|  |  |  |
| --- | --- | --- |
| **eMERGE Network: External Collaborator Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT322 | |
| **Submission Date** | 12/20/2018 | |
| **Project Title** | Early prediction of risk for Alzheimer’s disease and related dementia using data-driven, scalable analysis of electronic health records and genetic data. | |
| **Tentative Lead Investigator** *(first author)* | Ji-Hwan Park, PhD, Brookhaven National Laboratory | |
| **Tentative Senior Author**  *(last author)* | Jiook Cha, PhD, Department of Psychiatry, Columbia University | |
| **eMERGE Site Sponsor & Contact** | Chunhua Weng, PhD, Associate Professor of Biomedical Informatics  Columbia University, chunhua@columbia.edu | |
| **All Other Authors** | Hal Eol Cho, MD; Jong-Hun Kim, MD, PhD; Yaakov Stern, PhD; Melanie Wall, PhD; Shinjae Yoo, PhD; Hyoung-Seop Kim, MD, | |
| **Sites Participating** | Columbia University, Brookhaven National Laboratory, Korean National Health Insurance Service Ilsan Hospital | |
| **Background / Significance** | Early prediction of risk for Alzheimer’s disease using ubiquitous and affordable measures will help better intervention strategy to delay the onset of diseases. The advent of digitalization has led an exponential increase in the volume of electronic health records (EHR) or electronic medical records (EMR) containing individuals’ history of health and healthcare. This big EHR data, combined with the recent machine learning approach, may provide an unprecedented opportunity to test predictive modeling in AD. Compared with the tremendous efforts of developing predictive biomarkers of AD, however, little has been reported as to the utility of the longitudinal EHR in predicting key clinical outcomes of AD. | |
| **Outline of Project** | In this study we will develop and validate data-driven model to predict future incidence of Alzheimer’s disease (AD) using EHR from multiple sources. Using large scale (one million) EHR from the Korean National Health Insurance Service along with the eMERGE data, we will test whether the data-driven predictive model has generalizability across multiple sources. | |
| **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)* | |
| **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 first use the Korean National Health Insurance Service database containing insurance billing data, health check-up data, and socio-demographics. We will use a population-representative data from one million Korean with nine years of EHR. We will implement recurrent neural network to for data-driven feature engineering and deep phenotyping of EHR, and benchmark the model against conventional machine learning models. We will test the model in predicting 1,2,3,4 year subsequent future incident AD. Then, we will use the EHR data from eMERGE to test whether the recurrent neural network model trained on Korean EHR can be transferred to the US data. With the machine learning we will the eMERGE data as an independent validation dataset. | |
| **Ethical Considerations** | Korean data is de-identified by the Korean National Health Insurance Service and publicly available. Therefore, this is not considered a human subject study according to our IRB. | |
| **Available Funding or Resources** | NIMH K01 (PI: Jiook Cha), Million Veterans Project (site-PI: Shinjae Yoo) | |
| **Target Journal** | JAMA Neurology | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Project approval: 12/2018-1/2019  Project duration: 2-7/2019  Draft completion: 8-9/2019  Submission: 10/2019 | |

**\*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