Published in final edited form as:
Stroke. 2021 January ; 52(1): 181–189. doi:10.1161/STROKEAHA.120.030663.

Automated Electronic Phenotyping of Cardioembolic Stroke

Wyliena Guan, MA, MS[1,*], Darae Ko, MD, MSc[2,*], Shaan Khurshid, MD[1], Ana T. Trisini Lipsanopoulos, BS[1], Jeffrey M. Ashburner, PhD, MPH[3], Lia X. Harrington, PhD[1], Natalia S. Rost, MD, MPH[4], Steven J. Atlas, MD, MPH[3], Daniel E. Singer, MD[3], David D. McManus, MD[5], Christopher D. Anderson, MD, MMSc[4,6], Steven A. Lubitz, MD, MPH[1]

[1]Cardiovascular Research Center and Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston, MA, United States

[2]Section of Cardiovascular Medicine, Boston University School of Medicine, Boston, MA, United States

[3]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, United States

[4]Department of Neurology, Massachusetts General Hospital, Boston, MA, United States

[5]Division of Cardiovascular Medicine, University of Massachusetts Medical School, Worcester, MA, United States

[6]Henry and Allison McCance Center for Brain Health, Massachusetts General Hospital, Boston, MA, United States

Abstract

Background and Purpose: Oral anticoagulation is generally indicated for cardioembolic (CE) strokes, but not for other stroke etiologies. Consequently, subtype classification of ischemic stroke is important for risk stratification and secondary prevention. Because manual classification of ischemic stroke is time-intensive, we assessed the accuracy of automated algorithms for performing CE stroke subtyping using an electronic health record (EHR) database.

Methods: We adapted Trial of Org 10172 in Acute Stroke Treatment (TOAST) features associated with CE stroke for derivation in the EHR. Using administrative codes and echocardiographic reports within Mass General Brigham Healthcare Biobank (N = 13,079), we iteratively developed EHR-based algorithms to define the TOAST CE stroke features, revising regular expression algorithms until achieving positive predictive value (PPV) ≥ 80%. We compared several statistical algorithms for discriminating CE stroke using the feature algorithms applied to EHR data from 1598 patients with acute ischemic strokes from the Massachusetts

Supplemental Materials
Supplemental Methods
Supplemental Tables I–IX
Supplemental Figures I–V

Corresponding author: Steven A. Lubitz, MD, MPH; Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, 55 Fruit Street, GRB 109, Boston, MA 02114; P:617-643-7339; F:617-726-3852; slubitz@mgh.harvard.edu.

[*]Jointly contributed to the manuscript

General Hospital (MGH) Ischemic Stroke Registry (2002-2010) with previously adjudicated TOAST and Causative Classification of Stroke (CCS) subtypes.

Results: Regular expression-based feature extraction algorithms achieved a mean PPV of 95% (range 88-100%) across 11 echocardiographic features. Among 1598 patients from the MGH Ischemic Stroke Registry, 1068 had any CE stroke feature within pre-defined time windows in proximity to the stroke event. CE stroke tended to occur at an older age, with more TOAST-based comorbidities, and with atrial fibrillation (82.3%). The best model was a random forest with 92.2% accuracy and area under the receiver operating characteristic curve (AUC) of 91.1% (95% CI 87.5% - 93.9%). Atrial fibrillation, age, dilated cardiomyopathy, congestive heart failure, patent foramen ovale, mitral annulus calcification, and recent myocardial infarction were the most discriminatory features.

Conclusions: Machine-learning based identification of CE stroke using EHR data is feasible. Future work is needed to improve the accuracy of automated CE stroke identification and assess generalizability of electronic phenotyping algorithms across clinical settings.

Keywords

stroke; classification; embolism; thrombosis; anticoagulant; electronic health records; natural language processing; cardiovascular disease; cerebrovascular disease/stroke; ischemic stroke; risk factors; information technology

Introduction

Ischemic stroke is one of the leading causes of death in the US and is a major cause of serious disability.[1] Understanding different stroke etiologies and classifying strokes into their etiologic subtypes is critical for implementing effective secondary prevention strategies. In particular, distinguishing cardioembolic (CE) from non-CE stroke has important implications for management, since anticoagulation therapy is generally indicated for individuals with CE stroke to prevent recurrent events.[2]

Despite the importance of ischemic stroke subtyping, manual subtyping using classification systems such as the Trial of Org 10172 in Acute Stroke Treatment (TOAST)[3] is time-consuming and limited by uncertainty at the time of initial stroke presentation.[4] Moreover, ischemic stroke subtyping requires assessment of historical details, clinical findings, laboratory tests, electrocardiography, and imaging results by expert reviewers, making manual classification within large clinical databases laborious and unscalable. If sufficiently accurate, automated phenotyping of ischemic stroke subtypes using a limited set of electronic health record (EHR) data may facilitate stroke research efforts and augment clinical decision-making.

Methods

Data Sharing

Study data are available from the corresponding author upon reasonable request.

Approach overview

Our method consists of 2 major steps (Figure 1). First, we defined TOAST features associated with CE stroke and modified the CE stroke features to enable their derivation in the EHR. We then developed and internally validated algorithms to define presence of a CE stroke feature within EHR using administrative codes and free texts of echocardiogram reports. Second, we used the EHR-based CE stroke feature algorithms to determine association between presence of a CE stroke feature and clinician-adjudicated CE stroke. We used machine learning methods to build a classifier for CE stroke.

