|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT391 | |
| **Submission Date** | May 19, 2020 | |
| **Project Title** | Natural Language Processing and other Machine Learning Phenotyping Lessons Learned across the eMERGE network | |
| **Tentative Lead Investigator** *(first author)* | Jennifer Pacheco | |
| **Tentative Senior Author**  *(last author)* | Chunhua Weng, Yuan Luo, and Wei-Qi Wei | |
| **All Other Authors** | Luke Rasmussen, Al’ona Furmanchuk, Garrett E. Eickelberg, other authors from sites who designed, validated & implemented NLP & ML algorithms for eMERGE | |
| **Sites Participating** | Northwestern, Columbia, Geisinger, Harvard, Mayo, Vanderbilt, KPW/UW, and all other interested eMERGE sites | |
| **Background / Significance** | Natural Language Processing (NLP) and other machine learning (ML) promise to extract nuanced phenotypes from electronic health record (EHR) narratives. The wide adoption of popular NLP pipelines such as cTAKES and Metamap, and simpler NLP using regular expressions (RegEx) and rules, could make it easier to share NLP algorithms across different sites. Also, using standard document types, such as those definied by LOINC [Denny], could help specify which text documents to parse. However, the implementation of such algorithms across multiple sites with disparate EHR systems and reproducing their results remains challenging. For example, document types and structures vary across EHRs, and some sites have larger amounts of text than others. In addition, there are multiple methods for training, testing, validating, and implementing such algorithms across sites. For example, the algorithm could be trained and tested at 1 site and then subsequently just tested at another site and then the final model implemented at all other sites; or could be trained at 2-3 sites and then a final model implemented at all other sites; or a single site could gather all possible data from multiple sites along with a set of validated records from each site for training and testing, so that 1 site does all the training and testing and subsequent implementation. Furthermore, there are of course multiple NLP pipelines as mentioned previously, and multiple ML model types that could be used, from logistic regression to random forests to machine learning and in particular deep learning. | |
| **Outline of Project** | The aim of this paper is two fold. Firstly, we will summarize our primary findings and challenges faced during our recent practice of the implementation of five NLP algorithms within the eMERGE network, and possibly other eMERGE network phenotypes that similarily had versions with and without NLP/ML. Secondly, we will make recommendations of “best practices” and a “checklist of site readiness” from eMERGE sites based on our experiences in this pilot project to help achieve this goal at the next cycle of the eMERGE program.  We hypothesize the portability and efficiency of eMERGE NLP development hinges on concerted efforts at both the network level and at the site level, and on the level of human and technological resources available.  Across the eMERGE network several sites have developed NLP &/other ML algorithms for different phenotypes, which used different models and tools, and validated those models, in some of the aforementioned ways. We will compare and contrast these methods and tools retrospectively as implemented for these eMERGE algorithms and present our lessons learned for successfully creating and sharing NLP/ML algorithms across multiple sites. We will only be collecting metadata about phenotype algorithms that are already completed or already planned to be completed in phase III of eMERGE, and the metadata will be from existing sources, such as eMERGE phenotyping (Google) tracking sheets and PheKB. Any new metadata might be from a brief survey that sites would agree to complete to gather additional metadata not already gathered.  In particular we will do the following:   1. Decide on any other data to collect, & develop any other metrics needed for comparing, other than those listed in data & analysis sections below 2. Gather data from existing eMERGE Google tracking sheets and from sites in a Google doc(s) &/or sheet(s) 3. Qualitative assessment of process of developing, validating & implementing NLP/ML algorithms 4. Quantitative assessment of same, such as performance statistics 5. Compare results between sites and algorithms 6. Summarize findings | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | ☐Demographics  ☐ICD9/10 codes  ☐CPT codes  ☐Phecodes  ☐BMI | ☐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)*  All data will be collected where possible from existing Google tracking sheets & docs the Phenotyping Workgroup has been using to track the following information, & the document type LOINC codes. If data is not available from the tracking sheets, it is marked denoting that sites will need to provide the data\*\*:   * Quantitative data:   + Algorithm accuracy statistics from validations (PPV, etc.)   + \*\*How many notes were extracted & analyzed per algorithm per site, for how many subjects; and, if further filtering of notes had to be done to get algorithm to finish in time, how many notes extracted initially vs. in final run   + Descriptive statistics (# of cases & controls, etc.)   + Timelines inc. how long to develop, validate, implement; and also \*\*how long code/pipelines (cTAKEs, Metamap, Python, R, Ruby, etc.) took to run for each alg. at ea. site   + # of lines of code, if any * Qualitative data:   + Whether algorithm applied to adults &/or pediatric cohorts   + Level of estimated difficulty, & predicted scientific value of adding NLP (from initial nomination tracking sheet)   + Document types, inc. descriptions, structure, section headers desired & extracted, & if applicable, \*\*LOINC document types selected to be extracted, & mapping of note types to LOINC & any dacross sites   + NLP pipelines (cTAKES, Metamap, etc.), UMLS versions, other packages used inc. software code (Python, R, Ruby, KNIME, etc.), and type of NLP (RegEx or not, etc.)   + Issues\*\* identified by sites when sharing, inc. technical issues such as not being able to run code on enterprise servers, customization required at individual sites; differences in notes inc. different types & structures, need to filter further to reduce number of notes to analyze, etc.   \*\*Any other quantitative or qualitative statistics sites authors agree to collect | |
| **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:  NLP/ML algorithms in eMERGE. Primarily, COPD/ACO, CRS, FH, SLE (lupus), Arrhythmias/ECGs. Secondly, any other eMERGE phenotype algorithms that used NLP/ML, esp. if algorithm implemented w/o NLP/ML & then again w/ NLP/ML, & inc. pediatric sites, such as atopic dermatitis, PAD, RA.  ☐No | |
| **Planned Statistical Analyses** | * As mentioned above, accuracy statistics where not already calculated to assess algorithms’ performance. Comparison of those accuracy measures across algorithms by type of algorithm * Comparison of the time to develop, validate, & implement across sites by algorithm & type of algorithm * Comparison of numbers of subjects and notes analyzed, and number of resulting cases and controls, across sites by algorithm and type of algorithm * Any other analyses author agree to perform on any listed quantitative or additional quantitative data agree to collect | |
| **Ethical Considerations** | None. | |
| **Target Journal** | JAMIA | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | March - May 2020: sites complete implementation of phenotypes not already implemented  March - May 2020: sites assess how they developed, tested, trained, validated &/or implemented the NLP/ML algorithms using the defined metrics  May 31, 2020 – 1st draft with lessons from primary & secondary sites on development & validation, and initial lessons from all sites on implementations  June 4-7: 1st full draft (~1 week after sites complete implementations)  June 21, 2020: 2nd / final draft  June 30, 2020: submit to journal | |

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

**References**

Denny JC, Spickard III A, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. Journal of the American Medical Informatics Association. 2009 Nov 1;16(6):806-15.