Electronic Health Record-Derived Phenotyping Models to Improve Genomic Research in Stroke

Phyllis Mary Thangaraj

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Electronic Health Record-Derived Phenotyping Models to Improve Genomic Research in Stroke

Phyllis Mary Thangaraj

Stroke is a highly heterogeneous and complex disease that is a leading cause of death in the United States. The landscape of risk factors for stroke is vast, and its large genetic burden has yet to be fully discovered. We hypothesize that the small number of stroke variants recovered so far is due to 1) the vast phenotypic heterogeneity of stroke and 2) binary labeling of stroke genome-wide association study (GWAS) participants as cases or controls. Specifically, genome-wide association studies accumulate hundreds of thousands to millions of participants to acquire adequate signal for variant discovery. This requires time-consuming manual curation of cases and controls often involving large-scale collaborations. Genetic biobanks connected to electronic health records (EHR) can facilitate these studies by using data routinely captured during clinical care like billing diagnosis codes. These data, however, do not define adjudicated cases and controls, with many patients falling somewhere in between. There is an opportunity to use machine learning to add nuance to these definitions. We hypothesize that an expanded definition of disease by incorporating correlated diseases and risk factors from EHR data will improve GWAS power. We also hypothesize that granularly subtyping stroke using unsupervised learning methods can provide insight into stroke etiology and heterogeneity. In Chapter 1, we described the motivation for building upon current phenotyping methods for subtyping and genome-wide association studies to improve GWAS power. In Chapter 2, using patients from Columbia-New York Presbyterian (NYP) Hospital, we built and evaluated machine learning

models to identify patients with acute ischemic stroke based on 75 different case-control and classifier combinations. In chapter 3, we compared two data-driven and unsupervised methods, non-negative matrix factorization (NMF) and hierarchical poisson factorization, to subtype stroke patients and determined whether any of the subtypes correlate to stroke severity. In chapter 4, we estimated the heritability of acute ischemic stroke by treating the patient probabilities assigned by the machine learning phenotyping models for acute ischemic stroke in chapter 2 as a quantitative trait and mapping the probabilities to Columbia-NYP EHR-generated pedigrees. We also applied our machine learning phenotyping algorithm method, which we call QTPhenProxy, to venous thromboembolism on Columbia eMERGE Consortium patients and ran a genome-wide association study using the model probabilities as a quantitative trait. Finally, we applied QTPhenProxy to subjects in the UK Biobank for stroke and 14 other diseases and ran genome-wide association studies for each disease. We found that our machine learned models performed well in identifying acute ischemic stroke patients in the Columbia-NYP EHR and in the UK Biobank. We also found some NMF-derived subtypes that were significantly correlated with stroke severity. We were underpowered in the eMERGE venous thromboembolism cohort GWAS and did not recover any known or new variants. Finally, we found that QTPhenProxy improved the power of GWAS of stroke and several subtypes in the UK Biobank, recovered known variants, and discovered a new variant that replicates in a previous stroke GWAS. Our results for QTPhenProxy demonstrate the promise of incorporating large but messy sets of data, such as the electronic health record, to improve signal in genome-wide association studies.

Table of Contents

List of Tables

List of Figures

Acknowledgements

It takes a village to raise a PhD student, and I am thankful to have so many people in my life who have shared with me their support, expertise, and enthusiasm to make my degree possible.

I first would like to thank my PhD adviser, Nicholas Tatonetti, for being a wonderful and supportive mentor to me. From day one, your enthusiasm for uncovering the latest mysteries in data science has fueled my excitement for the field and sharing my work with others. I have learned so much during my PhD, and a large share of it is from your endless encouragement, tireless commitment to teaching, and positive environment you have cultivated in the lab. Thank you also to the current and former members of the Tatonetti Lab, who have been true friends and collaborators: Nicholas Giangreco, Anna Basile, Theresa Koleck, Katie Brown, Rami Vanguri, Kayla Quinnies, Joseph Romano, Alexandre Yahi, Ola Jacunski, Mary Boland, Tal Lorberbaum, Yun Hao, Fernanda Polubriaginof, Jenna Kefeli, Vijendra Ramlall, Kai Chen, and Deidre Gregory.

I next would like to thank my thesis committee, George Hripcsak, Mitchell Elkind, Adler Perotte, and Tuuli Lappalainen. Thank you for your insightful advice and comments; I have always enjoyed our discussions during my thesis committee meetings and am so grateful for the time you have taken to guide me in my education and for your mentorship.

Thank you to the Medical Scientist Training Program at Columbia for being my home for the last seven years. It has truly been a privilege to be a part of this program, and I thank the directors Steve Reiner, Ron Liem, Patrice Spitalnik, and Michael Shelanski for making the program a warm and talented community of physician-scientists. Thank you to Jeffrey Brandt, Rebecca Spurr, and Kaitlyn Matthews for your tireless administrative efforts to keep me afloat in

the program.

Thank you to the Integrated Cellular, Molecular, and Biomedical Sciences Program and Systems Biology Department for your support, and thank you to Zaia Sivo for keeping me on track throughout my PhD years. Thank you to the Department of Biomedical Informatics for hosting me throughout my PhD. I have greatly enjoyed my time here, and thank you to Rosemary Vasquez for your help with scheduling.

Thank you to the National Heart, Lung, and Blood Institute for funding my F30 fellowship.

Thank you to my mentors at Yale: Meg Urry, for paving the way for women in physics in our department, Thierry Emonet, for introducing me to the interdisciplinary science of mathematical biology, and Simon Mochrie, for guiding me in my first long term research project.

Thank you to my mentors from the Lab of Biological Modeling at the National Institutes of Health, Carson Chow and Shashaank Vattikuti, for advising me in a fascinating computational neuroscience research project and introducing me to the field of genetics. Thank you also to the NIH Post-baccalaureate Intramural Research Training Award for your support and excellent advising for applying to graduate school.

I would not be where I am without the support of my dear friends from college and beyond: Christina Lin, Lianna Valdes, Shelagh Mahbubani, Corinna Li, Zuzana Culakova, Dan Lao, Young Lim, Rahul Dalal, and Prashanth Selvaraj.

Thank you also to my friends at Columbia, who have made the ups and downs of the program a lot of fun to brave with: Chioma Madubata, Elizabeth Balough, Lia Boyle, Lisa Grossman, Janet Woollen, Dave Thomas, Scott Kanner, Ben Schrank, Andrew Hollar, my MD/PhD entering class of 2013, the Bard Hall Players, the wonderful students, faculty, and staff of the Department of Biomedical Informatics, the CMS AI challenge team, and the Phenotyping Working Group.

Especially thank you to my best friend since high school, Akriti Gupta, who has supported me through thick and thin.

Thank you to Marc Tuozzolo, for your humor and constant positive encouragement.

Thank you to Beverly Wrede for your mentorship and support throughout my childhood.

Thank you to my cousins, aunts, uncles, and extended family for many fond memories of get-togethers and supporting me throughout all of my endeavors, including coming to my shows.

And finally, thank you to my wonderful parents, who from day one have been the best parents I could ask for.

Dedication

Dedicated to the memory of my grandparents, Mariam and K.C. Philip and Mary and M.A. Thangaraj, all trailblazers in their fields.

Dedicated to the honor of my parents, Elizabeth and Arun Thangaraj, who have been an unending source of enthusiasm, support, and love throughout my life.

Chapter 1: Introduction

Stroke is a leading cause of death and the top cause of disability in the US [1]. Almost 800,000 people in the United States have a stroke every year, and over one sixth of these result in death[2, 3]. In addition, the rate of decline of stroke death has decreased in recent years and even increased in certain communities, such as younger patients, the Hispanic community, and the south[3, 4]. A stroke is characterized by an acute focal loss of neurological function and is primarily caused by loss of blood flow to a specific area of the brain. In 80% of strokes, loss of blood flow is due to a blockage, which is known as ischemic stroke, while in 20% of strokes, blood loss is due to a leaking or burst vessel, which is known as a hemorrhagic stroke [5]. There are many identifiable risk factors for stroke, including various metabolic, cardiovascular, and coagulative diseases, medications, lifestyle, and demographics. In addition, "triggers" such as pollution, infection, and inflammatory disorders can precipitate the acute event [5]. Accurate determination of the etiology of disease is essential for risk stratification and optimal treatment, but this can be difficult as up to 35% of strokes are of undetermined cause [6, 7]. Most of the unidentified risk, up to 40%, is thought to be genetic [8]. Initial stroke GWAS found few variants of genome-wide significance with thousands of participants[9, 8, 10]. The largest stroke genome-wide association study (GWAS) consisted of a cohort of 520,000 participants, including 67,162 cases [11]. Even with this large number of cases, the study only found thirty-two genome-wide significant variants for stroke, many fewer than the hundreds to thousands of variants found for related stroke risk factors [12]. Since stroke requires both genetic risk factors and environmental stimuli to manifest[5], patients with genetic predisposition for the disease but without a stroke event would be considered controls. This type of stringent classification may lead to loss of power [13]. A recent study improved predictability of acute ischemic stroke by using a polygenic risk score that incorporated known risk factor variants in addition to known acute ischemic stroke variants [12]. This highlights the potential for incor-

porating medical data into case assignment to improve genetic signal and reduce information loss from treating potential genetically susceptible subjects with known risk factors as controls.

Genome-wide association studies were first introduced as an unbiased answer to family linkage and specific gene studies [14]. Initially, family linkage studies identified heritable diseases that aggregated within families, but they discovered variants at a scale too small and biased towards specific genes found within the studied families [13, 15]. From the earliest GWAS, the goal of finding variants that contribute to disease required identification of as many people with the disease as possible [14, 16, 17]. These variants, however, only accounted for a portion of heritability[14, 15, 17]. Later studies argue that much of this was due to overly conservative significance thresholds or causal variants not inherited with the marker single nucleotide polymorphism (SNP)[13, 18, 19]. These causal variants may have been expressed at low frequency, which supports the need to maximize the sample size and to develop high throughput methods for identifying these cohorts[13, 17, 16].

Therefore, an essential task for genetic studies is to define who has the disease[13]. In a genetic context, it requires choosing the best representation of the genotype, or genetic makeup of the subject, through the phenotype, or physical manifestation of the genotype[20, 21]. The phenotype can be described through appearance, characterization, behavior, and acute events stemming from gene and environment interactions[22]. The field of phenomics and the human phenome project developed from the recognition that missing heritability could be partially attributed to heterogeneity in the phenotype[16, 23]. With cohorts for genetic studies ballooning in size, a high-throughput and accurate way to phenotype cohorts is essential[23]. High-throughput identification of cases and controls can be difficult, however, due to time-consuming chart review and incompleteness of medical records. Current GWAS require a large number of well-adjudicated cases for statistical power[14, 17]. Convention argues that a cleaner phenotype at appropriate granularity for the study disease results in a stronger genetic signal[17, 13]. This overlooks patients with a genetic predisposition for the disease but may not have yet presented with the phenotype. In addition, the stroke phenotype from individual to individual is heterogeneous, and its description is dependent on the

data available. For example, stroke manifested in two different patients as hemiplegia and CT imaging positive for a cerebral artery blockage may be caused by entirely different etiologies of the disease, such as atrial fibrillation and stasis in the heart versus atherosclerosis. In most research settings, it can be defined simply as a single International Classification of Diseases (ICD) code for stroke[24]. A more holistic description of the phenotype can be defined within the electronic health record, which has a rich assortment of information to describe a patient's medical timeline.

The electronic health record has a long and successful history of use in genomic research even though its primary purpose is for clinical coordination and billing[25, 26]. The UK Biobank is a prospective study of 500,000 adults, aged 40-69, containing comprehensive questionnaires, physical measurements, imaging, EHR data, genotyping, and exome data[27, 28]. This resource is openly used worldwide to study the genetics of thousands of diseases, social determinants, and biological markers[29]. Some limitations of the UK Biobank include lack of visit dates and lack of diversity within the biobank, where 95% of participants are of European ancestry. This prevents the incorporation of the medical timeline into patient phenotyping, which is possible in biobanks directly linking EHR data. In addition, the lack of diversity can lead to missed common and rare variants found in other ancestral backgrounds[15, 30, 31]. More recent biobanks, such as the All of Us Research Program in the United States, are aiming to fill the gap in diversity[30]. The electronic Medical Records and Genomics Network (eMERGE) consortium is a large collaboration across nine major academic medical centers combining biobanks of patients with their electronic health record data. They also develop and maintain pheKB, a library of fifty-five rule-based phenotyping algorithms which have been validated across other sites with a positive predictive value of at least 90%. Their large cross-institution data set and stringent phenotyping have resulted in hundreds of publications including many successful genome-wide association studies[32, 33, 34, 35]. Some of the difficulty in replication of phenotypes across sites, however, are due to implementation of the algorithms. Many require natural language processing and inclusion/exclusion criteria that may be recorded differently across sites. PheKB algorithms translated into the Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership Common

Data Model (OMOP CDM) transferred better across institutions over the original phenotype algorithms[36, 37]. The OHDSI OMOP CDM is a world-wide community-driven standardized model for organizing electronic health record data and mapping the data to vocabularies and terminologies for research[38]. Major healthcare centers around the world have formatted their EHR data to the OMOP CDM, leading to observational research studies with hundreds of millions of patients and over a billion data points[39, 40, 41]. The OMOP CDM provides the base necessary to develop high-throughput phenotyping within the EHR.

There are biases unique to phenotyping with EHR data since its primary uses are for patient care and billing. Often, the data are incomplete and missing not at random. Diagnoses codes are issued and tests ordered when doctors need them, and data completeness depends on patient interactions with healthcare, such as having insurance or choosing not to seek care [42, 26, 25, 43, 44]. This can lead to a bias in algorithms assessing patient disease risk towards patients who frequently seek health care[45]. In addition, since only a portion of people with a given disease seek help, generalizing outcomes from the patients can be problematic[26]. Finally, diagnosis codes can often be chosen for reimbursement purposes rather than actual diagnosis, and diagnosis code use can change over time, leading to inaccuracies in phenotyping[43, 25].

Despite these caveats, successful phenotyping has combined specific data domains and vocabularies within the OMOP CDM[36, 37]. The data contains seven domains, including conditions, procedures, measurements (lab values), drugs (medications), devices, observations (such as social history), and visits (type of visit, such as inpatient)[46]. Not all domains are populated with every CDM implementation, but the most common are conditions, procedures, measurements, drugs, and visits. Within conditions, which represent patient diagnoses, the International Classification of Disease codes, version 9 and 10 (ICD9 and ICD10) and Systematized Nomenclature of Medicine (SNOMED) codes are most commonly used[46]. ICD9 and ICD10 codes tend to be more sensitive over specific, while Current Procedural Terminology (CPT) codes, used for procedures, are more specific but have low recall across sites[47]. In addition, for diseases with easily detectable symptoms, such as stroke, the less likely a false positive diagnosis label will occur[42]. For diseases

that are diagnosed or confirmed by lab values, the measurement domain is shown to have high precision, and medications are also a good confirmatory data type[47].

Not only can phenotypes be richly described by the information in the EHR, but high throughput screening for diseases worth pursuing through GWAS can be performed. Highly heritable diseases, for example, have a strong genetic susceptibility[48]. Several papers have developed high-throughput methods for estimating disease heritability without genetic data. Studies have inferred genetic overlap between disease co-occurrences in millions of patients[49], estimated high-throughput heritability through building familial relationships through insurance claims[50], and developed an algorithm that creates EHR-derived pedigrees by inferring familial relationships through emergency contact information[51]. These methods estimate the heritability of hundreds of diseases without genetic information and rely on single insurance billing diagnosis codes to phenotype the patients. Another study further teased apart genetic versus environmental contribution to phenotypic variance using insurance claims and phenotyped cases using phecodes, which are a manually agglomerated group of ICD9 and ICD10 codes to represent a phenotype[52, 53].

The heterogeneous nature of diseases such as stroke is another source of loss of power in GWAS, suggesting a need to conduct individual genetic studies on disease subtypes[17, 54]. Traditionally, subtyping a patient requires a time consuming and laborious process of integrating multiple facets of data, including medical notes, labs, and imaging reports with medical experience by physicians[55]. There is an opportunity for an unsupervised data-driven subtyping method to complement the physician-defined approach to phenotyping. The most common unsupervised subtyping categories are clustering, dimensionality reduction, topological data analysis, and deep learning[54]. Factorization and dimensionality reduction methods are common data-driven approaches to identify subtypes[56, 57, 58, 59, 60]. They can model interactions between items while also de-emphasizing missing data. Non-negative matrix factorization is a dimensionality reduction method to learn representation of parts. The method imposes non-negative constraints on its factorization, so all components are additive[61]. Mixed membership models such as Latent Dirichlet Allocation and semi-supervised mixed membership models have identified different disease phe-

notypes from heterogeneous data in the electronic health record[62, 63]. In addition, Bayesian factorization methods such as hierarchical Poisson factorization have been used in the past to subtype customers for recommender systems because of their ability to capture interactions of a user with a set of items while also taking into account sparsity of data, large variation in interaction frequency, and deemphasizing the weight of zeros[64]. Levitin et. al. applied hierarchical Poisson factorization to determine novel transcriptional patterns of genes in single cell RNA-seq data to separate different cell types within tissues and predict patient survival from glioblastoma[65].

In addition, deep learning has been applied to the subtyping field. Deep learning uses neural networks to process large amounts of features and find novel patterns within the data[66]. Denoising autoencoders (DAs) are a type of neural network that reconstructs an input through a hidden layer, compressing thousands of features, or inputs, to a more compact representation[67, 68]. DAs have been used to predict disease susceptibility to many different diseases[69], to predict survival in ALS patients[68], and to successfully separate subtypes of simulated patients[67]. Reducing phenotypic heterogeneity can improve genetic signal within subtypes[58]. Li et. al[70] found three novel subtypes of Type 2 diabetes applying topological data analysis to clinical data enriched for subtype-specific genetic variants. This reinforces the usefulness of subtyping to increase genetic homogeneity of cohorts.

In this thesis, we explore three aims to improve phenotyping for genetic studies, particularly when applied to stroke. In Aim 1, we develop a phenotyping algorithm to differentiate between acute ischemic stroke patients and other similar diseases (Chapter 2). In Aim 2, we compare two unsupervised learning algorithms to subtype acute ischemic stroke with more granularity and with less manual labor than current methods (Chapter 3). Finally, in Aim 3, we develop QTPhenProxy, a quantitative trait proxy that uses the phenotyping algorithms' probability output from Aim 1 to improve the power of estimating heritability and genome-wide significant variants within Columbia's electronic health record and eMERGE data set (Chapter 4) and the UK Biobank (Chapter 5).

Chapter 2: Comparative analysis, applications, and interpretation of EHR-based stroke phenotyping methods

2.1 Introduction

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[55]. This process is often manually performed and time-consuming, and can reveal misclassification errors[71, 55].

One simple approach to identify acute ischemic stroke (AIS) is the diagnosis-code based algorithm created by Tirschwell and Longstreth[24]. Identifying every AIS patient using these criteria, however, can be difficult due to the inaccuracy and incompleteness of diagnosis recording through insurance billing[24, 72, 44]. Past studies have shown that the positive predictive value (PPV), sensitivity, and specificity of identifying stroke using ICD9 codes varies widely depending on cohort and data available [24, 73, 74, 75, 76, 77, 78]. 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[79]. Phenotyping features from the EHR have been traditionally culled and curated by experts to manually construct algorithms[80], but machine learning (ML) techniques present the

potential advantage of automating this process of feature selection and refinement[81, 82, 83, 84]. Recent machine learning approaches have also combined publicly available knowledge sources with EHR data to facilitate feature curation[85, 86]. 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 identify such cohorts using automated feature selection and imperfect case definitions in a high-throughput manner[87, 88, 89, 90]. Within these ML methods, the selection of cases and controls using diagnosis codes can significantly affect model performance[91]. Feature size can also influence the utility of the model. Structured medical data, in particular, have hierarchical organization that can be utilized for the grouping of similar features and have improved classification performance[92, 93]. Finally, interpretation of the ML phenotyping model relies on a calibration assessment[94]. Calibration provides a way to understand the clinical utility of the phenotyping model.

Two stroke phenotyping algorithms have used machine learning to enhance the classification performance of a diagnosis-code based AIS phenotyping algorithm[77, 78, 95, 96]. Ni et al. manually curated hundreds of features and confirmed that machine learning classifiers outperformed stroke ICD9 codes and research nurses' review on all metrics except recall[77]. Imran et al. developed a knowledge-driven phenotyping algorithm to define a gold standard of stroke cases from physician review and found that logistic regression trained with only ICD9 codes for acute ischemic stroke as features outperformed ICD9 AIS codes defined by Tirschwell and Longstreth in sensitivity and specificity[78]. 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 trained on cases defined by diagnostic codes with manually-curated cohorts. We also applied key methods to optimize the robustness of these models to missing data, the calibration to ensure a clinically meaningful model output, and the number of features to improve generalizability. We then applied the models to all 6.4 million patients in our EHR to estimate the prevalence of potential stroke patients that do not have ICD codes for AIS. We tested one of our best-performing models in an independent test set, the UK Biobank, to evaluate its ability to detect self-reported AIS patients without the requisite ICD codes. Our phenotyping method utilizes machine learning classifiers with minimal data processing to increase the number of stroke patients recovered within the EHR and reduces the time and effort needed to find them for research studies.

2.2 Objective

The aim of this chapter is to develop and evaluate a machine learning algorithm to phenotype ischemic stroke patients in the EHR for research purposes. We hypothesize that the addition of EHR data will identify features in addition to ICD9 and ICD10 AIS codes that can phenotype these patients in a data-agnostic manner.

2.3 Methods

2.3.1 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. Figure 1 shows the overall workflow of training and testing the models, the models' evaluation, and its testing in an independent test set.

2.3.2 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 include structured medical data such as conditions, procedures, medication orders, lab measurement values, visit type, demographics, and observations. 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]. This includes patients from the CUIMC stroke service (Figure 2.1, Table 2.1) that were part of a larger group of patients with acute cerebrovascular diseases, 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 (Phyllis Thangaraj (PT) and Benjamin Kummer (BK)) each manually reviewed 50 patients' charts for a total of 100 patients from this cohort to determine baseline false positive rates.

2.3.3 Patient Population

We defined three case groups. We first included all patients from the CUIMC stroke service with recorded with 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 codes 434.x1, 433.x1, 436 (where x is any number), and the code is in the primary diagnostic position[24]. Our dataset did not specify the diagnostic position of codes. We also included ICD10 code equivalents, I63.xxx or I67.89, with the ICD10 codes determined from ICD9 from Centers for Medicaid and Medicare Services (CMS) General Equivalence Mappings[97]. Because patients with cerebrovascular event diagnosis codes such as transient ischemic attack or migraine aura with cerebral infarction may have suffered AIS but may not have an attached AIS diagnosis code, we also created a group of cases with cerebrovascular disease-related ICD codes defined by the *ICD-9-Clinical Modification (CM)* Clinical Classifications Software tool (CCS). This includes

ICD9 codes 430-438 and 346.xx as well as their ICD10 equivalents, I60.xxx-I66.xxx, R29.xxx, and G43.xxx (cohort C)[98]. We then defined four control groups (Figure 2.1, Table 2.1). First, we defined a control group of patients without AIS-related diagnosis codes (I). Since cerebrovascular disease is a major risk factor for stroke[5] 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).

2.3.4 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 and ICD10 codes for phenotyping[99, 100], we mapped diagnoses and procedure features from ICD9, ICD10, and *Current Procedural Terminology 4* (CPT4) codes to SNOMED concept IDs and used *RxNorm* IDs for medication prescriptions. We identified patients with Hispanic ethnicity using an algorithm combining race and ethnicity codes[51]. 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. Importantly, 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, there is a large imbalance between cases and controls, which can lead to a machine

Variable	Identification	N Samples
Total Patients	CUIMC CDW Person ID	6,377,222
Diagnosis Codes	ICD9,ICD10,SNOMED	140,300,457
Procedure Codes	ICD9,ICD10,CPT,SNOMED	64,383,775
Prescription Orders	RxNorm	40,759,814
Training Categories		
Stroke Service Cases (S)	Seen by NYP Stroke Service	4,484
Tirschwell Criteria AIS Cases (T)	ICD9:434.x1,433.x1,ICD10:I63.xxx	79,306
CCS Cerebrovascular Cases (C)	ICD9:346.6x,430,431,432.x,433.xx	181,698
AIS Mimetic Diseases Controls (N)	ICD9:191.x,225.x,340,250.0,431	8,438
No Stroke Controls (I)	No (T) Codes	5,243,646
No Cerebrovascular Disease Controls (C)	No (C) Codes	5,149,975
Cerebrovascular disease, w/o AIS Controls (CI)	(T) codes, No (C) codes	102,435
Random set of patients Controls (R)	With >=1 ICD9 or ICD10 diagnosis code	5,396,172
UK Biobank		
Total Subjects	With diagnoses codes, procedure codes, medication prescriptions, or demographics	384,208
Tirschwell Criteria AIS Cases (T)	ICD10:I63.xxx,I64.x,(41202, 41204)	4,922
No Cerebrovascular Disease Controls (C)	No (C) Codes (41202, 41204)	312,500
Subjects with AIS but no diagnosis codes	Date of AIS (42008) - (T) cases	163

Table 2.1: **Select Structured Data and Sample Case/Controls for models available in CUIMC Common Data Warehouse.**

learning classifier that undersamples cases[91]. By randomly sampling controls in a 1:1 case to control ratio, we created a balanced dataset. In addition, we set the maximum sample size to 16,000 patients in order to control the size of the feature set.

2.3.5 Model development

Using all 15 case-control combinations, we trained 75 models using logistic regression classifiers with L1 and elastic net regularization as well as random forest, AdaBoost, gradient boosting, and neural network classifiers on the gathered features. We chose these classifiers to compare a variety of feature-to-outcome relationships: linear (logistic regression), ensemble (random forest,

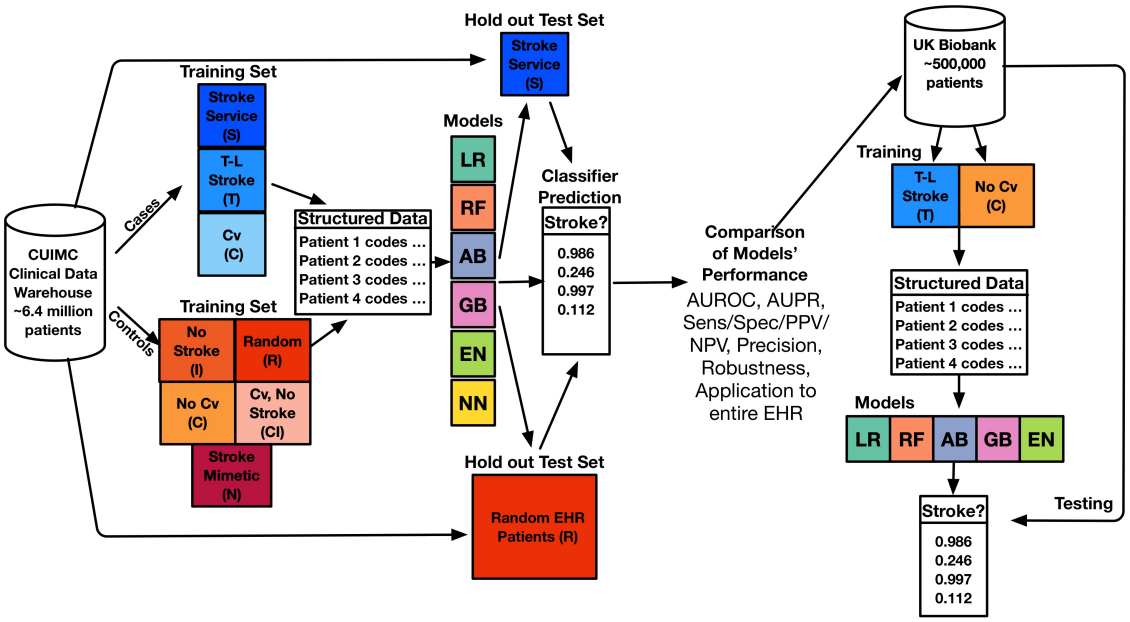


Figure 2.1: **Schematic of Model Training, Testing, Evaluation, and Application to UK Biobank.** Cases: 1) Physician curated, stroke service patients (S), 2) Patients with stroke insurance billing codes (I), 3) Patients with cerebrovascular disease billing codes (C) Controls: 1) Stroke mimetic neurological diseases (N), 2) Sample of patients in EHR without stroke code (I), 3) Sample of patients in EHR without cerebrovascular disease codes (C), 4) Sample of patients in EHR with cerebrovascular disease codes but without stroke codes or not gold standard (CI) 5) Random sample in EHR (R). Case: Control ratio was 1:1 and 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). Area Under the Receiver Operating Curve (AUROC), Area under the Precision Recall Curve (AUPRC), Sensitivity (Sens), Specificity (Spec), Positive Predictive Value (PPV), Negative Predictive Value (NPV).

AdaBoost, gradient boosting), and non-linear (neural network). We tuned the models' hyperparameters using 10-fold cross validation. We ran a grid search for the hyperparameters of maximum tree depth, number of estimators, L1-L2 ratio, penalty parameters, subsampling rate, learning rate, momentum, and dropout. Outside of the default parameters, we used a max tree depth of 100, 1000 estimators, and square root maximum feature number for the random forest models. For L1 Logistic regression, we used an inverse of regularization strength of 0.1, and for elastic net regularization, we used a penalty parameter of 0.01 and L1-L2 ratio parameter of .01, and "log" loss. For the boosting algorithms, we used a learning rate of 0.1, 1000 estimators, and for gradient boosting we also subsampled at a rate of 0.5, max depth as 10, and square root maximum feature number. The neural network model was comprised of two layers, the first with 64 neurons, relu activator,

and l1 kernel regularizer. The second layer contained two neurons and a softmax activator. Learning was compiled by stochastic gradient descent with a learning rate of 0.01 and momentum of 0.9, nesterov=True. We included a dropout of data in the first layer at a rate of 0.3. Loss was calculated by categorical cross-entropy. We then determined a probability threshold for each model from the training set. Within the validation set of each training fold, controls were bootstrapped to form a 100:1 control to case ratio to represent the prevalence of AIS in the population[1]. Precision and recall were then calculated from the bootstrapped set. To determine the optimal threshold to maximize precision while maintaining a high recall, we calculated maximum F scores at different βs using the following equation:

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall}$$

where β = 1.0, 0.5, 0.25, and 0.125. Using the probability threshold determined from cross-fold validation, we then calculated the maximum F1 score, sensitivity, specificity, positive predictive value, negative predictive value, and precision on a holdout set of 1000 patients from the stroke service and 100,000 non-overlapping randomly selected patients. We chose this test ratio to imitate the prevalence of AIS in the general population. The models were evaluated on the test set with area under the receiver operating curve (AUROC) and average precision score (AP), a proxy for area under the precision-recall curve. All models were programmed using the Python sklearn scientific computing package (Python Software Foundation, www.python.org)[101]. We then aggregated common features found in the top ten in importance or beta coefficient weight for each model, and we evaluated the contribution of each feature to each model by comparing its prevalence in the cases with its prevalence in the controls and as a function of its importance (or weight) in the model.

2.3.6 Robustness of model

To test the robustness of the models, we trained on 100%, 90%, 70%, 50%, 30%, 10%, 5%, 1%, 0.5%, and 0.1% of the original training set. We then evaluated the performance of each training size on the test set through AUROC and AP. We resampled the set 10 times at each training size and averaged the results.

2.3.7 Calibration

The classifiers assign a probability of being an AIS case to each patient. In order to interpret the meaning of these model probabilities and their applicability to other patients, we developed an empirical calibration method for the models. For each training fold during 10-fold cross validation, we first sorted the probabilities by value. We then averaged a bin of 100 patients intervals based on their assigned probability (predicted probability). Within each bin, we calculated the proportion of cases (actual probability). We repeated this sliding bin, averaging for every 10 patients, creating an empirical function between the training set predicted probabilities and actual probabilities averaged across training folds. We then repeated the same averaging of predicted and actual probabilities for the sorted test set. We calibrated the test set predicted probabilities using the empirical function generated from the training folds. To evaluate calibration success, particularly for identifying cases, we calculated the root mean squared error (RMSE) of all calibrated test probabilities larger than 0.1. We made this cut off because many of the poorly calibrated models have many points near zero, giving a false low RMSE score. Models without any calibrated probabilities larger than 0.1 were given an RMSE of 'N/A'.

2.3.8 Feature collapsing

Reduction of feature size is also essential for reproducibility of the models in other hospital settings. Our approach collapses features together by utilizing the hierarchical structure of ICD9, ICD10, and CPT4 diagnosis codes and procedure codes, and ATC ingredient definitions for medications. We mapped ICD9 and ICD10 diagnosis and procedure codes and drug prescription order

ingredients to SNOMED CT concept IDs. About 40% of the features could not be mapped automatically to SNOMED concepts, so we kept codes and ingredients that did not have concept IDs as separate features. We then mapped ICD9 codes to ICD10 codes using General Equivalence Mappings [97]. We mapped all ICD9 and ICD10 diagnosis and procedure codes and CPT4 procedure codes to the Clinical Classifications System defined by the Health Cost and Utilization Project [98, 102, 103, 104, 105]. The CCS has flat level codes (most granular), and multilevel codes level 1-3 (least to most granular). We mapped all codes to the flat, multilevel level 1 and multilevel level 2 codes. We then mapped drug prescription ingredients to ATC codes by matching corresponding RxNorm codes to ATC codes. Some drugs required manual mapping to ATC codes. All in all, we successfully mapped 70% of 45 million prescription orders to ATC codes. We then reduced prescription order features to the five different levels of ATC, which include, in increasing order of specificity, anatomical systems (14), therapeutic subgroups (78), pharmacological subgroups (159), chemical subgroups (364), and chemical substances (895). We then reran the training sets with collapsed features in three combinations: flat CCS level with chemical substrates (the most granular), multilevel CCS level 2 with pharmacological subgroups, and multilevel CCS level 1 with anatomical systems (the least granular). We chose these pairings to maintain similar ratios of drugs and ICD9 and ICD10 codes. We evaluated the robustness of hierarchical feature collapsing using the method described above. We then compared performance across levels of granularity using AUROC and average precision score, which is a proxy for area under the precision-recall curve.

2.3.9 Generalized linear model to determine feature category contribution to models

To further explore the contribution of conditions, procedures, drugs, and demographics, to the models' classification performance, we ran a multivariate generalized linear model (GLM) on the stroke probabilities generated from each feature category. We retrained the models solely using features from each category. We also added a fifth category, which comprised the ICD9 codes that make up the T-L criteria and ICD10 equivalent codes (see Table 2.1). We added this category

to compare the performance of the T-L criteria to the other features in AIS patient classification because our models do not include the T-L criteria in training. We then generated AIS probabilities for the holdout test set using each of the retrained models. Finally, we combined the model probabilities from each of the five models and ran the generalized linear model with binomial distribution to determine which categories significantly ($p < .05$) contributed to case determination and their corresponding beta coefficients.

2.3.10 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 codes as defined by T-L criteria and associated ICD10 codes.

2.3.11 External validation of acute ischemic stroke patient classification in the UK Biobank

The UK Biobank is a prospective health study of over 500,000 participants, ages 40-69, with comprehensive EHR and genetic data[106]. Given that this dataset contains 4,922 patients with an AIS related ICD10 code, similar to our T case cohort, and 163 patients without AIS related ICD10 codes, the UK Biobank can evaluate our machine learning models' ability to recover potential AIS patients that lack AIS-related ICD10 codes. 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 AIS cases without AIS ICD10 codes 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).

2.4 Results

2.4.1 Study cohort

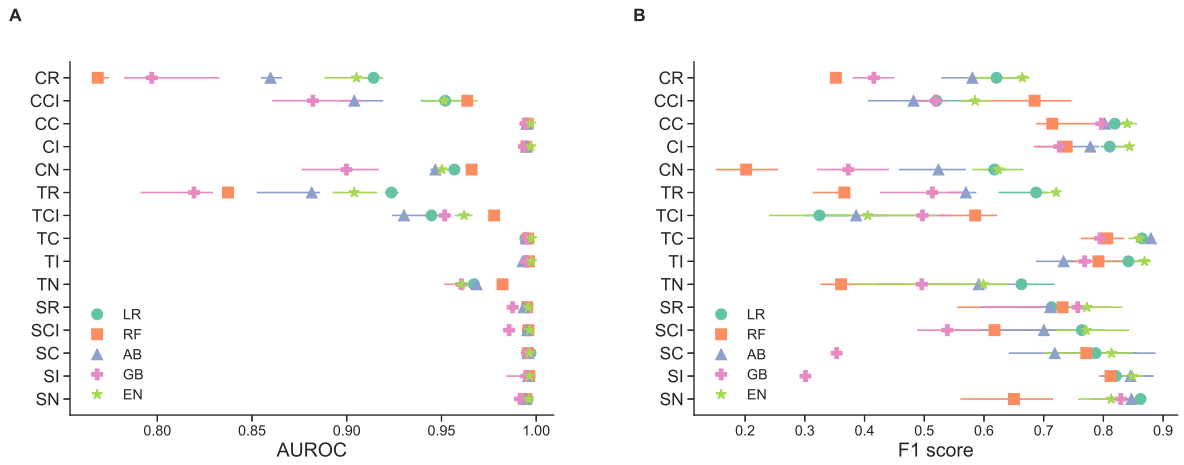
Table 1 presents the data and the total number of patients available for each set of cases and controls used in the training and internal and external validation parts of this study. Out of the CUIMC EHR, which has a total of 6.4 million patients, we extracted 4,844 stroke service patients, which we found to have a 4-16% false positive rate for stroke. Table 2.2 presents demographic characteristics.

Variable	Case- S	Case- T	Case- C		
N	3473	6376-8000	6271-8000		
Gender, Female	1709 (49.2%)	3351-4222 (51.8-52.8%)	3298-4278 (51.9-53.1%)		
Age > 50	3115 (89.7%)	5367-6747 (83.6-84.3%)	5194-6570 (81.4-82.8%)		
Race/Ethnicity					
Black or African American	283 (8.15%)	330-348 (4.18-5.18%)	268-355 (3.94-4.43%)		
White	742 (21.4%)	1083-1364 (16.2-17.0%)	1263-1654 (19.7-20.7%)		
Hispanic or Latino	597 (17.2%)	496-677 (7.78-8.46%)	469-653 (7.48-8.16%)		
Other	40 (1.15%)	43-69 (0.588%-0.862%)	64-76 (0.825-1.02%)		
Unknown/Declined to answer	1811(52.1%)	4425-5646 (69.4-70.6%)	4204-5431 (66.6-67.9%)		
	Control- N	**Control- C**	**Control- I**	**Control- CI**	**Control- R**
N	3465-6346	3473-8000	3473-8000	3473-8000	3458-7920
Gender, Female	2233-4081 (64.3-64.4%)	1927-4492 (55.5-56.1%)	1945-4493 (55.4-56.2%)	1828-4298 (52.6-53.7%)	1970-4547 (56.2-57.4%)
Age > 50	1746-3225 (50.4-50.8%)	1029-2366 (29.4-29.6%)	1040-2458 (29.9-30.7%)	2781-6454 (80.0-80.7%)	1950-4524 (56.4-57.1%)
Race/Ethnicity					
Black or African American	145-277 (4.18-4.36%)	116-271 (3.26-3.39%)	129-295 (3.54-3.71%)	130-342 (3.74-4.28%)	200-495 (5.78-6.27%)
White	591-1113 (17.1-17.5%)	300-734 (8.50-9.18%)	348-725 (8.76-10.0%)	759-1848 (21.8-23.0%)	814-2021 (23.5-25.5%)
Hispanic or Latino	276-482 (7.60-7.96%)	232-559 (6.51-6.99%)	233-515 (5.79-6.70%)	283-630 (7.10-8.15%)	436-1111 (12.6-14.0%)
Other	30-38 (0.599-0.866%)	28-60 (0.662-0.806%)	29-71 (0.738-0.888%)	33-87 (0.950-1.09%)	65-144 (1.75-1.88%)
Unknown/Declined to answer	2424-4440 (70.0%)	2789-6470 (79.7-80.9%)	2734-6470 (78.7-80.9%)	2265-5155 (64.4-65.2%)	1943-4293 (52.8-56.2%)

Table 2.2: **Demographics of Case-control cohorts**

2.4.2 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. Logistic regression classifiers with L1



(a) AUROC

(b) F1 score

Figure 2.2: **Performance of select models on holdout test set** ((a): AUROC, (b): F1).(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 Method 2.3.3 and Table 2.6 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, see Method 2.3.5

penalty gave the best 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 2.2a).

Across all classifier types, the TC model had the best average F1 score (0.832 ± 0.0383), which is a measure of the harmonic mean of precision and recall. Logistic regression models with L1 penalty (LR) and Elastic Net penalty (EN) had the best classifier average F1 score (0.705 ± 0.146 and 0.710 ± 0.134 respectively) (Figure 2.2b). Stroke service cases gave the highest average precision (0.932 ± 0.0536), while cases identified through AIS codes and controls without cerebrovascular disease or the AIS 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 (Table 2.3). Precision at top 50 and 100 patient probabilities was very high, 0.99-1.00, for all classifiers except EN. Precision at top 500 patient probabilities was high except for the TCI, TR, CN, CCI, and CR models (Table 2.4).

Case/ Control Combo	LR Sens	RF Sens	AB Sens	GB Sens	EN Sens	LR Spec	RF Spec	AB Spec	GB Spec	EN Spec	LR PPV	RF PPV	AB PPV	GB PPV	EN PPV	LR NPV	RF NPV	AB NPV	GB NPV	EN NPV
SN	0.749	0.522	0.753	0.699	0.715	1.000	0.999	1.000	0.996	0.999	0.007	0.005	0.007	0.007	0.007	0.997	0.995	0.998	0.997	0.997
SI	0.842	0.748	0.923	0.766	0.825	0.998	0.998	0.997	0.996	0.998	0.008	0.007	0.009	0.008	0.008	0.998	0.997	0.999	0.998	0.998
SC	0.941	0.791	0.959	0.836	0.902	0.996	0.997	0.994	0.996	0.996	0.009	0.008	0.010	0.008	0.009	0.999	0.998	1.000	0.998	0.999
SCI	0.575	0.450	0.524	0.428	0.651	1.000	1.000	1.000	0.998	1.000	0.006	0.004	0.005	0.004	0.006	0.996	0.995	0.995	0.994	0.997
SR	0.586	0.549	0.546	0.522	0.643	1.000	1.000	1.000	1.000	1.000	0.006	0.005	0.005	0.005	0.006	0.996	0.996	0.995	0.995	0.996
TN	0.525	0.331	0.461	0.434	0.388	1.000	0.997	0.999	0.998	1.000	0.005	0.003	0.005	0.004	0.004	0.995	0.993	0.995	0.994	0.994
TI	0.743	0.692	0.576	0.614	0.705	0.999	0.999	0.999	0.999	1.000	0.007	0.007	0.006	0.006	0.007	0.997	0.997	0.996	0.996	0.997
TC	0.822	0.737	0.827	0.784	0.781	0.999	0.999	0.999	0.998	0.999	0.008	0.007	0.008	0.008	0.008	0.998	0.997	0.998	0.998	0.998
TCI	0.205	0.381	0.329	0.439	0.250	1.000	1.000	0.997	0.999	1.000	0.002	0.004	0.003	0.004	0.002	0.992	0.994	0.993	0.994	0.993
TR	0.602	0.248	0.522	0.370	0.552	0.997	0.999	0.996	0.998	0.999	0.006	0.002	0.005	0.004	0.005	0.996	0.993	0.995	0.994	0.996
CN	0.471	0.181	0.394	0.273	0.382	0.999	0.993	0.998	0.999	0.999	0.005	0.002	0.004	0.003	0.004	0.995	0.992	0.994	0.993	0.994
CI	0.736	0.746	0.706	0.667	0.752	0.999	0.996	0.998	0.998	0.999	0.007	0.007	0.007	0.007	0.007	0.997	0.997	0.997	0.997	0.998
CC	0.833	0.827	0.797	0.780	0.798	0.998	0.995	0.998	0.998	0.999	0.008	0.008	0.008	0.008	0.008	0.998	0.998	0.998	0.998	0.998
CCI	0.347	0.486	0.496	0.352	0.395	1.000	1.000	0.992	0.998	1.000	0.003	0.005	0.005	0.004	0.004	0.994	0.995	0.995	0.994	0.994
CR	0.596	0.274	0.538	0.424	0.649	0.995	0.996	0.995	0.997	0.982	0.006	0.003	0.005	0.004	0.007	0.996	0.993	0.995	0.994	0.996

Table 2.3: **Sensitivity, specificity, positive predictive value, and negative predictive value of models on holdout test set.** See Table 2.6 for case-control, model, and evaluator abbreviations' definitions. Blue values > 0.85, yellow values between 0.65 and 0.85, red values < 0.65.

