

Original Paper

Natural Language Processing and Machine Learning for Identifying Incident Stroke From Electronic Health Records: Algorithm Development and Validation

Yiqing Zhao, PhD; Sunyang Fu, MHI; Suzette J Bielinski, PhD; Paul A Decker, MSc; Alanna M Chamberlain, PhD; Veronique L Roger, MD; Hongfang Liu, PhD; Nicholas B Larson, PhD

Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

Corresponding Author:

Nicholas B Larson, PhD
Department of Health Sciences Research
Mayo Clinic
205 3rd Ave SW
Rochester, MN, 55905
United States
Phone: 1 507 293 1700
Email: Larson.Nicholas@mayo.edu

Abstract

Background: Stroke is an important clinical outcome in cardiovascular research. However, the ascertainment of incident stroke is typically accomplished via time-consuming manual chart abstraction. Current phenotyping efforts using electronic health records for stroke focus on case ascertainment rather than incident disease, which requires knowledge of the temporal sequence of events.

Objective: The aim of this study was to develop a machine learning–based phenotyping algorithm for incident stroke ascertainment based on diagnosis codes, procedure codes, and clinical concepts extracted from clinical notes using natural language processing.

Methods: The algorithm was trained and validated using an existing epidemiology cohort consisting of 4914 patients with atrial fibrillation (AF) with manually curated incident stroke events. Various combinations of feature sets and machine learning classifiers were compared. Using a heuristic rule based on the composition of concepts and codes, we further detected the stroke subtype (ischemic stroke/transient ischemic attack or hemorrhagic stroke) of each identified stroke. The algorithm was further validated using a cohort (n=150) stratified sampled from a population in Olmsted County, Minnesota (N=74,314).

Results: Among the 4914 patients with AF, 740 had validated incident stroke events. The best-performing stroke phenotyping algorithm used clinical concepts, diagnosis codes, and procedure codes as features in a random forest classifier. Among patients with stroke codes in the general population sample, the best-performing model achieved a positive predictive value of 86% (43/50; 95% CI 0.74-0.93) and a negative predictive value of 96% (96/100). For subtype identification, we achieved an accuracy of 83% in the AF cohort and 80% in the general population sample.

Conclusions: We developed and validated a machine learning–based algorithm that performed well for identifying incident stroke and for determining type of stroke. The algorithm also performed well on a sample from a general population, further demonstrating its generalizability and potential for adoption by other institutions.

(*J Med Internet Res* 2021;23(3):e22951) doi: [10.2196/22951](https://doi.org/10.2196/22951)

KEYWORDS

stroke; natural language processing; electronic health records; machine learning

Introduction

Stroke is a syndrome involving a rapid loss of cerebral function with vascular origin [1]. The loss of function can result in deep coma or subarachnoid hemorrhage. There are two broad

categories of stroke: hemorrhagic and ischemic stroke [2]. Hemorrhage is caused by bleeding within the skull cavity, while ischemia is characterized by inadequate blood to supply a part of the brain. Stroke identification is an important outcome for various cardiovascular studies [3-5]. However, a challenge with stroke ascertainment is the inconsistent use of International

Classification of Diseases (ICD) codes [6], which may result in inaccurate code-based ascertainment of cases [7]. Therefore, the time-consuming process of electronic health record (EHR) abstraction remains the gold standard of stroke ascertainment [8,9].

Machine learning has recently gained popularity for its ability to classify patients or make predictions on various aspects of diseases. In contrast to manually curated algorithms based on domain expertise, machine learning is a data-driven approach that can be trained on large data sets to identify and leverage complex feature relationships and improve classification and prediction tasks thereby. In terms of stroke, machine learning algorithms have been applied to predict future stroke cases [10], mortality and recurrent strokes [11,12], and treatment outcomes [13,14]. Most existing phenotyping algorithms have been developed to only differentiate between cases and noncases of diseases [15-18]; however, ascertaining incident disease (ie, first occurrence of disease) in a population is a more difficult task [8,19,20]. A recent study by Ni et al [21] examined potential predictive features of stroke occurrence including demographic, clinical, and diagnostic characteristics of patients. The authors found that diagnostic tests for stroke, such as computed tomography (CT) and magnetic resonance imaging (MRI), contributed to most of the model performance, and that the optimal feature set included imaging findings, signs and symptoms, interventions, emergency department assessments, findings from angiography and carotid ultrasound tests, ICD codes, substance use (smoking, alcohol, and street drugs) characteristics, and demographics. However, features such as signs and symptoms, substance use characteristics, and demographics may not be specific enough for disease ascertainment, as there is a high prevalence of stroke-like symptoms among people without a diagnosis of stroke [22]. In addition, incorporating too many features in the model may result in overfitting without appropriate regularization. Another study [7] also used ICD and Current Procedural Terminology (CPT) [23] codes as features to classify positive, possible, and negative stroke cases. However, stroke-related clinical concepts (including both disease name concepts and symptom concepts) in unstructured clinical notes were not included in this model.

Rapid adoption of EHRs has enabled secondary use of the EHR data in epidemiological research [24-26]. Previous studies noted the existence of bias using a single type of EHR data (ie, diagnosis codes) [27-29]. To avoid this bias, the Electronic Medical Records and Genomics (eMERGE) consortium [30,31] has piloted the development of EHR-based phenotyping algorithms using multiple types of EHR data [32-34]. This has given rise to a number of phenotyping algorithms that use both structured EHR data (eg, demographics, diagnosis and procedure codes, laboratory test results, and medications) and unstructured EHR data (eg, clinical notes, imaging reports, and discharge summaries) [35-38]. However, the eMERGE consortium algorithms are typically focused on identifying cases and noncases rather than characterizing a new-onset (ie, incident) disease in a population. Moreover, extracting information from unstructured clinical text is a nontrivial task that involves natural language processing techniques [39-41].

In our paper, we address existing challenges for stroke ascertainment, specifically for incident stroke. Our research objective is to develop and validate a machine learning-based phenotyping algorithm to identify incident stroke and detailed stroke subtypes based on three major EHR-derived data elements: clinical concepts extracted from clinical notes; ICD, Ninth Revision (ICD-9) diagnosis codes; and CPT procedure codes.

