ORIGINAL RESEARCH

A phenotyping algorithm to identify acute ischemic stroke accurately from a national biobank: the Million Veteran Program

Tasnim F Imran,[1–3,*] Daniel Posner,[1,4,*] Jacqueline Honerlaw,[1] Jason L Vassy,[1,2] Rebecca J Song,[1] Yuk-Lam Ho,[1] Steven J Kittner,[5] Katherine P Liao,[1,2] Tianxi Cai,[1,6] Christopher J O'Donnell,[1,2] Luc Djousse,[1,2] David R Gagnon,[1,4] J Michael Gaziano,[1,2] Peter WF Wilson,[7,8] Kelly Cho[1,2]

On behalf of the VA Million Veteran Program

[1]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Cooperative Studies Program, VA Boston Healthcare System, Boston, MA, USA; [2]Department of Medicine, Division of Aging, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [3]Department of Medicine, Cardiology Section, Boston Medical Center, Boston University School of Medicine, Boston, MA, USA; [4]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA; [5]Department of Neurology, Baltimore VA Medical Center and University of Maryland School of Medicine, Baltimore, MD, USA; [6]Harvard T. H. Chan School of Public Health, Boston, MA, USA; [7]Atlanta VA Medical Center, Decatur, GA, USA; [8]Department of Medicine, Division of Cardiovascular Disease, Emory University School of Medicine, Atlanta, GA, USA

*These authors contributed equally to this work

Correspondence: Kelly Cho
Massachusetts Veterans Epidemiology Research and Information Center, 150 South Huntington Avenue, Boston, MA 02130, USA
Tel +1 857 364 4523
Email Kelly.Cho@va.gov

Background: Large databases provide an efficient way to analyze patient data. A challenge with these databases is the inconsistency of ICD codes and a potential for inaccurate ascertainment of cases. The purpose of this study was to develop and validate a reliable protocol to identify cases of acute ischemic stroke (AIS) from a large national database.

Methods: Using the national Veterans Affairs electronic health-record system, Center for Medicare and Medicaid Services, and National Death Index data, we developed an algorithm to identify cases of AIS. Using a combination of inpatient and outpatient ICD9 codes, we selected cases of AIS and controls from 1992 to 2014. Diagnoses determined after medical-chart review were considered the gold standard. We used a machine-learning algorithm and a neural network approach to identify AIS from ICD9 codes and electronic health-record information and compared it with a previous rule-based stroke-classification algorithm.

Results: We reviewed administrative hospital data, ICD9 codes, and medical records of 268 patients in detail. Compared with the gold standard, this AIS algorithm had a sensitivity of 91%, specificity of 95%, and positive predictive value of 88%. A total of 80,508 highly likely cases of AIS were identified using the algorithm in the Veterans Affairs national cardiovascular disease-risk cohort (n=2,114,458).

Conclusion: Our algorithm had high specificity for identifying AIS in a nationwide electronic health-record system. This approach may be utilized in other electronic health databases to accurately identify patients with AIS.

Keywords: acute ischemic stroke, algorithm, large databases, big data, administrative health data, cerebrovascular accident

Introduction

Stroke is the second-most common cause of death worldwide and also a leading cause of significant morbidity.[1] The major types of stroke – ischemic, intracerebral hemorrhage, and subarachnoid hemorrhage – differ in their risk factors and clinical presentation. Even within these major types of stroke, there are subtypes with different risk factors and prognosis.[1] Therefore, studies with a large number of stroke cases are needed to examine associations with clinical outcomes and to characterize the phenotypes of these patients. Cohort studies have been used for such clinical research, but generally have relatively few outcome events and may lack the breadth of clinical data that is available in health-care databases.

The Veterans Health Administration (VHA) has one such health-care database, with over 15 years of electronic health records (EHRs) for millions of veterans. The records enable a multitude of observational research studies, but EHRs often lack the

precise and uniform diagnoses required for health research. In practice, this means many risk factors and health events must be inferred from the records' structured elements, such as standardized codes (ICD9-CM) and drug prescriptions, and unstructured elements, such as the narrative text contained in radiology and pathology reports. However, there are often inaccuracies in coding and limitations in the types of data elements available.[2–6] This leads to challenges in reliably distinguishing among major subphenotypes of stroke based on ICD9-CM codes alone (specifically ischemic and hemorrhagic), which have very different risk factors and pathophysiologic mechanisms. As such, it is essential to develop algorithms to identify specific types of stroke accurately. Furthermore, highly accurate phenotyping algorithms are urgently needed in ongoing and future genomic and mechanistic studies in megabiobanks like the Million Veteran Program (MVP), the UK Biobank, and the All of Us Program.

In this paper, we present a machine-learning algorithm for identifying acute ischemic stroke (AIS) from structured elements of VHA patient records, ie, a "structured stroke algorithm".[1] The machine-learning algorithm can be used for both longitudinal and case–control studies to identify cases of AIS and to quantify uncertainty in stroke classification via predicted stroke probabilities. The predicted stroke probabilities are used to develop a specialized algorithm for case–control studies, and uncertain stroke cases are removed from the analysis. Finally, we provide suggestions for choosing among competing algorithms and demonstrate that using health-care databases to ascertain stroke cases is practical, low-cost, and high-yield.

Methods
Data sources
Using the national Veterans Affairs (VA) EHR system, Centers for Medicare and Medicaid Services, and National Death Index data, we put together a cohort of patients with AIS. The VHA system is the largest single-payer health-care entity in the US, containing over 144 hospitals and 1,221 outpatient centers.[8] Local hospital and clinic data, including inpatient, outpatient, laboratory values, and vital signs, are stored in a central VHA corporate data warehouse. Our current study includes creation and validation of cases of AIS from the national VHA cohort. We also present stroke characteristics of the patient population in a cardiovascular disease (CVD)-risk cohort and the MVP cohort.

Cardiovascular disease and risk factors: the CVD in veterans study
The purpose of the CVD in veterans study is to determine risk for CVD, including heart disease and stroke, in veterans by examining existing EHR data. The VA CVD-risk cohort includes any veteran who has received care and had a cholesterol measurement at the VA from 2002 onwards. Prediction equations will be developed from the experience of the veterans who receive care across the US. The three specific aims of the CVD events in veterans study are to estimate risk for first CVD events in veterans, estimate risk for first CVD in veterans with diabetes mellitus, and estimate risk for recurrent CVD events in veterans who have already developed CVD. Implementation of this strategy should lead to veteran-specific health care that is individualized in terms of total CVD-risk management.

