|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT450 | |
| **Submission Date** | 5/24/22 | |
| **Project Title** | Evaluation of the utility of the All of Us Research Program participant data to construct an ancestry calibration model for the eMERGE IV PRS analytical pipeline | |
| **Tentative Lead Investigator** *(first author)* | Chris Kachulis | |
| **Tentative Lead Investigator Email Address** | Christopher Kachulis <ckachuli@broadinstitute.org> | |
| **Tentative Senior Author**  *(last author)* | Niall Lennon | |
| **All Other Authors** | Open to eMERGE and AoU investigators as appropriate based on content | |
| **Sites Participating** | CC (Broad) | |
| **Background / Significance** | Leveraging the diversity of the AoU dataset provides the opportunity to create a more accurate model for improved representation of participants from different ancestries in the high risk category for PRS across the eMERGE IV conditions | |
| **Outline of Project** | This report will evaluate the performance and utility of a population reference panel created from AoU participant genotyping data. The AoU cohort data has several theoretical advantages for use in this way for the eMERGE PRS pipeline: the AoU genotyping data was created with the same array, same manifest, same cluster file as the eMERGE data will use; the AoU cohort has better population representation (including more admixed individuals) than the prior panels used for this purpose (1000G data). Our analysis will dig in to the practical implications of creating a population reference panel in the AoU data and will compare to 1000G performance across the 10 eMERGE conditions. | |
| **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)*** | *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)*  **The only eMERGE data needed is the PRS model information (sites, weights) for the 10 emerge conditions.** | |
| **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** | Model creation. Ancestry adjustment comparisons to 1000G data. | |
| **Ethical Considerations** | No participant-level data (eMERGE or AoU) will be imported into or exported out of existing secure storage locations | |
| **Target Journal** | TBD | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Initial analyses: Completed  Repeat analysis in test chort: 2-3 months  Preprint prior to ASHG 2022 (That is October 2022).  Submission: November 2022 | |

**\*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