Comparative analysis, applications, and interpretation of electronic health record-based stroke phenotyping methods

Phyllis Thangaraj
Columbia University Medical Center https://orcid.org/0000-0003-3968-1563
Benjamin R Kummer
Icahn School of Medicine at Mount Sinai
Tal Lorberbaum
Columbia University Medical Center
Mitchell S.V. Elkind
Columbia University Medical Center
Nicholas P Tatonetti (nick.tatonetti@columbia.edu)
Columbia University https://orcid.org/0000-0002-2700-2597

Research

Keywords: phenotyping algorithms, acute ischemic stroke, machine learning, electronic health record studies

Posted Date: November 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-16812/v4

Version of Record: A version of this preprint was published on December 7th, 2020. See the published version at https://doi.org/10.1186/s13040-020-00230-x.

Abstract

Background: Accurate identification of acute ischemic stroke (AIS) patient cohorts is essential for a wide range of clinical investigations. Automated phenotyping methods that leverage electronic health records (EHRs) represent a fundamentally new approach cohort identification without current laborious and ungeneralizable generation of phenotyping algorithms. We systematically compared and evaluated the ability of machine learning algorithms and case-control combinations to phenotype acute ischemic stroke patients using data from an EHR.

Materials and Methods: Using structured patient data from the EHR at a tertiary-care hospital system, we built and evaluated machine learning models to identify patients with AIS based on 75 different case-control and classifier combinations. We then estimated the prevalence of AIS patients across the EHR. Finally, we externally validated the ability of the models to detect AIS patients without AIS diagnosis codes using the UK Biobank.

Results: Across all models, we found that the mean AUROC for detecting AIS was 0.963±0.0520 and average precision score 0.790±0.196 with minimal feature processing. Classifiers trained with cases with AIS diagnosis codes and controls with no cerebrovascular disease codes had the best average F1 score (0.832±0.0383). In the external validation, we found that the top probabilities from a model-predicted AIS cohort were significantly enriched for AIS patients without AIS diagnosis codes (60-150 fold over expected).

Conclusions: Our findings support machine learning algorithms as a generalizable way to accurately identify AIS patients without using process-intensive manual feature curation. When a set of AIS patients is unavailable, diagnosis codes may be used to train classifier models.

Background

Stroke is a complex disease that is a leading cause of death and severe disability for millions of survivors worldwide.[1] Accurate identification of stroke etiology, which is most commonly ischemic but encompasses several other causative mechanisms, is essential for risk stratification, optimal treatment, and support of clinical research. While electronic health records (EHR) are an emerging resource that can be used to study stroke patients, identification of stroke patient cohorts using the EHR requires the integration of multiple facets of data, including medical notes, labs, imaging reports, and medical expertise of neurologists. This process is often manually performed and time-consuming, and can reveal mis-classification errors.[2] One simple approach to identify acute ischemic stroke (AIS) is the diagnosis-code based algorithm created by Tirschwell and Longstreth.[3] However, identifying every AIS patient using these criteria can be difficult due to the inaccuracy and incompleteness of diagnosis recording through insurance billing. [3–5] Additionally, this approach prevents the identification of AIS patients until after hospital discharge, thereby limiting the clinical usability of identification algorithms in time-sensitive situations, such as in-hospital care management, research protocol enrollment, or acute treatment.

Reproducibility and computability of phenotyping algorithms stem from the use of structured data, standardized terminologies, and rule-based logic.[6] Phenotyping features from the EHR have been traditionally culled and curated by experts to manually construct algorithms, [7] but machine learning techniques present the potential advantage of automating this process of feature selection and refinement.[8-11] Recent machine learning approaches have also combined publicly available knowledge sources with EHR data to facilitate feature curation.[12,13] Additionally, while case and control phenotyping using EHR data has also relied on a small number of expert curated cohorts, recent studies have demonstrated that ML approaches can expand upon and identify such cohorts using automated feature selection and imperfect case definitions in a high-throughput manner.[14-18] Studies have also shown that case and control selection with diagnosis codes can significantly affect model performance, the hierarchical organization of structured medical data can be utilized for feature reduction and model performance improvement, and calibration is essential for understanding the clinical utility of a phenotyping model.[19-22] Stroke phenotyping algorithms have also used machine learning to enhance the classification performance of a diagnosis-code based AIS phenotyping algorithm.[23-26] However, while ML models present an opportunity to automate identification of AIS patients (i.e. phenotyping) with commonly accessible EHR data and develop new approaches to etiologic identification and subtyping, the optimal combination of cases and controls to train such models remains unclear.

Given the limitations of manual and diagnosis-code cohort identification, we sought to develop phenotypic classifiers for AIS using machine learning approaches, with the objective of specifically identifying AIS patients that were missing diagnosis codes. Additionally, considering the challenge of identifying true controls in the EHR for the purpose of model training, we also attempted to determine the optimal grouping of cases and controls by selecting and comparing model discriminatory performance with multiple case-control group combinations. We also sought to contrast model training based on cases defined by diagnostic code with that using manually-curated cohorts. Our phenotyping method utilizes machine learning classifiers with minimal data processing to increase the number of stroke patients recovered within the EHR and reduce the time and effort needed to find them for research studies.

Results

Study cohort

Table 1 presents the data and the total number of patients available for each set of cases and controls used in the training and internal and external validation parts of this study. Out of the Columbia University Irving Medical Center (CUIMC) Clinical Data Warehouse (CDW), which has a total of 6.4 million patients, we extracted 4,844 stroke service patients, which we found to have a 4-16% false positive rate for stroke through manual review. Supplementary Table 2 presents demographic characteristics for the training sets, and Supplementary Table 3 and Supplementary Table 4 present demographic and feature category coverage for the testing sets.

Variable	Identification	N Samples
Total Patients	CUIMC CDW Person ID	6,377,222
Diagnosis Codes	ICD9-CM, ICD10-CM, SNOMED	140,300,457
Procedure Codes	ICD9-CM, ICD10-CM, CPT, SNOMED	64,383,775
Prescription Orders	RxNorm	40,759,814
Training Categories: Cases		
(S) Cases: Stroke Service Patients	Seen by NYP Stroke Service	4,484
(T) Cases: AIS Tirschwell Criteria	ICD9-CM: 434.x1, 433.x1, ICD10-CM: I63.xxx	79,306
(C) Cases: CCS Cerebrovascular Disease	ICD9-CM: 346.6x,430, 431, 432.x, 433.xx	181,698
Training Categories: Controls		
(N) Controls: AIS Mimetic Diseases	ICD9-CM: 191.x, 225.x, 340, 250.0, 431	8,438
(I) Controls: Without AIS Tirschwell Criteria	No (T) Codes	5,243,646
(C) Controls: Without CCS Cerebrovascular Disease	No (C) Codes	5,149,975
(CI) Controls: With CCS Cerebrovascular disease, w/o AIS Tirschwell Criteria	(C) codes, No (T) codes	102,435
(R) Random set of patients	With >=1 ICD9-CM or ICD10-CM diagnosis code	5,396,172

Table 1: Select Structured Data and Sample Case/Controls for models available in Columbia University Irving Medical Center Common Data Warehouse. NYP= New York Presbyterian, AIS= Acute Ischemic Stroke, CCS = Clinical Classifications Software

Algorithm performance

We trained 75 models using all combinations of cases, controls, and model types after excluding 15 neural network models due to poor performance (architecture described in supplemental methods). Logistic regression classifiers with L1 penalty gave the best area under the receiving operator curve (AUROC) performance (0.913-0.997) and the best average precision score (0.662-0.969), followed by logistic regression classifiers with elastic net penalty (Figure 2, Supplementary Table 5).

Across all classifier types, the models using the T-C case-control combination had the best average F1 score (0.832±0.0383), whereas logistic regression models with L1 penalty (LR) and elastic-net penalty had the best classifier average F1 score (0.705±0.146 and 0.710±0.134 respectively) (Figure 2B, Supplementary Table 8). Use of cases from the CUIMC stroke service gave the highest average precision

(0.932±.0536), while cases identified through AIS diagnosis codes and controls without cerebrovascular disease or acute ischemic stroke (AIS)-related diagnosis codes (TC, TI) gave high precision as well (0.896±0.0488 and 0.918±0.0316, respectively). The sensitivity of the models ranged widely, between 0.18 and 0.96, while specificity narrowly ranged between 0.993-1.0 (Supplementary Table 9).

