|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT303 | |
| **Submission Date** | August 20, 2018 | |
| **Project Title** | Classifying design patterns of EHR-driven phenotyping algorithms | |
| **Tentative Lead Investigator** *(first author)* | Yizhen Zhong (Northwestern University) | |
| **Tentative Senior Author**  *(last author)* | Yuan Luo (Northwestern University) | |
| **All Other Authors** | Luke Rasmussen, Justin Starren | |
| **Sites Participating** | (All sites are invited to participate) | |
| **Background / Significance** | To facilitate consistent and reproducible phenotyping algorithm development, eMERGE has previously proposed “phenotype design patterns”. A phenotype design pattern is intended to address a commonly seen issue in the use of EHR data, and provide guidance to a solution. The initial proposed set of phenotype design patterns was developed by a manual review of existing phenotypes, and leveraged expert opinion. The process was time-consuming, and would not scale as more phenotype algorithms are defined. Thus, automated approaches to identify candidate patterns would advance our understanding on best practices in the field of phenotyping. Preliminary work conducted by Zhong et al. explored bag-of-words model or embedding models in order to classify candidate phenotype patterns. This will expand upon these initial findings. | |
| **Outline of Project** | * Develop a richly annotated set of phenotype algorithm descriptions * Evaluate the performance of state-of-art algorithms on phenotype design patterns classification * Evaluate the performance of state-of-art algorithm on phenotype design patterns clustering | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI  (No specific variables are requested, please see Other Desired Data) | 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)*  Written artifacts for phenotypes in PheKB, including those in development  Status of phenotype as of date downloaded for annotation | |
| **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** | Apply natural language processing on phenotyping algorithms. Build one-vs-the-rest classifier for each design pattern. Apply various classification algorithms (e.g. Support vector) and evaluate the overall performance with weighted F1 score. Apply various clustering algorithms (e.g. K-means) and evaluate with adjusted mutual information. | |
| **Ethical Considerations** | Individual participant/patient data is not used for this study. | |
| **Target Journal** | Data scale is small, probably a conference paper | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | August 20, 2018 – Begin project  September 15, 2018 – Complete annotations and reconciliation  September 30, 2018 – Complete analyses  October 10, 2018 – First draft of manuscript to co-authors | |

**\*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