Definition of CE stroke based on the TOAST criteria

The TOAST classification system provides criteria for classifying ischemic stroke to be one of five subtypes: 1) large artery, 2) lacunar, 3) cardioembolism, 4) stroke of other determined origin, and 5) stroke of undetermined etiology.[3] The criteria for CE stroke considers a list of potential cardiac sources of embolic stroke. We made several practical modifications to the TOAST CE stroke features to enable their derivation in EHR. First, we combined "mechanical prosthetic valve" and "bioprosthetic cardiac valve" into "mechanical or biological prosthetic valve" *a priori* given concerns that billing codes lack adequate resolution to discriminate between the two features. Second, we considered mitral stenosis and atrial fibrillation as separate features, rather than using the categories, "mitral stenosis with atrial fibrillation" and "mitral stenosis without atrial fibrillation", as originally described in TOAST. Third, we subordinated "lone atrial fibrillation" to "atrial fibrillation" given our concern about the clinical validity of lone atrial fibrillation as a distinct condition.[5] Fourth, we combined patent foramen ovale and atrial septal defects since both have the potential to facilitate paradoxical emboli. The final set of our CE stroke features included both the high and medium-risk features from the TOAST criteria with the aforementioned modifications and include the following: mechanical or bioprosthetic valve, mitral stenosis, mitral valve prolapse, mitral annulus calcification, left atrial appendage thrombus, left atrial turbulence, sick sinus syndrome, recent myocardial infarction (<4 weeks prior to or after stroke), myocardial infarction (>4 weeks, <6 months after stroke), left ventricular thrombus, dilated cardiomyopathy, congestive heart failure, akinetic left ventricular segment, hypokinetic left ventricular segment, atrial myxoma, infective or nonbacterial thrombotic endocarditis, atrial septal aneurysm, and patent foramen ovale or atrial septal defect (Table 1, Supplemental Table I). We did not attempt to distinguish between medium and high-risk TOAST CE features in analyses.

Extraction of CE stroke features from EHR

We utilized the Mass General Brigham Biobank[6] to iteratively develop algorithms for EHR-based ascertainment of TOAST CE stroke features. The Mass General Brigham Biobank comprises EHR data from 30,716 volunteers. For the current study, we collected longitudinal EHR data using the Research Patient Database Repository[6] from 13,158 patients enrolled in the Mass General Brigham Biobank with echocardiogram reports as of December 2018. The Mass General Brigham Institutional Review Board approved all study activities. Informed consent was obtained from all subjects, their legally authorized representatives, or waived via protocol-specific allowance.

We derived CE stroke features from administrative codes consisting of International Classification of Disease (ICD) 9 or10 codes and Current Procedural Terminology codes (Supplemental Table II) as well as from free texts of echocardiogram reports. For features based on free texts, we seeded the algorithms with clinically informed language from a domain expert (D.K., S.A.L.) and then applied natural language processing (NLP) text mining and regular-expression methods to the entire corpus of available echocardiograms for feature extraction (Supplemental Table III). Text-mining focused on identifying informative words and text strings (which we refer to as "phrases") in reports that would indicate the presence of a TOAST CE stroke feature. Using R to partially imitate the functionality of Voyant Tools,[7] our methodology can efficiently summarize textual descriptions of clinical features from tens of thousands of documents into about 50 representative phrases useful for devising regular expressions logic for feature extraction.

We used clinical domain knowledge to facilitate the development of positive, negation, and neutral rules as well as the development of additional logic to search for phrases under specific report sections to indicate the presence of a CE stroke feature. Subsequently, we manually reviewed 50 reports for which the algorithm indicated the presence of a modified TOAST CE stroke feature, for a total of 550 charts (50 charts x 11 features). We then iteratively revised algorithm components and repeated the sampling in 50 independent charts to achieve a positive predictive value of ≥ 80% for each feature, based on expert review (D.K., S.A.L.). Test characteristics for the NLP-derived feature algorithms are summarized in the Supplemental Table IV. R version 3.4 and R packages 'stringr', 'lubridate', 'tokenizers', 'corpus', 'tm', and 'quanteda' were used for the development of the NLP regular expressions algorithms. The R scripts used to create a clean data corpus and extract features associated with CE stroke are available as open source (https://github.com/sag129/cardioembolic_stroke_subtyping).

Classification of CE stroke

We used the MGH Ischemic Stroke Registry[8–11] to build a classifier for CE stroke using machine learning methods. The MGH Ischemic Stroke Registry is a prospective hospital-based observational registry consisting of patients with acute ischemic stroke. We included only individuals in whom stroke subtypes were adjudicated. Stroke adjudication was performed retrospectively according to the TOAST criteria or CCS system[12] by trained neurologists independent of the treating physicians, utilizing all available clinical data at the time of delayed adjudication including continuous inpatient telemetry monitoring, echocardiograms, and other diagnostic testing. For the present study, we included 1,598 patients from the MGH Ischemic Stroke Registry with events that occurred between 2002 and 2010, of which 1,468 patients were adjudicated using TOAST. When TOAST adjudications were missing, CCS adjudication was used (N=130), which has previously been shown to be highly correlated with TOAST.[9] To maximize the validity of our algorithm, we considered CE strokes as those classified as "Definite Cardioembolic" by TOAST, or "Cardio-Aortic Embolism Evident" or "Cardio-Aortic Embolism Probable" by CCS criteria in our primary analysis. Strokes that were classified as "Possible Cardioembolic" by TOAST criteria or "Cardio-Aortic Embolism Possible" by CCS criteria as well as all other strokes not meeting the criteria for CE stroke were considered as non-CE stroke. Supplemental

Table V describes the number of patients in the MGH Stroke Registry observed to have one of various types of echocardiogram reports (N with at least any echocardiogram = 1491, 93.3%).

Statistical Analysis

We plotted the frequencies and intersections of CE stroke features among individuals with and without CE strokes using UpSet plots.[13] Machine learning classification methods utilized to distinguish between CE stroke and non-CE stroke (*i.e.*, large artery atherosclerosis, lacunar, other, and undetermined) included univariable logistic regression, multivariable logistic regression (Supplemental Table VIII), logistic regression regressed against indicator predictors counts (≥1, ≥2, and ≥3 features), logistic regression regressed against the count of features present during stroke, K-nearest neighbors (KNN), Support vector machines (SVM), Classification and Regression Tree (CART) decision tree, and Random Forest (RF) approaches.

We performed bootstrapping 50 times to estimate each model's performance in terms of accuracy, positive predictive value, negative predicted value, sensitivity, specificity, F1 score, and C-statistic (*i.e.*, AUC). In each bootstrap iteration, the data were randomly split into 70% for training and 30% for testing. To prevent overfitting, 5-fold cross-validation was performed and repeated 10 times within the training dataset. The best performing classification model was chosen based on accuracy and AUC. All analyses were performed using the R statistical and programming language v.3.4.0, including packages 'dplyr', 'UpSetR', 'caret', 'e1071', and 'ROCR'.