We also evaluated the AUROC and maximum F1 Score using a hold-out test set of Tirschwell (T) criteria cases and a random selection of (I) controls. We trained on S, T, and C cases and C controls, and found AUROC of 0.932-0.937 for the TC and CC trained sets and 0.69-0.87 for the SC trained sets. We also see a maximum F1 score of 0.351-0.432 for the TC and CC trained sets and 0.298-0.321 for the SC trained sets (Supplementary Figure 35).

Feature importance

We found the most commonly chosen features associated with stroke diagnosis were procedures used in evaluation of AIS, including extra- and intra-cranial arterial scans, computerized tomography (CT) scans and magnetic resonance imaging (MRI) of the brain, and MR angiography (Figure 3A). We also found that all 75 models relied on incremental contributions from many different features (Figure 3B, Supplementary Figures 20-34).

Internal validation in institutional EHR

We applied the 75 models to the entire CUIMC EHR with at least one diagnosis code, totaling between 5,324,725 and 5,315,923 patients depending on the case/control set. We found that the results varied widely across models, but most predicted a prevalence of between 0.2-2% of patients in the EHR were AIS patients. The models with controls with cerebrovascular disease codes but no AIS codes predicted the lowest prevalence of AIS patients, and found 50.3-100% of the proposed patients had AIS diagnosis codes. The models with the best performance and robustness, 1) stroke service cases and controls without cerebrovascular disease codes and 2) cases with AIS codes and controls without cerebrovascular disease codes with 1) Logistic Regression and L1 Penalty classifier and 2) Adaboost classifier, had sensitivities between 0.822-0.959, specificities 0.994-0.999, and estimated AIS prevalence in the EHR ranging between 1.3-2.0% (Supplementary Table 9, Table 2). Within these proposed AIS patients, 37.7-41.4% had an AIS diagnosis code (Table 2).

Case/ Control Combo	LR EHR Prev	RF EHR Prev	AB EHR Prev	GB EHR Prev	EN EHR Prev	LR with AIS codes	RF with AIS codes	AB with AIS codes	GB with AIS codes	EN with AIS codes
SN	0.7	0.7	1.0	1.3	0.7	41.3	32.2	35.6	29.0	26.4
SI	1.1	2.0	1.5	1.7	1.1	40.5	23.0	35.7	29.8	27.1
SC	1.3	1.7	1.5	1.8	1.3	37.7	25.4	37.9	30.8	28.5
SCI	0.2	0.1	0.2	0.3	0.2	83.1	82.6	76.9	72.2	63.5
SR	0.2	0.2	0.3	0.5	0.2	75.4	63.2	68.8	58.2	48.9
TN	0.9	0.8	0.9	1.0	0.9	44.7	28.5	47.2	35.6	22.5
TI	1.6	2.3	1.4	4.7	1.6	43.8	31.4	47.9	21.8	8.10
TC	1.7	2.7	2.0	1.6	1.7	41.4	28.2	39.0	43.1	32.6
TCI	0.1	0.0	0.1	0.1	0.1	94.6	96.1	85.9	95.3	79.0
TR	0.8	0.8	0.8	0.4	0.8	46.1	40.0	44.0	61.4	31.1
CN	1.3	1.3	1.3	1.0	1.3	34.0	17.1	33.5	31.5	21.4
CI	2.0	3.3	1.9	1.9	2.0	37.5	24.2	39.5	39.8	39.9
CC	2.3	3.3	2.2	2.1	2.3	35.6	25.3	37.2	37.1	29.9
CCI	0.0	0.0	0.1	0.0	0.0	97.5	100	50.3	92.8	74.2
CR	1.0	0.9	0.9	0.7	1.0	37.3	35.6	37.7	42.6	29.6

Table 2. Prevalence of acute ischemic stroke patients identified by each classifier across the EHR and proportion of those patients with T-L criteria. Prev=prevalence. See Supplementary Table 1 for case-control and model abbreviations' definitions.

UK Biobank Variables	Identification	N samples
Subject Data	Diagnosis codes (ICD10), procedure codes (OCPS4), medication prescriptions (Mapped to RxNorm), or demographics	384,208
(T) Cases: AIS Tirschwell Criteria	ICD10 I63.xxx, I64.x (41202,41204)	4,922
(C) Controls: Without Cerebrovascular Disease	No (C) Codes (41202,41204)	312,500
Self-reported AIS but no diagnosis codes	Date of AIS (42008), no AIS Tirschwell Criteria	163

Table 3. Select Structured Data and Case/Control criteria for external validation in the UK Biobank.

External validation

We evaluated the performance of the TC models on identifying 2,624 patients without AIS ICD10 codes. The top 50, 100, 500, and 2,624 probabilities had a precision of over 29%, and up to 80% (Figure 4). Since within the test set only 0.5% of the patients had AIS, this translates to a 60-150-fold increase in AIS detection over random choice.

Discussion

Using a feature-agnostic, data-driven approach with minimal data transformation, we developed models that identify acute ischemic stroke (AIS) patients from commonly-accessible EHR data at the time of patient hospitalization without making use of AIS-related ICD9 and ICD10 codes as defined by Tirschwell and Longstreth. In demonstrating that AIS patients can be recovered from other EHR-available structured clinical features without AIS codes, this approach is in contrast to previous machine learning phenotyping algorithms, which have relied on manually curated features or use AIS-related diagnosis codes as the sole nonzero features in their models. [23,24,3]

Cases and controls for training of phenotyping algorithms can be challenging to identify and define given the richness of available EHR data. From the sparsity of diagnosis codes in the EHR, it follows that patients lacking an AIS-related diagnosis code may not always be considered as a control in stroke cohorts. Similarly, it is difficult to determine whether patients with cerebrovascular diseases, which can serve as risk factors for AIS, or share genetic and pathophysiologic underpinnings with AIS should be considered controls. Additionally, due to the prevalence of AIS mimics, cohort definitions based on diagnosis code criteria may be unreliable. In light of the problems in defining patient cohorts from EHR data, we found marked differences in classifying performance across 15 different case-control training sets. While training with cases from the CUIMC stroke service cases identified stroke patients most accurately and with the highest precision and recall, we also found that training with cases identified from AIS codes with controls from either 1) no cerebrovascular disease or 2) no AIS codes afforded high

precision (Supplementary Table 5). These findings suggest that a manually curated cohort may not be necessary to train the phenotyping models, and the AIS codes may be enough to define a training set. Using these models, we also increased our AIS patient cohort by 60% across the EHR, suggesting that the AIS codes themselves are not sufficient to identify all AIS patients.

We found that stroke evaluation procedures, such as a CT scan or MRI, were important features in many of the models, which corroborates with a previous study. [23] Since none of these models use AIS diagnosis codes as features, this suggests that procedures may serve as proxies for them when identifying AIS cohorts. In some cases, the AIS code will only be added during outpatient follow up. For example, while in the stroke service set, 13.5% of cases did not have AIS codes in the inpatient setting but did in the outpatient setting, and 90% of these patients had had a CT scan of the head. We also found evidence that procedures provided a significant contribution to classification in the models in supplementary analysis (Supplementary Methods, Results, and Supplementary Figure 4).

We found that as measured by AUROC and AP, discriminatory performance of the random forest, logistic regression with L1 and elastic net penalties, and gradient boosting models was robust, even when up to 95% of the training set was removed. These findings showed that a training set size as small as 70-350 samples can maintain high performance, depending on the model.

Our results from traditional model performance and robustness evaluations show that our best machine learning phenotyping algorithm used Logistic Regression with L1 penalty or AdaBoost classifiers trained with controls without any cerebrovascular disease-related codes and a stroke service case population. However, we found that a similar model performed comparably well using cases identified by AIS-related diagnosis codes, suggesting that these models do not require manual case curation for high performance. In addition, our validation study in the UK Biobank detected AIS patients without ICD10-CM codes up to 150-fold better than random selection.

