**eMERGE Network Proposal for Analysis**

Project/Manuscript Concept Sheet

|  |  |
| --- | --- |
| **Reference Number** | NT163 |
| **Submission Date** | Revised: 6.21.2017; Original: June 11, 2015 |
| **Project Title** | The identification of adverse events in the eMERGE PGx cohort using the electronic health record data and machine learning analytical techniques, and assessing the contribution of genetic variation in the 84 pharmacogenes. |
| **Tentative Lead Investigator (first author)** | Jared Erwin |
| **Tentative Senior Author (last author)** | Gail Jarvik, David Crosslin |
| **All other authors** | Ian Stanaway, Iftikhar Kullo, Josh Denny, Joshua Smith, Sarah Stallings, Peggy D. Robertson, David S. Carrell, Adam S. Gordon, Mariza de Andrade, Paul K. Crane, Michael O. Dorschner, Eric B. Larson, Deborah A. Nickerson, Jodell Jackson  ***Any eMERGE II or III authors interested in participating***. |
| **Sites Involved** | University of Washington/Group Health, Seattle, WA;  Mayo Clinic, Rochester, MN;  Marshfield Clinic Research Foundation, Marshfield, WI;  Northwestern University, Chicago, IL;  Vanderbilt University, Nashville, TN;  Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD; and  Broad Institute of Harvard & MIT; Cambridge, MA.  Essentia Health; Duluth, MN.  Geisinger, Danville, PA.  The Mt. Sinai Hospital; New York, NY. |
| **Background / Significance** | Some 9,000 participants in the eMERGE Network were sequenced with the targeted Pharmacogenomics Research Network sequence platform (PGRNseq), thus linking electronic health records (EHR) to pharmacogenetic variant data to ultimately return actionable results. PGRNseq contains the coding regions, UTRs, and 2kb upstream for 84 pharmacogenes.  Our goal is to identify adverse events in the eMERGE PGx cohort using the electronic health record data and machine learning analytical techniques, and assess the contribution of genetic variation in the 84 pharmacogenes.  We will analyze Rx data and diagnosis code data using the Drug Event Base provided by Joshua Smith from Vanderbilt. This will identify a set of possible adverse drug events for all subjects with appropriate data. These adverse events can be used both in unsupervised machine learning to look for interesting grouping among the clinical data and the pharmacogenetic variation. The adverse events could also be used as phenotypes to explore association to the pharmacogentic data.  Illustration of early results can be found here looking at the first 4 PCs of the ADE correlation matrix:  http://rpubs.com/feordin/ADE |
| **Outline of Project** | 1. PGRNseq 9000 eMERGE participants 2. Annotate variants with SnpEff and SeattleSeq 3. Anaylize Rx and diagnosis codes with adverse event database 4. Explore possible AEs and clinical data with unsupervised machine learning (ML) 5. Merge possible adverse events (AE), demog, and genetic data 6. If ML analysis shows interesting grouping, focus on those AEs for association 7. Merge in demographic, generate genetically determined ancestry 8. Assess trends of association of AE’s with 84 PGx genes. Also, stratify by ancestry and assess trends. 9. Summarize data and write manuscript |
| **Desired**  **Variables (essential for analysis**  **indicated by \*)** | **Phenotypes:**   1. PheWAS codes, Rx data, AEs identified by merging Rx data, diagnosis codes and Drug Evidence Base   **Covariates:**   1. Basic demog data will be derived from the CC, and ancestry will be computed using PCA. |
| **Desired data** | PGRNseq data (ongoing), annotated target sequence, Rx data, demog, diagnosis codes (ICD9 codes) |
| **Planned Statistical Analyses** | 1. Generate PCA from PGx data using SNPRelate. 2. Assess patterns of AE’s using k-means clustering and other unsupervised clustering algorithms. 3. Add single variant, gene, and pathway variation data to assess the contribution when clustering |
| **Ethical considerations** | There are no physical risks involved. |
| **Target Journal(s)** |  |
| **Milestones\*\*** | Project duration: 3 months;  First draft: August, 2017 |

**\*\*** This section should include: Timeline for completion of project, including approval, project duration, first and second draft of the paper and submission.