Case/ Control Combo	LR Precision @50	RF Precision @50	AB Precision @50	GB Precision @50	EN Precision @50	LR Precision @100	RF Precision @100	AB Precision @100	GB Precision @100	EN Precision @100	LR Precision @500	RF Precision @500	AB Precision @500	GB Precision @500	EN Precision @500	LR Precision @NCases	RF Precision @NCases	AB Precision @Ncases	GB Precision @Ncases	EN Precision @Ncases
SN	1.00	1.00	1.00	1.00	0.86	1.00	1.00	1.00	0.92	0.75	1.00	1.00	1.00	0.99	0.87	0.76	0.79	0.79	0.79	0.67
SI	1.00	1.00	1.00	1.00	0.89	1.00	1.00	1.00	0.97	0.81	1.00	1.00	1.00	0.99	0.90	0.97	0.99	1.00	1.00	0.91
SC	1.00	1.00	1.00	0.99	0.88	1.00	1.00	1.00	0.97	0.81	1.00	1.00	1.00	1.00	0.91	0.80	0.80	0.80	0.80	0.73
SCI	1.00	1.00	1.00	1.00	0.92	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	0.91	0.72	0.77	0.79	0.81	0.78
SR	1.00	1.00	1.00	1.00	0.91	1.00	1.00	1.00	0.99	0.83	1.00	1.00	1.00	1.00	0.91	1.00	1.00	1.00	1.00	0.89
TN	1.00	1.00	1.00	0.98	0.76	0.68	0.72	0.70	0.58	0.49	1.00	1.00	0.98	0.89	0.74	0.90	0.95	0.94	0.81	0.63
TI	1.00	1.00	1.00	0.99	0.87	1.00	1.00	1.00	0.97	0.84	0.99	1.00	0.99	0.92	0.77	1.00	0.99	0.98	0.92	0.77
TC	1.00	1.00	1.00	0.99	0.88	1.00	1.00	1.00	0.96	0.83	1.00	1.00	1.00	0.99	0.88	1.00	0.99	0.98	0.92	0.81
TCI	1.00	1.00	1.00	1.00	0.79	1.00	0.99	0.99	0.97	0.80	0.74	0.64	0.79	0.62	0.64	0.98	0.97	0.96	0.84	0.65
TR	1.00	1.00	1.00	0.96	0.68	0.99	0.97	0.88	0.58	0.39	0.83	0.73	0.73	0.73	0.55	0.99	0.99	0.98	0.82	0.54
CN	1.00	1.00	1.00	0.90	0.69	0.13	0.15	0.15	0.18	0.21	0.9	0.81	0.72	0.66	0.57	0.95	0.90	0.86	0.64	0.47
CI	1.00	1.00	1.00	0.99	0.85	1.00	1.00	0.99	0.89	0.73	1.00	1.00	0.99	0.95	0.80	0.97	0.98	0.97	0.89	0.76
CC	1.00	1.00	1.00	0.99	0.84	1.00	1.00	1.00	0.93	0.79	1.00	1.00	1.00	0.98	0.82	0.95	0.96	0.97	0.93	0.78
CCI	1.00	1.00	1.00	0.98	0.78	1.00	1.00	1.00	0.98	0.81	0.63	0.70	0.69	0.21	0.32	0.99	0.96	0.91	0.69	0.50
CR	1.00	1.00	1.00	0.88	0.64	1.00	0.88	0.82	0.48	0.35	0.71	0.43	0.35	0.58	0.53	0.99	0.98	0.95	0.71	0.49

Table 2.4: **Precision at top 50, 100, and number of known cases for each classifier.** Blue values > 0.99, yellow values between 0.9 and 0.99, orange values between 0.7 and 0.9, red values < 0.7.

2.4.3 Feature importances

We aggregated common features found in the top ten in importance or beta coefficient weight for each model, as seen in Figure 2.3A. We found the most commonly chosen features were procedures used in evaluation of AIS, including extra and intra-cranial arterial scans, CT scans and MRIs of the brain, and MR angiography. We also found age over 50 to be a top common demographic feature, and Aspirin and atorvastatin (a cholesterol lowering medication) to be important medication features. Top diagnosis features included unspecified essential hypertension, history of transient ischemic attack or stroke without residual effect, pulmonary congestion and hypostasis, convulsions, and iatrogenic cerebrovascular infarction or hemorrhage. We also evaluated the contribution of each feature to each model by comparing its prevalence in the cases with its prevalence in the controls and as a function of its importance (or weight) in the model. Figure 2.3B plots

one of the 75 models, TC case-control combination with an adaboost classifier. We found that all 75 models relied on incremental contributions from many different features (figs. 2.3 and 2.24 to 2.38).

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

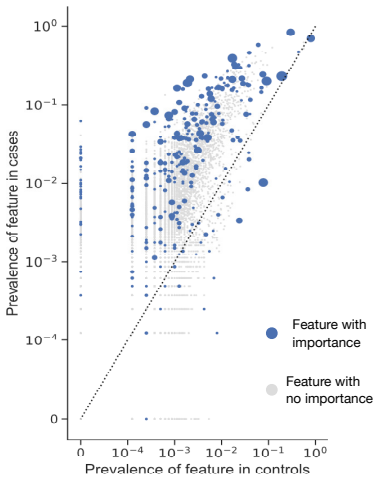


Figure 2.3: **A) Common top 10 features in the models,(B) Prevalence of features in cases vs controls in the TC AB model.** (A) After each of the 75 models were trained, we counted the number of times each feature was represented as one of the top ten by absolute coefficient weight, for methods like 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) 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.

2.4.4 Feature reduction and subsequent algorithm performance.

The number of features of the models ranged from 22,000-35,500 depending on case-control combination and classifier type. We evaluated the performance of collapsing features by ICD9 and ICD10 and Anatomical Therapeutic Chemical Classification System (ATC) code hierarchy, reducing the number of features per model to 72-1500 features, depending on the class size and collapsing level (see Method 2.3.8). As seen in Figure 2.4, we found a 9-27% reduction in AUROC performance, compared to a 90-94% reduction in average precision score.

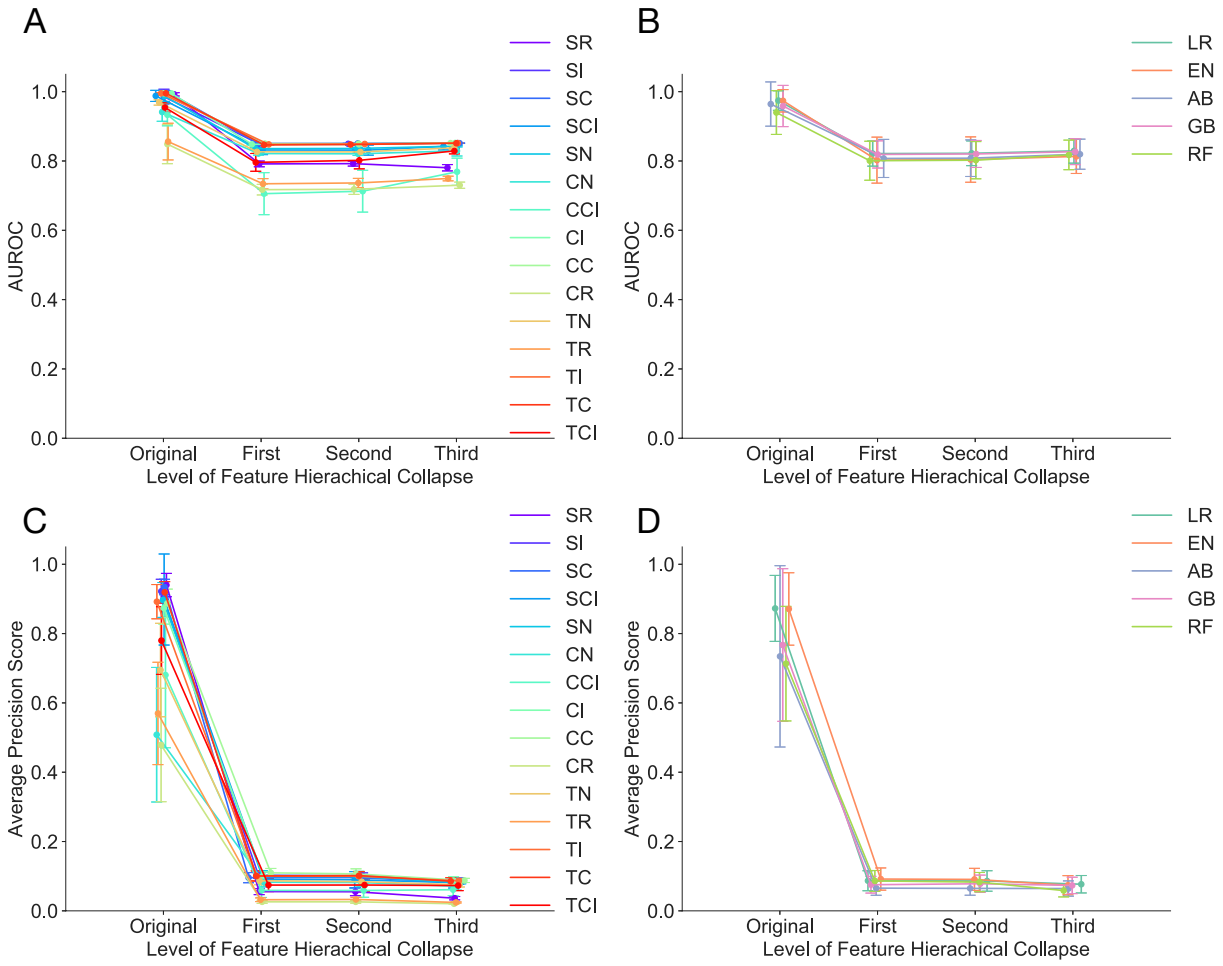


Figure 2.4: **Performance of models at increasing level of feature hierarchical collapse.** Logistic regression with L1 penalty (LR), Logistic Regression with elastic net penalty (EN), random forest (RF), AdaBoost (AB), gradient boosting (GB), and a two-layer neural network (NN). (a) The area under the receiver operating characteristic curve (AUROC) for each of the models versus the level of abstract with 'First' being the lowest level of abstraction and 'Third' being the highest. (b) AP, annotation described in (a).

2.4.5 Calibration

We calibrated the models to the training set results to produce meaningful probability estimates (See Method 2.3.5, Figure 2.5 and supplementary material). We found a wide range of success in calibration of the models. Uncalibrated, the AdaBoost classifier produced a narrow range of probabilities, between 0.4 and 0.6, while the other models ran a range of probabilities from 0 to 1. On average, test set calibration of the AdaBoost models had the lowest root mean squared error (RMSE). The AdaBoost models with controls without cerebrovascular disease codes (C) showed the lowest RMSE for each respective case set (RMSE=0.0917 (SC), 0.1212 (TC), and 0.1779 (CC)). Models with no calibrated predicted probabilities above 0.1 were given an RMSE of N/A, and this applied to 23 of the models. These models mostly had a case set of patients with any cerebrovascular disease code (C) or used a gradient boosting classifier. Notably, the logistic regression model with L1 penalty with S cases and random (R) controls or controls with cerebrovascular disease but no stroke (CI) had uncalibrated test set probabilities with low RMSE, but high RMSE after calibration (see figure 2.12, figure 2.13).

2.4.6 Robustness

We evaluated the robustness of the models to increasingly smaller training sets (Figure 2.6, 2.7). When stratified by classifier type, all classifier types except AdaBoost maintained at least 80% of its AUROC performance with 1% of the training size (70-160 samples) and maintained at least 70% of its AP performance with 5% of the training size (350-800 samples). When stratified by case, models with stroke service cases (S) maintained at least 96% of its AUROC performance with 1% of the training size (~70 samples) and maintained at least 94% of its AP performance with 5% of the training size (~350 samples). Other cases (T, C) maintained at least 92% of its AUROC performance and 77% of its AP performance with 5% of the training size (~350 samples). When stratified by control, models with stroke mimetic controls (N) maintained at least 99% of its AUROC performance and 92% of its AP performance with 5% of the training size (350-635 samples). Random patient controls (R) maintained 88% of its AUROC performance and 75% of

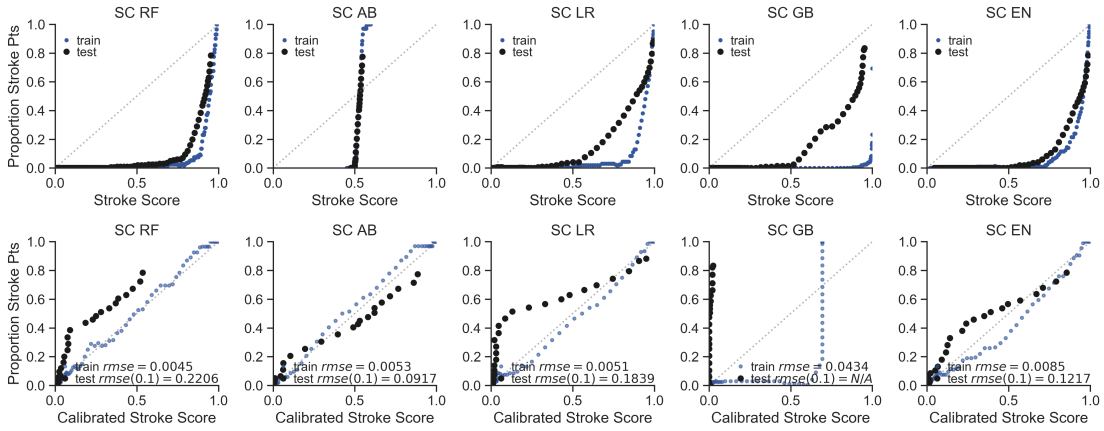


Figure 2.5: **Classifier type with stroke service cases and without cerebrovascular disease (SC) case-control combination varies in calibration success between stroke score, or model probability, and actual proportion of patients at each probability.** Top panel plots the proportion of stroke case patients within a bin of 100 patients with similar stroke scores for models trained with stroke service cases and controls without cerebrovascular disease ICD9 or 10 codes. The black circles plot the mean test stroke score across all training set folds, while the blue circles plot the training fold stroke scores combined. Classifier type varies from left to right:Random Forest (RF), AdaBoost (AB), Logistic Regression with L1 penalty (LR), Gradient Boosting (GB), Logistic regression with elastic net penalty (EN). Bottom panel plots the proportion of stroke patients within a bin of 100 test set patients with similar stroke scores versus the calibrated test scores at specific scores rounded to the nearest thousandth (Black dots). The test scores are calibrated from an empirical distribution determined from the training set 2.3.7. The blue dots plot the proportion of stroke patients within a bin of 100 bootstrapped training set patients with similar stroke scores versus the calibrated training set scores rounded to the nearest thousandth. The light grey dots show perfect calibration.

its AP performance with 5% of the training size (350-800 samples). Controls without T-L AIS codes (I) maintained at least 97% of its AUROC performance and 91% of its AP performance with 1% of the training size (70-160 samples) and 5% of the training size (350-800 samples), respectively. Controls without cerebrovascular disease codes (C) maintained at least 91% of its AUROC performance and 92% of its AP performance with 0.5% of the training size (35-80 samples) and 5% of the training size (350-800 samples), respectively. Finally, controls with cerebrovascular disease but without T-L AIS codes (CI) maintained at least 91% of its AUROC performance and 91% of its AP performance with 5% of the training size (700-1600 samples) and 10% of the training (1400-3200 samples), respectively. The SI case-control combination and EN classifier type model was removed from Figure 2.6,2.7 and the above summary because of very high standard deviation

(125-150% of value) that obscured the trends of the other models.

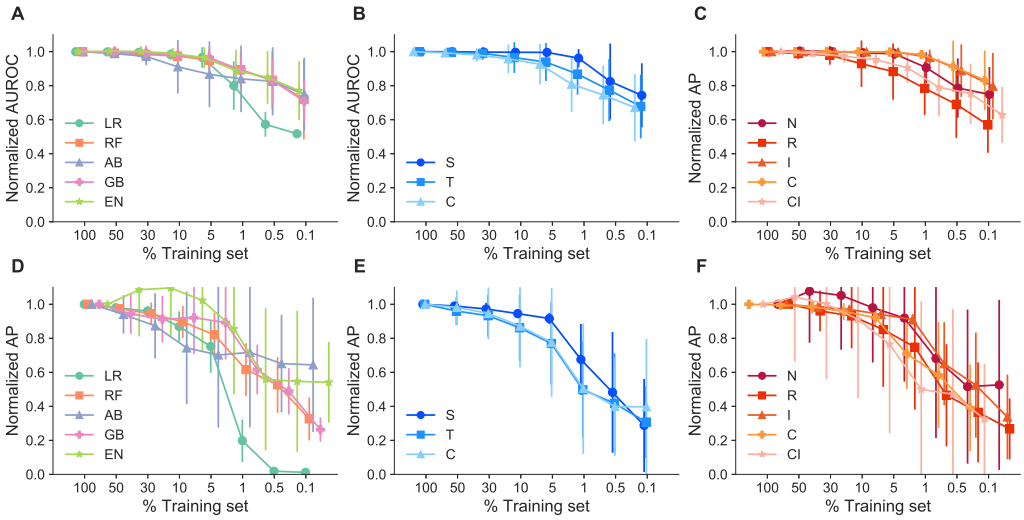


Figure 2.6: **Robustness of models. Top: Normalized area under the receiver operating curve (AUROC) across all case-control combinations.** A: stratified by classifier type (LR, RF, AB, GB, and EN), B: stratified by case type (first letter S,T,C), C: stratified by control type (second and third letters: N, R, I,C,CI). Bottom: Normalized Average precision score (AP) values across all case-control combinations, D: stratified by classifier type (LR, RF, AB, GB, and EN, E: stratified by case type (first letter S,T,C), F: stratified by control type (second and third letters: N, R, I,C,CI). Neural Network Models and EN GI model were removed from normalized graphs due to high variance and obfuscation of other models' trends.

2.4.7 Feature category contribution to models

We trained models on individual feature type categories (procedures, medications, conditions, and T-L AIS conditions) and compared their contributions to model fit using nested linear models (Method 2.3.9). We found that procedures contributed the most to model probabilities (Figure 2.8). 89% of the generalized linear model (GLM) models had a significant ($p < 0.05$) coefficient weight for the procedure feature category, followed by medication orders, which made a significant contribution in 59% of the models. The T-L AIS codes feature category, notably, was significant in 17% of the models. The beta coefficient of the T-L category was 2-20 times larger than the other category coefficients.

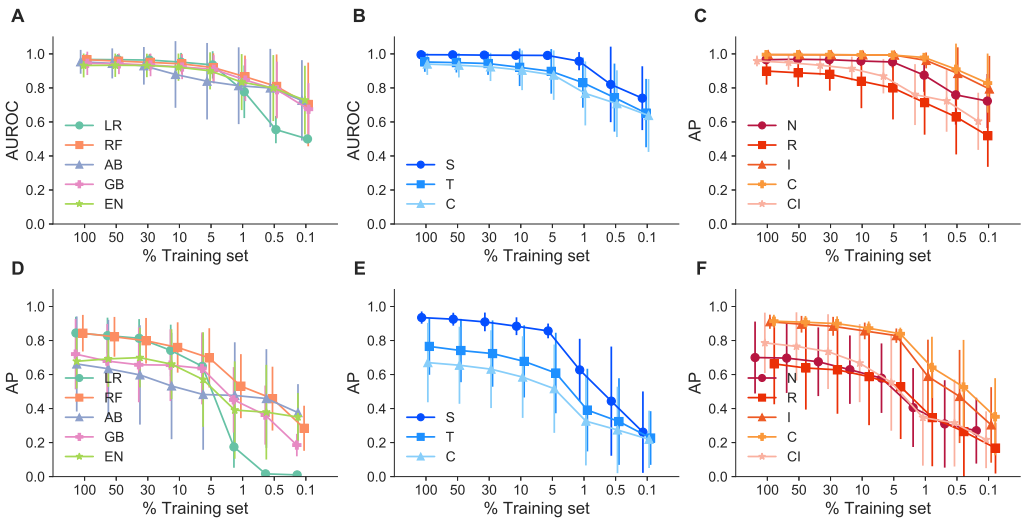


Figure 2.7: **Robustness of models.** Top panel, Left to Right: Area under the receiver operating curve (AUROC) across all case-control combinations, A: stratified by classifier type (LR, RF, AB, GB, and EN), B: AUROC across all models stratified by case type (first letter: S, T, C), C: AUROC across all case-control combinations, stratified by control type (second and third letters: N, R, I,C,CI). Bottom Panel, left to right: D: Average precision score (AP) values across all case-control combinations, stratified by model, E: AP across all models stratified by case type, F AP across all models stratified by control type. Neural Network Models and EN GI model were removed from normalized graphs due to high variance and obfuscation of other models' trends.

2.4.8 Internal validation in institutional EHR

We applied our models to the entire CUIMC EHR to estimate the prevalence of AIS patients (Method 2.3.10). We trained 75 models using 15 case-control combinations and 5 classifier types and then applied the models to the entire CUIMC EHR with at least one diagnosis code, totaling 5,324,725-5,315,923 patients depending on the case/control set. Based on the thresholds defined by the maximum F1 score from the training set (see Method 2.3.5), we determined the prevalence of AIS patients estimated in the EHR by each model. From these proposed cases, we also calculated the proportion of patients with an ICD9 or ICD10 code for AIS defined by T-L. We found that the results varied widely across models, but most predicted that a prevalence 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%

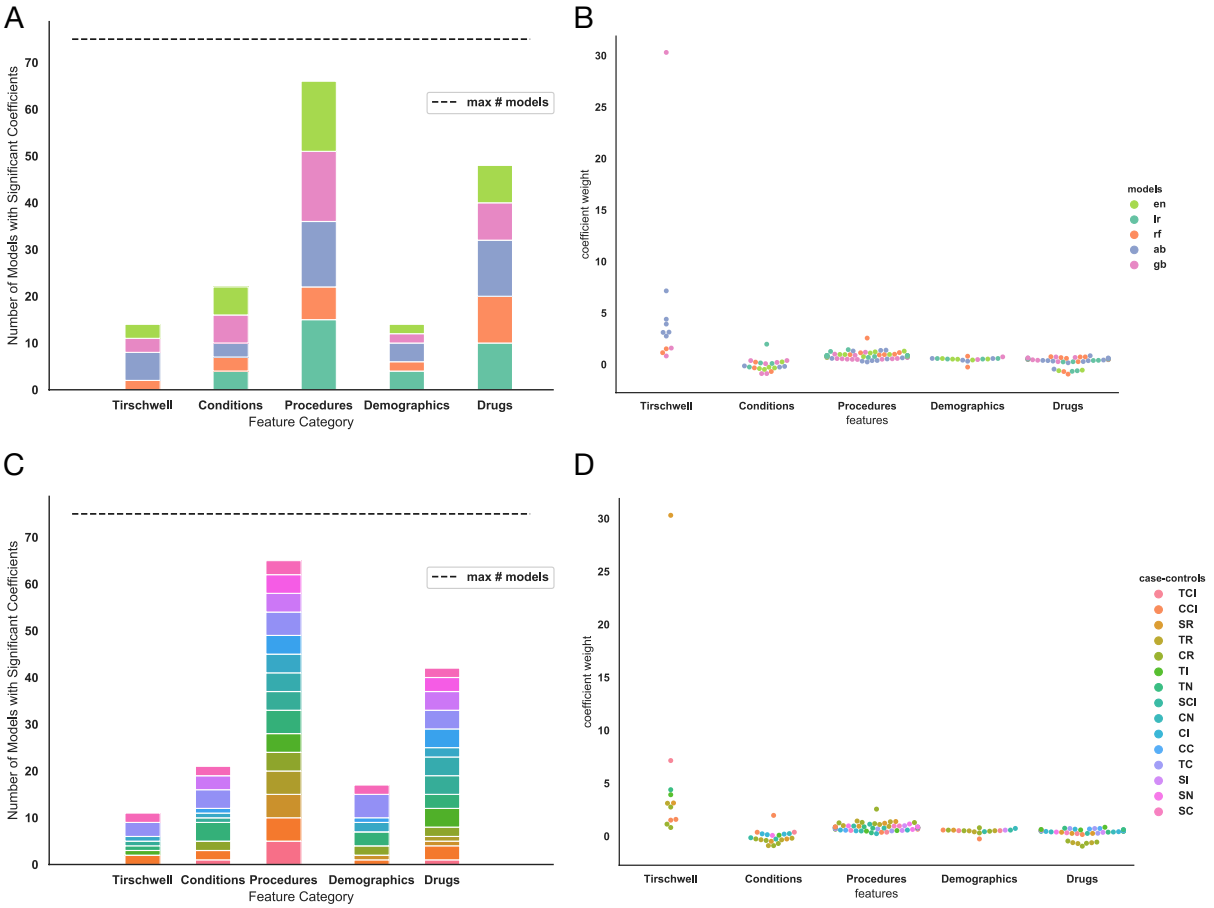


Figure 2.8: **Feature category contribution to model fits.** Top panel Left to right: A shows the total number of models with a significant coefficient weight for each feature category, stratified by classifier type. B shows the value of the significant coefficient weights, stratified by classifier type. Bottom panel Left to right: C shows the total number of models with a significant coefficient weight for each feature category, stratified by case-control type. D shows the value of the significant coefficient weights, stratified by case-control type. TCI Random Forest model and CCI EN model had significant coefficients in all categories, and all of these weights were greater than 10E14, so they would be plotted beyond the scope of the y axis in B and D.

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% (see Table 2.3, Table 2.5). Within these proposed AIS patients, 37.2-47.2% had an AIS diagnosis code (see Table 2.5).

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 2.5: **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 Table 2.6 for case-control and model abbreviations' definitions.

2.4.9 External validation of acute ischemic stroke patient classification in the UK Biobank

We evaluated the performance of the TC models to identify 2,624 patients without AIS ICD10 codes. As seen in Figure 2.9, the top 50, 100, 500, and 2,624 probabilities had a precision of over 29%, and up to 80%. 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.

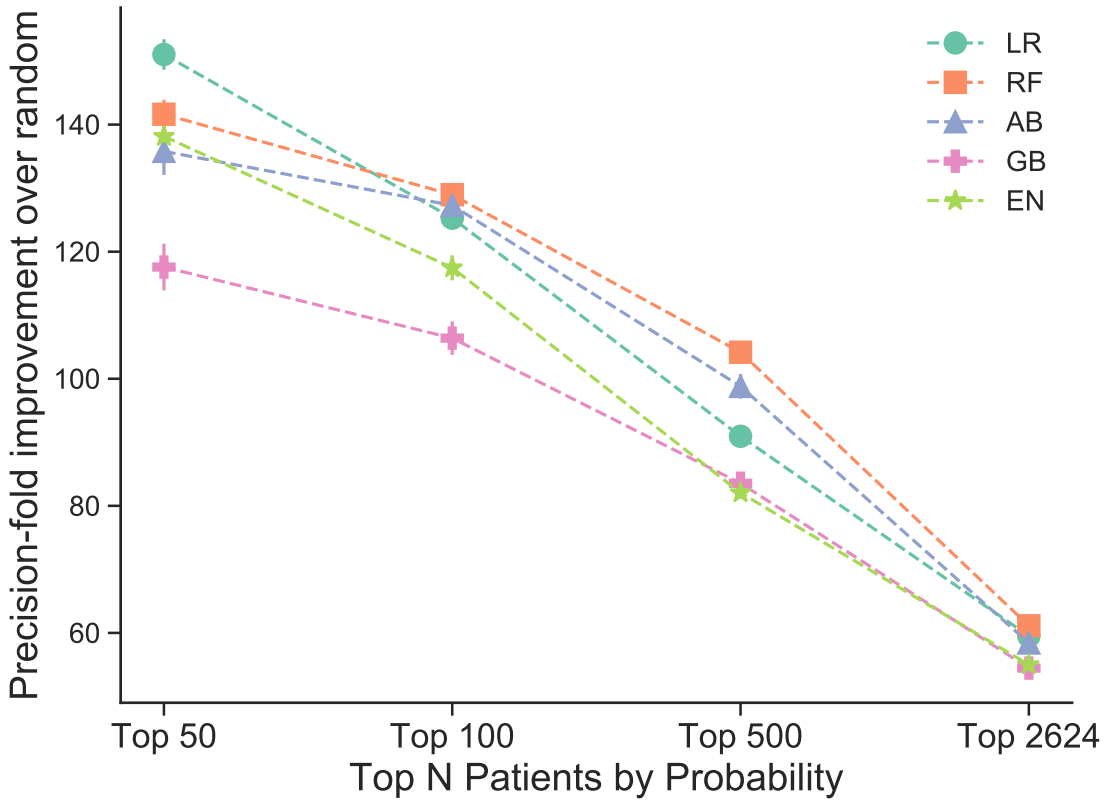


Figure 2.9: **Precision-fold over random sampling of acute ischemic stroke cases without related ICD10 codes at top 50, 100, 500, and 2,624 patient probabilities assigned by machine learning algorithms.** With 95% confidence intervals in error bars. See 2.6 for model abbreviations' definitions.

2.5 Discussion

2.5.1 Machine learned models are able to identify acute ischemic stroke patients without direct evidence.

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 used AIS-related diagnosis codes as the sole nonzero features in their models. [77, 78, 24]. We show that AIS patients can be recovered from other EHR-available structured clinical features excluding the T-L criteria.

2.5.2 Case and control choices are important for acute ischemic stroke phenotyping.

Case-control selection for 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 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 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. 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.

2.5.3 Procedures serve as a proxy for acute ischemic stroke diagnosis codes in model features.

We found that stroke evaluation procedures, such as a CT scan or MRI, were important features in many of the models. Since none of these models use AIS diagnosis codes as features, this suggests that procedures may serve as proxies for AIS cohort identification. In some cases, the AIS code will only be added during outpatient follow up. For example, 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 found that procedures provided a significant contribution ($p < .05$) to patient classification in 89% of the models, while the T-L diagnosis codes provided a significant contribution in 17% of the models (though the T-L beta coefficient was 2-20 times higher than the other categories). This suggests that procedures are important proxy features for the T-L AIS diagnosis codes (Figure 2.3).

2.5.4 Other diagnosis codes may be useful for phenotyping acute ischemic stroke.

In addition, iatrogenic cerebrovascular infarction or hemorrhage (SNOMED ID 44834124, ICD9 997.02) was found to be a top feature and is a code not within the cerebrovascular infarction ICD9 code class. Although this code does not distinguish between an ischemic and hemorrhagic stroke, it was found in 4% of cases and could be a useful addition for AIS cohort identification purposes. Other conditions that were highlighted by the models include hemiplegia and convulsions. This is consistent with features that may be coded during initial inpatient evaluation for a stroke. In some cases, the actual AIS will only be coded during outpatient followup.

2.5.5 Models showed robustness to reduction of training set size, but not with code-hierarchy-based feature reduction.

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. However, reducing feature set size by collapsing with CCS and ATC hierarchy resulted in large drops in performance, most likely due to the 95-99% drop in feature set size and high level description of the features. This suggests that the feature reduction by hierarchical collapse trains with excessively high-level features, leading to unsuccessful AIS identification. Feature size reduction of up to 60% through training set size reduction, however, maintained high performance of the models.

2.5.6 Calibration using an empirical function differentiates the models and may identify additional control sets.

We calibrated the output of the models so that the predicted probability could be meaningfully interpreted (see Method 2.3.7). After recalibration, the AdaBoost classifier type, in particular,had the lowest root mean squared error (RMSE) of probabilities predicted. AdaBoost models that were trained with controls without cerebrovascular disease produced the lowest RMSE. This suggests the need for a more stringent definition of controls than simply no AIS code diagnosis. AdaBoost and logistic regression models trained on stroke service patients as cases and stroke mimics as controls also performed well and showcases another important control set to consider. Some case-control sets, namely stroke service cases with random patients as controls and cerebrovascular disease without AIS codes as controls produced well-calibrated test sets before calibration by the boot-strapped training set but poor calibration in testing. This suggests that our training set may have over-weighted the importance of the controls. In addition, we evaluate the calibration performance based on the RMSE of the calibrated test set, when in the future we would like to evaluate based on the training set. Our study shows that calibration potential of each model may better discriminate model success for studies with a large control:case ratio than traditional evaluation methods such as AUROC.

2.5.7 Models can identify a large number of stroke patients without acute ischemic stroke diagnosis codes.

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 codes up to 150-fold better than random selection.

2.5.8 Limitations

This study has several limitations. First, we relied on noisy labels and proxies for training our models, as evidenced by the false positive rate of 4-16% that was determined by manual review. Without a gold standard set of cases, model performance is difficult to definitively evaluate. Second, we used only structured features contained within standard terminologies across the patients' entire timeline and did not use clinical notes. 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.

2.5.9 Strengths

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, it may be generalizable to other health systems outside CUIMC, whereas recent studies have relied on manually curated feature sets[77]. 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.

2.6 Conclusions and future directions

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. Specifically, our phenotyping model aims to place patients on a spectrum of being a stroke patient. The probability represents patients along this spectrum including those with direct evidence of stroke, patients who have predisposition to stroke but have not experienced the environmental trigger leading to stroke, patients with some risk factors for stroke but have not had a stroke, and patients who have very few to no risk factors for stroke and have not had a stroke. 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[107, 108, 109, 110, 111]. 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.

2.7 Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 41039. I thank Dr. Benjamin R. Kummer for help in chart review of stroke patients, providing the stroke service patient information, and helpful edits on the paper associated with this chapter[112].

I also thank Dr. Tal Lorberbaum for sharing helpful code and edits on the paper. I also thank Dr. Mitchell V. S. Elkind for providing the stroke service patient information and supervision guidance on this study and edits on the paper. I also 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.

2.8 Supplementary Materials

Abbreviation	Meaning
AIS	Acute Ischemic Stroke
T-L	Tirschwell- Longstreth
ICD9	International Classification of Disease, Version 9
ICD10	International Classification of Disease, Version 10
ATC	Anatomical Therapeutic Chemical Classification System
CCS	Clinical Classifications Software
LR	Logistic Regression with L1 penalty
EN	Logistic Regression with Elastic Net penalty
RF	Random Forest
AB	Adaboost
GB	Gradient Boosting
NN	Neural Network
CvD	Cerebrovascular Disease
SN	Stroke Service Cases, Stroke Mimetic Controls
SI	Stroke Service Cases, Controls without T-L codes for AIS
SC	Stroke Service Cases, Controls without ICD9 or ICD10 codes for CvD
SCI	Stroke Service Cases, Controls with ICD9 or ICD10 codes for CvD, and without T-L codes for AIS
SR	Stroke Service Cases, Random patients in the EHR as controls
TN	Cases with T-L codes for AIS, Stroke Mimetic Controls
TI	Cases with T-L codes for AIS, Controls without T-L codes for AIS
TC	Cases with T-L codes for AIS, Controls without ICD9 or ICD10 codes for CvD
TCI	Cases with T-L codes for AIS, Controls with ICD9 or ICD10 codes for CvD, and without T-L codes for AIS
TR	Cases with T-L codes for AIS, Random patients in the EHR as controls
CN	Cases with ICD9 or ICD10 codes for CvD, Stroke Mimetic Controls
CI	Cases with ICD9 or ICD10 codes for CvD, Controls without T-L codes for AIS
CC	Cases with ICD9 or ICD10 codes for CvD, Controls without ICD9 or ICD10 codes for CvD
CCI	Cases with ICD9 or ICD10 codes for CvD, Controls with ICD9 or ICD10 codes for CvD and without T-L codes for AIS
CR	Cases with ICD9 or ICD10 codes for CvD, Random patients in the EHR as controls
RMSE	Root Mean Squared Error
GLM	Generalized Linear Model
CMS	Centers for Medicare and Medicaid Services
AUROC	Area under the Receiver Operating Curve
AP	Average Precision Score
P	Precision
R	Recall
FB	F-score with beta coefficient B
EHR	Electronic Health Record
PPV	Positive Predictive Value
NPV	Negative Predictive Value
CUIMC	Columbia University Irving Medical Center
CDW	Common Data Warehouse

Table 2.6: **Abbreviations in this Study**

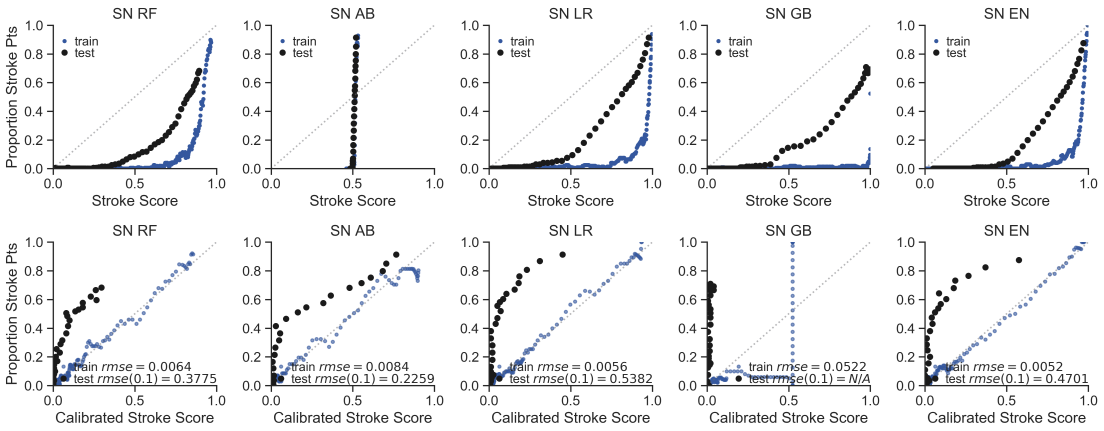


Figure 2.10: **Classifier type with Stroke Service Cases, Stroke Mimetic Controls (SN) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

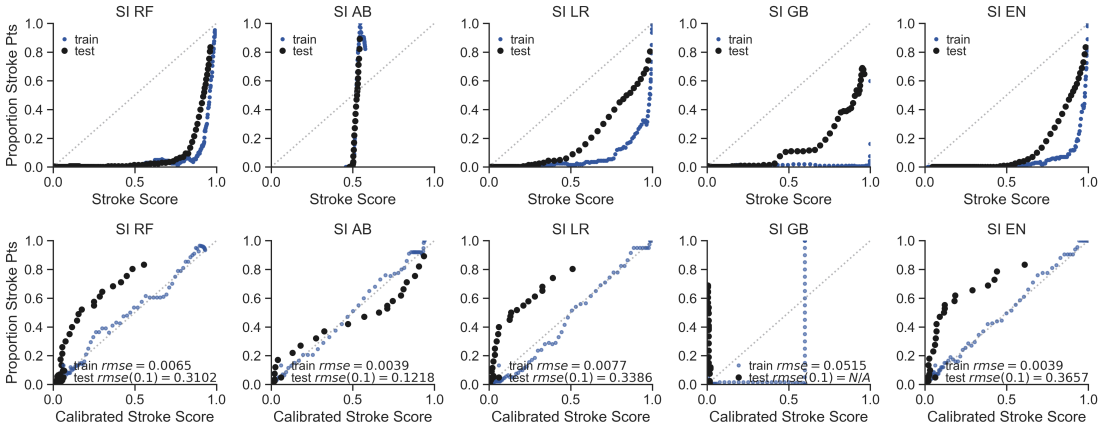


Figure 2.11: **Classifier type with Stroke Service Cases, Controls without T-L codes for AIS (SI) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

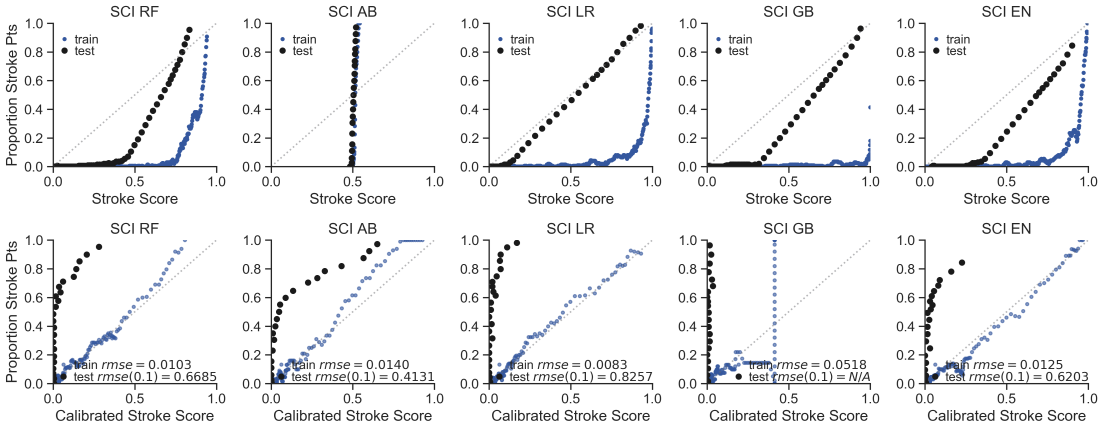


Figure 2.12: **Classifier type with Stroke Service Cases, Controls with ICD9 or ICD10 codes for CvD, and without T-L codes for AIS (SCI) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

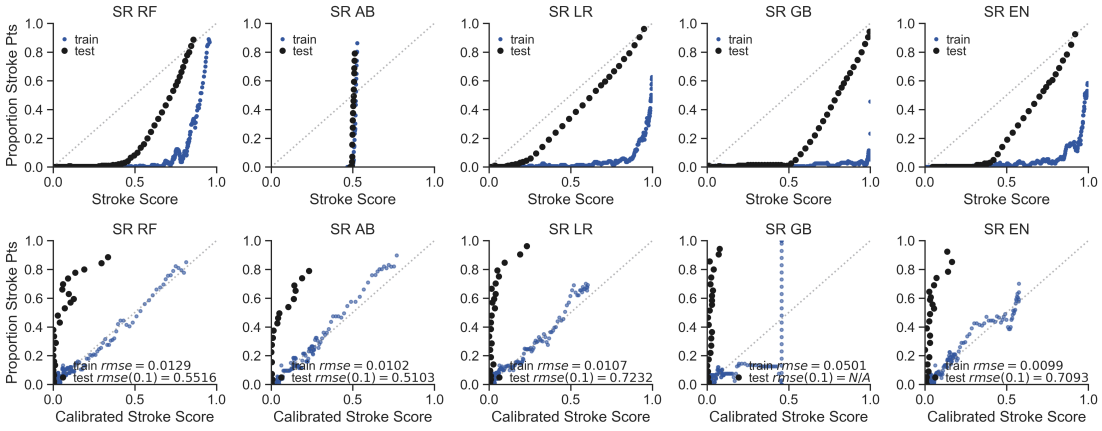


Figure 2.13: **Classifier type with Stroke Service Cases, Random patients in the EHR as controls (SR) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

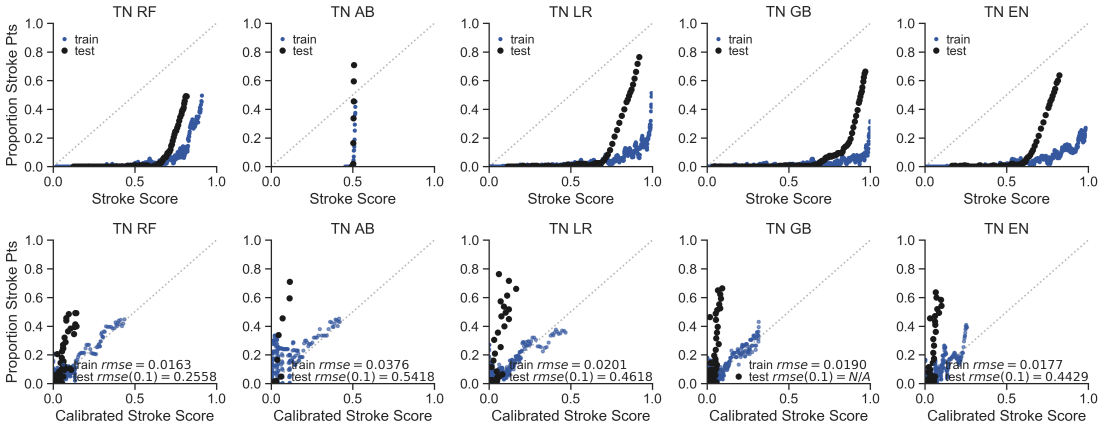


Figure 2.14: **Classifier type with Cases with T-L codes for AIS, Stroke Mimetic Controls (TN) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