In light of our findings, we recommend using machine learning models trained on all available structured EHR data, not just AIS diagnosis codes, to identify AIS patients. Previous studies required time-consuming manual curation of features or trained on only AIS codes, which would have missed AIS patients identified through a CT scan or MRI but without AIS diagnosis codes.[23,24] Our thorough investigation of feature importance shows that each feature contributes to the improved performance of the models. We also recommend restricting controls further to patients without cerebrovascular disease diagnosis codes, rather than just without AIS diagnosis codes to improve discriminatory ability. In addition, we show improved AUROC and specificity, and comparable sensitivity, precision, recall, and F1-score using SC and TC case-control sets, to previous studies.[23,24] Finally, as shown in Table 2, we show the vast potential for identifying AIS cases in the EHR that do not have an AIS diagnosis code.

This study has several limitations. First, we relied on noisy labels and proxies for training our models, as evidenced by our manual review false positive rate. Without a gold standard set of cases, model performance is difficult to definitively evaluate. We relied on pre-defined codes, the Tirschwell criteria, and

patients evaluated for stroke as our cases. We included a random set of patients as our holdout control test set for representation of all patients in the EHR. This is a limitation, however, because patients with Tirschwell criteria could be labeled as random controls. We addressed this by removing any Tirschwell criteria patients from the hold out controls. In general, the use of random controls could lead to overlapping of cases and controls, especially in common disease, but one can use known diagnostic codes for the disease to separate cases and controls. Our method importantly does not include any codes used in the case and control definitions in our machine learned features in order to identify other features involved in defining stroke patients. We do this so that our models are not reidentifying Tirschwell criteria, and instead are identifying novel features complementary to the criteria. This removal of overlapping cases and controls can also influence our calibration results described in the supplementary materials by changing the proportion of expected stroke cases at each probability score; however, this only amounted to a removal of 0.05% of overlapping patients. We also do see a marked decreased in F1 performance and a slight decrease in AUROC when testing on hold out Tirschwell criteria cases instead of Stroke Service cases in the Columbia EHR. This may be due to better documentation of structured EHR data, particularly procedures and medications, in Stroke Service patients as seen in Supplementary Table 4. However, in the UK Biobank, which used Tirschwell criteria cases as a holdout test set, we see high precision in identifying AIS patients over random. This would suggest reduction in sensitivity of our model. Second, we used only structured features contained within standard terminologies across the patients' entire timeline, and did not use clinical notes. In addition, the biases inherent in phenotyping with billing codes are a significant limitation. Often the data is missing not at random, and data completeness relies on patient interaction with the healthcare system, which can lead to ascertainment bias towards diagnoses and tests that doctors already suspect or patients who actively seek care and make generalizing outcomes from these patients difficult. [27-31] Diagnosis also often are chosen for reimbursement purposes rather than actual diagnosis, and diagnosis code use changes over time, leading to inaccuracies in phenotyping. [27,28] Given previous studies, however, it has been established that stroke can be identified by diagnosis codes with high sensitivity, specificity, and positive predictive value. [3,46] While clinical notes may contain much highly relevant information, they may also give rise to less reproducible and generalizable feature sets. Additionally, each feature contributed incrementally to high performance of the models and required minimal processing to acquire. Third, due to limitations of time and computational complexity, we did not exhaustively explore all possible combinations of cases and controls, including other potential AIS mimetic diseases. Despite these limitations, precision in the internal validation using the held-out set was high, and when applied to an external validation cohort, the developed models improved detection of AIS patients between 60 and 150-fold over random patient identification. Fourth, we did not study clinical implementation of the models. However, the discriminatory ability of the classifiers in the external validation suggest that although these models have not been implemented clinically, they may potentially be useful for improving the power of existing clinical and research study cohorts.

Our study benefits from several strengths. First, to address the current deficiencies in developing phenotyping algorithms, we developed an approach that demonstrates comparable discriminatory ability

of identifying patients with AIS to past methods but has the added benefit of using EHR data that is generally available during inpatient hospitalization. Second, our model features were composed of structured data that encompass a larger feature variety than purely ICD-code based algorithms. Third, because our model incorporated structured data from standard terminologies, they therefore may be generalizable to other health systems outside CUIMC, whereas recent studies have relied on manually curated feature sets.[23] Fourth, we examined several different combinations of cases, controls and classifiers for the purposes of training phenotyping models. Finally, our phenotype classifiers assign probability of having had an AIS, which moves beyond binary classification of patients to develop a more granular description of patient's disease state.

Conclusions

In addition to research tasks such as cohort identification, future models could focus on timely interventions such as care planning prior to discharge and risk stratification. We showed that structured data may be sufficiently accurate for classification, allowing for widespread usability of the algorithm. We also demonstrated the potential for using machine learning classifiers for cohort identification, which achieve high performance with many features acquired through minimal processing. In addition, patient cohorts derived using AIS diagnosis codes may obviate the need for manually-curated cohorts of patients with AIS, and procedure codes may be useful in identifying patients with AIS that may not have been coded with AIS-related diagnosis codes. We, and others, hypothesize that expanding cohort size by assigning a probability of disease may improve the power of heritability and genome-wide association studies.[32–37] Utilizing the structured framework present in many current EHRs, along with machine learning models may provide a generalizable approach for expanding research study cohort size.

Methods

Study design

In this study, we developed several machine learning phenotyping models for AIS using combinations of different case and control groups derived from our institution's EHR data. Use of Columbia patient data was approved by Columbia's institutional review board and UK Biobank data approved with UK Biobank Research Ethics Committee (REC) approval number 16/NW/0274. We also applied key methods to optimize number of features for generalizability, as well as calibration to ensure a clinically meaningful model output, and model robustness to missing data. To estimate the prevalence of potential AIS patients without AIS-related *International Classification of Diseases-Clinical Modification* (ICD-CM) codes, we then applied the developed models to all patients in our institutional EHR. Finally, we externally validated our best-performing model in an independent cohort from the UK Biobank to evaluate its ability to detect AIS patients without the requisite ICD codes. Figure 1 shows the overall workflow of training and testing the models, the models' evaluation, and its testing in an independent test set.

Data sources

We used data from patients in the Columbia University Irving Medical Center Clinical Data Warehouse (CUIMC CDW), which contains longitudinal health records of 6.4 million patients from CUIMC's EHR, spanning 1985-2018. The data are organized into tables and standardized vocabularies and terminologies in the format of the Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership Common Data Model (OMOP CDM).[38] The data include structured medical data such as conditions, procedures, medication orders, lab measurement values, visit type, demographics, and observations. This includes patients from the CUIMC stroke service (Figure 1, Table 1) that were part of a larger group of patients with acute cerebrovascular diseases and were prospectively identified upon admission to New York Presbyterian Hospital and recorded as part of daily research activities by a CUIMC stroke physician between 2011 and 2018. Two researchers (PT and BK) each manually reviewed 50 patients' charts for a total of 100 patients from this cohort to determine baseline false positive rates.

Patient population

We defined 3 case groups. We first included all patients from the CUIMC stroke service that were recorded as having AIS (cohort S). We then defined all patients in the CDW that met the Tirschwell-Longstreth (T-L) diagnosis code criteria for AIS (cohort T), which comprise ICD9-CM codes 434.x1, 433.x1, 436 (where x is any number) and the code is in the primary diagnostic position. [3] Our dataset did not specify the diagnostic position of codes. We also included ICD10-CM code equivalents, I63.xxx or I67.89, with the ICD10-CM codes being determined from ICD9-CM from Centers for Medicaid and Medicare Services (CMS) General Equivalence Mappings,] with a “10000” flag.[39] Because patients with cerebrovascular disease are also likely to have suffered AIS, but may not have an attached AIS-related diagnosis code, we also created a group of cases according to cerebrovascular disease-related ICD codes defined by the *ICD-9-Clinical Modification* (CM) Clinical Classifications Software tool (CCS), as well as their ICD10-CM equivalents (cohort C).[40]

We then defined 4 control groups (Figure 1, Table 1). First, we defined a control group of patients without AIS-related diagnosis codes (I). Due to the fact that cerebrovascular disease is a major risk factor for stroke,[41,42] and to test a more stringent control definition than that of group (I), we also defined an additional group without any of the CCS cerebrovascular disease codes defined in cohort (C). Then, we defined a control set using CCS cerebrovascular disease diagnosis codes other than AIS (CI). Because multiple clinical entities can present as AIS, we also defined a group of controls according to diagnosis codes for AIS mimetic diseases (N), including hemiplegic migraine (ICD9-CM 346.3), brain tumor (191.xx, 225.0), multiple sclerosis (340), cerebral hemorrhage (431), and hypoglycemia with coma (251.0). Finally, we identified a control group culled from a random sample of patients (R).