Methods

This study was approved by the Mayo Clinic Institutional Review Board (no. 17-008818) and is in accordance with the ethical standards mandated by the committee on responsible human experimentation. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Study Design

This was a predictive modeling study that used observational cohort data for training and validation. We employed an atrial fibrillation (AF) cohort, in which all incidences of stroke were manually ascertained in a previous study [4], to train and test our phenotyping algorithm for the date of incident stroke events. We then evaluated the generalizability of our algorithm in a general population cohort.

The AF Cohort

The AF cohort comprised a patient population from Olmsted County, Minnesota, USA [4,42]. Olmsted County is an area relatively isolated from other urban centers with only a few providers delivering most care to residents, primarily Mayo Clinic and Olmsted Medical Center [43-45]. Extracting all health care-related events was completed through the Rochester Epidemiology Project (REP), a records linkage system [43,44]. The REP is a records linkage system that allows retrieval of nearly all health care utilization and outcomes of residents living in Olmsted County. The electronic indexes of the REP include demographic information, diagnostic and procedure codes, health care utilization data, outpatient drug prescriptions, results of laboratory tests, and information about smoking, height, weight, and body mass index. ICD-9 codes and the Mayo Clinic electrocardiograms were obtained among adults aged ≥ 18 years from 2000 to 2014 to ascertain AF. Patients were identified by the presence of an ICD-9 code for stroke through March 31, 2015, and then validated by manual review of the EHR. Strokes were classified as ischemic strokes/transient ischemic attack or hemorrhagic strokes [4,46]. The first (incident) event of each type of stroke after the incident AF date was ascertained, regardless of whether a patient had a prior stroke. The AF cohort included 4914 validated patients with AF, 1773 of whom were screened for a possible stroke. Table 1 shows the cohort characteristics. Manual abstraction of the EHR validated the stroke code in 740 patients. Manual ascertainment of stroke and the dates of the events were used as a gold standard to train and test the stroke algorithm.

Table 1. Atrial fibrillation cohort characteristics.

Measure	Cohort (n=4914)	Screened (n=1773)
Gender, n (%)		
Female	2309 (46.99)	869 (49.01)
Male	2605 (53.01)	904 (50.99)
Age at diagnosis of AF^a (years), mean		
Female	76	80
Male	70	74
ICD-9 ^b diagnosis codes ^c , n	27,243	27,243

^aAF: atrial fibrillation.

^bICD-9: International Classification of Diseases, Ninth Revision.

^cICD retrieval was from AF incidence date to March 31, 2015. AF validations were from 2000 to 2014.

Candidate Predictive Features

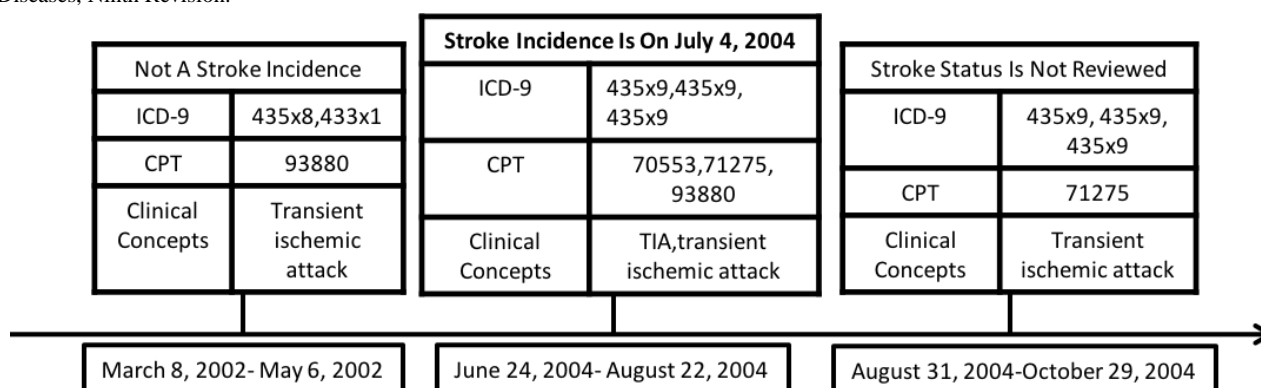
The proposed algorithm aimed to identify first (incident) stroke events within a certain time frame. The three major data elements we used were clinical concepts, ICD-9 codes, and CPT codes. To align with the manual review process, only codes and clinical notes from the AF incident date to March 31, 2015, were retrieved and processed. In our analyses, we constructed different models by varying the inclusion of CPT codes and symptom-related clinical concepts in the model feature set and compared different models' performances.

Both ICD-9 and CPT codes were extracted from the REP database. Clinical concepts were identified from the major and secondary problem list section of Mayo Clinic EHR, and from clinical notes from other REP sites using a natural language processing system, MedTagger [47]. Expert-provided vocabulary was adopted from a previous study [48] to extract clinical concepts from unstructured clinical notes. MedTagger enables a series of natural language processing processes, including regular expression matching and positive, negative, or probable identification with ConText [49,50], and is insensitive to upper and lower case. MedTagger is also able to determine if the extracted clinical concepts are referring to the patients or their family members, or if the extracted clinical concepts are in present tense and thus are referring to a current event rather than a past medical condition. We considered only documents with positive, present-tense stroke mentions that were referring

to patients themselves. Table S1 in [Multimedia Appendix 1](#) lists clinical concepts for 2 major stroke subtypes and stroke-related symptoms. Table S2 in [Multimedia Appendix 1](#) lists ICD-9 codes for 2 stroke subtypes and stroke-related symptoms. Table S3 in [Multimedia Appendix 1](#) lists the CPT codes used in the stroke algorithm.

Clinical concept dates were determined by the date of the clinical notes from which clinical concepts were extracted, while ICD-9 and CPT code dates were extracted from the REP. Each visit was characterized by clinical concepts, ICD-9, and CPT codes within a 60-day window. The visit date was determined by the earliest date of any of the 3 elements in the 60-day window. If visit dates were within a 60-day window of a confirmed stroke incidence date, they were considered positive instances; otherwise, they were considered negative instances. [Figure 1](#) demonstrates an example with an incident stroke on July 4, 2004. All visits were extracted and included in our data set if there was at least one key word or code during a 60-day window. Nurse abstractors reviewed every visit sequentially until they determined the incidence date to be July 4, 2004. All subsequent visits after a positive stroke incident were not reviewed and thus were not included in our analyses. Since the confirmed stroke incidence date fell in the date range of the third visit (June 24, 2004-August 22, 2004), we considered the combination of codes and clinical concepts in this visit to be predictive of a positive stroke incidence.