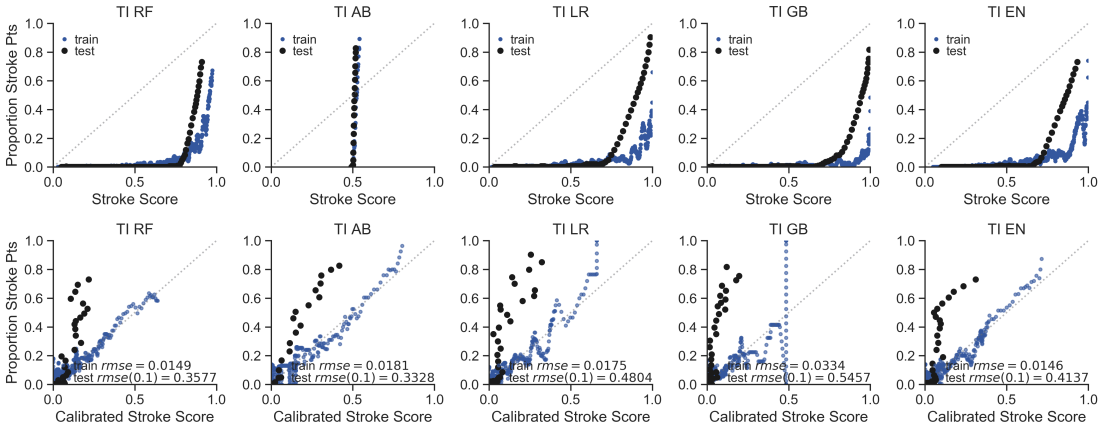


Figure 2.15: **Classifier type with Cases with T-L codes for AIS, Controls without T-L codes for AIS (TI) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

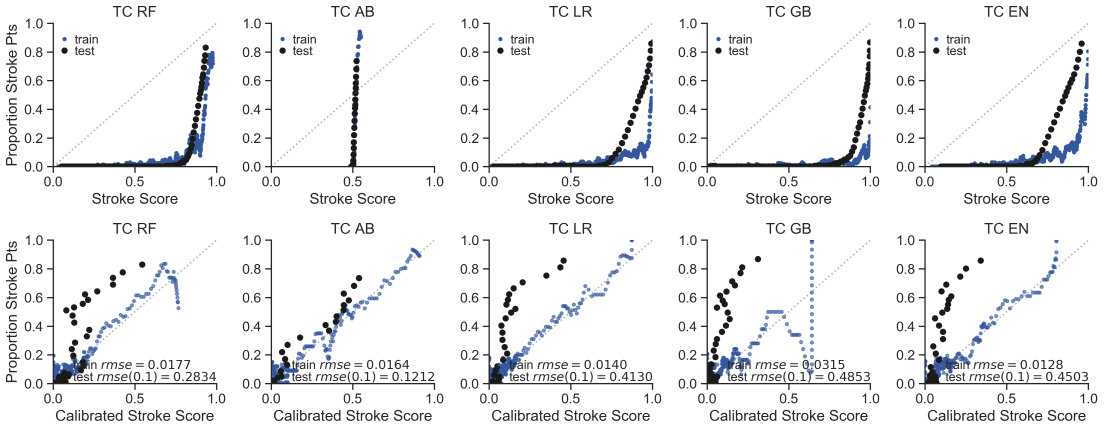


Figure 2.16: **Classifier type with Cases with T-L codes for AIS, Controls without ICD9 or ICD10 codes for CvD (TC) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

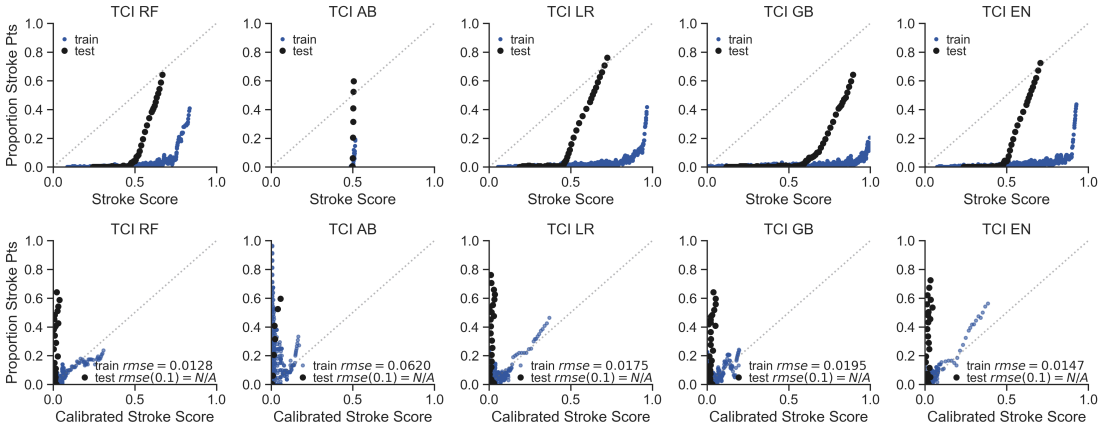


Figure 2.17: **Classifier type with Cases with T-L codes for AIS, Controls with ICD9 or ICD10 codes for CvD, and without T-L codes for AIS (TCI) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

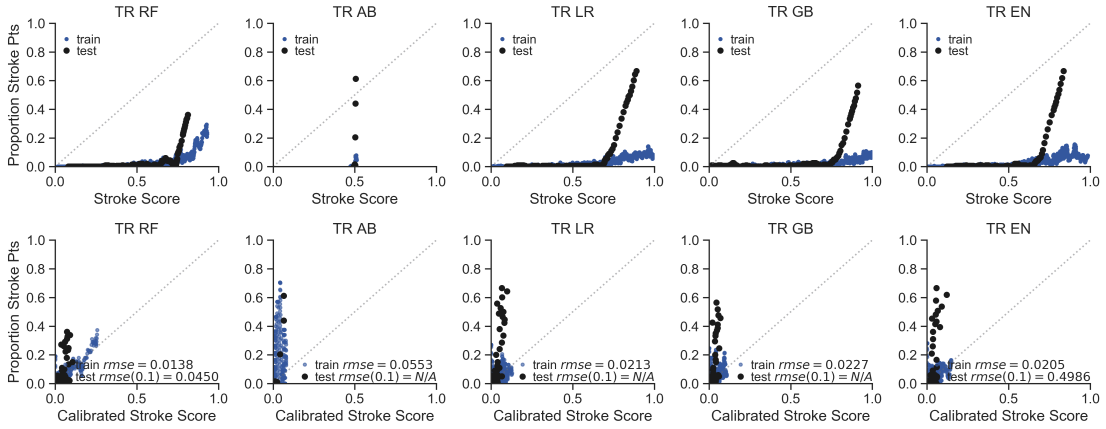


Figure 2.18: **Classifier type with Cases with T-L codes for AIS, Random patients in the EHR as controls (TR) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

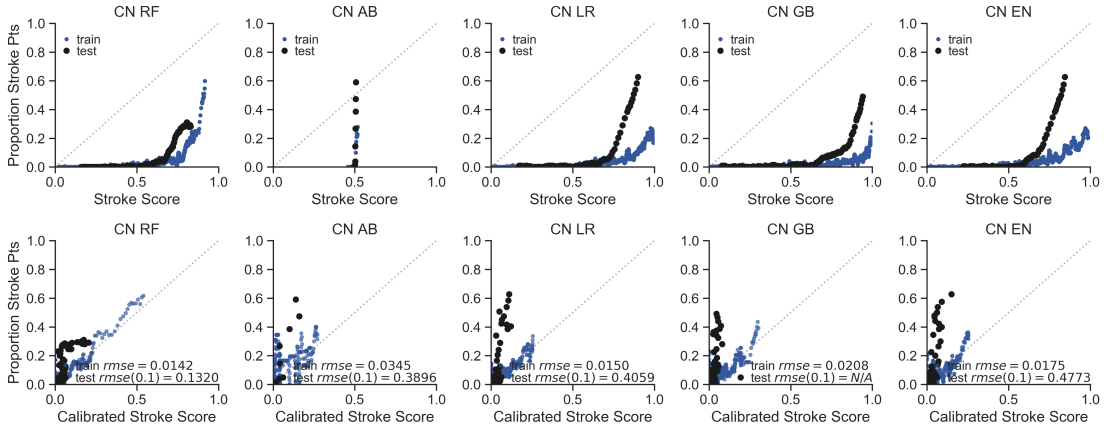


Figure 2.19: **Classifier type with Cases with ICD9 or ICD10 codes for CvD, Stroke Mimetic Controls (CN) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

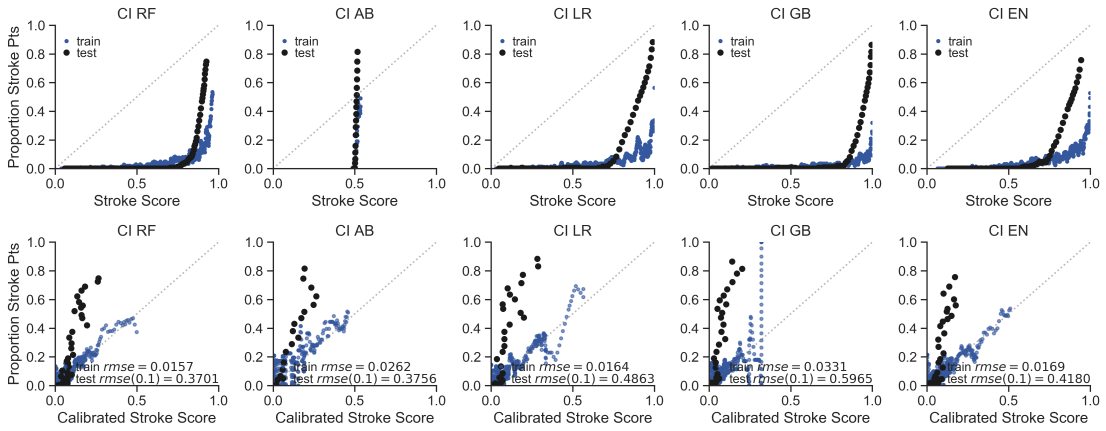


Figure 2.20: **Classifier type with Cases with ICD9 or ICD10 codes for CvD, Controls without T-L codes for AIS (CI) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

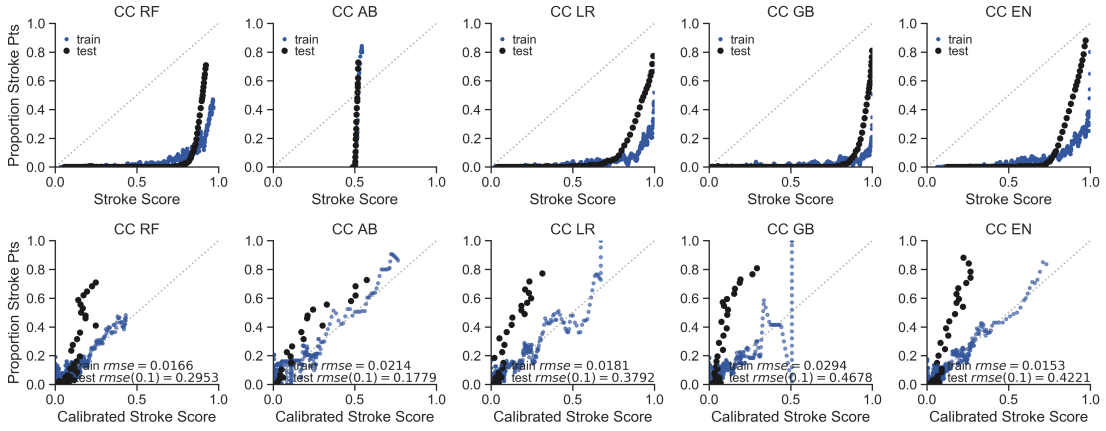


Figure 2.21: **Classifier type with Cases with ICD9 or ICD10 codes for CvD, Controls without ICD9 or ICD10 codes for CvD (CC) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

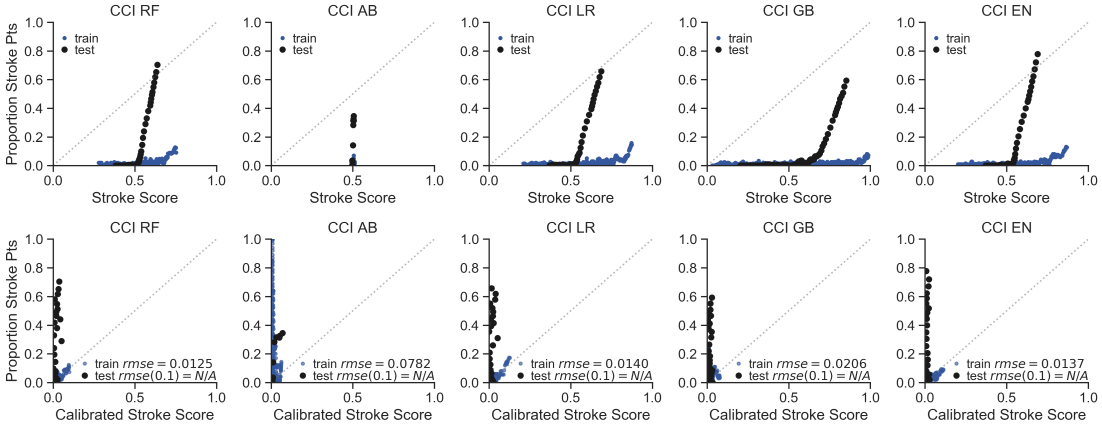


Figure 2.22: **Classifier type with Cases with ICD9 or ICD10 codes for CvD, Controls with ICD9 or ICD10 codes for CvD and without T-L codes for AIS (CCI) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

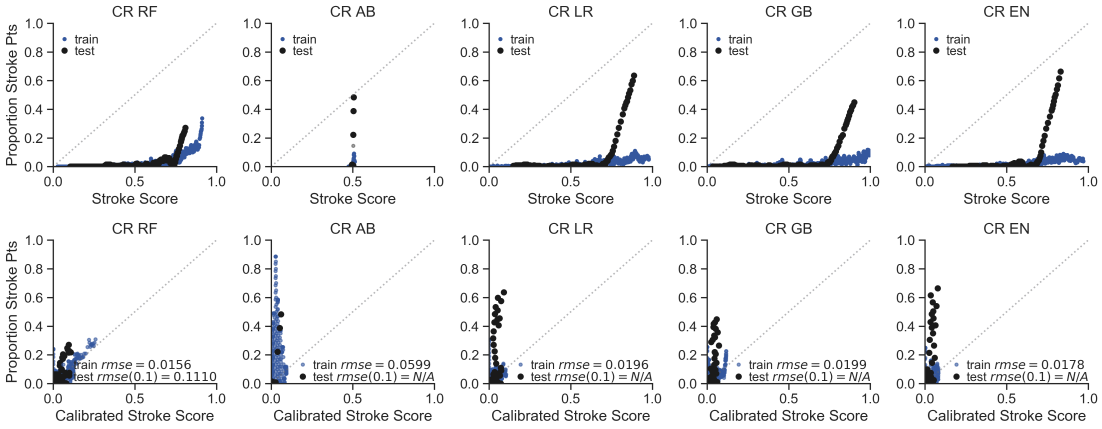


Figure 2.23: **Classifier type with Cases with ICD9 or ICD10 codes for CvD, Random patients in the EHR as controls (CR) case-control combination varies in calibration success between stroke score and actual proportion of patients at each probability.**

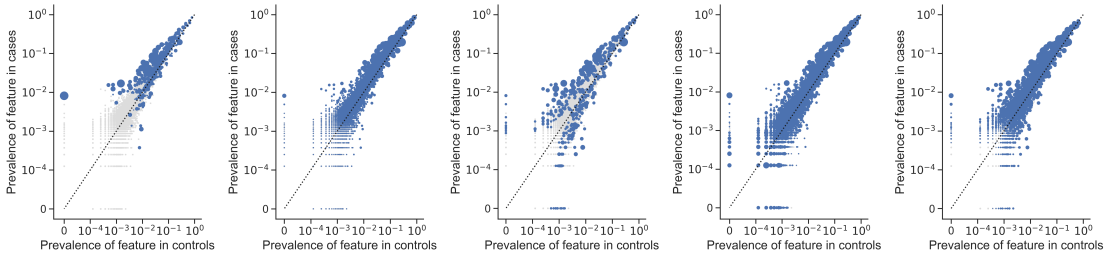


Figure 2.24: **Prevalence of features in cases vs controls in the TCI model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

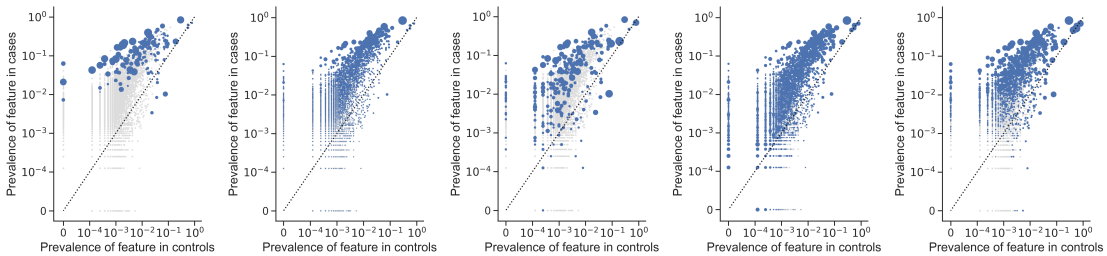


Figure 2.25: **Prevalence of features in cases vs controls in the TC model. Left to Right: LR, RF, AB, GB, and EN models.** Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

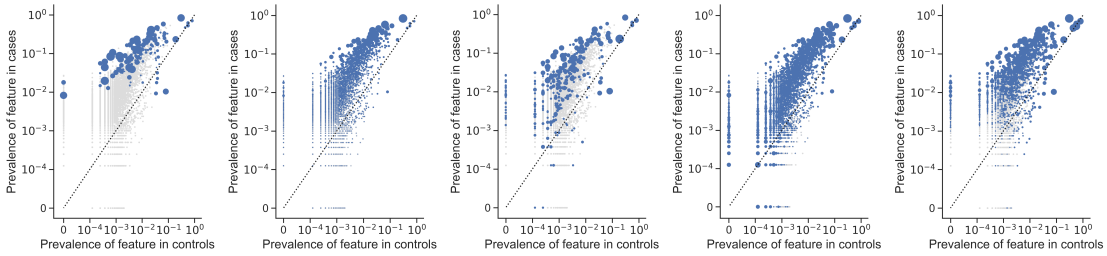


Figure 2.26: **Prevalence of features in cases vs controls in the TI model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance

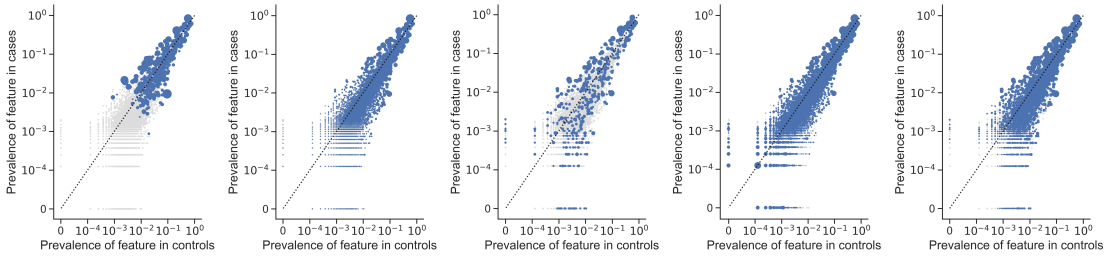


Figure 2.27: **Prevalence of features in cases vs controls in the TR model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

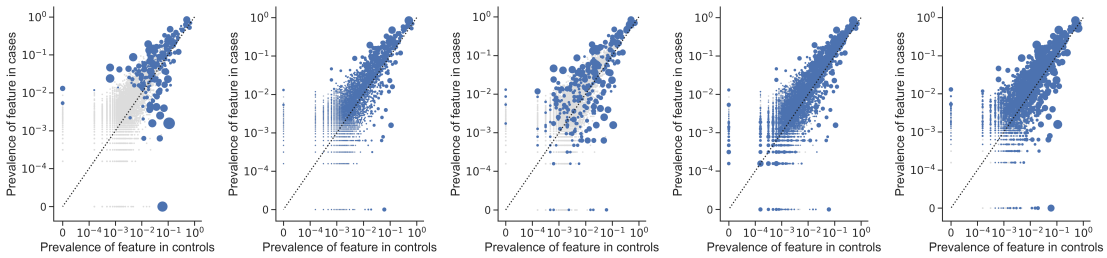


Figure 2.28: **Prevalence of features in cases vs controls in the TN model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

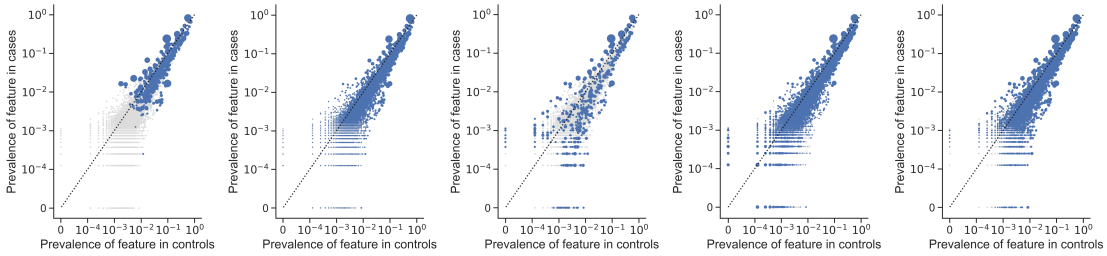


Figure 2.29: **Prevalence of features in cases vs controls in the CR model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

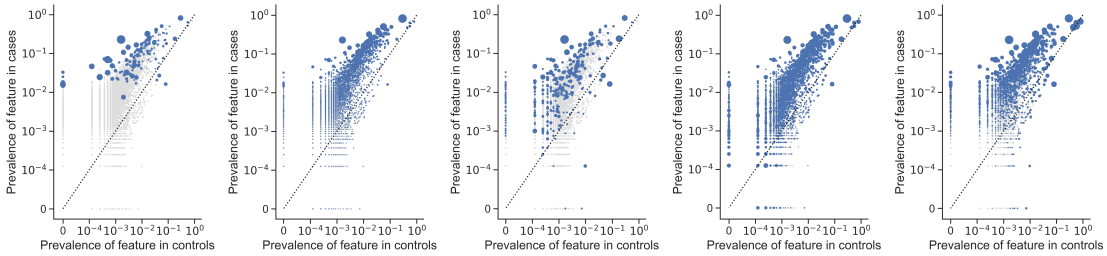


Figure 2.30: **Prevalence of features in cases vs controls in the CC model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

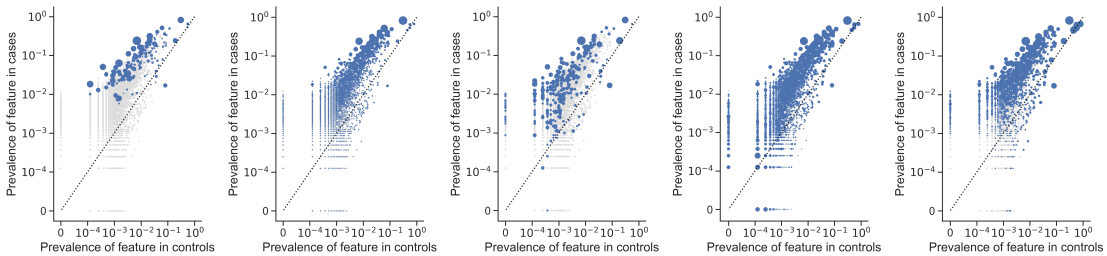


Figure 2.31: **Prevalence of features in cases vs controls in the CI model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

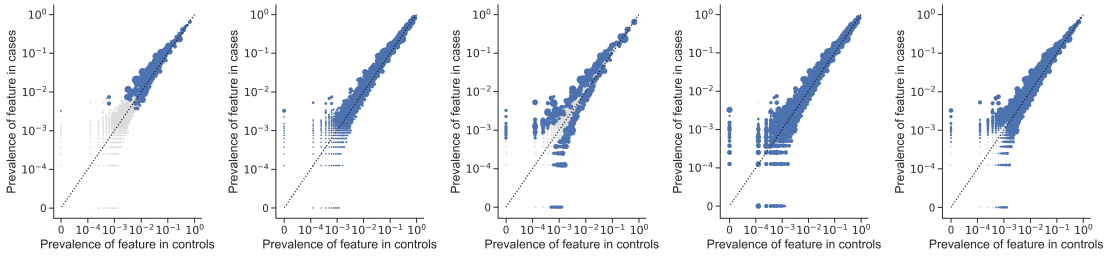


Figure 2.32: **Prevalence of features in cases vs controls in the CCI model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

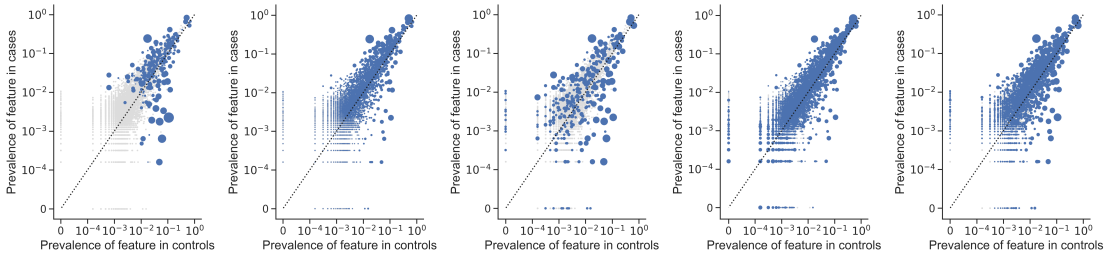


Figure 2.33: **Prevalence of features in cases vs controls in the CN model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

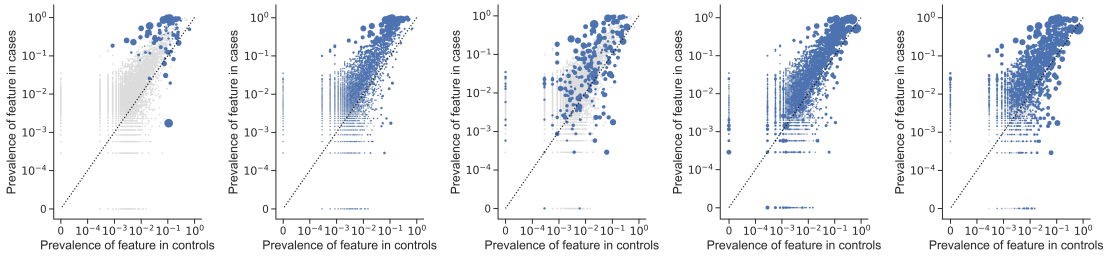


Figure 2.34: **Prevalence of features in cases vs controls in the SN model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

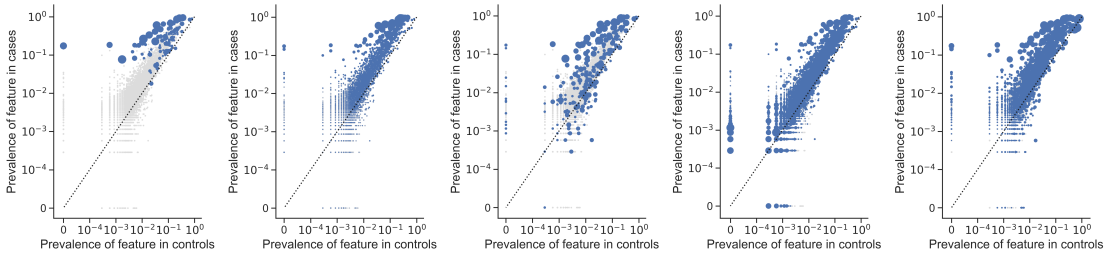


Figure 2.35: **Prevalence of features in cases vs controls in the SCI model.** Left to Right: LR, RF, AB, GB, and EN models.Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

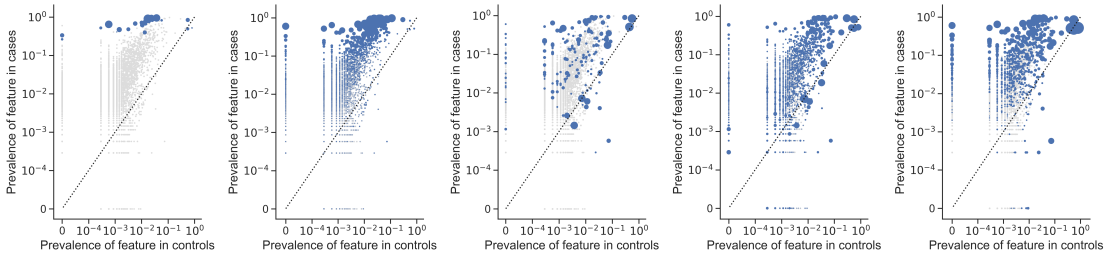


Figure 2.36: **Prevalence of features in cases vs controls in the SC model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

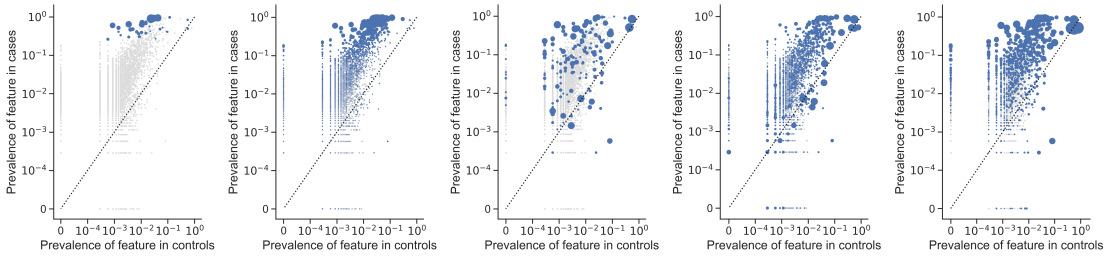


Figure 2.37: **Prevalence of features in cases vs controls in the SI model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

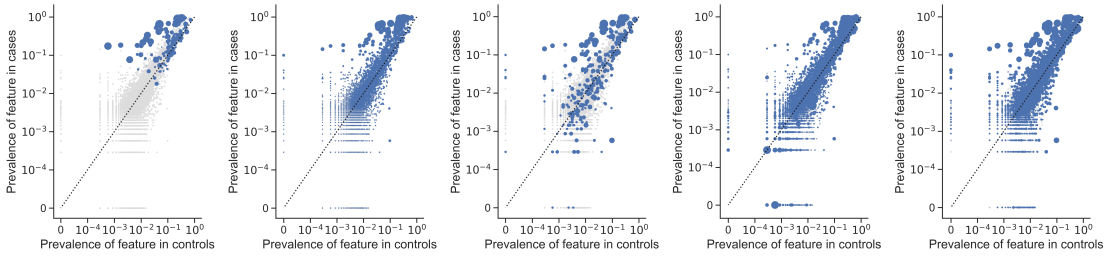


Figure 2.38: **Prevalence of features in cases vs controls in the SR model.** Left to Right: LR, RF, AB, GB, and EN models. Axes are on a logarithmic scale. Increasing size of blue dot correlates with higher feature importance or beta coefficient weight, depending on the classifier type. Gray dots are features with zero importance.

Chapter 3: Data-driven subtyping of acute ischemic stroke

3.1 Introduction

Although genome-wide association studies have uncovered many variants for a wide range of diseases, many still have a high proportion of missing heritability[17, 15, 113, 114]. This lost genetic signal may be due to phenotype heterogeneity[17]. In addition, the largest GWAS for stroke and its ischemic stroke subtypes found only thirty-two significant loci after carefully phenotyping each patient[11, 8, 9, 115]. This number of variants is much smaller than those found in other common diseases. This is partially due to the genetic heterogeneity of the disease across subtypes, which hinders the discovery of causative variants[9]. Accurate determination of the etiology of stroke is essential for genetic studies, in addition to risk stratification, optimal treatment, and prediction of outcomes. This can be difficult, however, because, as described in Chapter 1, there is a wide range of causes of stroke, with up to 40% labeled undetermined by traditional methods of subtyping[8, 6, 7, 4]. This highlights the need for a data-driven approach to subtyping in order to uncover the etiologies of strokes of unknown cause and patterns of distinct combinations of environmental and genetic risk factors leading to stroke. Traditionally, the Trial of Org 10172 Acute Stroke Treatment (TOAST) criteria is used to subtype ischemic stroke into the following categories: large artery atherosclerotic stroke, small artery lacunar stroke, cardioembolic stroke, stroke of other determined cause, and stroke of unknown cause[116]. These traditional criteria, however, have not been granular enough for treatment decisions. The ASCO (Atherosclerosis, Small vessel disease, Cardiac source, and Other) criteria provide percentages of evidence for each subtype without choosing the most probable cause[117]. The Causative Classification System (CCS1) for stroke provides a comprehensive algorithm combining medical and family history, demographics, imaging, and labs to determine the most likely cause of the stroke. It assigns patients multiple

pathogenic mechanisms based on their clinical picture and is currently found to be the most sensitive subtyping system[118, 119]. This classification method, however, is time consuming, since it requires a trained researcher to manually classify the patients[118]. We hypothesize that dividing ischemic stroke patients into more specific groups based on their clinical picture can provide data-driven insight into the causes of the disease, predict cohorts more likely to have recurrence or serious morbidity, and group people who are more likely to share similar gene variants.

Although the electronic health record (EHR) provides a rich amount of information about the medical history of each patient, identifying subtypes within the EHR is challenging because of the sparsity of the data, its ascertainment bias, and its optimization for insurance billing rather than research. Patients only interact with the EHR when sick or for a wellness check up, and which leads to a discontinuous and incomplete medical timeline[26, 44]. In addition, structured data is recorded primarily for medical communication and insurance billing purposes, so it is not optimized for research purposes[26, 72]. As seen in Table 3.1, the vast majority of cases are solely described as cerebral infarction due to unspecified cause. More information, such as risk factors, prior treatments, and procedures must be curated from the EHR to determine the subtype of stroke. In addition, optimal subtyping techniques need to take into account the sparsity and missingness of the data.

Factorization methods, such as non-negative matrix factorization (NMF), have been shown to identify interpretable latent topics, or subtypes, in text better than PCA or vector quantization[61]. NMF has successfully subtyped type 2 diabetes mellitus, hypertension and cardiovascular disease, but has not been applied to stroke[57, 58, 59, 60]. A limitation of NMF is that it can over-inflate the importance of missing data, not taking into account the false negative[65]. In addition, Bayesian factorization methods such as hierarchical Poisson factorization is well-suited for identifying subtypes of patients in the EHR. Similar to previous uses of the method in RNA-seq or click data, item use is sparse, there are some patients who are in the system much more frequently than others, and data are missing not-at random[65, 64]. We can apply these methods to identify groups, or topics, of risk factors before a stroke that may predict outcomes after the stroke. Deep learning methods

Concept name	ICD code	# patients
Cerebral artery occlusion, unspecified with cerebral infarction	434.91	4062
Cerebral infarction due to embolism of cerebral arteries	434.11	2524
Cerebral infarction, unspecified	I63.9	1185
Carotid artery obstruction	433.11	1113
Acute ill-defined cerebrovascular disease	I9:436	1076
Infarction - precerebral	I9:433.11	778
Cerebral infarction due to occlusion of precerebral artery	I63.50	297
Vertebral artery obstruction	433.21	125
Cerebral infarction due to embolism of middle cerebral artery	I63.411	125
Cerebral infarction due to thrombosis of middle cerebral artery	I63.312	113
Multiple and bilateral precerebral arterial occlusion	I9:433.31	68
Basilar artery occlusion	I9:433.01	56
Cerebrovascular disease	I67.89	33
Precerebral arterial occlusion	I9:433.81	15
Carotid artery occlusion	I10:I63.231	13
Cerebral infarct due to thrombosis of precerebral arteries	I10:I63.032	13
Cerebral infarction due to embolism of precerebral arteries	I63.10	10
No matching concept	I63.523	8
Cerebellar artery occlusion	I63.541	6
Carotid artery thrombosis	I10:I63.031	4
Cerebral infarction due to cerebral venous thrombosis, non-pyogenic	I63.6	4
Cerebral infarction due to occlusion of basilar artery	I10:I63.22	4
Carotid artery embolism	I10:I63.131	2
Basilar artery thrombosis	I10:I63.02	1
Basilar artery embolism	I10:I63.12	1
Vertebral artery thrombosis	I10:I63.011	1

Table 3.1: **Distribution of AIS subtypes using only AIS ICD9 or ICD10 codes**

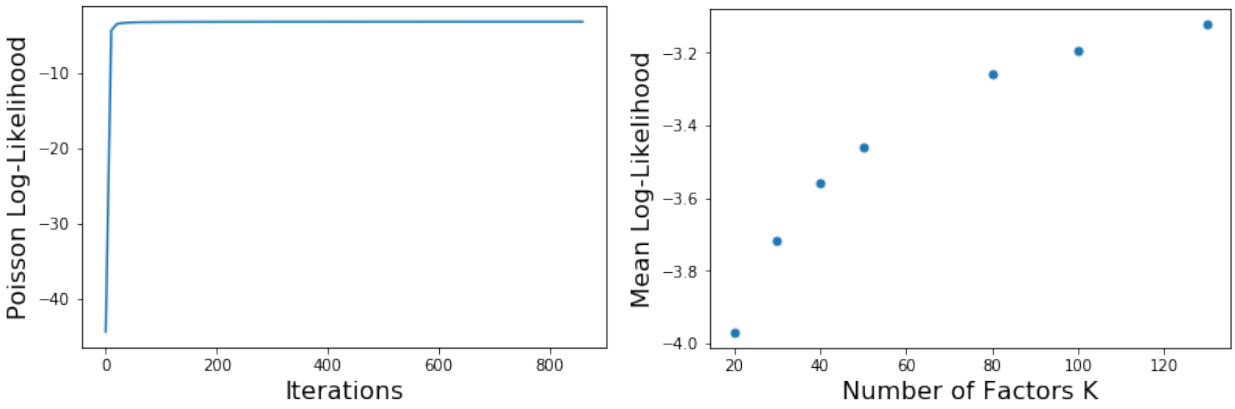
such as denoising autoencoders, as described in Chapter 1 also have been successful in subtyping disease, but interpretation and implementation can be difficult. In this study, we apply NMF and HPF to structured EHR data to identify topics of clinical features found in patients before their first acute ischemic stroke that are significant for predicting stroke severity.

3.2 Methods

Data extracted from the Columbia University Irving Medical Center (CUIMC) Clinical Data Warehouse (CDW) contained longitudinal health records of 6.4 million patients spanning from 1985-2018. We selected 4,386 patients with acute ischemic stroke (AIS) evaluated on the stroke

service at New York Presbyterian Hospital. We gathered diagnostic and procedure International Classification of Diseases (ICD) codes, medication prescriptions, race/ethnicity, age, and gender of each patient. AIS was defined by diagnosis codes within the Tirschwell-Longstreth criteria[24]. To study risk factors leading to the AIS, we only extracted data before the patient's first recorded stroke. We also manually extracted the National Institute of Health Stroke Severity Score from the unstructured medical notes of 488 of the patients seen on the stroke service for their AIS as a measure of severity. We used data structured in the OMOP CDM format in the EHR because this data is more easily accessible and reproducible at other hospitals, and requires little processing power compared to parsing notes from the EHR[36]. We ran non-negative matrix factorization of data from all 4,386 stroke patients before their first acute ischemic stroke. In order to reduce sparsity, we trained the model on features seen in at least 20 patients, which reduced the feature size from 14,618 to 2,264. In NMF, the patient by feature matrix (X) is factorized into a topic by feature matrix (H) and patient by topic (W) matrix. In H, each feature is assigned a weighting to every topic, and in W, every patient is assigned a weighting to every topic. We implemented the NMF module in the scikit-learn python package. We set the number of components, or topics, to 20, calculated loss by kullback-leibler distance, penalized the model with an even ratio of l1 and l2, and ran for 1000 iterations.

We then compared the non-negative matrix factorization method to a probabilistic method, hierarchical Poisson factorization (HPF). We implemented scHPF created by [65] and found at https://www.github.com/simslab/scHPF/tree/master. In our implementation, cells replaced patients, and structured medical data of the patients replaced genes. Patient scores for each factor and feature scores for each factor were calculated by sampling the product of variational distributions of latent variables from the model. We used the default settings for the hyperparameters. As seen in [65], the posterior probability of the model was inferred using Coordinate Ascent Variational Inference. Hyperparameters b′ and d′ were calculated empirically based on the ratio of mean to variance of items per patient or per medical feature, a′ and c′ were set to 1 and a and c were set to 0.3, and variational parameters ξ, β,θ, and η were sampled from gamma distributions whose rate



(a) Poisson Log-Likelihood vs. HPF run

(b) Mean Log-likelihood vs. Number of factors number

Figure 3.1: **(a) Poisson Log-Likelihood vs. Iteration for HPF shows convergence, k=100. (b) Mean Log-Likelihood vs. factor k size**

and shape were randomly initialized between 0.5 and 1.5 times its prior[65].

We ran the model for 7 different factor size k: 20, 30, 40, 50, 80, 100, and 130. Figure 3.1a shows a sample convergence of the model at k=100, and Figure 3.1b shows increased log-likelihood with increasing k size.

We ran the model 5 times with random initialization until the drop in loss was less than 0.001 percent, 1000 iterations were run, or the loss increased, whichever came first.

3.2.1 Determining stability of the topics

To determine the stability of the topics, we calculated the Tanimoto coefficient of the top 20 features between every topic between every NMF and HPF run (n=100 runs, 8 runs respectively).

3.2.2 Using topics to predict stroke severity

Because of the instability of the HPF topics, we only used NMF topics for severity analysis. For the 488 patients with stroke severity scores, we randomly divided the patients into 300 training cases and 188 testing cases. We then ran a univariate linear regression on each topic from NMF with stroke severity as the outcome. After correction by false discovery rate, we predicted the stroke severity of the test patient cases.

3.3 Results

3.3.1 Non-negative matrix factorization and hierarchical Poisson factorization topics

Each patient was assigned to the topic with their maximum W component found using NMF. Table 3.2 shows the top 10 scoring medical features for each factor. The maximum weight for 33% of patients was on a topic with general demographics such as white race, gender, age and lymphoma and colon cancer. The next largest proportion, 7.5% of patients, had maximum weight for the topic with essential hypertension, pure hypercholesterolemia, and osteoarthritis as its top three features, and 6.7% of patients had maximum weight for the topic with atrial fibrillation, essential hypertension, and aortic valve disorder as its top three features.

In comparison, as seen in Table 3.3, hierarchical Poisson factorization displayed more specialized topics, only covering 1-3 different disease processes and with no generalized hospitalization or procedure topics as seen using NMF. Topic 6 had the highest proportion of patient maximum scores, 26%, whose top scoring features included cancers and pregnancy. This group of patients were enriched for internal carotid artery dissection, acetaminophen, laxative, and vertebral artery dissection after stroke diagnosis. The next most common topic, 19, with 9% of patients included topics of joint disease and Parkinson's disease. These patients were enriched for procedures of the elderly including shingles vaccine administration, hearing aid fitting, ear wax removal, and physical therapy after stroke diagnosis. In general, for HPF, the enrichment of features in patients after stroke were similar to the disease processes of top scoring features of each factor.