Model features

From the CDW, we gathered race, ethnicity, age, sex, diagnostic and procedure insurance billing codes as well as medication prescriptions for all patients. We dichotomized each feature based on its presence or

absence in the data. Because *Systematized Nomenclature of Medicine* (SNOMED) concept IDs perform similarly to ICD9-CM and ICD10-CM codes for phenotyping, [43] we mapped diagnoses and procedure features from ICD9-CM, ICD10-CM, and *Current Procedural Terminology 4* (CPT4) codes to SNOMED concept IDs using the OHDSI OMOP mappings, and used *RxNorm* IDs for medication prescriptions. We identified patients with Hispanic ethnicity using an algorithm combining race and ethnicity codes.[44] The most recent diagnosis in the medical record served as the age end point and we dichotomized age as greater than or equal, or less than 50 years. We excluded from our feature set any diagnosis codes that were used in any case or control definitions. Because approximately 5 million patients exist in the CUIMC CDW without a cerebrovascular disease diagnosis code, we addressed this large resultant imbalance in cases and controls by randomly sampling controls to create a balanced, or 1:1 case to control ratio. In addition, we set the maximum sample size to 16,000 patients in order to control the size of the feature set. See Supplementary Methods for model development.

Internal validation using all EHR patients

To identify the number of patients classified as having AIS in our institutional EHR, we applied each of the 75 models to the entire patient population in the CUIMC CDW with at least one diagnosis code. We chose a probability threshold based on the maximum F1 score determined for each model from the training set. We also determined the percentage of patients that had AIS ICD9-CM codes as defined by T-L criteria and associated ICD10-CM codes.

External validation

The UK Biobank is a prospective health study of over 500,000 participants, ages 40-69, containing comprehensive EHR and genetic data.[45] Given that this dataset contains 4,922 patients with an AIS related ICD10 code, similar to our T case cohort criteria, and 163 patients with self-reported AIS, the UK Biobank can evaluate our machine learning models' ability to recover potential AIS patients that lack AIS-related ICD10 codes. In a systemic review, the UK Biobank Stroke Outcomes group found positive predictive value between 22-87% and negative predictive value between 88-99% for self-reported strokes.[46] One difference between the UK Biobank definition of the AIS related ICD10 codes and our definition is their addition of code I64, which translates as "Stroke, not specified as haemorrhage or infarction". We chose the most accurate and robust case-control combination from our models (cases defined by the T-L AIS codes (T) and controls without codes for cerebrovascular disease (C) in a 1:1 case-control ratio as our training set) to train the phenotyping model using conditions specified by ICD10 codes, procedures specified by OCPS4 codes, medications specified by RxNorm codes, and demographics as features, excluding features that were used to create the training and testing cohorts. We trained on half of the patients with AIS related ICD10 codes, and then tested our models on the rest of the UK Biobank data which included self-reported AIS cases and the other half of the patients with AIS related ICD10 codes. We added these patients to improve the power of detecting cases, and we removed the AIS related ICD10 codes from our feature set to prevent recovery of patients due to these codes. We

resampled the control set 50 times and evaluated the performance of the algorithm through AUROC, AP, and precision at the top 50, 100, 500 and 2,624 patients (ordered by model probability).

Declarations

Ethics approval and consent to participate

Use of Columbia patient data was approved by Columbia's institutional review board and UK Biobank data approved with UK Biobank Research Ethics Committee (REC) approval number 16/NW/0274. This research has been conducted using the UK Biobank Resource under Application Number 41039.

Consent for publication

Not applicable.

Availability of data and materials

Code for analysis and figure generation in this study is publicly available online at https://github.com/pthangaraj/Stroke-Phenotyping. Electronic health record data structured in OMOP OHDSI format is required for use.

Competing interests

None.

Funding

PMT is funded by F30HL14094601, and previously was funded by 5T32GM007367, 5R01GM107145 and 10T3TR002027. NPT is funded by R35GM131905 and was funded by 5R01GM107145.

Authors' contributions

PMT and NPT designed the study and drafted the original manuscript; BRK and MSE provided list of stroke service patients; BRK and PMT performed chart evaluation of stroke service patients; PMT and NPT performed analyses and wrote code. TL wrote code; MSE and NPT provided supervision; and PMT, BRK, TL, MSE, and NPT provided critical feedback on the manuscript.

Acknowledgements

An earlier version of this manuscript was included in the doctoral dissertation of Phyllis M. Thangaraj. We thank Dr. Patrick Ryan, Dr. Fernanda Polubriaginof, Dr. Theresa Koleck, Dr. Prashanth Selvaraj, Dr. Rami Vanguri, Dr. Joseph Romano, Alexandre Yahi, and Dr. Kayla Quinnies for their feedback and guidance.

References

1. Benjamin EJ, Virani SS, Callaway CW, et al. Heart Disease and Stroke Statistics—2018 Update: A Report From the American Heart Association. *Circulation*. 2018;137:e67–e492.

2. Arch AE, Weisman DC, Coca S, et al. Missed Ischemic Stroke Diagnosis in the Emergency Department by Emergency Medicine and Neurology Services. *Stroke*. 2016;47:668–673.

3. Tirschwell DL, Longstreth Jr WT. Validating Administrative Data in Stroke Research. *Stroke*. 2002;33:2465–2470.

4. Benesch C, Witter D, Wilder A, et al. Inaccuracy of the International Classification of Diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology*. 1997;49:660–664.

5. Weiskopf NG, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*. 2013;46:830–836.

6. Mo H, Thompson WK, Rasmussen LV, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *Journal of the American Medical Informatics Association*. 2015;22:1220–1230.

7. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assn*. 2014;21:221–230.

8. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assn*. 2013;20:117–121.

9. Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. *Amia Annu Symposium Proc Amia Symposium Amia Symposium*. 2011;2011:189–96.

10. Peissig, P, Costa, V, Caldwell, M, Rottscheit, C, Berg, R, Mendonca, E, and Page, D, “Relational machine learning for electronic health record-driven phenotyping,” Journal of Biomedical Informatics, vol. 52, 260–270, 2014.

11. Chen, Y, Carroll, R, Hinz, E, Shah, A, Eyler, A, Denny, J, and Xu, H, “Applying active learning to high-throughput phenotyping algorithms for electronic health records data,” Journal of the American Medical Informatics Association, vol. 20, no. e2, e253–e259, 2013.

12. Yu S, Chakrabortty A, Liao KP, et al. Surrogate-assisted feature extraction for high throughput phenotyping. *J Am Medical Informatics Assoc Jamia*. 2016;ocw135.

13. Ning W, Chan S, Beam A, et al. Feature Extraction for Phenotyping from Semantic and Knowledge Resources. *Journal of biomedical informatics*. 2019;103122.

14. Yu S, Ma Y, Gronsbell J, et al. Enabling phenotypic big data with PheNorm. *Journal of the American Medical Informatics Association : JAMIA*. 2017;

15. Agarwal V, Podchiyska T, Banda JM, et al. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association : JAMIA*. 2016;23:1166–1173.

16. Halpern Y, Horng S, Choi Y, et al. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association : JAMIA*. 2016;23:731–40.

17. Murray, SG, Avati, A, Schmajuk, G, and Yazdany, J. "Automated and flexible identification of complex disease: Building a model for systemic lupus erythematosus using noisy labeling.," Journal of the American Medical Informatics Association: JAMIA, vol. 26, no. 1, 61–65, 2019.

18. B. K. Beaulieu-Jones, C. S. Greene, and Pooled Resource Open-Access ALS Clinical Trials Consortium, "Semi-supervised learning of the electronic health record for phenotype stratification," Journal of Biomedical Informatics, vol. 64, pp. 168–178, 2016.

19. C. Walsh and G. Hripcsak, "The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions," Journal of Biomedical Informatics, vol. 52, pp. 418–426, Dec. 2014.

20. A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, and N. Elhadad, "Diagnosis code assignment: Models and evaluation metrics," Journal of the American Medical Informatics Association, vol. 21, no. 2, 231–237, 2014.

21. Y. Zhang, "A hierarchical approach to encoding medical concepts for clinical notes," Association for Computational Linguistics, 67–72, 2008.

