|  |  |
| --- | --- |
| **External Collaborator Proposal** *for* **eMERGE Network Analysis**  Project/Manuscript Concept Sheet | |
| **Reference Number** | NT432 |
| **Submission Date** | 10/1/2021 |
| **Tentative Lead Investigator** *(first author with contact information and affiliation)* | Ruowang Li, [ruowang@upenn.edu](mailto:ruowang@upenn.edu), University of Pennsylvania |
| **Tentative Senior Author**  *(last author)* | Marylyn D. Ritchie, [marylyn@upenn.edu](mailto:marylyn@upenn.edu), University of Pennsylvania;  Jason H. Moore, [jhmoore@upenn.edu](mailto:jhmoore@upenn.edu), University of Pennsylvania; |
| **eMERGE Site Sponsor & Contact** | University of Pennsylvania, Marylyn D. Ritchie |
| **Project Title** | Polygenic risk vector improves genetic risk predictions for cardio-metabolic diseases |
| **All Other Authors** | eMERGE participating sites, Xinyuan Zhang, Binglan Li, others TBD |
| **Other eMERGE Sites Involved** | eMERGE participating sites |
| **Background / Significance** | Genetic risk prediction plays an important role in determining the disease prevention and treatment strategies. For many complex diseases, multiple genes and genetic loci have been shown to be associated with disease risks. In most cases, Individual SNPs have only displayed modest predictability of disease risks; however, the weighted aggregations of many SNPs, or a polygenic risk score (PRS), have shown good prediction performances for many diseases including diabetes, obesity, cancer, and heart diseases. However, PRS is still limited in that it only explains a part of the total phenotype variance. Here, we propose to develop a new method called polygenic risk vector (PRV) that aims to capture additional genetic signals that can be used for disease risk prediction. We will apply our method to cardio-metabolic diseases. With high-quality genotypic data and well-documented electronic health records, eMERGE network phase III provides an ideal setting to validate our new method. |
| **Outline of Project** | 1. **Phenotype definition**   Phenotypes will be defined by applying “rule of three” on longitudinal ICD9 codes.   1. **Population stratification**   We plan to conduct method validation for European population  **Genomic analyses**   * 1. Calculate PRV on eMERGE participants to determine their genetic risks   2. Correlate participants’ genetic risks with their true disease status |
| **Desired Variables**  *(essential for analysis*  *indicated by* ***\*****)* | We seek the following variables for our analyses:  Primary ICD9 codes\*:   * Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279) * Diseases of circulatory system (390-459)   Confounder variables\*: age, sex, and race/ethnicity  Related phenotypes\*: blood lipid levels, (serum cholesterol levels), body-mass-index, |
| **Desired Data** | eMERGE-III HRC imputed data |
| **Planned Statistical Analyses** | 1. Perform quality control 2. Perform validation analysis of the new PRV method |
| **Ethical Considerations** | Genomics data and phenotypic data will be de-identified to protect confidentiality. |
| **Available Funding or Resources** |  |
| **Milestones\*\*** | 1. Complete QC by early July, 2021 2. Complete validation analyses for PRV by July, 2021 3. Write manuscript by August, 2021 |

***\*\**** *This section should include the timeline for completion of project, including: approval, project duration, first and second draft of the paper and submission.*