|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT342 | |
| **Submission Date** | 04/29/2019 | |
| **Project Title** | Comorbidity Clusters in Clinical Conditions: An Analysis of Electronic Health Record Data | |
| **Tentative Lead Investigator** *(first author)* | Ting He | |
| **Tentative Senior Author**  *(last author)* | Casey Overby Taylor | |
| **All Other Authors** | Johns Hopkins (Lemke KW, Richards TM, Chute CG, Weiner JP), Mayo Clinic (Arruda-Olson, Kullo I, Ye Z), Kaiser Permanente Washington(Carrell D), Vanderbilt (Denny JC, Wei W), Columbia (Hripcsak G, Kiryluk K, Larson EB), Marshfield (Peissig P), Geisinger (Walton NA), others that are interested | |
| **Sites Participating** | We propose a network-wide study invited to all point people from other sites. The analyses will be led by Johns Hopkins University. | |
| **Background / Significance** | The goal of this project is to investigate the patterns of co-morbidities and assess the different clustering methods which can stratify the severity among conditions. Prior work, ‘Pilot evaluation of ACGs for characterizing co-morbidities in eMERGE cohorts’ led by Casey Overby Taylor and Jonathan Weiner, which characterizing comorbidities using Johns Hopkins ACG system and comparing the results with other comorbidity methods.  Based on this research, we will (a) run clustering methods to identify comorbidity clusters among eMERGE cohorts, (b) define the best methods which can sperate the disease severity well, (c) check if the outcome still held when consideration of patient demographic information. This evaluation will be conducted with a range of selected eMERGE phenotypes. | |
| **Outline of Project** | 1. Clean and structure data for analysis 2. Describe the population characteristics 3. Define the interested eMERGE phenotypes and define the condition severity 4. Draw the population table 5. Apply clustering methods to identify comorbidity clusters among cases 6. Evaluate the clustering methods based on our stratify needs 7. Check the relationships when considering patient demographic groups (i.e., race and gender) into interested data 8. Prepare manuscript for submission | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  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  A subset of eMERGE phenotypes will be selected for this analysis.  No | |
| **Planned Statistical Analyses** | 1. Use a range of clustering methods such as k means and hierarchical clustering to find the comorbidity clusters within different case groups. 2. Generate the precise and recall as the major measurements of clustering methods which can stratify the condition severity. 3. Compare outcomes between different clustering methods | |
| **Ethical Considerations** | None | |
| **Target Journal** | Journal of Biomedical Informatics | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Complete data collection and data cleaning – end of Jan 2019 (already completed)  Complete data analyses – May 15th, 2019  Complete draft manuscript – end of May 2019  Final paper submission for Journal – end of June 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