22. C. G. Walsh, K. Sharman, and G. Hripcsak, "Beyond discrimination: A comparison of calibration methods and clinical usefulness of predictive models of readmission risk," Journal of Biomedical Informatics, vol. 76, pp. 9–18, Dec. 2017.

23. Ni Y, Alwell K, Moomaw CJ, et al. Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis. *Plos One*. 2018;13:e0192586.

24. Imran TF, Posner D, Honerlaw J, et al. A phenotyping algorithm to identify acute ischemic stroke accurately from a national biobank: the Million Veteran Program. *Clin Epidemiology*. 2018;10:1509–1521.

25. V. Abedi, N. Goyal, G. Tsivgoulis, N. Hosseinichimeh, R. Hontecillas, J. Bassaganya-Riera, L. Elijovich, J. E. Metter, A. W. Alexandrov, D. S. Liebeskind, and et al., "Novel screening tool for stroke using artificial neural network.," Stroke, vol. 48, no. 6, 1678–1681, 2017.

26. Z. Chen, R. Zhang, F. Xu, X. Gong, F. Shi, M. Zhang, and M. Lou, "Novel prehospital prediction model of large vessel occlusion using artificial neural network.," Frontiers in aging neuroscience, vol. 10, p. 181, 2018.

27. W. Hersh, M.Weiner, P. Embi, J. Logan, P. Payne, E. Bernstam, H. Lehmann, G. Hripcsak, T. Hartzog, J. Cimino, and J. Saltz, "Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research," Medical Care, vol. 51, Aug. 2013.

28. J. M. Overhage and L. M. Overhage, "Sensible use of observational clinical data," Statistical Methods in Medical Research, vol. 22, no. 1, pp. 7–13, Feb. 2013.

29. R. M. Kaplan, D. A. Chambers, and R. E. Glasgow, "Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias," Clinical and Translational Science, vol. 7, no. 4, pp. 342–346, 2014.

30. S. Schneeweiss and J. Avorn, "A review of uses of health care utilization databases for epidemiologic research on therapeutics," Journal of Clinical Epidemiology, vol. 58, no. 4, pp. 323–337, Apr. 2005.

31. N. G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, "Defining and measuring completeness of electronic health records for secondary use," Journal of biomedical informatics, vol. 46, no. 5, 830–836, 2013.

32. Weiskopf NG, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46:830–836.

33. Sinnott JA, Cai F, Yu S, et al. PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies. *Journal of the American Medical Informatics Association : JAMIA*. 2018;

34. Sinnott JA, Dai W, Liao KP, et al. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human genetics*. 2014;133:1369–82.

35. Bastarache L, Hughey JJ, Hebbring S, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science*. 2018;359:1233–1239.

36. Son JH, Xie G, Yuan C, et al. Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes. *American journal of human genetics*. 2018;103:58–73.

37. Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. *Journal of the American Medical Informatics Association : JAMIA*. 2017

38. Reich, C., and Ryan, P.B., and Belenkaya, R., Natarajan,K., and Blacketer, C. OMOP Common Data Model v6.0 Specifications. https://github.com/OHDSI/CommonDataModel/wiki Accessed September 2019

39. 2018 ICD-10 CM and GEMs. U.S. Centers for Medicare & Medicaid Services. https://www.cms.gov/medicare/coding/icd10/2018-icd-10-cm-and-gems.html. Accessed February 2018.

40. HCUP CCS-Services and Procedures. Healthcare Cost and Utilization Project (HCUP). March 2017. Agency for Healthcare Research and Quality. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed March 2019.

41. Boehme AK, Esenwa C, Elkind M. Stroke Risk Factors, Genetics, and Prevention. *Circ Res*. 2017;120:472–495.

42. Benjamin EJ, Blaha MJ, Chiuve SE, et al. Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. Circulation. 2017;135(10):e146. Epub 2017 Jan 25

43. Hripcsak G, Levine ME, Shang N, et al. OUP accepted manuscript. *J Am Med Inform Assn*. 2018;

44. Polubriaginof F, Vanguri R, Quinnies K, et al. Disease Heritability Inferred from Familial Relationships Reported in Medical Records. *Cell*. 2018;173:1692-1704.e11.

45. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *Plos Med*. 2015;12:e1001779.

46. Woodfield, R., Group, U. B. S. O., Group, U. B. F. and O. W. & Sudlow, C. L. M. Accuracy of Patient Self-Report of Stroke: A Systematic Review from the UK Biobank Stroke Outcomes Group. *PLOS ONE* **10**, e0137538 (2015).

Figures

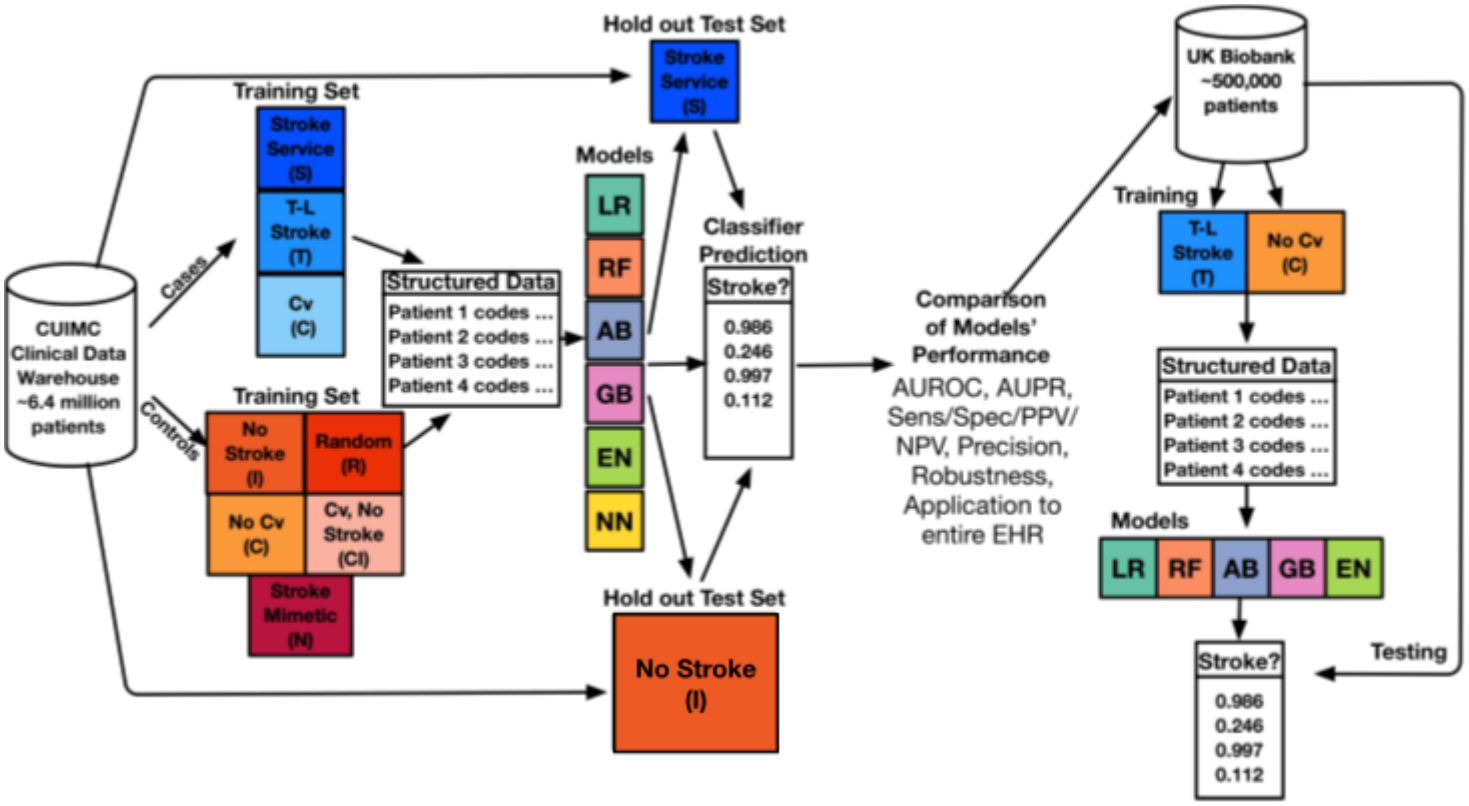

Figure 1

Schematic of Model Training, Testing, Evaluation, and Application to UK Biobank. See methods for case/control abbreviations. Case: Control ratio was 1:1, subjects overlapping in the case and control definitions were removed from the control set, and subjects overlapping between the training and testing sets were removed from the testing set before any training or testing. Models included Random Forest (RF), Logistic Regression with L1 penalty (LR), Neural Network (NN), Gradient Boosting (GB), Logistic Regression with Elastic Net Penalty (EN) and Adaboost (AB). AUROC: Area Under the Receiver Operating Curve, AUPR: Area under the Precision-Recall Curve, Sens: Sensitivity, Spec: Specificity, PPV: Positive Predictive Value, NPV: Negative Predictive Value.