Results

Among the 30,716 individuals in the Mass General Brigham Biobank, we observed 3,144 transesophageal echocardiogram reports, 27,965 transthoracic echocardiogram reports, and 19,518 echocardiogram reports that were not clearly labeled as transthoracic or transesophageal, for a total of 50,627 echocardiographic reports spanning 13,079 unique individuals. The mean age among those with an echocardiogram was 61.5 ± 15.7 years, 48.9% were women, and 2,058 (15.7%) had at least one ICD9 or 10 diagnosis code stroke. The performance of the CE stroke feature algorithms is depicted in the Supplemental Table IV, with a mean positive predictive value of 95% (range 88-100%).

Baseline characteristics of the 1,598 patients from the MGH Ischemic Stroke Registry are described in Table 2. The mean age was 66.9 years and 61.9% were men. CE stroke (32.6% CE vs. 67.4% non-CE) tended to occur at an older age (73.0 ± 13.1 vs. 64.0 ± 14.2), concurrently with a greater number of TOAST CE stroke features (Supplemental Figure IV). As expected, atrial fibrillation was the most common source for CE stroke, and it was more common among individuals with CE than non-CE stroke (82.3% vs. 13.3%) (Figure 2).

Performance of the various models tested is shown in Figure 3. The best performing model was a random forest, with an accuracy of 92.2% (95% CI 89.7% - 94.2%) and AUC of 91.1% (95% CI 87.5% - 93.9%). SVM, Logistic Regression, and CART similarly performed well with accuracies ≥ 87% and AUC ≥ 85%. Results from multivariable logistic regression show that atrial fibrillation, hypokinetic left ventricular segment, and infective endocarditis were associated with increased likelihood of CE stroke, whereas male sex and atrial septal aneurysm were associated with a lower likelihood (Supplemental Table VIII). In the random forest model, the most important features were atrial fibrillation and age as shown in Figure 4. Patterns of features present among individuals with and without CE stroke are displayed in Supplemental Figure IV.

Discussion

In a sample of over 1,500 acute ischemic stroke patients with gold-standard stroke subtyping by expert adjudicators, we observed that electronic phenotyping using diagnosis codes, procedure codes, and regular expression-based features extracted from echocardiogram reports was effective for identifying relevant features for discriminating CE from non-CE strokes. Our findings suggest that automated electronic subtyping of ischemic strokes as CE is feasible in large datasets using algorithms that utilize a parsimonious subset of medical record data.

Overall, the use of machine learning has grown more widespread in stroke subtyping research. Garg et al. 2019 used admission, progress, discharge vascular neurology notes, and radiology reports to classify ischemic strokes into 5 TOAST subtypes.[14] In contrast, our approach used different data types – ICD codes, procedure codes, and echocardiogram reports – which were not dependent on stroke neurologist assessments, to derive a set of features upon which machine learning classifiers were applied. Moreover, Garg et al. used an approach that combined both machine learning to identify relevant words/phrases most associated with improved stroke subtype classification and a principal components analysis to create highly informative features from combinations of word/phrase terms to create a feature set. Such complexity may minimize interpretability and portability to other data sets. We used pre-identified clinical features, applied clinical domain knowledge in extracting those features, and then performed machine learning classification to identify associations between features and stroke subtypes. Nevertheless, our study in tandem with that of Garg et al. demonstrates the feasibility of electronic phenotyping of stroke subtypes. Depending on the proposed application of an algorithm, methods might be selected which favor either accuracy or portability.

Our study has clinical and research implications. First, our findings further add to the growing body of literature suggesting that scalable solutions to electronic phenotyping of ischemic stroke may be achievable using EHR data. Ultimately, our efforts may enable extension of stroke subtyping beyond those with gold-standard labels. For example, there are thousands of additional ischemic strokes that have not had gold-standard phenotyping in our and other EHRs, to which electronic algorithms could be applied to improve sample sizes for stroke subtype-specific research. Our methods could also be extended to include additional stroke subtypes, such as large artery stroke, lacunar occlusion, and cryptogenic stroke. Such efforts may aid initiatives to understand the biological mechanisms of stroke subtypes, develop improved image detection algorithms for stroke subtyping, and facilitate research on health outcomes related to specific stroke subtypes.

Second, our findings suggest that decision support using electronic phenotyping algorithms for ischemic stroke may be feasible in the future. Recent research suggests it is feasible to apply machine learning models to provide rapid, highly accurate and consistent assessment of echocardiograms and radiology images.[15, 16] Incorporating output from artificial intelligence-assisted point-of-care classifiers might aid in the appropriate and efficient utilization of diagnostic resources (*e.g.*, implantable loop recorders) and therapeutic management of stroke survivors. Indeed, some data suggest that the use of forms of cardiac rhythm monitoring may be underutilized following strokes.[17] Future research is warranted to determine the potential utility of implementing automated stroke classification algorithms to augment clinical subtyping.

Our study had several limitations. The electronic algorithm developed in this study needs to be tested and validated in other study samples outside of the MGH Ischemic Stroke Registry. We used a relatively small sample (N = 1598) with gold-standard adjudications for algorithm validation. Moreover, our training set was sparse for some features (e.g. nonbacterial endocarditis) and this may limit the generalizability of our model to stroke datasets rich with such features. We acknowledge that our CE stroke subtyping requires data integrated over time and after 90 days following ischemic stroke. In our study, we included text from echocardiography reports, but did not include elements, such as neuroimaging reports which may identify other features of stroke subtypes, vascular imaging reports which may identify vascular stenoses, or laboratory data to identify hyperlipidemia. Moreover, further refinement of features based solely on ICD codes could minimize feature misclassification.

