eMERGE Network: Manuscript Concept Sheet

Reference Number *(to be assigned by CC)*	NT395
Submission Date	06/28/2020
Project Title	Natural language processing for system lupus erythematosus and its sub-phenotype identification using Electronic health records
Tentative Lead Investigator *(first author)*	Yu Deng
Tentative Lead Investigator Email Address	yu.deng@northwestern.edu
Tentative Senior Author *(last author)*	Yuan Luo
All Other Authors	Theresa Walunas, Jennifer Pacheco, Rosalind Ramsey-Goldman, Abel Kho and others from interested eMERGE sites
Sites Participating	Northwestern University, Vanderbilt and other interested eMERGE sites
Background / Significance	System lupus erythematosus (SLE) is a heterogeneous autoimmune disease that have diverse manifestations. Computational phenotyping using electronic health records on SLE serves as an important foundation to SLE GWAS study. Early efforts on SLE phenotyping focused on a rules-based algorithm using structured data only based on the Systemic Lupus International Collaborating Clinics Classification Criteria for SLE which is comprised of 17 criteria divided into clinical and immunologic domains. In this study, we expanded the previous rule-based algorithm by including Natural Language Processing (NLP) component. We further developed algorithms to detect important sub-phenotypes in lupus: oral ulcer, renal and arthritis sub-phenotypes. We compared the NLP algorithm's performance with the performance of the previous algorithm that used structured data only.
Outline of Project	1. Develop computational phenotyping algorithm using ICD codes and NLP techniques for lupus phenotype and its sub-phenotypes: oral ulcers, arthritis and renal disorder 2. Compare algorithm performance (sensitivity vs specificity) between the above algorithms 3. Validate new (NLP) algorithm at secondary site 4. Implementation of validated (NLP) algorithm by other eMERGE sites
Desired Data - Common Variables* *(Available from the CC)*	☒ Demographics ☒ ICD9/10 codes ☒ 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)* The lupus phenotype is dependent on the following lab data, and for the NLP, the text of encounter notes and kidney pathology reports. All text & lab tests are part of the SLE phenotype developed within eMERGE. Anti-NA, anti-Smith, anti-dsDNA are also part of the autoimmune disease phenotype. Autoantibodies: (anti-Smith, anti-phospholipid, anti-dsDNA) Low Complement Direct Coombs Test Anti-Nuclear Antibody WBC (to include leukocytes and thrombocytes)
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):HLA from PGRNSeq and imputed from other sets
Does project pertain to an existing eMERGE Phenotype?	☒ Yes, if so please list Phase III, Systemic lupus erythematosus phenotype ☐ No
Planned Statistical Analyses	1. Rule-based algorithm to predict lupus phenotype using components from regular expressions and structured data 2. Penalized logistic regression to predict renal sub-phenotype using NLP component (metamap concept, regular expression pattern) and structured data components (ICD diagnosis codes, lab values) 3. Rule-based algorithm to predict arthritis sub-phenotype using components from regular expressions and structured data 4. F-measures, PPV, NPV, sensitivity, specificity to measure model performance.
Ethical Considerations	None
Target Journal	BMC Medical Informatics and Decision Making or JAMIA
Milestones *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)*	- 01/2020: Develop phenotyping algorithm for SLE, arthritis, oral ulcers renal disorder - complete - 05/2020: Validation of algorithm with Vanderbilt - complete - 07/2020: Implementation of algorithm by remaining eMERGE sites - 07/15/2020: Manuscript 1st draft completion and send to co-authors with 2 week deadline for co-authors to reply with suggested edits - 08/6/2020: Manuscript final draft completion and send to co-authors with 1 week deadline for approval by all co-authors - 08/13/2020: Manuscript submission

*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

- 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