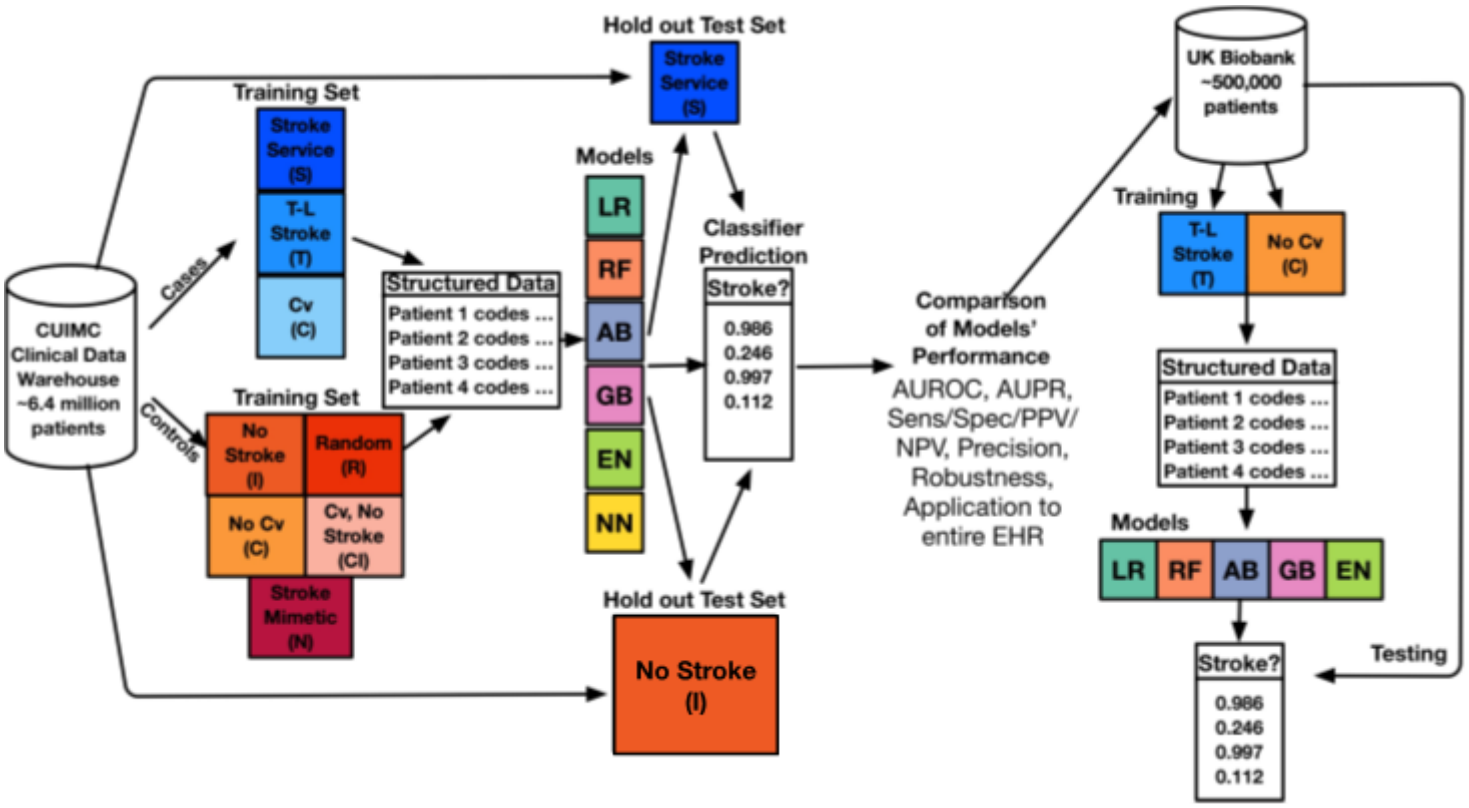

Figure 1

Schematic of Model Training, Testing, Evaluation, and Application to UK Biobank. See methods for case/control abbreviations. Case: Control ratio was 1:1, subjects overlapping in the case and control definitions were removed from the control set, and subjects overlapping between the training and testing sets were removed from the testing set before any training or testing. Models included Random Forest (RF), Logistic Regression with L1 penalty (LR), Neural Network (NN), Gradient Boosting (GB), Logistic Regression with Elastic Net Penalty (EN) and Adaboost (AB). AUROC: Area Under the Receiver Operating Curve, AUPR: Area under the Precision-Recall Curve, Sens: Sensitivity, Spec: Specificity, PPV: Positive Predictive Value, NPV: Negative Predictive Value.

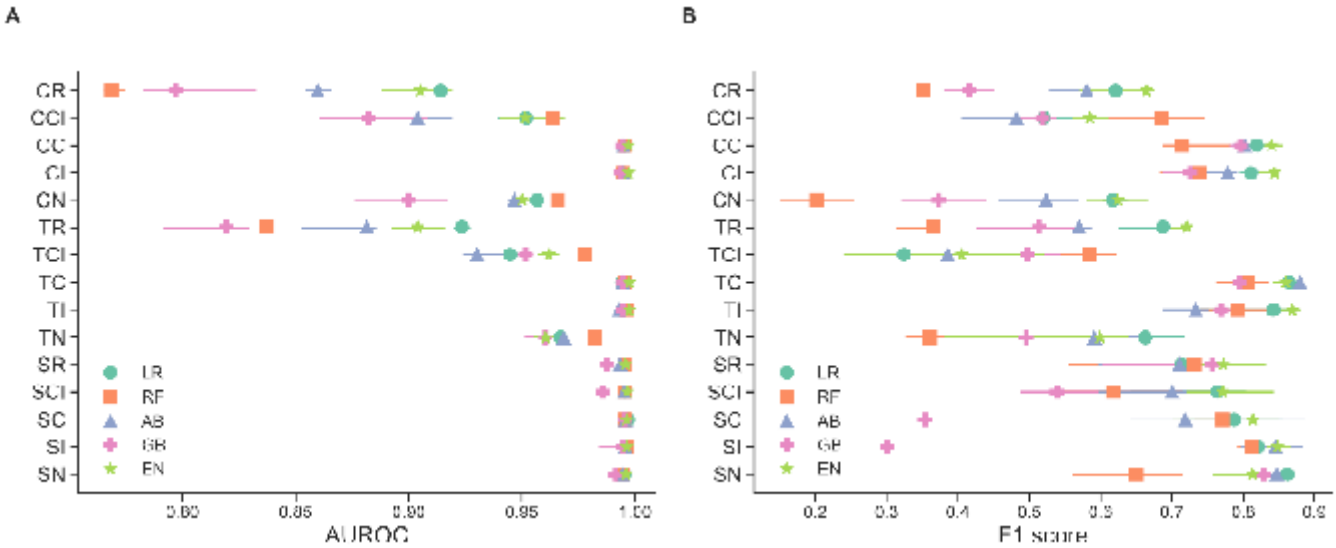

Figure 2

Performance of select models on Stroke Service holdout test set ((a): AUROC (circle: median, bars: 50% CI), (b): F1 (circle: median, bars: 50% CI)). Different combinations of cases and controls are shown on the y-axis. (LR) logistic regression with l1 penalty, (RF) random forest, (AB) AdaBoost, (GB) gradient boosting, (EN) logistic regression with elastic net penalty. Different combinations of cases and controls are shown on the y-axis. Cases (first letter) may be one of cerebrovascular (C), T-L (T), or Stroke Service (S). Controls (second and third letters) may be one of random (R), cerebrovascular disease but no AIS code (CI), no cerebrovascular disease (C), no AIS code (I), or a stroke mimetic disease (N), See Methods and Supplementary Table 1 for definitions of sets. Threshold to compute the F1 score on the testing set was chosen as the threshold that yielded the maximum F1 in cross-validation on the training set (Methods, Supplementary Table 10).

