

A. SPECIFIC AIMS

Stroke remains the primary source of morbidity and mortality associated with atrial fibrillation (AF), despite major advances in prevention – direct acting oral anticoagulants and left atrial appendage occlusion.²⁻⁴ While effective stroke-prevention strategies are available for patients appropriately identified to be at significant risk,^{4, 6} optimal implementation of these treatments is limited by (1) rudimentary stroke risk stratification tools, and (2) disparities in care and outcomes of AF.^{1, 7-12, 14-20} Strokes still occur among patients with AF who are misclassified as low-risk or fail to receive appropriate therapies due to healthcare disparities.^{8, 21} Thus, **there remains a critical need for personalized and equitable stroke risk stratification among patients with AF, in order to optimally implement contemporary stroke-prevention therapies.**^{12, 13}

The overarching goal of this proposal is to build on recent R56 support to develop a portable, equitable, personalized risk-stratification tool to improve AF-related stroke prevention, a major NIH priority.^{12, 13} Our objectives are to: (i) discover new stroke risk factors for patients with AF, incorporating social determinants of health (SDoH) with millions of health record covariates, using an innovative comorbidity discovery framework (Poisson Binomial Comorbidity [PBC]);^{5, 7} (ii) combine these new risk factors with established ones using machine-learning (ML), in order to determine their network structure and provide explainability; and (iii) develop, deploy, and test a personalized stroke risk stratification tool for AF patients across different health systems in a disparity-aware fashion. Our central hypothesis is that stroke risk stratification can be improved through methods that: leverage all available data, including SDoH; capture and quantify synergies among known and newly discovered risk factors in socioeconomic context; and can be ported to other health systems, adapting to different populations. The rationale for this project is that current AF-related stroke risk management lacks the precision and awareness required to optimally implement treatments because it does not adequately account for (1) population diversity, (2) SDoH and disparities, (3) synergistic interactions among risk factors, and (4) novel, emerging risk factors. We will attain our overall objectives through the following specific aims:

Aim 1. Discover new clinical and socioeconomic relationships that influence stroke risk in patients with AF. *We hypothesize that the PBC approach will identify previously unrecognized factors and relationships associated with AF-related stroke.* Our PBC approach to co-variate discovery⁷ is specifically designed for big-data; it automatically controls for confounding on a per-patient basis and explores temporal connections, allowing us to conduct feature selection among millions of elements. PBC retains the power to identify relationships between factors and outcomes of interest, contrasting with stratified sampling.

Approach: We will deploy PBC on PCORnet datasets augmented with uniquely available SDoH. This will allow for primary discovery, cross validation, and investigations of the role of cross-site variance in the discovery process. The identified clinical and SDoH variables will be included in ML-based risk calculation machinery.

Aim 2. Develop a socially conscious, ML-based machinery for calculating personalized stroke risk among patients with AF. *We hypothesize that SDoH interact synergistically with some clinical variables more than others, and that there exist subgroups of patients for whom certain SDoH factors are critical for accurate risk stratification.* Our preliminary results show that many risk factors for stroke are conditionally-dependent upon one another; that their combined effects are not simply additive, and that impacts can vary widely depending on various SDoH. These synergies may explain how SDoH drive disparities in care and outcomes.

Approach. Building on recent work,^{5, 7, 22} we will use Probabilistic Graphical Models (PGMs) to combine known and novel variables (Aim 1) for interpretable personalized stroke risk stratification. PGMs are explainable ML tools that capture and quantify synergistic interactions among conditionally dependent variables (not captured by predictive models). They can provide more comprehensive, personalized, and equitable risk estimates.⁵

Aim 3. Benchmark an ML-based stroke risk stratification across a diverse cohort of health systems within PCORnet and discover biases and drivers of downstream care disparities. *We hypothesize that site-specific differences in healthcare environments and patient SDoH impact (a) performance of stroke-risk stratification tools and (b) disparities in care and outcomes.*

Approach: We will use ML-based discovery methods to calculate risk and compare to the clinical standard (CHA₂DS₂-VASc) within and across sites. Unlike traditional modeling where replication failures are terminal, the explainability of PGMs enables easy identification of variables driving performance, including SDoH. This will be used to improve model transportability and to better understand biases and downstream disparities.

Successful completion of this project will (1) increase our understanding of how biases and care disparities impact stroke risk across institutions, and (2) inaugurate a new era in precision medicine for patients with AF, providing truly personalized risk stratification for stroke that is, (a) portable, (b) disparity-aware, and (c) equitable. The resulting web-based electronic decision support tool will be poised for deployment in a future, pragmatic prospective trial of emerging stroke prevention strategies in AF.

B. RESEARCH STRATEGY

B.1. SIGNIFICANCE

Atrial fibrillation (AF) is the most common cardiac arrhythmia, and its prevalence is rising dramatically.^{23, 24} Nearly 6 million US adults have AF and it leads to significant morbidity and mortality, and more than \$6B in annual hospitalization costs alone.²⁴⁻²⁶ The primary source of AF-related morbidity and mortality is stroke, and AF-related strokes are clinically more devastating than strokes of other etiologies.^{8, 27} Systemic oral anticoagulation via vitamin K antagonism (e.g., warfarin), with an inherent risk of bleeding, has been the historic standard for stroke prevention in these patients. However, there have been significant improvements in strategies to prevent AF-related strokes, as well as to reduce sequelae of anticoagulation (i.e., bleeding).

B.1.2. While Stroke Prevention Therapies Have Improved, Stroke Risk Stratification Has Not

Since 2010, four novel, direct-acting oral anticoagulants (DOACs) have been approved for stroke prevention in AF.²⁸⁻³¹ These drugs have lower overall risk of bleeding with at least equivalent reduction in stroke compared with traditional warfarin therapy.⁴ The reduced bleeding risk of DOACs, and other agents in development, suggests they may be appropriate for patients at lower stroke risk (**Figure 1**).^{1, 20, 32} However, current risk tools are incapable of accurately identifying intermediate versus truly low stroke risk individuals (see §B.1.3); up to 7,000 yearly strokes might still occur among ~1 million ‘low-risk’ patients.^{1, 20, 33 34}

Furthermore, some patients either (a) cannot tolerate anticoagulation, or (b) have residual stroke risk despite anticoagulation. In this setting, left atrial appendage occlusion (LAAO) devices are available for the prevention of stroke among patients with AF.^{2, 3, 6} An emerging role for LAAO may include adjunctive therapy among patients at very high risk of AF-related stroke, in combination with DOACs. Yet we lack a precise risk stratification mechanism for patients with AF. In sum, current risk stratification tools limit optimal implementation of therapies.¹²

B.1.3. Current Stroke Risk Stratification for Patients with AF is not Precise, Portable, or Personalized

Contemporary, guideline-recommended risk stratification for stroke among patients with AF is guided by the one-year stroke risk estimated by the CHA₂DS₂-VASc score^{33, 34} – an ordinal score from 0 (lowest risk) to 9 (highest risk) and including 7 factors: sex, age, and history of heart failure, vascular disease, hypertension, diabetes, and/or prior stroke. It was primarily developed in a European registry of 1,084 patients with AF, including a total of only 25 stroke events.³⁵ Despite endorsements across international guidelines,^{33, 34} there are well-known limitations to the CHA₂DS₂-VASc score: (1) its derivation population is inherently limiting due to size and lack of diversity; (2) some components (e.g., sex) have not held up as robust risk factors in broader studies;³⁶ and (3) precision of the risk prediction is poor, particularly for patients with calculated low scores. For example, guidelines have moved away from sex as a risk factor, as it was not found in subsequent studies to be a robust predictor of AF-related stroke;^{34, 36} an observation supported by our own, preliminary analyses (see §B.3.4).

Precision: identifying subgroups with increased risk. Current guidelines recommend use of oral anticoagulation among patients with CHA₂DS₂-VASc score of ≥2 (excluding sex; **Figure 1**).^{33, 34} For patients with a score of 1 (excluding sex), anticoagulation may be considered, but there are little data guiding these decisions. Stroke risk may be elevated among some patients with scores in these ranges – these ‘**inappropriately low-risk**’ patients may have undiagnosed or unrecognized risk factors (e.g., chronic kidney disease), or social determinants of health (SDoH) placing them at higher risk, yet often are not treated with anticoagulation. Even in the CHA₂DS₂-VASc derivation cohort, the 1-year stroke rate estimate ranged from 0 to 4.7% for scores <2.³⁵ More precise data on stroke risk in contemporary cohorts with AF have been limited by either high anticoagulation treatment rates and/or low stroke event rates.²¹ And while numerous alternative stroke risk stratification

Table 1. Common abbreviations and defined terms used in this proposal.

AF: atrial fibrillation	GROUSE: Greater Plains Collaborative Reusable Observable Unified Study Environment
CDM: common data model	LAAO: left atrial appendage occlusion
Comorbidity: co-occurring medical diagnoses, procedures, medications, or any SDoH that are statistically associated with a clinical variable	PBC: Poisson binomial comorbidity discovery
DOAC: direct-acting oral anticoagulant	PGM: probabilistic graphical model
EHR: electronic health record	PCORnet: Patient-Centered Clinical Research Network
GPC: Greater Plains Collaborative	SDoH: social determinants of health
	UHealth: University of Utah Health

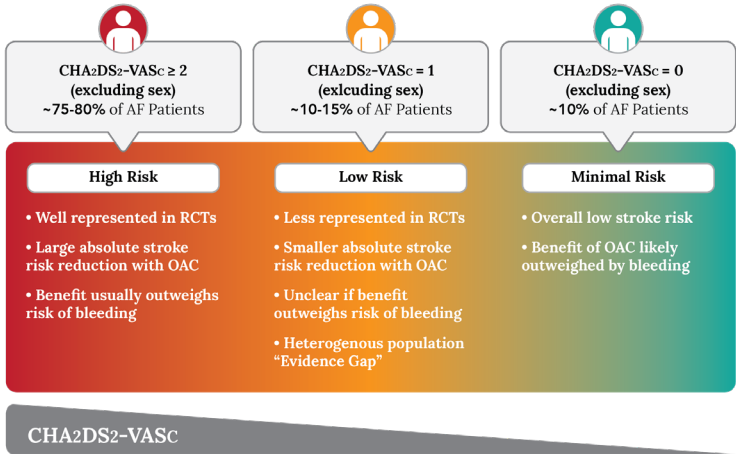


Figure 1. Overview of current stroke risk assessment and treatments for patients with AF.¹

approaches have been studied, none has been adopted to further stratify ‘low’ CHA₂DS₂-VASc patients.³⁷⁻⁴⁶ There remains no robust, guideline-recommended tool to identify patients who may be currently mislabeled as ‘low risk’.^{12, 20} Additionally, strokes still occur among patients with increased CHA₂DS₂-VASc scores despite appropriate oral anticoagulation therapy – at an *average* of nearly 2% per year.^{4, 42} Prospective identification of these patients with substantial ‘**residual risk**’ due to unrecognized risk factors and SDoH remains elusive.^{40, 47}

Personalization & Portability: adapting risk models to different populations. The CHA₂DS₂-VASc score to forecast one-year stroke risk is a very basic calculation of broad risk categories, designed for simple implementation. Despite the wide availability of electronic health records (EHR) and data warehouses, no additional patient or system-specific data are routinely integrated into stroke risk stratification for patients with AF. Recognized shortcomings of the CHA₂DS₂-VASc score include: (a) no means for integration of disease severity (e.g., heart failure); (b) no means to address synergy (i.e., conditional dependencies) between and among risk factors; and (c) the score is blind to variability in risk among different cohorts, particularly where disparities result from SDoH. Numerous emerging strategies have been proposed to improve on CHA₂DS₂-VASc for the prediction of stroke associated with AF, including imaging characteristics,⁴⁸ serum biomarkers,^{43, 46} and non-portable, ‘black-box’ machine-learning approaches.⁴⁹⁻⁵¹ Yet there remain significant limitations to these approaches, including: (a) challenges in ‘explainability’; (b) necessity for additional, potentially burdensome testing (e.g., biomarkers, imaging); (c) lack of transportability to diverse cohorts; (d) limited precision; and (e) not leveraging the full breadth of EHR data in terms of millions of factors available for consideration.⁵ We will specifically overcome these limitations by discovering drivers of risk across varying populations to create an explainable personalized and portable risk model that addresses disparities in treatment and outcomes.

B.1.4. Despite Known Disparities in Care and Outcomes among Patients with AF, SDoH Are Not Included in Stroke Risk Assessments

Limitations in stroke risk stratification can be magnified across diverse populations of patients with AF, where disparities in care are well-described (**Table 2**).¹³ For example, Black patients are at a lower risk of developing AF, yet Black patients with AF are less likely to receive guideline-recommended stroke-prevention therapy in the form of oral anticoagulation (particularly newer drugs).^{8, 14, 15, 19} Consistently, variables such as insurance status are associated with access to more complex interventions – uptake of DOACs was frequently dictated by cost.⁵²⁻⁵⁴ More broadly, SDoH (other than race) have been found to correlate with AF incidence and treatment even in locales with universal healthcare access and funding: rates of AF diagnosis can vary by geography, education, and marital status, often independent of reported income level.^{9, 10}

Outcomes for AF also vary according to SDoH. Black patients with AF have been reported to have a 3-fold higher risk of AF-related stroke and an increased risk of mortality, compared with Whites.^{17, 18} This may be in part due to Black patients’ worse anticoagulation control, a powerful predictor of clinical outcomes.^{11, 16} Even in universal healthcare settings, lower socioeconomic status is associated with both lower rates of interventions, and worse clinical outcomes including mortality.⁵⁵ The mechanisms driving these variable outcomes remain under-studied and, crucially, SDoH are not incorporated into routine, clinical decision-making.¹³ In the current proposal, we refer to methodological approaches that are socially-conscious (SDoH are candidate variates), in order to develop tools that are socially-aware (informed of the impact of SDoH), with the goal of making care more equitable.^{12, 13}

Table 2. Disparities in care and outcomes of patients with atrial fibrillation. ^{7-9,12-19}
Black patients with atrial fibrillation <ul style="list-style-type: none"> • Less likely to receive guideline-recommended therapy <ul style="list-style-type: none"> - Less likely to receive direct-acting anticoagulants • Higher risk of stroke and mortality
Less-insured patients with atrial fibrillation <ul style="list-style-type: none"> • Lower access to direct-acting anticoagulants • Less likely to undergo complex procedures
Other social determinants of health (geography, education, social) <ul style="list-style-type: none"> • Less likely to be diagnosed with atrial fibrillation • Less treatment of atrial fibrillation • Higher mortality

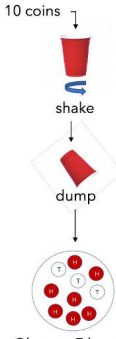
B.1.5. Significance of the Expected Research Contribution

This work represents a significant departure from standard approaches to stroke risk stratification for patients with AF. Using (a) novel, explainable, ML approaches and (b) unique, individual-level, structured SDoH data across multiple sites, we aim to uncover stroke risk factors and their relationships that would otherwise remain undiscoverable. Moreover, we have designed these analyses with an eye towards implementation – we will be able to deploy a risk tool that is adapted to the target population, and that can describe the incremental value, if any, of including SDoH in the risk assessment for AF-related stroke. This will allow for more personalized and precise risk assessment, and it will improve our understanding of disparities in treatments and outcomes among patients with AF.

