**eMERGE Network Proposal for Analysis**

Manuscript Concept Sheet

|  |  |
| --- | --- |
| **Reference Number** | NT295 |
| **Submission Date** | June 22, 2018 |
| **Project Title** | Exploring the vagueness of definitions in a phenotype algorithm |
| **Tentative Lead Investigator (first author)** | Anika Ghosh |
| **Tentative Senior Author (last author)** | Luke Rasmussen |
| **All other authors** | Jennifer A. Pacheco, Chunhua Weng, Jyoti Pathak |
| **Sites Involved** | Northwestern, Columbia, Cornell, & any other interested eMERGE sites |
| **Background / Significance** | When defining how to identify all patients with or without a certain phenotype, the opportunity for vagueness in the phenotype algorithm definition exists and can potentially lead to miscommunication or misinterpretation errors. This vagueness can occur in the discussions between a clinician-researcher and a data analyst (acting as a query mediator)[1,2], and also when a data analyst instructs another analyst how to write the query. One example is asking for patients with a diagnosis of diabetes, but not specifying the actual codes from a medical terminology that are considered correct to indicate that diagnosis for a specific context. Another example is “elevated troponin” may correspond to different troponin thresholds that differ in clinical and research settings. Another example is requesting all patients be at least 40 years of age, but not specifying when that criteria should be met (e.g. as of the time the query is executed vs. at the time of first diagnosis)[3].  The risk of vagueness when translating a narrative phenotype to an executable one can be mitigated using common data models (CDMs) and harmonized terminologies (such as with i2b2 or OHDSI). In these cases, executable code is shared across institutions, so no reinterpretation is needed. However, even phenotype definitions created for a CDM may potentially have errors due to ambiguity translating the requirements into the definition itself.  Although eMERGE is moving to use a CDM, we have a wealth of experience with the creation, interpretation, and implementation of narrative phenotype algorithms. For this project, we propose to focus on understanding the source of and methods for mitigation of ambiguity and vagueness within phenotype algorithm definitions (AKA pseudocode) which are then given to other analysts to implement at another institution, considering as well why and how multiple phenotype definitions may exist[4].  We note that examples of identifying and resolving vagueness can be found on PheKB. Sites post phenotype algorithms created by a query mediator, yet implementers of those queries still have questions due to the ambiguity in them, and clarification is needed on how to implement. Each distinct question asked could be seen as a distinct source of ambiguity or vagueness in the algorithm.  From this project, we propose to develop a set of best practices guidelines to remove ambiguity and vagueness during the authoring of phenotypes. Previous work within eMERGE described phenotype “design patterns” which could help in reducing vagueness, but are not alone a complete source address this issue[5]. Clearing up vagueness will not only help when other analysts implement the algorithm on their own database, but it will also help the person writing the pseudo code to make sure the definition is not vague but complete and as unambiguous as possible. We believe this can reduce the amount of errors that arise from incorrectly selected cases and/or controls. |
| **Outline of Project** | 1. Literature review of vagueness and assumptions already noted 2. Draft a phenotyping algorithm authoring guide with a checklist of rules based on these assumptions and ambiguity to avoid when writing “pseudo code” or a query in a CDM for other users to implement 3. Assemble a group of experts to go over the rules and add/delete as needed 4. Select and split up phenotypes on PheKB, into enrichment and validation sets. We will use around 75% of phenotypes in the enrichment set to guide the development of rules. 5. Quantify the degree of vagueness in the enrichment set (i.e., # of statements that are vague or can cause varied interpretations and need to be further specified) 6. Test guide and rules against the validation set to see if the questions asked on PheKB could have been addressed using the guide in the phenotype authoring 7. Data analyses (see below) 8. Write paper |
| **Desired**  **Variables (essential for analysis**  **indicated by \*)** | * From PheKB:   + \*All algorithm definition files and metadata for phenotypes from all eMERGE phases with validated status & non-eMERGE phenotypes on PheKB that are publicly available   + \* Comments/questions and examples of varying interpretations posted on PheKB for those phenotypes   + Data Dictionaries for those same phenotypes, to test for ambiguity in the dictionaries as well |
| **Desired data** | * All definition data on phenotype algorithms selected for this project from PheKB: See above list of variables * Queries &/or other software code (SQL, SAS, R, KNIME, Ruby, etc.) developed to implement phenotype algorithms (de-identified if necessary) |
| **Planned Statistical Analyses** | * Basic Summary Statistics such as the number & types of phenotype algorithms and data dictionaries, types of data used, etc. * Semi-qualitative analysis of the results of testing the rules |
| **Ethical considerations** | None |
| **Target Journal** | Journal of Biomedical Informatics or JAMIA |
| **Milestones\*\*** | * June 2018 – approval of this concept sheet by eMERGE * End of August 2018 – Literature Review and Rough draft of rules * End of November 2018 – Expert Review of the rules * End of January 2019 – Enrichment of the rules using PheKB * End of April 2019 - Evaluating the rules using PheKB * End of June 2019 – Writing the journal article * End of July 2019 – Review of first draft of article by all co-authors * Mid August 2019 –Review of second (final) draft and approval to submit by all co-authors * End of August 2019 – Submit article |

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

**References**

1. Hruby GW, Boland MR, Cimino JJ, Gao J, Wilcox AB, Hirschberg J, Weng C. Characterization of the biomedical query mediation process; Proceedings of AMIA 2013 Clinical Research Informatics Summit; 18–22 March 2013; San Francisco, CA. pp. 89–93.
2. Hruby GW, Rasmussen LV, Hanauer D, Patel V, Cimino JJ, Weng C. A Multi-Site Cognitive Task Analysis for Biomedical Query Mediation. *International journal of medical informatics*. 2016;93:74-84. doi:10.1016/j.ijmedinf.2016.06.006.
3. Musen, M., et al., Knowledge engineering for a clinical trial advice system: uncovering errors in protocol specification. Bull Cancer, 1987. 74(3):: p. 291-6.
4. Richesson RL et al., A comparison of phenotype definitions for diabetes mellitus. JAMIA. 2013;20(e2):e319-e326.
5. Rasmussen LV, Thompson WK, Pacheco JA, Kho AN, Carrell DS, Pathak J, Peissig PL, Tromp G, Denny JC, Starren JB. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. Journal of biomedical informatics. 2014 Oct 1;51:280-6.