Conclusions

Our study demonstrates that automated electronic phenotyping can be utilized to ascertain features from a limited set of EHR data for discriminating CE from non-CE stroke subtypes. Our findings further demonstrate that feature extraction using NLP, and novel statistical classification techniques, may be utilized to create algorithms that may enable automated phenotyping. Future efforts are warranted to assess the portability of stroke subtyping algorithms such as the one outlined here in this study, the utility of such algorithms for facilitating stroke research in large-scale datasets, and their potential to augment clinical decision support.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Sources of Funding:

Dr. Ko is supported by American College of Cardiology Foundation/Merck Research Fellowship in Cardiovascular Diseases and Cardiometabolic Disorders. Dr. Lubitz is supported by NIH grant 1R01HL139731, and Drs. Lubitz and Anderson are supported by American Heart Association (AHA) 18SFRN34250007. Dr. Anderson is also supported by NIH grant R01NS103924 and the Massachusetts General Hospital Center for Neuroscience and Henry and Allison McCance Center for Brain Health. Dr. McManus is supported by R01HL126911, R01HL137734, R01HL137794, R01HL135219, R01HL136660, U54HL143541, and 1U01HL146382 from the National Heart, Lung and Blood Institute. Dr. Khurshid is supported by NIH T32HL007208. This work was supported by an investigator-initiated industry sponsored research award from Boehringer Ingelheim (BIPI QIE2016001).

Boehringer Ingelheim Pharmaceuticals, Inc (BIPI) had no role in the design, analysis or interpretation of the results in this study; BIPI was given the opportunity to review the manuscript for medical and scientific accuracy as it relates to BIPI substances, as well as intellectual property considerations.

Disclosures:

Dr. Lubitz receives sponsored research support from Bristol Myers Squibb / Pfizer, Bayer AG, Boehringer Ingelheim, and Fitbit; has consulted for Bristol Myers Squibb / Pfizer and Bayer AG; and participates in a research collaboration with IBM. Dr. Harrington participates in a research collaboration with Bayer AG. Dr. Anderson receives sponsored research support from Bayer AG and has consulted for ApoPharma, Inc. Dr. Singer receives research support from Bristol-Myers Squibb and has consulted for Boehringer Ingelheim, Bristol-Myers Squibb, Johnson and Johnson, Merck, and Pfizer. Dr. McManus receives sponsored research grant support from Bristol-Myers Squibb, Boehringher-Ingelheim, Pfizer, Flexcon, Fitbit, Philips Healthcare, Biotronik, has received consultancy fees from Bristol-Myers Squibb, Pfizer, Flexcon, Boston Biomedical Associates, Rose Consulting, Samsung, Heart Rhythm Society, and received non-financial support from Samsung and Apple. Dr. Atlas has received research support from Bristol Myers Squibb / Pfizer and Boehringer Ingelheim and has consulted for Fitbit. The remainder of the authors report no disclosures.

Abbreviations:

AUC	area under the receiver operating characteristic curve
CCS	Causative Classification of Stroke
CE	cardioembolic
EHR	electronic health record
ICD	International Classification of Disease
CPT	Current Procedural Terminology
MGH	Massachusetts General Hospital
NLP	natural language processing
PPV	positive predictive value
TOAST	Trial of Org 10172 in Acute Stroke Treatment

References

1. Heron m. Deaths: Leading causes for 2017. National vital statistics reports; vol 68, no 6 Hyattsville, md: National center for health statistics 2019.
2. Kernan WN, Ovbiagele B, Black HR, Bravata DM, Chimowitz MI, Ezekowitz MD, et al. Guidelines for the prevention of stroke in patients with stroke and transient ischemic attack: A guideline for healthcare professionals from the american heart association/american stroke association. Stroke. 2014;45:2160–2236 [PubMed: 24788967]
3. Adams HP Jr., Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. Toast. Trial of org 10172 in acute stroke treatment. Stroke. 1993;24:35–41 [PubMed: 7678184]
4. Adams HP Jr., Biller J. Classification of subtypes of ischemic stroke: History of the trial of org 10172 in acute stroke treatment classification. Stroke. 2015;46:e114–117 [PubMed: 25813192]
5. Wyse DG, Van Gelder IC, Ellinor PT, Go AS, Kalman JM, Narayan SM, et al. Lone atrial fibrillation. Does it Exist? 2014;63:1715–1723
6. Weiss ST, Shin MS. Infrastructure for personalized medicine at partners healthcare. J Pers Med. 2016;6
7. Sinclair S, Rockwell G. Voyant tools: Reveal your texts. 2016

8. Loci associated with ischaemic stroke and its subtypes (sign): A genome-wide association study. The Lancet. Neurology 2016;15:174–184 [PubMed: 26708676]

9. Ay H, Arsava EM, Andsberg G, Benner T, Brown RD Jr., Chapman SN, et al. Pathogenic ischemic stroke phenotypes in the ninds-stroke genetics network. Stroke. 2014;45:3589–3596 [PubMed: 25378430]

10. McArdle PF, Kittner SJ, Ay H, Brown RD, Meschia JF, Rundek T, et al. Agreement between toast and ccs ischemic stroke classification. The NINDS SiGN Study. 2014;83:1653–1660

11. Kim GM, Park KY, Avery R, Helenius J, Rost N, Rosand J, et al. Extensive leukoaraiosis is associated with high early risk of recurrence after ischemic stroke. Stroke. 2014;45:479–485 [PubMed: 24370756]

12. Arsava EM, Ballabio E, Benner T, Cole JW, Delgado-Martinez MP, Dichgans M, et al. The causative classification of stroke system: An international reliability and optimization study. Neurology. 2010;75:1277–1284 [PubMed: 20921513]