Million Veteran Program
The MVP was launched to establish a national, representative, and longitudinal study of veterans for genomic and nongenomic research that combines data from self-reported surveys, EHRs, and biospecimens. Details of this VA-based megabiobank have been published.[9] Briefly, the VHA population consists of approximately 21 million veterans, of whom 8.9 million are users of the VHA. The MVP is an ongoing study that is recruiting at a steady state at approximately 50 VA medical facilities throughout the country. The source population consists of those who are active users of the VHA with a valid mailing address and able to provide informed consent (~6.9 million veterans). The target number of study participants for the MVP is at least 1 million veterans from the eligible pool. As of May 2017, 570,946 veterans have enrolled in MVP (359,149 with a baseline survey). Among those with genotyping data (352,708), 240,230 had a baseline survey with self-reported demographics. Of these, 92% are men with a mean age of 64.5±12.4 years, 57.9% are aged 50–69 years, 74.9% are White, and 19.1% African-American (based on self-reported race). Ultimately, the goal is to link clinical data in the VA EHR in the computerized patient-record system with MVP questionnaires and a MVP genome-wide association study (GWAS), to allow integration of genetic testing in the future, thus contributing to precision medicine. The Institutional Review Board at the VA Boston Healthcare System approved this study.

Selection of cases for review
A total of 300 veterans were sampled from a random subset of 1 million veterans who came to the VA for an outpatient or

inpatient visit (labeled "data mart" in Figure 1). The sample size was selected based on the average time per chart review and limit of 40 hours per chart reviewer (30 minutes/chart × 300 charts × two reviewers per chart/seven reviewers = 42 hours per reviewer). The data mart was used to reduce the computation and memory burden. The first 100 veterans were randomly sampled from the data mart. The 100 veterans had comparable demographics, comorbidity profiles, and visit histories compared to general VA users. Of these 100 veterans, chart reviewers found four veterans with AIS (cases) and 96 veterans without AIS or other strokes (controls). ICD9 codes 430 and 433–437 were completely absent in controls. To gather enough cases of AIS, we randomly selected 200 veterans from the data mart for stroke ICD9 codes 430 and 433–437 (labeled "stroke mart" in Figure 1), resulting in 81 controls (including two transient ischemic attacks [TIAs] and five hemorrhagic strokes), 87 definite AIS cases, and 32 possible AIS cases. We tried several selection criteria to yield the most AIS cases: an inpatient or outpatient ICD9 code, an inpatient ICD9 code and multiple outpatient codes, or other combinations thereof (Table S1). Each chart was reviewed by two of seven reviewers (six physicians and one nurse), with disagreements settled by a third reviewer when necessary. Because charts were sometimes reviewed by more than two reviewers, we assessed interrater reliability with the Fleiss–Cuzick (κ-like) measure of agreement,[10] which is comparable to Cohen's κ[19] and measures interrater agreement in excess of agreement between completely random ratings. Fleiss's κ was computed from three categories: non-AIS (including

hemorrhagic strokes and TIAs), possible AIS, and definite AIS. Agreement was somewhat low (κ=0.612), due to missing primary data in possible AIS cases, ie, no documentation of initial presentation of stroke and/or imaging data available to be certain whether the patient had truly had a stroke.

Chart-review validation

We used a Microsoft Access database on the Vinci platform to perform chart reviews. The records included both structured and unstructured elements. This database included a list of charts with patient demographics, including date of birth, date of death, sex, and race; ICD9 codes for stroke diagnosis, date of the ICD9 code, code description, and visit type (outpatient or inpatient); text-integration utility or unstructured notes or clinician notes extracted from the medical record (these included mainly admission notes, discharge summaries, notes by neurologists or cardiologists, and some progress notes), note title, and the number of notes in view; underlying cause of death from the National Death Index (data collected by the Centers for Disease Control from death certificates); ICD9 codes from Centers for Medicare and Medicaid Services billing records for non-VA providers; and radiology notes (which included radiology-procedure name and description of findings).

Figure 2 illustrates the flowchart of the chart-review process, and Figure S1 depicts the organization of the database platform. In addition to the ICD codes, chart review for AIS was validated using the American Heart Association/American Stroke Association definition of AIS. AIS is defined as
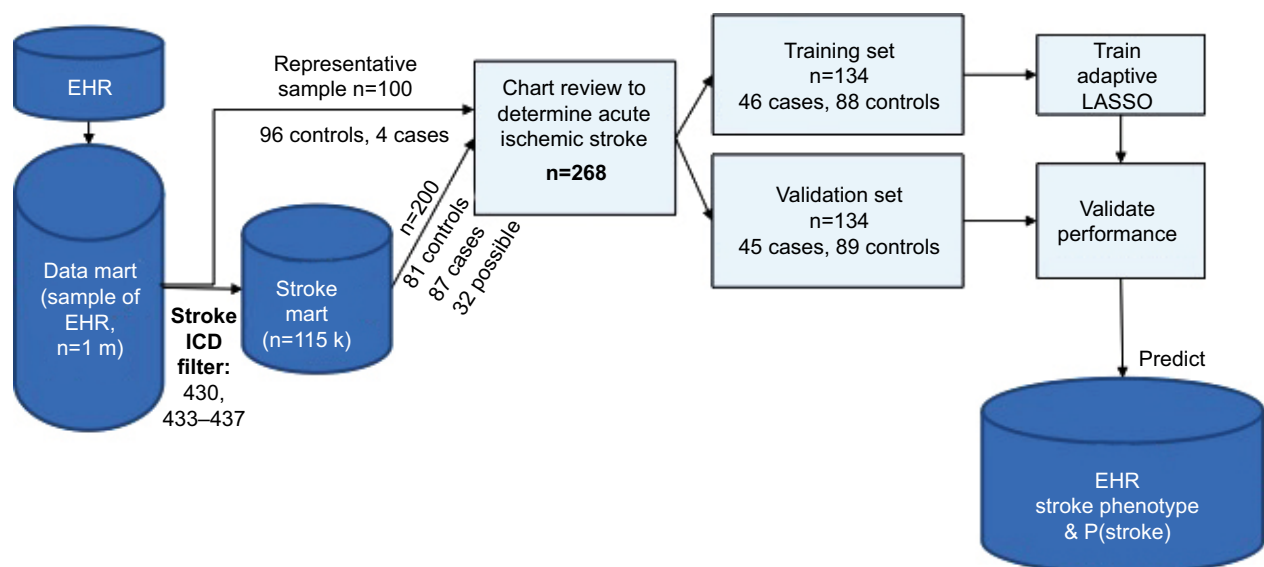
Figure 1 Stroke-classification flowchart for chart reviews.
Abbreviations: EHR, electronic health record; P(stroke), probability of stroke.

