|  |  |
| --- | --- |
| **eMERGE Network: Proposal for Analysis**  Project/Manuscript Concept Sheet | |
| **Reference Number** | NT294 |
| **Submission Date** | June 14, 2018 |
| **Project Title** | Framework for Assessing EHR-Based Phenotype Algorithm Complexity |
| **Tentative Lead Investigator** *(first author)* | Luke Rasmussen |
| **Tentative Senior Author**  *(last author)* | Yuan Luo |
| **All Other Authors** | Jennifer Pacheco, Anika Ghosh, Theresa Walunas, Abel Kho, Jyoti Pathak |
| **Sites Involved** | Northwestern, Cornell & any other interested institutions |
| **Background / Significance** | As more EHR-based phenotypes have been proposed, developed and implemented across multiple institutions as part of eMERGE (and beyond), we have recognized that not all phenotype algorithms are created equal. The time to develop a single phenotype algorithm within an institution, as well as the time for another institution to implement the same phenotype algorithm may vary greatly.  While we understand that some phenotype algorithms are more “complex” than others, this notion of complexity can be measured in different ways. First, a certain amount of work needs to go into pulling the data elements that are fed into the phenotype algorithm. Second, there is a notion of complexity in the implementation of the phenotype logic using the input data elements. Finally, the data elements requested as part of a data dictionary represent another level of complexity (as some co-variates may be a phenotype by themselves).  In this work, we propose to develop a framework by which a complexity metric of a phenotype algorithm may be measured, and provide context around the interpretation of this metric. This builds upon existing work with clinical quality measures [1] and clinical trials eligibility [2], as well as an earlier analysis of eMERGE phenotypes [3,4]. We will assess the metric across implementations of existing eMERGE 1-3 phenotypes, which will also allow us to better understand any changes in complexity of phenotypes over time.  [1] Dorr DA, et al. “From simply inaccurate to complex and inaccurate: complexity in standards-based quality measures”  [2] Ross J, et al. “Analysis of Eligibility Criteria Complexity in Clinical Trials”  [3] Conway M, et al. “Analyzing the Heterogeneity and Complexity of Electronic Health Record Oriented Phenotyping Algorithms”  [4] Thompson WK, et al. “An Evaluation of the NQF Quality Data Model for Representing Electronic Health Record Driven Phenotyping Algorithms” |
| **Outline of Project** | 1. Further refine the proposed complexity metric (initial version presented in Desired Variables) 2. Using the complexity metric definition, use a structured form to quantify/classify the associated attributes for a phenotype definition (number of phenotypes depends on how many individuals are involved and interested – ideally a minimum of 20) 3. Analyze results of the classification |
| **Desired Variables**  *(essential for analysis*  *indicated by* ***\*****)* | \*For each phenotype algorithm analyzed for this study, we will classify:   * Input complexity   + For each data element that is required to calculate a phenotype, classify:     - Type of data element (mapped to OMOP CDM domains)     - Value set used to identify data element       * Vocabulary(ies) used       * Number of total elements after full expansion (if not fully expanded)     - Anticipated modality of the data (structured, free text requiring NLP, handwritten notes requiring OCR)     - Is event/location information needed?     - Is date/time information needed?     - Is associated provider information needed?     - Are associated values (e.g., laboratory value, medication dose/route) needed? * Algorithm logic complexity   + Estimated cyclomatic complexity of pseudocode   + Max depth of Boolean operators   + Number of temporal relationships   + Number of negation/exclusion operators   + Number of arithmetic comparators for observed values (e.g., labs, meds)   + Number of comparators used for counts of data elements * Covariate (data dictionary) complexity   + For each variable within the data dictionary     - If only a structured element, apply classification from Input complexity     - If a higher-level/derived element, apply classification from Input complexity + Algorithm logic complexity   \*Each co-author involved is asked to classify at least two phenotype algorithms to ensure we have sufficient coverage. |
| **Desired Data** | * Phenotype algorithm definition as pseudocode or other higher-level representation (e.g., HQMF) * Data dictionary(ies) |
| **Planned Statistical Analyses** | Summary statistics showing variation across the axes and domains of complexity that we measure. |
| **Ethical Considerations** |  |
| **Target Journal** | Journal of Biomedical Informatics |
| **Milestones\*\*** | June 2018 – Submit concept sheet  End of August 2018 – Complete review and proposal of complexity metric  End of October 2018 – Complete review of phenotypes using metric (co-authors)  End of December 2018 – Complete analyses of results  End of February 2019 – First draft of article complete  End of March 2019 – Review and feedback sent by co-authors  End of April 2019 – Second draft completed  End of May 2019 – Submit manuscript |

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