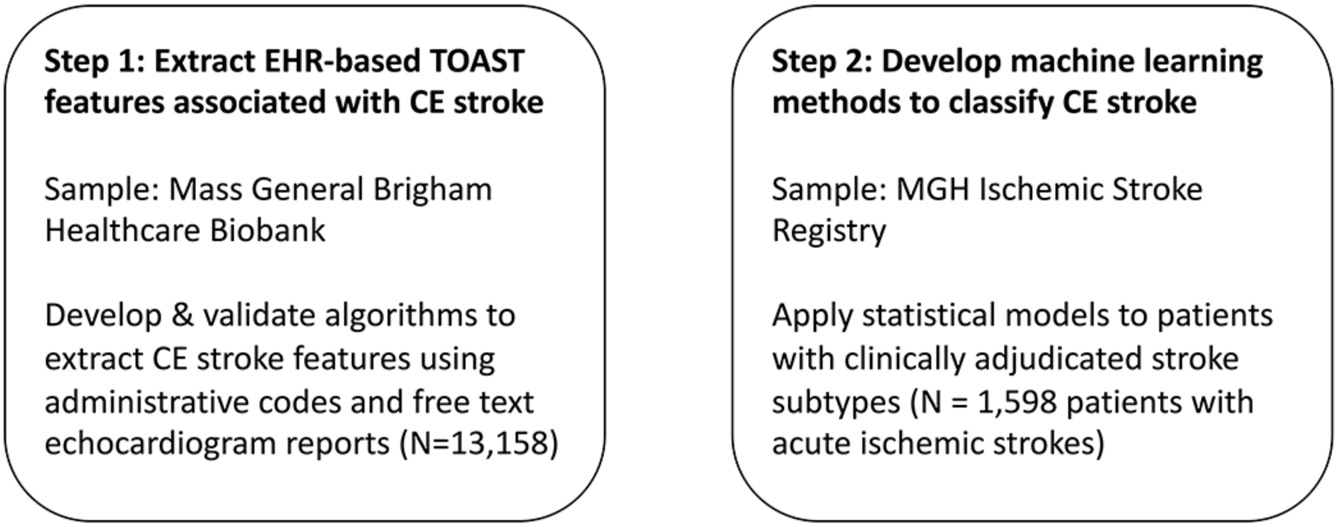
13. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. Upset: Visualization of intersecting sets. IEEE Trans Vis Comput Graph. 2014;20:1983–1992 [PubMed: 26356912]

14. Garg R, Oh E, Naidech A, Kording K, Prabhakaran S. Automating ischemic stroke subtype classification using machine learning and natural language processing. J Stroke Cerebrovasc Dis. 2019;28:2045–2051 [PubMed: 31103549]

15. Alsharqi M, Woodward WJ, Mumith JA, Markham DC, Upton R, Leeson P. Artificial intelligence and echocardiography. Echo Res Pract. 2018;5:R115–r125 [PubMed: 30400053]

16. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp. 2018;2:35 [PubMed: 30353365]

17. Lip GY, Hunter TD, Quiroz ME, Ziegler PD, Turakhia MP. Atrial fibrillation diagnosis timing, ambulatory ecg monitoring utilization, and risk of recurrent stroke. Circ Cardiovasc Qual Outcomes. 2017;10



Step 1: Extract EHR-based TOAST features associated with CE stroke

Sample: Mass General Brigham Healthcare Biobank

Develop & validate algorithms to extract CE stroke features using administrative codes and free text echocardiogram reports (N=13,158)

Step 2: Develop machine learning methods to classify CE stroke

Sample: MGH Ischemic Stroke Registry

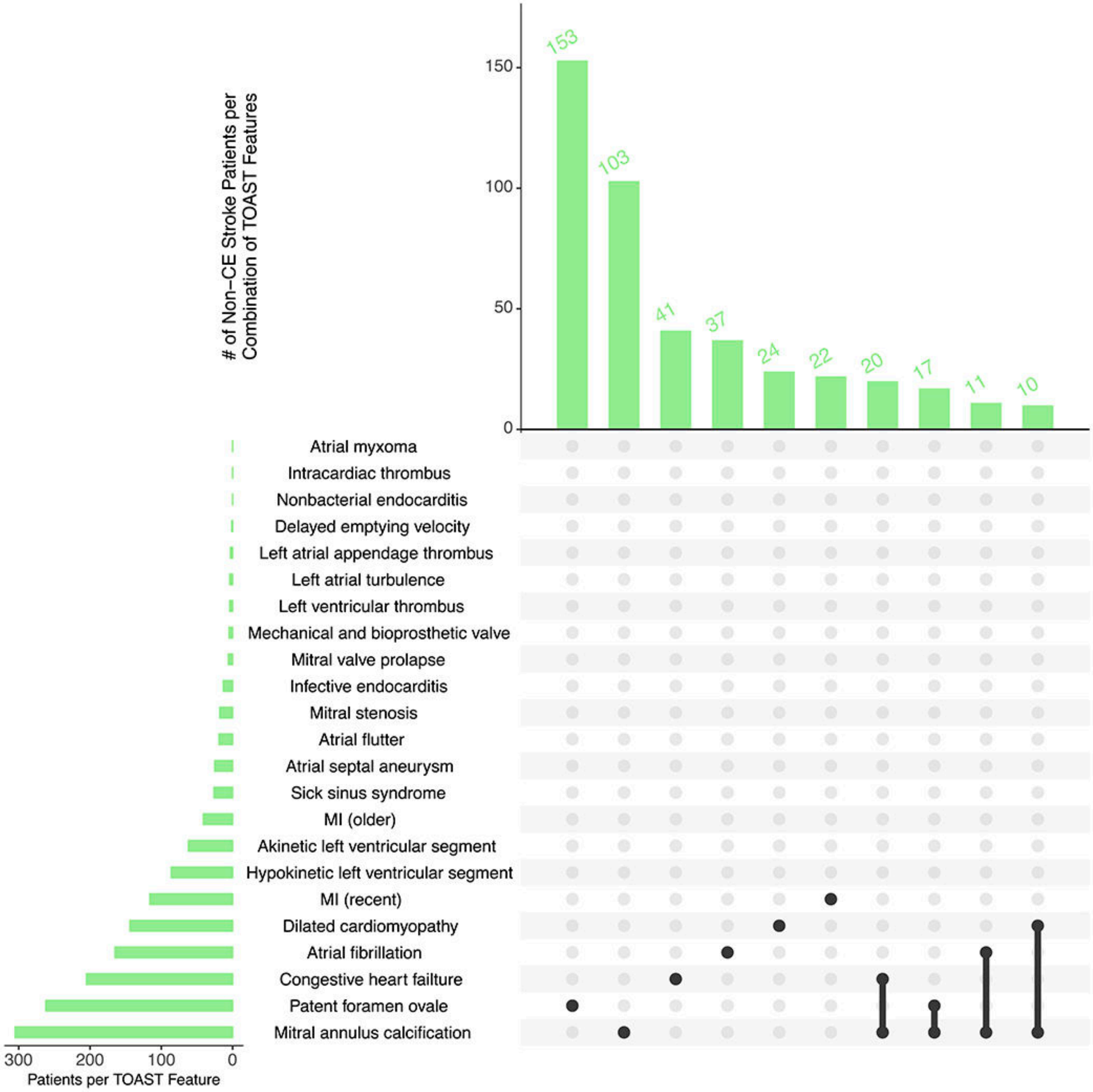
Apply statistical models to patients with clinically adjudicated stroke subtypes (N = 1,598 patients with acute ischemic strokes)

Figure 1.