an episode of neurological dysfunction that is due to a focal cerebral, spinal, or retinal infarction. This can be based on pathological evidence, radiographic evidence, or other objective evidence of focal ischemia in a vascular territory or clinical evidence of focal ischemic injury based on symptoms that last >24 hours, with other etiologies excluded. Silent stroke or "historical stroke" was defined as central nervous system infarction with imaging evidence of infarction, but without a clear history of acute neurological symptoms.[20]

Below these data elements, a reviewer's comments section was provided. For instance, if it was not clear whether the patient had had a stroke from the data provided or if the patient was ruled out for a stroke, these comments would be listed in this section. The next section included classification (stroke event, historical stroke, family history of stroke, no event, relevant medication for stroke prevention, or "unable to decide" if it was not clear whether the patient had had an acute stroke from the data provided). If the chart reviewer selected stroke event, he/she would then be prompted to enter the classification type of the stroke event: AIS, hemorrhagic

stroke, stroke unspecified, or TIA. In addition to stroke event, the reviewer was asked to specify the date of the event in the format of month, day, and year. If the date was unknown, this field was left blank. The reviewer also noted whether the stroke was a first or second event. Differing reviews for the same patient were resolved through discussion, which produced a single consensus label.

After the date, a field titled "source of decision" appeared, which prompted the reviewer to select which data sources were used from the aforementioned sources. Multiple sources could be selected here. In order to improve this database platform for subsequent chart reviews, the reviewer was asked to list supportive text or keywords that were helpful in making his/her decision.

Algorithm development and statistical analysis

The 300 charts were assigned a label of "no AIS" (n=177), "possible AIS" (n=32), or "definite AIS" (n=91), where "no AIS" included hemorrhagic strokes (n=5) and TIAs (n=2).
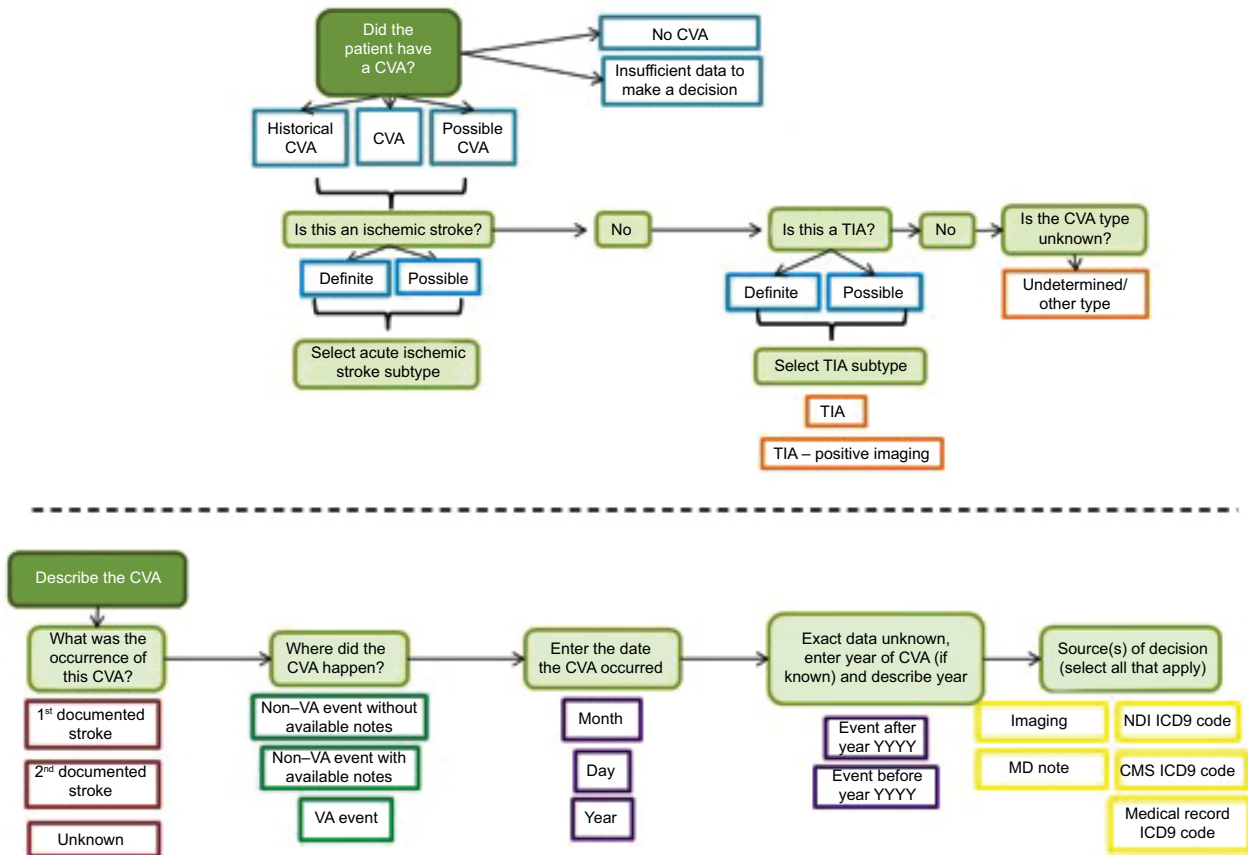

Figure 2 Development of structured acute ischemic stroke algorithm.
Abbreviations: CMS, Centers for Medicare and Medicaid Services; CVA, cerebrovascular accident; NDI, National Death Index; TIA, transient ischemic attack; VA, Veterans Affairs.

Charts labeled "possible AIS" had insufficient evidence of AIS, ie, relevant physician notes were present, but primary imaging data or a clinical exam at diagnosis absent. We treated possible AIS charts as missing data, excluded them from training the algorithm, and later compared them against other stroke categories using algorithm-predicted stroke probabilities (Figure 3).

