**eMERGE Network Proposal for Analysis**

Manuscript Concept Sheet

|  |  |
| --- | --- |
| **Submission Date** | 9/19/2016 |
| **Project Title** | A Machine Learning Approach to EHR Phenotyping of (Migraine & other) Headaches in Adults  |
| **Tentative Lead Investigator (first author)** | Jen Pacheco |
| **Tentative Senior Author (last author)** | Will Thompson |
| **All other authors**  | Firas Wehbe, Abel Kho, Megan J Roy-Puckelwartz, Dimitri Krainc, Erin Gustafson, Kathryn Jackson |
| **Sites Involved** | Northwestern, other interested eMERGE sites |
| **Background / Significance** | Migraine is manifested by headache that is usually unilateral and frequently associated with nausea, vomiting, and photophobia. It is a significant health problem, with a prevalence of 15% and a cumulative lifetime incidence of 43% in women and 18% in men. Migraine leads to significant economic burden in the US, with the annual loss of paid work was equivalent to about 20 days per migraine sufferer. GWA studies have identified a number of loci, however missing heritability remains a significant problem with these studies. Identification of new genes that contribute to development of migraines will facilitate development of targeted therapies for a subset of migraines and also offer an opportunity for development of preventive treatments (1). There are multiple clinical subtypes of migraine, including migraine with aura, migraine without aura, migraine subtypes based on pathophysiology, and cluster headaches, & a number of these have been studied by GWAS (2). |
| **Outline of Project** | 1. Identify by ICD codes (for ex., ICD-9 codes starting w/ 346) potential headache patients
2. Create a training & testing data set of the headache signs and symptoms of those patients by chart review
3. Use 1 or more existing published criteria to mark the reviewed patients as definitely, probably, or likely not having migraine headache
4. Use NLP of encounter notes, and possibly geocoding, to extract clinical and environmental factors related to headaches, and to extract headache signs, symptoms and sub-types.
5. Use supervised machine learning techniques to classify patients as cases/non-cases.
6. Use unsupervised machine learning techniques to cluster headache patients into sub-populations esp. w.r.t. migraine headaches.
7. Also compare known genotypes associated with migraines to all the PheWAS phenotypes for subjects with those variants, to determine if PheWAS can connect other types of headaches to migraines.
 |
| **Desired****Variables (essential for analysis****indicated by \*)** | 1. \*headache signs, sub-type, symptoms, and severity
2. Comorbidities
3. \*Treatments including medications
 |
| **Desired data** | 1. \*demographics
2. \*headache sub-type & severity
3. comorbidities
4. \*Treatments including medications
5. Geocoded data
 |
| **Planned Statistical Analyses** | Supervised machine learning (precision/recall/F-measure), Unsupervised clustering (statistical analysis of distribution of features to determine quality of sub-populations), Natural language processing for signs and symptoms (precision/recall/F-measure) |
| **Ethical considerations** | None. |
| **Target Journal** | JAMIA, and/or Neurological Diseases clinical specialty journal |
| **Milestones\*\*** | 1. 12/2017 : ready for secondary validation
2. 3/2018: secondary validation complete
3. 5/2018: run at all participating sites
4. 6/30/2018: 1st draft sent
5. 8/24/2018: 2nd draft sent
6. 8/31/2018: submit to journal
 |

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

**References**

1. Cader MZ. The molecular pathogenesis of migraine: new developments and opportunities. Hum Mol Genet. 2013:ddt364.

2. Schurks M. Genetics of migraine in the age of genome-wide association studies. The journal of headache and pain. 2012;13(1):1-9. Epub 2011/11/11. doi: 10.1007/s10194-011-0399-0. PubMed PMID: 22072275; PubMed Central PMCID: PMCPMC3253157.