|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT434 | |
| **Submission Date** | 10/04/2021 | |
| **Project Title** | Federated transfer learning methods for constructing PRS in ancestrally diverse populations | |
| **Tentative Lead Investigator** *(first author)* | Tian Gu | |
| **Tentative Lead Investigator Email Address** | tiangu@hsph.harvard.edu | |
| **Tentative Senior Author**  *(last author)* | Rui Duan [rduan@hsph.harvard.edu](mailto:rduan@hsph.harvard.edu) Harvard University  Jordan Smoller [jsmoller@partners.org](mailto:jsmoller@partners.org) Mass General Brigham; Harvard University | |
| **All Other Authors** | Ruiqi Lyu, eMERGE participating sites, others TBD | |
| **Sites Participating** | Open to all eMERGE participating sites | |
| **Background / Significance** | Polygenic risk scores (PRS) have gained prominence as a promising tool to advance precision medicine. However, amounting evidence has shown that PRS accuracy decreases with increasing genetic distance between the discovery and target cohorts, demonstrating low generalizability and transferability across ancestral populations. Due to such lack of representation of non-EA populations in large genomics datasets, the performance of PRS in non-European populations is generally much poorer than the performance in European ancestry (EA) populations, particularly in African ancestry (AA) populations, which has the potential to further perpetuate and even exacerbate known health disparities. Despite some recent efforts in bridging the gap of health disparity via inclusive data collection strategies, it remains challenging from the methodological side to optimize prediction model performance in ancestrally diverse populations accounting for issues such as the substantial amount of heterogeneity in the genetic architectures, LD structures and minor allele frequencies; the lack of representation of non-European population; PRS constructions for admixed populations. | |
| **Outline of Project** | In this project, we will develop a methodology framework for constructing PRS models in ancestrally diverse populations with improved accuracy, generalizability and transferability, and reduce potential biases and disparities due to the lack of representation of non-European populations. We will develop federated transfer learning methods that leverage the similarities in genetic architectures across diverse populations and integrate larger bodies of data from multiple healthcare institutions. We will apply our method to cardio-metabolic diseases, and mental disorders. With high-quality genotypic data and well-documented electronic health records, eMERGE network phase III provides an ideal setting to validate our new method. | |
| **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 |
| **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** | 1. Perform quality control 2. Perform validation analysis of the new PRS method | |
| **Ethical Considerations** | None | |
| **Target Journal** | Depends on results | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Gather data from coordinating center: 10-12/2021  Conduct statistical analyses: 12-1/2021  Write manuscript: 1/2021  Circulate and submit manuscript: 2-3/2021 | |

**\*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