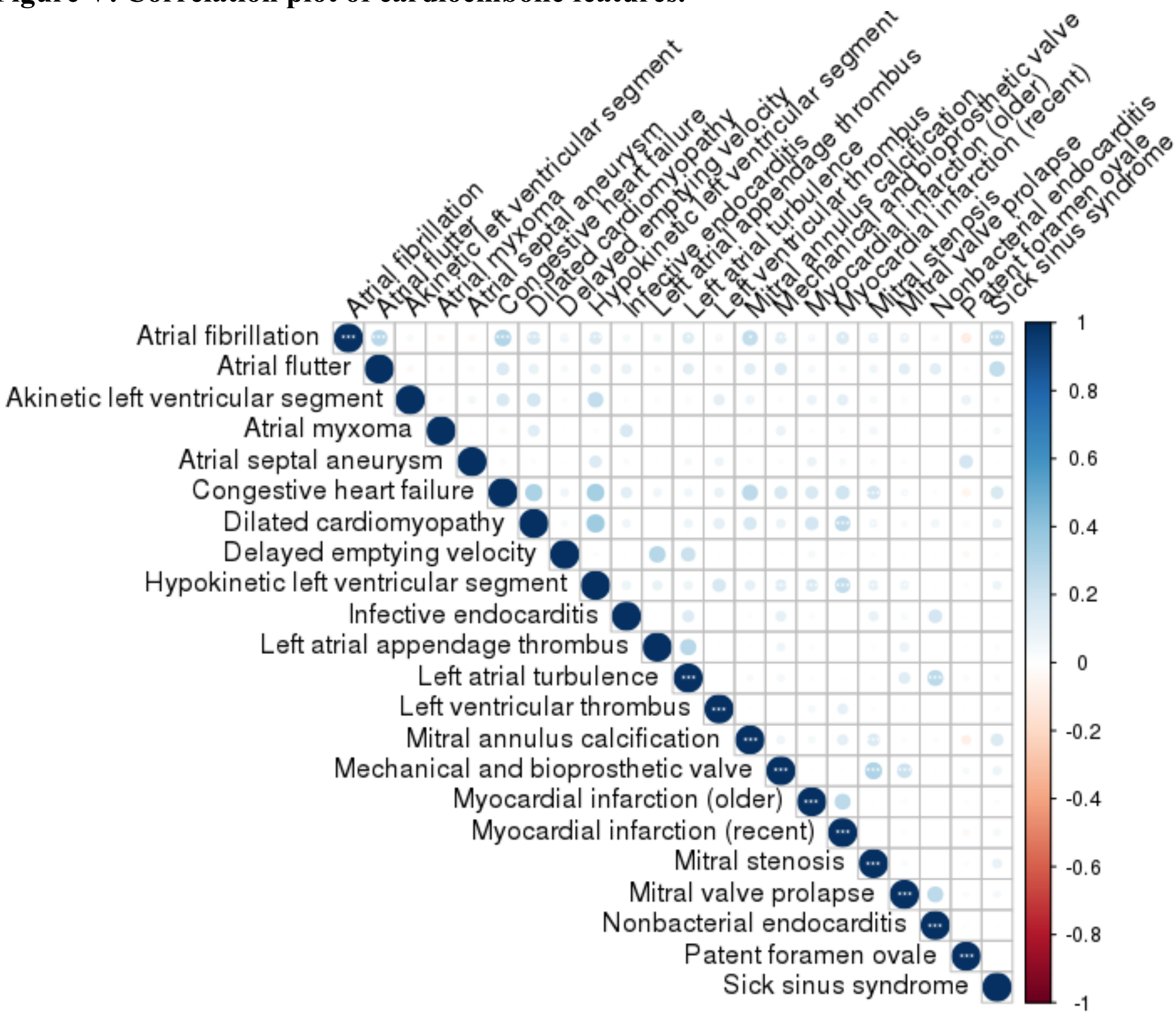
Previous research has developed a regular expressions algorithm for extracting left ventricular ejection fraction.[1] Extracted left ventricular ejection fraction quantities less than or equal to 40 qualifies as hypokinetic left ventricular segment (also called "reduced ejection fraction").

Figure IV. Total number of cardioembolic stroke features per patient.



Histogram of the total number of TOAST cardioembolic stroke features per patient. The histogram shows that cardioembolic stroke patients tend to have a greater number of TOAST cardioembolic stroke features per patient meanwhile non-cardioembolic stroke patients tend to have 0-2 total number of TOAST cardioembolic stroke features.

Figure V. Correlation plot of cardioembolic features.



Correlation plot of TOAST cardioembolic features. Pairs of variables with correlation greater than 0.3 were Atrial fibrillation and Cardiac heart failure; Cardiac heart failure and Hypokinetic left ventricular segment; Dilated cardiomyopathy and Hypokinetic left ventricular segment; Mitral annulus calcification and Mitral stenosis; and Mechanical and bioprosthetic valve and Myocardial infarction (recent). Correlation matrix did not show that any pair of features having correlation greater than 0.4.