Figure 3.2 shows hierarchical clustering maps of patient scores across all topics for NMF and HPF. As seen in the figure, NMF assigns smaller weights to most factors for each patient compared to the patient scores sampled from the HPF model. The HPF model appears to identify clearer clusters of patients compared to NMF.

Factor	Highest scoring before-stroke features	Topic Summary
1	Dizziness and giddiness,Abdominal pain,Hyperlipidemia,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low complexity,Type 2 diabetes mellitus,Chest pain,Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only,Osteoarthritis,Pure hypercholesterolemia,Essential hypertension	**Heart disease, hypertension, DM2, treatment and procedure**
2	Critical care, evaluation and management of the critically ill or critically injured patient; first 30-74 minutes,Past history of procedure,Subsequent hospital care, per day, for the evaluation and management of a patient, which requires at least 2 of these 3 key components: A detailed interval history; A detailed examination; Medical decision making of high complexity. Counseling and/or coor,Abnormal breath sounds,Pleural effusion,Cardiomyopathy,Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only,Radiologic examination, chest; single view, frontal,Subsequent hospital care, per day, for the evaluation and management of a patient, which requires at least 2 of these 3 key components: An expanded problem focused interval history; An expanded problem focused examination; Medical decision making of moderate complexity. Congestive heart failure	**Heart failure, critical care, pulmonary issues**
3	Acute renal failure syndrome,Chronic kidney disease,End stage renal disease,Anemia of chronic renal failure,Subsequent hospital care, per day, for the evaluation and management of a patient, which requires at least 2 of these 3 key components: An expanded problem focused interval history; An expanded problem focused examination; Medical decision making of moder,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o,Benign essential hypertension,Disorder of transplanted kidney,History of renal transplant,Long-term drug therapy	**Kidney disease**
4	High risk drug monitoring status,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low,Metoprolol,Atrial flutter,Subsequent stage of staged operation,Warfarin,Mitral valve disorder,Essential hypertension,Aortic valve disorder,Atrial fibrillation	**Arrythmia, valve disorder, related procedure and treatment**
5	pneumococcal capsular polysaccharide type 12F vaccine,pneumococcal capsular polysaccharide type 5 vaccine,pneumococcal capsular polysaccharide type 15B vaccine,Radiologic examination, chest, 2 views, frontal and lateral,Docusate,Sodium Chloride,Radiologic examination, chest; single view, frontal,Acetaminophen,Subsequent hospital care, per day, for the evaluation and management of a patient, which requires at least 2 of these 3 key components: An expanded problem focused interval history; An expanded problem focused examination; Medical decision making of moderate complexity. Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only	**Hospitalization, ECG, Vaccine**
6	Renal osteodystrophy,Chronic kidney disease stage 5,Chronic kidney disease stage 4,Hemodialysis procedure with single evaluation by a physician or other qualified health care professional,Dialysis procedure,Hypertensive renal disease with renal failure,Type 2 diabetes mellitus,Chronic kidney disease,Anemia of chronic renal failure,End stage renal disease	**End stage renal disease and treatment, T2DM**
7	Protein; electrophoretic fractionation and quantitation, serum,Spinal stenosis of lumbar region,Aspirin,Atrial fibrillation,Essential hypertension,Anemia,Current tobacco non-user (CAD, CAP, COPD, PV) (DM) (IBD),Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o,Multiple myeloma,Eligible clinician attests to documenting in the medical record they obtained, updated, or reviewed the patient's current medications	**multiple myeloma, anemia, arrhythmia, hypertension**
8	Dysthymia,Group psychotherapy (other than of a multiple-family group),Gynecologic examination,Recurrent major depressive episodes,Finding relating to psychosocial functioning,Interview and evaluation, described as limited,Headache,Anxiety disorder,Depressive disorder,Human immunodeficiency virus infection	**psychiatric disorders, HIV**
9	Disorder due to type 2 diabetes mellitus,Glucose,Glucagon,Insulin Glargine,Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only,Insulin, Aspart, Human,Metformin,Type 1 diabetes mellitus,Type 2 diabetes mellitus,Type 2 diabetes mellitus	**Diabetes mellitus, type 1 and 2**
10	Chest pain,Radiologic examination, chest, 2 views, frontal and lateral,Dyspnea,Chronic obstructive lung disease,Injection or infusion of other therapeutic or prophylactic substance,Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only,Emergency department visit for the evaluation and management of a patient, which requires these 3 key components: An expanded problem focused history; An expanded problem focused examination; and Medical decision making of moderate complexity. Counseling ,Chest pain,Emergency department visit for the evaluation and management of a patient, which requires these 3 key components: A detailed history; A detailed examination; and Medical decision making of moderate complexity. Counseling and/or coordination of care with o,Asthma	**chest pain, asthma, ECG**
11	White,Adult health examination,Primary malignant neoplasm of colon,F,M,Unknown,Malignant lymphoma,Primary malignant neoplasm of respiratory tract,Adult health examination,age	**General demographics and cancer**
12	No matching concept,Therapeutic procedure, 1 or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility,Closed fracture of distal end of radius,age,Hispanic,Secondary hypertension,Ophthalmological services: medical examination and evaluation, with initiation or continuation of diagnostic and treatment program; Intermediate, established patient,Group psychotherapy (other than of a multiple-family group),Nuclear senile cataract,Recurrent major depressive episodes	**depression, ophthalmological service, hispanic, bone fracture**
13	Depressive disorder,Metformin,Diabetic oculopathy,Renewal of prescription,Screening for malignant neoplasm of breast,Interview and evaluation, described as limited,Psoriasis,Pure hypercholesterolemia,Type 1 diabetes mellitus,Type 2 diabetes mellitus	**Depression, diabetes, breast cancer, psoriasis, osteoarthritis**
14	Coronary arteriosclerosis,Chronic ischemic heart disease,Congestive heart failure,Left heart failure,Type 2 diabetes mellitus,H/O: heart recipient,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low,Hyperlipidemia,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o,Essential hypertension	**hypertension, hyperlipidemia, diabetes mellitus 2, heart failure**
15	Neoplasm of uncertain behavior of skin,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o,Actinic keratosis,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A problem focused history; A problem focused examination; Straightforward medical decision making. Counselin,Low back pain,Arthralgia of the lower leg,Level IV - Surgical pathology, gross and microscopic examination Abortion - spontaneous/missed Artery, biopsy Bone marrow, biopsy Bone exostosis Brain/meninges, other than for tumor resection Breast, biopsy, not requiring microscopic evaluation of surgica,Arthropathy of knee joint,Primary malignant neoplasm of prostate,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low complexity	**Joint and bone disease, hyperlipidemia**
16	clopidogrel,Coronary bypass graft finding,Preinfarction syndrome,Aspirin,Chronic ischemic heart disease,Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only,Hyperlipidemia,Chest pain,Coronary arteriosclerosis,Coronary arteriosclerosis in native artery	**Biopsy, skin disease, cancer**
17	Acute respiratory failure,Attention to catheter,Pneumonia,Subsequent hospital care, per day, for the evaluation and management of a patient, which requires at least 2 of these 3 key components: A detailed interval history; A detailed examination; Medical decision making of high complexity. Counseling and/or coor,Atelectasis,Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only,Abnormal breath sounds,Pleural effusion,Critical care, evaluation and management of the critically ill or critically injured patient; first 30-74 minutes,Radiologic examination, chest; single view, frontal	**Acute pulmonary issue, Pneumonia**
18	Subsequent stage of staged operation,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low,Acute pancreatitis,Endocarditis,Obstruction of bile duct,Chronic pancreatitis,Abdominal pain,Chronic hepatitis C,Cirrhosis - non-alcoholic,Diseases of mitral and aortic valves	**Heart disease, kidney disease, pneumonia**
19	Osteoarthritis,Osteoporosis,Hypothyroidism,Gynecologic examination,Gynecologic examination,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: An expanded problem focused history; An expanded problem focused examination; Medical decision making of low,Low back pain,Breast lump,Screening for malignant neoplasm of breast,Primary malignant neoplasm of female breast	**GU evaluation, breast cancer, heart valve disease, pancreatitis**
20	Eligible clinician attests to documenting in the medical record they obtained, updated, or reviewed the patient's current medications,Rheumatoid arthritis,Respiratory symptom,Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report,Post-inflammatory pulmonary fibrosis,Chronic pulmonary heart disease,Heritable pulmonary arterial hypertension,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A comprehensive history; A comprehensive examination; Medical decision making of high complexity. Counseling,Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A detailed history; A detailed examination; Medical decision making of moderate complexity. Counseling and/o,Dyspnea	**Pulmonary embolism, pulmonary heart disease, medical exam, karposi's sarcoma**

Table 3.2: **Top ten features found from structured medical data from 4,368 AIS patients before their first recorded stroke using Non-negative Matrix Factorization, number of components=20**

Factor	Highest scoring before-stroke features	Topic Summary
1	'Status_asthmaticus', 'IgE-mediated_allergic_asthma', 'Acute_exacerbation_of_chronic_obstructive_airways_disease', 'Acute_exacerbation_of_chronic_asthmatic_bronchitis', 'Alcohol_withdrawal_syndrome', 'Pressurized_or_nonpressurized_inhalation_treatment_for_acute_airway_obstruction_for_therapeutic_purposes_and/ or_for_diagnostic_purposes_such_as_sputum_induction_with_an_aerosol_generator,_nebulizer,_metered_dose_inhaler_or_intermittent_positive_pressure_b', 'Respiratory_medication_administered_by_nebulizer', 'Alcohol_abuse', 'Exacerbation_of_asthma', 'Asthma'	**Airway disease and treatment, alcohol abuse**
2	'Obstructive_hydrocephalus', 'Complete_transposition_of_great_vessels', 'Tetralogy_of_Fallot', 'Primary_malignant_neoplasm_of_thyroid_gland', 'Epilepsy', 'Primary_malignant_neoplasm_of_prostate', 'Subarachnoid_hemorrhage', 'Primary_malignant_neoplasm_of_kidney', 'Cerebral_hemorrhage', "Parkinson's_disease"	**Cancer, brain hemorrhage, congenital abnormality, parkinson's disease**
3	'Hemoglobin_SS_disease_without_crisis', 'Black', 'Unknown', 'M', 'Hb_SS_disease', 'Primary_malignant_neoplasm_of_rectosigmoid_junction', 'Primary_malignant_neoplasm_of_sigmoid_colon', 'Primary_malignant_neoplasm_of_rectum', 'Amyloidosis', 'Primary_malignant_neoplasm_of_colon'	**Sickle cell disease, cancer, male, black or unknown race**
4	'pneumococcal_capsular_polysaccharide_type_6B_vaccine', 'pneumococcal_capsular_polysaccharide_type_1_vaccine', 'pneumococcal_capsular_polysaccharide_type_10A_vaccine', 'pneumococcal_capsular_polysaccharide_type_23F_vaccine', 'pneumococcal_capsular_polysaccharide_type_7F_vaccine', 'pneumococcal_capsular_polysaccharide_type_8_vaccine', 'pneumococcal_capsular_polysaccharide_type_2_vaccine', 'pneumococcal_capsular_polysaccharide_type_19A_vaccine', 'pneumococcal_capsular_polysaccharide_type_19F_vaccine', 'pneumococcal_capsular_polysaccharide_type_9V_vaccine'	**Flu Vaccine**
5	'H/O:_artificial_joint', 'Acquired_spondylolisthesis', 'Closed_fracture_of_neck_of_femur', 'Strain_of_rotator_cuff_capsule', 'Spinal_stenosis', 'Closed_fracture_of_upper_end_of_humerus', 'Closed_fracture_of_distal_end_of_radius', 'Actinic_keratosis', 'Spinal_stenosis_of_lumbar_region', 'Rheumatoid_arthritis'	**Joint disease and skin disease**
6	'Single_live_birth', 'Prolonged_depressive_adjustment_reaction', 'Other_group_therapy', 'Caretaking/parenting_skills_education,_guidance,_and_counseling', 'Depressed_bipolar_I_disorder', 'Patient_currently_pregnant', 'Personality_disorder', 'Borderline_personality_disorder', 'Group_psychotherapy_(other_than_of_a_multiple-family_group)', 'Recurrent_major_depressive_episodes'	**Pregnancy and psychiatric disorders**
7	'Malignant_neoplasm_of_uterus', 'Gangrenous_disorder', 'Ulcer_of_foot', 'Skin_ulcer_of_calf', 'Ankle_ulcer', 'Peripheral_circulatory_disorder_associated_with_type_2_diabetes_mellitus', 'Atherosclerosis_of_native_arteries_of_the_extremities', 'Atherosclerosis_of_native_arteries_of_the_extremities', 'Ulcer_of_foot', 'Atherosclerosis_of_native_arteries_of_the_extremities'	**Uterine cancer and peripheral vascular disease**
8	'Valvular_endocarditis', 'Aneurysm_of_thoracic_aorta', 'Chronic_pulmonary_edema', 'Acute_and_subacute_bacterial_endocarditis', 'Atelectasis', 'Cardiogenic_shock', 'Heritable_pulmonary_arterial_hypertension', 'Pleural_effusion', 'Psoriasis', 'Aortic_valve_disorder'	**Acute and chronic heart and pulmonary disease**
9	'No_matching_concept', 'Malignant_neoplasm_of_upper_lobe,_bronchus_or_lung', 'Injection_or_infusion_of_cancer_chemotherapeutic_substance', 'Adverse_effect,_due_to_correct_medicinal_substance_properly_administered', 'Adverse_reaction_to_drug', "Kaposi's_sarcoma", 'Multiple_myeloma', 'Malignant_lymphoma', 'Primary_malignant_neoplasm_of_respiratory_tract', 'Human_immunodeficiency_virus_infection'	**Cancer, adverse drug effects, HIV infection**
10	'Ophthalmological_services:_medical_examination_and_evaluation,_with_initiation_or_continuation_of_diagnostic_and_treatment_program;_intermediate,_established_patient', 'Fibrocystic_disease_of_breast', 'Lateral_epicondylitis', 'Chronic_angle-closure_glaucoma', 'Vaccination_required', 'Visual_field_examination,_unilateral_or_bilateral,_with_interpretation_and_report;_extended_examination_(eg,_Goldmann_visual_fields_with_at_least_3_isopters_plotted_and_static_determination_within_the_central_30,_or_quantitative,_automated_threshold_perim', 'Nonexudative_age-related_macular_degeneration', 'Exudative_age-related_macular_degeneration','Ophthalmological_services:_medical_examination_and_evaluation,_with_initiation_or_continuation_of_diagnostic_and_treatment_program;_comprehensive,_established_patient,_1_or_more_visits', 'Nuclear_senile_cataract'	**Eye diseases and procedures and breast disease**
11	'Complication_of_transplanted_liver', 'Cholangitis', 'Portal_hypertension', 'Chronic_viral_hepatitis_B_without_delta-agent', 'Malignant_neoplasm_of_liver', 'Chronic_pancreatitis', 'Obstruction_of_bile_duct', 'Chronic_hepatitis_C', 'H/O:_liver_recipient', 'Cirrhosis_-_non-alcoholic'	**Chronic liver disease**
12	'Angina_pectoris', '92980', 'Acute_myocardial_infarction', 'Insertion_of_drug-eluting_coronary_artery_stent(s)', 'Acute_subendocardial_infarction', 'Coronary_bypass_graft_finding', 'Percutaneous_transluminal_coronary_angioplasty_[PTCA]', 'Left_heart_cardiac_catheterization', 'Coronary_arteriosclerosis_in_native_artery', 'Patient_post_percutaneous_transluminal_coronary_angioplasty'	**Cardiac arrest and treatment**
13	'Septic_shock', 'Pneumothorax', 'Postoperative_shock', 'Feeding_difficulties_and_mismanagement', 'Gammopathy', 'Complication_of_transplanted_lung', 'Trauma_and_postoperative_pulmonary_insufficiency', 'Postoperative_shock', 'H/O:_lung_recipient', 'Acute_respiratory_failure'	**Shock, acute lung disease**
14	'History_of_clinical_finding_in_subject', 'Primary_malignant_neoplasm_of_central_portion_of_female_breast', 'Malignant_neoplasm_of_upper-inner_quadrant_of_female_breast', 'Overlapping_malignant_neoplasm_of_female_breast', 'Congenital_malposition_of_heart', 'Malignant_neoplasm_of_upper-outer_quadrant_of_female_breast', 'Carcinoma_in_situ_of_breast', 'Atrial_fibrillation', 'Diseases_of_mitral_and_aortic_valves', 'Primary_malignant_neoplasm_of_female_breast'	**brest cancer, atrial fibrillation, valve disease**
15	'Blood_pressure_has_a_systolic_value_of_<_140_and_a_diastolic_value_of_<_90', 'High_risk_drug_monitoring_status', 'Chronic_idiopathic_pulmonary_fibrosis', 'Idiopathic_pulmonary_fibrosis', 'Bmi_is_documented_above_normal_parameters_and_a_follow-up_plan_is_documented', 'Type_2_diabetes_mellitus_without_complication', 'Past_history_of_procedure', 'Essential_hypertension', 'Hyperlipidemia', 'Prothrombin_time'	**Hypertension,**
16	'93734', 'Initial_insertion_of_transvenous_leads_[electrodes]_into_atrium_and_ventricle', 'Initial_insertion_of_dual-chamber_device', 'Malignant_tumor_of_ascending_colon', 'Polymyalgia_rheumatica', 'Programming_device_evaluation_(in_person)_with_iterative_adjustment_of_the_implantable_device_to_test_the_function_of_the_device_and_select_optimal_permanent_program med_values_with_analysis,_review_and_report_by_a_physician_or_other_qualified_health_care_', 'Atrioventricular_block', 'Cardiac_pacemaker_in_situ', '93731', 'Diverticular_disease_of_colon'	**Intestinal and Pancreatic disease**
17	'Hemoglobin;_glycosylated_(A1C)', 'Diabetes_outpatient_self-management_training_services,_group_session_(2_or_more),_per_30_minutes', 'rosiglitazone', 'Type_2_diabetes_mellitus', 'Diabetic_oculopathy', 'Onychomycosis_due_to_dermatophyte', 'Disorder_due_to_type_2_diabetes_mellitus', 'Disorder_due_to_type_2_diabetes_mellitus', 'Type_1_diabetes_mellitus', 'Type_2_diabetes_mellitus'	**Acute and Chronic Pulmonary disease**
18	'Past_history_of_procedure', 'Hyperparathyroidism_due_to_renal_insufficiency', 'Chronic_kidney_disease_stage_5', 'Renal_osteodystrophy', 'Dialysis_procedure', 'Hemodialysis_procedure_with_single_evaluation_by_a_physician_or_other_qualified_health_care_professional', 'Disorder_of_transplanted_kidney', 'Anemia_of_chronic_renal_failure', 'End_stage_renal_disease', 'History_of_renal_transplant'	**Peripheral vascular disease**
19	'Microscopic_hematuria', 'Retention_of_urine', 'Benign_prostatic_hypertrophy_with_outflow_obstruction', 'Complex_uroflowmetry_(eg,_calibrated_electronic_equipment)', 'Measurement_of_post-voiding_residual_urine_and/or_bladder_capacity_by_ultrasound,_non-imaging', 'Raised_prostate_specific_antigen', 'Paraplegia', 'Nocturia', 'Malignant_tumor_of_urinary_bladder', 'Incomplete_emptying_of_bladder'	**End stage renal disease and treatment**
20	'Biopsy_of_heart', 'Programming_device_evaluation_(in_person)_with_iterative_adjustment_of_the_implantable_device_to_test_the_function_of_the_device_and_select_optimal_permanent_program med_values_with_analysis,_review_and_report_by_a_physician_or_other_qualified_health_care_', 'Acute_on_chronic_systolic_heart_failure', 'Endomyocardial_biopsy', 'Cardiac_transplant_disorder', 'Automatic_implantable_cardiac_defibrillator_in_situ', 'Paroxysmal_ventricular_tachycardia', 'Past_history_of_procedure', 'H/O:_heart_recipient', 'Cardiomyopathy'	**Joint disease and parkinson's disease**

Table 3.3: **Top ten features found from structured medical data from 4,368 AIS patients before their first recorded stroke using hierarchical Poisson factorization, number of components=20**

3.3.2 NMF topics are stable

The NMF topics were more stable than HPF topics. Out of the expected NMF topic matches, 98% had a Tanimoto coefficient of 1. Out of the expected HPF topic matches, 24% had a Tanimoto coefficient above 0.5.

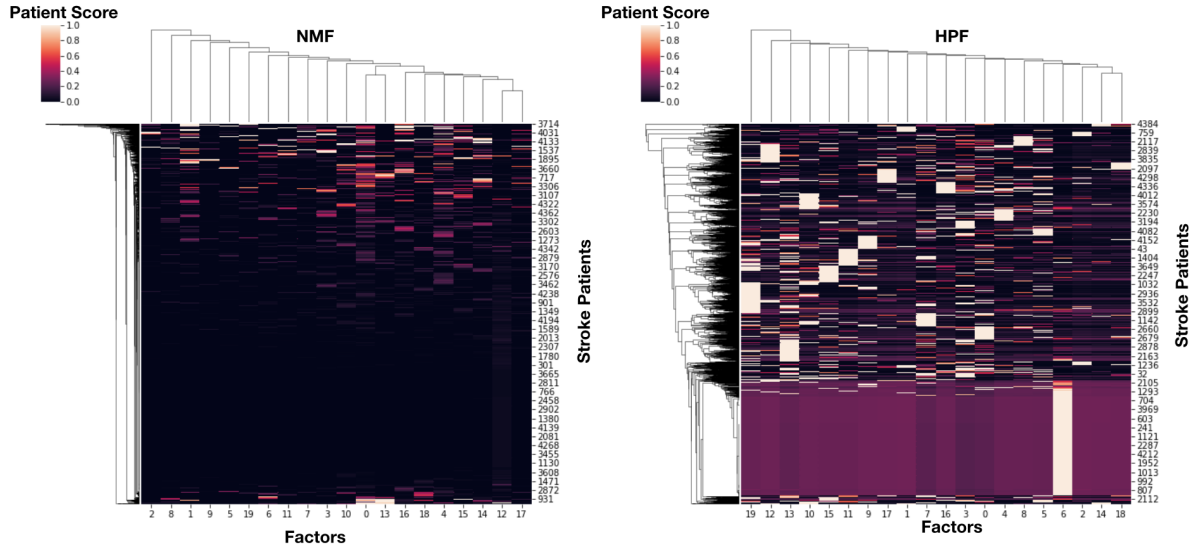


Figure 3.2: **Hierarchical Clustering of Patient Scores for each factor, K=20, Left= NMF, Right=HPF**

3.3.3 Topics were significantly correlated with stroke severity

After running a univariate linear regression for each topic on different randomized training samples of 300 patients with stroke severity scores, we found four topics significantly correlated with stroke severity in at least one run. Table 3.4 shows the performance of the four topics on predicting stroke severity in the test set. The average stroke severity score in the test set was 8.85 ± 8.71. The topic with atrial fibrillation as its highest weighted feature was significantly correlated with severity for 35% of the trials with the lowest p-value of 2.3E-05, and its top features included valve disorders, essential hypertension, warfarin, atrial flutter, metoprolol, and high risk drug monitoring status. The patients highest weighted for this topic had the highest average NIHSS score of 12.1 ± 10.9. The topic with type 2 diabetes mellitus as its highest weighted feature was significantly correlated with severity in 5% of the trials with the lowest p-value of 0.0012, and its top features included type 1 diabetes mellitus, ECG and medications for diabetes. The patients highest weighted for this topic had an average NIHSS score of 8.53 ± 7.10. The topic with a chest x-ray as its highest weighted feature was significantly correlated with severity in 2% of the trials with the lowest p-value of 0.0088, and its top features included critically ill patient, pleural

effusion, abnormal breath sounds, ECG, atelectasis, medical decision making of high complexity, pneumonia, and acute respiratory failure. The patients highest weighted for this topic had the second highest average NIHSS score of 11.4 ± 7.87. The final significantly correlated topic with AIS, in 2% of the trials and with a lowest p-value 0.043, had age as its highest weighted feature, and its top features included adult health exam, various cancer diagnoses, unknown and white race, and gender.

	Atrial Fibrillation	**Diabetes Mellitus**	**General demographics and cancer**	**Critical Respiratory Failure**
β coefficient	2.3 ± 1.3	2.0 ± 0.74	2.6 ± 0.3	3.0 ± 0.73
FDR-corrected p-value(min, max)	(2.3e-5, 0.048)	(0.0012, 0.037)	(0.043, 0.048)	(0.0088, 0.042)
Pct. significant trials	35%	5%	2%	2%
Mean absolute error	6.6 ± 0.21	6.7 ± 0.14	7.1 ± 0.21	7.1 ± 0.18
Root mean squared error	8.2 ± 0.32	8.3 ± 0.20	9.3 ± 0.29	8.9 ± 0.26
Mean NIHSS	12 ± 10	9.8 ± 10	8.5 ± 7.7	11 ± 7.9

Table 3.4: **Topics derived from Non-negative matrix factorization significantly correlated with stroke severity**

3.4 Discussion

Subtyping using factorization methods produced topics enriched for important risk factors before the patients' first acute ischemic stroke. Non-negative matrix factorization topics were more stable compared to HPF topics, and several were significantly correlated with stroke severity. The topic with the most significant runs also had patients with the highest mean severity score. The top weighted feature in this topic was atrial fibrillation, which is associated with more severe strokes[120]. The mean absolute error of the severity predictions was between 6.6-7.2, which is

smaller than the difference between mild and severe strokes on the NIHSS scale (10-30 points). This topic also represents a traditional TOAST subtype of acute ischemic stroke, cardio-embolic stroke. Other topics that correlated with severity represented more granular etiologies of stroke than those found using the TOAST subtype. One included a top weighted feature of diabetes mellitus 2, a known risk factor for stroke[5]. In addition, a third topic included highly weighted features such as acute respiratory failure and pneumonia. Respiratory infections such as influenza have been shown to trigger cardiovascular events such as myocardial infarction[121] and stroke[122, 5].

3.4.1 Limitations

This study had several limitations. We did not complete hyperparameter tuning for the HPF model, which could be a reason why the HPF model topics had low stability across runs. Each HPF run had a long run time, making the NMF model more feasible. In addition, at up to 120 topics, the HPF mean log-likelihood had not plateaued, suggesting that more topics are needed for greater stability. Such a large number of topics however, does not translate to feasible subtypes for acute ischemic stroke. In addition, although the NMF topics were stable, the most common topic was a general demographics and cancer topic. Although the general demographic features of the topic were non-specific, cancer is a risk factor for stroke[123]. In addition, other topics that did correlate with severity, such as patients with acute respiratory distress and pneumonia and those with atrial fibrillation represent important AIS patient subsets. Finally, we did not determine the genetic heterogenity of the subjects within the subtypes, which would require a larger sample size and future study.

3.5 Conclusions

This study implemented non-negative matrix factorization (NMF) and hierarchical Poisson factorization (HPF) to identify groups of risk factors within patients that predict acute ischemic stroke severity. Both NMF and HPF identified subgroups of patients with specific disease processes, though NMF topics were preserved across runs more stably than HPF topics. Factorization

approaches to analyze data available in the EHR can identify previously unrecognized ischemic stroke subtypes, unbiased by clinical preconceptions of etiology. Such subtypes correlate with stroke severity and include conventionally recognized clusters, such as stroke associated with atrial fibrillation and other cardiac diseases, but also novel subtype clusters, such as stroke associated with respiratory failure and pneumonia. Future directions include improving the stability of topics within the HPF models, studying other neuro-psychological sequelae of stroke, and incorporating unstructured notes into this model to improve the richness of the data. A semi-supervised approach of projecting factors from the models on physician-subtyped patients may provide a link between the data-derived factors and clinically meaningful AIS subtypes such as those defined by TOAST and ASCO[63, 116, 117]. Finally, we would like to test whether these more granular topics separate stroke patients into less genetically heterogeneous subtypes by estimation of heritability of each subtype using RIFTEHR and solarSTRAP (discussed in Chapters 4 and 5).

3.6 Acknowledgements

I would like to thank Dr. Mitchell Elkind and Dr. Benjamin Kummer for providing the stroke service patients subtyped in this study. I would also like to thank Dr. Mitchell Elkind for helpful comments in this study.

Chapter 4: Expansion of Case/Control cohorts by application of machine learning models to the EHR: Applications to heritability and genetics within Columbia and the eMERGE dataset

4.1 Introduction

Large genetic repositories connected to electronic health records (EHR) form the basis of local and national precision medicine initiatives. These linked repositories promise the ability to perform thousands of simultaneous genetic studies using diagnoses, procedures, and treatment responses that are routinely captured by the EHR. Not all patients have genetic data available, however, limiting what actually can be studied. There is opportunity to use machine learning methods to derive quantitative proxy variables and expand the sample population. We hypothesized that the output of a supervised machine learning classifier can be used as a proxy variable for disease classification and be an efficient strategy for expanding a study cohort. We tested this hypothesis in two genetics studies that would benefit from cohort expansion: heritability estimation and a genome-wide association study.

Heritability, often measured through family studies, provides a quantitative measure for the genetic component of a disease[15]. Conducting family or genome-wide association studies for every disease is a time-consuming and expensive process; therefore, a systematic manner to identify diseases with a genetic component is needed. Polubriaginof et al. developed a high- throughput corollary to the genetic estimate of heritability, termed observational heritability[51]. Without genetic data, they suggested which diseases are heritable and should be studied genetically. In contrast to traditional estimation of heritability, observational heritability is measured using families derived from the EHR. These families were extracted using an algorithm, RIFTEHR, which uses emergency contact information to infer familial relationships between patients at the hospi-

tal[51]. They found and inferred relationships within 223,307 families and estimated heritability using a bootstrapped version of SOLAR (SOLARstrap), a program that uses identity by descent calculations to estimate trait heritability from pedigrees[51, 124]. They were able to estimate the heritability of over 500 diseases, though they attempted to estimate 2500 disease heritabilities. Since the acquisition of disease case/control status was from patient electronic health records, this heritability analysis is not structured like a tradition genetic study, in which there is a proband and search for family members with or without the disease[51, 124]. Only family members that interact with the hospital system have a disease status, which leads to an ascertainment bias and a reduced cohort size[51, 124, 26]. Phenotyping may also have affected the success of disease heritability estimation. Each case was phenotyped by a single ICD9 code for the disease, and each control did not have the ICD9 code for the disease or any disease code within the same Clinical Classification System category[98]. Acute ischemic stroke heritability, for example, was not able to be estimated by RIFTEHR and SOLARstrap. As demonstrated in Chapter 2, single diagnosis codes for disease may not be enough to identify all cases, and it is not clear whether patients without a stroke diagnosis can be considered a control. Similarly, it is difficult to determine whether family members with only cerebrovascular disease or metabolic disease should be considered controls since there is significant enrichment of genetic variants from coronary artery disease, hypertension, and atrial fibrillation in stroke patients[125]. Therefore, we hypothesized that heritability estimates could benefit from assigning patients on a spectrum of disease from 0 to 1. Specifically, instead of treating stroke as a dichotomous trait, we converted it to a quantitative one– the probability that a patient had an acute stroke.

Genome-wide association studies move beyond the family linked studies to identify common variants affecting complex traits. Ischemic stroke, in particular, has a high heritability estimate, 37.9% ± 5.2%, and the most recent GWAS identified over 20 associated variants[10, 11]. To achieve adequate power, this study required over 40,000 cases and 520,000 participants overall. These cohort sizes, with genetic data, are generally only achievable through large meta-analyses. We hypothesized that our cohort expansion method could improve the power of genome-wide as-

sociation studies with a fraction of those cases. We tested this hypothesis on a disease with a validated phenotyping algorithm applied to a small number of cases within a large dataset. The Electronic Medical Records and Genomics (eMERGE) consortium developed a database of rule-based algorithms, PheKB, to identify cases and controls of dozens of phenotypes [34, 126]. The algorithms are designed by physicians and bioinformaticians and validated at one or more major medical centers with a positive predictive value of at least 90%. The algorithms vary in computational complexity, from extracting diagnosis and procedure billing codes and lab values to applying natural language processing to medical notes. In addition, the phenotypes are validated within the eMERGE consortium, which consists of subjects across nine study sites in the US. As a proof of concept that the conversion of dichotomous to quantitative traits can identify variants of a disease, we implemented a phenotype that was well characterized in PheKB, venous thromboembolism (VTE). To demonstrate the utility of this quantitative trait proxy in a genetics study, we trained classifiers on algorithmically phenotyped cases and controls of venous thromboembolism (VTE) using EHR-extracted features. We then tested the classifiers on VTE cases and controls in the EHR with genotyping data. Finally, to better understand the effect of converting a binary trait to quantitative on power, we simulated a complex trait. We used an additive model, as suggested through previous studies, as the mode of inheritance of causal traits [127], and varied the effect sizes, penetrance, and allele frequencies.

In our heritability estimation study, we show that our stroke phenotyping models can estimate stroke heritability, while traditional binary classification of stroke could not. In our genome-wide association study of venous thromboembolism in the Columbia eMERGE cohort, we were underpowered to recover variants. Our simulation studies, however, suggest that power may be minimally improved by converting binary traits to quantitative ones.

4.2 Methods

4.2.1 Estimating heritability in stroke with phenotyping model probabilities

We applied the probabilities generated from our phenotyping models to estimate the heritability of stroke without genetic data (Figure 4.1). We ran SOLARstrap against the model probabilities identified from each case-control and classifier model. For every model, we estimated the heritability 200 times, sampling 2000 families for each run. This bootstrapping method reduces ascertainment bias in the heritability estimation. If we were to estimate the heritability using all families available in the EHR, the estimate would be influenced by the most complete families with the disease in the EHR. By bootstrapping the families, we reduce inflated heritability estimation, as shown in [51]. We compared heritability estimates with and without taking into account shared environment within households. Finally, we compared the number of SOLARstrap runs that successfully converged using the original model probabilities to the number of runs that converged after multiplying the probabilities by 100.

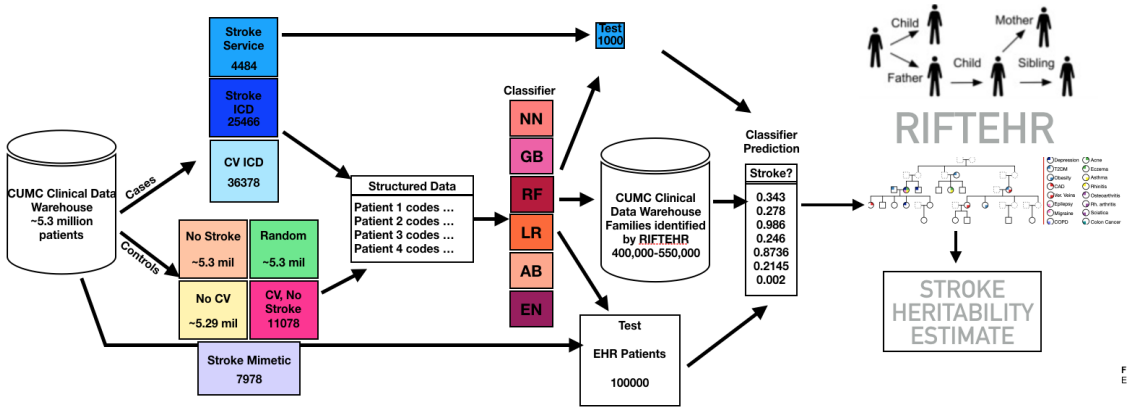


Figure 4.1: **Schematic of Stroke Model Training and Application to heritability estimation.** Cases:1) Physician curated, gold standard stroke patients (G), 2) Patients with stroke insurance billing codes (I), 3) Patients with cerebrovascular disease billing codes (C) Controls: 1) Stroke mimetic neurological diseases (N), 2,3) Sample of patients in EHR without stroke code (I,S), 4) Sample of patients in EHR without cerebrovascular disease codes (C), 5) Sample of patients in EHR with cerebrovascular disease codes but without stroke codes or not gold standard (CI) 6) Random sample in EHR (R). Case:Control ratio was between 1:1 and 1:2, models included Random Forest (RF), Logistic Regression (LR), Neural Network (NN), Gradient Boosting (GB), Elastic Net (EN) and Adaboost (AB).

4.2.2 Venous Thromboembolism Phenotyping Model Development

From the Columbia group of the eMERGE consortium, we received 3,071 patients' MRNs linked to the New York Presbyterian Hospital EHR. Given the small number of patients and the low effect sizes of known stroke variants, we did not use stroke as our test disease[11]. We instead chose a disease with a rule-based phenotyping algorithm from the PheKB database due to the PheKB algorithms' high positive predictive value and reproducibility across institutes[126]. Seven of the eMERGE PheKB algorithms were applied to the Columbia eMERGE group patients: venous thromboembolism, chronic kidney disease, colorectal cancer, autism, rheumatoid arthritis, type II diabetes, and heart failure. Because of the small sample size of patients, we filtered for diseases with known associated variants with an odds ratio of greater than 2 using the EBI-GWAS catalog and the largest number of cases[128]. Chronic kidney disease had the most cases, 2231, but only 288 controls, which would have lead to a large case/control imbalance. We chose venous thromboembolism (VTE), which is a thrombotic disease and a leading cause of cardiovascular death[129, 130]. It is often caused by commonly inherited variants leading to a hypercoaguable state[129]. The disease has 5 high effect size variants, increasing the chance of recovery through GWAS[131, 130, 33, 129, 132]. The algorithm was developed at the Mayo Clinic and had 100% positive predictive value and 95% negative predictive value at their institute[133]. When applied to the Columbia eMERGE dataset, the algorithm assigned 419 as cases for VTE, 2,551 as controls, and 101 as undetermined. We collected genotyping data on 2,065 subjects, 1058 of which were male, and 1007 female. (VTE cases, n=289, controls, n=1721) We then trained several machine learning models on the patients without genetic data and with a definite case-control label (VTE cases, n=130, controls, n=830). From the EHR, we pulled conditions, procedures, drugs, and all demographics of each patient, assigning each as a binary presence or absence of the feature to each patient. Age was also treated as a binary variable, where greater than 50 was considered 1. We trained logistic regression models with L1 (LR) and elastic net (EN) penalties, a random forest (RF) model, gradient boosting (GB) model, adaboost (AB) model, and a neural network (NN) on the 960 patients without genotyping data, evaluating with 10-fold cross validation. Hyperparame-

ters were the same as described in Method 2.3.5. In addition, we compared the performance of 1) using feature counts instead of binary presence or absence of each feature, and 2) only including features with 2 or more counts in the model. We then tested the original binary feature models on the patients with genetic data.

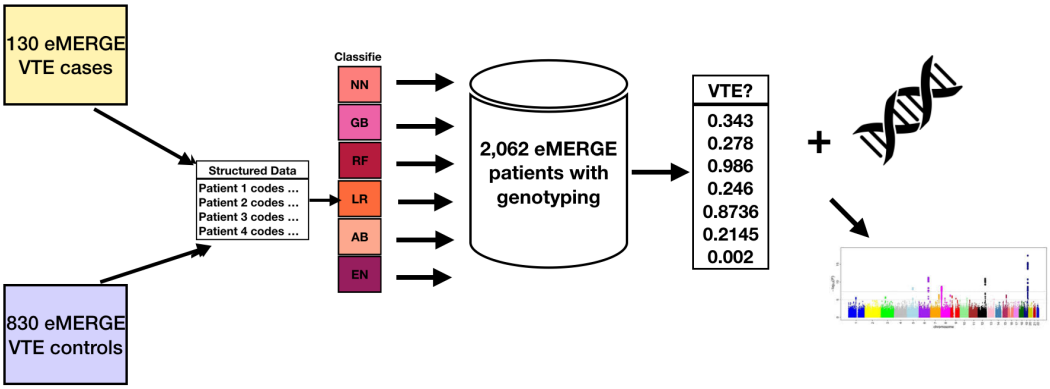


Figure 4.2: **Schematic of Model Training and Application in GWAS.** Cases: eMERGE Venous Thromboembolism cases without genetic data, Controls: eMERGE Venous Thromboembolism controls without genetic data. Models included Random Forest (RF), Logistic Regression (LR), Neural Network (NN), Gradient Boosting (GB), Elastic Net (EN) and Adaboost (AB) and are trained on structured medical data and applied to 2,062 patients with genotyping data. Quantitative trait GWAS using the model probabilities for the patients was then run.

4.2.3 VTE GWAS implementation

After assigning model probabilities for VTE for Columbia eMERGE subjects with genetic data, we ran genome-wide association studies. The genotyping data was imputed by the University of Michigan Imputation Server (MIS) using the Haplotype Reference Consortium (HRC1.1) in build 37[35]. Quality control on the genotyping data conducted by Dr. Anna Basile included filtering for SNPs with a minor allele frequency of greater than 1%, Hardy-Weinberg equilibrium test p-value $> 10^{-10}$, and imputation INFO scores of greater than 0.3. She then ran linkage disequilibrium pruning on the variants with an r^2 threshold of 0.7 and PCA analysis. The top five principal components explained over 98% of the variance, so we used these 5 PCs as covariates in addition to sex, age, incidence of stroke, and incidence of myocardial infarction. We ran a Mann-Whitney U test comparing the cases and controls for each covariate and found significant differences for all except

for PCs 2, 4, and 5. We removed these PCs from the covariates. We then ran a logistic regression GWAS on the 2,062 patients using Plink, where cases and controls were identified from the PheKB algorithm[134]. We then ran four linear regression genome-wide association studies on the same patients' genotyping data using different quantitative phenotypes: 1) the model probabilities from the neural network (NN) model, 2) model probabilities from the random forest (RF) model, and 3) and 4), the log base 10 of the NN and RF model probabilities. We chose the RF model since it gave the best performance (see Result 4.3.2) and the NN model since its model probabilities were the most normally distributed. We also compared the log base 10 probabilities since their distributions were more normal. We determined whether the known VTE variants were recovered by the binary or quantitative studies by first converting the known VTE variants from the EBI-GWAS database to chromosome positions from the chr38:12 build to the chr37.p13 build[128]. The genetic data was imputed by the University of Michigan, and we were given chr:bp as variant identifiers rather than SNP name. We mapped the variants using the NCBI Genome Remapping Service[135].

