**eMERGE Network Proposal for Analysis**

Project/Manuscript Concept Sheet

|  |  |
| --- | --- |
| **Reference Number** | NT255 |
| **Submission Date** | September 5, 2017 |
| **Project Title** | Deep Phenotyping in electronic health records for facilitating diagnosis of genetic disorders |
| **Tentative Lead Investigator (first author)** | Chunhua Weng |
| **Tentative Senior Author (last author)** | Kai Wang |
| **All other authors** | Jung Hoon Son, Gangcai Xie, Chi Yuan, Tian Kang, David Fasel, Lyudmila Ena, George Hripcsak, Carol Friedman; other eMERGE investigators interested in the project. |
| **Sites Involved** | All sites that will contribute phenotype data and join the collaboration |
| **Background and Significance** | Electronic health records (EHRs) capture rich, fine-grained phenotypic manifestations. The integration of detailed phenotypic information and clinical exome data promises to facilitate phenotype-driven diagnosis of monogenic diseases. However, how to best extract and synthesize phenotypes from heterogeneous EHR using both structured and unstructured data remains a major hurdle, so detailed nuanced phenotype information is typically unused or underused when physicians need to prioritize clinical exome tests for ordering.  Here we propose a high-throughput high-fidelity EHR phenotype extraction and analysis framework using two steps: one for HPO concept extraction and normalization from EHR narratives supplemented by related structured data for labs and medications followed by the other for prioritizing disease genes based on the HPO concepts. The method can expedite clinical diagnoses and increase diagnostic yield using clinical exomes by incorporating EHR-derived gene ranking information. We want to leverage the collaboration of the eMERGE network for supporting the phenotyping and phenotype-driven diagnoses and knowledge management for rare genetic disorders. |
| **Outline of Project** | Aim I. Perform natural language processing on EHR narratives (by genetic counselors or clinical geneticists) and integrate this data with structured EHR data represented by terminology standards to extract human phenotypes for genetic syndromes  Aim II. Compare causal genes from clinical exome tests with ranked list of genes predicted by phenotype-driven gene prioritization |
| **Desired**  **Variables (essential for analysis**  **indicated by \*)** | Clinical variables:   * Age * Gender * Race * Ethnicity * Diagnostic labs which generate positive genetic report |
| **Desired data** | \*Relevant EHR narratives (including genetic counseling notes and molecular pathology reports with phenotypes observed and documented by genetic counselors) written before genetic diagnosis, for all undiagnosed diseases that requires clinical exome to obtain a genetic diagnosis  \*Disease-specific lab and medication data  \*Causal genes from positive diagnosis in clinical report |
| **Planned Statistical Analyses** | Examination of the causal genes among all genes ranked by phenotype analysis  The analysis will be adjusted for age, sex, and disease categories.  The analyses will be conducted using a combination of R packages, NLP software tools and custom developed software tools. |
| **Ethical considerations** | None |
| **Target Journal** | * American Journal of Human Genetics * PLOS Genetics |
| **Milestones** | September 2017: Proposal submission  October-December 2017: Phenotype data aggregation and analyses in internal datasets  January 2018-May 2018 Phenotype data aggregation and analysis in the entire eMERGEseq cohort  June-August 2018: Manuscript draft  August 2018: Manuscript submission |