B.2. INNOVATION

B.2.1. We Developed a Novel and Rigorous Approach to Risk Stratification in Health System Cohorts

There is strong evidence demonstrating heterogeneity of risk in AF populations – risk of stroke has been found to vary by demographics, comorbid disease, frequency of AF, and after interventions.⁵⁶⁻⁵⁹ Attempts to incrementally improve stroke risk stratification using classic regression modeling with stable parameters applied across all AF populations^{43, 46, 48} or leveraging large EHR datasets using alternative machine-learning methods⁴⁹⁻⁵¹ have limitations in terms of the ability to (a) model conditional dependencies across a landscape of relevant factors and (b) provide clear explainability.⁶⁰⁻⁶⁸ In response, we have developed novel and rigorous approaches to feature selection, landscape discovery, and risk calculation using EHRs.⁵



10 coins
shake
dump
Observe 7 heads, 3 tails

The Binomial test (AKA χ^2 , for large N)

$$\Pr(X \geq k) = 1 - \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i}$$

Every coin (or patient) must have same probability of outcome, e.g. $p = 0.5$

The Poisson Binomial test

$$\Pr(X \geq k) = 1 - \sum_{i=0}^k \sum_{A \in \mathcal{P}_i} \prod_{j \in A} p_j \prod_{j \in A^c} (1-p_j)$$

Every coin (or patient) can have a 'personal' probability, of outcome, e.g. $p_1 = 0.5, p_2 = 0.61, p_3 = 0.21 \dots$ etc.

Box 1. The Binominal vs. the Poisson Binomial test. For ten coins shaken from a cup, a Binomial test gives probability of observing, say, >6 heads assuming every coin has the same probability of heads (e.g., 0.5). In contrast, the Poisson binomial does not have this limitation; each coin can have an individual probability of heads. This is more appropriate for patients, where each has a different probability of an outcome (e.g., stroke). *The PBC approach makes it possible to model personalized patient risks during comorbidity discovery.* See ⁷ for details.

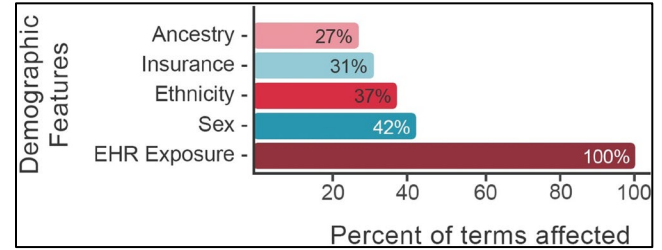


Figure 2: Percent of medical terms influenced by various demographic features in the UHealth cohort. Demographic variables used in the comorbidity discovery process on the y-axis. Percent of all comorbidities influenced by a given demographic feature on the x-axis; e.g., sex influences 42% and EHR exposure 100% of diagnoses, procedures, and medications in UHealth EHRs. EHR exposure includes age, length of medical record history, and number of visits. PBC uses logistic regression to model how each patient's demographics impacts probability of having a medical term. See ref ⁷ for details.⁸⁻¹⁰

Traditional approaches to risk factor selection, including the CHA₂DS₂-VASc, have started with review of comorbid conditions within small curated cohorts.^{35, 69, 70} In point of fact, the CHA₂DS₂-VASc score was based primarily on a cohort of only 1,084 patients, with only 25 stroke events.³⁵ The contemporary availability of broad and deep EHR data provides ready access to nearly all relevant health data associated with an outcome. **Yet traditional analytic methods for these data assume that each patient has a disease probability equal to the population incidence rate.** Confounding variables are usually controlled for by sub-setting the data (i.e., stratification), at significant expense in terms of statistical power.^{7, 71, 72}

Our group has developed a novel and rigorous approach to risk-factor identification and prediction across healthcare populations using high-dimensional EHR data.^{5, 7, 73} Leveraging modern, high-powered computational workflows, we have developed a Poisson binomial-based approach to comorbidity discovery (PBC; **See Box 1**).⁷ Note: to simplify discussion, throughout this proposal, we refer to co-occurring medical di-

agnoses, procedures, medications, or any SDoH that may influence risk using the single, blanket term, **comorbidity**. The PBC approach models how each patient's demographics impact their probability of an outcome (**Figure 2**, details in ref ⁷). These personalized probabilities are then used to calculate pairwise expectations and P-values between specific co-morbidities under the Poisson binomial distribution, rather than the hypergeometric and binomial distributions that are used for Fisher's exact test and the χ^2 test, respectively. Moreover, this approach can also be used to temporally order comorbidities and to determine the significance of directionalities. We have shown that the PBC approach provides a considerable advantage over traditional, stratification-based approaches.⁷ We compared PBC to stratification-based statistics to understand

Table 3. Poisson binomial comorbidity (PBC) approach consistently demonstrated superior statistical power to identify factors associated with AF and stroke in University of Utah Health data.

	PBC p-value			X ² p-value		
	N=1,538,059	N=95,407	N=9,525	N=1,538,059	N=95,407	N=9,525
No Features	1E-31020	1E-1715	1E-203	1E-31020	1E-1715	1E-203
+ sex	1E-31017	1E-1955	1E-215	1E-16657	1E-1125	1E-147
+ age	1E-25448	1E-1589	1E-200	1E-1304	1E-88.3	1E-13.1
+ ancestry	1E-14381	1E-628	1E-73.1	1E-15.72	1	1
+ ethnicity	1E-11357	1E-806	1E-110	1E-12.25	1	1
+ insurance	1E-11533	1E-771	1E-83	1E-2.68	1	1
+ span	1E-11325	1E-698	1E-84.1	1E-1.75	1	1

Table 3. Progressively smaller matched random UHealth samples, such that each cohort is a subset of the precursor. Cells are p-values for the association between AF and stroke, by PBC or χ^2 . P-values less than the Bonferroni corrected alpha ($1E-9.5$) in red bold; cells above the significance threshold are black. Stratum filters apply to the features' column, row by row: no filters, female, age 50–59 years, white, non-Hispanic, commercial insurance, ≥ 2 years of medical history (span).⁵

the impact of specific covariates on the relationship between AF and stroke among University of Utah Health (UHealth) patients.⁵ As shown in **Table 3**, despite decreasing cohort size, the PBC approach consistently demonstrated superior statistical power to identify factors associated with AF and stroke. The result is **increased power for discovering new factors, which we will leverage to explore the relationships among AF, stroke, demographics, comorbid diseases, medications, and SDoH.**

Figure 3 illustrates the power and scalability of the PBC approach for *ab initio* discovery of comorbid clinical and demographic variables.⁵ **Figure 3A** visualizes the UHealth Patient Disease Network discovered using PBC. It summarizes all significant comorbid relationships from the millions of factors among 1.6 million patients. Each node (circle) is a diagnosis, procedure, or medication. The edges (lines) between each node denote statistically significant comorbid relationships discovered using PBC. **Figure 3B** shows a small subgraph of the network centered on Heart Transplant. Here, nodes correspond to diagnoses (black), procedures (red), and medications (blue). Edges are temporally ordered comorbidities with arrows denoting direction. Edges are labeled with marginal transition probabilities (e.g., patient flux). For example, within UHealth 4.9% of patients diagnosed with viral myocarditis will eventually undergo heart transplantation, and 54% of those patients will be prescribed tacrolimus post-transplant.

Probabilistic Graphical Models (PGMs)⁵ are then used to calculate the joint distribution of the multiple variables that the PBC has identified as having a substantial adjusted association with a specific outcome, thereby describing the surrounding landscape that influences the outcome (so called ‘multi-morbidity’ calculations). For more on these points see ref ^{5, 7}. **Unlike neural nets, PGMs can more readily provide fully interpretable ML.**^{5, 60-68} These features combine to make PGMs ideal for teasing apart the intertwined, conditionally-dependent impacts of comorbidities and SDoH upon stroke. **Figure 4** illustrates how pairwise comorbid relationships discovered by PBC can be recast into a PGM wherein the edges now represent conditional dependencies between variables. The utility of PGMs for risk calculations is best illustrated by problems of complex *multi-morbid* landscape, where conditional dependencies between variables interact to further modulate outcome risk. For example, while a patient with cardiomyopathy has an 86-fold increased risk for heart transplant, and a patient prescribed milrinone is at 175-fold increased risk, a cardiomyopathy patient *with* a milrinone prescription is 407 times more likely to receive a heart transplant – a risk greater than the sum of the parts (e.g., 175 + 86 ≠ 407). **The ability of PGMs to capture conditional dependencies (synergies) between variables is a major advantage and innovation for risk calculations.**

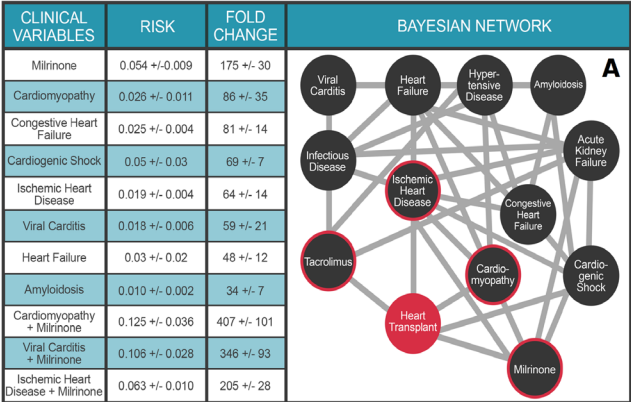


Figure 4: Probabilistic Graphical Model (PGM) for Heart Transplant (n = 1.6 million patients). Variables chosen based on Bonferroni-corrected codes preceding and significantly associated with heart transplant. Each node represents a diagnosis, procedure, or medication code and each edge a conditional dependence between nodes.⁵ Note that because the edges in the PGM shown denote conditional dependencies, rather than transition probabilities as in Figure 3B, the PGM’s topology is necessarily different from the trajectory topology in Figure 3B.^{5, 8-13}

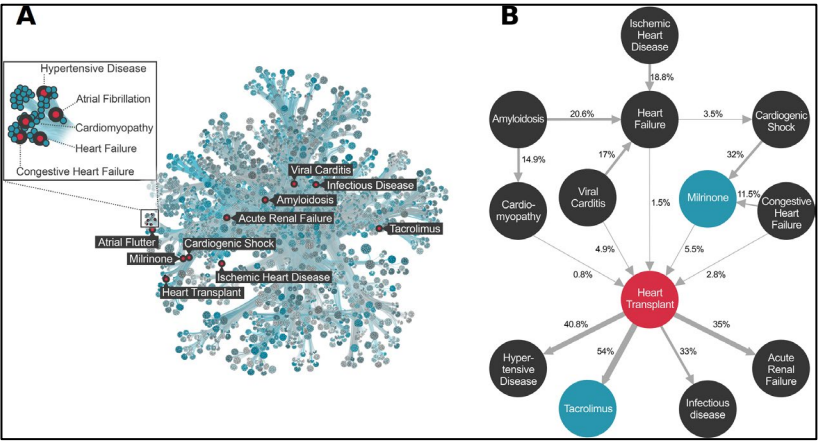


Figure 3: PBC-based Comorbidity Network for UHealth. Panel A: Graphical representation of 39,055 diagnosis codes, 5,716 procedure codes, and 1,764 medication codes comprising 50 million pairwise-comorbidities. Panel B: Term trajectory for Adult Heart Transplant. Edges are temporally ordered comorbidities (Bonferroni alpha = 10E-9.5), and labeled with transition probabilities (e.g., patient flux).⁵

Another strength of PGMs is that they enable risk stratification using variables belonging to different modalities — diagnoses, procedures, demographics, laboratory studies, etc. **Figure 5** provides an example using the outcome of sinoatrial node dysfunction (SND). This PGM includes variables from multiple modalities: self-reported ancestry, sex, and ethnicity; diagnoses (AF, coarctation of the aorta); measurements such as obesity; and even the insurance type. Note that lifetime risk of UHealth patients for SND with a prior diagnosis of AF is equal for Caucasians and African American patients. Risk, however, doubles for Hispanic patients (see ⁵ for additional details). The ability to combine

diverse datatypes for risk calculations is a major asset for our SDoH-based investigations. Our proposal leverages a diverse collection of SDoH variables available through our collaborations with PCORnet and Acxiom, a commercial, consumer data company (see LOS: Acxiom, LLC and Dr. Russ Waitman). We will use PBC to discover new clinical and SDoH comorbidities, and PGMs to discover how they synergize and impact stroke outcomes.

B2.2. Commercial-Based Individualized Data Elements are an Untapped Resource for Social Determinants of Health (SDoH).

In the digital era, EHRs are only one source of data for precision medicine. In order to better-understand SDoH, data elements describing the individual's life *outside* the healthcare system are required.⁷⁴ The most common approach to SDoH is the use of grouped geographic data – zip codes or regional characteristics that summarize a group.^{74, 75} Yet while characteristics such as poverty are associated with mortality, merely living in an area associated with poverty is not.^{74, 76} In order to better disentangle the relationships between health and SDoH, specific, individual data elements are needed.^{13, 77} For example, individual employment and marital history have been associated with stroke risk.⁷⁸ However, these elements are historically collected via patient surveys – an inefficient and biased approach.