Covariates considered for the algorithm are presented in Table 1. Only stroke-specific (or CVD-related) procedure
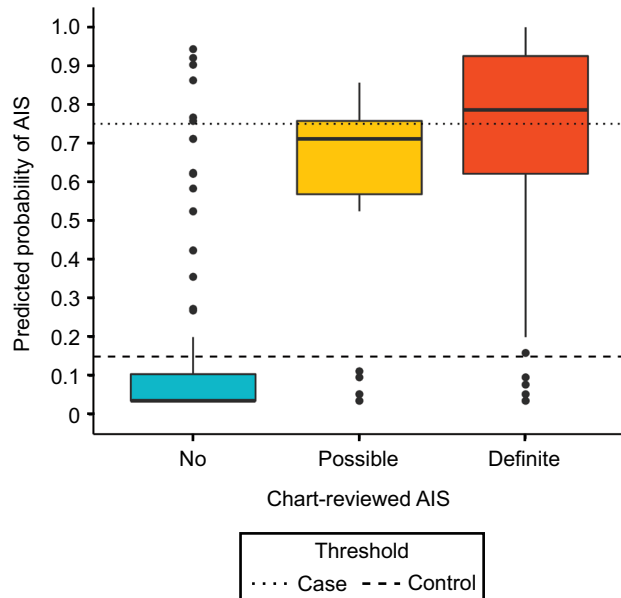
Figure 3 Predicted probabilities of stroke based on charts reviewed.
Notes: Thresholds optimized for largest n (excluding $P_{control}<P<P_{case}$) with Cohen's $\kappa>0.9$ between algorithm labels and review labels.
Abbreviation: AIS, acute ischemic stroke.

codes, medication, such clinical parameters as blood pressure, and cause of death were included. Categorical covariates, such as individual ICD9 codes, were summarized with counts (number of recorded codes) and indicators (0, absence; 1, presence of a comorbid condition). Each medication was coded as three covariates: indication of use, total dosage, and mean daily dosage. Maximum and mean systolic blood pressure (mmHg) were also included.

Algorithm development is presented in Figure 1. First, the data (n=268) were randomly split into a training set (n=134, 46 cases and 88 controls) and validation set (n=134, 45 cases, 89 controls). Several adaptive least absolute shrinkage and selection operator (LASSO)[11,12] models, ie, regularized logistic regressions, were trained over 50 repetitions of ten-fold cross-validation, where regularization parameter λ and adaptive parameter γ (0.5, 1, or 2) were tuned to minimize cross-validated binomial deviance. Adaptive LASSO was chosen over other machine-learning algorithms because of its feature selection and transparency. Cross-validation was also used to choose the most informative set of predictors. In total, four sets of predictors were considered: the Tirschwell and Longstreth[7] classifier and all accompanying ICD9 codes that occurred in at least ten training observations, two sets with additional variables added to the baseline set, and a final set with all predictors (Table 1). The algorithm with the lowest deviance was then used to predict the probability of AIS and no AIS in the validation data. R version 3.4.3 was used for all analyses.[13]

Classification metrics (sensitivity, specificity, positive predictive value [PPV], area under the curve) were evaluated

Table 1 List of predictors and variable importance in the acute ischemic stroke algorithm

Predictors	Description	β (log odds)
Intercept	Baseline risk (without predictors)	−3.288
Tirschwell and Longstreth classifier	AIS if 433.x1, 434, or 436*	2.286
Tirschwell and Longstreth ICD9	436	1.496
	Log (number of 434.91)	1
	433.x1	0
Other AIS ICD9 codes	Log (number of 433, 434, 436, 437.0, and 437.6)	0.586
ICD9, ICD10 codes	All CVD-related,# excluding above	0+
CPT codes$	MRI or CT brain/neck angiography	0+
CMS codes‡	All CVD and stroke-related	0+
Medication	All stroke-related	0+
Vitals	Systolic blood pressure	0+
Cause of death	Stroke, CVD-related, other	0+

Notes: *Excluding 434.x0, 430 (subarachnoid hemorrhage), and 431 (intracerebral hemorrhage). Counts of 433.x1 were unimportant after inclusion of the Tirschwell and Longstreth[7] classifier. +Not included in the optimal predictor set or estimated 0 by adaptive LASSO. #Diabetes (250), CHD-related (410–415, 427), cerebrovascular disease (430-438), hypotension (458), syncope (780.2), TIA Hx (V12.54), fall Hx (V15.88), aortocoronary bypass (V45.81), coronary angioplasty (V45.82). $Current procedural terminology (CPT) codes are used in the VHA for reporting medical services and procedures. ‡Procedure codes used by the Centers for Medicare and Medicaid Services (CMS).
Abbreviations: AIS, acute ischemic stroke; CVD, cardiovascular disease; MRI, magnetic resonance imaging; CT, computed tomography; LASSO, least absolute shrinkage and selection operator.

on the validation set for three models: the Tirschwell and Longstreth rule-based algorithm,[7] the best machine-learning model with default probability threshold (ie, longitudinal algorithm), and the best machine-learning model with optimal probability thresholds (ie, case–control algorithm). The default probability threshold assigned AIS to subjects with predicted probabilities >0.5 and no AIS to subjects with predicted probabilities <0.5. Optimal probability thresholds included a case threshold, above which subjects were predicted cases, and a control threshold, below which subjects were predicted controls. A subject between the case and control thresholds was considered "uncertain" and excluded when assessing the model's classification performance. The case and control thresholds were optimized in the training data to yield the largest sample size (excluding "uncertain" cases), with high agreement between algorithm and reviewer labels (Cohen's κ>0.9).

Clinical characteristics of patients with acute ischemic stroke in the nationwide VA cohort

Demographic characteristics

Age and race were extracted from the VHA electronic database, noted at the time of diagnosis of AIS. We categorized race according to standard categories already present in the VHA system: White, American Indian/Alaska native, Asian, Black/African-American, native Hawaiian or other Pacific Islander and other.

Cardiovascular risk factors and comorbid conditions

Smoking status was predicted from VHA lifestyle surveys and EHRs. Each subject was assigned one of three categories: current, past, or present smoker. To report smoking accurately, we used a probabilistic model to predict smoking status using structured data from the EHR, which included smoking-related health factors, diagnosis codes, medications, and outpatient-clinic specialty.[9] The MVP's self-reported smoking categories of never, former, and current from the baseline and lifestyle surveys were used as the gold standard. LASSO regression with tenfold cross validation was used to select the most meaningful predictors of smoking status and to apply a penalty to prevent overfitting. β-Coefficients were used to calculate the predicted probability of being a never, former, or current smoker. Subjects were assigned to the smoking group for which they had the highest predicted probability.

Body-mass index was calculated from weight in kilograms divided by height in meters squared, extracted from

medical records at the time of AIS diagnosis. Hypertension, hyperlipidemia, diabetes mellitus, atrial fibrillation, COPD, coronary artery disease, peripheral vascular disease, congestive heart failure, chronic kidney disease, chronic liver disease, pulmonary embolism, and deep-vein thrombosis were extracted from the patients' medical records using ICD9 codes (either one inpatient or two outpatient codes).