Figure 1. Inclusion of clinical concepts and codes on a patient visit timeline. CPT: Current Procedural Terminology; ICD-9: International Classification of Diseases, Ninth Revision.

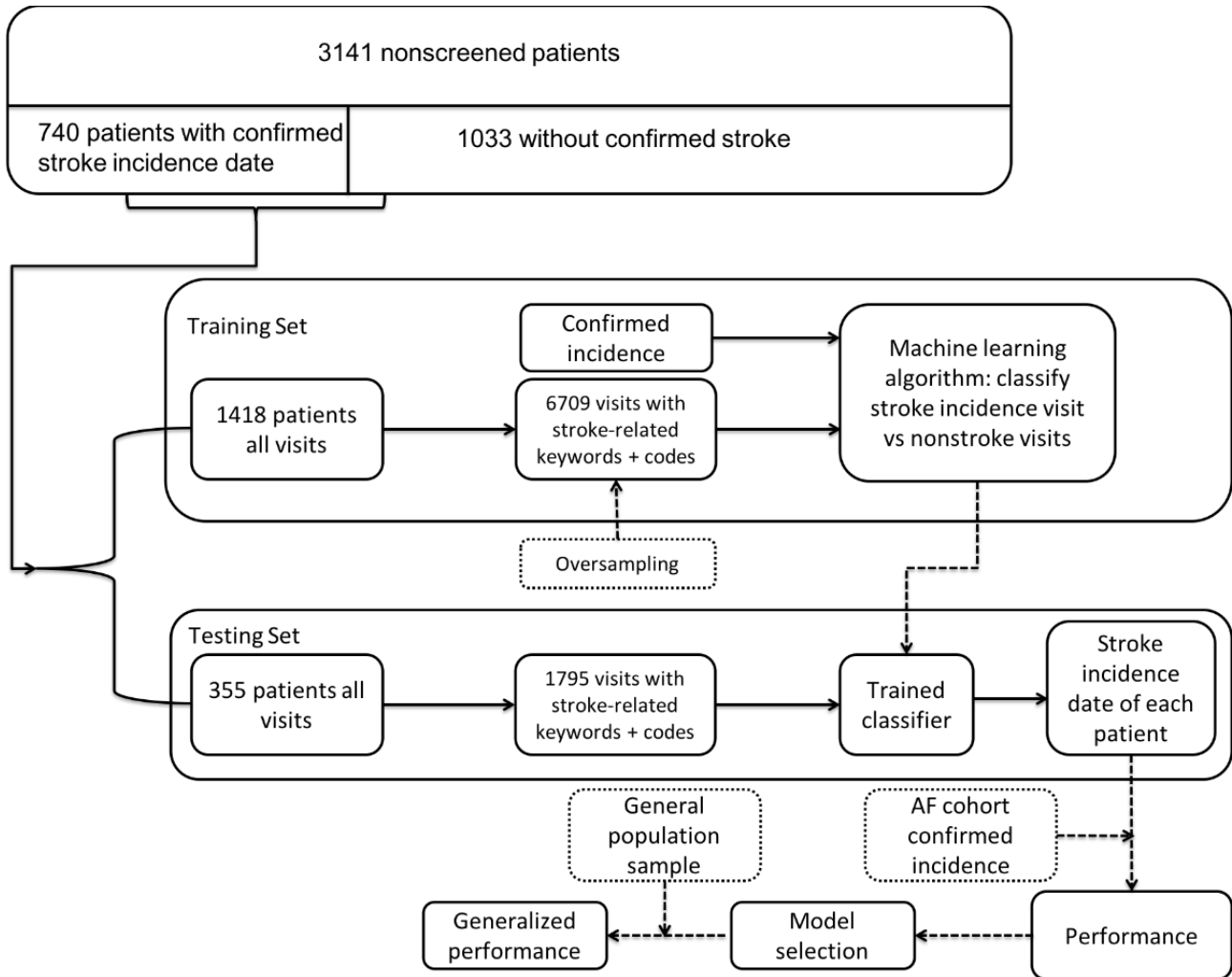


Data Analysis

After incident stroke was confirmed, visits afterwards were not reviewed by abstractors and thus excluded from our overall data set. Figure 2 shows the workflow of the algorithm training and testing process. We created a data set with 9130 confirmed visits (with stroke vs nonstroke labels) among the 1773 patients. In total, there were 746 stroke visits and 8384 nonstroke visits. The stroke incidence count (n=746) was larger than the number of patients with confirmed stroke incidence (740) because incidence dates for different subtypes of stroke (ischemic stroke/

transient ischemic attack and hemorrhagic stroke) were all recorded, such that patients might have had multiple incidence dates. We included data from a randomly selected 79.98% of our screened patients (1418/1773 patients; 7253 visits) as a training set and the remaining 20.02% of our screened patients (355/1773 patients; 1877 visits) were retained as an independent testing set. Due to the outcome imbalance in the data set (positive:negative ratio of about 1:10), we used the synthetic minority oversampling technique [51] to create oversampled training data sets with an oversampling percentage of 1000%.

Figure 2. Stroke algorithm training and testing workflow. AF: atrial fibrillation.



We considered two machine learning classifiers, logistic regression and random forest [52], to train our phenotyping models. Logistic regression served as a baseline modeling algorithm. Random forest was also chosen because of its high performance with structured input features and better model flexibility. We also considered the influence of feature groups by varying the inclusion of CPT codes and symptom terms in the input feature set. The hyperparameter tuning of the machine learning models was performed using 10-fold cross-validation. The performance metrics adopted for the machine learning task in the test set were precision, recall, and F score. The oversampling and machine learning modeling training and testing processes were implemented in Weka 3 (University of Waikato) [53]. Additional statistical summaries were performed using the R statistical software version 3.6.2 (The R Foundation

for Statistical Computing). Quantitative variables are summarized as means, while nominal variables are expressed by counts and percentages.