4.3 Results

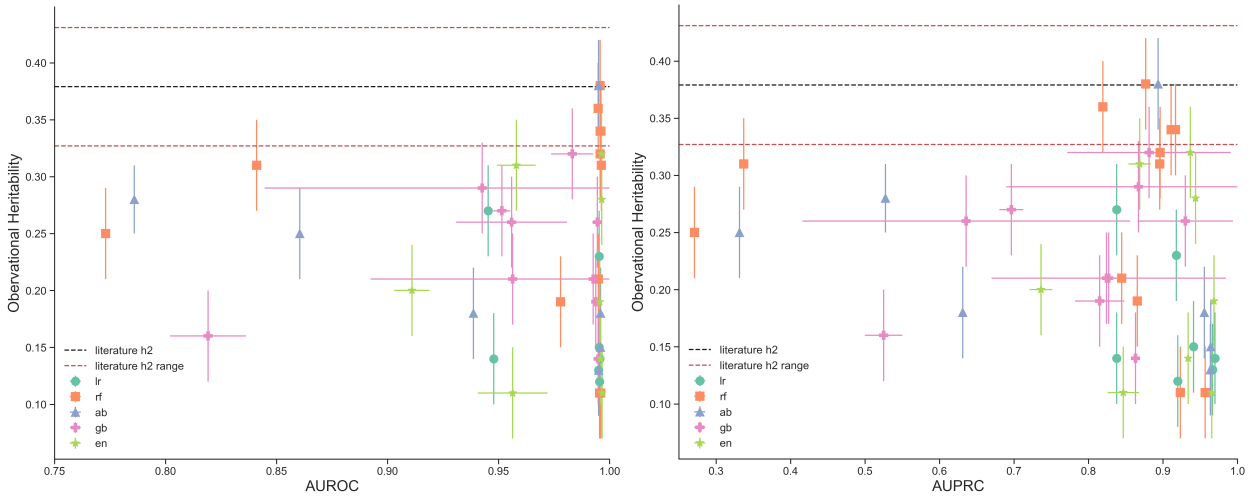
4.3.1 Heritability estimates using the models as quantitative traits

Using the model probabilities from our phenotyping models, we estimated the observational heritability of acute ischemic stroke to be 0.16-0.28, depending on the model used. We used a cutoff of at least 30/200 converging runs and a Proportion of Significant Attempts (POSA) score of 0.7. As shown in Polubriaginof et. al, the POSA score measures the proportion of converged runs that are statistically significant. The more runs that converged and are statistically significant, the more representative the heritability of all samples. Adjusting by household effect reduced the heritability estimates by an average of 6.2%. Random Forest models had the most case-control sets successfully estimate heritability (12/15). All other classifier types increased the proportion of successful heritability estimates among case-control sets (LR from 6/15 to 7/15, AB from 1/15 to 7/15, GB from 4/15 to 11/15, and EN, from 7/15 to 8/15) by multiplying the model probabilities by 100. Though the number of successful estimates increased, mean values did not change apprecia-

bly, with an increase of 0.1-0.2%. The RF models estimated the highest mean heritability for acute ischemic stroke, 0.27 ± 0.090, LR had a mean heritability estimate of 0.17 ± 0.053, AB, 0.22 ± 0.08, GB, 0.23 ± 0.053, and EN, 0.21 ± 0.081 The TCI case-control set successfully estimated heritability across all classifier types. The CC case-control set with random forest or adaboost classifier had the highest heritability estimates of 0.38 ± 0.04. The TI, TC, and CI case-control sets with random forest classifier also had high heritability estimates of 0.34 ± 0.04, 0.34 ± 0.04 and 0.36 ± 0.04 respectively. The SCI and CCI case-control sets had the lowest mean heritability estimates of 0.12 ± 0.013 and 0.12 ± 0.015 respectively. Figure 4.3a shows the observational heritability estimates compared to model area under the receiver operating curve, and figure 4.3b shows the estimates compared to area under the precision-recall curve. Neither shows an appreciable relationship between heritability estimates and performance. However, some high performing random forest and adaboost models have heritability estimates within the literature heritability estimate range for acute ischemic stroke. The curated case (S) models had the lowest average heritability estimates 0.20 ± 0.08, followed by the CCS cerebrovascular disease case models (C), 0.24 ± 0.09, and finally the AIS diagnosis code cases (T) 0.25 ± 0.07. The no ischemic stroke diagnosis code controls models had the highest heritability estimates (0.28 ± 0.059) followed by the no CCS cerebrovascular disease control models (0.26 ± 0.097). The N and CI control models had the lowest heritability estimates (0.18 ± 0.033, 0.18 ± 0.070 respectively) and the random R control models had a mean heritability estimate of 0.22 ± 0.070. Overall, our average heritability estimates were 36% lower than the literature estimate of 0.379 ± 0.052.

4.3.2 Performance of the VTE Phenotyping Algorithms

Cross-validation performance: The random forest model had the highest performance with an AUROC of 0.84, area under the precision-recall curve (AUPRC) was 0.54 and maximum F1 score was 0.51. The LR model had an AUROC of 0.82 and AUPRC of 0.48, AB model had an AUROC of 0.79 and AUPRC of 0.47, EN model had an AUROC of 0.77 and AUPRC of 0.47, GB had an AUROC of 0.71 and AUPRC of 0.40, and NN model had an AUROC of 0.79 and AUPRC of 0.41.



(a) Observational heritability vs. AUROC

(b) Observational heritability vs AUPRC

Figure 4.3: **Observational heritability estimates versus phenotyping model performance.** (a) shows the observational heritability estimates versus area under the receiver operating curve (AUROC) and literature heritability (h2) range (b) shows the observational heritability estimates versus area under the receiver precision-recall curve (AUPRC) and literature heritability (h2) range

Changing the features from binary to counts or only using features with 2 or more counts resulted on average in only a small (0.01-0.02) drop in AUROC.

4.3.3 Genome-wide association study for VTE

We were unable to recover any known VTE variants to nominal 0.05 significance in the binary or the quantitative GWAS. We also were unable to find any variants with genome-wide significance. Only 59 out of 129 known VTE variants were found in the eMERGE genotyping array, which had 6.8 million variants. A Mann-Whitney U test of the ranks of the VTE variants by P-value compared to the other variants showed an AUROC of 0.6 for the binary trait GWAS, and 0.575 for the quantitative trait GWAS using the log probability of the random forest model. Logarithm of the p-values of the binary trait GWAS results were minimally correlated with the quantitative trait results with an r^2 of 0.30.

4.3.4 Simulation of traits

The quantitative VTE score represents a probability of a patient's classification as a VTE patient based off their medical history. The probability is derived from a model trained on the EHR-derived medical history of PheKB determined cases and controls. We were unable to rediscover the known VTE variants using binary or quantitative traits, most likely due to a relatively small sample size. To further test whether quantitative traits could recover variants as well as binary traits, we ran a simulation on a single variant (Figure 4.4). We tested whether quantitative traits would be more robust to phenotyping noise than binary traits. We created a sample of 10,000 patients, with a minor allele frequency of 0.1, penetrance of 0.3 of the variant leading to disease, and general population disease risk of 0.05. We then randomly assigned patients both variant and disease status based on the rates described above. We then added noise at levels between .025 and .5 to the disease labels and calculated the odds ratios and p-values of the variant effect size using a Fisher's exact test.

We simulated 100 binary features for each patient, and assigned the features randomly at rates between 0-1 for patients with the variant and between 0-0.5 for patients without the variant. We include a wider range of feature assignment rates for those with the variant because we imagine there would be specific tell tale features, such as a medication or a condition that may more commonly be seen in the patients with the variants. We then trained a random forest classifier on 75% of patients and their 100 features to predict the remaining patients' disease status. We then fit a linear regression on the disease probabilities assigned by the model to variant status for each patient and extracted the beta coefficient and p-value. We then swapped the disease labels of the training set at different noise rates between 0.025-0.5. We again predicted on the remaining 25%, fit a linear regression to the disease probabilities identified by the model to determine variant status of each patient, and extracted the beta coefficients and their p-values.

4.3.5 Simulation of traits results

We found that penetrance of the variant was directly proportional to the performance of the classifier (Figure 4.5A). We also found that increased noise was inversely proportional to the per-

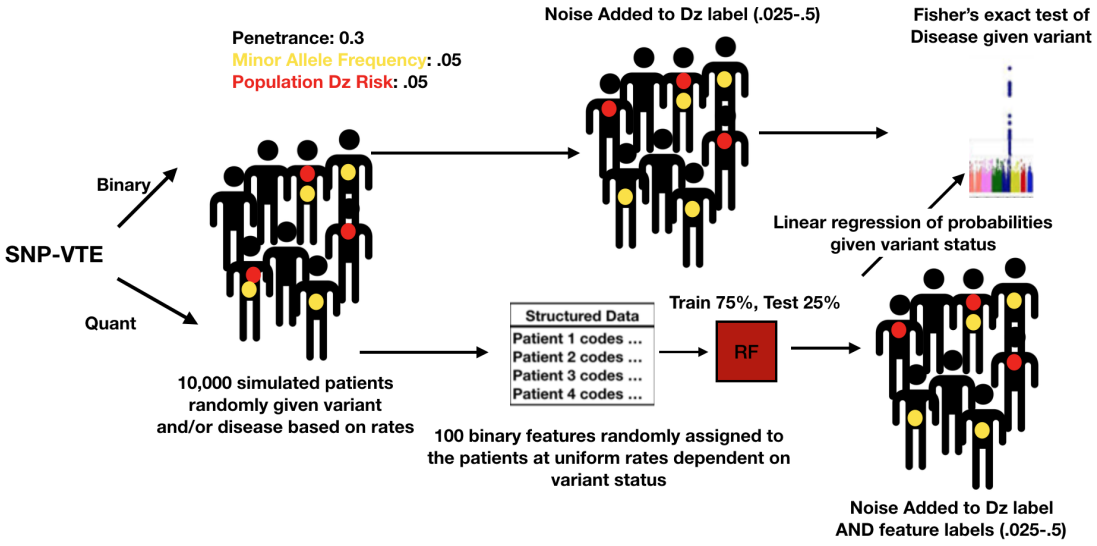


Figure 4.4: **Schematic of Disease-Trait Power Simulation**

formance of the classifier (Figure 4.5B). Within the binary trait simulation results, we found that a penetrance of 0.3 gave an odds ratio of 8.1, and so used this value for the rest of the simulations. We found that for the binary trait simulation, as noise increased, the odds ratio decreased towards 1 as did the standard error. P values also were very low at low noise and moved towards insignificant with noise at rates of 40% and 50% (Figure 4.5C). For the quantitative simulation, the p value also increased towards insignificant at a rate of 0.5, and was lower than the binary trait p values at all noise levels. Although converted odds ratios from the beta coefficients of the linear regression using the quantitative trait were lower than the binary trait odds ratio, both held genome wide significance until a noise rate of 50%.

4.4 Discussion

4.4.1 The models can estimate observational heritability at a lower average value than the literature estimate.

In Chapter 2, we developed a method to create a probability of disease for the acute ischemic stroke phenotype. The RIFTEHR and SOLARstrap methods in Polubriaginof et. al. estimated the

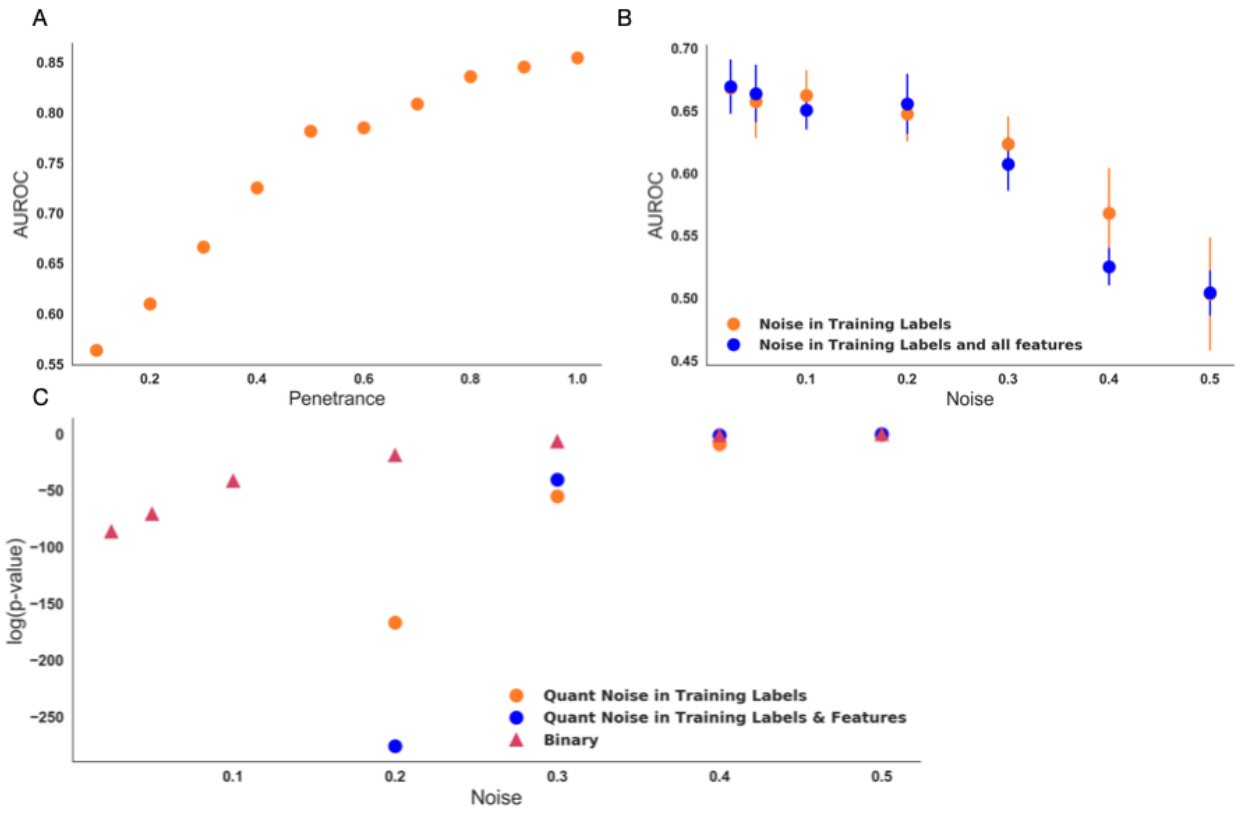


Figure 4.5: **Results of Simulation Study, Varying Penetrance and Noise.** (A) Area Under the Receiver Operating Curve (AUROC) vs. Penetrance for simulated quantitative trait, (B) Area Under the Receiver Operating Curve (AUROC) vs. Noise for simulated quantitative trait. Orange represents results from noise only in training labels while blue represents results from noise in both the training labels and feature assignment. (C) Log base 10 of P value versus Noise. Red triangles represent the binary trait p-values, orange circles represent results from noise only in training labels while blue circles represent results from noise in both the training labels and feature assignment.

heritability of over 500 diseases using only EHR data. For over 2000 diseases, however, including stroke, the methods were unable to estimate heritability. We hypothesized that our models could recover some of these heritability estimates by expanding case/control assignment to a probability of disease and subsequently converting the binary trait to a quantitative one. For many of our models, especially those using random forest classifiers, the RIFTEHR and SOLARstrap methods successfully converged. The average heritability estimate was 36% lower than the known heritability estimate. In general, RIFTEHR and SOLARstrap estimate heritabilities 20% lower than the literature values[51]. Our models' heritability estimates were lower and did not appear to have a correlation between performance and heritability estimate. Some high performing random forest

and adaboost models did estimate heritability within the literature range. In addition, the CCS cerebrovascular disease (C) or ischemic stroke diagnosis code (I) cases and no cerebrovascular disease or no ischemic stroke controls had the highest heritability estimates overall. This suggests that expanding the case set by training on CCS cerebrovascular disease classifications or diagnosis codes for stroke captured the disease heritability across families more successfully than training on a curated data set. Although further study is needed to optimize the use of model probabilities to convert binary to quantitative traits, we show that we can recover observational heritability estimates for initially under-powered phenotypes, such as acute ischemic stroke.

4.4.2 Genome-wide association studies of venous thromboembolism were underpowered in the Columbia eMERGE dataset.

Our study was underpowered to discover genome-wide significant VTE variants using either traditional binary phenotyping or model probability phenotyping. All prior VTE GWAS studies had at least 1,500 cases of European ancestry or 400 cases of African ancestry, while our study only had 289 cases, half of which were African ancestry and the other half European. Although venous thromboembolism has known variants with high odds ratios, we were unable to recover these variants to nominal significance. Simulated trait studies suggest that conversion from a dichotomous to quantitative trait will result in minimal to potentially improved power to re-identify established causal variants in GWAS. We plan to run future studies with larger cohort size to demonstrate that quantitative proxy traits can improve performance by increasing the effective sample size.

4.5 Conclusions

Using EHR-derived model probabilities as quantitative traits to replace traditional case-control assignment in genetic studies requires further analysis in larger cohorts, as demonstrated in our under-powered venous thromboembolism genome-wide association study. The quantitative trait proxy method did estimate observational stroke heritability, however, while traditional case-control assignment failed. Our simulated study also suggests potential p-value improvement when convert-

ing binary traits to quantitative, though future evaluation in a large cohort is needed.

4.6 Acknowledgements

I would like to thank Dr. Anna O. Basile for running the quality control analysis, linkage disequilibirum pruning, principal component analysis, and genetic ancestry analysis on the Columbia eMERGE dataset for the VTE study. I also thank her for helping determine the optimal test disease. I would also like to thank Dr. Ning Shang for applying the VTE PheKB algorithm to the Columbia eMERGE dataset. I would also like to thank Dr. Fernanda Polubriaginof for providing an algorithm for identifying ethnicity in the electronic health record and for helpful discussions in using RIFTEHR. Finally, I would like to thank Dr. Krzysztof Kiryluk, Dr. Ali Gharavi, Dr. Chunhua Weng, Dr. Mitchell Elkind, and Dr. George Hripcsak for helpful discussions and access to the Columbia eMERGE genomic data and PheKB algorithm assignment.

Chapter 5: QTPhenProxy, a supervised machine learning model that leverages Electronic Health Record data to improve power in genome-wide association studies in the UK Biobank

5.1 Introduction

Genome-wide association studies accumulate hundreds of thousands to millions of participants to acquire adequate signal for variant discovery. High-throughput identification of cases and controls can be difficult, however, due to time-consuming chart review and incompleteness of medical records. Current disease genome-wide association studies develop case-control sets in which power relies on a large number of pure cases, and the missingness of EHR data could prevent some cases from discovery in a high-throughput manner[13, 44, 136]. In addition, extreme case-control imbalance in biobanks can lead to increased type 1 error when running linear mixed model genome-wide association analysis[137]. An incorporation of additional accessible EHR data could improve case curation sensitivity. In addition, many diseases such as stroke result from a combination of gene and environmental interactions, and there is significant overlap with comorbidities in genome-wide significant variants[11]. Therefore, it is difficult to confirm every person without the event is a control, suggesting the utility of a disease likelihood assignment[13, 138].

The definition of the disease phenotype influences the success of detecting a genetic signal since power is generally calculated by number of cases and controls [17, 139]. We propose that including EHR information about comorbidities[140] and other health information about the subject in trait assignment can improve the power of GWAS. Past studies have shown that incorporating diagnosis count improved the power of genetic studies, and the addition of patient questionnaires and genetic correlations to hospital records improved detection of cases[141, 136, 142, 143]. Incorporating EHR data to develop a probability of suicide attempt also improved the power of its

genome-wide association study [143]. We argue that including several modalities of health data to estimate assignment of case probability can improve the power of genomic studies. For example, the most successful genome-wide association study for stroke required 40,585 cases and 406,111 controls[11]. We hypothesize that we can discover genome-wide significant variants associated with stroke with a fraction of those cases (4,354) by incorporating EHR information into a quantitative trait assignment.

In this study, we use machine learning methods to expand sample cohorts by assigning every patient a probability of disease. As described in Chapter 2, the probability represents patients along a spectrum of the disease including those who have experienced stroke, those who are predisposed to stroke but have not experienced a stroke, and those who have not had a stroke and do not have stroke risk factors. We hypothesize that the output of a supervised machine learning classifier, trained on the EHR data of a small number of known cases and controls, can be used as a proxy variable for stroke and will be an efficient strategy for expanding cohort size. We demonstrate that our quantitative proxy trait can improve power over its respective binary trait in ischemic stroke. We also show that the new variants discovered are known in similar cardiovascular and neurological diseases. We find up to 13 LD independent loci that pass genome-wide significance and conditional analysis, and the majority of the associated genes are known to be associated with stroke or cardiovascular disease. For stroke and its subtypes, ischemic stroke, subarachnoid hemorrhage, and intracerebral hemorrhage, QTPhenProxy recovered known and discovered new stroke variants with an order of magnitude fewer cases than traditional genome-wide association studies.

5.2 Methods

5.2.1 QTPhenProxy Phenotyping Model.

We gathered clinical features of primary ICD10 diagnosis codes, OCPS4 procedure codes (UK Biobank code 42100), medications (UK Biobank code 20003), race/ethnicity (UK Biobank code 1001), and age. 2018 served as the age end point. We mapped the OCPS4 procedure codes (UK

Biobank code 240) to SNOMED-CT codes and the UK Biobank medication codes to RxNorm codes by name[99, 144]. We also gathered data of self-reported and first occurrence of all stroke, ischemic stroke, subarachnoid stroke, and intracerebral hemorrhage, asthma, COPD, myocardial infarction, ST-elevation Myocardial infarction, and Non ST-elevation Myocardial Infarction. We chose to study all 18 diseases that had a validated algorithmically defined outcome (category 42) in the UK Biobank: the above stroke subtypes, myocardial infarction and two main subtypes: STEMI and NSTEMI, asthma, COPD, dementia and three main subtypes: frontotemporal dementia, vascular dementia, and Alzheimer's disease, motor neuron disease (also known as ALS), and Parkinson's disease and three subtypes: Parkinsonism, Progressive Supranuclear Palsy, and Multiple Systems Atrophy. We made a large matrix in which each extracted EHR feature was a binary variable based on the presence or absence of the feature. We dichotomized age as greater than or equal or less than 50 years. For each disease, we then defined the cases from the ICD10 code combinations described in the UK Biobank's phenotyping algorithm[145]. We then trained 5 different classifiers: 1) logistic regression with elastic net penalty, 2) logistic regression with L1 penalty, 3) random forest, 4) Adaboost, and 5) gradient boosting classifiers on 50% of the cases and an equal number of controls. Controls were identified as subjects without any ICD10 codes within the same category as the Clinical Modification Clinical Classifications Software tool[102]. We then applied the trained algorithm to the whole UK Biobank, resulting in a model probability or quantitative trait proxy for each subject and each disease. For comparison, another phenotype file with binary assignment of case and control for each disease was prepared as well. Table 5.4 shows the number of cases available for each disease.

5.2.2 Evaluation of QTPhenProxy Model Performance.

To evaluate the performance of the QTPhenProxy model, we determined its ability to recover cases when the defining ICD10 codes are removed. We trained 50% of the cases and an equal number of controls on the clinical features described above using the EN and RF classifiers. We chose these classifiers because of their overall high performance and because their probability

assignment distributions were continuous (Figure 5.2). We removed the ICD10 codes used to define the cases and controls in our feature set. We then tested the model on all other subjects in the UK Biobank, which included known cases, and for some diseases such as Ischemic Stroke and Myocardial Infarction, self-reported cases that did not have an ICD10 code for the disease. We then evaluated the recovery of 1) cross-validation and 2) the holdout test set through Precision at top 50, 100, 500, and number of cases, area under the receiver operating curve, area under the precision recall curve, and maximum F1 score.

5.2.3 Genotyping and Imputation.

UK Biobank subjects that were of White British descent, were in the UK Biobank PCA calculations and therefore without 3rd degree and above relatedness, and were without aneuploidy were used in this study, totalling 337147 subjects (181,032 females and 156,115 males)[146, 28]. Of the nearly 500,000 participants, approximately 50,000 subjects were genotyped on the UK BiLEVE Array by Affymetrix while the rest were genotyped using the Applied Biosystems UK Biobank Axiom Array with over 800,000 markers. The arrays share 95% marker coverage. Initially, we ran a QC extracting markers with an MAF>0.5%, INFO score > 0.3, and Hardy-Weinberg equilibrium test mid-p-value > 10^{-10} using all subjects, which we will refer to as QC1. QC1 was run for GWAS of all 18 diseases. We then re-ran the QC, which we will refer to as QC2, more stringently by extracting markers with an MAF>1%, INFO score > 0.8, and Hardy-Weinberg equilibrium test mid-p-value > 10^{-5} using Plink2[147]. UKBB version 3 Imputation combined the Haplotype Research Consortium with the UK10K haplotype resource using the software IMPUTE4[146].

5.2.4 Genome-wide Association Analysis.

The binary trait and QTPhenProxy probabilities were compared by running two separate association analyses. For both analyses, covariates included age at 2018, sex, first 10 principal components, and the genotyping array the sample was carried out on. In QC1, the original PCs determined by the UK Biobank QC were used. In QC2, we calculated the PCs using the method

described in Method 5.2.11. QC2 GWAS was only run for stroke, ischemic stroke, subarachnoid hemorrhage, and intracerebral hemorrhage GWAS. For the binary trait GWAS, a logistic regression was run, adjusted with the aforementioned covariates. For comparison, the QTPhenProxy probabilities were quantile normalized and run under a linear regression adjusting for the same covariates. We also permuted the probabilities within the phenotyping files and ran additional GWAS 10 times to ensure the signal was correlated with the phenotype.

5.2.5 Mapping variants to known disease variant marker sets and mapping marker sets to disease systems.

The EBI-GWAS catalogue has a database of all published GWAS[128]. We extracted over 2,000 disease marker sets conducted on populations with European ancestry. MedDRA is a standardized medical vocabulary developed by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH)[148]. All terms in the vocabulary can be mapped to its highest system level, which includes 27 different organ systems and other general and lab studies such as social circumstance and investigations. Using the NCBO annotator, we mapped the names of the EBI-GWAS disease marker sets to the MedDRA System Organ Classes level[149].

5.2.6 Assessing the specificity of the QTPhenProxy-derived variants.

To assess the disease specificity of the genome-wide significant variants, we first calculated the proportion of genome-wide significant variants in each of the EBI-GWAS disease marker sets. We then aggregated the marker sets together by System Organ Class to evaluate the systems enriched for genome-wide significant variants. We ordered the marker sets in each class by proportion of genome-wide significant variants and divided by the number of marker sets in each class. We also compared the proportion of variants of varying significance of the marker sets related to each disease with 1) the other marker sets related to the same System Organ Class as the disease and 2) all other marker sets. We stratified each comparison by significance value, between 0.05 and

5E-08.

5.2.7 Evaluation of recovery of known variants

For each disease, we gathered EBI-GWAS marker sets that contained the disease in its name. These represent known variants of each disease. We then extracted the p-value from either the binary trait logistic regression or QTPhenProxy linear regression. We then ran a t-test comparing the negative log base 10 p-values of the binary trait with QTPhenProxy GWAS. We also ran a t-test comparing the difference between the binary and QTPhenProxy log base 10 p-values for the known ischemic stroke variants and an equivalent number of random variants.

5.2.8 Refinement of discovered variants by QTPhenProxy using conditional analysis

At each LD-independent locus, the SNP with lowest p-value may not be the variant that causes the most phenotypic variation within the area[150]. Therefore, we applied GCTA-COJO, a conditional analysis that takes into account lead SNPs and the LD structure of a sample of the population, to our genome-wide association results[150]. We randomly sampled 10,000 subjects from the UK-Biobank for the linkage disequilibrium calculation[151]. From the GCTA-COJO results, we then mapped each locus to its nearest gene using dbSNP and the UCSC Genome Browser accessed at http://genome.ucsc.edu/[152, 153]. For intergenic loci, we chose the 1-2 nearest genes that were at most 10,000 kbps away.

5.2.9 Correlation of QTPhenProxy GWAS beta coefficients to Binary trait GWAS Odds Ratio

We calculated the Pearson correlation between the beta-coefficients of the QTPhenProxy GWAS and log of the odds ratios of the Binary trait GWAS. In order to account for noise, we calculated the correlation with variants with different levels of significance in the QTPhenProxy models. P-value cutoffs included 1, 0.05,0.0005,5e-06,and 5e-08.

5.2.10 Simulation of Conversion of QTPhenProxy trait to Binary trait and Conversion of beta coefficients to odds ratios

In order to validate our method of converting binary traits to quantitative traits, we ran simulations and tested the correlation between the two methods. Using SOLAR, a software package for estimating heritability using identity by descent calculations, we simulated a quantitative trait with one quantitative trait locus with two alleles and a nearby marker locus with two alleles[124]. We first removed all related individuals with a resulting cohort of 4,195 subjects. For the simulation, we varied the frequency for the causal minor allele and a marker minor allele from 0.05-0.45 in increments of 0.010, the mean quantitative trait value for the heterozygous genotype from 5-45 in increments of 10 and the homozygous genotypes' mean±50, the standard deviation of the quantitative trait from 5-20 in increments of 4, and the recombinant fraction from 0.01-0.10 in increments of 0.02. After simulating the quantitative trait distribution, we then normalized the trait to several distributions: standard normal, normal distribution with mean 0 and standard deviation 10, and mean 50 and standard deviation 10. We compared distributions because we quantile normalized our QTPhenProxy trait values before running the genome-wide associations studies. We then converted each simulation to a binary trait using liability thresholding[154]. Liability thresholding was implemented as follows: We determined a quantitative trait value as a threshold based off the prevalence of the simulated trait, which we varied from 2.5-20% in increments of 5%. Any subject above this threshold is labeled a case and the rest, controls, in the binary trait phenotype. We then ran linear or logistic regressions using the python package *statsmodels* between the simulated quantitative trait or the binary trait and the subjects' genotypes for the marker and causal loci[155]. We developed a conversion formula for the beta coefficients to odds ratios by linearly regressing the correlation between the simulated effect sizes. We then converted the beta-coefficients to odds ratios of the UK Biobank GWAS results by multiplying the beta coefficient by the average slope and intercept and then taking the exponential of the result.

5.2.11 PCA

In order to confirm the cases were well distributed within the data and to determine the number of principal components to use as covariates, we conducted PCA. For each of the four diseases, we first pruned the variants used for PCA by running a sliding window of size 100 kbps, 5 variant step size, and r^2 threshold of 0.1. We then combined the chromosomes, extracting only the pruned variants, using Plink2 and cat-bgen software[156]. We then ran PCA with Plink2[147] and evaluated the PCs to be used as covariates using a skree plot. Within the main PCs, we plotted them against each other, highlighting the distribution of the cases.

5.2.12 LD Score Regression and evaluation of genomic inflation

We determined the lambda genomic correction and LD score regression coefficient using the software LDSC[157]. We used the disease GWAS summary statistics and European LD scores precomputed from 1000 genomes by the Alkes group[157]. QQ plots were plotted using qqman[158]. To determine the relationship between genomic inflation and minor allele frequency, we binned all variants in the stroke GWAS by minor allele frequency into 30 bins. We then calculated the genomic inflation of the p-values of the variants in each bin.

5.2.13 Genetic Correlation of QTPhenProxy with MEGASTROKE and Coronary Artery Disease GWAS

We measured the genetic correlation of the QTPhenProxy EN stroke model with QC1 quality control with the MEGASTROKE all stroke GWAS summary statistics[11] and coronary artery disease GWAS summary statistics from [159]. We calculated genetic correlation using the LDSC software[157].

5.3 Results

5.3.1 QTPhenProxy Model Performance

We trained models with 5 different classifier types and the 18 disease phenotypes as defined by the UK Biobank Algorithmically Defined Outcomes rubric[145]. Random forest models overall showed the best area under the receiver operating curve and Adaboost models gave the best area under the precision recall curve on the hold-out test set. Overall stroke, followed by ischemic stroke, showed high precision at top 50 patients ordered by probability (Figure 5.1, tables 5.1, 5.2 and 5.5). We chose the probabilities from the EN and RF models to run the genome-wide association analyses because the distribution of their probabilities was continuous and included values from 0-1 (Figure 5.2).

Disease	Model	AUROC	AUPRC	Maximum F1
Ischemic Stroke	rf	0.953(0.000279)	0.248(0.00201)	0.326(0.00229)
	ab	0.950(0.000504)	0.246(0.00465)	0.316(0.00335)
	lr	0.946(0.000409)	0.239(0.0021)	0.322(0.00196)
	gb	0.952(0.000686)	0.210(0.00479)	0.294(0.00319)
	en	0.943(0.000362)	0.209(0.00193)	0.295(0.00199)
SAH Stroke	rf	0.875(0.000746)	0.144(0.00299)	0.242(0.00423)
	ab	0.859(0.00186)	0.171(0.0062)	0.271(0.00807)
	lr	0.850(0.00115)	0.151(0.00419)	0.291(0.00509)
	gb	0.863(0.00149)	0.113(0.00379)	0.203(0.00513)
	en	0.863(0.000958)	0.109(0.00377)	0.198(0.00497)
ICH Stroke	rf	0.903(0.000924)	0.0483(0.00143)	0.116(0.00253)
	ab	0.900(0.00171)	0.0480(0.00196)	0.114(0.00305)
	lr	0.879(0.00145)	0.0270(0.000709)	0.0762(0.00137)
	gb	0.900(0.00123)	0.0365(0.00125)	0.0935(0.00227)
	en	0.903(0.00118)	0.0329(0.000755)	0.0861(0.00146)
Stroke	rf	0.899(0.000404)	0.326(0.0016)	0.375(0.00147)
	ab	0.906(0.000711)	0.335(0.00423)	0.380(0.00262)
	lr	0.905(0.000528)	0.314(0.00143)	0.368(0.00154)
	gb	0.903(0.00047)	0.292(0.00245)	0.355(0.00151)
	en	0.896(0.000448)	0.282(0.00113)	0.350(0.000962)

Disease	Model	AUROC	AUPRC	Maximum F1
NSTEMI	rf	0.943(0.000192)	0.168(0.000887)	0.263(0.0012)
	ab	0.94(0.000469)	0.195(0.00169)	0.286(0.00153)
	lr	0.941(0.000234)	0.189(0.00102)	0.281(0.00112)
	gb	0.937(0.00617)	0.173(0.0111)	0.262(0.0127)
	en	0.94(0.000244)	0.173(0.000975)	0.258(0.00099)

Disease	Model	AUROC	AUPRC	Maximum F1
STEMI	rf	0.977(0.000154)	0.193(0.00203)	0.275(0.00208)
	ab	0.978(0.000287)	0.223(0.00227)	0.308(0.00206)
	lr	0.977(0.000262)	0.211(0.00169)	0.296(0.00149)
	gb	0.973(0.00694)	0.204(0.012)	0.283(0.0145)
	en	0.974(0.000216)	0.181(0.00142)	0.26(0.00159)

Disease	Model	AUROC	AUPRC	Maximum F1
MI	rf	0.935(0.000197)	0.356(0.001)	0.424(0.000676)
	ab	0.941(0.000156)	0.384(0.00123)	0.465(0.00106)
	lr	0.941(0.000153)	0.367(0.000968)	0.44(0.00087)
	gb	0.94(0.000161)	0.365(0.00105)	0.434(0.000826)
	en	0.931(0.000181)	0.335(0.000798)	0.392(0.000694)

Disease	Model	AUROC	AUPRC	Maximum F1
MSA	rf	0.952(0.00188)	0.00639(0.00107)	0.0317(0.00395)
	ab	0.945(0.00324)	0.00393(0.000431)	0.0203(0.00272)
	lr	0.822(0.00886)	0.00125(3.59e-05)	0.00347(5.82e-05)
	gb	0.951(0.00224)	0.00439(0.00057)	0.022(0.00231)
	en	0.952(0.00189)	0.00353(0.000161)	0.0176(0.0015)

Disease	Model	AUROC	AUPRC	Maximum F1
Motor Neuron	rf	0.927(0.00256)	0.0349(0.00328)	0.111(0.00752)
	ab	0.91(0.00428)	0.0162(0.00137)	0.0684(0.00503)
	lr	0.795(0.00795)	0.0217(0.00318)	0.1(0.00826)
	gb	0.917(0.00333)	0.0165(0.00167)	0.0646(0.00521)
	en	0.909(0.00398)	0.0151(0.00144)	0.0629(0.00491)

Disease	Model	AUROC	AUPRC	Maximum F1
Parkinson's	rf	0.886(0.0013)	0.063(0.00302)	0.139(0.00371)
	ab	0.863(0.00172)	0.0667(0.00477)	0.153(0.00694)
	lr	0.842(0.00188)	0.0549(0.00323)	0.14(0.00562)
	gb	0.866(0.00144)	0.0234(0.00113)	0.0705(0.00242)
	en	0.87(0.00164)	0.0264(0.00104)	0.0819(0.00241)

Disease	Model	AUROC	AUPRC	Maximum F1
	rf	0.908(0.000947)	0.0659(0.00119)	0.138(0.00175)
	ab	0.901(0.00141)	0.0606(0.00258)	0.124(0.00382)
Parkinsonism	lr	0.898(0.00119)	0.0583(0.000865)	0.121(0.00117)
	gb	0.903(0.00101)	0.0515(0.000953)	0.112(0.00182)
	en	0.897(0.00114)	0.0494(0.00069)	0.112(0.00121)
Disease	Model	AUROC	AUPRC	Maximum F1
	rf	0.915(0.000342)	0.2(0.00144)	0.266(0.00137)
	ab	0.915(0.000431)	0.224(0.00306)	0.292(0.00182)
COPD	lr	0.915(0.000339)	0.245(0.00116)	0.303(0.00107)
	gb	0.915(0.000321)	0.215(0.00146)	0.283(0.00133)
	en	0.906(0.000415)	0.197(0.000952)	0.264(0.00111)
Disease	Model	AUROC	AUPRC	Maximum F1
	rf	0.806(0.000366)	0.165(0.000643)	0.23(0.000714)
	ab	0.804(0.000447)	0.185(0.00118)	0.254(0.000653)
Asthma	lr	0.808(0.000415)	0.197(0.000627)	0.254(0.000564)
	gb	0.807(0.000323)	0.176(0.000695)	0.238(0.000683)
	en	0.79(0.000501)	0.146(0.000474)	0.211(0.000552)
Disease	Model	AUROC	AUPRC	Maximum F1
	rf	0.93(0.00228)	0.0158(0.00127)	0.0652(0.00345)
	ab	0.906(0.00334)	0.0105(0.00101)	0.0516(0.00376)
Vascular Dementia	lr	0.887(0.00374)	0.00842(0.000719)	0.0486(0.00442)
	gb	0.922(0.00245)	0.0116(0.00082)	0.0539(0.00368)
	en	0.919(0.00294)	0.0137(0.000906)	0.0615(0.00337)
Disease	Model	AUROC	AUPRC	Maximum F1
	rf	0.877(0.00215)	0.00989(0.000578)	0.0468(0.00245)
	ab	0.846(0.00367)	0.00697(0.000334)	0.0371(0.00215)
Alzheimer's	lr	0.808(0.00269)	0.00453(0.000162)	0.0305(0.00116)
	gb	0.857(0.00321)	0.00675(0.000255)	0.0359(0.00199)
	en	0.857(0.00285)	0.00646(0.000277)	0.0348(0.00205)
Disease	Model	AUROC	AUPRC	Maximum F1
	rf	0.895(0.00101)	0.0536(0.00112)	0.115(0.00181)
	ab	0.905(0.00106)	0.09(0.0042)	0.192(0.00907)
All Dementia	lr	0.887(0.0011)	0.119(0.00235)	0.243(0.00355)
	gb	0.905(0.000731)	0.0567(0.000921)	0.116(0.00162)
	en	0.878(0.0014)	0.0441(0.000976)	0.105(0.00182)

Table 5.1: **Phenotyping Models Performance.** *rf*: Random Forest, *ab*= Adaboost, *lr*=Logistic regression with L1 penalty, *gb*=Gradient Boosting, *en*=Logistic Regression with elastic net penalty models. *AUROC*: Area under the Receiver Operating Curve, *AUPRC*: Area under Precision-Recall curve. 95% confidence intervals

Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.739(0.0132)	0.682(0.0105)	0.544(0.00575)	0.322(0.00246)
	ab	0.71(0.0192)	0.674(0.0154)	0.527(0.0117)	0.309(0.00405)
IscStroke	lr	0.801(0.0143)	0.672(0.00997)	0.483(0.00560)	0.314(0.00205)
	gb	0.609(0.0305)	0.544(0.0246)	0.425(0.0179)	0.282(0.00714)
	en	0.74(0.0141)	0.63(0.00946)	0.435(0.00476)	0.29(0.00197)
	rf	0.607(0.0178)	0.55(0.0106)	0.308(0.00623)	0.235(0.00381)
	ab	0.627(0.0185)	0.538(0.0144)	0.329(0.0102)	0.266(0.00801)
SAHStroke	lr	0.535(0.0187)	0.5(0.0147)	0.356(0.00722)	0.271(0.00630)
	gb	0.532(0.0203)	0.442(0.0176)	0.249(0.00759)	0.197(0.00557)
	en	0.543(0.0188)	0.441(0.0136)	0.244(0.00773)	0.192(0.00523)
	rf	0.263(0.013)	0.216(0.00968)	0.124(0.00328)	0.110(0.00265)
	ab	0.203(0.0185)	0.171(0.0124)	0.108(0.00400)	0.0997(0.00352)
ICHStroke	lr	0.109(0.0121)	0.100(0.00713)	0.0742(0.00252)	0.0677(0.0027)
	gb	0.137(0.0146)	0.123(0.00838)	0.0908(0.00279)	0.0834(0.00240)
	en	0.123(0.0110)	0.115(0.00761)	0.0835(0.00262)	0.0768(0.00204)
	rf	0.974(0.00548)	0.941(0.00506)	0.822(0.00339)	0.374(0.0015)
	ab	0.940(0.0120)	0.910(0.0120)	0.813(0.0131)	0.378(0.00265)
Stroke	lr	0.929(0.00793)	0.872(0.00678)	0.775(0.00413)	0.365(0.00164)
	gb	0.868(0.0324)	0.818(0.0309)	0.696(0.0180)	0.351(0.00166)
	en	0.848(0.00829)	0.836(0.00706)	0.694(0.00394)	0.345(0.000996)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.367(0.0151)	0.35(0.0117)	0.293(0.00471)	0.228(0.00161)
	ab	0.454(0.0224)	0.428(0.016)	0.361(0.00636)	0.258(0.00208)
NSTEMI	lr	0.498(0.0128)	0.452(0.00926)	0.346(0.00383)	0.251(0.00117)
	gb	0.4(0.0451)	0.363(0.0379)	0.296(0.028)	0.224(0.0189)
	en	0.494(0.0136)	0.438(0.00987)	0.329(0.00347)	0.23(0.00129)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.376(0.019)	0.38(0.0137)	0.336(0.00505)	0.248(0.00252)
	ab	0.446(0.0192)	0.429(0.0146)	0.365(0.00649)	0.27(0.0026)
STEMI					

	lr	0.45(0.0116)	0.459(0.0087)	0.355(0.00515)	0.253(0.00195)
	gb	0.399(0.0326)	0.388(0.03)	0.341(0.0248)	0.245(0.0174)
	en	0.365(0.0117)	0.363(0.00947)	0.328(0.00404)	0.233(0.0015)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.706(0.0139)	0.677(0.0103)	0.623(0.00441)	0.414(0.000878)
	ab	0.773(0.0107)	0.755(0.00615)	0.668(0.00409)	0.449(0.00128)
MI	lr	0.772(0.00873)	0.749(0.00669)	0.685(0.00369)	0.419(0.000945)
	gb	0.746(0.0159)	0.751(0.0125)	0.684(0.00506)	0.414(0.000937)
	en	0.884(0.00827)	0.82(0.0062)	0.706(0.00302)	0.381(0.000827)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.0256(0.00608)	0.0164(0.00332)	0.0102(0.00112)	0.0185(0.00401)
	ab	0.008(0.00314)	0.007(0.00237)	0.00756(0.00114)	0.00765(0.00264)
MSA	lr	0.0036(0.00213)	0.0018(0.00106)	0.00036(0.000213)	0.00222(0.00131)
	gb	0.0128(0.0038)	0.01(0.00235)	0.0066(0.000916)	0.0109(0.00271)
	en	0.0032(0.00203)	0.0058(0.00176)	0.0058(0.000817)	0.00469(0.0018)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.164(0.0154)	0.112(0.0089)	0.041(0.00221)	0.0965(0.00728)
	ab	0.058(0.00832)	0.0526(0.00636)	0.0346(0.00249)	0.0491(0.0057)
Motor Neuron	lr	0.0968(0.0181)	0.0632(0.00948)	0.0449(0.00521)	0.0567(0.00727)
	gb	0.07(0.00949)	0.0552(0.00618)	0.0311(0.00227)	0.0518(0.00502)
	en	0.0568(0.00779)	0.0492(0.00531)	0.031(0.0024)	0.0469(0.00505)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.394(0.0231)	0.309(0.0142)	0.157(0.00497)	0.133(0.00369)
	ab	0.33(0.0263)	0.274(0.02)	0.167(0.00887)	0.147(0.00706)
Parkinson's	lr	0.331(0.0215)	0.26(0.0149)	0.154(0.00693)	0.135(0.00578)
	gb	0.12(0.0107)	0.102(0.00862)	0.0668(0.00329)	0.0617(0.00258)
	en	0.129(0.012)	0.117(0.00671)	0.0793(0.00327)	0.0725(0.00277)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.279(0.0177)	0.261(0.0115)	0.191(0.00419)	0.131(0.00218)
	ab	0.316(0.0202)	0.276(0.015)	0.18(0.00716)	0.117(0.00368)
Parkinsonism	lr	0.214(0.0116)	0.198(0.00881)	0.156(0.00365)	0.112(0.00144)
	gb	0.208(0.0134)	0.185(0.0105)	0.129(0.00372)	0.0977(0.00187)
	en	0.188(0.0115)	0.19(0.00905)	0.122(0.00319)	0.102(0.00144)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.784(0.017)	0.732(0.0123)	0.567(0.00564)	0.261(0.00162)
	ab	0.776(0.0194)	0.738(0.0188)	0.607(0.0139)	0.288(0.00201)
COPD	lr	0.878(0.00797)	0.831(0.00528)	0.681(0.00474)	0.301(0.00116)

	gb	0.766(0.0138)	0.719(0.0114)	0.584(0.0062)	0.279(0.00134)
	en	0.762(0.00879)	0.742(0.00626)	0.575(0.00457)	0.261(0.00109)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.538(0.0183)	0.542(0.0125)	0.517(0.00457)	0.225(0.000823)
	ab	0.743(0.0162)	0.708(0.0103)	0.582(0.00752)	0.251(0.000822)
Asthma	lr	0.831(0.00915)	0.808(0.00623)	0.677(0.00463)	0.251(0.000607)
	gb	0.761(0.0167)	0.725(0.0118)	0.599(0.00578)	0.233(0.000811)
	en	0.703(0.00917)	0.627(0.00856)	0.502(0.00402)	0.201(0.000615)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.0612(0.0081)	0.0532(0.00556)	0.033(0.00176)	0.0493(0.00461)
	ab	0.0432(0.00802)	0.041(0.00579)	0.0255(0.00175)	0.0382(0.00473)
Vascular Dementia	lr	0.034(0.0108)	0.0268(0.00702)	0.0187(0.00317)	0.0226(0.0059)
	gb	0.052(0.00844)	0.0434(0.00507)	0.0268(0.00168)	0.0394(0.00417)
	en	0.056(0.00815)	0.0528(0.00557)	0.0304(0.00162)	0.0488(0.00417)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.0556(0.00847)	0.0506(0.00462)	0.0311(0.00189)	0.0386(0.0029)
	ab	0.0388(0.00631)	0.0332(0.00448)	0.0231(0.0017)	0.028(0.00259)
Alzheimer's	lr	0.0156(0.00499)	0.0134(0.00374)	0.0139(0.00213)	0.0121(0.00255)
	gb	0.03(0.00568)	0.0302(0.00409)	0.0223(0.00156)	0.0277(0.00219)
	en	0.0236(0.00681)	0.0268(0.00465)	0.0214(0.00182)	0.0239(0.00234)
Disease	Model	Prec at 50	Prec at 100	Prec at 500	Prec at N
	rf	0.323(0.0125)	0.281(0.00994)	0.177(0.00397)	0.112(0.00193)
	ab	0.247(0.0144)	0.235(0.0127)	0.198(0.00998)	0.168(0.00964)
All Dementia	lr	0.38(0.0138)	0.351(0.00989)	0.26(0.00573)	0.211(0.00496)
	gb	0.178(0.0134)	0.178(0.00971)	0.146(0.0035)	0.106(0.00173)
	en	0.132(0.0121)	0.14(0.00831)	0.138(0.00381)	0.0975(0.00189)