Clinical and laboratory values

Systolic and diastolic blood pressures are reported in mmHg. Total cholesterol, high-density-lipoprotein cholesterol, low-density-lipoprotein cholesterol, triglycerides, and hemoglobin A_{1c} were extracted from medical visits after the diagnosis of AIS. Two clinicians adjudicated laboratory values for accuracy. Estimated glomerular filtration rate was calculated from serum creatinine using the Chronic Kidney Disease Epidemiology Collaboration equation.[14]

Medications

Aspirin, clopidogrel, warfarin, statins, β-blockers, angiotensin-converting-enzyme inhibitors, and angiotensin-receptor blockers were obtained from prescription codes within the VA healthcare system. Two independent clinician reviewers adjudicated these medications. Of note, many patients at the VA obtain aspirin as an over-the-counter medication instead of through a prescription, and thus the reported percentage of aspirin use in our table is likely an underestimation.

Results

The final 268 "no stroke" and "definite stroke" chart-reviewed labels, in addition to the EHR elements (Table 1), were used to train the machine-learning algorithms. The algorithm assigned a diagnosis of AIS when probability of "definite stroke" exceeded a specified threshold. Similarly, a diagnosis of no stroke was assigned when probability of "definite stroke" fell under another specified threshold (Table 2). Two algorithms were developed: one for longitudinal cohort studies and one for case–control studies. For longitudinal studies, in which a stroke phenotype was necessary for every patient, case and control thresholds at 0.5 were most accurate. For case–control studies, in which we were able to exclude uncertain stroke cases and controls, a case threshold of 0.75 and control threshold of 0.15 yielded the highest sensitivity and specificity, excluding uncertain patients ($0.15 \leq P \leq 0.75$).

The best-performing model was adaptive LASSO regression with hyperparameters $\lambda=0.016$ and $\gamma=1$ trained with the second predictor set. The final predictors selected by LASSO were the Tirschwell and Longstreth classifier ($\beta_1=2.29$),

Table 2 Classification performance in the validation set (n=134)

Algorithm	Stroke*	No stroke	Sensitivity	Specificity	PPV	AUC
Tirschwell and Longstreth[#]			0.889 (0.81–0.96)[§]	0.83 (0.76–0.90)	0.727 (0.62–0.84)	
Longitudinal cohort	P[‡]>0.5	P<0.5	0.844 (0.75–0.94)	0.875 (0.82–0.93)	0.776 (0.68–0.87)	0.926 (0.89–0.96)
Case–control	P≥0.75	P≤0.15	0.906 (0.81–0.97)	0.946 (0.90–0.99)	0.879 (0.78–0.97)	0.948 (0.90–0.98)

Notes: *Decision rule for classifying acute ischemic stroke; [#]from Tirschwell and Longstreth;[7] [‡]predicted from classification model; [§]performance measure (bootstrapped 95% CI).
Abbreviations: PPV, positive predictive value; AUC, area under the curve.

indicator for ICD9 codes 436 (β_2=1.50), natural log of number of ICD9 code 434.91 (β_3=1.0), and log of number of ischemic stroke-related ICD9 codes 433, 434, 436, 437.0, and 437.6 (β_4=0.59). The predicted probability for subjects with no predictors was 0.036, positive Tirschwell and Longstreth classifier 0.269, only Tirschwell and Longstreth classifier and one instance of ICD9 434.91 0.424, and only Tirschwell and Longstreth classifier and ICD9 436 0.621. Additional ICD9 codes (433, 434, 436, 437.0, and 437.6) conferred additional risk above these baseline risks.

Our case–control algorithm had sensitivity of 91%, specificity of 95%, and PPV of 88%. It was slightly more sensitive than Tirschwell and Longstreth's algorithm (0.906 vs 0.889) and much more specific (0.946 vs 0.83), leading to a superior PPV (0.879 vs 0.727). The probability threshold improved on the Tirschwell and Longstreth algorithm by excluding subjects with uncertain stroke diagnoses (ie, predicted stroke probabilities between 0.15 and 0.75), but the exclusions led to fewer cases and controls. The longitudinal cohort algorithm had comparable classification performance to the Tirschwell and Longstreth algorithm, with lower sensitivity (0.844 vs 0.889), but higher specificity (0.875 vs 0.83) and PPV (0.776 vs 0.727).

Table 3 contains sample sizes and characteristics, including demographic, clinical variables, and comorbid conditions, when the case–control algorithm was applied to the VHA-wide cohort. Characteristics of patients with predicted AIS cases from the chart reviews (60 cases), the MVP (3,423 cases/323,122) and the VA-wide CVD-risk cohort (80,508 cases/2,114,458) are shown. For the VA-wide CVD-risk cohort, the mean age was 65±11 years, mean body-mass index 29 kg/m[2], 98% of patients were men, and 81% were White. The prevalence of smoking (current or past) was 82%, hypertension was 76%, and hyperlipidemia was 61%. In contrast, the MVP cohort had a slightly lower cumulative incidence of AIS, tended to be younger, with a mean age of 56 years, and had lower smoking prevalence of 41%. From the year 2000 onwards, the CVD-risk cohort had an estimated crude 15-year cumulative incidence of 80.8 cases/1,000

persons using the Tirschwell and Longstreth algorithm and 75.8 cases/1,000 persons using the longitudinal cohort algorithm. From 2000, the MVP cohort had estimated crude 15-year cumulative incidence of 25.7 cases/1,000 persons and 21.1 cases/1,000 persons with the Tirschwell and Longstreth and longitudinal cohort algorithms, respectively. Tables S2 and S3 provide the incidence per year based on the rule-based algorithm[7] and the longitudinal cohort algorithm in the national VA CVD-risk cohort and the MVP cohort.

Discussion

Our algorithm demonstrates that patients with AIS can be identified with both high sensitivity and specificity in a large national database. We present three different approaches to this task: a rule-based algorithm,[7] a machine-learning algorithm for longitudinal cohort studies, and a machine-learning algorithm for case–control studies that excludes patients with uncertain stroke diagnoses. The rule-based Tirschwell and Longstreth algorithm[7] yields high sensitivity and sample size, while the longitudinal cohort algorithm has increased specificity and PPV. The case–control algorithm has the highest specificity and PPV, but produces no classification for patients with uncertain stroke diagnoses and cannot be applied to longitudinal cohorts. The machine-learning algorithm produces probabilities, which can be used to recalibrate the model for different (non-VA) populations.[15]

Ideally, power analyses should be run to choose among the three models and determine the optimal trade off of sensitivity and specificity for each study. If a power analysis is unavailable, we suggest applying strict probability thresholds with high specificity to populations where disease prevalence is low. Both false positives and false negatives will reduce the effect size of a risk factor, but false positives and algorithm specificity have a higher impact when stroke is less prevalent. Finding the appropriate balance of sensitivity and specificity is especially important when effect sizes are small, such as in GWASs. High specificity and sample size are essential for GWASs, due to most genetic variants having small effects on complex phenotypes, such as stroke.

