Making NLP-derived data actionable within the OHDSI ecosystem

Michael Gurley[1], Kyle Zollo-Venecek[2], Andrew Williams[2], Daniel Smith[3], Robert Miller[4], Vipina Kuttichi Keloth[5], Hua Xu[5].
[1]Northwestern University, [2]Tufts University, [3]Winship Cancer Institute of Emory University, [4]Minderoo Foundation, [5]Yale University

Background

It is widely asserted[1] "that about 80% of medical data remains unstructured and untapped after it is created". The OHDSI community has created innovative methodological techniques to partially compensate for missing data within real-world data sets. Even with these techniques, obtaining a portion of this missing data would make the OHDSI community's ability to generate real-world evidence even more robust. Emerging natural language processing (NLP) technology like large language models and knowledge graphs makes the extraction of data trapped in narrative texts a more realistic possibility. Currently, NLP-derived data does not have an actionable home within the OHDSI ecosystem. The NOTE_NLP table is not integrated into Atlas or any of the OHDSI methods libraries. Consequently, NLP-derived data cannot be integrated into OHDSI network studies without resorting to ad hoc workarounds. To remedy this deficiency, the OHDSI NLP Working Group[2] has submitted a proposal[3] to the OHDSI CDM Working Group that creates conventions to enable NLP-derived data to be more widely used within the OHDSI community.

Methods

The proposal includes the following conventions and DDL changes:

- Deposit NLP-derived data into the standard OMOP clinical event tables. (CONDITION_OCCURRENCE, PROCEDURE_OCCURRENCE, DRUG_EXPOSURE, DEVICE_EXPOSRURE, VISIT_OCCURRENCE, MEASUREMENT, OBSERVATION).
- The domain of the concept entered into the current NOTE_NLP.note_nlp_concept_id dictates the placement in the appropriate clinical event table. A concept in the 'Condition' domain goes to CONDITION_OCCURRENCE, a concept in the 'Drug' domain goes to DRUG_EXPOSRUE and so on.
- Set the '_type_concept_id' field in the clinical event table to indicate the NLP-derived provenance of the clinical fact. Use the 'NLP' type concept.[4]
- Link entries to the NOTE_NLP table via the addition of a polymorphic foreign key to the NOTE_NLP table: NOTE_NLP.nlp_event_id and NOTE_NLP.nlp_event_field_concept_id. The NOTE_NLP table falls back to the role of a metadata resource, recording the evidentiary provenance of the NLP-derived fact: a link to the clinical document and the snippet of text supporting the NLP-derived clinical event assertion.

The OHDSI NLP Working Group has also created a NOTE_NLP_MODIFIER extension table to assist ETLers in converting NLP outputs to clinical event tables. NOTE_NLP does not contain enough data to support transfer from it to the clinical event tables. For example, clinical event dates, unit concepts, or value fields. An ETLer would require access to the full data model of the NLP pipeline to move data from NOTE_NLP to the clinical event tables. To remedy this, the NLP pipeline can add modifying rows in NOTE_NLP_MODIFIER to fill out the details not present in NOTE_NLP. This enables full translation from NOTE, NOTE_NLP, and NOTE_NLP_MODIFIER into the OMOP clinical event tables. The OHDSI NLP Working Group is developing a SQL-agnostic script that will implement the translation generically.

See Figure 1 for a representation of the OMOP NLP data model.

Figure 1. OMOP NLP data model with extension table.

Results

To validate the utility of the NOTE_NLP proposal, OHDSI NLP Work Group members at Northwestern, Tufts, and Emory are performing a proof of concept (POC). The POC involves compiling the DDL changes, running a local NLP pipeline to extract a target set of variables (ICDO3 site, ICDO3 histology, and WHO grade) for brain tumor patients against inside/outside surgical pathology reports, and ETLing the NLP outputs in adherence to the guidance of the NOTE_NLP proposal into an OMOP instance. The POC participants will then run an analytic package to compare counts of cohorts having condition concepts that include histology based on EHR discrete diagnoses versus NLP-derived pathology-confirmed diagnoses; the analytic package will also compare differences in initial diagnosis date for brain tumor condition concepts based on EHR discrete diagnoses versus NLP-derived pathology-confirmed diagnoses.

Conclusion

The necessity of incorporating unstructured data into the OHDSI ecosystem is of paramount importance in keeping OHDSI at the forefront of observational research. The NOTE_NLP proposal provides the necessary conventions and structural changes to enable NLP-derived data to participate in the generation of real-world evidence. The NOTE_NLP proposal does not provide prescriptive guidance on the NLP stack to use at a site or the validation methodology to apply to NLP-derived data. The proposal outlines conventions a site should adhere to if they seek to deposit NLP-derived data in an OMOP instance. Future work will focus on creating prescriptive guidance and validation methodologies.

References/Citations

1. Kong, Hyoun-Joong, Managing Unstructured Big Data in Healthcare System. Health Inform Res. 2019 Jan; 25(1): 1–2.
2. Keloth, Vipina K., et al. "Representing and Utilizing Clinical Textual Data for Real World Studies: An OHDSI Approach." Journal of Biomedical Informatics (2023): 104343.
3. https://docs.google.com/document/d/1yjHFwYJN1Xz8QzrWN2uYPCLasmsZCDmF-bZKUNb_H7E/edit

4. See here: https://athena.ohdsi.org/search-terms/terms/32858