|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT351 | |
| **Submission Date** | 06/17/2019 | |
| **Project Title** | eMERGE PheWAS catalog of GWAS SNPs | |
| **Tentative Lead Investigator** *(first author)* | Todd Edwards, Vanderbilt | |
| **Tentative Senior Author**  *(last author)* | Joshua Denny, Vanderbilt | |
| **All Other Authors** | WeiQi Wei, Qiping Feng, Digna Velez Edwards, Jacob Keaton, Jacklyn Hellwege, Dan Roden | |
| **Sites Participating** | All eMERGE participating sites | |
| **Background / Significance** | **What we know about polygenic risk score (PRS)**. In the past decade, genome-wide association (GWA) studies have provided insights to complex human diseases and facilitated development of novel therapeutics. Recently, international collaborations enabled larger GWAS and increasingly accurate estimation of genetic contribution to complex diseases. By leveraging the available GWAS results, polygenic risk score (PRS), which sum the disease-associated SNPs weighted by their effect size, are a measure of an individual’s susceptibility to disease.1 Although initially used to evaluate each patient’s clinical risk to develop a disease, PRS was proved to have much wider applications, in both biomedical research and clinical practice. For example, the causal relationship between HDL-C levels and CVD risk was a topic of intensive debate. In a Mendelian Randomization study, HDL-C PRS failed to associate with risk of myocardial infarction, even though the HDL-C PRS significantly correlated with measured HDL-C level. 2  **Large EHR-based biobanks** are important for genetic studies. These biobanks, linked DNA samples to a de-identified version of the patient’s EMR that contains longitudinal health care information, from which a large number of cases and controls for specific diseases can be rapidly identified. Furthermore, enabled by the richness of clinical phenotypes, large EHR-based biobanks provide us with an unprecedented opportunity to scan a wide range of clinical phenotypes simultaneously  **PheWAS as an approach to answer clinical questions** The phenome wide association study (PheWAS) was first introduced in 2010. 3 It is an approach that is essentially the obverse of GWAS. PheWAS uses an unbiased approach to ask what phenotypes are associated with one or a few genotypes (PRS), or what phenotypes are associated with another single phenotype. The ability of PheWAS to replicate previously known genotype-phenotype associations provided proof-of-concept 3. Since then, PheWAS has been widely adopted for discovering novel pleiotropic genotype-phenotype or phenotype-phenotype relationships 4,5.  **PRS x PheWAS to answer clinical questions** Combining PRS and PheWAS approaches will provide insight into diseases and their underlying biological mechanisms. Investigators working with the UK biobank have developed PRS for 162 traits and tested their association with 551 traits.6 This elegant work demonstrate that PRS x PheWAS approach can provide a rich resource to infer causal relationships between diseases, such as HDL-C and cardiovascular disease. Yet, we believe that PRSs have even wider applications. | |
| **Outline of Project** | We will construct a PRS for each phecode trait using BioVU data and evaluate the predictive properties of each PRS using the phecode outcomes in the eMERGE dataset. This will provide evidence for which EHR diagnoses are predictable via PRS, both including and excluding other known clinical risk factors. We will also conduct a PheWAS for each PRS using eMERGE resources. | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  •Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279)  •Mental disorders (290-319)  •Diseases of circulatory system (390-459)  •Diseases of the genitourinary system (580-629)  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  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)* | |
| **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** | All proposed genetic data have been QC’d and imputed to HRC reference panels.  We will construct PRSs for phecode outcomes in BioVU using PRSice, Plink, or LD predict. This is an active area of methodological development and we will monitor the literature for best practices. We will additionally stratify by race and build models separately for white and black BioVU participants. We will consider using cross-validation for building PRS models, although we may not have the data density to use this approach for many of the phecode outcomes.  We will evaluate the predictive properties of the PRS models in logistic regression using area under the receiver-operator curve (AUC), with the corresponding C-statistic, confidence interval, and p-value. We will also calculate and report the sensitivity, specificity, PPV, NPV, accuracy, and Matthews correlation coefficient for each model.  Additionally, we will evaluate each phecode PRS with a PheWAS of the phecode outcomes in eMERGE using logistic regression. Results will be compared with the Bonferroni threshold to determine significance. | |
| **Ethical Considerations** |  | |
| **Target Journal** | TBD | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | 1. Complete QC by early 2019  2. Complete analyses by middle 2020  4. Write manuscript in 2020 | |

**\*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