|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT418 | |
| **Submission Date** | 3/18/2021 | |
| **Project Title** | Polygenic risk score diagnostics- which features drive PRS performance? | |
| **Tentative Lead Investigator** *(first author)* | Daniel Hui and Brenda Xiao (PhD students in Dr. Ritchie’s lab)  [dahui@pennmedicine.upenn.edu](mailto:dahui@pennmedicine.upenn.edu), [brendax@pennmedicine.upenn.edu](mailto:brendax@pennmedicine.upenn.edu) | |
| **Tentative Lead Investigator Email Address** | [marylyn@pennmedicine.upenn.edu](mailto:marylyn@pennmedicine.upenn.edu) | |
| **Tentative Senior Author**  *(last author)* | Marylyn Ritchie | |
| **All Other Authors** | Shefali Verma, Dokyoon Kim, Sarah Tishkoff – any eMERGE authors who are interested | |
| **Sites Participating** | We welcome participation from all eMERGE sites. | |
| **Background / Significance** | Polygenic risk scores (PRS) are an important analytic technology to aggregate the genetic effects on a trait across the genome. When performing a PRS, there are a number of features that are part of the analysis:   * Dataset in which GWAS summary statistics (effect estimates) are derived (training dataset) * Reference panel in which LD is derived (LD dataset) * Study dataset where the PRS is being evaluated (test dataset)   One of our key questions for this study is: What is the impact of each of these features when conducting a PRS analysis in datasets of mixed ancestry?  Through the analyses in this study, we want to evaluate the impact of ancestry and sample size of the training dataset and the impact of ancestry of the LD dataset on the performance of PRS in different tests datasets.  To maximize sample size, we will use BMI as a quantitative trait for these analyses. The purpose of this paper is to provide details on which of these features have an impact on PRS performance and provide lessons learned through this evaluation. | |
| **Outline of Project** | To conduct this investigation of the features that drive performance of the PRS, we have designed the following study:   * We will use three different training datasets to derive the GWAS summary statistics: UK Biobank full European ancestry dataset, GIANT trans-ancestry meta-analysis results (Locke et al 2015 PMID: [25673413](https://www.ncbi.nlm.nih.gov/pubmed/25673413), and a down-sampled UK Biobank dataset matching sample size and ancestry distribution of GIANT. * We will use five different LD panel reference datasets: 1) 1000 Genomes European superpopulation, 2) all 1000 Genomes (trans-ancestry), 3) 5000 individuals from the testing dataset, 4) 5000 individuals from UKBB European ancestry, 5) 5000 individuals from UKBB trans-ancestry. * We will use PRsice as the PRS methodology. * We will generate PRS in 12 testing datasets from eMERGE   + eMERGE trans-ancestry all ages   + eMERGE trans-ancestry adults (age >=18)   + eMERGE trans-ancestry teens (age from 13 to 17)   + eMERGE trans-ancestry children (age <13)   + eMERGE European ancestry all ages   + eMERGE European ancestry adults (age >=18)   + eMERGE European ancestry teens (age from 13 to 17)   + eMERGE European ancestry children (age <13)   + eMERGE African ancestry all ages   + eMERGE African ancestry adults (age >=18)   + eMERGE African ancestry teens (age from 13 to 17)   + eMERGE African ancestry children (age <13) * We will compare the R2 across all of these analyses to determine which of the features lead to statistically significantly different performance of PRS | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  Geocoding 2015 ACS variables’  Other: Case/Control status on Phase I and Phase II phenotypes |
| **Other Desired Data *(Available from participating sites)*** | We have all of the data needed for this project. | |
| **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 BMI – we already have the data  No | |
| **Planned Statistical Analyses** | PRSice is the PRS methodology selected  Wilcoxon will be used to compare the results of the PRS analysis in the different train, LD reference, and test data | |
| **Ethical Considerations** | None | |
| **Target Journal** | Genetic Epidemiology or Human Genetics | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | * Complete analyses – April 2021 * Draft manuscript to share with coauthors – July 2021 * Submit manuscript – August 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
* 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