Table 3 Baseline characteristics of populations with predicted acute ischemic stroke (strict algorithm)

	Chart reviews	Million Veteran Program	CVD-risk cohort
Predicted strokes (n)/population size (n)	60/199	3,423/323,122	80,508/2,114,458
Age (years), mean ± SD	59.0±11.4	56.4±9.7	64.8±11
<30 years	0	25 (0.7%)	147 (0.2%)
30–49.99 years	5 (9.8%)	694 (20.5%)	5,741 (8.2%)
50–59.99 years	23 (45.1%)	1,570 (46.5%)	19,305 (27.4%)
60–69.99 years	7 (13.7%)	776 (23%)	19,453 (27.7%)
>70 years	16 (31.4%)	313 (9.3%)	25,687 (36.5%)
Men	51 (100%)	3,241 (97.5%)	67,583 (98.1%)
Race			
White	30 (63.8%)	2,433 (75.4%)	53,902 (81.3%)
American Indian/Alaska native	0	17 (0.5%)	370 (0.6%)
Asian	0	21 (0.7%)	444 (0.7%)
Black/African-American	17 (36.2%)	735 (22.8%)	10,785 (16.3%)
Native Hawaiian or other Pacific Islander	0	21 (0.7%)	824 (1.2%)
Cardiovascular risk factors			
Smoking, current or past	38 (90.5%)	1,329 (41.4%)	66,227 (82.3%)
Body-mass index (kg/m^2)	28.8±6.4	30±5.5	28.8±5.4
Hypertension (%)	49 (81.7%)	2,667 (77.9%)	61,084 (75.9%)
SBP (mmHg)	143±27.2	138.9±23	139.8±22.8
DBP (mmHg)	79.7±15.1	79.6±13.6	76.1±13
Hyperlipidemia (%)	31 (51.7%)	2,391 (69.9%)	48,837 (60.7%)
Total cholesterol (mg/dL)	168.8±44.2	170.7±42.8	173.9±43.4
HDL cholesterol (mg/dL)	43±14.8	41.7±12.2	42.4±12.9
LDL cholesterol (mg/dL)	98.8±33.9	99.4±37.4	101.4±36
Triglycerides (mg/dL)	146.2±131.5	160.1±129.2	157.9±124.2
Diabetes mellitus (%)	25 (41.7%)	1,233 (36%)	27,628 (34.3%)
HbA$_{1c}$ (mmol/mol)	6.8±1.7	6.7±1.6	6.8±1.5
eGFR (mL/min/1.73 m^2)	64.9±27	74.4±20.5	67.8±20.7
Medication			
*Aspirin	42 (70%)	1,289 (37.7%)	26,035 (32.3%)
Clopidogrel	28 (46.7%)	843 (24.6%)	20,230 (25.1%)
tPA: alteplase or reteplase	2 (3.3%)	6 (0.2%)	75 (0.1%)
Warfarin	12 (20%)	558 (16.3%)	12,186 (15.1%)
Statins	42 (70%)	2,458 (71.8%)	50,712 (63%)
β-Blockers	33 (55%)	1,546 (45.2%)	35,017 (43.5%)
ACE inhibitors/ARBs	34 (56.7%)	1,600 (46.7%)	37,514 (46.6%)
Comorbid conditions			
Atrial fibrillation	8 (13.3%)	406 (11.9%)	10,336 (12.8%)
COPD	15 (25%)	420 (12.3%)	12,599 (15.6%)
Coronary heart disease	30 (50%)	1,101 (32.2%)	29,604 (36.8%)
Peripheral vascular disease	10 (16.7%)	263 (7.7%)	7,146 (8.9%)
Congestive heart failure	15 (25%)	224 (6.5%)	6,620 (8.2%)
Chronic kidney disease	7 (11.7%)	299 (8.7%)	6,164 (7.7%)
Chronic liver disease	0	51 (1.5%)	678 (0.8%)
Deep-vein thrombosis	0	19 (0.6%)	209 (0.3%)
Pulmonary embolism	1 (1.7%)	31 (0.9%)	536 (0.7%)

Notes: Ages computed at a baseline year of 2002. Descriptive statistics for continuous variables computed using first lab values within a year following the first stroke event. Dichotomous variables, such as medications, are positive if any records found within a year of first stroke event. *Aspirin is taken by many patients as an over the counter and/or non-VA medication instead of a prescription, and thus the reported percentage in this table is an underestimation of aspirin use.
Abbreviations: ACE, angiotensin converting enzyme; ARBs, angiotensin-receptor blockers; CVD, cardiovascular disease; DBP, diastolic blood pressure; eGFR, estimated glomerular filtration rate; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SBP, systolic blood pressure; tPA, tissue plasminogen activator.

Misclassification of cases and controls would reduce the effect and require an exponentially larger sample size to maintain adequate power.

In the absence of a power analysis, we suggest the algorithm with the fewest false positives and false negatives overall for choosing between a more sensitive algorithm and a more specific algorithm. The more sensitive algorithm will have (sensitivity gain) $\times p \times$ n fewer false negatives and (sensitivity decrease) $\times (1-p) \times$ n more false positives than the more specific algorithm, where p is the prevalence and n the population size. In other words, use the more sensitive algorithm when:

$$\frac{gain\ in\ sensitivity}{decrease\ in\ specificity} > \frac{1-p}{p}$$

If every patient needs a clear classification, such as when AIS is used to define another phenotype or is used in a longitudinal study, we recommend identifying stroke with simpler rule-based algorithms (such as Tirschwell and Longstreth)[7] or our longitudinal cohort algorithm. Table 3 shows that the sensitivity:specificity ratio was 3.54 for the Tirschwell and Longstreth algorithm compared to the longitudinal cohort algorithm, so the rule-based Tirschwell and Longstreth algorithm[7] would be preferred over our longitudinal cohort algorithm when stroke prevalence exceeds 22%.

