|  |  |  |
| --- | --- | --- |
| **eMERGE Network: External Collaborator Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT435 | |
| **Submission Date** | Oct 10,2021 | |
| **Project Title** | Network-based comorbidity risk score for prediction using biobank-scaled PheWAS data | |
| **Tentative Lead Investigator** *(first author)* | Yonghyun Nam | |
| **Tentative Senior Author**  *(last author)* | Anurag Verma\*, Dokyoon Kim\* | |
| **eMERGE Site Sponsor & Contact** | Anurag Verma, Marylyn D. Ritchie, University of Pennsylvania | |
| **All Other Authors** | Sang-Hyuk Jung, Vivek Sriram, Hong-Hee Won, Jae-Seung Yun, Brenda Xiao | |
| **Sites Participating** | TBD | |
| **Background / Significance** | The prediction of an individual’s disease risk is an essential part of precision medicine and will be required to improve public healthcare. One of the most popular method for predicting disease risk, the polygenic risk score (PRS), has shown promising results for identifying high-genetic risk individuals for various diseases. However, a major weakness of the PRS is its focus on a single trait for the estimation of genetic risk scores. Conventional PRSs fail to consider genetic relationships across multiple diseases, even though multiple diseases will usually afflict patients at once or in succession. To overcome these difficulties, we developed a novel method of network-based comorbidity risk scores (netCRS) to quantify associations among multiple phenotypes from phenome-wide association studies (PheWAS) | |
| **Outline of Project** | We see to replicate the proposed network-based comorbidity risk scoring algorithms constructed by UKBiobank PheWAS data.  For selected five phenotypes (type 2 diabetes, BMI, hypertension, myocardial infarction, coronary artery disease, chronic kidney disease), we will define case and control sample based on PheCodes and then run following analyses:   1. Polygenic risk scores (PRSs) 2. Risk prediction of selected phenotypes with proposed method (netCRS) 3. Association analysis of netCRS and PRS | |
| **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)*** | *If it is possible to extract ICD code, CPT code and lab measure encounter date, it would be very useful for the analyses to identify maternal age for maternal health phenotypes.* | |
| **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** | 1. Graph-based semi-supervised learning models for predicting individual disease risk scores  2. Generalized Regression and mixed linear models adjusting for age, gender, and PCs for GWAS and PheWAS analyses. | |
| **Ethical Considerations** | N/A | |
| **Available Funding or Resources** |  | |
| **Target Journal** | TBD | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Analysis to complete by December 2021 | |

**\*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
* Other: Case/Control status on Phase I and Phase II phenotype: only on GWAS dataset participants