More broad and detailed individual data are available in the form of consumer preference data. While these data types have been long-used for marketing purposes, consumers have provided guidance on the use of such data for research purposes.⁷⁹ In a limited study of volunteered receipt data, Danish household diabetes risk was found to correlate with consumer grocery purchasing.⁸⁰ Larger, broader repositories of such data represent an untapped source of potentially-informative SDoH data to better-understand stroke risk. A **major innovation** in our approach is the inclusion of data from Acxiom, LLC – a commercial, consumer data and marketing company. Acxiom houses individualized data for 2 billion people worldwide and includes elements not available in any health system EHR data, such as household demographics, finances, spending patterns, and leisure interests (thousands of variables). The current proposal will include 85 items pre-selected by the Greater Plains Collaborative Research (GPC) Network as especially informative social determinants of health. Linkage of the Acxiom data to PCORnet site data will be facilitated through collaboration with GPC lead investigator, Dr. Russ Waitman, PhD (see §B.3.2, and LOS). We will apply our novel analytic methods, for the first time adding individual consumer data, to address major gaps in risk stratification for AF-related stroke.

In summary, the proposed research is innovative and a significant departure from prior work because it:

- Leverages a novel, portable, and highly personalized approach to risk stratification:
 - PBC discovery allows for a massively scalable means to tease apart relationships among high-dimensional covariates, including but not limited to SDoH.
 - PGMs enable the simultaneous modeling of influential factors that form the surrounding landscape of the outcome. They facilitate the identification of multilevel dependencies and synergies among the factors, as well as the interpretation of probabilistic dependencies and patterns that contribute to outcome prognosis.
- Includes data from multiple, diverse sites, addressing disparity and portability at the development stage:
 - Diverse cohorts *at derivation and validation* phases, to guide model development at all stages.
 - Instead of applying the proposal to data from a single site (UHealth), we will make it portable and modify it to accommodate various populations.
- Incorporates uniquely available, structured, and individualized data on SDoH:
 - Acxiom consumer data add phenotyping never used in this space.
 - Individualized data offer the best opportunity to understand SDoH, risk, and disparities thereof.
- Will be executed by accomplished investigators with highly-complementary skillsets:
 - Our team includes internationally renowned researchers in AF and stroke, personalized medicine, bioinformatics, biostatistics, computational biology, administrative data research, and

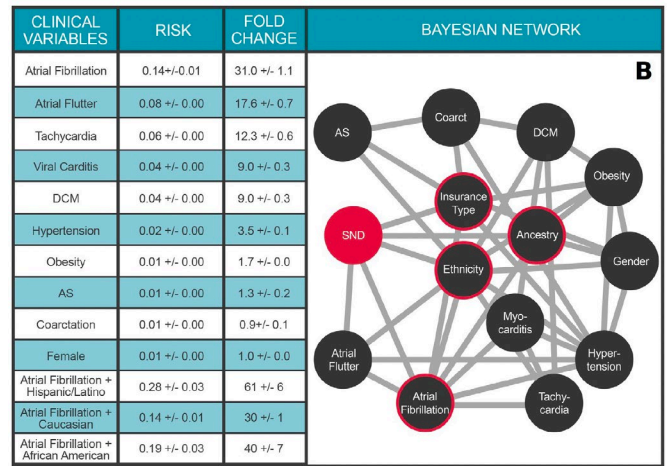


Figure 5. A PGM for Adult Sinoatrial Node Dysfunction (SND). Each node represents a diagnosis or procedure, each edge a conditional dependence between nodes. N = 1.6 million patients. Clinical variable terms represent CCS billing codes. DCM: Dilated cardiomyopathy; AS: Aortic stenosis; Coarct: Coarctation of the aorta.⁵

health care disparities and bias (see §B.3.1, below).

B.3. APPROACH

B.3.1. Team

The proposed research will be led by Dr. Steinberg (**PI**), a clinical cardiac electrophysiologist and recognized authority in the space of AF-related stroke risk assessment and treatment.^{81, 82} Dr. Yandell (**Co-I**) brings to the team decades of experience as a computational scientist and leadership as regards the use of PGMs for risk stratification.^{5, 7} Dr. Tristani-Firouzi (**Co-I**) will oversee the development of tools to enable precision medicine approaches to AF-related stroke risk stratification.^{73, 83-85} Dr. Xu (**Co-I**) will provide statistical support for method development as well as computational implementation. We are joined by experienced collaborators to ensure smooth data flow from the respective sources (**Dr. Spinka**, Data Scientist and lead for the GPC Reusable Observable Unified Study Environment (GROUSE) project; and **Dr. Waitman**, PI of the GPC). Additionally, we have engaged and garnered the support of key thought leaders: **Dr. Eric Peterson**, Vice Provost and Senior Associate Dean for Clinical Research and Vice President for Health System Research at University of Texas Southwestern Medical Center, and a world-renowned cardiologist and researcher focusing on clinical studies that leverage large datasets to improve health through more personalized care; and **Dr. Kevin Thomas**, Vice Dean for Diversity, Equity and Inclusion at Duke University and a practicing cardiac electrophysiologist, who has devoted his career to improving disparities in care for patients with cardiovascular disease (see LOS).

B.3.2. Partners and Data Sources

Patient-Centered Clinical Research Network: The Greater Plains Collaborative

The national Patient-Centered Clinical Research Network (**PCORnet**) is an integrated partnership of large clinical research networks. As a participating center in PCORnet, UHealth belongs to the Greater Plains Collaborative (GPC), a collection of 13 medical centers across 8 states encompassing 21 million patients.⁸⁶ Eight of these PCORnet sites, comprising ~**17.9M** patients, have already committed to allow use of their data for the current proposal (see **Table 4**; LOS from sites and GPC Coordinating Center). Collaboration with PCORnet sites will be directly facilitated by the GPC Principal Investigator, Russ Waitman, PhD (see LOS). Access and analysis of these data will be performed through the Greater Plains Collaborative Reusable Observable Unified Study Environment (GROUSE).⁸⁷ Each site's data is available within GROUSE in a standard common data model (CDM) format, facilitating analyses across sites. GROUSE access will be facilitated by collaboration with

Table 4. Summary of PCORnet sites and patient populations so far committed to participation in our study. Race percentages do not sum to 100% due to other categories (other races, multiple races, missing data).
Total patients: 17,896,789

GPC Site	Patients	Age (mean)	American Indian or Alaska Native	Asian	Black or African American	Native Hawaiian/ Pacific Islander	White
Allina Health System	2,511,597	44	0.76%	4.02%	8.55%	0.28%	80.62%
Medical College of Wisconsin	1,524,707	47	0.35%	2.17%	14.43%	0.08%	68.34%
University of Iowa	1,258,443	43	Not Yet Available				
University of Kansas	2,935,242	47	0.13%	0.80%	4.63%	0.06%	35.38%
University of Missouri	2,144,807	49	0.08%	1.21%	4.63%	0.04%	58.47%
University of Texas Southwestern	4,491,723	39	0.20%	3.38%	12.50%	0.14%	44.41%
University of Utah	1,414,837	42	0.86%	2.57%	1.97%	0.33%	71.06%
University of Nebraska	1,615,433	50	0.32%	1.20%	4.53%	0.13%	44.82%

a GPC investigator experienced in the use of these data, **Christine Spinka, PhD**.

GPC Patient Stakeholder Engagement: The GPC Rapid Community and Engagement (Rapid PACE) resource consists of trained patient research advocates from diverse communities across GPC sites. We have already engaged these informed patient stakeholders and incorporated their feedback on the current proposal and will continue to leverage their expertise through each phase of the research (see Rapid PACE LOS).

Axiom, LLC

Household and individual socioeconomic data are widely used in consumer marketing, with some compiled from public records and lifestyle/interest data self-reported from consumer surveys.⁸⁸⁻⁹⁰ Axiom houses individualized data on 2 billion people worldwide (including all ~260 million Americans), elements not available through any health system. These include data on household demographics, finances, spending patterns and preferences, employment, and leisure interests across thousands of collected elements. The current proposal will include 85 pre-selected items related to social determinants of health that have been determined by our GPC colleagues as especially informative SDoH. With R56 support, we have already linked some of these

data to the local UHealth population (using names and dates of birth; see §B.3.4). Links between Acxiom data to other PCORnet site data will be facilitated through collaboration with GPC lead investigator, Dr. Russ Waitman, PhD (see **Figure 6**; see LOS). Dr. Waitman’s team also has already independently linked Acxiom data to >80% of their local site population, also. Data for > 99% of all US adults are available through Acxiom. We thus anticipate very high linkage rates in the remaining PCORnet sites’ data. As part of this process, they will integrate Acxiom variables by having the sites send finder files directly to Acxiom. They will ensure quality control of the data linkage and integration into the CDM, including an evaluation of missing data rates in our PCORnet population. They will also evaluate ‘inter-rater’ reliability agreement between EHR and Acxiom for demographic variables such as race and ethnicity which are also in many patients’ EHR records. Finally, we will use the resulting analytic dataset, including EHR data linked to the key 85 Acxiom data elements for all available AF patients, to pursue variable selection for our PGMs.

B.3.3. Inclusion Criteria and Definitions

Across all aims and datasets, our disease-based cohort, exposure covariates, and outcome variables will be defined consistent with prior use of EHR data in clinical research for AF, by our group and others.⁹¹⁻⁹⁵ In brief, we will require adequate observation time, in the form of run-in periods and follow-up periods of at least 1 year each. AF exposure will be defined by the presence of >1 AF international classification of diseases (ICD) diagnostic codes within 1 year, including at least one from an outpatient encounter. We will exclude patients with valvular AF (i.e., mechanical valve and/or mitral stenosis), as their stroke risk is uniquely elevated and the CHA₂DS₂-VASc score is not applicable. The index date is defined by the first AF diagnosis code, at which time baseline comorbid diseases and prior interventions are measured. Comorbid diseases and outcomes (i.e., stroke) are defined by ICD code combinations previously validated in administrative data research in cardiovascular disease.^{91, 92} In addition to the medical components of the CHA₂DS₂-VASc score, they include other clinical diagnoses that may be associated with stroke risk, such as (but not limited to) chronic kidney disease, pulmonary disease, dementia, malignancy, liver disease, and blood dyscrasias.

Incident outcome events (stroke) are measured based on *primary* ICD codes for emergency room or inpatient encounters, thus differentiating a new event from documentation of a prior event. As AF is both a causative and correlative risk factor for stroke, all stroke types are included, an approach consistent with CHA₂DS₂-VASc derivation and the field at large.^{35, 96-103} We have previously used these definitions for analyses within UHealth data,^{5, 95, 104-106} as well as within Medicare and PCORnet datasets.^{100, 107, 108} Rates of AF in major health systems are consistently 2-3%, yielding an overall AF cohort across sites of >350,000.^{109, 110} The rate of stroke among UHealth patients with AF is 5-6%; given the overall sample size of AF patients, we consider that there are sufficient patterns in the observed dataset to adequately represent the distribution of stroke risk.

Throughout the proposal, we use several terms that are variably defined in different literatures across epidemiology, biostatistics, clinical research, and machine-learning and computational science. For clarity and consistency, we have defined our usage in **Table 5**.

B.3.4. Preliminary Data

With R56 support (#R56HL168264), we have used the PBC and PGM methods described above to construct preliminary models of AF-related stroke risk and SDoH among 1.6 million patients within UHealth. The preliminary results of single-site PBC discovery are shown in **Figure 7**. *These are constructed using a PGM approach, where the*

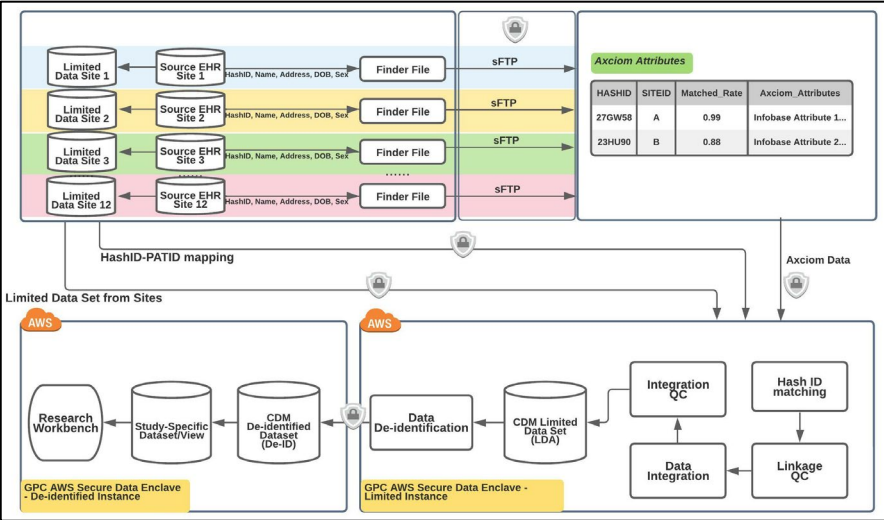


Figure 6. Data flow for linkage between Greater Plains Collaborative and Acxiom (Figure courtesy Russ Waitman, PhD and Xing Song, PhD). GPC: Greater Plains Collaborative; AWS: Amazon Web Services.

