|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT402 | |
| **Submission Date** | 08/11/2020 | |
| **Project Title** | Identify disease causing factors for RA by high-throughput cluster analyses | |
| **Tentative Lead Investigator** *(first author)* | Rachel Knevel / Marc P. Maurits | |
| **Tentative Lead Investigator Email Address** | rknevel@bwh.harvard.edu | |
| **Tentative Senior Author**  *(last author)* | Beth Karlson / Soumya Raychaudhuri | |
| **All Other Authors** | TBD | |
| **Sites Participating** | We welcome participation of all sites. | |
| **Background / Significance** | In many diseases with unclarified etiology patients exhibit a high variety in symptoms, treatment response and disease outcome. A tantalizing idea is that this clinical heterogeneity is the result of a variety in etiological mechanisms. An example of such a disease is rheumatoid arthritis (RA) where the large variation in disease etiology and progression is still unexplained. Whilst patients with RA are diagnosed with the same disease, prognosis and treatment response significantly differ.[1, 2] The idea that RA exists of multiple unknown subsets is a widely supported notion as well as that multiple factors together contribute to the development of this disease.[3-5]  Recent advances in the field of big data offer novel opportunities to study the clinical heterogeneity underlying many multifactorial diseases. First of all, databases containing longitudinal data on a large number of individuals, collected free of prior hypotheses, are available globally in the form of Electronic Medical Care Records (EMR),[6, 7] which provides a useful tool to obtain coded disease data by translating the high dimensional ICD code structure to phenotypic codes (PheCodes).[8] Secondly, dimensionality reduction techniques such as tSNE and cluster identification methods such as kNN allow unsupervised clustering of samples on high dimensional data.[9, 10] Such approaches have been proven extremely instrumental to comprehensively explore the cell to cell heterogeneity observed in single cell analyses,[11] and have made significant contributions in the discovery of novel cell subsets. We hypothesize that clustering methods from single cell bioinformatic can be similarly useful for observational clinical data.  We aim to identify subset of RA with different, possible etiologic, factors associated with disease development. For this, we will develop a high-throughput pipeline for data analysis. Once we have identified the clusters, we will use the known RA genetic risk variants to identify whether the patients within a cluster share the same risk variants and have different risks than patients from other clusters. | |
| **Outline of Project** | The project will undertake several analyses:  - Development of pipeline for dimensionality reduction and cluster identification  - Harmonization of eMERGE datasets.  - Clustering of harmonized data  - Validating clusters of common phenotypes with lab and medication  - Identification of clusters in patients with RA  - Validation of clusters with known RA risk variants | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  Geocoding 2015 ACS variables  Other: Case/Control status |
| **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 Rheumatoid arthritis  No | |
| **Planned Statistical Analyses** | Test phase  1. Knn clustering  2. tSNE visualization  Validation in other cohorts  3.HARMONY for harmonization  4.Repeat step 1 and 2  5. Test genetic association with identified clusters | |
| **Ethical Considerations** | There are no additional risks involved than the known risk of eMERGE data collection. Informed consent is obtained on all patients to provide blood for DNA analyses. The phenotypic and genetic data will be stored at a secured location in the data storage system at Partners Enterprise Research Information System (ERIS). No data will be shared with unauthorized third parties. Patient’s identity will not be compromised by the proposed analysis. We will also abide by the eMERGE guidelines in this regard. | |
| **Target Journal** | JAIMA, Annals of Rheumatic Diseases | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | Total Duration of the study: 12 months  Method development: 4 months  First submission methodology paper: Oct 2020  Phenotype collection: 2 months  Analysis within RA: 6 months  Draft of manuscript to authors: July 2021  First submission: August 2021 | |

**\*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