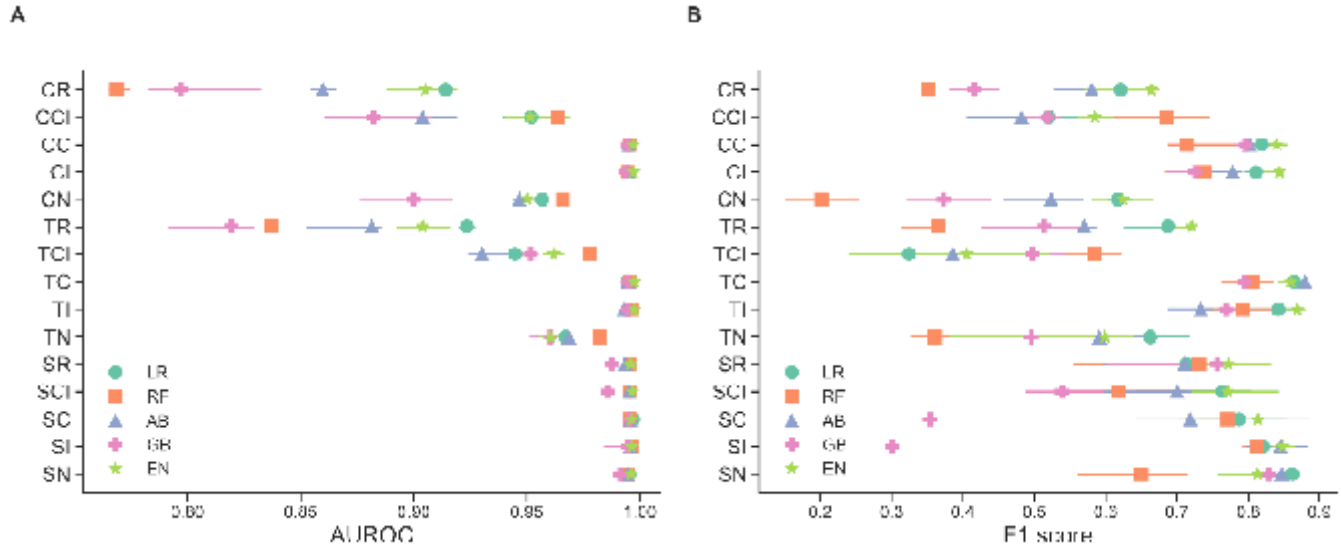

Figure 2

Performance of select models on Stroke Service holdout test set ((a): AUROC (circle: median, bars: 50% CI), (b): F1 (circle: median, bars: 50% CI)). Different combinations of cases and controls are shown on the y-axis. (LR) logistic regression with l1 penalty, (RF) random forest, (AB) AdaBoost, (GB) gradient boosting, (EN) logistic regression with elastic net penalty. Different combinations of cases and controls are shown on the y-axis. Cases (first letter) may be one of cerebrovascular (C), T-L (T), or Stroke Service (S). Controls (second and third letters) may be one of random (R), cerebrovascular disease but no AIS code (CI), no cerebrovascular disease (C), no AIS code (I), or a stroke mimetic disease (N), See Methods and Supplementary Table 1 for definitions of sets. Threshold to compute the F1 score on the testing set was chosen as the threshold that yielded the maximum F1 in cross-validation on the training set (Methods, Supplementary Table 10).

A

Feature (SNOMED ID)	% Models	Ave Freq. Cases	Ave Freq. Controls
CT head or brain; no contrast (2211327)	64%	54%	14%
Age > 50	57%	85%	50%
Duplex scan of extracranial arteries; bilateral (2313975)	55%	28%	8%
Aspirin (1112807)	44%	61%	18%
MRI, Brain and Brain stem; no contrast (2211351)	37%	36%	7%
Transcranial Doppler study intracranial arteries (2313977)	33%	12%	1%
Atorvostatin (1545958)	32%	48%	11%
Unspecified essential hypertension (44821949)	31%	64%	26%
MR angiography, neck; no contrast (2211348)	27%	55%	4%
Level IV - Surgical pathology, gross, and microscopic... (2213283)	27%	28%	24%
Pulmonary congestion and hypostasis (44825477)	20%	30%	3%
History of TIA or stroke w/o residual effect (45576200)	19%	6%	0.7%
Convulsions (44823442)	16%	3%	2%
Iatrogenic cerebrovascular infarction or hemorrhage (44834124)	13%	3%	0.3%

B

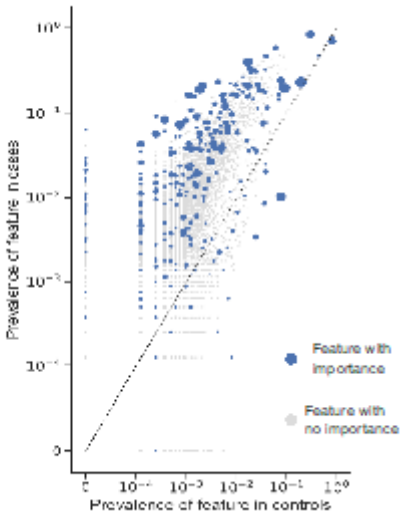

Figure 3

A: Common top 10 features in the models. After each of the 75 models were trained, we counted the number of times each feature was represented as one of the top ten by absolute coefficient weight, for methods like logistic regression, or by feature importance, for methods like random forest. Above are features from this analysis along with the proportion of models in which they were in the top ten (% Models), the average frequency in the cases (Ave. Freq. Cases) and the average frequency in the controls (Ave. Freq. Controls). B: Prevalence of features in cases vs controls in the TC AB model. Axes were on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

A

Feature (SNOMED ID)	% Models	Ave Freq. Cases	Ave Freq. Controls
CT head or brain; no contrast (2211327)	64%	54%	14%
Age > 50	57%	85%	50%
Duplex scan of extracranial arteries; bilateral (2313975)	55%	28%	8%
Aspirin (1112807)	44%	61%	18%
MRI, Brain and Brain stem; no contrast (2211351)	37%	36%	7%
Transcranial Doppler study intracranial arteries (2313977)	33%	12%	1%
Atorvostatin (1545958)	32%	48%	11%
Unspecified essential hypertension (44821949)	31%	64%	26%
MR angiography, neck; no contrast (2211348)	27%	55%	4%
Level IV - Surgical pathology, gross, and microscopic... (2213283)	27%	28%	24%
Pulmonary congestion and hypostasis (44825477)	20%	30%	3%
History of TIA or stroke w/o residual effect (45576200)	19%	6%	0.7%
Convulsions (44823442)	16%	3%	2%
Iatrogenic cerebrovascular infarction or hemorrhage (44834124)	13%	3%	0.3%

B

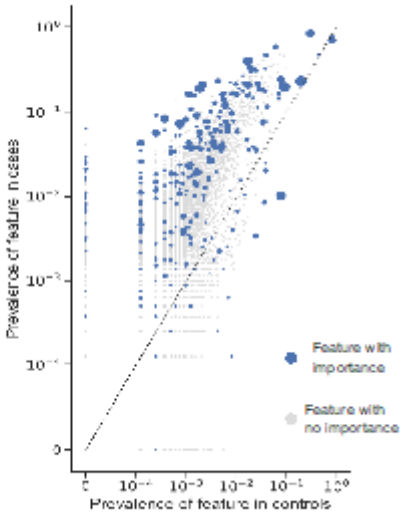

Figure 3

A: Common top 10 features in the models. After each of the 75 models were trained, we counted the number of times each feature was represented as one of the top ten by absolute coefficient weight, for

methods like logistic regression, or by feature importance, for methods like random forest. Above are features from this analysis along with the proportion of models in which they were in the top ten (% Models), the average frequency in the cases (Ave. Freq. Cases) and the average frequency in the controls (Ave. Freq. Controls). B: Prevalence of features in cases vs controls in the TC AB model. Axes were on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

