|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT463 | |
| **Submission Date** | 11/11/2022 | |
| **Project Title** | Risk factors that affect performance of polygenic risk scores across diverse cohorts | |
| **Tentative Lead Investigator** *(first author)* | Daniel Hui (PhD student in Dr. Ritchie’s lab)  [dahui@pennmedicine.upenn.edu](mailto:dahui@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** | Any eMERGE authors who are interested. | |
| **Sites Participating** | We welcome participation from all eMERGE sites. | |
| **Background / Significance** | Polygenic risk scores (PRS) provide individualized genetic estimates of a phenotype by aggregating genetic effects across hundreds or thousands of loci, typically from genome-wide association studies (GWAS). In recent years it has become increasingly apparent that transferability of PRS performance across different cohorts is poor. Most analyses to-date have focused on ancestry differences being the main driver of this lack of transferability. A growing body of evidence has demonstrated that PRS performance and effect estimates are influenced by differences in certain environmental (classically termed “gene-environment” effects or interactions) or personal-level covariates – different phenotypes seem to be differently affected by these covariates, with most evidence being for adiposity traits such as body mass index (BMI). Several gaps in current knowledge exist about these covariate-specific effects:  -Many analyses have assessed only a handful of these covariates, due to the myriad of choices possible in typical large-scale biobanks  -Little investigation has been done to systematically understand why certain covariates affect PRS performance, with such knowledge being useful to reduce the potential search for variables that impart context-specific effects  -Most studies investigating PRS-covariate interactions have been in European ancestry individuals; unfortunately, comparing differences in PRS performance and prediction while controlling for differences in ancestry versus differences in context has not been assessed in previous studies. Moreover, covariate-specific effects are notorious for having poor replicability in human genetics studies, and previous PRS-covariate interaction studies have been predominantly performed in the UK Biobank (UKBB), which the majority of individuals are aged 40-69 (i.e., excluding young adults) who are overall healthier than those from other e.g., hospital-based, cohorts as well as predominantly European ancestry  -PRS performance is often assessed using linear models and in isolation of clinical covariates, which in practice would often be available. Machine learning models often have increased performance over linear models and are capable of modeling complex relationships and interactions between variables, which may serve to increase predictive performance especially given evidence for PRS-covariate specific effects | |
| **Outline of Project** | Using genetic data with linked-phenotypic records, we estimated the effects of covariate stratification and interaction on performance and effect estimates of PRS for BMI (PRSBMI). These analyses were done across four datasets: UKBB, Penn Medicine BioBank (PMBB) (15), Electronic Medical Records and Genomics (eMERGE) dataset (16), and Genetic Epidemiology Research on Adult Health and Aging (GERA). These datasets include two ancestry groups (N=491,111 European ancestry (EUR), N=21,612 African ancestry (AFR)), and 62 covariates (25 present in multiple datasets) typically associated with cardiometabolic health and adiposity (including blood lipids, socioeconomic status, blood pressure, diet, physical activity, alcohol intake, smoking, and lung function, among others). We performed four main analyses across all cohorts and ancestries:  -Calculated the difference in PRSBMI R2 among individuals stratified by different quintiles of these variables (different groups for binary variables) and assessed whether a variable’s association with BMI could explain its effect on PRSBMI performance across groups.  -Assessed the significance of PRSBMI-covariate interaction terms, and their increases to model R2 over models only using main effects  -Correlated main effects, interaction effect, and R2 differences  -Used machine learning models and evaluated their increase in performance over linear models, even those that included regularization and interaction terms | |
| **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, HDL, LDL, Triglycerides, Total cholesterol  No | |
| **Planned Statistical Analyses** | -R2 values were bootstrapped to measure statistically significant differences between quintiles of covariates  -Significance of PRSBMI-covariate interaction terms and their increases to model R2 over models only using main effects was performed using linear regression  -Main effects, interaction effects, and maximum R2 differences between quintiles were correlated across all cohorts weighted by sample size  -Neural networks were compared to L1-regularized linear regression, with interaction terms and main effects only | |
| **Ethical Considerations** | None | |
| **Target Journal** | Nature Communications | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | -December 2022, share manuscript with collaborators  -January 2023, submit manuscript | |

**\*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