Table 5. Mathematical definitions for terms used in this proposal.		
Risk Calculation	Abbreviation	Equation
Relative Risk	RR	$\frac{P(\text{Outcome} \mid \text{Specified Criteria}) \text{ in General Population}}{P(\text{Outcome} \mid \text{No Specified Criteria}) \text{ in General Population}}$
Absolute Risk	AR	$\frac{\# \text{ of Individuals with Outcome in General Population}}{\# \text{ of Individuals in General population}}$
Absolute Risk Ratio	ARR	$\frac{P(\text{Outcome} \mid \text{Specified Criteria}) \text{ in General Population}}{\text{Absolute Risk of Outcome in General Population}}$

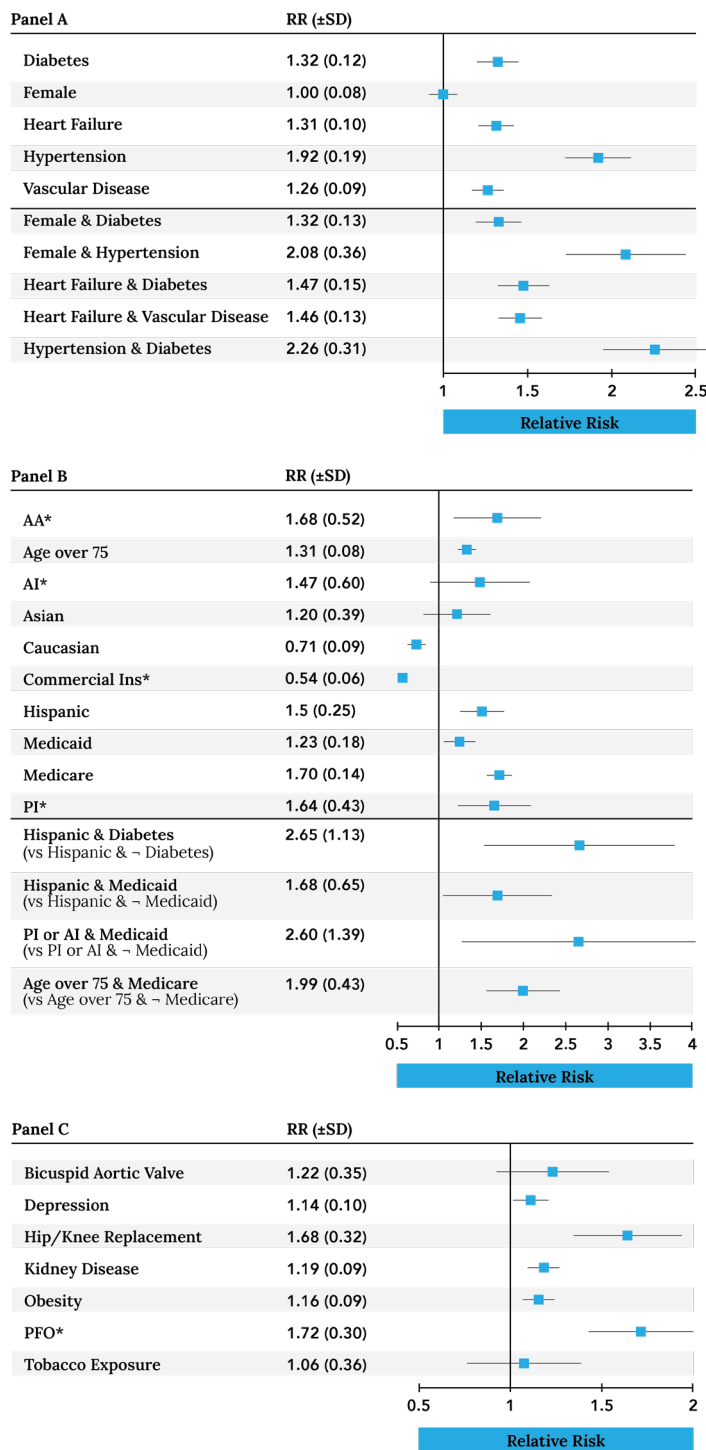


Figure 7. PGM-Based Relative Risks for Stroke among UHealth Patients with AF. **Panel A:** Various CHA₂DS₂-VASc factors (top half all represent CHA₂DS₂-VASc scores of 1; bottom scores of 2). **Panel B:** Impact of selected SDoH on stroke risk. **Panel C:** Risk imparted by selected factors discovered *ab-initio* using PBC. *AA= African Ancestry (Black), AI= American Indian, Ins = Insurance, PI= Pacific Islander, \neg = absence of variable, SD = Standard Deviation, PFO = Patent Foramen Ovale.

We will deploy PBC on PCORnet datasets augmented with uniquely available SDoH. This will allow for primary discovery, cross validation, and investigations of the role of cross-site variance in the discovery process. PBC-identified clinical and SDoH variables will be included in ML-based risk calculation machinery (Aim 2).

PBC deployment on PCORnet data: The Common Data Model (CDM) is the fundamental pillar that supports

probability calculations rely on the identified conditional dependencies. Our results highlight the limitations of over-simplifying risk with CHA₂DS₂-VASc components – risk factors such as diabetes (1.32x) and hypertension (1.92x) do not yield similar probabilities of stroke, though they are accorded the same value in the CHA₂DS₂-VASc score (1 point each). Moreover, among patients with CHA₂DS₂-VASc scores in the treatment ‘twilight’ zone (~2), there is wide variability in relative risk (1.32x-2.26x).

These results also demonstrate that SDoH, such as race, ethnicity, and insurance status can impact stroke risk (**Figure 7B**). For example, Medicaid patients of American Indian or Pacific Islander ancestry have ~2x risk of stroke compared with Hispanic patients. More granular SDoH data will help tease out these effects – **we have already acquired data from Acxiom for our UHealth site and linked them to >500k patients, demonstrating the feasibility of this process**. Preliminary integration of Acxiom data into the PBC discovery infrastructure demonstrates factors such as multi-generational housing, education, occupation, and household children are strongly associated with AF-related stroke risk ($p < 1E-6$). Additionally, a variety of other variables may be associated with AF-related stroke risk in different patient populations, which have not been previously recognized (**Figure 7C**). While the mechanisms of these preliminary findings are beyond the scope of this discussion, they show that our PBC and PGM methodologies provide means for their discovery and to quantify their impacts, crucial first steps towards better understanding stroke risk (Aim 1).

Importantly, event rates for some groups have wide margins of error, owing to small populations even across our entire UHealth site. For example, the Black population in UHealth is low, whereas UHealth is relatively enriched with persons of Pacific Islander heritage. This illustrates the value of multisystem data for refined risk calculations, yielding cohorts that are orders of magnitude larger than any single system (**Figure 8**). Data from multiple PCORnet sites, with linked Acxiom SDoH elements, will provide the power to pursue the proposed aims, specifically geared towards understanding SDoH in traditionally under-represented groups (c.f. **Table 4**).

B.3.5. Aim 1: Discover new clinical and socioeconomic relationships that influence stroke risk in patients with AF. We hypothesize that the PBC approach will identify previously unrecognized factors and relationships associated with AF-related stroke.

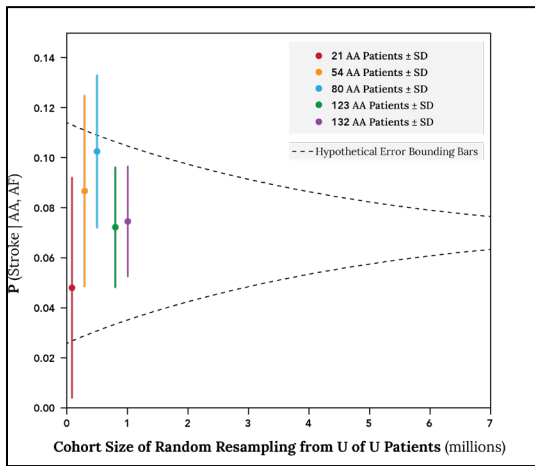


Figure 8: Standard Deviation (SD) of Stroke Risk in Patients of African Ancestry (AA) decreases with increasing number of observations. Bounding bars for SD show accuracy grows with cohort size, and SD decreases. Presently, SD is large for all data points, as AA individuals are uncommon in UHealth. The GPC collaboration (c.f. **Table 4**) will increase this number more than 10-fold.

between a comorbidity (e.g., diagnosis, procedure, SDoH) and the probability of having a stroke. As medical terms are included or excluded from year to year, and coding practices vary over time, we will also use the date of the patient's last visit to inform the temporal order between factors and outcome. The result will be a patient-comorbidity landscape which describes the significant associations between every possible comorbidity factor (c.f. **Figure 3A**) and stroke. Using these results, we will select features for inclusion in the interpretable risk stratification tools (see Aims 2 & 3) based on clinical relevance and strength of association, and absolute risk. Candidates will include Acxiom data elements of SDoH, in addition to clinical data in the PCORnet CDM (described above).

Effect of therapy: A key limitation of recent alternative approaches to improving risk stratification for stroke has been the widespread adoption of oral anticoagulation and thus lack of an 'untreated' population. Through our approach, we can include chronic oral anticoagulation as a specific data element, with temporality; this will allow us to understand stroke risk before and after treatment, across the population.⁷

Cross Validation. The multisite nature of the PCORnet dataset affords powerful opportunities for validation, and investigations of how stroke risks vary across institutions. We will perform primary discovery on the overall dataset, then study replicability at the site level. These analyses will allow us to explicitly model site as a variable in our PGMs (see Aim 3 for details). Briefly, we will use standard cross-validation procedures within and between PCORnet sites to explore differences in risk magnitudes, reproducibility, and variance for outcomes, the contributions of comorbid clinical factors, and for SDoH variables. We will use the results of these analyses to control for false discovery rates during PBC-based feature discovery (see refs ^{5, 7}, for methods).

B.3.6. Aim 2. Develop a socially conscious, ML-based machinery for calculating personalized stroke risk among patients with AF. We hypothesize that SDoH interact synergistically with some clinical variables more than others, and that there exist subgroups of patients for whom certain SDoH factors are critical for accurate risk stratification.

Building on our recent work,^{5, 7} we will use PGMs to combine known and novel clinical and SDoH variables (Aim 1) for personalized stroke risk stratification. PGMs are explainable ML tools that capture and quantify synergistic interactions among conditionally dependent variables (poorly captured by current regression models). As our previous work showed, PGMs can provide more accurate, personalized, and equitable risk estimates.⁵

PCORnet, empowering exchange of standardized data sets that facilitate large-scale, multi-site research. The PCORnet CDM leverages standard terminologies and coding systems for healthcare, including ICD/SNOMED (diagnoses, clinical observations, and observation-qualifiers), CPT (procedures), RxNORM (medications) and LOINC (labs). We will comply with established governance requirements and procedures for data requests from the PCORnet sites, including data use, patient rights and protections, misconduct and deviations, and data request oversight.

Acxiom: Our contract with Acxiom will allow use of up to 1 million individuals' records – enough to include all available patients with AF from the participating PCORnet sites. These records will be linked using unique identifiers by the GPC coordinating center, University of Missouri (see LOS). This will allow us to augment the PCORnet data with novel features that may represent SDoH – we will use PBC to determine which Acxiom variables are significantly associated with stroke outcomes (a sample is shown in **Table 6**).

Patients with AF from all participating PCORnet sites will be included. Using PBC,⁷ we will model the effects of confounding variables on the stroke outcome, including how each patient's demographics (c.f. **Figure 2**) significantly impacts the relationship

Table 6. Sample of Acxiom data elements included in the 85 GPC selected features.

Description	Match/Coverage	Granularity
Indicates whether a household is owner or renter occupied.	100.00%	Household
Indicates the years a household has lived at their address.	100.00%	Household
Indicates the known number of children in the household.	100.00%	Household
Indicates whether anyone in the household is married.	100.00%	Household
Indicates the number of adults (18+) living in the household.	100.00%	Household
Indicates the estimated household income in narrow ranges.	100.00%	Household
Resource based index of household spending capacity rather than strictly a measure of available assets.	95.07%	Person
Indicates the highest known education of the person	77.95%	Person
Indicates ancestry of the person	99.98%	Person

Developing Probabilistic Graphical Models (PGMs)

PGMs provide means to determine the combined effects of multiple comorbidities on an outcome and the synergies between the comorbidities, showing an interpretable landscape of dependencies that lead to stroke.⁵ Building upon our recently published approach,^{5,7} we will develop PGMs for the outcome of stroke among patients with AF at each PCORnet site, using features selected using PBC and validated by AF experts. We learn the structures of the PGMs using the python3 package *pomegranate*,¹¹¹ which provides a Bayesian Information Criterion-based DP-A* exact structure search algorithm for creation of Bayesian networks.¹¹²⁻¹¹⁴ The search algorithm explores the applicable space of conditional dependencies in order to discover the optimal network structure for the data. Parameter learning for this optimal network will be accomplished using the loopy belief propagation algorithm.¹¹⁵ We use the same package for inference and multi-morbidity risk calculations. The visual interpretation is designed with Python3 package *graph_tool* and D3.js Java library.¹¹⁶

Features that are judged to be of clinical relevance will be used as inputs to learn the PGM structure and infer risk. Demographic characteristics are described in a categorical data format, (e.g., ancestry, ethnicity, or insurance type) or “present/absent” binary variables in case of medical diagnoses and procedures. The absence of an entry in a medical diagnosis indicates the absence of a disease state that could potentially be present, whereas the presence of an entry indicates the existence of both an assessment and a condition. Because of the data reporting format, the method does not differentiate between status unknown (no diagnose) and negative diagnoses, which opens possibilities for statistical enhancements in sensitivity analysis. Continuous variables (e.g., age, vital signs) will be optimally dichotomized as described below. As part of the PBC process (see Aim 1), temporalized order is imposed on the data to ensure potential risk factors precede the outcome of interest (stroke), prior to learning the PGM structure. When trained on binary temporalized data (specifying temporal order of events based on occurrence window), PGMs are forced to learn temporal conditional probabilities. Missing data are handled inherently by the PGM structure learning process – the method identifies the best structure over the dataset including patients with missing values under the statistical assumption of missing at random.^{117, 118} Therefore, no patients are excluded due to missing data. The result is a network model of the joint probability distribution of the multimorbidity network.¹¹⁹

Discovering optimal values for PGM risk calculation

Although there is no formal requirement excluding the use of continuous variables in PGM construction, for practical applications using finite datasets, continuous variables often need to be dichotomized. We have developed means to optimally dichotomize continuous measures. Briefly, a grid search is used to evaluate the performance of the nascent PGM using receiver operating characteristic curves (ROCs). For example, we recently constructed a PGM (**Figure 9**) using data from a cohort of children with severe congenital heart disease that included mental (MDI) and physical (PDI) neurodevelopmental outcomes, growth (LAZ), sex, and damaging genetic variants related to heart development or known syndromic genes. For each value of MDI/PDI (range 50-120), we calculated the joint conditional probability of an individual having low MDI/PDI given the presence or absence of other conditions (variables) for each network. These conditional probability distributions were used to evaluate the accuracy of each candidate network to classify an individual as low or high MDI/PDI score, allowing us to select the optimal score resulting in the highest area under a ROC (AUC). This approach discovered that a score of 70 for both MDI and PDI gives the best AUC scores (0.97 for PDI and 0.88 for MDI). A score of 70 corresponds to 2 standard deviations below the mean, consistent with clinical intuition for a cut-off. We apply a similar approach to dichotomize key continuous variables that drive stroke in AF.