Table 5.2: **Precision at top 50, 100, 500, and N cases probabilities of phenotyping models.** *rf*: Random Forest, *ab*= Adaboost, *lr*=Logistic regression with L1 penalty, *gb*=Gradient Boosting, *en*=Logistic Regression with elastic net penalty models. *Prec*: Precision. 95% confidence intervals

Disease	Model	Sensitivity	Specificity
	rf	0.347(0.00483)	0.996(0.000133)
	ab	0.369(0.00744)	0.995(0.000256)
Ischemic Stroke	lr	0.362(0.00528)	0.995(0.000127)
	gb	0.368(0.00868)	0.994(0.000354)

	en	0.332(0.00624)	0.995(0.000195)
SAH Stroke	rf	0.215(0.00503)	0.999(5.47e-05)
	ab	0.272(0.00924)	0.999(7.14e-05)
	lr	0.266(0.00764)	0.999(6.34e-05)
	gb	0.204(0.00598)	0.999(9.43e-05)
	en	0.201(0.00597)	0.999(0.000126)
ICH Stroke	rf	0.133(0.00715)	0.998(0.000140)
	ab	0.167(0.0113)	0.998(0.000219)
	lr	0.177(0.0211)	0.995(0.000742)
	gb	0.141(0.00950)	0.998(0.000268)
	en	0.146(0.0123)	0.997(0.000398)
Stroke	rf	0.377(0.00414)	0.989(0.000286)
	ab	0.390(0.00559)	0.989(0.000379)
	lr	0.399(0.00499)	0.987(0.000408)
	gb	0.388(0.00462)	0.986(0.000373)
	en	0.393(0.00425)	0.985(0.000336)

Disease	Model	Sensitivity	Specificity
NSTEMI	rf	0.425(0.007)	0.987(0.000314)
	ab	0.426(0.00737)	0.989(0.000312)
	lr	0.423(0.00663)	0.989(0.000284)
	gb	0.45(0.0192)	0.983(0.00473)
	en	0.387(0.00569)	0.989(0.000275)

Disease	Model	Sensitivity	Specificity
STEMI	rf	0.414(0.00813)	0.991(0.000265)
	ab	0.501(0.0112)	0.99(0.000356)
	lr	0.506(0.00956)	0.99(0.000297)
	gb	0.49(0.0177)	0.986(0.0052)
	en	0.416(0.00949)	0.99(0.000372)

Disease	Model	Sensitivity	Specificity
MI	rf	0.498(0.00399)	0.98(0.000349)
	ab	0.541(0.00383)	0.982(0.000319)
	lr	0.546(0.00269)	0.978(0.000224)
	gb	0.532(0.00443)	0.979(0.000357)
	en	0.466(0.00355)	0.979(0.000375)

Disease	Model	Sensitivity	Specificity
MSA	rf	0.0625(0.0153)	0.999(0.000229)

	ab	0.0593(0.0122)	0.999(0.00035)
	lr	0.653(0.0335)	0.939(0.00315)
	gb	0.0491(0.0115)	0.999(0.000299)
	en	0.0662(0.0131)	0.999(0.000346)

Disease	Model	Sensitivity	Specificity
	rf	0.0943(0.0061)	1(4.16e-05)
	ab	0.117(0.0123)	0.999(0.00014)
Motor Neuron	lr	0.151(0.0198)	0.999(0.000125)
	gb	0.0924(0.00995)	1(9.9e-05)
	en	0.0984(0.00965)	0.999(0.000119)

Disease	Model	Sensitivity	Specificity
	rf	0.136(0.00503)	0.999(0.000112)
	ab	0.161(0.00823)	0.999(0.000113)
Parkinson's	lr	0.153(0.00564)	0.999(0.000114)
	gb	0.105(0.00817)	0.997(0.000343)
	en	0.116(0.00616)	0.998(0.000219)

Disease	Model	Sensitivity	Specificity
	rf	0.179(0.00758)	0.995(0.000343)
	ab	0.157(0.00857)	0.995(0.000347)
Parkinsonism	lr	0.196(0.00976)	0.993(0.00053)
	gb	0.197(0.00923)	0.992(0.00053)
	en	0.165(0.00634)	0.994(0.000366)

Disease	Model	Sensitivity	Specificity
	rf	0.309(0.0054)	0.987(0.000477)
	ab	0.316(0.00507)	0.989(0.000365)
COPD	lr	0.327(0.00458)	0.989(0.000327)
	gb	0.317(0.00506)	0.988(0.000383)
	en	0.293(0.00504)	0.988(0.000419)

Disease	Model	Sensitivity	Specificity
	rf	0.278(0.00434)	0.959(0.00122)
	ab	0.287(0.00346)	0.965(0.000893)
Asthma	lr	0.288(0.00376)	0.965(0.000908)
	gb	0.289(0.00388)	0.959(0.00108)
	en	0.284(0.00375)	0.949(0.00118)

Disease	Model	Sensitivity	Specificity
	rf	0.102(0.0104)	0.999(0.000112)
	ab	0.08(0.0102)	0.999(0.000142)
Vascular Dementia			

	lr	0.113(0.0128)	0.999(0.000332)
	gb	0.0874(0.0089)	0.999(0.000135)
	en	0.0888(0.0102)	0.999(0.000102)
Disease	Model	Sensitivity	Specificity
	rf	0.062(0.00475)	0.999(0.000121)
	ab	0.0574(0.00727)	0.999(0.000213)
Alzheimer's	lr	0.0786(0.0065)	0.998(0.000177)
	gb	0.0526(0.00631)	0.999(0.000205)
	en	0.0612(0.00618)	0.999(0.000219)
Disease	Model	Sensitivity	Specificity
	rf	0.126(0.00416)	0.996(0.000232)
	ab	0.309(0.0264)	0.993(0.00122)
All Dementia	lr	0.296(0.00669)	0.996(0.000244)
	gb	0.183(0.0102)	0.993(0.000512)
	en	0.155(0.00745)	0.994(0.000491)

Table 5.3: **Sensitivity and Specificity of phenotyping models.** *rf*: Random Forest, *ab*= Adaboost, *lr*=Logistic regression with L1 penalty, *gb*=Gradient Boosting, *en*=Logistic Regression with elastic net penalty models. 95% confidence intervals

5.3.2 Variants recovered by QTPhenProxy for all stroke, ischemic stroke, subarachnoid hemorrhage, intracerebral hemorrhage, and improvement over traditional binary method using the QC1 markers and principal components

The binary trait logistic regression genome-wide association study for stroke, ischemic stroke, and subarachnoid hemorrhage recovered no variants with genome-wide significance. Binary trait logistic regression genome-wide association study for intracerebral hemorrhage recovered 2 variants with genome-wide significance. Quantitative trait linear regression using QTPhenProxy probabilities for stroke recovered 120 genome-wide significant SNPs with 16 LD-independent loci and 1215 genome-wide significant SNPs with 39 LD-independent loci, for ischemic stroke recovered 202 genome-wide significant SNPs with 15 LD-independent loci, and 1266 genome-wide significant SNPs with 46 LD-independent loci, for subarachnoid hemorrhage recovered 69 genome-wide

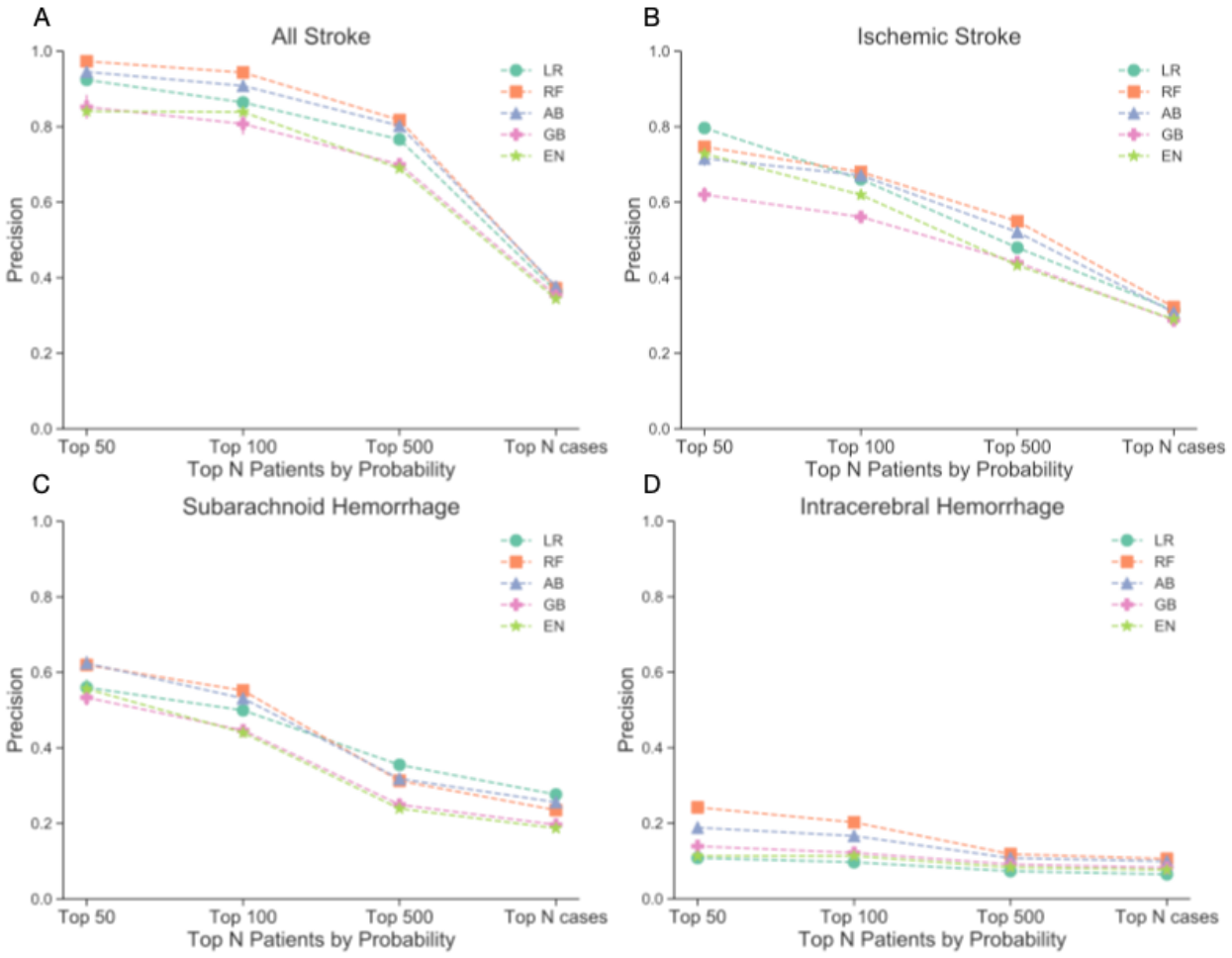


Figure 5.1: **Precision at top 50, 100, 500, and N cases probabilities assigned by machine learning algorithms on hold out test set.** A. All Stroke, B. Ischemic Stroke, C. Subarachnoid Hemorrhage, D. Intracerebral Hemorrhage. Circles represent precision for the Logistic Regression model with L1 penalty, squares: Random Forest model, triangles: Adaboost model, cross: Gradient boosting model, and star Logistic Regression model with Elastic Net penalty.

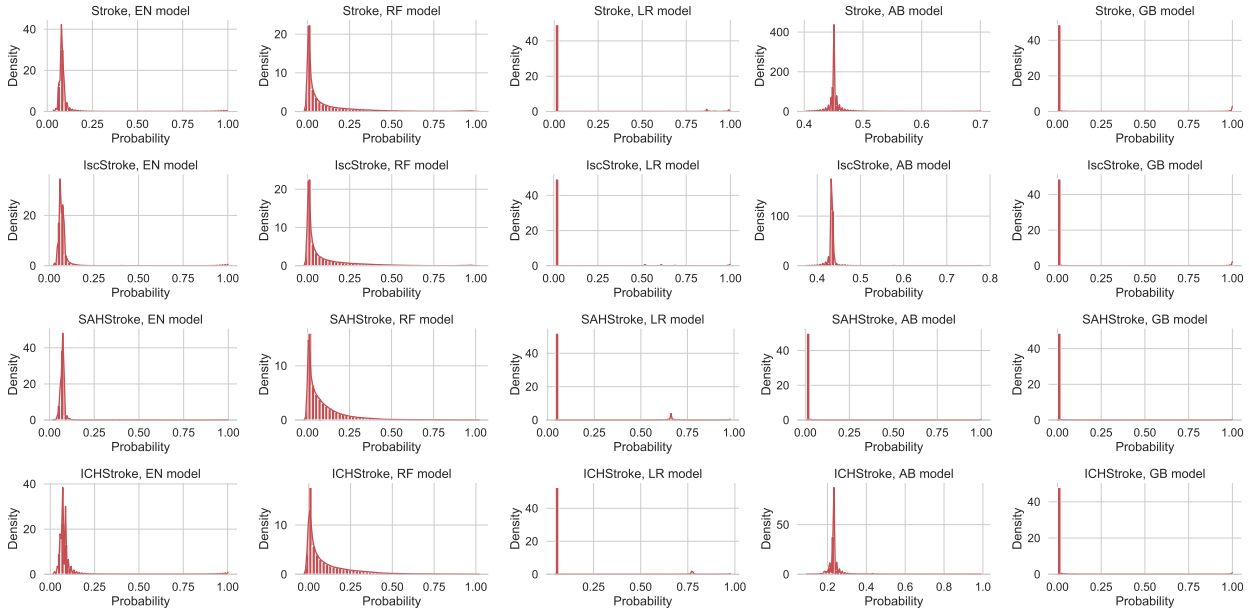


Figure 5.2: **Model probability distributions assigned by machine learning algorithms.** *RF*: Random Forest model, *EN*: Logistic Regression with elastic net penalty model, *LR*: Logistic Regression model with L1 penalty, *AB*: Adaboost model,*GB*: Gradient boosting model, *IscStroke*: Ischemic Stroke phenotype,*SAHStroke*: Subarachnoid hemorrhage phenotype,*ICHStroke*: Intracerebral hemorrhage phenotype

significant SNPs with 4 LD-independent loci, and 722 genome-wide significant SNPs with 40 LD-independent loci, and for intracerebral hemorrhage recovered 146 genome-wide significant SNPs with 8 LD-independent loci, and 1059 genome-wide significant SNPs with 54 LD-independent loci using the EN and RF classifiers, respectively (Table 5.4). We show the comparison of genome-wide significant variants between the QTPhenProxy EN model and Binary trait GWAS for stroke in a Hudson plot (Figure 5.3)[160]. Out of known ischemic stroke variants, both models also recovered 3 known variants with genome-wide significance, and the EN model recovered 21/49 of the known variants, equivalent to a sensitivity of 0.428, while the traditional binary trait recovered 15/49, (sensitivity=0.306), and RF model recovered 14/49 (sensitivity=0.286) with nominal p-value of 0.05 (Table 5.5). For all stroke, sensitivity of known stroke variants was 0.333 using the EN model, 0.261 using the RF model, and 0.202 using the binary trait. Subarachnoid hemorrhage did not have a specific EBI-GWAS disease marker set, and the EBI-GWAS disease marker set for intracerebral hemorrhage only consisted of one variant. The difference in p-values between the

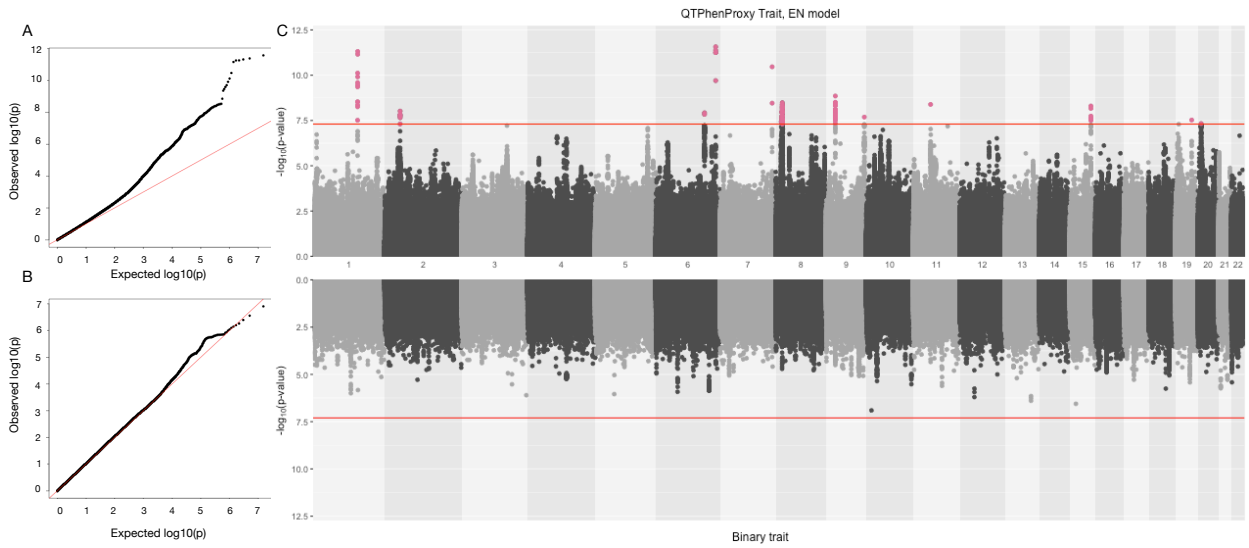


Figure 5.3: **QQ Plots and Hudson plot of QTPhenProxy genome-wide association analysis, EN Model with Binary trait genome-wide association analysis for Stroke, using QC1 quality control. A. Q-Q Plot for QTPhenProxy, EN Model GWAS.** B. Q-Q Plot for Binary trait GWAS. C. Top Manhattan plot is for QTPhenProxy, bottom plot for binary trait. Variants with p-value < 5e-08 are highlighted in pink, and the dashed lines are at the same value.

binary method and QTPhenProxy method for known ischemic stroke variants was significantly increased compared to the same number of random variants using the EN classifier (two sample t test, t=2.43, p=0.0184 for EN classifier and t=1.74, p=0.0876 for RF classifier). In addition, QTPhenProxy with EN classifier showed a significant decrease in p-value of all known ischemic stroke variants compared to traditional binary method (two sample t-test, t=-2.1 p=0.0367) while QTPhenProxy with RF classifier did not show a significant decrease (t=-1.48,p=0.144). The difference in p-values between the binary method and QTPhenProxy method for known all stroke variants was significantly increased compared to the same number of random variants for the EN classifier (t=2.80, p=0.00638) but not significant for the RF classifier (t=0.737,p=0.463). In addition, QTPhenProxy with EN classifier showed a decrease in p-value of all known stroke variants compared to traditional binary method (two sample t-test, t=-1.87 p=0.0639) while QTPhenProxy with RF classifier did not show a significant decrease (t=-1.05,p=0.298).

Disease	N Cases	Bin Hits	RF Hits	RF AUROC	EN Hits	EN AUROC
Asthma	21491	8548	3946	0.805	3117	0.79
COPD	8290	82	2544	0.915	2088	0.906
All cause Dementia	2279	36	550	0.85	91	0.878
Alzheimer's disease	357	72	1356	0.876	196	0.856
Vascular dementia	178	34	1832	0.93	611	0.919
Frontotemporal Dementia	39	6	1655	NA	0	NA
Motor neuron disease	173	1	1652	0.927	48	0.908
Myocardial Infarction	9407	487	2126	0.935	940	0.931
STEMI	3386	82	1806	0.977	563	0.974
NSTEMI	3932	137	2922	0.943	459	0.94
Parkinsonism	2321	3	881	0.908	41	0.897
Parkinson's disease	934	3	997	0.885	3	0.87
Progressive supranuclear palsy	14	34	3789	NA	40	NA
Multiple System Atrophy	120	14	1386	0.952	15	0.952
Stroke	4354	0	1215	0.899	120	0.896
Ischemic stroke	3308	0	1266	0.952	202	0.943
Intracerebral hemorrhage	581	2	1059	0.903	146	0.903
Subarachnoid hemorrhage	665	0	722	0.875	69	0.862

Table 5.4: **Number of genome-wide significant variants.** *N Cases*: number of cases. *hits*: Number of genome-wide significant variants, *Bin*: binary trait, *RF*: Random Forest model, *EN*: Logistic Regression with elastic net penalty model.

Disease	Bin Sens	RF Sens	EN Sens
Asthma	0.667	0.458	0.513
COPD	0.254	0.182	0.2
All cause Dementia	0	1	1
Alzheimer's disease	0.285	0.155	0.173
Myocardial Infarction	0.789	0.71	0.736
Parkinsonism	0.473	0.231	0.0842
Stroke	0.217	0.217	0.261
Ischemic stroke	0.312	0.312	0.333

Table 5.5: **Proportion of known stroke variants that reach nominal significance for each model.** *N Cases*: number of cases, *sens*: sensitivity measure for each model GWAS and binary trait GWAS, measures what proportion of known stroke variants reach genome-wide significance in each association test. *Bin*: binary trait, *RF*: Random Forest model, *EN*: Logistic Regression with elastic net penalty model.

5.3.3 Variants recovered by QTPhenProxy for all stroke, ischemic stroke, subarachnoid hemorrhage, intracerebral hemorrhage, and improvement over traditional binary method using the QC2 markers and principal components

The binary trait logistic regression genome-wide association study for stroke, ischemic stroke, and subarachnoid hemorrhage recovered no variants with genome-wide significance. Binary trait logistic regression genome-wide association study for intracerebral hemorrhage recovered 2 variants with genome-wide significance. Quantitative trait linear regression using QTPhenProxy probabilities for stroke recovered 7 LD-independent loci, 3 LD-independent loci, for ischemic stroke, 3 LD-independent loci for subarachnoid hemorrhage, and 3 LD-independent loci for intracerebral hemorrhage using the EN classifier. We show the comparison of genome-wide significant variants between the EN and Binary in a Hudson plot (Figure 5.4)[160]. Out of known ischemic stroke variants, both models also recovered 3 known variants with genome-wide significance, and the EN model recovered 16/49 of the known variants, equivalent to a sensitivity of 0.326, while the traditional binary trait recovered 15/49, (sensitivity=0.306), and RF model recovered 15/49 (sensitivity=0.306) with nominal p-value of 0.05 (Figure 5.5). For all stroke, sensitivity of known stroke variants was 0.265 using the EN model, 0.220 using the RF model, and 0.220 using the binary trait. The t-tests referred to in Result 5.3.2 did not show significance.

5.3.4 Conditional analysis refines candidate variants to mostly lead some nearby SNPs.

Conditional analysis of the GWAS with QC1 quality control for the QTPhenProxy EN model for all stroke identified 13 candidate variants with genome-wide significance, all of which are novel, though some of the nearest genes to these loci have been identified in previous studies through nearby loci. There were 12 loci identified for the stroke subtype ischemic stroke, 3 loci for subarachnoid hemorrhage, and 9 loci for intracerebral hemorrhage (Table 5.6). Several loci overlapped across stroke and some subtypes. Position 46705193 on chromosome 11, which is in an intronic section of *ARHGAP1*[152], showed genome-wide significance in stroke, ischemic stroke, and intracerebral hemorrhage. A missense mutation at locus rs6025 in the *F5* gene showed

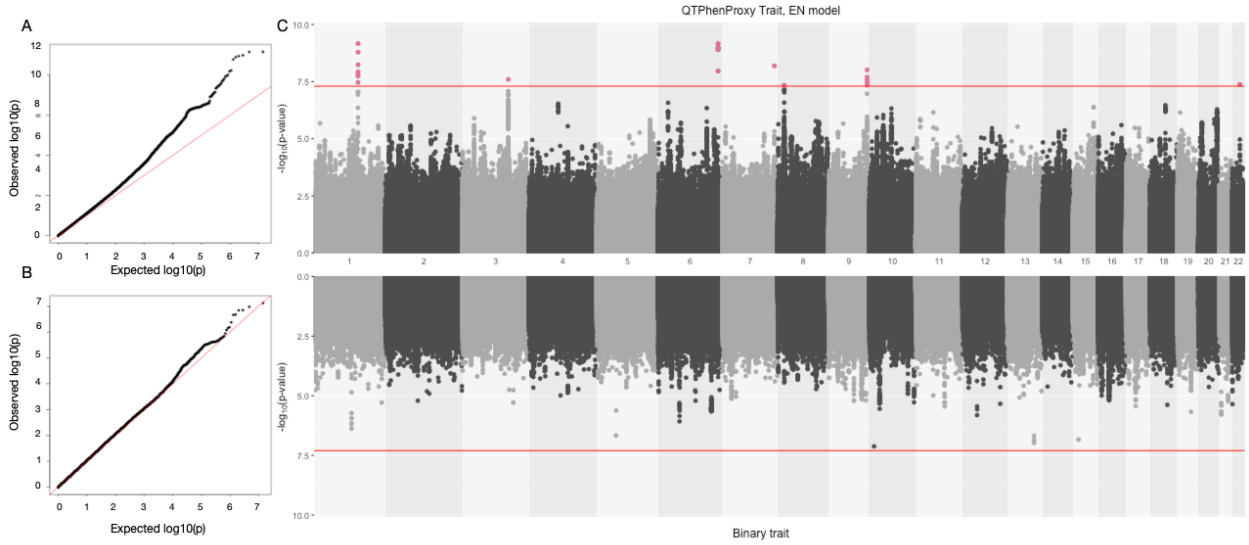


Figure 5.4: **QQ Plots and Hudson plot of QTPhenProxy genome-wide association analysis, EN Model with Binary trait genome-wide association analysis for Stroke, using QC2 quality control.** QQ Plots and Hudson plot of QTPhenProxy genome-wide association analysis, EN Model with Binary trait genome-wide association analysis for Stroke, using QC2 quality control. QQ Plots and Hudson plot of QTPhenProxy genome-wide association analysis, EN Model with Binary trait genome-wide association analysis for Stroke, using QC2 quality control. A. Q-Q Plot for QTPhenProxy, EN Model GWAS. B. Q-Q Plot for Binary trait GWAS. C. Top Manhattan plot is for QTPhenProxy, bottom plot for binary trait. Variants with p-value < 5e-08 are highlighted in pink, and the dashed lines are at the same value.

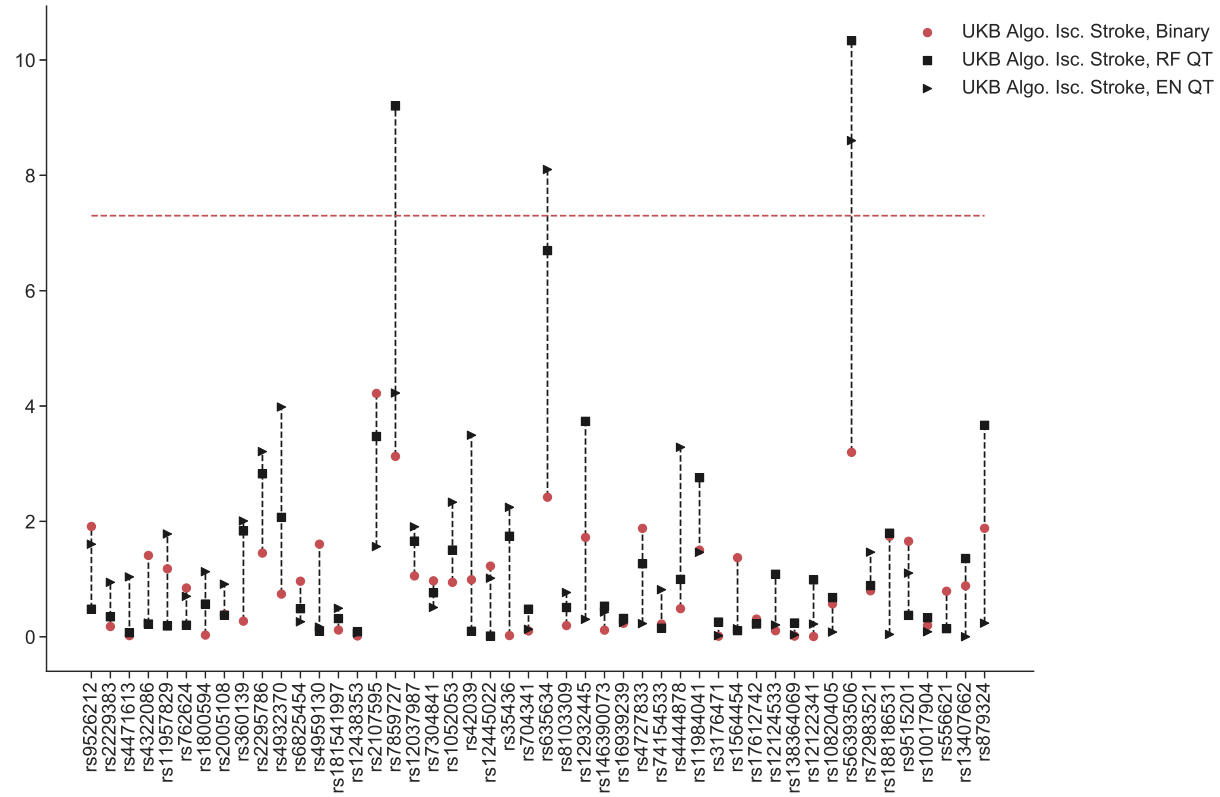


Figure 5.5: **QTPhenProxy recovers known ischemic stroke variants. Results from QTPhenProxy GWAS with QC2 quality control.** Horizontal axis shows Ischemic Stroke variants catalogued in EBI-GWAS that have shown genome-wide significance in previous studies. Black markers represent p-values of variants recovered by QTPhenProxy models (square=RF,Triangle=EN).

genome-wide significance across stroke, intracerebral hemorrhage, and locus rs1894692, which is near the *F5* gene[152], showed genome wide significance in ischemic stroke. Additional loci that showed genome-wide significance in both all stroke and ischemic stroke included those that are intronic to *LPA* and *NOS3* and nearby to *APOC1P1*/*APOC1* and *CDKN2A*/*CDKN2B*[152]. Finally, all stroke and intracerebral hemorrhage shared additional loci that showed genome-wide significance after conditional analysis, which were nearby *LOC105377992* and *RPS4X9*[152]. All stroke genome-wide significant variants also included intronic loci in the *MTA3*, *SOX7*, *ABO*, *FURIN*, and *PLCB1* genes[152]. Ischemic stroke genome-wide significant variants included intronic loci in the *LPAL2* gene, intergenic loci in the *PITX2*, *LDLR*, and an insertion mutation in the *ABO* gene[152]. Subarachnoid hemorrhage genome-wide significant variants included an intronic locus in the *ABO* gene. Finally, intracerebral hemorrhage genome-wide significant variants included a missense mutation loci in the *ABCG8* gene, and intronic loci in the *NOS3* and *XKR6* genes[152]. The variants near or in *LOC105377992*, *NOS3*, *CDKN2A*/*CDKN2B*, *LPA*, *LDLR*, *ABCG8*, and *RPS4X9* were below nominal significance of 0.05 in the MEGASTROKE stroke and ischemic stroke GWAS of European ancestry (Table 5.7)[11].

Phenotype	rsID	Chr.	Gene	Location	Allele	RAF	Beta	Converted OR	P value
Stroke	rs6025	1q24.2	*F5*	missense	T/C	0.0231	0.0492 (0.0421-0.0564)	1.09 (1.08-1.10)	**4.87e-12**
Stroke	rs8179838	2p21	*MTA3*	intronic	T/C	0.621	0.0126 (0.0104-0.0149)	1.02 (1.02-1.03)	**9.63e-09**
Stroke	rs6923947	6q22.32	*LOC105377992*	intergenic	A/G	0.447	0.0123 (0.0101-0.0144)	1.02 (1.02-1.02)	**1.19e-08**
Stroke	rs55730499	6q25.3	*LPA*	intronic	T/C	0.0817	0.0273 (0.0234-0.0312)	1.05 (1.04-1.06)	**2.70e-12**
Stroke	rs3918226	7q36.1	*NOS3*	intronic	T/C	0.0814	0.0263 (0.0223-0.0302)	1.05 (1.04-1.05)	**3.44e-11**
Stroke	rs6991641	8p23.1	*SOX7*	intronic	G/C	0.403	0.013 (0.0108-0.0152)	1.02 (1.02-1.03)	**3.29e-09**
Stroke	rs1333049	9p21.3	*CDKN2A*	intergenic	C/G	0.481	0.0129 (0.0108-0.0151)	1.02 (1.02-1.03)	**1.41e-09**
Stroke	9:136138765	9q34.3	*ABO*	intronic	G/GC...	0.184	0.0155 (0.0128-0.0183)	1.03 (1.02-1.03)	**2.05e-08**
Stroke	11:46705193	11p11.2	*ARHGAP1*	intronic	C/CT	0.28	0.0143 (0.0119-0.0167)	1.02 (1.02-1.03)	**4.10e-09**
Stroke	rs59065675	15q26.1	*FURIN*	intronic	C.../C	0.47	0.0127 (0.0105-0.0149)	1.02 (1.02-1.03)	**5.00e-09**
Stroke	rs814573	19q13.32	*APOC1*	intergenic	T/A	0.188	0.0156 (0.0128-0.0184)	1.03 (1.02-1.03)	**2.97e-08**
Stroke	20:862958	20p12.3	*PLCB1*	intronic	TA/T	0.76	0.0138 (0.0113-0.0164)	1.02 (1.02-1.03)	**4.73e-08**
IscStroke	rs1894692	1q24.2	*F5*	intergenic	G/A	0.0218	0.0546 (0.0478-0.0614)	1.1 (1.09-1.11)	**8.22e-16**
IscStroke	rs369787256	4q25	*PITX2*	intergenic	C/T	0.0905	0.0223 (0.0188-0.0257)	1.04 (1.03-1.05)	**1.17e-10**
IscStroke	rs55730499	6q25.3	*LPA*	intronic	T/C	0.0817	0.0272 (0.0236-0.0307)	1.05 (1.04-1.05)	**2.49e-14**
IscStroke	rs117733303	6q25.3	*LPAL2*	intronic	G/A	0.0187	0.0428 (0.0356-0.05)	1.08 (1.06-1.09)	**2.69e-09**

IscStroke	rs3918226	7q36.1	*NOS3*	intronic	T/C	0.0814	0.0209 (0.0173-0.0246)	1.04 (1.03-1.04)	**6.81e-09**
IscStroke	rs2205258	8q23.3	*LOC107986968*	intronic	C/T	0.83	0.0151 (0.0125-0.0177)	1.03 (1.02-1.03)	**6.76e-09**
IscStroke	rs8176719	9q34.2	*ABO*	insertion	TC/T	0.34	0.0145 (0.0124-0.0165)	1.02 (1.02-1.03)	**2.17e-12**
IscStroke	rs1333049	9p21.3	*CDKN2A*	intergenic	C/G	0.481	0.0132 (0.0112-0.0151)	1.02 (1.02-1.03)	**1.31e-11**
IscStroke	11:46705193	11p11.2	*ARHGAP1*	intronic	C/CT	0.28	0.0122 (0.01-0.0145)	1.02 (1.02-1.02)	**3.22e-08**
IscStroke	rs814573	19q13.32	*APOC1*	intergenic	T/A	0.188	0.0172 (0.0146-0.0198)	1.03 (1.02-1.03)	**1.85e-11**
IscStroke	rs118068660	19p13.2	*LDLR*	intergenic	C/T	0.901	0.0193 (0.0159-0.0227)	1.03 (1.03-1.04)	**1.43e-08**
IscStroke	rs12151925	20p12.2	*FAT1P1*	intergenic	C/T	0.222	0.0129 (0.0105-0.0152)	1.02 (1.02-1.03)	**4.26e-08**
SAHStroke	rs6025	1q24.2	*F5*	missense	T/C	0.0231	0.0673 (0.0593-0.0753)	1.12 (1.11-1.14)	**4.61e-17**
SAHStroke	rs36058710	9q34.2	*ABO*	intronic	CT/C	0.287	0.0166 (0.0139-0.0194)	1.03 (1.02-1.03)	**1.03e-09**
SAHStroke	rs34850248	12q12	*LINC02400*	intergenic	A/AC	0.0714	0.0268 (0.0221-0.0314)	1.05 (1.04-1.06)	**9.95e-09**
ICHStroke	rs6025	1q24.2	*F5*	missense	T/C	0.0231	0.0480 (0.0411-0.0549)	1.09 (1.07-1.1)	**2.85e-12**
ICHStroke	rs11887534	2p21	*ABCG8*	missense	G/C	0.935	0.0271 (0.0229-0.0313)	1.05 (1.04-1.06)	**7.90e-11**
ICHStroke	rs372302634	3q22.3	*ESYT3-MRAS*	intergenic	T/T...	0.457	0.0121 (0.01-0.0142)	1.02 (1.02-1.02)	**8.51e-09**
ICHStroke	rs6923947	6q22.32	*LOC105377992*	intergenic	A/G	0.447	0.0127 (0.0106-0.0148)	1.02 (1.02-1.03)	**9.5e-10**
ICHStroke	rs2077111	6q22.33	*RPS4XP9*	intergenic	G/A	0.447	0.012 (0.00995-0.0141)	1.02 (1.02-1.02)	**8.89e-09**
ICHStroke	rs1808593	7q36.1	*NOS3*	intronic	G/T	0.25	0.0131 (0.0107-0.0155)	1.02 (1.02-1.03)	**4.14e-08**
ICHStroke	rs7013277	8p23.1	*XKR6*	intronic	C/A	0.396	0.0116 (0.00947-0.0137)	1.02 (1.02-1.02)	**4.6e-08**
ICHStroke	11:46705193	11p11.2	*ARHGAP1*	intronic	C/CT	0.28	0.0133 (0.011-0.0156)	1.02 (1.02-1.03)	**1.35e-08**

Table 5.6: **Genome-wide significant variants discovered by QTPhenProxy, EN Model, using QC1 quality control.** Cytogenic position was determined using [153]. *rsID*: variant id, *Chr*: cytogenic position, *Location*: location of variant relative to the gene, *Allele*: risk/reference alleles, *RAF*: risk allele frequency in population, *Beta*: Beta coefficient *OR*: Odds Ratio, 95% Confidence Intervals

5.3.5 Conditional analysis refines candidate variants to mostly lead some nearby SNPs.

Conditional analysis of the GWAS with QC2 quality control for the QTPhenProxy EN model for all stroke identified 7 candidate variants with genome-wide significance, all which are novel, though some of the nearest genes to these loci have been identified in previous studies through nearby loci. There were 3 loci identified for the stroke subtype ischemic stroke, 3 loci for subarachnoid hemorrhage, and 3 loci for intracerebral hemorrhage (Table 5.7). Almost all of the genome-wide significant variants overlapped with those found running the GWAS using QC1 quality control. New variants that were genome-wide significant for all stroke included a dele-

tion variant that was intergenic to the gene *ATP5MC1P3*, an intronic variant of the *MRSA* gene, and a 3'UTR variant in *KCNJ4*[152]. In addition, several genome-wide significant variants in the MEGASTROKE stroke and ischemic stroke GWAS were near the variants that were genome-wide significant in the QTPhenProxy EN GWAS. Almost all of those nearby MEGASTROKE genome-wide significant variants replicated to at least nominal significance of 0.05 in the QTPhenProxy EN GWAS (Table 5.8).

Phenotype	rsID	Chr.	Gene	P value	MEGASTROKE P value
Stroke	rs1894692	1q24.2	*F5-SLC19A2*	**6.87E-10**	0.435
Stroke	rs8179838	2p21	*MTA3*	8.36E-06	0.515
Stroke	rs769407520	3q22.3	*ATP5MC1P3*	**2.50E-08**	not found
Stroke	rs6923947	6q22.32	*LOC105377992*	4.58E-07	0.04332
Stroke	rs55730499	6q25.3	*LPA*	**6.85E-10**	0.05915
Stroke	rs3918226	7q36.1	*NOS3*	**6.45E-09**	0.00134
Stroke	8:10206921	8p23.1	*MSRA*	**4.62E-08**	not found
Stroke	rs1333049	9p21.3	*CDKN2A-CDKN2B*	5.02E-06	1.85E-07
Stroke	9:136138765	9q34.3	*ABO*	**9.73E-09**	not found
Stroke	11:46705193	11p11.2	*ARHGAP1*	7.06E-07	not found
Stroke	rs59065675	15q26.1	*FURIN*	4.06E-07	not found
Stroke	rs814573	19q13.32	*APOC1-APOC1P1*	2.52E-05	0.412
Stroke	20:862958	20p12.3	*PLCB1*	1.95E-06	not found
Stroke	rs2269608	22q13.1	*KCNJ4*	**4.23E-08**	0.212
IscStroke	rs1894692	1q24.2	*F5-SLC19A2*	**1.90E-11**	0.229
IscStroke	rs369787256	4q25	*PITX2-MIR297*	3.69E-06	not found
IscStroke	rs55730499	6q25.3	*LPA*	**1.26E-09**	0.0316
IscStroke	rs117733303	6q25.3	*LPAL2*	0.000119	0.422
IscStroke	rs3918226	7q36.1	*NOS3*	3.59E-06	0.00118
IscStroke	rs2205258	8q23.3	*LOC107986968*	7.16E-06	0.415
IscStroke	9:1361387	9q34.2	*ABO*	**4.99E-10**	not found
IscStroke	rs1333049	9p21.3	*CDKN2A-CDKN2B*	3.01E-06	1.09E-06
IscStroke	11:46705193	11p11.2	*ARHGAP1*	3.85E-05	not found

IscStroke	rs814573	19q13.32	*APOC1-APOC1P1*	2.20E-06	0.716
IscStroke	rs118068660	19p13.2	*LDLR*	1.28E-06	1.14E-05
IscStroke	rs12151925	20p12.2	*LOC101929413-FAT1P1*	0.000107	0.104
SAHStroke	rs6025	1q24.2	*F5*	**6.34E-18**	0.453
SAHStroke	rs36058710	9q34.2	*ABO*	**1.59E-10**	not found
SAHStroke	rs34850248	12q12	*LINC02400*	**7.18E-09**	not found
ICHStroke	rs6025	1q24.2	*F5*	**1.52E-10**	0.453
ICHStroke	rs11887534	2p21	*ABCG8*	**4.29E-08**	6.85E-04
ICHStroke	rs372302634	3q22.3	*ESYT3-MRAS*	**4.03E-08**	not found
ICHStroke	rs6923947	6q22.32	*LOC105377992*	1.47E-07	0.00433
ICHStroke	rs2077111	6q22.33	*RPS4XP9*	1.96E-06	0.0372
ICHStroke	rs1808593	7q36.1	*NOS3*	2.96E-06	0.565
ICHStroke	rs7013277	8p23.1	*XKR6*	2.06E-05	0.156
ICHStroke	11:46705193	11p11.2	*ARHGAP1*	3.59E-06	not found

Table 5.7: **Genome-wide significant variants discovered by QTPhenProxy, EN Model, using QC2 quality control.** Also includes QC2 p-value for genome-wide significant variants from QC1 GWAS and p-values for variants from the MEGASTROKE stroke or ischemic stroke european GWAS [11]. Cytogenic position was determined using [153]. *rsID*: variant id, *Chr*: cytogenic position

5.3.6 Correlation between effect sizes of QTPhenProxy and traditional binary trait analysis

We determined the correlation between the effect sizes of the binary trait GWAS and QTPhen-Proxy GWAS. Pearson correlation of the beta coefficients and log of the odds ratios increased when restricting to variants with small p-values (max r^2=0.70) (Figure 5.9). Too few variants, such as with p-values > 5e-08 resulted in a decreased Pearson correlation.