Validation Cohort

We evaluated the generalizability of our model on a sample from a general population cohort of 71,429 patients. This cohort consisted of individuals sampled in Olmsted County, Minnesota on January 1, 2006, with an age ≥30 years and with no prior history of cardiovascular disease. We applied the best performing model based on the leave-out test set to this entire population cohort to generate incident stroke predictions. We then randomly selected 50 patients from those who had no stroke-related features (ie, de facto negative stroke predictions), 50 patients from those who were shown to have negative stroke

predictions, and 50 patients from those who were shown to have positive stroke predictions and a predicted incident stroke for evaluation. This verification-based sampling strategy allowed for estimates of positive and negative predictive values (PPVs and NPVs, respectively) by conditioning on algorithm predictions. Under these conditions ($n=50$), the half-width of the 95% Wilson score CI for the PPVs and separate NPVs would be approximately 0.1 for a true value of 0.85.

All 150 patient cases were reviewed by 1 nurse abstractor to confirm incident stroke, which served as our gold standard. We recorded model prediction outputs on all patient visits in the 150-patient validation set. We combined visit-level true predictions to generate patient-level incidence predictions by saving only the earliest date of positive predictions as stroke incidences. We compared patient-level incidence predictions with our gold standard. True prediction in our evaluation meant the date of the predicted incident stroke was within 60 days of the abstracted stroke date. A 2 x 2 confusion matrix was used to calculate performance scores for prediction evaluation. Model performance metrics included PPV and NPV using manual evaluation as the gold standard and patient-level predictions to calculate true positives, false positives, true negatives, and false negatives. The uncertainty of these performance estimates was calculated using Wilson score 95% CI for proportions.

In addition, we developed heuristic rules to distinguish stroke subtype (ischemic stroke/transient ischemic attack or hemorrhagic stroke) of each identified stroke incidence by

analyzing the composition of keyword or code input feature sets (in a window of 60 days). We counted the number of keywords or codes for each ischemic stroke/transient ischemic attack and hemorrhagic stroke. If an input feature set contained more keywords or codes for ischemic stroke/transient ischemic attack, then this incidence was considered an ischemic stroke incidence; otherwise, it was considered a hemorrhagic stroke incidence. We only evaluated correct incident stroke predictions from the previous step in the evaluation data set with manually ascertained subtypes as the gold standard. Accuracy was calculated to measure performance of the subtype identification.

Results

Model Selection and Subtype Identification

Table 2 shows the algorithm performance measured on the test set for 8 models run on 4 input combinations and 2 classifiers (logistic regression and random forest). The random forest classifier outperformed the logistic classifier regardless of the feature sets used. Inclusion of CPT codes as features improved the performance for the random forest model with F score increased from 0.836 (Model 3) to 0.905 (Model 1). However, in the logistic model, the inclusion of CPT codes slightly improved the F score from 0.772 (Model 4) to 0.793 (Model 2). Using comparisons to all features (Model 1 and 2) and excluding the symptom terms (Model 6 and 7) achieved better F score (values italicized in Table 2).

Table 2. Stroke algorithm performance.

Model	ICD-9 ^a	Clinical concept	CPT ^b	Classifier	Precision	Recall	F score
1	Yes	Symptoms + disease concepts	Yes	Random forest	0.912	0.906	0.905
2	Yes	Symptoms + disease concepts	Yes	Logistic	0.807	0.795	0.793
3	Yes	Symptoms + disease concepts	No	Random forest	0.835	0.845	0.836
4	Yes	Symptoms + disease concepts	No	Logistic	0.791	0.777	0.772
5	Yes	Disease-only concept	Yes	Random forest	<i>0.920</i>	<i>0.915</i>	<i>0.915</i>
6	Yes	Disease-only concept	Yes	Logistic	0.809	0.798	0.796
7	Yes	Disease-only concept	No	Random forest	0.856	0.847	0.846
8	Yes	Disease-only concept	No	Logistic	0.779	0.767	0.763

^aICD-9: International Classification of Diseases, Ninth Revision.

^bCPT: Current Procedural Terminology.

Model Generalizability

Table 3 shows the distribution of stroke features in the AF cohort and the general population cohort. The AF cohort had a higher proportion of stroke-related codes and concepts. Results from the evaluation of the 150 selected patient records are presented in Table 4. Prediction performance corresponded to a PPV of

0.86 (95% CI 0.74-0.93), an NPV without ICD codes of 1.00 (95% CI 0.92-1.00), and an NPV with codes of 0.92 (95% CI 0.90-0.98). No strokes were observed among patients with no eligible stroke ICD codes. For subtype characterization, we achieved an accuracy of 80% (95% CI 0.68-0.89) in the general population sample.

Table 3. Patient feature distribution post-AF.

Stroke feature distribution	AF ^a screened		AF nonscreened (n=3141), n (%)	AF cohort total (n=4914), n (%)	Olmsted County cohort (N=71,429), n (%)
	Stroke (n=740), n (%)	No stroke (n=1033), n (%)			
ICD-9 ^b +CPT ^c +CC ^d	654 (88.37)	379 (36.69)	0 (0)	1033 (21.02)	2726 (3.82)
ICD-9+CPT	66 (8.92)	596 (57.70)	0 (0)	662 (13.47)	1018 (1.42)
ICD-9+CC	9 (1.22)	12 (1.16)	0 (0)	21 (0.43)	48 (0.067)
CPT+CC	0 (0)	0 (0)	167 (5.32)	167 (3.40)	1595 (2.23)
ICD-9	11 (1.49)	46 (4.45)	0 (0)	57 (1.16)	194 (0.27)
CPT	0 (0)	0 (0)	1736 (55.27)	1736 (35.33)	17,433 (24.40)
CC	0 (0)	0 (0)	11 (0.35)	11 (0.24)	566 (0.79)
None	0 (0)	0 (0)	1227 (39.06)	1227 (24.97)	47,849 (66.99)

^aAF: atrial fibrillation.

^bICD-9: International Classification of Diseases, Ninth Revision.

^cCPT: Current Procedural Terminology.

^dCC: clinical concepts.

Table 4. Generalizability analysis results from the Olmsted County cohort.

Gold standard	Stroke algorithm prediction (N=150)		
	Negative (n=100)		Positive (n=50)
	No ICD-9 ^a codes (n=50)	Predicted no stroke (n=50)	
Stroke	0	4	43
No Stroke	50	46	7

^aICD-9: International Classification of Diseases, Ninth Revision.