Overview of electronic phenotyping for CE stroke

Figure 1 is a flow diagram delineating the sequential steps for our approach.

CE: cardioembolic; EHR: electronic health record; MGH: Massachusetts General Hospital



# of Non–CE Stroke Patients per Combination of TOAST Features
0
50
100
150
153
103
41
37
24
22
20
17
11
10
Atrial myxoma
Intracardiac thrombus
Nonbacterial endocarditis
Delayed emptying velocity
Left atrial appendage thrombus
Left atrial turbulence
Left ventricular thrombus
Mechanical and bioprosthetic valve
Mitral valve prolapse
Infective endocarditis
Mitral stenosis
Atrial flutter
Atrial septal aneurysm
Sick sinus syndrome
MI (older)
Akinetic left ventricular segment
Hypokinetic left ventricular segment
MI (recent)
Dilated cardiomyopathy
Atrial fibrillation
Congestive heart failture
Patent foramen ovale
Mitral annulus calcification
300
200
100
0
Patients per TOAST Feature

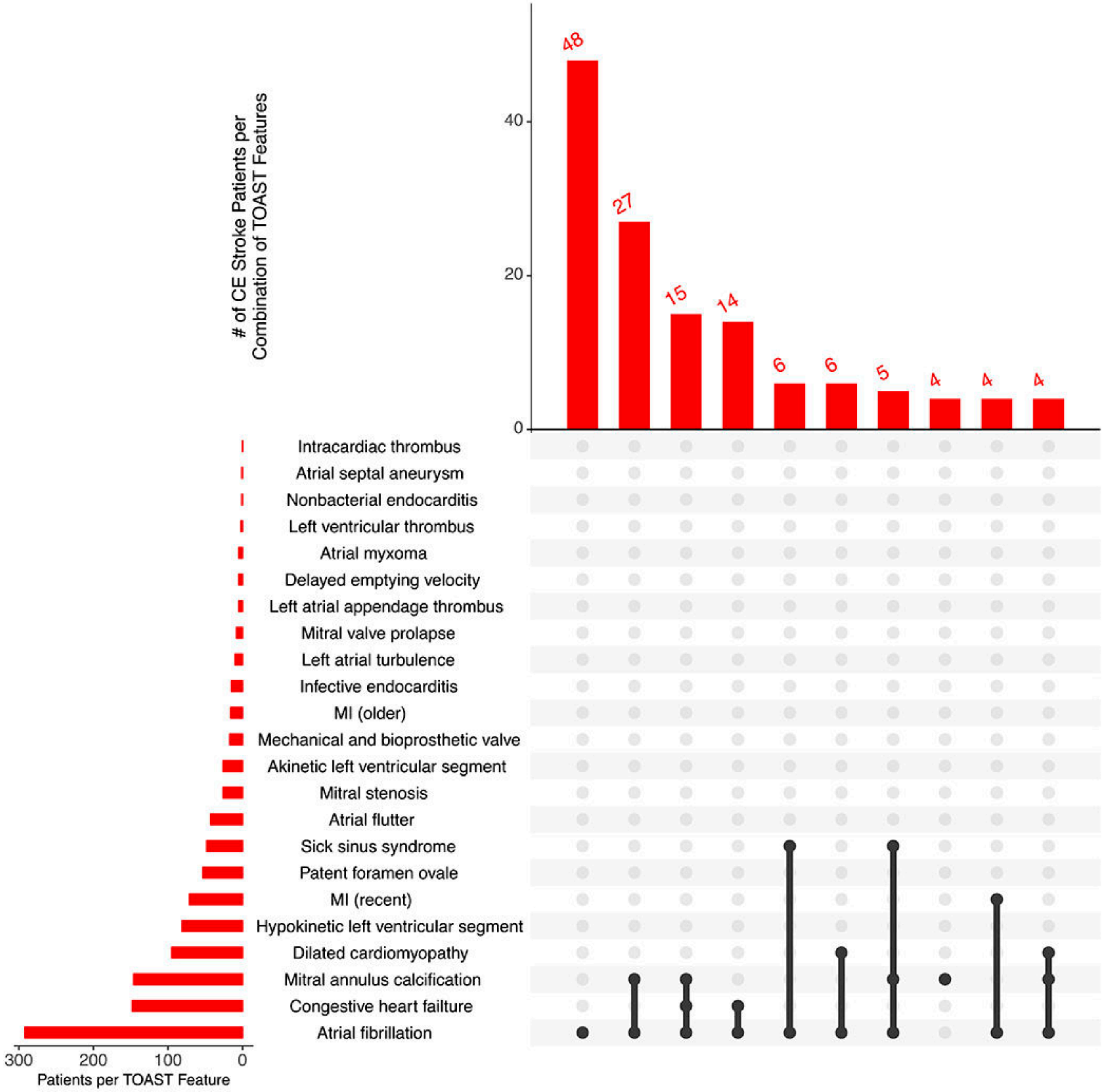


Figure 2.

Frequency of non-CE (A) and CE (B) strokes per combination of TOAST cardioembolic stroke features.

The top UpSet plot depicts overall frequency of the top features for non-CE stroke, and the bottom depicts overall frequency of the top features for CE stroke. Black lines connecting multiple features indicate the presence of a combination of features.

