|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT388 | |
| **Submission Date** | 04/27/2020 | |
| **Project Title** | Data-driven Phenotyping and Prediction of LVEF Changes in Heart Failure Patients using Machine Learning | |
| **Tentative Lead Investigator** *(first author)* | Prakash Adekkanattu | |
| **Tentative Senior Author**  *(last author)* | Jyotishman Pathak | |
| **All Other Authors** | Prakash Adekkanattu, Joseph Kabariti, Parag Goyal, Faraz Ahmad, Zhenxing Xu, Guoqian Jiang, Iftikhar Kullo, Yuan Luo, Luke V. Rasmussen, Jennifer A. Pacheco, Richard C. Kiefer, Pascal S. Brandt, Jei Xu, Fei Wang, Thomas R. Campion, Jr, Jyotishman Pathak | |
| **Sites Participating** | Weill Cornell, Northwestern University, Mayo Clinic | |
| **Background / Significance** | Left ventricular ejection fraction (LVEF) is an important predictor of mortality in heart failure (HF) patients. A predictive model for LVEF changes overtime would assist physicians on prognosis and treatment options. | |
| **Outline of Project** | \*\* This is an expansion of the eMERGE Heart Failure phenotype but includes the extraction and analysis of additional data elements. eMERGE sites are welcome to participate, but it will require additional phenotyping effort. \*\*  Participating sites will perform the following tasks:   * Extract additional data elements, including running EchoExtractor NLP pipeline to extract left ventricular ejection fraction form echocardiograms, if the data is not already existing in EHR. * Provided MS SQL queries will be modified and run at each individual site to identify heart failure patients as defined by the phenotype definition. * Provided MS SQL queries will be modified and run at each individual site to collect characteristics(demographic, vitals, labs/measurements, comorbidities, medications) for the patient cohort defined through the previous queries. * Provided R script will be modified and run at each individual site to analyze and prepare the data. * Conduct chart review of 100-150 charts to establish a local gold standard * Provided Python code will be modified and run locally to evaluate the machine learning prediction models * Contribute to summarization of results at each site and participate in manuscript writing | |
| **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)*  Laboratory results: BNP  Medications for heart failure  Ejection fraction (derived from structured data or using NLP) | |
| **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: Heart failure (<https://phekb.org/phenotype/heart-failure-hf-differentiation-between-preserved-and-reduced-ejection-fraction>)  No | |
| **Planned Statistical Analyses** | P-value by chi square test on patient characteristics for the sub-phenotypes of heart failure: HFpEF, HFmrEF, and HFrEF | |
| **Ethical Considerations** | Retrospective longitudinal data analysis of heart failure patients, aged 18 years or older, in accordance with individual IRB protocol at each participating site. | |
| **Target Journal** | Circulation: Heart Failure, JAMA Cardiology, Journal of Cardiac Failure, …? | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Feasibility assessment and regulatory approval (1 month)  Code modification and testing (1 months)  Data collection and analysis (2 months)  Draft completion and submission ( 1 months) | |

**\*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