Discussion

Principal Findings

The rapid expansion of information available in EHRs opens new opportunities to combine structured and unstructured data for research. Advances in machine learning methods and tools facilitate the combination of multimodal clinical data for effective development of phenotyping algorithms. However, performance of stroke electronic phenotyping algorithms varies by stroke subtypes [25] and phenotyping tasks (ie, case vs noncase or incident stroke phenotyping). Our previous study showed that when naïve ICD codes with clinical concept matching were used, stroke incidence identification had a PPV of 60.6% while case-versus-noncase identification had a much higher PPV of 88.7% [20].

In this study, we included clinical concepts extracted from clinical notes along with ICD-9 and CPT codes for incident stroke ascertainment. The rationale to add CPT codes is that diagnosis of stroke usually needs to be confirmed by imaging evidence and will probably be followed by therapeutic procedures. Thus, the addition of CPT codes in the model could potentially help to reduce the information redundancy effect by distinguishing between past and current events recorded in clinical notes. Our algorithm closely resembles the ascertainment

process (chart review) of clinicians, which uses multiple types of EHR data (eg, diagnoses and procedure codes, unstructured clinical notes) in a parsimonious manner. Due to the redundancy and temporal ambiguity in unstructured clinical notes, we needed to construct a data set with sufficient and interpretable features from multimodal clinical data.

We found that the random forest generated better results, while the addition of CPT codes improved overall performance. This may be because imaging procedures, especially head CT or MRI, are critical in the diagnosis of stroke. Therefore, CPT codes of such procedures can be important indicators for distinguishing between incident and historical events. In addition, ICD codes and therapeutic procedures can vary significantly between incident and recurrent events. Meanwhile, we observed that the additions of stroke-related symptom concepts were not helpful for the phenotyping task. This may be due to the fact that our stroke incidence ascertainment depends largely on the ubiquitous nature of many stroke-related symptoms: they may be stroke-related but not necessarily stroke specific. Additionally, ascertainment requires well-documented evidence, such as imaging or imaging reports. Without properly recorded evidence, patients are not likely to be ascertained as stroke.

Our generalizability evaluation demonstrates that models trained using a specific disease cohort for incident stroke ascertainment

can generalize well to a general patient population. This is very encouraging given there are many existing patient cohorts available. Secondary use of these patient cohorts would be a cost-effective way for developing machine learning–based phenotyping algorithms. The study also illustrates that incorporating structured EHR data, such as CPT codes, can effectively distinguish incident stroke mentions from historical events in the clinical notes.

One limitation of our study is the dependence of domain experts to provide relevant clinical concepts, ICD-9 codes, and CPT codes. In the future, we will explore advance feature engineering approaches to identify those relevant concepts or codes automatically or semiautomatically. We are also aware that our imbalance cohort data and oversampling strategies might have introduced overfitting. Although our evaluation in the general population proved the performance of the algorithm, in the future, we can adopt a case–control matching strategy to deal with imbalanced data and mitigate the potential overfitting issue.

In addition, new treatment strategies (mechanical thrombectomy) to treat stroke have been in the market in recent years, and thus the features used in our algorithm could have different weights for predictions of events in different temporal settings. A more precise strategy could consider using different features for prediction tasks in different time frames, where variations in clinical knowledge and care path have been considered.

Conclusions

In conclusion, the high prevalence of stroke and the lack of an efficient algorithm to confirm incident stroke events necessitate the development of an effective and interpretable algorithm to identify incident stroke occurrences. In this paper, we described our efforts to develop and validate an EHR-based algorithm that accurately identifies incident stroke events and goes beyond typical case-versus-noncase stroke identification. Our algorithm's good performance in a general population sample demonstrates its generalizability and potential to be adopted by other institutions.

Acknowledgments

The authors thank the other investigators, the staff, and the participants of the stroke phenotyping study for their valuable contributions. The stroke phenotyping study was conducted by a collaborative team of researchers from the Department of Health Sciences Research at Mayo Clinic and used 2 cohorts: the AF cohort and the Olmsted County cohort. This study was supported by grants from the National Institutes of Health (no. R01 HL136659 and R21 AG062580) and the American Heart Association (no. 11SDG7260039), and was made possible using the resources of the Rochester Epidemiology Project (no. R01 AG034676).

Authors' Contributions

YZ had full access to all study data and takes responsibility for the integrity of the data and the accuracy of the data analysis. All authors were responsible for study concept; design, acquisition, analysis, or interpretation of data; critical revision of the manuscript for important intellectual content; and administrative, technical, or material support. YZ was responsible for the drafting of the manuscript. YZ and NBL were responsible for statistical analysis. SB and AMC were responsible for obtaining funding. HL, SB, and NBL were responsible for study supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1

ICD-9, CPT codes, and clinical concepts for two major stroke subtypes and stroke-related symptoms.
[\[DOCX File, 34 KB-Multimedia Appendix 1\]](#)

References

1. Bonita R. Epidemiology of stroke. *Lancet* 1992 Mar 08;339(8789):342-344. [doi: [10.1016/0140-6736\(92\)91658-u](https://doi.org/10.1016/0140-6736(92)91658-u)] [Medline: [1346420](https://pubmed.ncbi.nlm.nih.gov/1346420/)]
2. Clinical diagnosis of stroke subtypes. UptoDate. URL: <https://www.uptodate.com/contents/clinical-diagnosis-of-stroke-subtypes> [accessed 2020-07-22]
3. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. *PLoS One* 2014;9(11):e110900 [FREE Full text] [doi: [10.1371/journal.pone.0110900](https://doi.org/10.1371/journal.pone.0110900)] [Medline: [25369203](https://pubmed.ncbi.nlm.nih.gov/25369203/)]
4. Chamberlain AM, Brown RD, Alonso A, Gersh BJ, Killian JM, Weston SA, et al. No decline in the risk of stroke following incident atrial fibrillation since 2000 in the community: a concerning trend. *J Am Heart Assoc* 2016 Jun 13;5(6):e003408 [FREE Full text] [doi: [10.1161/JAHA.116.003408](https://doi.org/10.1161/JAHA.116.003408)] [Medline: [27412902](https://pubmed.ncbi.nlm.nih.gov/27412902/)]
5. Scannapieco FA, Bush RB, Paju S. Associations between periodontal disease and risk for atherosclerosis, cardiovascular disease, and stroke. A systematic review. *Ann Periodontol* 2003 Dec;8(1):38-53. [doi: [10.1902/annals.2003.8.1.38](https://doi.org/10.1902/annals.2003.8.1.38)] [Medline: [14971247](https://pubmed.ncbi.nlm.nih.gov/14971247/)]
6. Slee VN. The International Classification of Diseases: ninth revision (ICD-9). *Ann Intern Med* 1978 Mar;88(3):424-426. [doi: [10.7326/0003-4819-88-3-424](https://doi.org/10.7326/0003-4819-88-3-424)] [Medline: [629506](https://pubmed.ncbi.nlm.nih.gov/629506/)]