Phenotype	Gene	rsID	QTPhenProxy p-value
Stroke	*LPA*	rs56393506	1.09E-08
Stroke	*NOS*	rs1799983	not found
Stroke	*CDKN2B-CDKN2A*	rs7859727	6.94E-05
Stroke	*ABO*	rs635634	1.07E-07
Stroke	*FURIN*	rs4932370	2.55E-05
Stroke	*COL4A1*	rs9521634	0.00375
Stroke	*LOC105372798*	rs720470	0.88
Ischemic stroke	*PITX2-MIR297*	rs13143308	not found
Ischemic stroke	*LPA*	rs56393506	2.50E-09
Ischemic stroke	*NOS3*	rs1799983	not found
Ischemic stroke	*CDKN2B-CDKN2A*	rs7859727	5.95E-05
Ischemic stroke	*ABO*	rs635634	7.93E-09
Ischemic stroke	*LDLR*	rs8103309	0.171
Subarachnoid Hemorrhage	*ABO*	rs635634	2.22E-06
Intracerebral Hemorrhage	*NOS3*	rs1799983	not found

Table 5.8: **QTPhenProxy EN Model GWAS using QC2 quality control P-value of variants that were genome-wide significant in MEGASTROKE Stroke and Ischemic Stroke GWAS.**

5.3.7 QTPhenProxy results for other diseases.

For myocardial infarction, traditional binary trait method recovered 487 variants while quantitative trait linear regression using QTPhenProxy probabilities recovered 940 genome-wide significant SNPs and 2126 genome-wide significant hits using the EN and RF classifiers, respectively. Out of the 487 variants determined by traditional methods, 204 variants overlapped in the EN QTPhenProxy linear regression. In contrast, for COPD, traditional binary trait method recovered 82 variants while QTPhenProxy method recovered 2544 and 2088 variants using RF and EN classifiers, respectively. Out of the 82 variants determined by traditional method, none were recovered using the QTPhenProxy method.

5.3.8 Specificity analysis of genome-wide significant variants using EBI-GWAS marker sets

From the over 2,000 EBI-GWAS disease variant marker sets mapped to organ systems, we calculated the proportion of markers in each set that were found to be at genome-wide significance

by our QTPhenProxy model with QC2 markers and principal components. We found that the organ systems with markers sets with the highest proportion of genome-wide significant variants for stroke included vascular disorders, investigations, psychiatric disorders, and general disorders and administration site conditions (Figure 5.6). The enriched disease marker sets in the Investigations class included the lab values lipoprotein A levels, lipoprotein a levels adjusted for apolipoprotein A isoforms, blood protein levels, and white blood cell counts. The enriched disease marker sets in the General disorders and administration site conditions included aortic valve stenosis and allergy, while the Psychiatric disorders enriched marker sets were response to statins (LDL change) and venous thromboembolism. For ischemic stroke, the top enriched disease marker sets included those described for stroke and also activated partial thromboplastin time, coagulation factor levels, protein biomarkers, soluble levels of adhesion molecules, pancreatic cancer, and urinary metabolites.

We also found GWAS results from diseases other than stroke to be enriched with corresponding EBI-GWAS disease marker sets. QTPhenProxy EN and RF Model GWAS using QC1 quality control showed increased proportion of significant variants for myocardial infarction (MI), COPD, and Asthma in corresponding EBI-GWAS disease and System Organ Class marker sets compared to all other marker sets, but not in Alzheimer's or Parkinson's disease. Representative plots (Figure 5.7 for MI, COPD, and Parkinson's EN show increased median proportion significant variants in disease and system organ class marker sets compared to other marker sets for MI and COPD but not Parkinson's disease. This difference is seen when the threshold for significance is 0.05 or 0.005.

5.3.9 LD score regression intercept, genomic inflation, and evaluation of genomic inflation

Using QC1 quality control, the genomic inflation for QTPhenProxy EN model for Stroke was 1.133, while using the QC2 quality control the genomic inflation for the same model was 1.134. For binary traits using QC2 quality control, genomic inflation for stroke was 1.027. Using the QTPhenProxy EN model with QC2 quality control, the genomic inflation for Ischemic stroke was

Figure 5.6: **Disease categories that are enriched for variants discovered by QTPhenProxy genome-wide association study of Stroke.** X-axis is the top percentile of marker sets in each category, Y-axis is the proportion of variants in marker sets that overlap with QTPhenProxy genome-wide significant variants. Each shape corresponds to a disease category. In color are top disease categories: Square=Investigations, Triangle=Vascular disorders, Cross=Nervous system disorders, X=General disorders and administration site conditions, Star=cardiac disorders, and Diamond=Blood and lymphatic system disorders.

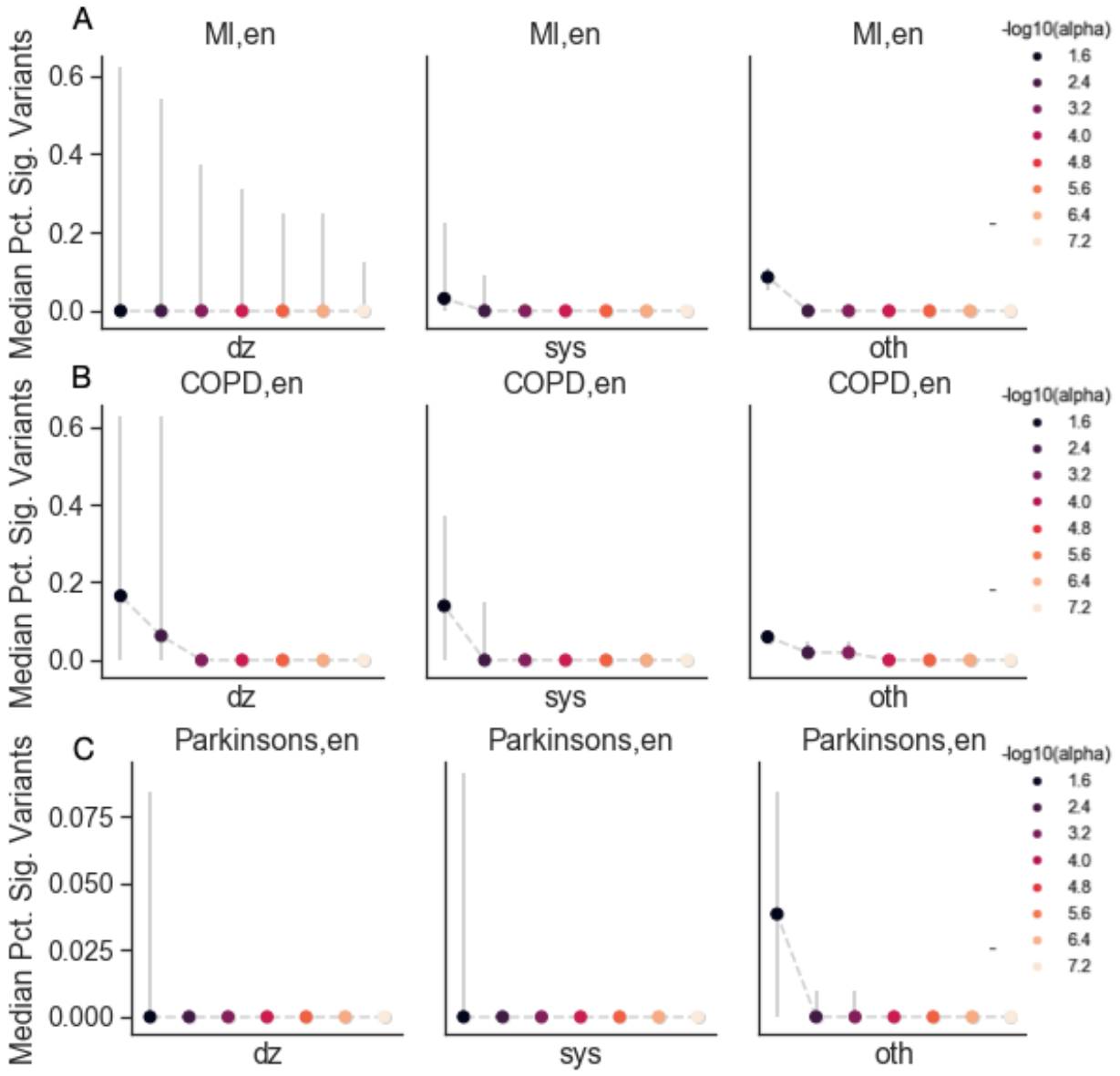


Figure 5.7: **Median proportion of significant variants stratified by disease, system organ class, and other disease markers for QTPhenProxy EN model genome-wide association study of MI, COPD, and Parkinson's Disease with QC1 quality control.** Colors correspond to negative log10 of p-value significance threshold (-log10(alpha)). Panel A corresponds to Myocardial Infarction (MI) GWAS, Panel B, Chronic Obstructive Pulmonary Disease (COPD) GWAS, and Panel C, Parkinson's Disease. *en*= logistic regression model with elastic net penalty. First column of each panel represents disease (dz) marker sets, second column, System Organ Class (sys) marker sets, and third column other (oth) disease marker sets.

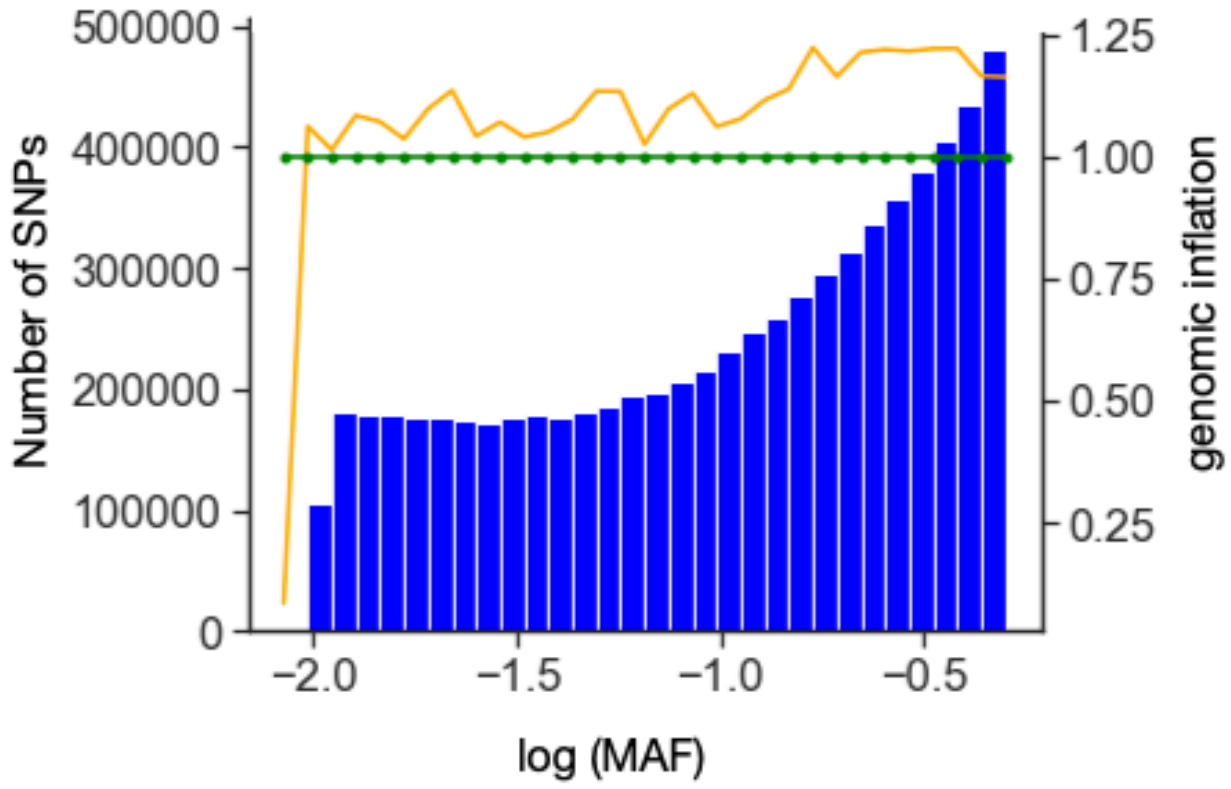


Figure 5.8: **Genomic Inflation within bins of variants with similar minor allele frequencies for QTPhenProxy EN Model for Stroke with QC2 quality control.** Orange line shows the genomic inflation of each bin of variants, green line shows 1.0 genomic inflation, and blue bars show numbers of SNPs (variants) in each bin, binned by log(MAF), or log(minor allele frequency.)

1.118, for subarachnoid hemorrhage was 1.122, and for intracerebral hemorrhage was 1.114. LD score regression intercept for QTPhenProxy EN model with QC2 quality control for stroke was 1.037 ± 0.0088, for ischemic stroke, 1.031 ± 0.0077, for subarachnoid hemorrhage, 1.012 ± 0.008, and for intracerebral hemorrhage, 1.051 ± 0.0081. We found that more common variants, or those with higher minor allele frequencies, had higher genomic inflation than rarer variants (Figure 5.8).

5.3.10 Genetic Correlation of QTPhenProxy with MEGASTROKE and Coronary Artery Disease GWAS

We measured a genetic correlation of 0.64 (p-value 1.8E-23) between the QTPhenProxy EN Model for stroke, QC1 quality control, and the MEGASTROKE all stroke GWAS and a genetic

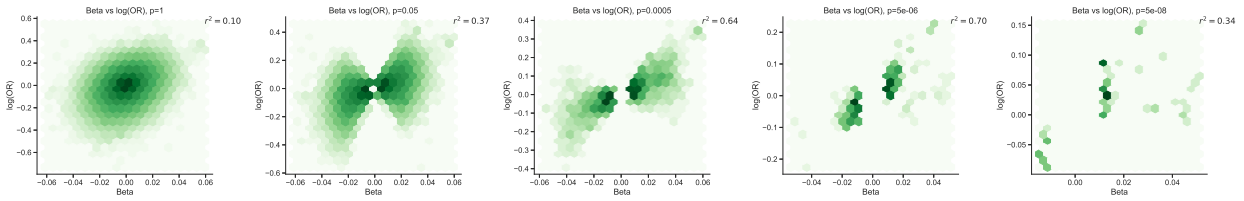


Figure 5.9: **Correlation between QTPhenProxy, EN Model GWAS and Binary trait GWAS effect size.** Pearson correlation is recorded in top right corner. From left to right, variants included decreases by restricting p-value.

correlation of 0.60 (p-value 1.4E-19) between the QTPhenProxy EN Model for stroke, QC2 quality control, and the MEGASTROKE all stroke GWAS. We also found a genetic correlation of -0.1 (p-value 0.0065) between the MEGASTROKE all stroke GWAS and coronary artery disease GWAS, and we found a genetic correlation of -0.0048 (p-value 0.88) between the QTPhenProxy EN Model for stroke, QC1 quality control, and coronary artery disease GWAS.

5.3.11 Simulation of Conversion of Quantitative trait to Binary trait shows similar correlation of effect sizes to empirical data.

We simulated the effect of a quantitative trait locus and nearby marker both on the original quantitative trait and the binary trait converted from the quantitative trait using liability thresholding. We found high correlation of effect sizes between the beta coefficients and log odds ratios of the quantitative trait variants with p-value < 0.005 and their respective binary trait variant effect sizes (Pearson correlation, r^2=0.82) (Figure5.9). Lower p-values were not tested because of too few qualifying variants. After standard normalizing the quantitative trait, we also found the slope of its correlation with the log binary trait to be stable across the parameters, with a mean of 1.63 for each parameter except for marker allele frequency (1.35) and prevalence (1.77) (Figure 5.10).

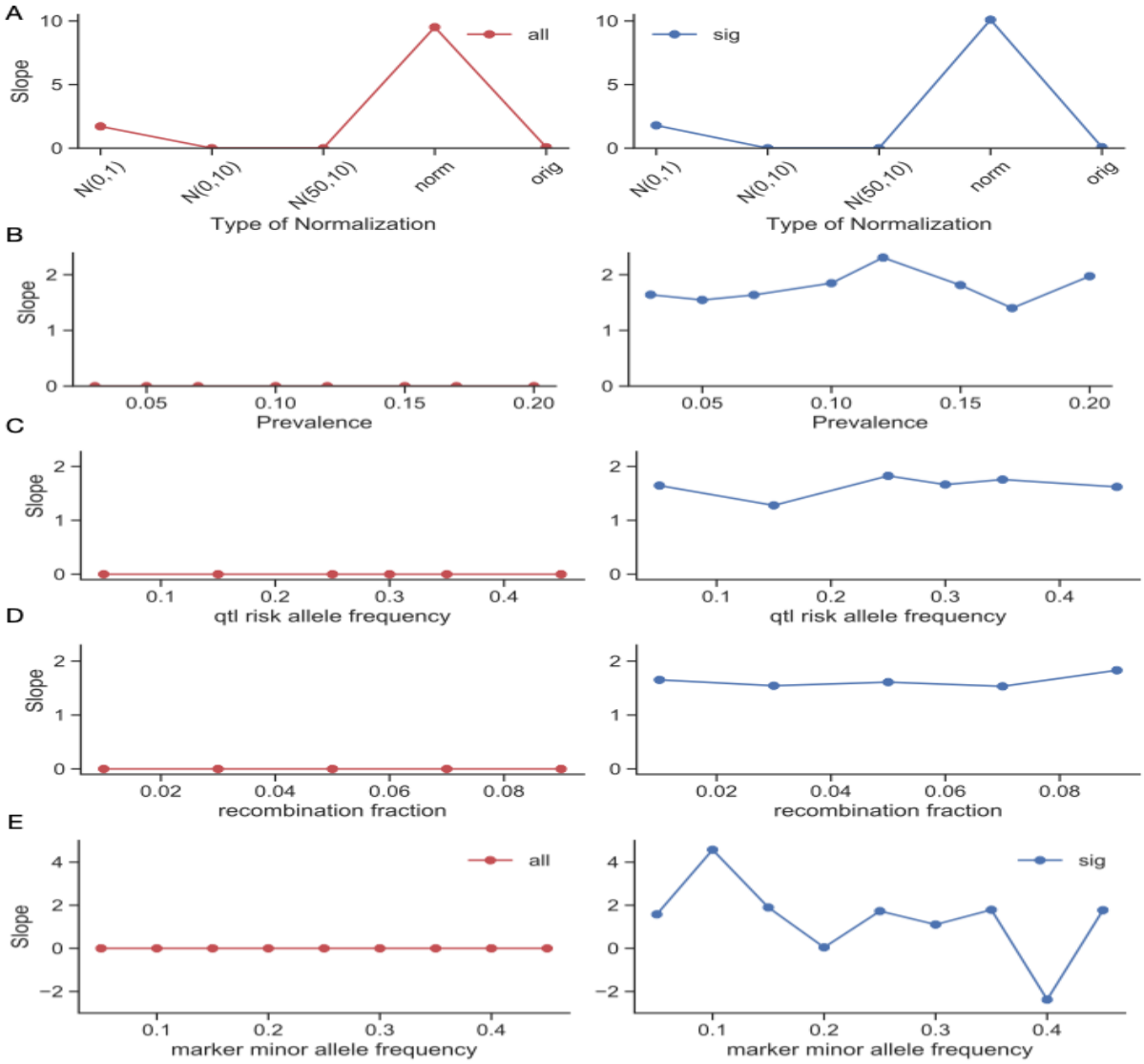


Figure 5.10: **Slope of correlation between beta from simulated quantitative trait and log(odds ratio) of simulated binary trait with varying simulation parameters.** Left panels calculate slope with all variants, right panels calculate slope with variants with p-value <0.005. (A) varies the transformation of the probability distribution, where *N(0,1)* is a normal distribution with mean 0, variance 1, *N(0,10)* is a normal distribution with mean 0, variance 10, *N(10,50)* is a normal distribution with mean 10, variance 50, *norm* is normalized by the maximum value, and *orig* is the original distribution. (B) varies the prevalence of the trait, (C) varies the causal allele frequency, (D) varies the recombination fraction between the causal allele and marker allele, and (E) varies the marker minor allele frequency.

5.4 Discussion

5.4.1 QTPhenProxy can identify patients with stroke using EHR data other than the disease diagnosis code

For all stroke and its subtype ischemic stroke, the machine learning models trained to assign QTPhenProxy probabilities performed well (greater than 90% AUROC, greater than 30% maximum F1 score, 74-97% precision at 50). The models trained on subarachnoid hemorrhage and intracerebral hemorrhage cases performed similarly with AUROC but with lower maximum F1 score and precision at 50. This may be due to the number of cases available for the two subtypes to train on, which was an order of magnitude smaller (600 cases versus 3300-4300 cases). Importantly, we only trained on half of the cases and tested on the other half, which equates to 300-2200 cases, depending on the disease. From this small training set, we were able to assign a probability of disease to all 500,000 subjects in the UK Biobank. This is a fraction of the over 40,000 cases required to power the most recent genome-wide association study for stroke (MEGASTROKE).

5.4.2 QTPhenProxy discovers many new variants and recovers known disease variants to genome-wide significance

Training on half the number of cases of each disease and assigning a probability from the trained models to each UK Biobank subject resulted in 3-13 loci with genome-wide significance using QTPhenProxy. In contrast, traditional binary trait GWAS using all disease cases resulted in the discovery of 0 genome-wide significant loci for stroke, ischemic stroke, and subarachnoid hemorrhage and 2 loci for intracerebral hemorrhage. In addition to new loci discovery, QTPhenProxy recovered known disease variants to a significance level of 0.05 with a sensitivity better than binary trait GWAS. In addition to recovering known variants, overall, the effect size of the known variants in the binary trait GWAS was correlated with QTPhenProxy effect sizes for variants with low p-values. These results suggest that the QTPhenProxy method can recover relevant variants with fewer cases than traditional methods for stroke.

5.4.3 Simulation of quantitative trait and corresponding binary trait further supported the correlation of effect sizes between the two methods.

Our simulations showed high correlation between quantitative trait beta coefficients and binary trait log odds ratio, which is similar to our empirical findings in the UK Biobank. We also show that the correlation slope is relatively stable across all the different simulation parameters, suggesting that there is a set correlation between quantitative traits and binary traits created from them. Further simulation will need to determine the effects of multiple loci on correlation between quantitative and corresponding binary traits.

5.4.4 Variants discovered for stroke are enriched in disease marker sets for vascular and neurological disease, and variants discovered for other diseases were enriched for disease and system specific markers.

As a specificity measure for the variants discovered by QTPhenProxy, we found the EN model had the highest proportion of overlapping variants with EBI-GWAS marker sets related to vascular disorders and associated lab values. QTPhenProxy variants improve the power of detecting variants related to diseases that are co-morbid or risk factors for stroke. We found that for myocardial infarction, Asthma, and COPD, there were increased proportion of variants with a p-value significance of 0.05 or lower for disease marker sets corresponding to the related disease or system organ class compared to other disease marker sets. This was not the case for Alzheimer's and Parkinson's disease. This suggests that for some diseases, the variants discovered by QTPhenProxy were specific for the study disease and its system.

5.4.5 QTPhenProxy has high genetic correlation with the MEGASTROKE GWAS

We found high genetic correlation (0.60-0.64) between the MEGASTROKE all stroke GWAS and QTPhenProxy EN Model stroke GWAS. In addition, we found little genetic correlation(-0.1 and -0.0048 respectively) between the MEGASTROKE and a recent coronary artery disease (CAD) GWAS ([159]) and QTPhenProxy and the same CAD GWAS. The same genetic correlation

calculation between MEGASTROKE and CAD was reported in [11] as a large genetic correlation of around 0.5 with a p-value of 1E-20; therefore, further work will be needed to resolve this discrepancy. The high genetic correlation between MEGASTROKE and QTPhenProxy suggests that both studies highlight shared genetic underpinnings and predicts that risk factors for stroke GWAS should have positive genetic correlation with QTPhenProxy as it has for MEGASTROKE GWAS in the past. [11].

5.4.6 Low LD score regression intercepts relative to genomic inflation suggests high polygenicity

Our genome-wide association studies using QTPhenProxy had moderate genomic inflation, slightly above 1.1, but low LD score regression intercepts near 1.0. The corresponding binary trait GWAS had genomic inflation below 1.05, suggesting minimal population stratification. Since the population for the binary trait GWAS was the same as that used for QTPhenProxy, this suggests that polygenicity, rather than population stratification, is the cause for genomic inflation. Polygenicity, or the contribution of small effects of many genes to a phenotype, may be the more likely cause[157]. In addition, Bulik-Sullivan et. al. argue that genomic inflation can increase with sample size when there is polygenicity, and LD score regression intercept is more robust in distinguishing inflated p-values from polygenicity. As seen in [161], increased genomic inflation with common variants over rare variants suggests polygenicity. We show a similar upward trend in Figure 5.8. Another sign of true signal over inflated signal is the evidence of causal variants in linkage disequilibrium, as seen in a Manhattan or Hudson plot [161]. Figures figs. 5.3 and 5.4 show genome-wide significant variants in linkage disequilibrium with each other. Finally, this study reports results from using two different quality controls in the genome-wide association analysis, QC1, and more stringent, QC2. Results using QC1 showed more variants with genome-wide significance and discovered more new variants for stroke than the QC2 GWAS. This may lead one to believe that more stringent quality control led to reduced inflation of p-values. However, the genomic inflation of the p-values from the QC1 GWAS was the same as the genomic inflation of the p-values from the QC2 GWAS (1.134 vs 1.133), with the QC2 GWAS having a slightly higher

LD score regression intercept value (1.0288 vs 1.0369). This suggests that the p-values from the QC1 GWAS may not be overly inflated.

5.4.7 QTPhenProxy replicates known stroke variants and discovers variants within cardiovascular disease genes

Several of the genes discovered by the QTPhenProxy GWAS, using QC1 or QC2 quality control have been associated with stroke, including *NOS3*, *FURIN*, *PITX2*, *CDK2NB*, *LDLR*, and *ABO*[11, 162]. *NOS3*, in particular, was discovered through meta-analysis of the MEGASTROKE results with the UK Biobank[162]. We were able to replicate this association using only the QTPhenProxy method on the UK Biobank, and not traditional binary trait analysis. *NOS3* has been shown to be related to hypertension either through salt excretion regulation in the kidney[163] or regulation of vascular relaxation in endothelial cells[164]. Other discovered genes by QTPhenProxy are also associated with related cardiovascular diseases. The *F5* gene codes for an essential coagulation factor and mutations that can lead to increased thrombosis or hemorrhage, depending on the mutation[165, 166, 167]. *LPA* (Lipoprotein-A) codes for a protein that can contribute to atherosclerosis[168, 169]. Mutations in the *PLCB1* gene, which encodes for phospholipase synthesis, have been found in epilepsy and seizures[170]. The *LDLR* gene codes for the low-density lipoprotein receptor, which is involved in cholesterol production[171], and the *APOC1/APOC1P1* genes code for parts of apoliprotein C1, which are involved in high density lipoprotein metabolism [172]. *ARHGAP1*, a gene coding for Rho GTPase activating protein 1 has been associated with cancer phenotypes and activation of hypoxic and inflammatory pathways[173, 174, 175]. The *ABCG8* gene has been associated with combined gwas of lipids and inflammation, lipid levels, and gallstone disease[176]. *EYST3* and *XKR6* gene mutations have been associated with coronary artery disease and ischemic stroke respectively in Asian populations, and *MRAS* gene mutations have been associated with coronary artery disease in two populations[177, 178, 179, 180]. These results suggest that QTPhenProxy has replicated genome-wide significant mutations in genes known to be related with stroke and discovered those associated with related risk factors such as coronary artery

neurological diseases. Out of the 13 newly discovered variants with genome-wide significance, five are intronic or in nearby genes that have been found in previous studies (MEGASTROKE), and five more have been discovered in previous GWAS studies of related cardiovascular diseases. The variants within and near the *F5* gene had the highest genetic signal for stroke and all of its subtypes, even though it was not replicated in the MEGASTROKE GWAS. Out of all the new variants discovered by the QTPhenProxy EN model using the GWAS QC2 quality control, the rs11887534 variant, a missense mutation within the *ABCG8* gene, replicated to a p-value significance of 6.85E-04 in the MEGASTROKE stroke GWAS of over 40,000 European ancestry subjects. Using a tenth of the number of cases, QTPhenProxy discovered a variant within a new gene that replicated in MEGASTROKE, and discovered variants within known stroke genes.

5.4.8 Limitations

There are several limitations with this method. First, out of the five machine learning models used in QTPhenProxy, only two provided probabilities along a continuous scale. Adaboost, gradient boosting, and logistic regression using L1 penalty models, although with high performance, assigned probabilities in discrete bins. These distributions violate the genome-wide association linear regression assumption of normal distribution of the quantitative trait. Although the probabilities produced by EN and RF models were not initially normally distributed, they were continuous, and could be adjusted with quantile normalization. In addition, the GWAS results from QTPhenProxy using the RF model resulted in many more hits than the EN model, and reduced sensitivity to known disease variants. Further study will be required to understand why one model gave more sensitive and specific results than the other, and whether there was p-value inflation in the random forest models. In addition, from our preliminary studies, the sensitivity of recovering known loci for diseases other than stroke and ischemic stroke did not improve using QTPhenProxy. This may be because the training model for phenotyping patients was optimized for stroke[112]. Since stroke is an acute event than can be identified with high accuracy in the electronic health record, this method may not translate as well to other diseases, such as chronic illnesses. Another

potential limitation of the study is that only about half of known loci for stroke are replicated with significance of less than 0.05 in the QTPhenProxy method, even though new variants are suggested.

5.4.9 Conclusions

We have developed a method, QTPhenProxy, that we have shown improves the power of genome-wide association studies in stroke and three of its subtypes: ischemic stroke, subarachnoid hemorrhage, and intracerebral hemorrhage with an order of magnitude fewer cases than required for traditional genome-wide association studies of the same diseases. Converting dichotomous traits to quantitative ones could result in improvement of power by incorporating electronic health record information for subjects who may have genetic susceptibility to stroke but may not have experienced a stroke yet. Previous studies have shown that for diseases with low prevalence, there can be a reduction of power using logistic regression for binary trait GWAS compared to linear regression for quantitative trait GWAS[19]. Recently, [12] showed that the inclusion of ischemic stroke risk factors' genome-wide significant SNPs in polygenic risk score improves prediction of ischemic stroke. This supports our idea that inclusion of risk factor information into the phenotype can help detect genetically susceptible subjects. We plan to test the correlation between our QTPhenProxy probabilities with the metaGRS polygeneic risk scores from [12] in the UK Biobank. We have shown that there is high genetic correlation between MEGASTROKE and QTPhenProxy. We also show that with as few as 2200 stroke subjects we can recover known variants of stroke and discover new variants that have been linked to cardiovascular and nervous system diseases. This method could be useful for studies with a small set of cases and without access to large meta-analyses. We also have suggested new variants that warrant further replication in other groups. QTPhenProxy shows the benefits of incorporating electronic health record data to convert traditional binary traits to quantitative in improving GWAS power.

5.5 Acknowledgements

I would like to thank Dr. Krzysztof Kiryluk and Dr. Anna O. Basile for helpful discussions. MedDRA® trademark is registered by IFPMA on behalf of ICH. The MEGASTROKE project received funding from sources specified at http://www.megastroke.org/acknowledgments.html

Conclusion

Stroke is a highly heterogeneous and heritable disease that is a leading cause of mortality and disability. The etiology of strokes are varied and complex, caused by a combination of modifiable, such as lifestyle, and non-modifiable, such as genetics, risk factors and environmental influences. Elucidating the genetics of stroke and related diseases, such as coronary artery disease, has led to a better understanding of the genetics of the disease and potential avenues for treatment[181, 162, 11]. For stroke, however, large scale genome-wide association studies have discovered far fewer variants than related risk factors such as blood pressure and atrial fibrillation[12]. Past studies have suggested that this is due to the phenotypic heterogeneity of stroke[8, 11]. Recent studies in polygenic risk scores of stroke have also found improved predictive power of stroke patients when incorporating known variants of risk factors for stroke. This suggests that there are variants still to be discovered for stroke, and they most likely overlap with related diseases. In an era of large scale biobanks paired with electronic health record data, we hypothesized that a high-throughput method of incorporating medical information of stroke patients into a phenotypic score could bolster the power of genomic studies.

In this thesis, we combined supervised machine learning with large scale EHR and biobank databases to develop a high-throughput phenotyping method that improves the power of stroke genome-wide association studies. In Chapter 2, we develop this high-throughput phenotyping method, and find, across several supervised machine learning phenotyping models trained on a

small cohort of cases and controls, high performance of acute ischemic stroke (AIS) case identification with minimal feature processing. We built upon previous stroke machine learning phenotyping algorithms by using a data agnostic approach to identify features such as diagnostic procedure codes that could be included to identify stroke patients. In our external validation of the method in the UK Biobank, we also found that the top probabilities from a model-predicted AIS cohort were significantly enriched for AIS patients without AIS diagnosis codes. Our findings support machine learning algorithms as a generalizable way to accurately identify AIS patients without using process-intensive manual feature curation. In chapter 3, we applied factorization methods to subtype acute ischemic stroke patients, a potential avenue to reduce phenotypic heterogenity in stroke. We found non-negative matrix factorization to produce more stable subtypes over hierarchical Poisson factorization, and we found several subtypes significantly correlated to stroke severity. These subtypes highlighted known risk factors for stroke, including atrial fibrillation, diabetes mellitus, and acute respiratory failure. We then tested whether the probabilities generated from our supervised machine learning phenotyping models, which we called QTPhenProxy, could be used as phenotypic scores for genetic studies. This required the use of the phenotypic score as a quantitative trait. In Chapter 4, we showed that some of our high performing QTPhenProxy models estimated the heritability of AIS within the range of literature values and others underestimated AIS heritability, while traditional case/control assignment of AIS was unable to estimate AIS heritability. Our venous thromboembolism genome-wide association study was under-powered to show whether converting a phenotype from a binary to quantitative trait could improve the power of genome-wide association studies. In Chapter 5, however, in the UK Biobank database, we showed that QTPhenProxy can discover genome-wide significant variants associated with stroke with an order of magnitude fewer cases (4,354) than the most recent stroke GWAS, MEGASTROKE (40,585). We found up to 7 LD independent loci that pass genome-wide significance while the binary definition yielded no genome-wide significant hits. The majority of the discovered variants were near genes known to be associated with stroke. In addition, we found a novel locus in the *ABCG8* gene that replicates in MEGASTROKE. We introduced a method that

may improve the power of genome-wide association studies when limited curated cohort data are available within a healthcare system.

This thesis also lays the groundwork for future studies along several avenues. Although we began to understand what the probabilities assigned from the phenotyping models mean through calibration and performance, further study is warranted through validation of the models by physicians. In addition, although we used supervised machine learning models for ease of reproducibility and implementation, incorporation of timeline, natural language processing of notes, and application of deep learning models could improve performance of the models. A future exploration for subtyping stroke would determine whether the subtypes found have reduced phenotypic heterogeneity compared to traditional subtypes of stroke, and thus lead to reduced genetic heterogeneity and clearer genetic signal. A first step would expand the subtyping cohort to all stroke patients in the electronic health record and estimate the heritability using RIFEHR and SOLARstrap[51]. In addition, we could compare other subtyping methods such as topological data analysis and deep learning methods such as denoising autoencoders to the factorization methods to determine the best method for subtyping stroke.

We would like to generalize QTPhenProxy for more diseases. Many diseases, such as endometrosis or autoimmune diseases, are undiagnosed or misdiagnosed, and may benefit from a supervised machine learning algorithm trained on EHR data to identify potential patients[182, 88, 183, 184] and improve the power of their genetic studies. To start, we need to further explore why QTPhenProxy did not improve the power of other disease GWAS in the UK Biobank. This would include determining which aspects of the model are most informative, whether it be the effect size or penetrance of the variant, the phenotypic heterogeneity of the disease, the duration of the disease, the performance of the model, or the the number of cases available. In addition, polygenic risk scores are an important recent addition to the genomics field, and recently have been successfully applied to stroke[12]. The score is comprised of the effect sizes of variants known to be associated with the study disease through GWAS[185]. An important future direction of this thesis would be to determine the correlation between the polygenic risk score of stroke with the

phenotypic probabilities assigned by QTPhenProxy in the UK Biobank. This would suggest that the incorporation of EHR information into the trait plays a similar role to incorporating variants from risk factors, as shown in [12].

There is still a long way to go to fully identify the genetic burden of stroke. Increased cohort size to improve power and disease subtypes to reduce phenotypic heterogeneity are established methods for identifying variants. To keep up with the hundreds of thousands of future subjects, automated methods of cohort and subtype identification are needed. In addition, new methods to improve genomic power, such as considering traditional binary traits as quantitative by utilizing a phenotypic score, should be further explored and expanded upon. This thesis demonstrates the promise of incorporating large but messy sets of data, such as the electronic health record, to improve signal in genome-wide association studies.

References

[1] E. J. Benjamin, S. S. Virani, C. W. Callaway, A. M. Chamberlain, A. R. Chang, S. Cheng, S. E. Chiuve, M. Cushman, F. N. Delling, R. Deo, and et al., "Heart disease and stroke statistics-2018 update: A report from the american heart association.," Circulation, vol. 137, no. 12, e67–e492, 2018.

[2] Benjamin Emelia J., Blaha Michael J., Chiuve Stephanie E., Cushman Mary, Das Sandeep R., Deo Rajat, de Ferranti Sarah D., Floyd James, Fornage Myriam, Gillespie Cathleen, Isasi Carmen R., Jiménez Monik C., Jordan Lori Chaffin, Judd Suzanne E., Lackland Daniel, Lichtman Judith H., Lisabeth Lynda, Liu Simin, Longenecker Chris T., Mackey Rachel H., Matsushita Kunihiro, Mozaffarian Dariush, Mussolino Michael E., Nasir Khurram, Neumar Robert W., Palaniappan Latha, Pandey Dilip K., Thiagarajan Ravi R., Reeves Mathew J., Ritchey Matthew, Rodriguez Carlos J., Roth Gregory A., Rosamond Wayne D., Sasson Comilla, Towfighi Amytis, Tsao Connie W., Turner Melanie B., Virani Salim S., Voeks Jenifer H., Willey Joshua Z., Wilkins John T., Wu Jason HY., Alger Heather M., Wong Sally S., and Muntner Paul, "Heart Disease and Stroke Statistics—2017 Update: A Report From the American Heart Association," Circulation, vol. 135, no. 10, e146–e603, Mar. 2017.

[3] Q. Yang, "Vital Signs: Recent Trends in Stroke Death Rates — United States, 2000–2015," MMWR. Morbidity and Mortality Weekly Report, vol. 66, 2017.

[4] N. Maaijwee, L. Rutten-Jacobs, P. Schaapsmeerders, E. van Dijk, and F.-E. de Leeuw, "Ischaemic stroke in young adults: Risk factors and long-term consequences," Nature Reviews Neurology, vol. 10, no. 6, 315–325, 2014.

[5] A. Boehme, C. Esenwa, and M. Elkind, "Stroke risk factors, genetics, and prevention.," Circulation research, vol. 120, no. 3, 472–495, 2017.

[6] G. Jickling, B. Stamova, B. Ander, X. Zhan, D. Liu, S.-M. Sison, P. Verro, and F. Sharp, "Prediction of cardioembolic, arterial, and lacunar causes of cryptogenic stroke by gene expression and infarct location," Stroke, vol. 43, no. 8, 2036–2041, 2012.

[7] A. Nouh, M. Hussain, T. Mehta, and S. Yaghi, "Embolic strokes of unknown source and cryptogenic stroke: Implications in clinical practice," Frontiers in Neurology, vol. 7, p. 37, 2016.

[8] H. Markus and S. Bevan, "Mechanisms and treatment of ischaemic stroke—insights from genetic associations," Nature Reviews Neurology, vol. 10, no. 12, 723–730, 2014.

[9] N. (SiGN) and I. (ISGC), "Loci associated with ischaemic stroke and its subtypes (sign): A genome-wide association study.," The Lancet. Neurology,

[10] S. Bevan, M. Traylor, P. Adib-Samii, R. Malik, N. L. Paul, C. Jackson, M. Farrall, P. M. Rothwell, C. Sudlow, and M. Dichgans, "Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations," Stroke, vol. 43, no. 12, 3161–3167, 2012.

[11] R. Malik, G. Chauhan, M. Traylor, M. Sargurupremraj, Y. Okada, A. Mishra, L. Rutten-Jacobs, A.-K. Giese, S. W. van der Laan, S. Gretarsdottir, and et al., "Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes," Nature Genetics, vol. 50, 1–14, 2018.

[12] G. Abraham, R. Malik, E. Yonova-Doing, A. Salim, T. Wang, J. Danesh, A. S. Butterworth, J. M. M. Howson, M. Inouye, and M. Dichgans, "Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke," Nature Communications, vol. 10, no. 1, p. 5819, Dec. 2019.

[13] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn, "Genome-wide association studies for complex traits: Consensus, uncertainty and challenges," Nature Reviews Genetics, vol. 9, no. 5, pp. 356–369, May 2008.

[14] P. M. Visscher, W. G. Hill, and N. R. Wray, "Heritability in the genomics era â concepts and misconceptions," Nature Reviews Genetics, vol. 9, no. 4, pp. 255–266, Apr. 2008.

[15] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," Nature, vol. 461, no. 7265, pp. 747–753, Oct. 2009.

[16] D. Houle, D. R. Govindaraju, and S. Omholt, "Phenomics: The next challenge," Nature Reviews Genetics, vol. 11, no. 12, pp. 855–866, Dec. 2010.

[17] B. Maher, "Personal genomes: The case of the missing heritability.," Nature, vol. 456, no. 7218, 18–21, 2008.

[18] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, "Common SNPs explain a large proportion of the heritability for human height," Nature Genetics, vol. 42, no. 7, pp. 565–569, Jul. 2010.

[19] N. Zaitlen, B. Pasaniuc, N. Patterson, S. Pollack, B. Voight, L. Groop, D. Altshuler, B. E. Henderson, L. N. Kolonel, L. L. Marchand, K. Waters, C. A. Haiman, B. E. Stranger, E. T. Dermitzakis, P. Kraft, and A. L. Price, "Analysis of case-control association studies with known risk variants," Bioinformatics, vol. 28, no. 13, pp. 1729–1737, Jul. 2012.

[20] W Johannsen, "The genotype conception of heredity1," The American Naturalist, no. 45, pp. 129–159, 1911.