Multimorbidity Networks for Risk Calculations and Electronic Decision Support

The multimorbidity network derived from a PGM allows calculation of an outcome based on the actual rates within the target population – patients with AF within a specific health system. A sample PGM for AF-related stroke within a small, population is shown in **Figure 10**. The PGM provides an inference engine capable of answering $O(3^n)$ personalized conditional risk queries, where n denotes the number of features (for this example, $n = 10$) describing a patient's

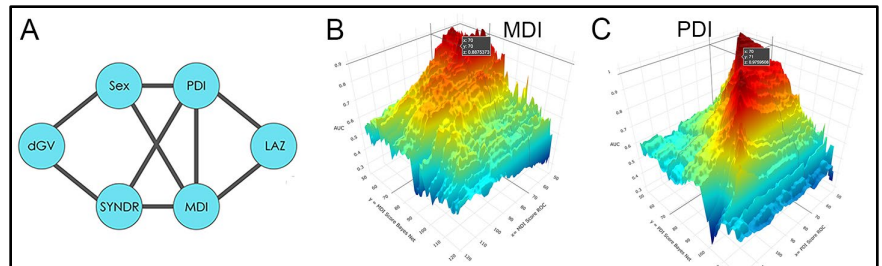


Figure 9. Optimal Dichotomization of Continuous Variables. Probabilistic graphical model (PGM) including the continuous variables MDI and PDI (A) with heat maps (B, C) maximizing area under a receiver operating curve (Z-axis), as a function of outcomes posterior probabilities and continuous values for mental (MDI) and physical (PDI) neurodevelopmental outcomes.

condition, and the base of the exponent is 3, because in case of binary EHR data there are three states for each node that can be specified in the network query for calculating conditional probabilities of interest: present, absent, or status unknown.⁵

We have millions of EHR elements at our disposal. As we have previously demonstrated, at these scales it is possible to explicitly categorize missing data as a state in the model (see **Figure 10**, U = factor status unknown).⁵ These facts together with our access to **All Payer Claims** data, the comprehensiveness of Acxiom's SDoH data, and PGMs methods,^{117, 118} make us well positioned to carry out the proposed aims.

Once a net is created, it is added to a searchable library. Users interact with a net as a web-based 'app', to toggle states and select an outcome of interest.¹¹ A major strength of this approach is the ability of these web-based tools to transform enormous collections of EHR data into compact, portable machines for outcomes research, with no exchange of PHI. They can serve as the primary intervention for a future pragmatic clinical trial of alternative risk stratification strategies for stroke prevention decisions in patients with AF. They could even serve as the foundation for ultimate deployment with Fast Healthcare Interoperability Resources (FHIR) for direct accessibility through the EHR.¹²⁰⁻¹²² See ref ⁵ for more.

B.3.7. Aim 3. Benchmark an ML-based stroke risk stratification across a diverse cohort of health systems using PCORnet and use it to discover biases and drivers of downstream care disparities. We hypothesize that site-specific differences in healthcare environments and patient SDoH impact (a) performance of stroke-risk stratification tools and (b) disparities in care and outcomes.

We will compare our ML-based solution to the clinical standard (CHA₂DS₂-VAsC), within and across sites. Unlike traditional modeling where replication failures are terminal, the explainability of PGMs (1) enables easy identification of variables (including SDoH) driving differences in performance, and (2) facilitates improved communication between physicians and patients. This information will be used to improve our models and to better understand biases and downstream care disparities, and to revise and improve our models for better transportability.^{123, 124} For example, this can be done by modeling 'site' within a PGM.

Overcoming Shortcomings of CHA₂DS₂-VAsC Categorization: We will primarily compare the PGM results to the CHA₂DS₂-VAsC score because: (1) it is the current, guideline-recommended score;³⁴ (2) emerging scores often require collection of additional data elements, limiting portability and generalizability.^{41, 43, 44, 46, 48} If resources permit, we will compare PGM results to other, emerging schema.^{45, 46, 125} We will compare the performance of PGM-based risk versus CHA₂DS₂-VAsC score among the overall population, and among specific subgroups of interest (e.g., untreated patients with low/intermediate CHA₂DS₂-VAsC scores). For example, a generally-accepted threshold for stroke risk warranting treatment is $\geq 2\%/yr$.^{28-31, 35, 126} patients with low risk CHA₂DS₂-VAsC score but with PGM risk above such a threshold demonstrate clinically-relevant 'discordance'.¹²⁷ Evaluation criteria will include AUC and misclassification rate. Furthermore, we will leverage PGM interpretability to explore which components of the PGM drive risk.^{5, 119}

Identifying Biases of Current Approaches: Because they are interpretable, our PGM risk calculations will provide insights into biases and disparities. First, using the above categorization of 'discordance' with CHA₂DS₂-VAsC, we can identify features driving PGM risk and their association with SDoH – *are there groups systematically mislabeled as low risk?* These are groups in whom the CHA₂DS₂-VAsC score has potentially contributed to disparities (**Figure 7**, preliminary examples).

At a broader level, these PGMs will facilitate a more comprehensive understanding of the interactions between risk factors and how SDoH can affect risk of stroke (c.f. discussion of **Figure 7**, especially **7B**). Moreover, such findings may yield opportunities for risk monitoring. Lastly, the PGMs will also allow us to identify and understand site-specific variability in risk, addressing

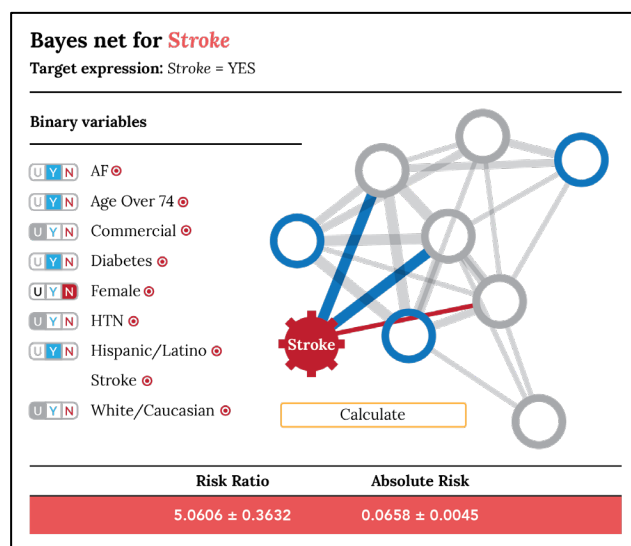


Figure 10: Proof-of-Principle Web-Based Interface to PGM for SDoH Impacts on Stroke Risk. Our software for web-based PGM tools provides easy access for users to explore the full space of conditional dependencies. While there are directional relationships in this directed acyclic graph, arrows have been omitted to avoid confusion with biological causality. Y = present. N = not present. U = status unknown. This PGM was trained on 100,000 random UHealth records. In this example, the projected risk is for a male patient with AF, >74 years of age, Hispanic, and has Diabetes.

transportability problems:^{123, 124} the ‘profile’ of a high-risk patient at each PCORnet site may differ, and this would highlight the need for dynamic risk models, trained for and adapted to local populations.

Ensuring Downstream Validity, Portability, and Transportability: Our Web-based mechanism for PGM dissemination will ensure easy access by collaborators and scientific colleagues.⁵ In order to ensure our risk calculations can be transported to sites and systems without granular SDoH data, a key deliverable will be the understanding of which patients’ risk is influenced by SDoH elements, and how. Based on those findings, we can quantify uncertainty of future risk calculations in these patients, which may be dependent on SDoH.

B.3.8. Potential Problems & Alternative Strategies

The majority of our SDoH data will come from Acxiom consumer datasets, which cover nearly every American adult. As race/ancestry are nearly universal in Acxiom, there is an opportunity to cross-validate those data with the self-reported data on race/ancestry from PCORnet sites across entire populations. Furthermore, as confirmation of the Acxiom, LLC data, we may consider benchmarking these data against other metrics of SDoH, such as: community indices; rural-urban categorization; socioeconomic categorization; and environmental factors based on PCORnet data derived from the Centers for Medicare and Medicaid Services data. While some of these elements are not necessarily individualized (and hence, not the primary approach), they could be used to provide external validation of data integrity. We note that discrepancies can be explicitly modeled by the PGM, with frequently-discrepant variables becoming less determinative for individual risk.

We expect to find differences among PCORnet sites, resulting from variability in demographics, and differences in health and care of the underlying populations. This is an advantage of our innovative approach, which provides means for disparity-aware risk assessment, and means to discover and quantify transportability issues. Indeed, we envision discovery and quantification of these trends as a major deliverable of this proposal.

We also expect that our approach will improve on the shortcomings of CHA₂DS₂-VASc score. However, if the model does not perform as anticipated, our methods can also provide insight into discrepancies (interpretable ML). For example, if the CHA₂DS₂-VASc risk factors remain the most strongly associated with stroke risk, our model will enable us to explore inter-dependence of these factors and how those relationships might vary from site to site. In short, understanding if and how our predictions fail is a key aspect to improving care.

This research is based on EHR data—while we have a strong track record of rigorous AF research using such data,^{91-93, 100, 106} they may be prone to biases and/or inaccuracies. We will test the stability of our findings with sensitivity analyses around exposure and outcome definitions (e.g., different ICD codes, position, temporality).

B.3.9. Sex as a Biological Variable

Patients of all biological sexes will be included here and reported in all publications. Notably, female sex was initially found to be a significant risk factor for AF-related stroke.³⁵ Subsequent analyses failed to replicate this risk³⁶ and guidelines no longer include sex.^{33, 34} Our preliminary PGM analyses confirm this finding (**Figures 7A and 10**). Our proposal positions us to better understand the relationship between sex and AF-related stroke, as we explore co-dependent risks and synergies among factors. For example, it is possible female sex is associated with increased risk for only *some* patients. We will be uniquely positioned to understand this variability.

B.3.10. Impact and Future Directions

Our ultimate goal is to inaugurate a new era for AF-related stroke risk assessment, using EHR data and PGMs to provide personalized, explainable, and equitable risk measurement tools that are portable. We have not yet addressed *bleeding risk* associated with anticoagulation, for several reasons: (1) bleeding outcomes are notoriously more difficult to accurately define compared with stroke;¹²⁸⁻¹³⁰ (2) unlike stroke, objective bleeding risk estimation has not been part of AF treatment guidelines;³⁴ and (3) emerging stroke prevention therapies do not carry traditional bleeding risk (e.g., LAAO).⁶ Future studies may include innovative approaches to understand bleeding risk, as well as pragmatic, randomized clinical trials of stroke-prevention strategies using our risk assessment tools, with an eye toward electronic decision support. This is a key step in moving from the mere collection of massive EHR data, to leveraging data to improve health.^{131, 132}

B.4. TIMELINE

Our timeline is 5 years, as shown in **Table 7**. The first year will be devoted primarily to data acquisition and linking among PCORnet sites and Acxiom, while risk factors are refined in UHealth. Following these initial steps, Aims 2 and 3 can proceed nearly simultaneously.

Table 7. Timeline of proposed aims.					
	YR 1	YR 2	YR 3	YR 4	YR 5
Aim 1: Risk Factor Identification					
• Data preparation and linking	●	●			
• Data analyses	●	●	●		
• Write and publish manuscripts		●	●	●	●
Aim 2: AI-Based Risk Calculation Machinery					
• Data preparation, and algorithm development		●	●		
• Algorithm testing		●	●	●	
• Write and publish manuscripts			●	●	●
Aim 3: Model Benchmarking					
• Data Analyses			●	●	●
• Write and publish manuscripts				●	●

