Name:	Ning Shang
Affiliation:	Department of Biomedical Informatics, Columbia University
Email:	ns3026@columbia.edu
Presentation type (select one):	Poster

Exploring Data Representation of Parsed Unstructured Clinical Data in Phenotyping Variable Retrieval

Ning Shang, PhD, Alexandre Yahi, MS, George Hripcsak, MD, MS
Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract

In this project, we analyze phenotyping algorithm requirements for using unstructured clinical text, and we explore what elements from parsed clinical notes can be used for retrieving the variables. Fourteen eMERGE phenotypes using unstructured clinical text are analyzed and implemented. Our goal is to determine what information from parsed unstructured clinical data should be included in common data model from phenotyping perspective.

Introduction

Electronic phenotyping uses data from Electronic Health Records (EHRs) and provides computational definitions of phenotypes and consequently supports clinical research and genomic medicine. Refining phenotyping algorithms is a complex and iterative process, requiring inputs from domain experts as well as data scientists. The algorithm uses both structured data (e.g., demographics, diagnoses, procedures, medications, and laboratory tests) and unstructured clinical text (e.g., pathology reports, visit notes)[1]. However, the implemented phenotypes are not portable since each institution may use different source data models and deal with different terminologies. To assist the portability, we have developed parameterized and modularized queries. Using parameterization and modularization, terminology and data schemas are separated from the query logic[2]. With this strategy, we have implemented thirty-one eMERGE phenotypes not only on our local clinical data warehouse (CDW) schema data but also on OMOP Common Data Model (CDM) schema data (V4, can be downloaded from http://phekb.org). In this process, we confirmed that OMOP CDM provides all the key data elements of structured data for phenotyping. However, there is a gap when representing unstructured data for phenotyping. In this project, we discuss our phenotyping implementation experiences dealing with unstructured data and provide our perspective on representing parsed unstructured data in CDM.

Methods

For eMERGE phenotypes that require the mining of unstructured clinical text, we investigated what are the key data elements required from eMERGE phenotyping algorithms. At the same time, we used the natural language processing (NLP) tool cTAKES to parse the unstructured clinical data for our eMERGE cohort (3086 subjects). To meet the phenotyping tasks, the extracted clinical information in the parsed clinical data were identified and annotated. To enable full-text indexing and efficient searching, the parsed clinical data are indexed and stored in Lucene for later information retrieval for phenotyping. Retrieved phenotyping variables are stored in a database. Integrated phenotyping variables information in a database can be directly accessed by the parameterized and modularized phenotyping queries (Figure 1).

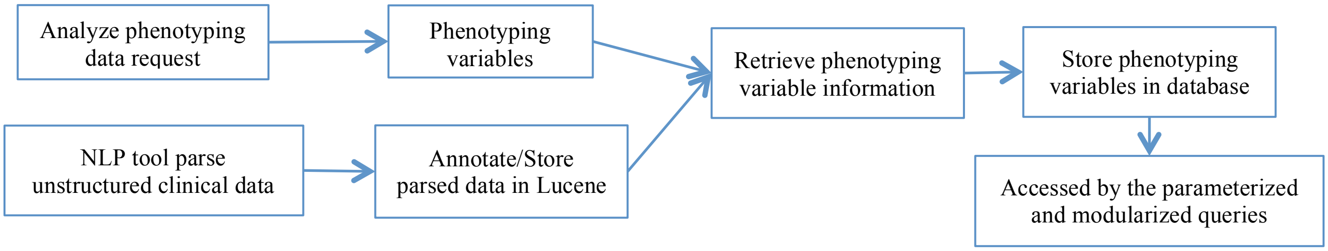
Analyze phenotyping data request
Phenotyping variables
Retrieve phenotyping variable information
Store phenotyping variables in database
NLP tool parse unstructured clinical data
Annotate/Store parsed data in Lucene
Accessed by the parameterized and modularized queries

Figure 1. Phenotyping workflow over unstructured clinical data.