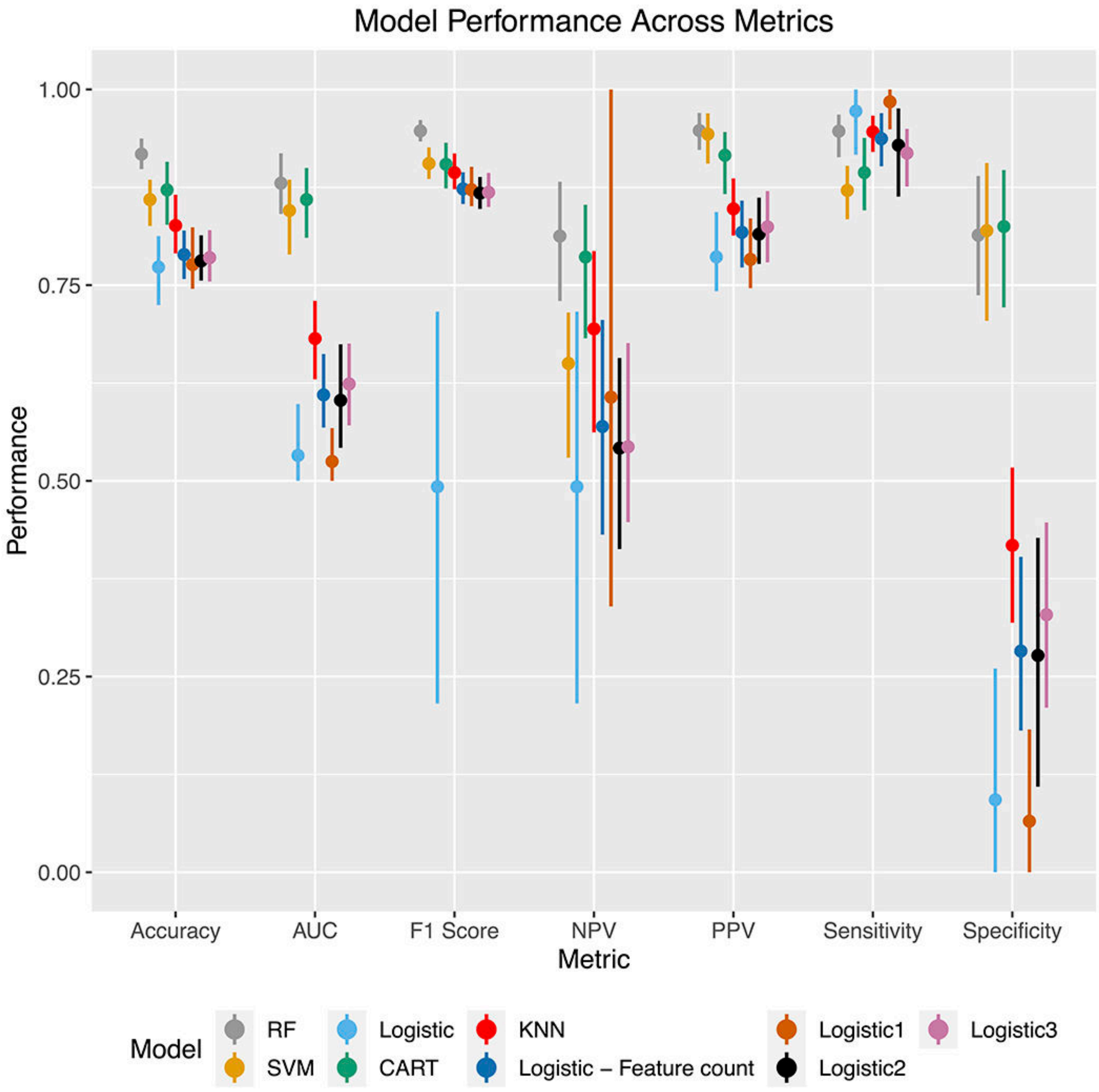


Figure 3.

Performance of various algorithms used for TOAST classification of cardioembolic stroke. Displayed from left to right are the algorithms with the best to worst performance based on accuracy. SVM denotes “Support Vector Machine”. CART denotes “Classification and Regression Tree”. KNN denotes “K Nearest Neighbors”. “Logistic regression” denotes multivariate logistic regression. “Logistic – feature count” denotes a logistic regression model regressed on the number of features present in the stroke event. “Logistic - ≥ 1 feature” to “Logistic - ≥ 3 feature” models respectively denote a logistic regression model regressed on a binary covariate indicating the presence of 1 or more features to 3 or more features.