BIBLIOGRAPHY/REFERENCES

1. McIntyre WF, Linz D. Atrial fibrillation and stroke: who is low risk and what are we going to do about it? *Eur Heart J*. 2022;43:3539-3541. PubMed PMID: 35265990.
2. Holmes DR, Reddy VY, Turi ZG, Doshi SK, Sievert H, Buchbinder M, Mullin CM, Sick P, Investigators PA. Percutaneous closure of the left atrial appendage versus warfarin therapy for prevention of stroke in patients with atrial fibrillation: a randomised non-inferiority trial. *Lancet*. 2009;374:534-42. PubMed PMID: 19683639.
3. Reddy VY, Doshi SK, Sievert H, Buchbinder M, Neuzil P, Huber K, Halperin JL, Holmes D, Investigators PA. Percutaneous left atrial appendage closure for stroke prophylaxis in patients with atrial fibrillation: 2.3-Year Follow-up of the PROTECT AF (Watchman Left Atrial Appendage System for Embolic Protection in Patients with Atrial Fibrillation) Trial. *Circulation*. 2013;127:720-9. PubMed PMID: 23325525.
4. Ruff CT, Giugliano RP, Braunwald E, Hoffman EB, Deenadayalu N, Ezekowitz MD, Camm AJ, Weitz JI, Lewis BS, Parkhomenko A, Yamashita T, Antman EM. Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: a meta-analysis of randomised trials. *Lancet*. 2014;383:955-62. PubMed PMID: 24315724.
5. Wesolowski S, Lemmon G, Hernandez EJ, Henrie A, Miller TA, Weyhrauch D, Puchalski MD, Bray BE, Shah RU, Deshmukh VG, Delaney R, Yostl HJ, Eilbeck K, Tristani-Firouzi M, Yandell M. An explainable artificial intelligence approach for predicting cardiovascular outcomes using electronic health records. *PLOS Digit Health*. 2022;1:e0000004. PubMed PMID: 35373216; PMCID: PMC8975108.
6. Freeman JV, Varosy P, Price MJ, Slotwiner D, Kusumoto FM, Rammohan C, Kavinsky CJ, Turi ZG, Akar J, Koutras C, Curtis JP, Masoudi FA. The NCDR Left Atrial Appendage Occlusion Registry. *J Am Coll Cardiol*. 2020;75:1503-1518. PubMed PMID: 32238316; PMCID: PMC7205034.
7. Lemmon G, Wesolowski S, Henrie A, Tristani-Firouzi M, Yandell M. A Poisson binomial-based statistical testing framework for comorbidity discovery across electronic health record datasets. *Nat Comput Sci*. 2021;1:694-702. PubMed PMID: 35252879; PMCID: PMC8896515.
8. Essien UR, Holmes DN, Jackson LR, 2nd, Fonarow GC, Mahaffey KW, Reiffel JA, Steinberg BA, Allen LA, Chan PS, Freeman JV, Blanco RG, Pieper KS, Piccini JP, Peterson ED, Singer DE. Association of Race/Ethnicity With Oral Anticoagulant Use in Patients With Atrial Fibrillation: Findings From the Outcomes Registry for Better Informed Treatment of Atrial Fibrillation II. *JAMA cardiology*. 2018;3:1174-1182. PubMed PMID: 30484833; PMCID: PMC6583087.
9. Wandell P, Carlsson AC, Gasevic D, Holzmann MJ, Arnlov J, Sundquist J, Sundquist K. Socioeconomic factors and mortality in patients with atrial fibrillation-a cohort study in Swedish primary care. *Eur J Public Health*. 2018;28:1103-1109. PubMed PMID: 29746622; PMCID: PMC6241208.
10. Wodschow K, Bihrmann K, Larsen ML, Gislason G, Erbsoll AK. Geographical variation and clustering are found in atrial fibrillation beyond socioeconomic differences: a Danish cohort study, 1987-2015. *Int J Health Geogr*. 2021;20:11. PubMed PMID: 33648527; PMCID: PMC7923319.
11. Yong C, Azarbal F, Abnoui F, Heidenreich PA, Schmitt S, Fan J, Than CT, Ullal AJ, Yang F, Phibbs CS, Frayne SM, Ho PM, Shore S, Mahaffey KW, Turakhia MP. Racial Differences in Quality of Anticoagulation Therapy for Atrial Fibrillation (from the TREAT-AF Study). *Am J Cardiol*. 2016;117:61-8. PubMed PMID: 26552504.
12. Go AS, Al-Khatib SM, Desvigne-Nickens P, Bansal N, Bushnell CD, Fang MC, Freeman JV, Gage BF, Hanke T, Hylek EM, Lopes RD, Noseworthy PA, Reddy VY, Singer DE, Thomas KL, True Hills M, Turakhia MP, Zieman SJ, Cooper LS, Benjamin EJ. Research Opportunities in Stroke Prevention for Atrial Fibrillation: A Report From a National Heart, Lung, and Blood Institute Virtual Workshop. *Stroke*. 2023;54:e75-e85. PubMed PMID: 36848427; PMCID: PMC9995163.
13. Benjamin EJ, Thomas KL, Go AS, Desvigne-Nickens P, Albert CM, Alonso A, Chamberlain AM, Essien UR, Hernandez I, Hills MT, Kershaw KN, Levy PD, Magnani JW, Matlock DD, O'Brien EC, Rodriguez CJ, Russo AM, Soliman EZ, Cooper LS, Al-Khatib SM. Transforming Atrial Fibrillation Research to Integrate Social Determinants of Health: A National Heart, Lung, and Blood Institute Workshop Report. *JAMA cardiology*. 2023;8:182-191. PubMed PMID: 36478155.
14. Meschia JF, Merrill P, Soliman EZ, Howard VJ, Barrett KM, Zakai NA, Kleindorfer D, Safford M, Howard G. Racial disparities in awareness and treatment of atrial fibrillation: the REasons for Geographic and Racial Differences in Stroke (REGARDS) study. *Stroke*. 2010;41:581-7. PubMed PMID: 20190000; PMCID: PMC2885129.

15. Gbadebo TD, Okafor H, Darbar D. Differential impact of race and risk factors on incidence of atrial fibrillation. *Am Heart J.* 2011;162:31-7. PubMed PMID: 21742087; PMCID: PMC3137277.
16. Golwala H, Jackson LR, 2nd, Simon DN, Piccini JP, Gersh B, Go AS, Hylek EM, Kowey PR, Mahaffey KW, Thomas L, Fonarow GC, Peterson ED, Thomas KL, Outcomes Registry for Better Informed Treatment for Atrial Fibrillation R. Racial/ethnic differences in atrial fibrillation symptoms, treatment patterns, and outcomes: Insights from Outcomes Registry for Better Informed Treatment for Atrial Fibrillation Registry. *Am Heart J.* 2016;174:29-36. PubMed PMID: 26995367.
17. Naderi S, Rodriguez F, Wang Y, Foody JM. Racial disparities in hospitalizations, procedural treatments and mortality of patients hospitalized with atrial fibrillation. *Ethn Dis.* 2014;24:144-9. PubMed PMID: 24804358.
18. Magnani JW, Norby FL, Agarwal SK, Soliman EZ, Chen LY, Loehr LR, Alonso A. Racial Differences in Atrial Fibrillation-Related Cardiovascular Disease and Mortality: The Atherosclerosis Risk in Communities (ARIC) Study. *JAMA cardiology.* 2016;1:433-41. PubMed PMID: 27438320; PMCID: PMC5347977.
19. Durrani AF, Soma S, Althouse AD, Leef G, Qin D, Saba S. Impact of Race on Outcome of Patients Undergoing Rhythm Control of Atrial Fibrillation. *J Immigr Minor Health.* 2018;20:14-19. PubMed PMID: 28066862.
20. Komen JJ, Pottegard A, Mantel-Teeuwisse AK, Forslund T, Hjemdahl P, Wettermark B, Hallas J, Olesen M, Bennie M, Mueller T, Carragher R, Karlstad O, Kjerpeseth LJ, Klungel OH. Oral anticoagulants in patients with atrial fibrillation at low stroke risk: a multicentre observational study. *Eur Heart J.* 2022;43:3528-3538. PubMed PMID: 35265981; PMCID: PMC9547505.
21. Jackson LR, 2nd, Kim S, Fonarow GC, Freeman JV, Gersh BJ, Go AS, Hylek EM, Kowey PR, Mahaffey KW, Singer D, Thomas L, Blanco R, Peterson ED, Piccini JP, Sr., Outcomes Registry for Better Informed Treatment of Atrial Fibrillation P, Investigators. Stroke Risk and Treatment in Patients with Atrial Fibrillation and Low CHA(2)DS(2)-VASc Scores: Findings From the ORBIT-AF I and II Registries. *J Am Heart Assoc.* 2018;7:e008764. PubMed PMID: 30369317; PMCID: PMC6201408.
22. Zimmerman RM, Hernandez EJ, Watkins WS, Blue N, Tristani-Firouzi M, Yandell M, Steinberg BA. An Explainable Artificial Intelligence Approach for Discovering Social Determinants of Health and Risk Interactions for Stroke in Patients With Atrial Fibrillation. *Am J Cardiol.* 2023;201:224-226. PubMed PMID: 37385178; PMCID: PMC10529447.
23. Go AS, Hylek EM, Phillips KA, Chang Y, Henault LE, Selby JV, Singer DE. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. *JAMA.* 2001;285:2370-5. PubMed PMID: 11343485.
24. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke.* 1991;22:983-8. PubMed PMID: 1866765.
25. Kim MH, Johnston SS, Chu BC, Dalal MR, Schulman KL. Estimation of total incremental health care costs in patients with atrial fibrillation in the United States. *Circ Cardiovasc Qual Outcomes.* 2011;4:313-20. PubMed PMID: 21540439.
26. Wolowacz SE, Samuel M, Brennan VK, Jasso-Mosqueda JG, Van Gelder IC. The cost of illness of atrial fibrillation: a systematic review of the recent literature. *Europace.* 2011;13:1375-85. PubMed PMID: 21757483.
27. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Cheng S, Delling FN, Elkind MSV, Evenson KR, Ferguson JF, Gupta DK, Khan SS, Kissela BM, Knutson KL, Lee CD, Lewis TT, Liu J, Loop MS, Lutsey PL, Ma J, Mackey J, Martin SS, Matchar DB, Mussolino ME, Navaneethan SD, Perak AM, Roth GA, Samad Z, Satou GM, Schroeder EB, Shah SH, Shay CM, Stokes A, VanWagner LB, Wang NY, Tsao CW, American Heart Association Council on E, Prevention Statistics C, Stroke Statistics S. Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association. *Circulation.* 2021;143:e254-e743. PubMed PMID: 33501848.
28. Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, Pogue J, Reilly PA, Themeles E, Varrone J, Wang S, Alings M, Xavier D, Zhu J, Diaz R, Lewis BS, Darius H, Diener HC, Joyner CD, Wallentin L, Committee R-LS, Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med.* 2009;361:1139-51. PubMed PMID: 19717844.
29. Giugliano RP, Ruff CT, Braunwald E, Murphy SA, Wiviott SD, Halperin JL, Waldo AL, Ezekowitz MD, Weitz JI, Spinar J, Ruzyllo W, Ruda M, Koretsune Y, Betcher J, Shi M, Grip LT, Patel SP, Patel I,

- Hanyok JJ, Mercuri M, Antman EM, Investigators EA-T. Edoxaban versus warfarin in patients with atrial fibrillation. *N Engl J Med*. 2013;369:2093-104. PubMed PMID: 24251359.
30. Granger CB, Alexander JH, McMurray JJ, Lopes RD, Hylek EM, Hanna M, Al-Khalidi HR, Ansell J, Atar D, Avezum A, Bahit MC, Diaz R, Easton JD, Ezekowitz JA, Flaker G, Garcia D, Geraldes M, Gersh BJ, Golitsyn S, Goto S, Hermosillo AG, Hohnloser SH, Horowitz J, Mohan P, Jansky P, Lewis BS, Lopez-Sendon JL, Pais P, Parkhomenko A, Verheugt FW, Zhu J, Wallentin L, Committees A, Investigators. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med*. 2011;365:981-92. PubMed PMID: 21870978.
 31. Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, Breithardt G, Halperin JL, Hankey GJ, Piccini JP, Becker RC, Nessel CC, Paolini JF, Berkowitz SD, Fox KA, Califf RM, Investigators RA. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med*. 2011;365:883-91. PubMed PMID: 21830957.
 32. Piccini JP, Caso V, Connolly SJ, Fox KAA, Oldgren J, Jones WS, Gorog DA, Durdil V, Viethen T, Neumann C, Mundl H, Patel MR, Investigators P-A. Safety of the oral factor Xla inhibitor asundexian compared with apixaban in patients with atrial fibrillation (PACIFIC-AF): a multicentre, randomised, double-blind, double-dummy, dose-finding phase 2 study. *Lancet*. 2022;399:1383-1390. PubMed PMID: 35385695.
 33. Hindricks G, Potpara T, Dagres N, Arbelo E, Bax JJ, Blomstrom-Lundqvist C, Boriani G, Castella M, Dan GA, Dilaveris PE, Fauchier L, Filippatos G, Kalman JM, La Meir M, Lane DA, Lebeau JP, Lettino M, Lip GYH, Pinto FJ, Thomas GN, Valgimigli M, Van Gelder IC, Van Putte BP, Watkins CL, Group ESCSD. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): The Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. *Eur Heart J*. 2021;42:373-498. PubMed PMID: 32860505.
 34. Joglar JA, Chung MK, Armbruster AL, Benjamin EJ, Chyou JY, Cronin EM, Deswal A, Eckhardt LL, Goldberger ZD, Gopinathannair R, Gorenek B, Hess PL, Hlatky M, Hogan G, Ibeh C, Indik JH, Kido K, Kusumoto F, Link MS, Linta KT, Marcus GM, McCarthy PM, Patel N, Patton KK, Perez MV, Piccini JP, Russo AM, Sanders P, Streur MM, Thomas KL, Times S, Tisdale JE, Valente AM, Van Wagoner DR. 2023 ACC/AHA/ACCP/HRS Guideline for the Diagnosis and Management of Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*. 2024;149:e1-e156. PubMed PMID: 38033089.
 35. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*. 2010;137:263-72. PubMed PMID: 19762550.
 36. Friberg L, Benson L, Rosenqvist M, Lip GY. Assessment of female sex as a risk factor in atrial fibrillation in Sweden: nationwide retrospective cohort study. *BMJ*. 2012;344:e3522. PubMed PMID: 22653980; PMCID: PMC3365143.
 37. Chen Z, Bai W, Li C, Wang H, Tang H, Qin Y, Rao L. Left Atrial Appendage Parameters Assessed by Real-Time Three-Dimensional Transesophageal Echocardiography Predict Thromboembolic Risk in Patients With Nonvalvular Atrial Fibrillation. *J Ultrasound Med*. 2017;36:1119-1128. PubMed PMID: 28233335.
 38. Di Biase L, Santangeli P, Anselmino M, Mohanty P, Salvetti I, Gili S, Horton R, Sanchez JE, Bai R, Mohanty S, Pump A, Cereceda Brantes M, Gallinhouse GJ, Burkhardt JD, Cesarani F, Scaglione M, Natale A, Gaita F. Does the left atrial appendage morphology correlate with the risk of stroke in patients with atrial fibrillation? Results from a multicenter study. *J Am Coll Cardiol*. 2012;60:531-8. PubMed PMID: 22858289.
 39. Dudzinska-Szczerba K, Michalowska I, Piotrowski R, Sikorska A, Paszkowska A, Stachnio U, Kowalik I, Kulakowski P, Baran J. Assessment of the left atrial appendage morphology in patients after ischemic stroke - The ASSAM study. *Int J Cardiol*. 2021;330:65-72. PubMed PMID: 33524464.
 40. Freedman B, Martinez C, Katholing A, Rietbrock S. Residual Risk of Stroke and Death in Anticoagulant-Treated Patients With Atrial Fibrillation. *JAMA cardiology*. 2016;1:366-8. PubMed PMID: 27438123.
 41. Go AS, Reynolds K, Yang J, Gupta N, Lenane J, Sung SH, Harrison TN, Liu TI, Solomon MD. Association of Burden of Atrial Fibrillation With Risk of Ischemic Stroke in Adults With Paroxysmal Atrial

Fibrillation: The KP-RHYTHM Study. *JAMA cardiology*. 2018;3:601-608. PubMed PMID: 29799942; PMCID: PMC6145663.