7. Imran TF, Posner D, Honerlaw J, Vassy JL, Song RJ, Ho YL, et al. A phenotyping algorithm to identify acute ischemic stroke accurately from a national biobank: the Million Veteran Program. *Clin Epidemiol* 2018;10:1509-1521 [[FREE Full text](#)] [doi: [10.2147/CLEP.S160764](https://doi.org/10.2147/CLEP.S160764)] [Medline: [30425582](https://pubmed.ncbi.nlm.nih.gov/30425582/)]
8. Coull A, Silver L, Bull L, Giles M, Rothwell P, Oxford Vascular (OXVASC) Study. Direct assessment of completeness of ascertainment in a stroke incidence study. *Stroke* 2004 Sep;35(9):2041-2045. [doi: [10.1161/01.STR.0000137605.48864.2f](https://doi.org/10.1161/01.STR.0000137605.48864.2f)] [Medline: [15256682](https://pubmed.ncbi.nlm.nih.gov/15256682/)]
9. Thangaraj P, Kummer B, Lorberbaum T, Elkind M, Tatonetti N. Comparative analysis, applications, and interpretation of electronic health record-based stroke phenotyping methods. *BioData Min* 2020 Dec 07;13(1):21 [[FREE Full text](#)] [doi: [10.1186/s13040-020-00230-x](https://doi.org/10.1186/s13040-020-00230-x)] [Medline: [33372632](https://pubmed.ncbi.nlm.nih.gov/33372632/)]
10. Moons KGM, Bots ML, Salonen JT, Elwood PC, Freire de Concalves A, Nikitin Y, et al. Prediction of stroke in the general population in Europe (EUROSTROKE): Is there a role for fibrinogen and electrocardiography? *J Epidemiol Community Health* 2002 Mar;56 Suppl 1:i30-i36 [[FREE Full text](#)] [doi: [10.1136/jech.56.suppl_1.i30](https://doi.org/10.1136/jech.56.suppl_1.i30)] [Medline: [11815642](https://pubmed.ncbi.nlm.nih.gov/11815642/)]
11. Ho KC, Speier W, El-Saden S, Liebeskind DS, Saver JL, Bui AAT, et al. Predicting discharge mortality after acute ischemic stroke using balanced data. *AMIA Annu Symp Proc* 2014;2014:1787-1796 [[FREE Full text](#)] [Medline: [25954451](https://pubmed.ncbi.nlm.nih.gov/25954451/)]
12. Peng S, Chuang Y, Kang T, Tseng K. Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination. *Eur J Neurol* 2010 Jul;17(7):945-950. [doi: [10.1111/j.1468-1331.2010.02955.x](https://doi.org/10.1111/j.1468-1331.2010.02955.x)] [Medline: [20136650](https://pubmed.ncbi.nlm.nih.gov/20136650/)]
13. Colak C, Karaman E, Turtay MG. Application of knowledge discovery process on the prediction of stroke. *Comput Methods Programs Biomed* 2015 May;119(3):181-185. [doi: [10.1016/j.cmpb.2015.03.002](https://doi.org/10.1016/j.cmpb.2015.03.002)] [Medline: [25827533](https://pubmed.ncbi.nlm.nih.gov/25827533/)]
14. Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS One* 2014;9(2):e88225 [[FREE Full text](#)] [doi: [10.1371/journal.pone.0088225](https://doi.org/10.1371/journal.pone.0088225)] [Medline: [24520356](https://pubmed.ncbi.nlm.nih.gov/24520356/)]
15. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet* 2011 Oct 07;89(4):529-542 [[FREE Full text](#)] [doi: [10.1016/j.ajhg.2011.09.008](https://doi.org/10.1016/j.ajhg.2011.09.008)] [Medline: [21981779](https://pubmed.ncbi.nlm.nih.gov/21981779/)]
16. Conway M, Berg R, Carrell D, Denny J, Kho A, Kullo I, et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. *AMIA Annu Symp Proc* 2011;2011:274-283 [[FREE Full text](#)] [Medline: [22195079](https://pubmed.ncbi.nlm.nih.gov/22195079/)]
17. Overby CL, Pathak J, Gottesman O, Haerian K, Perotte A, Murphy S, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. *J Am Med Inform Assoc* 2013 Dec;20(e2):e243-e252 [[FREE Full text](#)] [doi: [10.1136/amiainl-2013-001930](https://doi.org/10.1136/amiainl-2013-001930)] [Medline: [23837993](https://pubmed.ncbi.nlm.nih.gov/23837993/)]
18. Castro VM, Minnier J, Murphy SN, Kohane I, Churchill SE, Gainer V, International Cohort Collection for Bipolar Disorder Consortium. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *Am J Psychiatry* 2015 Apr;172(4):363-372 [[FREE Full text](#)] [doi: [10.1176/appi.ajp.2014.14030423](https://doi.org/10.1176/appi.ajp.2014.14030423)] [Medline: [25827034](https://pubmed.ncbi.nlm.nih.gov/25827034/)]
19. Feigin VL, Carter K. Editorial comment--Stroke incidence studies one step closer to the elusive gold standard? *Stroke* 2004 Sep;35(9):2045-2047. [Medline: [15331801](https://pubmed.ncbi.nlm.nih.gov/15331801/)]
20. Hollands GJ, Marteau TM, Fletcher PC. Non-conscious processes in changing health-related behaviour: a conceptual analysis and framework. *Health Psychology Review* 2016 Feb 16;10(4):381-394 [[FREE Full text](#)] [doi: [10.1080/17437199.2015.1138093](https://doi.org/10.1080/17437199.2015.1138093)]
21. Ni Y, Alwell K, Moomaw CJ, Woo D, Adeoye O, Flaherty ML, et al. Towards phenotyping stroke: leveraging data from a large-scale epidemiological study to detect stroke diagnosis. *PLoS One* 2018;13(2):e0192586 [[FREE Full text](#)] [doi: [10.1371/journal.pone.0192586](https://doi.org/10.1371/journal.pone.0192586)] [Medline: [29444182](https://pubmed.ncbi.nlm.nih.gov/29444182/)]
22. Howard VJ, McClure LA, Meschia JF, Pulley L, Orr SC, Friday GH. High prevalence of stroke symptoms among persons without a diagnosis of stroke or transient ischemic attack in a general population: the REasons for Geographic And Racial Differences in Stroke (REGARDS) study. *Arch Intern Med* 2006 Oct 09;166(18):1952-1958. [doi: [10.1001/archinte.166.18.1952](https://doi.org/10.1001/archinte.166.18.1952)] [Medline: [17030827](https://pubmed.ncbi.nlm.nih.gov/17030827/)]
23. How a CPT code becomes a code. American Speech-Language-Hearing Association. URL: <https://www.asha.org/uploadedFiles/How-A-CPT-Code-Becomes-A-Code.pdf> [accessed 2020-07-22]
24. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol* 2012 Dec;41(6):1625-1638 [[FREE Full text](#)] [doi: [10.1093/ije/dys188](https://doi.org/10.1093/ije/dys188)] [Medline: [23220717](https://pubmed.ncbi.nlm.nih.gov/23220717/)]
25. Woodfield R, Grant I, UK Biobank Stroke Outcomes Group, UK Biobank Follow-Up and Outcomes Working Group, Sudlow CLM. Accuracy of electronic health record data for identifying stroke cases in large-scale epidemiological studies: a systematic review from the UK Biobank Stroke Outcomes Group. *PLoS One* 2015;10(10):e0140533 [[FREE Full text](#)] [doi: [10.1371/journal.pone.0140533](https://doi.org/10.1371/journal.pone.0140533)] [Medline: [26496350](https://pubmed.ncbi.nlm.nih.gov/26496350/)]
26. Schuemie MJ, Sen E, 't JGW, van SEM, Sturkenboom MC, Kors JA. Automating classification of free-text electronic health records for epidemiological studies. *Pharmacoepidemiol Drug Saf* 2012 Jun;21(6):651-658. [doi: [10.1002/pds.3205](https://doi.org/10.1002/pds.3205)] [Medline: [22271492](https://pubmed.ncbi.nlm.nih.gov/22271492/)]