A recent study that validated ICD9 codes for AIS in a national health-insurance claims database in Taiwan found a PPV of 88%, sensitivity of 97%, and false-positive rate of 11.6%. The Taiwanese cohort had a PPV comparable to that of our study (88% vs 88%),[16] but a much higher false-positive rate (11.6% vs 5%).[16] In contrast to the previous algorithm described by Tirschwell and Longstreth,[7] which included an administrative database of discharge diagnoses for all patients hospitalized in Washington State (excluding the VA hospitals), our case–control algorithm had greater sensitivity (91% vs 89%) and much higher specificity (95% vs 83%). There are several plausible reasons that we may observe such differences. First, our case–control algorithm improves on specificity by excluding subjects with uncertain stroke diagnoses (those with predicted probabilities of 0.15–0.75). With the inclusion of only definite stroke cases, our specificity is increased. However, it yields a smaller number of cases compared to an algorithm that includes both probable and definite cases. For instance, some uncertain cases or cases of TIA may be coded as stroke and picked up in previous algorithms, but would not be classified as definite AIS in our algorithm. The population studied by Tirschwell and Longstreth[7] differed from our study in that it included patients in one state in hospitals other than the VA, whereas our study includes veterans in a nationwide cohort. The mean patient age (68 years in the validation subset) was slightly higher in the Tirschwell and Longstreth cohort than in ours (65 years). The percentage of women in this cohort was 44% compared to 2% in the VA. However, the overall underlying pathophysiology would be the same for the two populations. Moreover, the Tirschwell and Longstreth study was conducted in 2002, whereas our study is more recent, and thus there are differences in clinical diagnosis, imaging techniques, and management of stroke over time that affect these findings. The American Heart Association/American Stroke Association guidelines for the prevention of stroke in patients with stroke and TIA have been updated over the years to reflect these changes, with the latest in 2014.[17]

Our study has limitations. Although we have provided (bootstrapped) 95% CIs (Table 2), our small sample, lack of independent test set, chart-sampling criteria based on administrative codes, and exclusion of uncertain cases may lead to optimistic sensitivity and specificity. Some patients had AIS events at non-VA hospitals, but returned to the VHA system afterwards. Physicians noted these events and stroke-related complications, but the VHA records often lack the primary stroke data (imaging and clinical exam) that were collected at other institutions. We excluded these "possible" strokes when assessing classification performance, and the case–control algorithm excluded them using probability thresholds (Figure 3). While the algorithm does not require primary stroke data and might misclassify these cases in the population, Figure 3 shows that there is a positive trend in algorithm-predicted probability across the chart reviewed "no", "possible", and "definite" stroke groups. The medical charts chosen for review were based on ICD9-CM codes with a high probability for stroke. Those with less definite ICD codes were likely missed, thus decreasing sensitivity. However, this is likely to occur across all categories. Also, the time frame of our study was 1992–2014, and thus may not generalize to other periods.

Our study also has several strengths. First, we present a highly specific algorithm to analyze cases of AIS from a large national cohort, which can easily be applied to other databases as well. Careful implementation of the algorithm, such as removing uncertain stroke cases, resolves several study limitations of administrative data. Each stroke case also has an associated probability of being a true case, which can be used as its own phenotype in β-regression.[18] Consequently, researchers can utilize the large study population at little cost to quality and with minimal time and resources spent reviewing medical records. Also, in contrast to cohort studies commonly used in clinical research that may contain self-reported information or biomarker data, large clinical databases, such as the VHA database, have more detailed anthropometric, demographic, clinical, and biomarker data, often with large samples. Additionally, our gold-standard chart review incorporated brain-imaging data, which were lacking in many prior studies that developed stroke algorithms.[4] Such efforts as this to produce an accurate algorithm for AIS may be further utilized when linked with GWASs, thus paving the way for precision medicine in the future. Future developments will analyze unstructured data elements, such as applying

natural-language processing to free-form medical notes or using deep (machine) learning to differentiate between brain images of patients with and without stroke. Specifically, our group is working on developing a novel chart-review tool using natural-language processing for validating EHR-based phenotyping algorithms.

Acknowledgments

Disclosure

The authors report no conflicts of interest in this work.

References

1. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2095–2128.
2. Béjot Y, Mehta Z, Giroud M, Rothwell PM. Impact of completeness of ascertainment of minor stroke on stroke incidence: implications for ideal study methods. *Stroke*. 2013;44(7):1796–1802.
3. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *BMJ*. 2010;341:c4226–c4443.
4. Andrade S, Harrold L, Tjia J, et al. A systematic review of validated methods for identifying cerebrovascular accident or transient ischemic attack using administrative data. *Pharmacoepidemiol Drug Saf*. 2012;21(2 Pt 1):100–128.
5. Woodfield R, Grant I, Sudlow CL. 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.
6. Jones SA, Gottesman RF, Shahar E, Wruck L, Rosamond WD. Validity of hospital discharge diagnosis codes for stroke: the Atherosclerosis Risk in Communities Study. *Stroke*. 2014;45(11):3219–3225.
7. Tirschwell DL, Longstreth WT. Validating administrative data in stroke research. *Stroke*. 2002;33(10):2465–2470.
8. Department of Veterans Affairs. Department of Veterans Affairs: statistics at a glance. 2016. Available from: https://www.va.gov/vetdata/docs/Quickfacts/Homepage_slideshow_06_04_16.pdf. Accessed May 12, 2017.
9. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016;70:214–223.
10. Fleiss JL, Cuzick J. The reliability of dichotomous judgments: unequal numbers of judges per subject. *Appl Psychol Meas*. 1979;3(4):537–542.
11. Zou H. The adaptive LASSO and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418–1429.
12. Huang J, Ma S, Zhang CH. Adaptive LASSO for sparse high-dimensional regression models. *Stat Sin*. 2008;18(4):1603–1618.
13. R Core Team. R: a language and environment for statistical computing. 2016. Available from: https://www.R-project.org. Accessed August 2, 2018.
14. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med*. 2009;150(9):604–612.
15. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*. New York: Association for Computing Machinery; 2005:625–632.
16. Hsieh CY, Chen CH, Li CY, Lai ML. Validating the diagnosis of acute ischemic stroke in a National Health Insurance claims database. *J Formos Med Assoc*. 2015;114(3):254–259.
17. Kernan WN, Ovbiagele B, Black HR, et al. Guidelines for the prevention of stroke in patients with stroke and transient ischemic attack: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2014;45(7):2160–2236.
18. Ferrari SL, Cribari-Neto F. Beta regression for modelling rates and proportions. *J Appl Stat*. 2004;31(7):799–815.
19. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37–46.
20. Sacco RL, Kasner SE, Broderick JP, et al. An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2013;44(7):2064–2089.