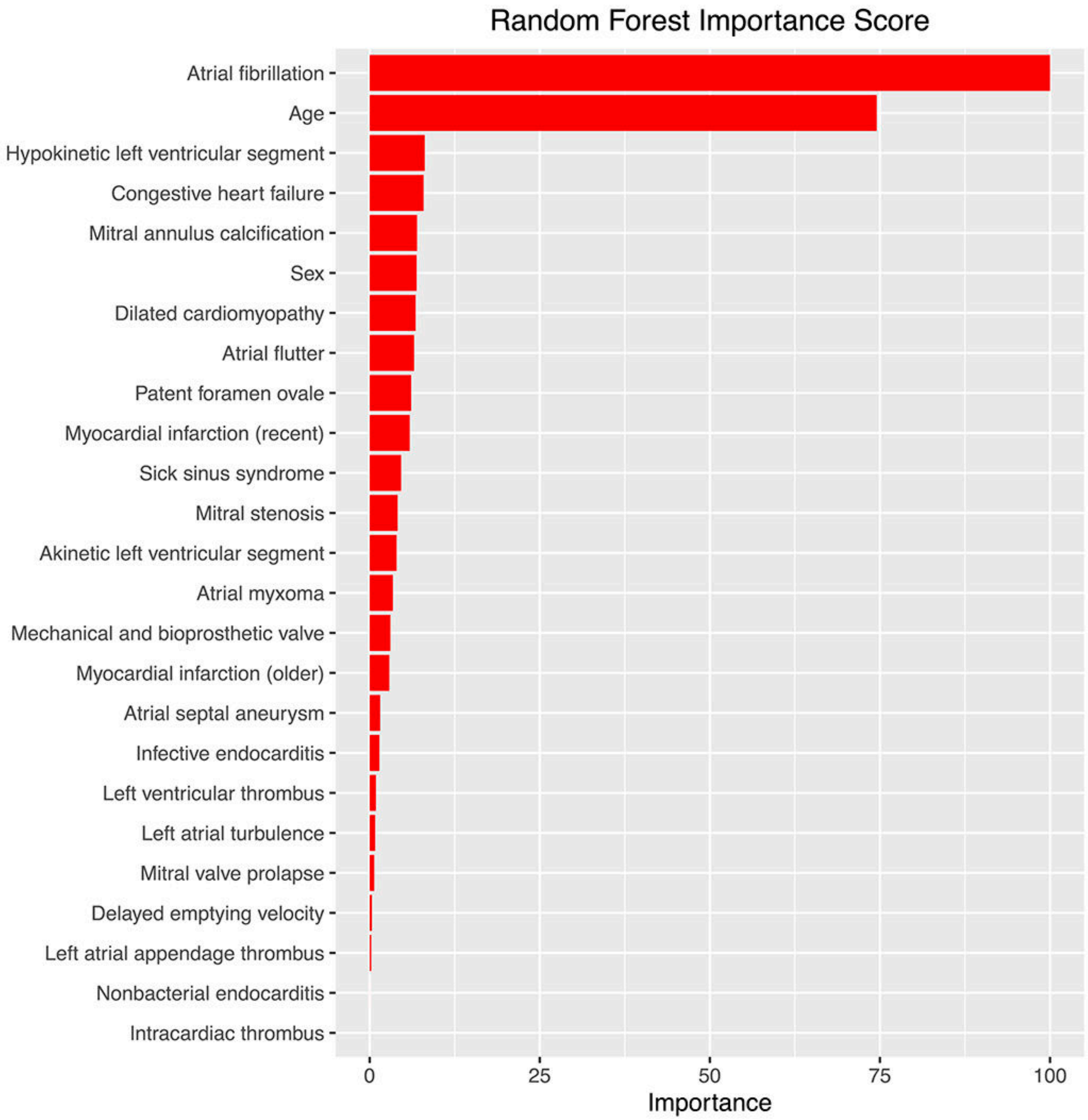


Figure 4.

Random forest importance scores identifying important variables for predicting TOAST cardioembolic stroke.

The best performing model was random forest, which showed that atrial fibrillation and age were the most important variables for discriminating cardioembolic stroke. Next most important were congestive heart failure, hypokinetic left ventricular segment, mitral annulus calcification, gender, and dilated cardiomyopathy.

Table 1.

Data sources and time window of cardioembolic stroke features

High-risk sources	**ICD codes**	**Procedure codes**	**Echocardiogram Report**
Mechanical or bioprosthetic cardiac valve[*]	✓	✓	
Mitral stenosis[*]	✓		
Atrial fibrillation (including lone atrial fibrillation)[*]	✓		
Left atrial/atrial appendage thrombus[†]			✓
Intracardiac thrombus[*]	✓		
Sick sinus syndrome[*]	✓		
Recent myocardial infarction (<4 weeks)[‡]	✓	✓	
Left ventricular thrombus[*]	✓		
Dilated cardiomyopathy[*]	✓		
Akinetic left ventricular segment[*]			✓
Atrial myxoma[*]	✓		
Infective endocarditis[*]	✓		
Medium-risk sources			
Mitral valve prolapse[*]	✓		✓
Mitral annulus calcification[*]	✓		✓
Left atrial turbulence (smoke)[†]			✓
Delayed emptying velocity[†]			✓
Atrial septal aneurysm[*]	✓		✓
Patent foramen ovale[*]	✓		✓
Atrial flutter[*]	✓		
Nonbacterial thrombotic endocarditis[†]	✓		
Congestive heart failure[*]	✓		
Hypokinetic left ventricular segment[*]			✓
Myocardial infarction[§]	✓	✓	

[*] Before or up to 90d after stroke;

[†] 90d before or after stroke;

[‡] <=4 weeks prior to stroke;

[§] between 6 months to 4 weeks prior to stroke

Table 2.

Baseline characteristics of the MGH ischemic stroke sample stratified by stroke mechanism

	Total sample	Non-CE stroke	CE stroke
Demographics	**N=1,598**	**N=1,243**	**N=355**
Male Sex	937 (58.6%)	755 (60.7%)	182 (51.2%)
Age	65.4 ± 15.4	63.3 ± 15.4	73.0 ± 13.1
TOAST CE stroke features			
Atrial fibrillation	457 (28.6%)	165 (13.3%)	292 (82.3%)
Atrial flutter	62 (3.9%)	19 (1.5%)	43 (12.1%)
Akinetic left ventricular segment	88 (5.5%)	62 (5%)	26 (7.3%)
Atrial myxoma	5 (0.3%)	0 (0%)	5 (1.4%)
Atrial septal aneurysm	26 (1.6%)	25 (2%)	1 (0.3%)
Congestive heart failure	353 (22.1%)	205 (16.5%)	148 (41.7%)
Dilated cardiomyopathy	239 (15%)	144 (11.6%)	95 (26.8%)
Delayed emptying velocity	6 (0.4%)	1 (0.1%)	5 (1.4%)
Hypokinetic left ventricular segment	167 (10.5%)	86 (6.9%)	81 (22.8%)
Infective endocarditis	28 (1.8%)	13 (1%)	15 (4.2%)
Intracardiac thrombus	0 (0%)	0 (0%)	0 (0%)
Left atrial appendage thrombus	8 (0.5%)	3 (0.2%)	5 (1.4%)
Left atrial turbulence	14 (0.9%)	4 (0.3%)	10 (2.8%)
Left ventricular thrombus	6 (0.4%)	4 (0.3%)	2 (0.6%)
Mitral annulus calcification	451 (28.2%)	305 (24.5%)	146 (41.1%)
Mechanical and bioprosthetic valve	22 (1.4%)	5 (0.4%)	17 (4.8%)
Myocardial infarction (later)	57 (3.6%)	41 (3.3%)	16 (4.5%)
Myocardial infarction (recent)	187 (11.7%)	116 (9.3%)	71 (20%)
Mitral stenosis	44 (2.8%)	18 (1.4%)	26 (7.3%)
Mitral valve prolapse	14 (0.9%)	6 (0.5%)	8 (2.3%)
Nonbacterial endocarditis	1 (0.1%)	0 (0%)	1 (0.3%)
Patent foramen ovale	315 (19.7%)	262 (21.1%)	53 (14.9%)
Sick sinus syndrome	74 (4.6%)	26 (2.1%)	48 (13.5%)

Data displayed as N (%) or Mean ± SD