|  |
| --- |
| **eMERGE Network: External Collaborator Manuscript Concept Sheet** |
| **Reference Number** *(to be assigned by CC)* | NT309 |
| **Submission Date** | 10/30/2018 |
| **Project Title** | An efficient algorithm for distributed analysis of pleiotropy in Electronic Health Record data |
| **Tentative Lead Investigator** *(first author)* | Ruowang Li, ruowang@upenn.edu , University of Pennsylvania |
| **Tentative Senior Author** *(last author)* | Yong Chen, ychen123@pennmedicine.upenn.edu ; Jason H. Moore, jhmoore@upenn.edu ; Marylyn D. Ritchie, marylyn@upenn.edu , University of Pennsylvania  |
| **eMERGE Site Sponsor & Contact** | University of Pennsylvania, Marylyn D. Ritchie |
| **All Other Authors**  | Rui Duan, Xinyuan Zhang, others TBD |
| **Sites Participating** | eMERGE participating sites |
| **Background / Significance** | Compared with GWAS, pleiotropic analysis can provide additional insights about the genetic connectivity among seemingly unrelated phenotypes. It is also advantageous to integrate relevant information from multiple EHRs to study pleiotropy because they can in effect increase study sample size and lead to increased power. However, due to identifiability and privacy concerns, patients’ genetic and clinical information are often heavily protected. Thus, we have developed a novel method Fast score test distributed pleiotropy (Fast-DP) to detect pleiotropy that can 1) Simultaneously test pleiotropic effects on multiple phenotypes 2) More computational efficient than the likelihood-ratio test in PheWAS 3) Distributedly integrate summary-level information from multiple EHRs. With high-quality genotypic data and well-documented electronic health records, eMERGE network phase III provides an ideal setting to validate our distributed method |
| **Outline of Project** | 1. **Phenotype definition**

Phenotypes will be defined by applying “rule of three” on longitudinal ICD9 codes.1. **Population stratification**

We plan to conduct pleiotropic association analyses for (1) European population, (2) African American population, respectively.1. **Genomic analyses**
	1. Perform Fast-DP analysis on eMERGE network participating sites to detect potential pleiotropy across multiple phenotypes

Apply functional genomic analysis on discovered pleiotropic variants to evaluate genetic architecture across these phenotypes |
| **Desired Data - Common Variables\*** *(Available from the CC)* | [x] Demographics *(age, sex, race/ethnicity)* [x] 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[x] BMI | [x] Common Variable Labs *(blood lipid levels, serum cholesterol levels)*[ ] 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** | [x] 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 control2.Perform Fast-DP analysis to detect potential pleiotropy3.Perform pathway analysis on discovered pleiotropy via gene set enrichment analysis |
| **Ethical Considerations** | Genomics data and phenotypic data will be de-identified to protect confidentiality. |
| **Available Funding or Resources** |  |
| **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 November, 20182. Complete analyses for identifying pleiotropic variants by December, 20183. Complete pathway analysis by early January, 20194. Write manuscript by January, 2019 |

**\*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