[21] M. R. Boland, G. Hripcsak, Y. Shen, W. K. Chung, and C. Weng, "Defining a comprehensive verotype using electronic health records for personalized medicine.," Journal of the American Medical Informatics Association: JAMIA, vol. 20, no. e2, e232–8, 2013.

[22] Nature Education, Phenotype / phenotypes | Learn Science at Scitable.

[23] N. Freimer and C. Sabatti, "The Human Phenome Project," Nature Genetics, vol. 34, no. 1, pp. 15–21, May 2003.

[24] D. Tirschwell and W. Longstreth, "Validating administrative data in stroke research," Stroke, vol. 33, no. 10, 2465–2470, 2002.

[25] 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.

[26] 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.

[27] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, G. McVean, S. Leslie, P. Donnelly, and J. Marchini, "Genome-wide genetic data on ~500,000 UK Biobank participants," Genetics, preprint, Jul. 2017.

[28] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O' Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini, "The UK Biobank resource with deep phenotyping and genomic data," Nature, vol. 562, no. 7726, pp. 203–209, Oct. 2018.

[29] B. Neale, UK Biobank, http://www.nealelab.is/uk-biobank, 2018.

[30] T. A. of Us Research Program Investigators, "The All of Us Research Program," New England Journal of Medicine, vol. 381, no. 7, pp. 668–676, Aug. 2019.

[31] T. G. P. Consortium, "An integrated map of genetic variation from 1,092 human genomes," Nature, vol. 491, no. 7422, pp. 56–65, Nov. 2012.

[32] M. D. Ritchie, J. C. Denny, D. C. Crawford, A. H. Ramirez, J. B. Weiner, J. M. Pulley, M. A. Basford, K. Brown-Gentry, J. R. Balser, D. R. Masys, J. L. Haines, and D. M. Roden, "Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record," The American Journal of Human Genetics, vol. 86, no. 4, pp. 560–572, Apr. 2010.

[33] J. A. Heit, S. M. Armasu, Y. W. Asmann, J. M. Cunningham, M. E. Matsumoto, T. M. Petterson, and M. D. Andrade, "A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q," Journal of Thrombosis and Haemostasis, vol. 10, no. 8, pp. 1521–1531, 2012.

[34] O. Gottesman, H. Kuivaniemi, G. Tromp, W. Faucett, R. Li, T. A. Manolio, and et. al., "The electronic medical records and genomics (emerge) network: Past, present, and future.," Genetics in Medicine, vol. 15, pp. 761–771, 2013.

[35] I. B. Stanaway, T. O. Hall, E. A. Rosenthal, M. Palmer, V. Naranbhai, R. Knevel, B. Namjou, R. J. Carroll, K. Kiryluk, A. S. Gordon, J. Linder, K. M. Howell, B. M. Mapes, F. T. J. Lin, Y. Y. Joo, M. G. Hayes, A. G. Gharavi, S. A. Pendergrass, M. D. Ritchie, M. de Andrade, D. C. Croteau, S. Raychaudhuri, S. T. Weiss, M. Lebo, S. S. Amr, D. Carrell, E. B. Larson, C. G. Chute, L. J. Rasmussen, M. J. Roy, P. Sleiman, H. Hakonarson, R. Li, E. W. Karlson, J. F. Peterson, I. J. Kullo, R. Chisholm, J. C. Denny, and G. P. Jarvik, "The eMERGE genotype set of 83,717 subjects imputed to ~40âmillion variants genome wide and association with the herpes zoster medical record phenotype," Genetic Epidemiology, vol. 43, pp. 63–81, 2019.

[36] G. Hripcsak, N. Shang, P. L. Peissig, L. V. Rasmussen, C. Liu, B. Benoit, R. J. Carroll, D. S. Carrell, J. C. Denny, O. Dikilitas, V. S. Gainer, K. M. Howell, J. G. Klann, I. J. Kullo, T. Lingren, F. D. Mentch, S. N. Murphy, K. Natarajan, J. A. Pacheco, W.-Q. Wei, K. Wiley, and C. Weng, "Facilitating phenotype transfer using a common data model," Journal of Biomedical Informatics, vol. 96, p. 103 253, Aug. 2019.

[37] N. Shang, C. Liu, L. V. Rasmussen, C. N. Ta, R. J. Caroll, B. Benoit, T. Lingren, O. Dikilitas, F. D. Mentch, D. S. Carrell, W.-Q. Wei, Y. Luo, V. S. Gainer, I. J. Kullo, J. A. Pacheco, H. Hakonarson, T. L. Walunas, J. C. Denny, K. Wiley, S. N. Murphy, G. Hripcsak, and C. Weng, "Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network," Journal of Biomedical Informatics, vol. 99, p. 103 293, Nov. 2019.

[38] C. Reich, P. Ryan, N. K. Belenkaya R., and C. Blacketer, Omop common data model v6.0 specifications, accessed:11.09.2019.

[39] M. A. Suchard, M. J. Schuemie, H. M. Krumholz, S. C. You, R. Chen, N. Pratt, C. G. Reich, J. Duke, D. Madigan, G. Hripcsak, and P. B. Ryan, "Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: A systematic, multinational, large-scale analysis," The Lancet, vol. 394, no. 10211, pp. 1816–1826, Nov. 2019.

[40] G. Hripcsak, P. B. Ryan, J. D. Duke, N. H. Shah, R. W. Park, V. Huser, M. A. Suchard, M. J. Schuemie, F. J. DeFalco, A. Perotte, J. M. Banda, C. G. Reich, L. M. Schilling, M. E. Matheny, D. Meeker, N. Pratt, and D. Madigan, "Characterizing treatment pathways at scale using the OHDSI network," Proceedings of the National Academy of Sciences, vol. 113, no. 27, pp. 7329–7336, Jul. 2016.

[41] M. R. Boland, P. Parhi, L. Li, R. Miotto, R. Carroll, U. Iqbal, P.-A. A. Nguyen, M. Schuemie, S. C. You, D. Smith, S. Mooney, P. Ryan, Y.-C. J. Li, R. W. Park, J. Denny, J. T. Dudley, G. Hripcsak, P. Gentine, and N. P. Tatonetti, "Uncovering exposures responsible for birth season â disease effects: A global study," Journal of the American Medical Informatics Association, vol. 25, no. 3, pp. 275–288, Mar. 2018.

[42] 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.

[43] 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.

[44] 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.

[45] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, Oct. 2019.

[46] O. H. D. S. a. Informatics, The Book of OHDSI. 2019, p. 63.

[47] J. C. Denny, "Chapter 13: Mining electronic health records in the genomics era," PLoS computational biology, vol. 8, no. 12, e1002823, 2012.

[48] R. M. Bilder, F. W. Sabb, T. D. Cannon, E. D. London, J. D. Jentsch, D. S. Parker, R. A. Poldrack, C. Evans, and N. B. Freimer, "Phenomics: The systematic study of phenotypes on a genome-wide scale," Neuroscience, Linking Genes to Brain Function in Health and Disease, vol. 164, no. 1, pp. 30–42, Nov. 2009.

[49] A. Rzhetsky, D. Wajngurt, N. Park, and T. Zheng, "Probing genetic overlap among complex human phenotypes," Proceedings of the National Academy of Sciences, vol. 104, no. 28, pp. 11 694–11 699, Jul. 2007.

[50] K. Wang, H. Gaitsch, H. Poon, N. J. Cox, and A. Rzhetsky, "Classification of common human diseases derived from shared genetic and environmental determinants," Nature genetics, vol. 49, no. 9, pp. 1319–1325, Sep. 2017.

[51] F. C. G. C. Polubriaginof, R. Vanguri, K. Quinnies, G. M. Belbin, A. Yahi, H. Salmasian, T. Lorberbaum, V. Nwankwo, L. Li, M. M. Shervey, and et al., "Disease heritability inferred from familial relationships reported in medical records.," Cell, vol. 173, no. 7, 1692–1704.e11, 2018.

[52] C. M. Lakhani, B. T. Tierney, A. K. Manrai, J. Yang, P. M. Visscher, and C. J. Patel, "Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes," Nature Genetics, vol. 51, no. 2, pp. 327–334, Feb. 2019.

[53] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford, "PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover geneâdisease associations," Bioinformatics, vol. 26, no. 9, pp. 1205–1210, May 2010.

[54] A. O. Basile and M. D. Ritchie, "Informatics and machine learning to define the phenotype," Expert Review of Molecular Diagnostics, vol. 18, no. 3, pp. 219–226, Mar. 2018.

[55] A. E. Arch, D. C. Weisman, S. Coca, K. V. Nystrom, C. R. Wira, and J. L. Schindler, "Missed ischemic stroke diagnosis in the emergency department by emergency medicine and neurology services.," Stroke, vol. 47, no. 3, 668–73, 2016.

[56] S. Lyalina, B. Percha, P. LePendu, S. V. Iyer, R. B. Altman, and N. H. Shah, "Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records," Journal of the American Medical Informatics Association: JAMIA, vol. 20, no. e2, e297–305, Dec. 2013.

[57] J. Ho, J. Ghosh, S. Steinhubl, W. Stewart, J. Denny, B. Malin, and J. Sun, "Limestone: High-throughput candidate phenotype generation via tensor factorization," Journal of Biomedical Informatics, vol. 52, 199–211, 2014.

[58] M. S. Udler, J. Kim, M. von Grotthuss, S. Bonàs-Guarch, J. B. Cole, J. Chiou, M. Boehnke, M. Laakso, G. Atzmon, B. Glaser, and et al., "Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis.," PLoS medicine, vol. 15, no. 9, e1002654, 2018.

[59] Y. Luo, C. Mao, Y. Yang, F. Wang, F. S. Ahmad, D. Arnett, M. R. Irvin, and S. J. Shah, "Integrating hypertension phenotype and genotype with hybrid non-negative matrix factorization," Bioinformatics, vol. 35, no. 8, pp. 1395–1403, Apr. 2019.

[60] J. Zhao, Y. Zhang, D. J. Schlueter, P. Wu, V. Eric Kerchberger, S. Trent Rosenbloom, Q. S. Wells, Q. Feng, J. C. Denny, and W.-Q. Wei, "Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: Cardiovascular disease case study," Journal of Biomedical Informatics, vol. 98, p. 103 270, Oct. 2019.

[61] D. Lee and S. H. Nature, "Learning the parts of objects by non-negative matrix factorization," Nature, 1999.

[62] R. Pivovarov, A. Perotte, E. Grave, J. Angiolillo, C. Wiggins, and N. Elhadad, "Learning probabilistic phenotypes from heterogeneous ehr data," Journal of Biomedical Informatics, vol. 58, 156–165, 2015.

[63] V. Rodriguez and A. Perotte, "Phenotype Inference with Semi-Supervised Mixed Membership Models," arXiv:1812.03222 [cs, q-bio, stat], Mar. 2019, arXiv: 1812.03222.

[64] P Gopalan, J. Hofman, and B. D. UAI, "Scalable recommendation with hierarchical poisson factorization.," UAI, 2015.

[65] H. Levitin, J. Yuan, Y. Cheng, F. Ruiz, E. Bush, J. Bruce, P. Canoll, A. Iavarone, A. Lasorella, D. Blei, and et al., "De novo gene signature identification from single-cell rna-seq with hierarchical poisson factorization," bioRxiv, p. 367 003, 2018.

[66] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[67] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in Proceedings of the 25th international conference on Machine learning - ICML '08, Helsinki, Finland: ACM Press, 2008, pp. 1096–1103, ISBN: 978-1-60558-205-4.

[68] 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.

[69] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," Scientific Reports, vol. 6, no. 1, pp. 1–10, May 2016.

[70] L. Li, W.-Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, and J. T. Dudley, "Identification of type 2 diabetes subgroups through topological analysis of patient similarity," Science translational medicine, vol. 7, no. 311, 311ra174, Oct. 2015.

[71] T. E. Madsen, J. Khoury, R. Cadena, O. Adeoye, K. A. Alwell, C. J. Moomaw, M. Erin, M. L. Flaherty, S. Ferioli, D. Woo, P. Khatri, J. P. Broderick, B. M. Kissela, and D. Kleindorfer, "Potentially missed diagnosis of ischemic stroke in the emergency department in the greater Cincinnati/Northern kentucky stroke study," Acad Emerg Med, vol. 23, no. 10, pp. 1128–1135, 2016.

[72] Benesch, Witter, Wilder, Duncan, Samsa, and Matchar, "Inaccuracy of the international classification of diseases (icd-9-cm) in identifying the diagnosis of ischemic cerebrovascular disease," Neurology, vol. 49, no. 3, 660–664, 1997.

[73] N. McCormick, V. Bhole, D. Lacaille, and J. A. Avina-Zubieta, "Validity of diagnostic codes for acute stroke in administrative databases: A systematic review.," PloS one, vol. 10, no. 8, e0135834, 2015.

[74] K. Olson, M. Wood, T Delate, L. Lash, J Rasmussen, A. Denham, and J. Merenich, "Positive predictive values of icd-9 codes to identify patients with stroke or tia.," The American journal of Managed Care, vol. 20, no. 2, 2014.

[75] T. E. Chang, J. H. Lichtman, L. B. Goldstein, and M. G. George, "Accuracy of icd-9-cm codes by hospital characteristics and stroke severity: Paul coverdell national acute stroke program.," Journal of the American Heart Association, vol. 5, no. 6, 2016.

[76] R. Woodfield, I. Grant, and C. L. Sudlow, "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, vol. 10, no. 10, e0140533, 2015.

[77] Y. Ni, K. Alwell, C. J. Moomaw, D. Woo, O. Adeoye, M. L. Flaherty, S. Ferioli, J. Mackey, F. De Los Rios La Rosa, S. Martini, and et al., "Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis.," PloS one, vol. 13, no. 2, e0192586, 2018.

[78] T. F. Imran, D. Posner, J. Honerlaw, J. L. Vassy, R. J. Song, Y.-L. L. Ho, S. J. Kittner, K. P. Liao, T. Cai, C. J. O'Donnell, and et al., "A phenotyping algorithm to identify acute ischemic stroke accurately from a national biobank: The million veteran program.," Clinical epidemiology, vol. 10, 1509–1521, 2018.

[79] H. Mo, W. Thompson, L. Rasmussen, J. Pacheco, G. Jiang, R. Kiefer, Q. Zhu, J. Xu, E. Montague, D. Carrell, and et al., "Desiderata for computable representations of electronic health records-driven phenotype algorithms," Journal of the American Medical Informatics Association, vol. 22, no. 6, 1220–1230, 2015.

[80] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. Embi, N. Elhadad, S. Johnson, and A. Lai, "A review of approaches to identifying patient phenotype cohorts using electronic

health records," Journal of the American Medical Informatics Association, vol. 21, no. 2, 221–230, 2014.

[81] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records.," Journal of the American Medical Informatics Association: JAMIA, vol. 20, no. 1, 117–21, 2013.

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

[83] R. Carroll, A. Eyler, and J. Denny, "Naïve electronic health record phenotype identification for rheumatoid arthritis.," AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, vol. 2011, 189–96, 2011.

[84] Y. Chen, R. Carroll, E. Hinz, A. Shah, A. Eyler, J. Denny, and H. Xu, "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.

[85] S. Yu, A. Chakrabortty, K. P. Liao, T. Cai, A. N. Ananthakrishnan, V. S. Gainer, S. E. Churchill, P. Szolovits, S. N. Murphy, I. S. Kohane, and et al., "Surrogate-assisted feature extraction for high-throughput phenotyping.," Journal of the American Medical Informatics Association: JAMIA, vol. 24, no. e1, e143–e149, 2017.

[86] W. Ning, S. Chan, A. Beam, M. Yu, A. Geva, K. Liao, M. Mullen, K. D. Mandl, I. Kohane, T. Cai, and et al., "Feature extraction for phenotyping from semantic and knowledge resources.," Journal of biomedical informatics, p. 103 122, 2019.

[87] S. Yu, Y. Ma, J. Gronsbell, T. Cai, A. Ananthakrishnan, V. Gainer, S. Churchill, P. Szolovits, S. Murphy, I. Kohane, and et al., "Enabling phenotypic big data with phenorm.," Journal of the American Medical Informatics Association: JAMIA, vol. 25, 2017.

[88] V. Agarwal, T. Podchiyska, J. M. Banda, V. Goel, T. I. Leung, E. P. Minty, T. E. Sweeney, E. Gyang, and N. H. Shah, "Learning statistical models of phenotypes using noisy labeled training data.," Journal of the American Medical Informatics Association: JAMIA, vol. 23, no. 6, 1166–1173, 2016.

[89] Y. Halpern, S. Horng, Y. Choi, and D. Sontag, "Electronic medical record phenotyping using the anchor and learn framework.," Journal of the American Medical Informatics Association: JAMIA, vol. 23, no. 4, 731–40, 2016.

[90] S. G. Murray, A. Avati, G. Schmajuk, and J. Yazdany, "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.

[91] 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.

[92] 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.

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

[94] 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.

[95] 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.

[96] 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.

[97] Center for Medicaid and Medicare Services, 2018 ICD-10 CM and GEMs, https://www.cms.gov/medicare/coding/icd10/2018-icd-10-cm-and-gems.html.

[98] Healthcare Cost and Utilization Project, CCS for ICD-10-CM, FY 2018 (October 2017), https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

[99] National Library of Medicine, UMLS SNOMED CT to ICD-10-CM Map.

[100] G. Hripcsak, M. E. Levine, N. Shang, and P. B. Ryan, "Effect of vocabulary mapping for conditions on phenotype cohorts.," Journal of the American Medical Informatics Association: JAMIA, 2018.

[101] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[102] Healthcare Cost and Utilization Project, Multi-Level Procedures CCS Categories, https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp#download.

[103] ——, Multi-Level Diagnoses CCS Categories, https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

[104] ——, CCS for ICD-10-PCS, FY 2018 (October 2017) , https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp#download.

[105] ——, 2018 CCS-Services and Procedures Software, https://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccscpt_downloading.jsp.

[106] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, and 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 medicine, vol. 12, no. 3, e1001779, 2015.

[107] J. A. Sinnott, W. Dai, K. P. Liao, S. Y. Shaw, A. N. Ananthakrishnan, V. S. Gainer, E. W. Karlson, S. Churchill, P. Szolovits, S. Murphy, and et al., "Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records.," Human genetics, vol. 133, no. 11, 1369–82, 2014.

[108] J. Sinnott, F. Cai, S. Yu, B. Hejblum, C. Hong, I. Kohane, and K. Liao, "Pheprob: Probabilistic phenotyping using diagnosis codes to improve power for genetic association studies.," Journal of the American Medical Informatics Association: JAMIA, 2018.

[109] L. Bastarache, J. Hughey, S. Hebbring, J. Marlo, W. Zhao, W. Ho, S. Driest, T. McGregor, J. Mosley, Q. Wells, and et al., "Phenotype risk scores identify patients with unrecognized mendelian disease patterns," Science, vol. 359, no. 6381, 1233–1239, 2018.

[110] J. H. Son, G. Xie, C. Yuan, L. Ena, Z. Li, A. Goldstein, L. Huang, L. Wang, F. Shen, H. Liu, and et al., "Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes.," American journal of human genetics, vol. 103, no. 1, 58–73, 2018.

[111] G. Hripcsak and D. Albers, "High-fidelity phenotyping: Richness and freedom from bias," Journal of the American Medical Informatics Association, 2017.

[112] P. M. Thangaraj, B. R. Kummer, T. Lorberbaum, M. V. S. Elkind, and N. P. Tatonetti, "Comparative analysis, applications, and interpretation of electronic health record-based stroke phenotyping methods," bioRxiv, 2019.

[113] A. Blanco-gomez, S. Castillo-Lluva, M. d. M. Saez-Freire, L. Hontecillas-Prieto, J. H. Mao, A. Castellanos-Martin, and J. Perez-Losada, "Missing heritability of complex diseases: Enlightenment by genetic variants from intermediate phenotypes," BioEssays, vol. 38, no. 7, pp. 664–673, 2016.

[114] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, "The mystery of missing heritability: Genetic interactions create phantom heritability," Proceedings of the National Academy of Sciences, vol. 109, no. 4, pp. 1193–1198, Jan. 2012.

[115] H. Ay, E. Arsava, G. Andsberg, T. Benner, R. D. Brown, S. N. Chapman, J. W. Cole, H. Delavaran, M. Dichgans, G. Engström, G. Eva, R. P. Grewal, K. Gwinn, C. Jern, J. Jordi, K. Jood, M. Katsnelson, B. Kissela, S. J. Kittner, D. O. Kleindorfer, D. L. Labovitz, S. Lanfranconi, J. Lee, M. Lehm, R. Lemmens, C. Levi, L. Li, A. Lindgren, H. S. Markus, M. P. F, O. Melander, B. Norrving, L. Peddareddygari, A. Pedersén, J. Pera, K. Rannikmäe, K. M. Rexrode, D. Rhodes, S. S. Rich, J. Roquer, J. Rosand, P. M. Rothwell, T. Rundek, R. L. Sacco, R. Schmidt, M. Schürks, S. Seiler, P. Sharma, A. Slowik, C. Sudlow, V. Thijs, R. Woodfield, B. B. Worrall, and J. F. Meschia, "Pathogenic ischemic stroke phenotypes in the NINDS-Stroke genetics network," Stroke, vol. 45, no. 12, pp. 3589–3596, 2014.

[116] Adams, Bendixen, Kappelle, Biller, Love, Gordon, and Marsh, "Classification of subtype of acute ischemic stroke. definitions for use in a multicenter clinical trial. toast. trial of org 10172 in acute stroke treatment.," Stroke, vol. 24, no. 1, 35–41, 1993.

[117] Amarenco, Bogousslavsky, L. Caplan, G. Donnan, and M. Hennerici, "New approach to stroke subtyping: The a-s-c-o (phenotypic) classification of stroke," Cerebrovascular Diseases, vol. 27, no. 5, 502–508, 2009.

[118] H. Ay, T. Benner, M. Arsava, K. Furie, A. Singhal, M. Jensen, C. Ayata, A. Towfighi, E. Smith, J. Chong, and et al., "A computerized algorithm for etiologic classification of ischemic stroke the causative classification of stroke system," Stroke, vol. 38, no. 11, 2979–2984, 2007.

[119] M. E. Arsava, J. Helenius, R. Avery, M. H. Sorgun, G. Kim, P. O. M, K. Park, J. Rosand, M. Vangel, and H. Ay, "Assessment of the predictive validity of etiologic stroke classification.," Jama Neurol, vol. 74, no. 4, p. 419, 2017.

[120] H. Lin, P. Wolf, K. M, A. Beiser, C. Kase, E. Benjamin, and D. RB, "Stroke severity in atrial fibrillation. the framingham study.," Stroke, vol. 27, no. 10, pp. 1760–4, 1996.

[121] J. C. Kwong, K. L. Schwartz, M. A. Campitelli, H. Chung, N. S. Crowcroft, T. Karnauchow, K. Katz, D. T. Ko, A. J. McGeer, D. McNally, D. C. Richardson, L. C. Rosella, A. Simor, M. Smieja, G. Zahariadis, and J. B. Gubbay, "Acute Myocardial Infarction after Laboratory-Confirmed Influenza Infection," New England Journal of Medicine, Jan. 2018.

[122] M. S. V. Elkind, P. Ramakrishnan, Y. P. Moon, B. Boden-Albala, K. M. Liu, S. L. Spitalnik, T. Rundek, R. L. Sacco, and M. C. Paik, "Infectious Burden and Risk of Stroke: The Northern Manhattan Study," Archives of Neurology, vol. 67, no. 1, pp. 33–38, Jan. 2010.

[123] B. B. Navi and C. Iadecola, "Ischemic stroke in cancer patients: A review of an underappreciated pathology," Annals of Neurology, vol. 83, no. 5, pp. 873–883, 2018.

[124] L. Almasy and J. Blangero, "Multipoint Quantitative-Trait Linkage Analysis in General Pedigrees," The American Journal of Human Genetics, vol. 62, no. 5, pp. 1198–1211, May 1998.

[125] Dichgans, Malik, Konig, Rosand, Clarke, Gretarsdottir, Thorleifsson, Mitchell, Assimes, Levi, and et al., "Shared genetic susceptibility to ischemic stroke and coronary artery disease," Stroke, vol. 45, no. 1, 24–36, 2014.

[126] J. Kirby, P. Speltz, L. Rasmussen, and et al., "Phekb: A catalog and workflow for creating electronic phenotype algorithms for transportability.," Journal of the American Medical Informatics Association, vol. 23, no. 6, pp. 1046–1052, 2016.

[127] T. J. C. Polderman, B. Benyamin, C. A. d. Leeuw, P. F. Sullivan, A. v. Bochoven, P. M. Visscher, and D. Posthuma, "Meta-analysis of the heritability of human traits based on fifty years of twin studies," Nature Genetics, vol. 47, no. 7, pp. 702–709, Jul. 2015.

[128] A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson, "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019," Nucleic Acids Research, vol. 47, no. D1, pp. D1005–D1012, Jan. 2019.

[129] D. Klarin, C. A. Emdin, P. Natarajan, M. F. Conrad, and S. Kathiresan, "Genetic Analysis of Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates Obesity as a Causal Risk Factor," Circulation. Cardiovascular genetics, vol. 10, no. 2, Apr. 2017.

[130] M. Germain, D. I. Chasman, H. de Haan, W. Tang, S. Lindstrom, L.-C. Weng, M. de Andrade, M. C. H. de Visser, K. L. Wiggins, P. Suchon, N. Saut, D. M. Smadja, G. Le Gal, A. van Hylckama Vlieg, A. Di Narzo, K. Hao, C. P. Nelson, A. Rocanin-Arjo, L. Folkersen, R. Monajemi, L. M. Rose, J. A. Brody, E. Slagboom, D. Aissi, F. Gagnon, J.-F. Deleuze, P. Deloukas, C. Tzourio, J.-F. Dartigues, C. Berr, K. D. Taylor, M. Civelek, P. Eriksson, B. M. Psaty, J. Houwing-Duitermaat, A. H. Goodall, F. Cambien, P. Kraft, P. Amouyel, N. J. Samani, S. Basu, P. M. Ridker, F. R. Rosendaal, C. Kabrhel, A. R. Folsom, J. Heit, P. H. Reitsma, D.-A. Tregouet, N. L. Smith, and P.-E. Morange, "Meta-analysis of 65,734 Individuals Identifies TSPAN15 and SLC44a2 as Two Susceptibility

Loci for Venous Thromboembolism," The American Journal of Human Genetics, vol. 96, no. 4, pp. 532–542, Apr. 2015.

[131] M. Germain, N. Saut, N. Greliche, C. Dina, J.-C. Lambert, C. Perret, W. Cohen, T. Oudot-Mellakh, G. Antoni, M.-C. Alessi, D. Zelenika, F. Cambien, L. Tiret, M. Bertrand, A.-M. Dupuy, L. Letenneur, M. Lathrop, J. Emmerich, P. Amouyel, D.-A. Tregouet, and P.-E. Morange, "Genetics of Venous Thrombosis: Insights from a New Genome Wide Association Study," PLOS ONE, vol. 6, no. 9, e25581, Sep. 2011.

[132] W. Tang, M. Teichert, D. I. Chasman, J. A. Heit, P.-E. Morange, G. Li, N. Pankratz, F. W. Leebeek, G. Pare, M. d. Andrade, C. Tzourio, B. M. Psaty, S. Basu, R. Ruiter, L. Rose, S. M. Armasu, T. Lumley, S. R. Heckbert, A. G. Uitterlinden, M. Lathrop, K. M. Rice, M. Cushman, A. Hofman, J.-C. Lambert, N. L. Glazer, J. S. Pankow, J. C. Witteman, P. Amouyel, J. C. Bis, E. G. Bovill, X. Kong, R. P. Tracy, E. Boerwinkle, J. I. Rotter, D.-A. Tregouet, D. W. Loth, B. H. C. Stricker, P. M. Ridker, A. R. Folsom, and N. L. Smith, "A Genome-Wide Association Study for Venous Thromboembolism: The Extended Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium," Genetic Epidemiology, vol. 37, no. 5, pp. 512–521, 2013.

[133] J. Heit, J. Pathak, J. Denny, G. Hinz, and Mayo Clinic, Venous Thromboembolism (VTE) | PheKB, 2012.

[134] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. d. Bakker, M. J. Daly, and P. C. Sham, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," The American Journal of Human Genetics, vol. 81, no. 3, pp. 559–575, Sep. 2007.

[135] National Center for Biotechnology Information, US National Library of Medicine, Coordinate remapping service: NCBI.

[136] C. DeBoever, Y. Tanigawa, M. Aguirre, G. McInnes, A. Lavertu, and M. A. Rivas, "Assessing digital phenotyping to enhance genetic studies of human diseases," Genetics, preprint, Aug. 2019.

[137] W. Zhou, J. B. Nielsen, L. G. Fritsche, R. Dey, M. E. Gabrielsen, B. N. Wolford, J. LeFaive, P. VandeHaar, S. A. Gagliano, A. Gifford, L. A. Bastarache, W.-Q. Wei, J. C. Denny, M. Lin, K. Hveem, H. M. Kang, G. R. Abecasis, C. J. Willer, and S. Lee, "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies," Nature Genetics, vol. 50, no. 9, pp. 1335–1341, Sep. 2018.

[138] J. Yang, N. R. Wray, and P. M. Visscher, "Comparing apples and oranges: Equating the power of case-control and quantitative trait association studies," Genetic Epidemiology, n/a–n/a, 2009.

[139] D. V. Zaykin and L. A. Zhivotovsky, "Ranks of Genuine Associations in Whole-Genome Scans," Genetics, vol. 171, no. 2, pp. 813–823, Oct. 2005.

[140] I. S. Kohane, "Using electronic health records to drive discovery in disease genomics," Nature Reviews Genetics, vol. 12, no. 6, pp. 417–428, Jun. 2011.

[141] J. A. Sinnott, F. Cai, S. Yu, B. P. Hejblum, C. Hong, I. S. Kohane, and K. P. Liao, "PheProb: Probabilistic phenotyping using diagnosis codes to improve power for genetic association studies," Journal of the American Medical Informatics Association, vol. 25, no. 10, pp. 1359–1365, Oct. 2018.

[142] J. Z. Liu, Y. Erlich, and J. K. Pickrell, "Case-control association mapping by proxy using family history of disease," Nature Genetics, vol. 49, no. 3, pp. 325–331, Mar. 2017.

[143] D. M. Ruderfer, C. G. Walsh, M. W. Aguirre, Y. Tanigawa, J. D. Ribeiro, J. C. Franklin, and M. A. Rivas, "Significant shared heritability underlies suicide attempt and clinically predicted probability of attempting suicide," Molecular Psychiatry, pp. 1–9, Jan. 2019.

[144] National Library of Medicine, UMLS Metathesaurus - RXNORM.

[145] C. Schnier, K. Bush, J. Nolan, C. L. M. Sudlow, and UK Biobank Outcome Adjuducation Group, Definitions of Stroke for UK Biobank Phase 1 Outcomes Adjudication, Aug. 2017.

[146] UK Biobank, "Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource," UK Biobank Website,

[147] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation PLINK: Rising to the challenge of larger and richer datasets," GigaScience, vol. 4, no. 1, p. 7, Dec. 2015.

[148] E. Brown, "Coding of Data â MedDRA and other Medical Technologies," in Clinical Data Management, Chichester, UK, Dec. 1999, p. 177, ISBN: 978-0-471-98329-3.

[149] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications," Nucleic Acids Research, vol. 39, no. suppl, W541–W545, Jul. 2011.

[150] J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, P. A. F. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, R. J. Loos, T. M. Frayling, M. I. McCarthy, J. N. Hirschhorn, M. E. Goddard, and P. M. Visscher, "Conditional and joint multiple-SNP analysis of GWAS summary

statistics identifies additional variants influencing complex traits," Nature Genetics, vol. 44, no. 4, pp. 369–375, Apr. 2012.

[151] H. R. Wells, M. B. Freidin, F. N. Zainul Abidin, A. Payton, P. Dawes, K. J. Munro, C. C. Morton, D. R. Moore, S. J. Dawson, and F. M. Williams, "GWAS Identifies 44 Independent Associated Genomic Loci for Self-Reported Adult Hearing Difficulty in UK Biobank," The American Journal of Human Genetics, vol. 105, no. 4, pp. 788–802, Oct. 2019.

[152] National Center for Biotechnology Information, National Library of Medicine., Database of single nucleotide polymorphisms (dbsnp). `http://www.ncbi.nlm.nih.gov/SNP/`, (dbSNP Build ID: GRCh37.p13), Bethesda, MD.

[153] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The Human Genome Browser at UCSC," vol. 12, pp. 996–1006, 2002.

[154] N. Risch, "Linkage Strategies for Genetically Complex Traits. 1. Multilocus Models," Am J Hum Genet., vol. 46, no. 2, pp. 222–228, 1990.

[155] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in 9th Python in Science Conference, 2010.

[156] G. Band and J. Marchini, "BGEN: A binary file format for imputed genotype and haplotype data," bioRxiv, p. 308 296, May 2018.

[157] B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale, "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies," Nature Genetics, vol. 47, no. 3, pp. 291–295, Mar. 2015.

[158] S. D. Turner, "Qqman: An R package for visualizing GWAS results using Q-Q and manhattan plots," bioRxiv, p. 005 165, May 2014.

[159] C. P. Nelson, A. Goel, A. S. Butterworth, S. Kanoni, T. R. Webb, E. Marouli, L. Zeng, I. Ntalla, F. Y. Lai, J. C. Hopewell, O. Giannakopoulou, T. Jiang, S. E. Hamby, E. Di Angelantonio, T. L. Assimes, E. P. Bottinger, J. C. Chambers, R. Clarke, C. N. A. Palmer, R. M. Cubbon, P. Ellinor, R. Ermel, E. Evangelou, P. W. Franks, C. Grace, D. Gu, A. D. Hingorani, J. M. M. Howson, E. Ingelsson, A. Kastrati, T. Kessler, T. Kyriakou, T. Lehtimäki, X. Lu, Y. Lu, W. März, R. McPherson, A. Metspalu, M. Pujades-Rodriguez, A. Ruusalepp, E. E. Schadt, A. F. Schmidt, M. J. Sweeting, P. A. Zalloua, K. AlGhalayini, B. D. Keavney, J. S. Kooner, R. J. F. Loos, R. S. Patel, M. K. Rutter, M. Tomaszewski, I. Tzoulaki, E. Zeggini, J. Erdmann, G. Dedoussis, J. L. M. Björkegren, H. Schunkert, M. Farrall, J. Danesh, N. J. Samani, H. Watkins, and P. Deloukas, "Association analyses based on false discovery rate implicate new loci for coronary artery disease," Nature Genetics, vol. 49, no. 9, pp. 1385–1391, Sep. 2017.

[160] A. Lucas, Hudson: R package, https://github.com/anastasia-lucas/hudson, Philadelphia,PA.

[161] D. Howrigan, L. Abbott, C. Churchhouse, and D. Palmer, UK Biobank, http://www.nealelab.is/uk-biobank, 2017.

[162] R. Malik, K. Rannikmae, M. Traylor, M. K. Georgakis, M. Sargurupremraj, H. S. Markus, J. C. Hopewell, S. Debette, C. L. M. Sudlow, M. Dichgans, and for the MEGASTROKE consortium and the International Stroke Genetics Consortium, "Genome-wide meta-analysis identifies 3 novel loci associated with stroke: MEGASTROKE and UK Biobank GWAS," Annals of Neurology, vol. 84, no. 6, pp. 934–939, Dec. 2018.

[163] Gao Yang, Stuart Deborah, Takahishi Takamune, and Kohan Donald E., "NephronâSpecific Disruption of Nitric Oxide Synthase 3 Causes Hypertension and Impaired Salt Excretion," Journal of the American Heart Association, vol. 7, no. 14, e009236, Jul. 2018.

[164] C. Farah, L. Y. M. Michel, and J.-L. Balligand, "Nitric oxide signalling in cardiovascular health and disease," Nature Reviews Cardiology, vol. 15, no. 5, pp. 292–316, May 2018.

[165] R. Asselta and F. Peyvandi, "Factor V deficiency," Seminars in Thrombosis and Hemostasis, vol. 35, no. 4, pp. 382–389, Jun. 2009.

[166] J. L. Kujovich, "Factor V Leiden thrombophilia," Genetics in Medicine: Official Journal of the American College of Medical Genetics, vol. 13, no. 1, pp. 1–16, Jan. 2011.

[167] D. A. Hinds, A. Buil, D. Ziemek, A. Martinez-Perez, R. Malik, L. Folkersen, M. Germain, A. Malarstig, A. Brown, J. M. Soria, M. Dichgans, N. Bing, A. Franco-Cereceda, J. C. Souto, E. T. Dermitzakis, A. Hamsten, B. B. Worrall, J. Y. Tung, METASTROKE Consortium, INVENT Consortium, and M. Sabater-Lleal, "Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis," Human Molecular Genetics, vol. 25, no. 9, pp. 1867–1874, 2016.

[168] The Emerging Risk Factors Collaboration, "Lipoprotein(a) Concentration and the Risk of Coronary Heart Disease, Stroke, and Nonvascular Mortality," JAMA : the journal of the American Medical Association, vol. 302, no. 4, pp. 412–423, Jul. 2009.

[169] Wang Long, Chen Juan, Zeng Ying, Wei Jie, Jing Jinjin, Li Ge, Su Li, Tang Xiaojun, Wu Tangchun, and Zhou Li, "Functional Variant in the SLC22a3-LPAL2-LPA Gene Cluster Contributes to the Severity of Coronary Artery Disease," Arteriosclerosis, Thrombosis, and Vascular Biology, vol. 36, no. 9, pp. 1989–1996, Sep. 2016.

[170] A.-S. Schoonjans, M. Meuwissen, E. Reyniers, F. Kooy, and B. Ceulemans, "PLCB1 epileptic encephalopathies; Review and expansion of the phenotypic spectrum," European journal of paediatric neurology, vol. 20, no. 3, pp. 474–479, May 2016.

[171] M. S. Brown and J. L. Goldstein, "Receptor-mediated endocytosis: Insights from the lipoprotein receptor system," Proceedings of the National Academy of Sciences of the United States of America, vol. 76, no. 7, pp. 3330–3337, Jul. 1979.

[172] S. Erqou, A. Thompson, E. D. Angelantonio, D. Saleheen, S. Kaptoge, S. Marcovina, and J. Danesh, "Apolipoprotein(a) Isoforms and the Risk of Vascular Disease: Systematic Review of 40 Studies Involving 58,000 Participants," Journal of the American College of Cardiology, vol. 55, no. 19, pp. 2160–2167, May 2010.

[173] M. B. Fessler, P. G. Arndt, S. C. Frasch, J. G. Lieber, C. A. Johnson, R. C. Murphy, J. A. Nick, D. L. Bratton, K. C. Malcolm, and G. S. Worthen, "Lipid Rafts Regulate Lipopolysaccharide-induced Activation of Cdc42 and Inflammatory Functions of the Human Neutrophil," Journal of Biological Chemistry, vol. 279, no. 38, pp. 39 989–39 998, Sep. 2004.

[174] K. Hashimoto, H. Ochi, S. Sunamura, N. Kosaka, Y. Mabuchi, T. Fukuda, K. Yao, H. Kanda, K. Ae, A. Okawa, C. Akazawa, T. Ochiya, M. Futakuchi, S. Takeda, and S. Sato, "Cancer-secreted hsa-miR-940 induces an osteoblastic phenotype in the bone metastatic microenvironment via targeting ARHGAP1 and FAM134a," Proceedings of the National Academy of Sciences, vol. 115, no. 9, pp. 2204–2209, Feb. 2018.

[175] X. Zhang, J. Guan, M. Guo, H. Dai, S. Cai, C. Zhou, Y. Wang, and Q. Qin, "Rho GTPaseâactivating protein 1 promotes apoptosis of myocardial cells in an ischemic cardiomyopathy model," Kardiologia Polska, vol. 77, no. 12, pp. 1163–1169, 2019.

[176] S. Ligthart, A. Vaez, Y.-H. Hsu, Inflammation Working Group of the CHARGE Consortium, PMI-WG-XCP, LifeLines Cohort Study, R. Stolk, A. G. Uitterlinden, A. Hofman, B. Z. Alizadeh, O. H. Franco, and A. Dehghan, "Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation," BMC Genomics, vol. 17, no. 1, p. 443, Dec. 2016.

[177] F. Jiang, Y. Dong, C. Wu, X. Yang, L. Zhao, J. Guo, Y. Li, J. Dong, G.-Y. Zheng, H. Cao, L. Jin, Y. Ren, W. Cheng, W. Li, X.-L. Tian, and X. Li, "Fine mapping of chromosome 3q22.3 identifies two haplotype blocks in ESYT3 associated with coronary artery disease in female Han Chinese," Atherosclerosis, vol. 218, no. 2, pp. 397–403, Oct. 2011.

[178] P.-F. Zheng, R.-X. Yin, G.-X. Deng, Y.-Z. Guan, B.-L. Wei, and C.-X. Liu, "Association between the XKR6 rs7819412 SNP and serum lipid levels and the risk of coronary artery

disease and ischemic stroke," BMC Cardiovascular Disorders, vol. 19, no. 1, p. 202, Aug. 2019.

[179] Y. Song, R. Ma, and H. Zhang, "The influence of MRAS gene variants on ischemic stroke and serum lipid levels in Chinese Han population," Medicine, vol. 98, no. 48, e18065, Nov. 2019.

[180] J. Erdmann, A. Grosshennig, P. S. Braund, I. R. Konig, C. Hengstenberg, A. S. Hall, P. Linsel-Nitschke, S. Kathiresan, B. Wright, D.-A. Tregouet, F. Cambien, P. Bruse, Z. Aherrahrou, A. K. Wagner, K. Stark, S. M. Schwartz, V. Salomaa, R. Elosua, O. Melander, B. F. Voight, C. J. O' Donnell, L. Peltonen, D. S. Siscovick, D. Altshuler, P. A. Merlini, F. Peyvandi, L. Bernardinelli, D. Ardissino, A. Schillert, S. Blankenberg, T. Zeller, P. Wild, D. F. Schwarz, L. Tiret, C. Perret, S. Schreiber, N. E. El Mokhtari, A. Schafer, W. Marz, W. Renner, P. Bugert, H. KlÃŒter, J. Schrezenmeir, D. Rubin, S. G. Ball, A. J. Balmforth, H.-E. Wichmann, T. Meitinger, M. Fischer, C. Meisinger, J. Baumert, A. Peters, W. H. Ouwehand, Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group, Myocardial Infarction Genetics Consortium, Wellcome Trust Case Control Consortium, Cardiogenics Consortium, P. Deloukas, J. R. Thompson, A. Ziegler, N. J. Samani, and H. Schunkert, "New susceptibility locus for coronary artery disease on chromosome 3q22.3," Nature Genetics, vol. 41, no. 3, pp. 280–282, Mar. 2009.

[181] A. M. Small, C. J. O'Donnell, and S. M. Damrauer, "Large-Scale Genomic Biobanks and Cardiovascular Disease," Current Cardiology Reports, vol. 20, no. 4, p. 22, Mar. 2018.

[182] M. McKillop, L. Mamykina, and N. Elhadad, "Designing in the Dark: Eliciting Self-tracking Dimensions for Understanding Enigmatic Disease," in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal QC, Canada: ACM Press, 2018, pp. 1–15, ISBN: 978-1-4503-5620-6.

[183] A. S. Faye, F. Polubriaginof, P. H. R. Green, D. K. Vawdrey, N. Tatonetti, and B. Lebwohl, "Low Rates of Screening for Celiac Disease Among Family Members," Clinical Gastroenterology and Hepatology, vol. 17, no. 3, pp. 463–468, Feb. 2019.

[184] J. A. O'Rourke, C. Ravichandran, Y. J. Howe, J. E. Mullett, C. J. Keary, S. B. Golas, A. R. Hureau, M. McCormick, J. Chung, N. R. Rose, and C. J. McDougle, "Accuracy of self-reported history of autoimmune disease: A pilot study," PLoS ONE, vol. 14, no. 5, May 2019.

[185] L. P. Sugrue and R. S. Desikan, "What Are Polygenic Scores and Why Are They Important?" JAMA, vol. 321, no. 18, pp. 1820–1821, May 2019.