27. Schellenbaum GD, Heckbert SR, Smith NL, Rea TD, Lumley T, Kitzman DW, et al. Congestive heart failure incidence and prognosis: case identification using central adjudication versus hospital discharge diagnoses. *Ann Epidemiol* 2006 Mar;16(2):115-122. [doi: [10.1016/j.annepidem.2005.02.012](https://doi.org/10.1016/j.annepidem.2005.02.012)] [Medline: [15964203](https://pubmed.ncbi.nlm.nih.gov/15964203/)]
28. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care* 2007 Jun;13(6 Part 1):281-288 [FREE Full text] [Medline: [17567225](https://pubmed.ncbi.nlm.nih.gov/17567225/)]
29. Ermenc B. Minimizing mistakes in clinical diagnosis. *J Forensic Sci* 1999 Jul;44(4):810-813. [Medline: [10432615](https://pubmed.ncbi.nlm.nih.gov/10432615/)]
30. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, eMERGE Team. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011 Jan 26;4:13 [FREE Full text] [doi: [10.1186/1755-8794-4-13](https://doi.org/10.1186/1755-8794-4-13)] [Medline: [21269473](https://pubmed.ncbi.nlm.nih.gov/21269473/)]
31. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* 2013 Oct;15(10):761-771 [FREE Full text] [doi: [10.1038/gim.2013.72](https://doi.org/10.1038/gim.2013.72)] [Medline: [23743551](https://pubmed.ncbi.nlm.nih.gov/23743551/)]
32. Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput Biol* 2012;8(12):e1002823 [FREE Full text] [doi: [10.1371/journal.pcbi.1002823](https://doi.org/10.1371/journal.pcbi.1002823)] [Medline: [23300414](https://pubmed.ncbi.nlm.nih.gov/23300414/)]
33. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012;19(2):212-218 [FREE Full text] [doi: [10.1136/amiainjnl-2011-000439](https://doi.org/10.1136/amiainjnl-2011-000439)] [Medline: [22101970](https://pubmed.ncbi.nlm.nih.gov/22101970/)]
34. Ritchie MD, Denny JC, Zuvich RL, Crawford DC, Schildcrout JS, Bastarache L, Cohorts for HeartAging Research in Genomic Epidemiology (CHARGE) QRS Group, et al. Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 2013 Apr 02;127(13):1377-1385 [FREE Full text] [doi: [10.1161/CIRCULATIONAHA.112.000604](https://doi.org/10.1161/CIRCULATIONAHA.112.000604)] [Medline: [23463857](https://pubmed.ncbi.nlm.nih.gov/23463857/)]
35. Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, Ramelson HZ, et al. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. *J Am Med Inform Assoc* 2011;18(6):859-867 [FREE Full text] [doi: [10.1136/amiainjnl-2011-000121](https://doi.org/10.1136/amiainjnl-2011-000121)] [Medline: [21613643](https://pubmed.ncbi.nlm.nih.gov/21613643/)]
36. Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, Heavirland J, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. *J Am Med Inform Assoc* 2012;19(5):859-866 [FREE Full text] [doi: [10.1136/amiainjnl-2011-000535](https://doi.org/10.1136/amiainjnl-2011-000535)] [Medline: [22437073](https://pubmed.ncbi.nlm.nih.gov/22437073/)]
37. Percha B, Nassif H, Lipson J, Burnside E, Rubin D. Automatic classification of mammography reports by BI-RADS breast tissue composition class. *J Am Med Inform Assoc* 2012;19(5):913-916 [FREE Full text] [doi: [10.1136/amiainjnl-2011-000607](https://doi.org/10.1136/amiainjnl-2011-000607)] [Medline: [22291166](https://pubmed.ncbi.nlm.nih.gov/22291166/)]
38. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc* 2011;18(5):601-606 [FREE Full text] [doi: [10.1136/amiainjnl-2011-000163](https://doi.org/10.1136/amiainjnl-2011-000163)] [Medline: [21508414](https://pubmed.ncbi.nlm.nih.gov/21508414/)]
39. Xu Y, Hong K, Tsujii J, Chang EI. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *J Am Med Inform Assoc* 2012;19(5):824-832 [FREE Full text] [doi: [10.1136/amiainjnl-2011-000776](https://doi.org/10.1136/amiainjnl-2011-000776)] [Medline: [22586067](https://pubmed.ncbi.nlm.nih.gov/22586067/)]
40. D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. *J Am Med Inform Assoc* 2011;18(5):607-613 [FREE Full text] [doi: [10.1136/amiainjnl-2011-000183](https://doi.org/10.1136/amiainjnl-2011-000183)] [Medline: [21697292](https://pubmed.ncbi.nlm.nih.gov/21697292/)]
41. Kahn MG, Weng C. Clinical research informatics: a conceptual perspective. *J Am Med Inform Assoc* 2012 Jun;19(e1):e36-e42 [FREE Full text] [doi: [10.1136/amiainjnl-2012-000968](https://doi.org/10.1136/amiainjnl-2012-000968)] [Medline: [22523344](https://pubmed.ncbi.nlm.nih.gov/22523344/)]
42. Chamberlain AM, Gersh BJ, Alonso A, Chen LY, Berardi C, Manemann SM, et al. Decade-long trends in atrial fibrillation incidence and survival: a community study. *Am J Med* 2015 Mar;128(3):260-7.e1 [FREE Full text] [doi: [10.1016/j.amjmed.2014.10.030](https://doi.org/10.1016/j.amjmed.2014.10.030)] [Medline: [25446299](https://pubmed.ncbi.nlm.nih.gov/25446299/)]
43. St Sauver JL, Grossardt BR, Yawn BP, Melton LJ, Rocca WA. Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project. *Am J Epidemiol* 2011 May 01;173(9):1059-1068 [FREE Full text] [doi: [10.1093/aje/kwq482](https://doi.org/10.1093/aje/kwq482)] [Medline: [21430193](https://pubmed.ncbi.nlm.nih.gov/21430193/)]
44. Rocca WA, Yawn BP, St Sauver JL, Grossardt BR, Melton LJ. History of the Rochester Epidemiology Project: half a century of medical records linkage in a US population. *Mayo Clin Proc* 2012 Dec;87(12):1202-1213 [FREE Full text] [doi: [10.1016/j.mayocp.2012.08.012](https://doi.org/10.1016/j.mayocp.2012.08.012)] [Medline: [23199802](https://pubmed.ncbi.nlm.nih.gov/23199802/)]
45. Rocca WA, Grossardt BR, Brue SM, Bock-Goodner CM, Chamberlain AM, Wilson PM, et al. Data Resource Profile: Expansion of the Rochester Epidemiology Project medical records-linkage system (E-REP). *Int J Epidemiol* 2018 Apr 01;47(2):368-368j [FREE Full text] [doi: [10.1093/ije/dyx268](https://doi.org/10.1093/ije/dyx268)] [Medline: [29346555](https://pubmed.ncbi.nlm.nih.gov/29346555/)]
46. Witt BJ, Brown RD, Jacobsen SJ, Weston SA, Yawn BP, Roger VL. A community-based study of stroke incidence after myocardial infarction. *Ann Intern Med* 2005 Dec 06;143(11):785-792. [doi: [10.7326/0003-4819-143-11-200512060-00006](https://doi.org/10.7326/0003-4819-143-11-200512060-00006)] [Medline: [16330789](https://pubmed.ncbi.nlm.nih.gov/16330789/)]