Supplementary materials

Notes

TIU notes					
ScrSSN	**Attending Physician Staff**	**Document Definition**	**Note Date**	**Report Text**	**Useful**
999999999	Physician1	NULL	1/1/1900		☐
999999999		123	12/26/2015	213948723498	☐

Record: 1 of 2 — No Filter — Search

Radiology

Radiology								
ScrSSN	**Exam Date**	**Interpreting Resident**	**Interpreting Staff**	**Procedure**	**Imaging Location**	**ReportText**	**ImpressionText**	**Useful**
999999999								☑

Record: 1 of 1 — No Filter — Search

Classification of stroke

Classification	Classification type	Date In Chart	Date Of Event	Source of decision	Supportive text from notes	Reviewer Comments
Stroke event	Acute Ischemic stroke (AIS)	9/1/2015	9/1/2015	Multiple	brain infarct	
Stroke event	Stroke - unspecified	8/26/2015	1/3/2010	Discharge Note	hx stroke	stroke type unclear
Family Hx of stroke	Stroke - unspecified	9/1/2015		Discharge Note	family hx stroke	

Classification

Stroke event

- Stroke event
- Family Hx of stroke
- Relevant meds for stroke
- No event
- Unable to decide

Classification type

Acute Ischemic stroke (AIS)

- Acute Ischemic stroke (AIS)
- Hemorrhagic stroke
- Stroke - unspecified
- Transient ischemic attack
- None

Figure S1 Database platform for the chart review process.
Note: In addition to what is presented in the above figure, the platform also included information on patient demographics, ICD-9 codes, and data from Centers for Medicare and Medicaid Services and the National Death Index.

Table S1 Process of case selection for review (n=300 charts)

Case selection process*

Selection criteria: Random selection of medical charts from the CVD Merit cohort with ICD-9 codes for acute ischemic stroke (AIS) (433.x, 434.x, 436)

Criteria	Count of charts meeting criteria in the national Veterans Affairs cohort
10 charts with only one inpatient ICD-9	2,710
10 charts with only one outpatient ICD-9	10,809
10 charts with one inpatient ICD-9 AND multiple outpatient ICD-9s	37,199
10 charts with no inpatient ICD-9 AND multiple outpatient ICD-9s	62,773
10 charts with multiple inpatient ICD-9s AND multiple outpatient ICD-9s	25,648

Note: *This process was replicated for each chart reviewer.

Table S2 Cumulative incidence of acute ischemic stroke (derived from rules based and statistical algorithm) from 2000 to 2015 in the national Veterans Cardiovascular Disease risk cohort

Year	Rules-based algorithm (Tirschwell)			Longitudinal cohort statistical algorithm		
	Count	Crude incidence	Incidence per 10,000 persons	Count	Crude incidence	Incidence per 10,000 persons
2000	13515	0.006407864	64.07864486	12359	0.005858918	58.58917605
2001	14792	0.007058559	70.5855855	13620	0.006494761	64.94760559
2002	16717	0.008033852	80.33852039	15599	0.007487083	74.87082754
2003	17637	0.008544632	85.44631736	16659	0.00805617	80.56170256
2004	17408	0.008506371	85.06371472	16594	0.00808991	80.89910428
2005	15232	0.007506932	75.06931788	14423	0.007088852	70.88852223
2006	14039	0.006971307	69.71307352	13318	0.006592482	65.92481858
2007	12189	0.006095149	60.95149133	11536	0.005748278	57.48277659
2008	8710	0.004382174	43.8217386	8164	0.004091562	40.9156198
2009	7817	0.003950198	39.50198293	7372	0.003709813	37.09813292
2010	7262	0.003684291	36.84291433	6839	0.003454407	34.54406781
2011	6876	0.003501359	35.01358839	6509	0.003299119	32.99118934
2012	6798	0.003473803	34.7380314	6367	0.003237828	32.37827508
2013	6295	0.003227982	32.27981653	5755	0.002936112	29.36112139
2014	5554	0.002857231	28.57231048	5183	0.002652073	26.52073355
2015	4571	0.002358269	23.5826911	4312	0.002212261	22.12261119

Table S3 Cumulative incidence of acute ischemic stroke (derived from statistical algorithm) from 2000 to 2015 in the Million Veteran Program

Year	Rules-based algorithm (Tirschwell)			Longitudinal cohort statistical algorithm		
	Count	Crude incidence	Incidence per 10,000 persons	Count	Crude incidence	Incidence per 10,000 persons
2000	274	0.000848	8.482394	234	0.000724	7.243977
2001	294	0.000911	9.109274	253	0.000784	7.83784
2002	315	0.000977	9.768835	275	0.000853	8.526074
2003	346	0.001074	10.74071	323	0.001002	10.02281
2004	363	0.001128	11.28054	325	0.001009	10.09499
2005	399	0.001241	12.41328	380	0.001182	11.8153
2006	399	0.001243	12.42871	364	0.001133	11.3312
2007	443	0.001382	13.81646	407	0.001268	12.68415
2008	487	0.001521	15.20977	434	0.001354	13.54278
2009	603	0.001886	18.86131	536	0.001675	16.74833
2010	729	0.002285	22.84557	666	0.002085	20.84533
2011	794	0.002494	24.93954	616	0.001932	19.32064
2012	930	0.002928	29.28433	641	0.002014	20.14368
2013	993	0.003136	31.35994	672	0.002116	21.16049
2014	923	0.002924	29.24097	692	0.002184	21.83647
2015	853	0.00271	27.1026	607	0.00192	19.19616

Table S4 Sensitivity and PPV of code-groups in chart-reviewed VA sample (N=300)

Code Group	ICD-9	Total	AIS	Possible AIS	ICH/SAH	TIA	No Stroke	Sensitivity	PPV
Tirschwell	433.x1, 434, 436	144	85	25	5	2	27	0.934	0.599
AHA/ASA	433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.01, 434.11, 434.91	72	50	6	2	1	13	0.549	0.704
ICH	431	8	4	0	4	0	0	0.800	0.571
SAH	430	3	2	0	1	0	0	0.200	0.500

Abbreviations: PPV, positive predictive value; VA, Veterans Affairs; AIS, acute ischemic stroke; ICH, intracranial hemorrhage; SAH, subarachnoid hemorrhage; TIA, transient ischemic attack; AHA/ASA, American Heart Association/American Stroke Association