42. Hart RG, Benavente O, McBride R, Pearce LA. Antithrombotic therapy to prevent stroke in patients with atrial fibrillation: a meta-analysis. *Ann Intern Med*. 1999;131:492-501. PubMed PMID: 10507957.
43. Hijazi Z, Lindback J, Alexander JH, Hanna M, Held C, Hylek EM, Lopes RD, Oldgren J, Siegbahn A, Stewart RA, White HD, Granger CB, Wallentin L, Aristotle, Investigators S. The ABC (age, biomarkers, clinical history) stroke risk score: a biomarker-based risk score for predicting stroke in atrial fibrillation. *Eur Heart J*. 2016;37:1582-90. PubMed PMID: 26920728; PMCID: PMC4875560.
44. Nedios S, Koutalas E, Kornej J, Rolf S, Arya A, Sommer P, Husser D, Hindricks G, Bollmann A. Cardiogenic Stroke Despite Low CHA(2) DS(2) -VASc Score: Assessing Stroke risk by Left Atrial Appendage Anatomy (ASK LAA). *J Cardiovasc Electrophysiol*. 2015;26:915-921. PubMed PMID: 26178767.
45. Piccini JP, Stevens SR, Chang Y, Singer DE, Lokhnygina Y, Go AS, Patel MR, Mahaffey KW, Halperin JL, Breithardt G, Hankey GJ, Hacke W, Becker RC, Nessel CC, Fox KA, Califf RM, Committee RAS, Investigators. Renal dysfunction as a predictor of stroke and systemic embolism in patients with nonvalvular atrial fibrillation: validation of the R(2)CHADS(2) index in the ROCKET AF (Rivaroxaban Once-daily, oral, direct factor Xa inhibition Compared with vitamin K antagonism for prevention of stroke and Embolism Trial in Atrial Fibrillation) and ATRIA (AnTicoagulation and Risk factors In Atrial fibrillation) study cohorts. *Circulation*. 2013;127:224-32. PubMed PMID: 23212720.
46. Singer DE, Chang Y, Borowsky LH, Fang MC, Pomernacki NK, Udaltsova N, Reynolds K, Go AS. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score. *J Am Heart Assoc*. 2013;2:e000250. PubMed PMID: 23782923; PMCID: PMC3698792.
47. Carlisle MA, Shrader P, Fudim M, Pieper KS, Blanco RG, Fonarow GC, Naccarelli GV, Gersh BJ, Reiffel JA, Kowey PR, Steinberg BA, Freeman JV, Ezekowitz MD, Singer DE, Allen LA, Chan PS, Pokorney SD, Peterson ED, Piccini JP, Patients OA, Investigators. Residual stroke risk despite oral anticoagulation in patients with atrial fibrillation. *Heart Rhythm O2*. 2022;3:621-628. PubMed PMID: 36589908; PMCID: PMC9795305.
48. Daccarett M, Badger TJ, Akoum N, Burgon NS, Mahnkopf C, Vergara G, Kholmovski E, McGann CJ, Parker D, Brachmann J, Macleod RS, Marrouche NF. Association of left atrial fibrosis detected by delayed-enhancement magnetic resonance imaging and the risk of stroke in patients with atrial fibrillation. *J Am Coll Cardiol*. 2011;57:831-8. PubMed PMID: 21310320; PMCID: PMC3124509.
49. Han L, Askari M, Altman RB, Schmitt SK, Fan J, Bentley JP, Narayan SM, Turakhia MP. Atrial Fibrillation Burden Signature and Near-Term Prediction of Stroke: A Machine Learning Analysis. *Circ Cardiovasc Qual Outcomes*. 2019;12:e005595. PubMed PMID: 31610712; PMCID: PMC8284982.
50. Li X, Liu H, Du X, Zhang P, Hu G, Xie G, Guo S, Xu M, Xie X. Integrated Machine Learning Approaches for Predicting Ischemic Stroke and Thromboembolism in Atrial Fibrillation. *AMIA Annu Symp Proc*. 2016;2016:799-807. PubMed PMID: 28269876; PMCID: PMC5333223.
51. Lip GYH, Tran G, Genaidy A, Marroquin P, Estes C, Landsheft J. Improving dynamic stroke risk prediction in non-anticoagulated patients with and without atrial fibrillation: comparing common clinical risk scores and machine learning algorithms. *European heart journal. Quality of care & clinical outcomes*. 2022;8:548-556. PubMed PMID: 33999139; PMCID: PMC9382661.
52. Huang C, Siu M, Vu L, Wong S, Shin J. Factors influencing doctors' selection of dabigatran in non-valvular atrial fibrillation. *J Eval Clin Pract*. 2013;19:938-43. PubMed PMID: 22834964.
53. Patel N, Deshmukh A, Thakkar B, Coffey JO, Agnihotri K, Patel A, Ainani N, Nalluri N, Patel N, Patel N, Patel N, Badheka AO, Kowalski M, Hendel R, Viles-Gonzalez J, Noseworthy PA, Asirvatham S, Lo K, Myerburg RJ, Mitrani RD. Gender, Race, and Health Insurance Status in Patients Undergoing Catheter Ablation for Atrial Fibrillation. *Am J Cardiol*. 2016;117:1117-26. PubMed PMID: 26899494.
54. Steinberg BA, Shrader P, Thomas L, Ansell J, Fonarow GC, Gersh BJ, Hylek E, Kowey PR, Mahaffey KW, O'Brien EC, Singer DE, Peterson ED, Piccini JP, Outcomes Registry for Better Informed Treatment of Atrial Fibrillation I, Patients. Factors associated with non-vitamin K antagonist oral anticoagulants for stroke prevention in patients with new-onset atrial fibrillation: Results from the Outcomes Registry for Better Informed Treatment of Atrial Fibrillation II (ORBIT-AF II). *Am Heart J*. 2017;189:40-47. PubMed PMID: 28625380.
55. Hagengaard L, Andersen MP, Polcwiartek C, Larsen JM, Larsen ML, Skals RK, Hansen SM, Riahi S, Gislason G, Torp-Pedersen C, Sogaard P, Kragholm KH. Socioeconomic differences in outcomes after

hospital admission for atrial fibrillation or flutter. *European heart journal. Quality of care & clinical outcomes*. 2021;7:295-303. PubMed PMID: 31560375.

56. Bunch TJ, May HT, Bair TL, Weiss JP, Crandall BG, Osborn JS, Mallender C, Anderson JL, Muhlestein BJ, Lappe DL, Day JD. Atrial fibrillation ablation patients have long-term stroke rates similar to patients without atrial fibrillation regardless of CHADS2 score. *Heart Rhythm*. 2013;10:1272-7. PubMed PMID: 23835257.
57. Chew DS, Li Z, Steinberg BA, O'Brien EC, Pritchard J, Bunch TJ, Mark DB, Patel MR, Nabutovsky Y, Greiner MA, Piccini JP. Arrhythmic Burden and the Risk of Cardiovascular Outcomes in Patients With Paroxysmal Atrial Fibrillation and Cardiac Implanted Electronic Devices. *Circ Arrhythm Electrophysiol*. 2022;15:e010304. PubMed PMID: 35089799.
58. Dalgaard F, North R, Pieper K, Fonarow GC, Kowey PR, Gersh BJ, Mahaffey KW, Pokorney S, Steinberg BA, Naccarrelli G, Allen LA, Reiffel JA, Ezekowitz M, Singer DE, Chan PS, Peterson ED, Piccini JP. Risk of major cardiovascular and neurologic events with obstructive sleep apnea among patients with atrial fibrillation. *Am Heart J*. 2020;223:65-71. PubMed PMID: 32179257; PMCID: PMC7214210.
59. Steinberg BA, Hellkamp AS, Lokhnygina Y, Patel MR, Breithardt G, Singer DE, Mahaffey KW, Fox KAA, Califf RM, Piccini JP. Higher risk of death and stroke in patients with persistent versus paroxysmal atrial fibrillation: results from the ROCKET AF trial. *Eur Heart J*; 2014.
60. Wang H, Cui Z, Chen Y, Avidan M, Abdallah AB, Kronzer A. Predicting Hospital Readmission via Cost-Sensitive Deep Learning. *IEEE/ACM Trans Comput Biol Bioinform*. 2018;15:1968-1978. PubMed PMID: 29993930.
61. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P, Yee H, Zhang K, Zhang Y, Flores G, Duggan GE, Irvine J, Le Q, Litsch K, Mossin A, Tansuwan J, Wang D, Wexler J, Wilson J, Ludwig D, Volchenboum SL, Chou K, Pearson M, Madabushi S, Shah NH, Butte AJ, Howell MD, Cui C, Corrado GS, Dean J. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18. PubMed PMID: 31304302; PMCID: PMC6550175.
62. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, Liu X, He Z. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Inform Assoc*. 2020;27:1173-1185. PubMed PMID: 32417928; PMCID: PMC7647281.
63. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep*. 2016;6:26094. PubMed PMID: 27185194; PMCID: PMC4869115.
64. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent Rep*. 2019;49:15-21. PubMed PMID: 30790315.
65. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise Qc. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20:310. PubMed PMID: 33256715; PMCID: PMC7706019.
66. Franz L, Shrestha YR, Paudel B. A deep learning pipeline for patient diagnosis prediction using electronic health records. *arXiv preprint arXiv:2006.16926*. 2020.
67. Heckerman D, Geiger D, Chickering DM. Learning bayesian networks: The combination of knowledge and statistical data. *arXiv preprint arXiv:1302.6815*. 2013.
68. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020;58:82-115. PubMed PMID: WOS:000516799200007.
69. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40:373-83. PubMed PMID: 3558716.
70. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care*. 1998;36:8-27. PubMed PMID: 9431328.
71. Siggaard T, Reguant R, Jorgensen IF, Haue AD, Lademann M, Aguayo-Orozco A, Hjaltelin JX, Jensen AB, Banasik K, Brunak S. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. *Nat Commun*. 2020;11:4952. PubMed PMID: 33009368; PMCID: PMC7532164.

72. Winter AC, Rist PM, Buring JE, Kurth T. Prospective comorbidity-matched study of Parkinson's disease and risk of mortality among women. *BMJ open*. 2016;6:e011888. PubMed PMID: 27670518; PMCID: PMC5051400.
73. Watkins WS, Hernandez EJ, Wesolowski S, Bisgrove BW, Sunderland RT, Lin E, Lemmon G, Demarest BL, Miller TA, Bernstein D, Brueckner M, Chung WK, Gelb BD, Goldmuntz E, Newburger JW, Seidman CE, Shen Y, Yost HJ, Yandell M, Tristani-Firouzi M. De novo and recessive forms of congenital heart disease have distinct genetic and phenotypic landscapes. *Nat Commun*. 2019;10:4722. PubMed PMID: 31624253; PMCID: PMC6797711.
74. Rezaeiahari M. Moving Beyond Simple Risk Prediction: Segmenting Patient Populations Using Consumer Data. *Front Public Health*. 2021;9:716754. PubMed PMID: 34336781; PMCID: PMC8319387.
75. Dawson S. Health care consumption and consumer social class: a different look at the patient. *J Health Care Mark*. 1989;9:15-25. PubMed PMID: 10303930.
76. Holt-Lunstad J, Smith TB, Layton JB. Social relationships and mortality risk: a meta-analytic review. *PLoS Med*. 2010;7:e1000316. PubMed PMID: 20668659; PMCID: PMC2910600.
77. Dzau VJ, McClellan MB, McGinnis JM, Burke SP, Coye MJ, Diaz A, Daschle TA, Frist WH, Gaines M, Hamburg MA, Henney JE, Kumanyika S, Leavitt MO, Parker RM, Sandy LG, Schaeffer LD, Steele GD, Jr., Thompson P, Zerhouni E. Vital Directions for Health and Health Care: Priorities From a National Academy of Medicine Initiative. *JAMA*. 2017;317:1461-1470. PubMed PMID: 28324029.
78. Malaeb D, Hallit S, Dia N, Cherri S, Maatouk I, Nawas G, Salameh P, Hosseini H. Effects of sociodemographic and socioeconomic factors on stroke development in Lebanese patients with atrial fibrillation: a cross-sectional study. *F1000Res*. 2021;10:793. PubMed PMID: 34504688; PMCID: PMC8383125.
79. Clarke H, Clark S, Birkin M, Iles-Smith H, Glaser A, Morris MA. Understanding Barriers to Novel Data Linkages: Topic Modeling of the Results of the LifeInfo Survey. *J Med Internet Res*. 2021;23:e24236. PubMed PMID: 33998998; PMCID: PMC8167605.
80. Sorensen KK, Nielsen EP, Moller AL, Andersen MP, Moller FT, Melbye M, Kolko M, Ejlskov L, Kober L, Gislason G, Starkopf L, Gerds TA, Torp-Pedersen C. Food purchases in households with and without diabetes based on consumer purchase data. *Prim Care Diabetes*. 2022;16:574-580. PubMed PMID: 35461790.
81. Steinberg BA, Piccini JP. Screening for Atrial Fibrillation With a Wearable Device. *JAMA*. 2018;320:139-141. PubMed PMID: 29998323.
82. Bunch TJ, Steinberg BA. Revisiting Rate versus Rhythm Control in Atrial Fibrillation - Timing Matters. *N Engl J Med*. 2020;383:1383-1384. PubMed PMID: 32865379.
83. Jou CJ, Arrington CB, Barnett S, Shen J, Cho S, Sheng X, McCullagh PC, Bowles NE, Pribble CM, Saarel EV, Pilcher TA, Etheridge SP, Tristani-Firouzi M. A Functional Assay for Sick Sinus Syndrome Genetic Variants. *Cell Physiol Biochem*. 2017;42:2021-2029. PubMed PMID: 28803248.
84. Tristani-Firouzi M. Revisiting the challenges of universal screening for long QT syndrome. *J Electrocardiol*. 2015;48:1053-7. PubMed PMID: 26355713.
85. Tristani-Firouzi M. The Long and Short of It: Insights Into the Short QT Syndrome. *J Am Coll Cardiol*. 2014;63:1309-1310. PubMed PMID: 24333498.
86. Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR. The Greater Plains Collaborative: a PCORnet Clinical Research Data Network. *J Am Med Inform Assoc*. 2014;21:637-41. PubMed PMID: 24778202; PMCID: PMC4078294.
87. Waitman LR, Song X, Walpitage DL, Connolly DC, Patel LP, Liu M, Schroeder MC, VanWormer JJ, Mosa AS, Anye ET, Davis AM. Enhancing PCORnet Clinical Research Network data completeness by integrating multistate insurance claims with electronic health records in a cloud environment aligned with CMS security and privacy requirements. *J Am Med Inform Assoc*. 2022;29:660-670. PubMed PMID: 34897506; PMCID: PMC8922172.
88. Experian. Experian audience lookbook. 2020. Available at: experian.com. Accessed 6/2/2022.
89. Acxiom. Healthcare Marketing—Predictive Analytics, Database Solutions, Strategy. Available at: acxiom.com. Accessed 6/2/2022.
90. Acxiom. The Power of Consumer and Lifestyle Data in Healthcare. Available at: acxiom.com. Accessed 6/2/2022.

91. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care*. 2005;43:480-5. PubMed PMID: 15838413.
92. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;43:1130-9. PubMed PMID: 16224307.
93. Steinberg BA, Li Z, O'Brien EC, Pritchard J, Chew DS, Bunch TJ, Mark DB, Nabutovsky Y, Greiner MA, Piccini JP. Atrial fibrillation burden and heart failure: Data from 39,710 individuals with cardiac implanted electronic devices. *Heart Rhythm*. 2021;18:709-716. PubMed PMID: 33508517; PMCID: PMC8096675.
94. Xian Y, Wu J, O'Brien EC, Fonarow GC, Olson DM, Schwamm LH, Bhatt DL, Smith EE, Suter RE, Hannah D, Lindholm B, Maisch L, Greiner MA, Lytle BL, Pencina MJ, Peterson ED, Hernandez AF. Real world effectiveness of warfarin among ischemic stroke patients with atrial fibrillation: observational analysis from Patient-Centered Research into Outcomes Stroke Patients Prefer and Effectiveness Research (PROSPER) study. *BMJ*. 2015;351:h3786. PubMed PMID: 26232340; PMCID: PMC4521370.
95. Zenger B, Zhang M, Lyons A, Bunch TJ, Fang JC, Freedman RA, Navaravong L, Piccini JP, Ranjan R, Spertus JA, Stehlik J, Turner JL, Greene T, Hess R, Steinberg BA. Patient-reported outcomes and subsequent management in atrial fibrillation clinical practice: Results from the Utah mEVAL AF program. *J Cardiovasc Electrophysiol*. 2020;31:3187-3195. PubMed PMID: 33124710; PMCID: PMC7749047.
96. Nadarajah R, Alsaeed E, Hurdus B, Aktaa S, Hogg D, Bates MGD, Cowan C, Wu J, Gale CP. Prediction of incident atrial fibrillation in community-based electronic health records: a systematic review with meta-analysis. *Heart*. 2022;108:1020-1029. PubMed PMID: 34607811; PMCID: PMC9209680.
97. Freeman JV, Wang Y, Akar J, Desai N, Krumholz H. National Trends in Atrial Fibrillation Hospitalization, Readmission, and Mortality for Medicare Beneficiaries, 1999-2013. *Circulation*. 2017;135:1227-1239. PubMed PMID: 28148599.
98. Graham DJ, Reichman ME, Wernecke M, Zhang R, Southworth MR, Levenson M, Sheu TC, Mott K, Goulding MR, Houston M, MaCurdy TE, Worrall C, Kelman JA. Cardiovascular, bleeding, and mortality risks in elderly Medicare patients treated with dabigatran or warfarin for nonvalvular atrial fibrillation. *Circulation*. 2015;131:157-64. PubMed PMID: 25359164.
99. Khazanie P, Liang L, Qualls LG, Curtis LH, Fonarow GC, Hammill BG, Hammill SC, Heidenreich PA, Masoudi FA, Hernandez AF, Piccini JP. Outcomes of medicare beneficiaries with heart failure and atrial fibrillation. *JACC. Heart failure*. 2014;2:41-8. PubMed PMID: 24622118; PMCID: PMC4174273.
100. Piccini JP, Sinner MF, Greiner MA, Hammill BG, Fontes JD, Daubert JP, Ellinor PT, Hernandez AF, Walkey AJ, Heckbert SR, Benjamin EJ, Curtis LH. Outcomes of Medicare beneficiaries undergoing catheter ablation for atrial fibrillation. *Circulation*. 2012;126:2200-7. PubMed PMID: 23019293; PMCID: PMC3500836.
101. Piccini JP, Mi X, DeWald TA, Go AS, Hernandez AF, Curtis LH. Pharmacotherapy in Medicare beneficiaries with atrial fibrillation. *Heart Rhythm*. 2012;9:1403-8. PubMed PMID: 22537885; PMCID: PMC3652262.
102. Lakshminarayan K, Solid CA, Collins AJ, Anderson DC, Herzog CA. Atrial fibrillation and stroke in the general medicare population: a 10-year perspective (1992 to 2002). *Stroke*. 2006;37:1969-74. PubMed PMID: 16809573.
103. Birman-Deych E, Radford MJ, Nilasena DS, Gage BF. Use and effectiveness of warfarin in Medicare beneficiaries with atrial fibrillation. *Stroke*. 2006;37:1070-4. PubMed PMID: 16528001.
104. Steinberg BA, Zhang M, Bensch J, Lyons A, Bunch TJ, Piccini JP, Siu A, Spertus JA, Stehlik J, Wohlfahrt P, Greene T, Hess R, Fang JC. Quantifying the Impact of Atrial Fibrillation on Heart Failure-Related Patient-Reported Outcomes in the Utah mEVAL Program. *J Card Fail*. 2022;28:13-20. PubMed PMID: 34324927; PMCID: PMC8748275.
105. Wohlfahrt P, Nativi-Nicolau J, Zhang M, Selzman CH, Greene T, Conte J, Biber JE, Hess R, Mondesir FL, Wever-Pinzon O, Drakos SG, Gilbert EM, Kemeyou L, LaSalle B, Steinberg BA, Shah RU, Fang JC, Spertus JA, Stehlik J. Quality of Life in Patients With Heart Failure With Recovered Ejection Fraction. *JAMA cardiology*. 2021;6:957-962. PubMed PMID: 33950162; PMCID: PMC8100912.

106. Steinberg BA, Turner J, Lyons A, Biber J, Chelu MG, Fang JC, Freedman RA, Han FT, Hardisty B, Marrouche NF, Ranjan R, Shah RU, Spertus JA, Stehlik J, Zenger B, Piccini JP, Hess R. Systematic collection of patient-reported outcomes in atrial fibrillation: feasibility and initial results of the Utah mEVAL AF programme. *Europace*. 2020;22:368-374. PubMed PMID: 31702780; PMCID: PMC7058971.
107. Jones WS, Mulder H, Wruck LM, Pencina MJ, Kripalani S, Munoz D, Crenshaw DL, Effron MB, Re RN, Gupta K, Anderson RD, Pepine CJ, Handberg EM, Manning BR, Jain SK, Girotra S, Riley D, DeWalt DA, Whittle J, Goldberg YH, Roger VL, Hess R, Benziger CP, Farrehi P, Zhou L, Ford DE, Haynes K, VanWormer JJ, Knowlton KU, Kraschnewski JL, Polonsky TS, Fintel DJ, Ahmad FS, McClay JC, Campbell JR, Bell DS, Fonarow GC, Bradley SM, Paranjape A, Roe MT, Robertson HR, Curtis LH, Sharlow AG, Berdan LG, Hammill BG, Harris DF, Qualls LG, Marquis-Gravel G, Modrow MF, Marcus GM, Carton TW, Nauman E, Waitman LR, Kho AN, Shenkman EA, McTigue KM, Kaushal R, Masoudi FA, Antman EM, Davidson DR, Edgley K, Merritt JG, Brown LS, Zemon DN, McCormick TE, 3rd, Alikhaani JD, Gregoire KC, Rothman RL, Harrington RA, Hernandez AF, Team A. Comparative Effectiveness of Aspirin Dosing in Cardiovascular Disease. *N Engl J Med*. 2021;384:1981-1990. PubMed PMID: 33999548; PMCID: PMC9908069.
108. Piccini JP, Hammill BG, Sinner MF, Jensen PN, Hernandez AF, Heckbert SR, Benjamin EJ, Curtis LH. Incidence and prevalence of atrial fibrillation and associated mortality among Medicare beneficiaries, 1993-2007. *Circ Cardiovasc Qual Outcomes*. 2012;5:85-93. PubMed PMID: 22235070; PMCID: PMC3332107.
109. Williams BA, Chamberlain AM, Blankenship JC, Hylek EM, Voyce S. Trends in Atrial Fibrillation Incidence Rates Within an Integrated Health Care Delivery System, 2006 to 2018. *JAMA network open*. 2020;3:e2014874. PubMed PMID: 32857147; PMCID: PMC7455855.
110. Steinberg BA, Li Z, Shrader P, Chew DS, Bunch TJ, Mark DB, Nabutovsky Y, Shah RU, Greiner MA, Piccini JP. Bimodal distribution of atrial fibrillation burden in 3 distinct cohorts: What is 'paroxysmal' atrial fibrillation? *Am Heart J*. 2022;244:149-156. PubMed PMID: 34838507; PMCID: PMC8727503.
111. Schreiber J. Pomegranate: fast and flexible probabilistic modeling in python. *The Journal of Machine Learning Research*. 2017;18:5992-5997.
112. Yuan C, Malone B, Wu X. Learning optimal Bayesian networks using A* search. Paper presented at: Twenty-Second International Joint Conference on Artificial Intelligence, 2011.
113. Koivisto M, Sood K. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*. 2004;5:549-573. PubMed PMID: WOS:000236327500005.
114. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*. 1995;20:197-243. PubMed PMID: WOS:A1995RX35400001.
115. Weiss Y. Correctness of local probability in graphical models with loops. *Neural Comput*. 2000;12:1-41. PubMed PMID: 10636932.
116. Peixoto TP. The graph-tool python library. *figshare*; 2014.
117. Scutari M. Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimized Implementations in the bnlearn R Package. *Journal of Statistical Software*. 2017;77:1-20. PubMed PMID: WOS:000399023300001.
118. Scutari M. Learning Bayesian Networks with thebnlearnRPackage. *Journal of Statistical Software*. 2010;35:1 - 22.
119. Pearl J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*: Morgan kaufmann; 1988.
120. Taber P, Radloff C, Del Fiore G, Staes C, Kawamoto K. New Standards for Clinical Decision Support: A Survey of The State of Implementation. *Yearb Med Inform*. 2021;30:159-171. PubMed PMID: 34479387; PMCID: PMC8416232.
121. Strasberg HR, Rhodes B, Del Fiore G, Jenders RA, Haug PJ, Kawamoto K. Contemporary clinical decision support standards using Health Level Seven International Fast Healthcare Interoperability Resources. *J Am Med Inform Assoc*. 2021;28:1796-1806. PubMed PMID: 34100949; PMCID: PMC8324242.
122. Kawamoto K, Kukhareva PV, Weir C, Flynn MC, Nanjo CJ, Martin DK, Warner PB, Shields DE, Rodriguez-Loya S, Bradshaw RL, Cornia RC, Reese TJ, Kramer HS, Taft T, Curran RL, Morgan KL, Borbolla D, Hightower M, Turnbull WJ, Strong MB, Chapman WW, Gregory T, Stipelman CH, Shakib JH, Hess R, Boltax JP, Habboushe JP, Sakaguchi F, Turner KM, Narus SP, Tarumi S, Takeuchi W, Ban H, Wetter DW, Lam C, Caverly TJ, Fagerlin A, Norlin C, Malone DC, Kaphingst KA, Kohlmann WK,

- Brooke BS, Del Fiore G. Establishing a multidisciplinary initiative for interoperable electronic health record innovations at an academic medical center. *JAMIA Open*. 2021;4:ooab041. PubMed PMID: 34345802; PMCID: PMC8325485.
123. Song X, Yu ASL, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, Hu Y, Liu M. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nat Commun*. 2020;11:5668. PubMed PMID: 33168827; PMCID: PMC7653032.
 124. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68:279-89. PubMed PMID: 25179855.
 125. Fox KAA, Virdone S, Pieper KS, Bassand JP, Camm AJ, Fitzmaurice DA, Goldhaber SZ, Goto S, Haas S, Kayani G, Oto A, Misselwitz F, Piccini JP, Dalgaard F, Turpie AGG, Verheugt FWA, Kakkar AK, Investigators G-A. GARFIELD-AF risk score for mortality, stroke, and bleeding within 2 years in patients with atrial fibrillation. *European heart journal. Quality of care & clinical outcomes*. 2022;8:214-227. PubMed PMID: 33892489; PMCID: PMC8888127.
 126. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA*. 2001;285:2864-70. PubMed PMID: 11401607.
 127. Steinberg BA, Kim S, Thomas L, Fonarow GC, Hylek E, Ansell J, Go AS, Chang P, Kowey P, Gersh BJ, Mahaffey KW, Singer DE, Piccini JP, Peterson ED, Outcomes Registry for Better Informed Treatment of Atrial Fibrillation I, Patients. Lack of concordance between empirical scores and physician assessments of stroke and bleeding risk in atrial fibrillation: results from the Outcomes Registry for Better Informed Treatment of Atrial Fibrillation (ORBIT-AF) registry. *Circulation*. 2014;129:2005-12. PubMed PMID: 24682387; PMCID: PMC4050636.
 128. Schulman S, Kearon C, Subcommittee on Control of Anticoagulation of the S, Standardization Committee of the International Society on T, Haemostasis. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. *J Thromb Haemost*. 2005;3:692-4. PubMed PMID: 15842354.
 129. Delate T, Jones AE, Clark NP, Witt DM. Assessment of the coding accuracy of warfarin-related bleeding events. *Thromb Res*. 2017;159:86-90. PubMed PMID: 29035718.
 130. Taggart M, Chapman WW, Steinberg BA, Ruckel S, Pregoner-Wenzler A, Du Y, Ferraro J, Bucher BT, Lloyd-Jones DM, Rondina MT, Shah RU. Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients. *JAMA network open*. 2018;1:e183451. PubMed PMID: 30646240; PMCID: PMC6324448.
 131. Olsen L, Aisner D, McGinnis JM. The learning healthcare system: workshop summary. 2007. PubMed PMID: 21452449.
 132. McGinnis JM, Fineberg HV, Dzau VJ. Advancing the Learning Health System. *N Engl J Med*. 2021;385:1-5. PubMed PMID: 34192452.