|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT387 | |
| **Submission Date** | 4/13/2020 | |
| **Project Title** | Polygenic risk for atrial fibrillation within the eMERGE III cohort | |
| **Tentative Lead Investigator** *(first author)* | Konrad Teodor Sawicki, MD PhD | |
| **Tentative Senior Author**  *(last author)* | Laura Rasmussen-Torvik, PhD | |
| **All Other Authors** | Jeremy Pasteris, BS, Megan Roy-Pucklewartz, PhD, Jennifer A. Pacheco, MS, Elizabeth McNally MD PhD | |
| **Sites Participating** | Northwestern University and other interested eMERGE sites | |
| **Background / Significance** | Atrial fibrillation (AF) is the most common cardiac arrhythmia, affecting over 30 million people worldwide. AF is associated with significant morbidity and mortality, including increased risk of stroke, heart failure, dementia, and death. The mechanisms underlying AF pathogenesis are complex, and the probability of developing AF is influenced by diverse clinical and genetic risk factors. Identification of individuals at increased genetic risk for AF may facilitate treatment and enhance screening strategies to prevent manifestation of the disease.  Over the past decade, there has been significant progress in understanding the genetics of AF and identifying the germline single nucleotide polymorphisms (SNPs) associated with the disease. Genome-wide association studies (GWAS) have identified over 100 genetic loci associated with AF, including genes for cardiac ion channels, transcription factors, sarcomeric proteins, calcium signaling, and cardiac development1. While initial work focused on identifying rare monogenic pathogenic variants that lead to several-fold increased risk, the vast majority of patients with AF do not harbor such mutations. For most individuals, the heritability of AF is complex and polygenic, driven by the cumulative effect of common variants across the genome.  Genome-wide polygenic scores (GPS) integrate information from millions of common DNA variants into a single measure of inherited susceptibility and has recently been shown to identify individuals who are at significantly increased risk of developing AF. GPS is an attractive biomarker because it is stable throughout life and may provide information about AF risk from an early age before the development of clinical risk factors and may help inform screening and therapeutic strategies.  In this study, we will use utilize both genome-wide and independent risk variant approaches to derive and validate polygenic scores for AF in the Electronic Medical Records and Genomics (eMERGE) network, a national biorepository consisting of over 15,000 individuals with genome-wide genotyping and whole-exome sequencing. We will assess if there are significant associations of each polygenic score with AF-related comorbidities, and if the addition of the polygenic score to clinical risk factors can better predict the incidence of AF and the risk of stroke in patients with AF. | |
| **Outline of Project** | We will first perform a GWAS to identify genetic variants for AF in the eMERGE network. Diagnosis of AF was based on electronic health record (EHR) information including inpatient International Classification of Disease (ICD-9, ICD-10) diagnosis codes. Because allele frequencies, linkage disequilibrium patterns, and effect sizes of polymorphisms vary with ancestry, we will stratify the GWAS by race/ethnicity (European ancestry, African American, Asian American, Latino). The GWAS will also be stratified by age and sex.  We will then use both genome-wide and independent risk variant approaches to derive polygenic scores for AF. For the genome-wide approach, we will utilize a previously published method to derive a GPS for AF2. Seven different scores will be derived, and the score with the best discriminative capacity will be determined based on maximal area under the receiver-operator curve (AUC) in a logistic regression model with AF as the outcome. Patients will be classified into GPS quintiles for AF. For the independent risk variant approach, we will utilize a previously published method using the most current optimal variant list for atrial fibrillation (166 variants) to calculate a polygenic risk score (PRS). We will test and validate the PRS for association with AF in eMERGE using multivariable logistic regression. Patients will be classified into GPS quintiles for AF.  Using polygenic scores derived from both genome-wide and individual risk variant approaches, we will stratify the polygenic scores based for AF in eMERGE by patient characteristics, including age, race/ethnicity, and body mass index. Logistic regression will be used to estimate the association of each polygenic score quintile with AF-related comorbidities, including stroke, heart failure, dementia and death. We will also determine if the addition of polygenic score to clinical AF risk factors based on the CHARGE-AF risk score (age, race, height, weight, systolic and diastolic blood pressure, current smoking, use of antihypertensive medication, diabetes, and history of myocardial infarction and heart failure) can better predict the incidence of AF. Finally, we will assess if the addition of the polygenic score to the CHA2DS2VASc score (age, sex, history of hypertension, heart failure, stroke, transient ischemic event, thromboembolism, vascular disease, and diabetes) better predicts the risk of stroke in patients with AF. Statistical analyses will be conducted using R software. | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  Other: Case/Control status on Phase I and Phase II phenotypes |
