|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT431 | |
| **Submission Date** | September 22, 2021 | |
| **Project Title** | PheWAS with colorectal cancer PRS developed in multi-ancestry cohort | |
| **Tentative Lead Investigator** *(first author)* | Elisabeth A. Rosenthal | |
| **Tentative Lead Investigator Email Address** | erosen@uw.edu | |
| **Tentative Senior Author**  *(last author)* | Gail P. Jarvik | |
| **All Other Authors** | David R. Crosslin, , Li Hsu, Ulrike Peters, Minta Thomas, Jihad Obeid, Brenda Mutai, Rich Green, Su Xian, any eMERGE investigator interested in the project. | |
| **Sites Participating** | All eMERGE sites | |
| **Background / Significance** | We are currently improving a previously developed a colorectal cancer (CRC) polygenic risk score (PRS) using multi-ancestry data. We will perform PheWAS using the improved CRC PRS to determine what phenotypes from the electronic health record (EHR) may be associated with it. | |
| **Outline of Project** | 1. Calculate improved CRC polygenic risk scores (PRS) in the current eMERGE GWAS data set.  2. Compare the performance of the improved score with the previous score.  3. Perform PheWAS against EHR derived phenotypes, adjusting for principal components of ancestry. Repeat within ancestry specific cohorts. | |
| **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)*** | *Phenotypes:*  *EHR derived phenotypes, CRC status from developed eMERGE CRC EHR algorithm, diagnosis codes, procedure codes, age, sex, bmi.*  *Covariates:*  *Demographic: Age, sex, self-identified race and decade of birth. Genetic: SNP array data, high penetrance variants for CRC-associated genes from eMERGESeq, when available.* | |
| **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. Measurement improvement in the PRS using AUC and OR.  2. Logistic regression and linear regression of EHR derived phenotypes against the PRS, adjusting for age, sex, and principal components of ancestry | |
| **Ethical Considerations** | None at this time | |
| **Target Journal** | Multiple manuscripts: AJHG, Cancer, Clinical Colorectal Cancer | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | 02/2021: calculate improved PRS  6/2022: Extract and QC phenotypes from EHR  12/2022: PheWAS of all derived phenotypes | |

**\*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