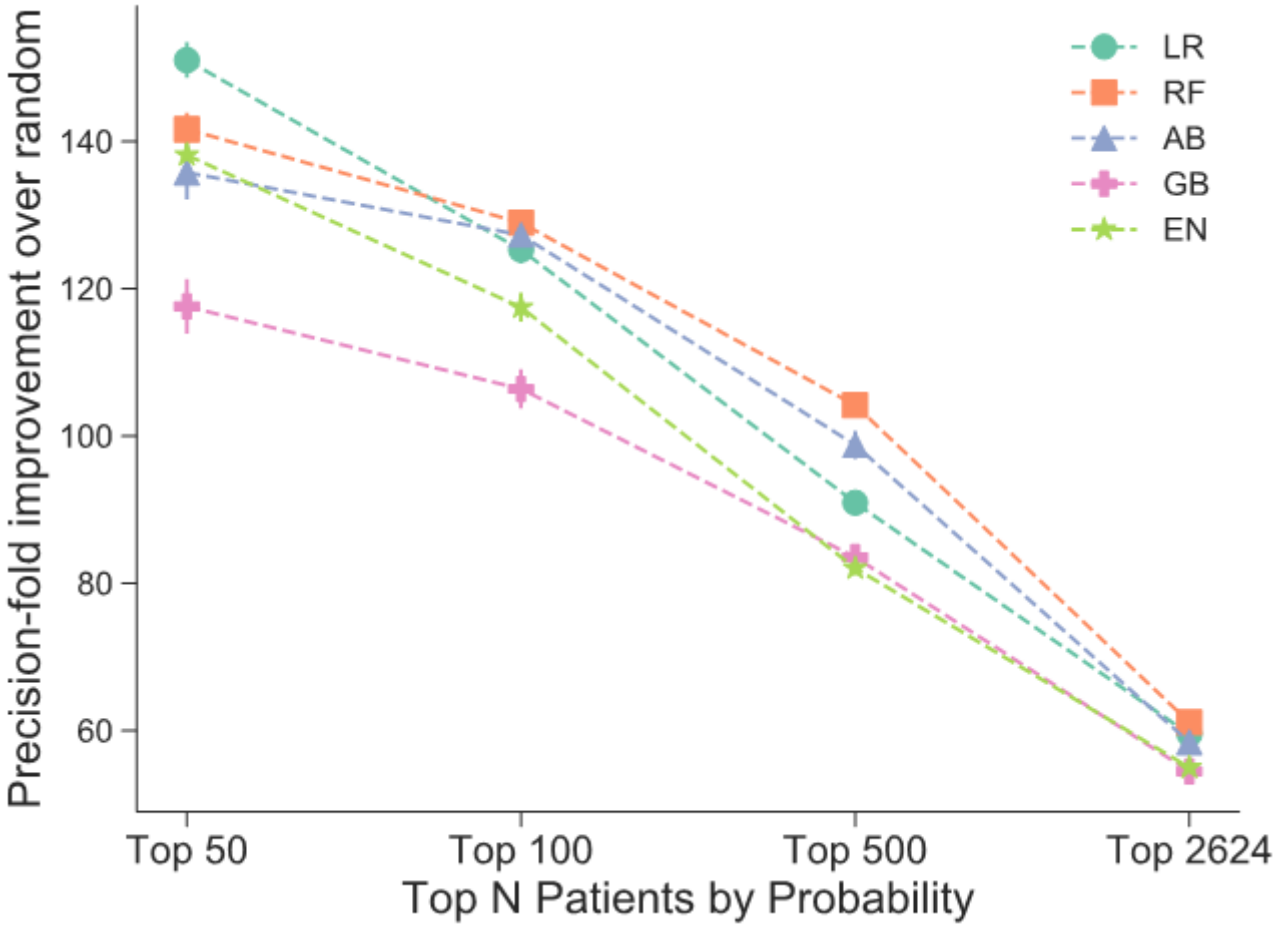

Figure 4

Precision-fold over random sampling of acute ischemic stroke cases without related ICD10 codes at top 50, 100, 500, and 2,624 patient probabilities assigned by machine learning algorithms. With 95% confidence intervals in error bars. See Supplementary Table 1 for model abbreviations' definitions.

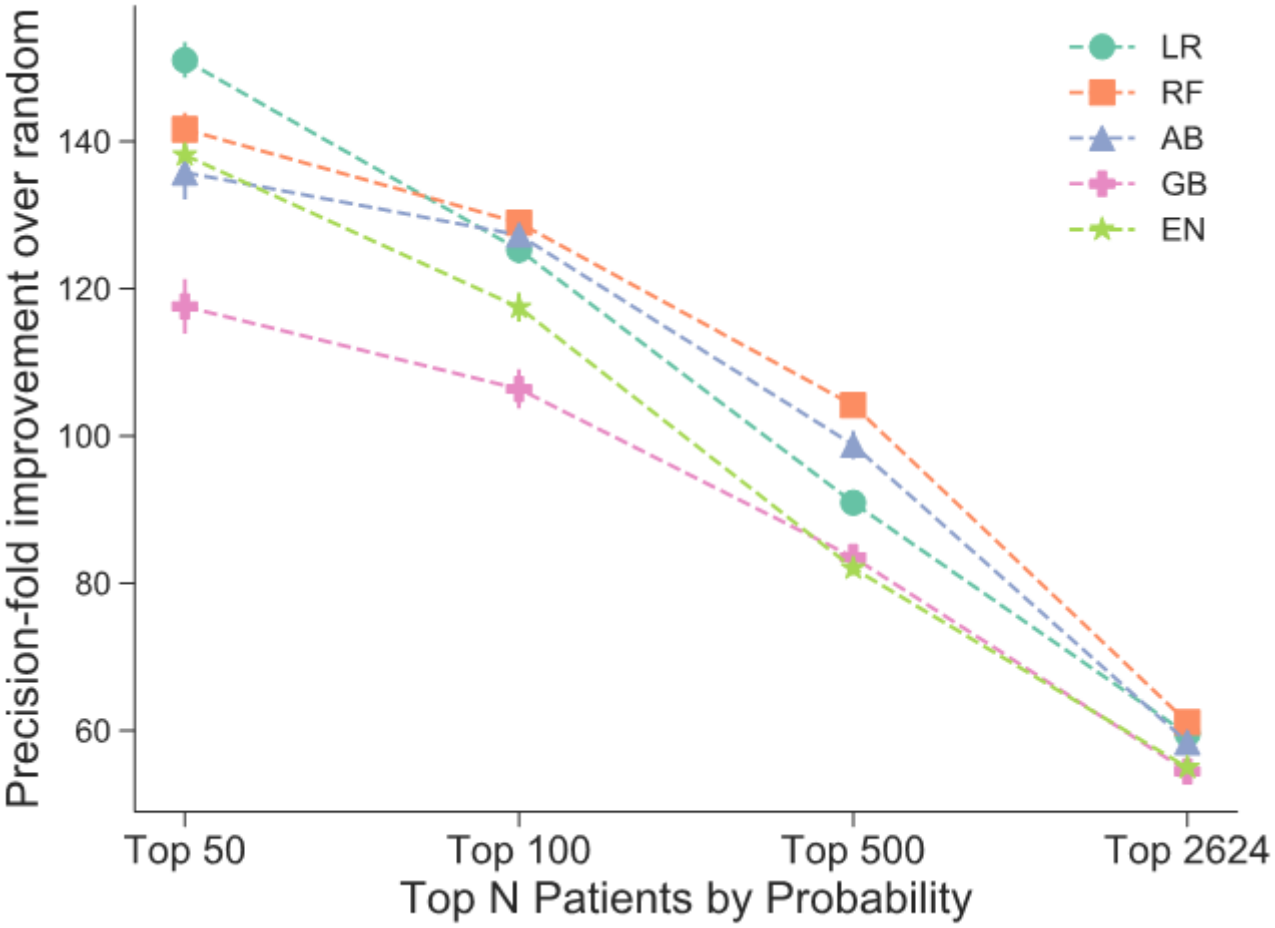

Figure 4

Precision-fold over random sampling of acute ischemic stroke cases without related ICD10 codes at top 50, 100, 500, and 2,624 patient probabilities assigned by machine learning algorithms. With 95% confidence intervals in error bars. See Supplementary Table 1 for model abbreviations' definitions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SUPPLEMENTARYMATERIALSBioDataMiningThangarajetalFinal.docx
- SUPPLEMENTARYMATERIALSBioDataMiningThangarajetalFinal.docx