Results

Fourteen eMERGE phenotypes (for example, ADHD, appendicitis, heart failure, VTE, etc.) make use of unstructured clinical text for identifying cohorts. Two major types of unstructured clinical text used are: reports (e.g. pathology, radiology) and visit notes (e.g. visit admission/progress/discharge/signout notes, attending/resident/primary/specialty notes, initial/follow-up notes). There are four types data request from eMERGE phenotypes (Table 1): (1) To check if a patient has a specific examination, which can be confirmed by if the relevant report exists. (2) To check if a condition related entity is documented in the notes. (3) Using notes to find if a relationship between condition entities are documented. (4) Extract important numeric measurements from reports.

Table 1. Major data request for eMERGE phenotyping algorithm.

Data request type	eMERGE example sample
Existence of specific report or note section	Presence of a Pathology Report [Appendicitis]. Must contain at least two Past Medical History sections and Medication lists (could substitute two non-acute clinic visits or requirement for annual physical) [Hypothyroidism].
Term/Concept mentioning in notes or specific sections	At least on diagnosis code for C. diff and at least one affirmative mention of C. diff infection (unqualified by negation, uncertainty, or historical reference) in progress notes [CDiff]. Retrieve DSM-IV Symptom criteria (Social Interaction/Communication/Behavior, Interests and Activities) terms from notes to confirm Autism [Autism]. Positive mention of HF in problem list through either NLP or structured data [Heart Failure].
Related terms mentioning in the same line or adjacent lines	Potential cases were identified if they contained at least one term from List 1 (terms identifying an ace-inhibitor, see below) AND List 2 (terms identifying cough, see below) one the same line (e.g., sentence) within the "Allergy section", "Medication section" or within the entire "Patient summary section" of the EMR [ACEIcough]. At least one non-negated "Disorder related terms" mention and "Anatomical site related terms" mention either in the SAME or adjacent sentences in a 'section of interest' [VTE].
Numeric values with/without temporal constraints	Exclude all patients with an Ejection Fraction (EF or LVEF) <35% within 1 year before or after meeting the CASE 1 definition [Resistant HTN]. In defining "Normal" ECG, QRSd between 65-120ms, ECG designed as "NORMAL", Heart Rate between 50-100, ECG Impression must not contain evidence of heart disease concepts [QRS].

Three main elements (Sentence, textsem.*, refsem.UmlsConcept) from cTAKES parsed clinical text and local developed section identification parser are used to retrieve phenotyping variables from above phenotyping data requests. By breaking down the clinical text into sections, lines, terms and linking UMLS concept to each term and indexing all these information, both raw data and parsed entity data are stored in Lucene for keywords search, concept search, entity relationship search, and multi-field search. Consequently, the Lucene indexed fields include note meta information (note file name, note code, note date, person id, etc.), section information (section raw data, section standardized name, section id), line information (line raw data, line id), text semantic information (raw text and cTAKES semantic modifiers of the text), and UMLS concept entity information (concept coded from which term string, CUI, other coding—SNOMED, RxNORM). Phenotyping variables retrieval from the index is a complicated and variable specific task, so we consider it independent from building phenotyping queries. The retrieval provides phenotyping variable information and is subsequently stored in the database as: variable id, person id, note type, note date, entity for retrieving the variable, the source data for the entity, the types of the entity (e.g. section, line, term string). From this note table, the parameterized and modularized query retrieve the variable information for phenotyping and so utilize the information extracted from clinical text.

Conclusion

This study presents a roadmap of the tools and methods used in our approach to implement eMERGE phenotype algorithms utilizing unstructured clinical data. Identified data requests from phenotype algorithms and relevant elements from parsed notes can be used as use cases and provide recommendations for data representation to OHDSI for extending the CDM for data extracted from clinical text.

References

1. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. Journal of the American Medical Informatics Association. 2013 Dec 1;20(e2):e206–11.
2. Shang N, Weng C, Hripcsak G. A method for enhancing the portability of electronic phenotyping algorithms: an eMERGE pilot study. 2016 AMIA Podium. Accepted.