47. Liu H, Bielinski SJ, Sohn S, Murphy S, Waghlikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. *AMIA Jt Summits Transl Sci Proc* 2013;2013:149-153 [[FREE Full text](#)] [Medline: [24303255](#)]
48. Mitchell-Box K, Braun KL. Fathers' thoughts on breastfeeding and implications for a theory-based intervention. *J Obstet Gynecol Neonatal Nurs* 2012;41(6):E41-E50. [doi: [10.1111/j.1552-6909.2012.01399.x](#)] [Medline: [22861175](#)]
49. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experienter, and temporal status from clinical reports. *J Biomed Inform* 2009 Oct;42(5):839-851 [[FREE Full text](#)] [doi: [10.1016/j.jbi.2009.05.002](#)] [Medline: [19435614](#)]
50. Chapman W, Chu D, Dowling J. editors. ConText: An algorithm for identifying contextual features from clinical text. 2007. URL: <https://dl.acm.org/doi/10.5555/1572392.1572408> [accessed 2020-07-21]
51. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic Minority Over-sampling Technique. *JAIR* 2002 Jun 01;16(6):321-357. [doi: [10.1613/jair.953](#)]
52. Liaw A, Weiner M. Classification and regression by randomForest. *R news* 2002;2(3):18-22.
53. Eibe F, Hall M, Witten I. The WEKA Workbench. In: *Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. Burlington, Massachusetts: Morgan Kaufmann; 4th edition (December 1, 2016); 2016.

Abbreviations

AF: atrial fibrillation
CPT: Current Procedural Terminology
CT: computed tomography
EHR: electronic health record
eMERGE: Electronic Medical Records and Genomics
ICD: International Classification of Diseases
ICD-9: International Classification of Diseases, Ninth Revision
MRI: magnetic resonance imaging
NPV: negative predictive value
PPV: positive predictive value
REP: Rochester Epidemiology Project

Edited by G Eysenbach, R Kukafka; submitted 27.07.20; peer-reviewed by C Colak, MO Kaya; comments to author 12.08.20; revised version received 25.08.20; accepted 20.01.21; published 08.03.21

Please cite as:

Zhao Y, Fu S, Bielinski SJ, Decker PA, Chamberlain AM, Roger VL, Liu H, Larson NB
Natural Language Processing and Machine Learning for Identifying Incident Stroke From Electronic Health Records: Algorithm Development and Validation
J Med Internet Res 2021;23(3):e22951
URL: <https://www.jmir.org/2021/3/e22951>
doi: [10.2196/22951](#)
PMID: [33683212](#)

©Yiqing Zhao, Sunyang Fu, Suzette J Bielinski, Paul A Decker, Alanna M Chamberlain, Veronique L Roger, Hongfang Liu, Nicholas B Larson. Originally published in the *Journal of Medical Internet Research* (<http://www.jmir.org>), 08.03.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the *Journal of Medical Internet Research*, is properly cited. The complete bibliographic information, a link to the original publication on <http://www.jmir.org/>, as well as this copyright and license information must be included.