| **Other Desired Data *(Available from participating sites)*** | *Please specifically list out any data elements that participating sites would collect or extract from clinical or other sources for this project (i.e. not common variables above)*  All data needed is in the existing common variable datasets identified above.  Optionally, if sites have, we would like systolic and diastolic blood pressure, smoking status, and use of antihypertensive medication. This could simply be from existing data extracts for data dictionaries from existing eMERGE phenotype algorithms such as: smkEver covariate from COPD algorithm, and blood pressure and yes/no if used anti-hypertensive meds from existing eMERGE cardiovascular (CRF, etc.) phenotype algorithms. | |
| **Desired Genetic Data** | eMERGE I-III Merged set (HRC imputed, GWAS)  eMERGE PGx/PGRNseq data set  eMERGEseq data set (Phase III)  eMERGE Whole Genome sequencing data set  eMERGE Exome chip data set  eMERGE Whole Exome sequencing data set  Other (not listed above) | |
| **Does project pertain to an existing eMERGE Phenotype?** | Yes, if so please list  No | |
| **Planned Statistical Analyses** | We will identify a primary dichotomous phenotype for GWAS based on ICD-9 and 10 codes and perform standard GWAS analysis, stratified by race.  For the genome-wide approach, we will utilize a previously published approach to derive a GPS for AF (Khera et al., Nature Genetics, 2018). Based on summary statistics from a recent large AF GWAS3 (17,931 cases and 115,142 controls) with 6.7 million variants, we will calculate a GPS for each individual in eMERGE. The LDPred computational algorithm, a Bayesian approach to calculate a posterior mean effect for each variant based on a prior and subsequent shrinkage based on linkage disequilibrium (<https://github.com/bvilhjal/ldpred>), will be used to generate seven candidate GPS as previously described2. The LDPred scores will be calculated based on values for the tuning parameter ρ, where ρ is the proportion of variants assumed to be causal for AF. Because ρ is unknown for any given disease, a range of ρ will be used: 1.0, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001. The GPS will be generated by multiplying the genotype dosage of each risk allele for each variant by its respective weight and then summing across all variants in the score using PLINK2 software, ultimately yielding a single continuous value for each individual. GPS for each individual will be corrected for age, sex, and genetic background (quantified by the 5 principal components of ancestry) using a residualized linear regression. Seven different scores will be derived, and the score with the best discriminative capacity will be determined based on maximal area under the receiver-operator curve (AUC) in a logistic regression model with AF as the outcome. Patients will be classified into GPS quintiles for AF.  For the independent risk variant approach, we will utilize a previously published method using the most current optimal variant list for atrial fibrillation (166 variants) to calculate a polygenic risk score (PRS). An inverse normal-transformed PRS for AF using summarized dosage-weighted risk estimates (beta coefficients) will be constructed from the list of 166 independent risk variants to yield a single continuous value for each individual. The full list of variants and beta coefficients can be found in Table 1. We will test and validate the PRS for association with AF in eMERGE using multivariable logistic regression. Patients will be classified into GPS quintiles for AF. | |
| **Ethical Considerations** | None | |
| **Target Journal** | Circulation | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | May 2020: Generate PRS  June 2020: perform PRS association analysis  Aug 2020: First Draft Manuscript circulated at NU  Sept 2020: Circulate to external co-authors  Oct 2020: Second/Final Draft Circulated to Co-Authors  Dec 2020: Submit for publication | |

**\*Common Variables available across all datasets:**

* Demographics: sex, year of birth, decade of birth, race, ethnicity
* Codes: (repeated values & age at event): ICD, CPT, Phecodes
* BMI: (repeated value & age at event) height, weight, BMI
* Labs: (lab name, repeated lab value & age at event) Serum total cholesterol, LDL, HDL, Triglycerides, Glucose fasting/non-fasting/unknown, & White Blood Cell count
* Medications: (medication name, repeated, & age at event) Cerivastatin sodium, Rosuvastatin, Simvastatin, Fluvastatin, Pravastatin, Lovastatin, Atorvastatin, & Pitavastatin
* Other: Case/Control status on Phase I and Phase II phenotype: only on GWAS dataset participants

References

1. Nielsen JB, Thorolfsdottir RB, Fritsche LG, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet.* 2018;50(9):1234-1239.

2. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* 2018;50(9):1219-1224.

3. Christophersen IE, Rienstra M, Roselli C, et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat Genet.* 2017;49(6):946-952.