| | |
|---|---|
| Name: | Fern FitzHenry |
| Affiliation: | Tennessee Valley Healthcare System |
| Email: | fern.fitzhenry@vanderbilt.edu |
| Presentation type (s): | Poster |


# OMOP CDM for Natural Language Processing: Piloting a VA NLP Data Set


**Fern FitzHenry[1,2], Olga V. Patterson[3,4] Jason Denton[1,2], Jesse Brannen[1], Ruth M. Reeves[1,2], Scott L. DuVall[3,4], Michael E. Matheny[1,2]**

**[1]Tennessee Valley Healthcare System, Veterans Affairs Medical Center, Nashville, TN; [2]Vanderbilt University, Nashville, TN; [3]VA Salt Lake City Health Care System, Salt Lake City, UT; [4]University of Utah, Salt Lake City, UT;**



**Abstract:** *In order to demonstrate the feasibility of hosting the results of Natural Language Processing (NLP) in a Common Data Model (CDM), we piloted a conversion of an ejection fraction dataset, which the national Department of Veterans Affairs (VA) healthcare network extracted using NLP, to the Observational Medical Outcomes Partnership (OMOP) common data model for NLP output. Ejection fraction results, like an estimated 70-80% of all clinical data, are often available in free text documents. CDMs need to accommodate information extracted from these free text documents. We describe a pilot implementation of an initial NLP CDM and the plans to upgrade to future versions.*


**Introduction:** The VA Informatics and Computing Infrastructure (VINCI) maintains and provides access to the national electronic health record extract in the OMOP common data model. Adding the output of NLP to the data set implemented within the OMOP CDM has been a priority, and an NLP use case for populating extracted values for left ventricular ejection fractions (LVEF) found in general notes and procedure notes (echocardiogram and radiology) was piloted. Recently, the Observational Health Data Sciences and Informatics (OHDSI) NLP workgroup published Note and NLP domains in the OMOP CDM.(1, 2) This study describes the pilot conversion of the LVEF NLP tool's output (3) into OMOP.

**Methods:** Since the VA began implementation of the electronic health record in the 1990s, there has been an accumulation of a very large repository of general note documents (~3.5 billion Text Integration Utilities [TIU] documents). In addition, ancillary procedure/result notes are also collected into the Corporate Data Warehouse (CDW). For this pilot project, the note data sources processed were general (TIU) notes, echocardiography reports, and radiology reports. The Salt Lake City VA team applied the LVEF NLP tool on the document to identify numeric LVEF results. Full implementation of the OMOP CDM versions related to NLP requires loading of both the source notes processed by the NLP system, as well as the data extracted via NLP - the notes to the NOTE table, and the NLP output to the NOTE_NLP table.(1) A note identifier is generated for each note as it is loaded into the NOTE table which is referenced in the NOTE_NLP table, linking the NLP results to the source note. However, space constraints on the CDM meant that the text of the notes could not be loaded into the CDM. We, therefore, opted to load only the metadata of the note, such as author characteristics, note title, and note creation date. The LOINC Document Ontology provided the vocabulary for identifying the document type for the source notes, while SNOMED provided the vocabulary for normalizing the NLP data extracted from the note content. The NOTE_NLP table stored the LVEF data extracted from the notes.

**Results:** We have processed and loaded documents containing LVEF into our development area across three document source categories (see table 1). There was a 2% ratio of LVEF findings to document counts for both Radiology and General (TIU) Notes. Echocardiology had a higher ratio, 68%.

Ejection fraction results are typically reported in a text-based format such as the following expressions:

1) "…LVEF ~50%..."

Table 1: Counts of Notes and NLP EF Findings (hits)

| Source | Count |
|---|---|
| Radiology Notes w/ NLP LVEF hits | 4,139,926 |
| Radiology Notes (Metadata) | 172,137,858 |
| Echocardiology Note w/ NLP LVEF hits | 1,133,795 |
| Echocardiology Notes (Metadata) | 1,676,747 |
| General TIU Note w/ NLP LVEF hits* | 925,252 |
| General TIU Notes (Metadata)** | 53,446,315 |

*Pilot:1 medical center loaded. Full set: 43,281,103

**Pilot:1 medical center loaded. Full set: 3,473,879,620

2) "…LVEF is 27% by Teichholz"
3) "…30-5% by visual estimation…"
4) "…severely depressed (est EF<30%)…"
5) "...left ventricular estimated ejection fraction is 45-50%..."

Although the CDM for NLP output captures the phrase expressing the LVEF finding, the model also specifies that this data type (representing the measurement value) be expressed as a single numerical value. Extracted LVEF phrases varied in the degree to which they conformed to the CDM specifications. The LVEF phrases listed above are illustrative of this variation. Phrases such as 2) express a single numerical value, 27%, whereas 1) and 4) include an indicator of less numerical precision. To load expressions representing ranges such as 3) and 5), the mid-point between the two values was determined. Certain phrases such as the type seen in 3) above present problematic values, since it is likely not the midpoint between 5 and 30, but the midpoint between 30 and 35 that was intended in the source note. With range expressions, we added a SNOMED concept qualifier, 4083205 "result comment", to alert the CDM user that the value was based on a range.

In the 5.1 and 5.2 OMOP CDM version of the NLP Notes domain, the only date that would be associated with LVEF findings was the note creation date of the source note document. Procedure notes in the dataset, (echocardiology and radiology) reporting the original results of the procedure were the only documents likely to have the correct date of the LVEF finding. Otherwise, among the general TIU notes, the LVEF was likely to be found as part of patient history or ongoing findings. Section identification was not available from the NLP output, so identifying the note type of the source document associated with the LVEF finding was important to determining the probability of an accurate LVEF date.

In the OMOP NLP CDM, the LOINC document ontology was used to characterize document type based on the VA national standardized title set. The VA participated in building this ontology and has largely transitioned to LOINC compliant standard document titles. We found a VA internal identifier for document type but were not able to find a crosswalk from this code to the LOINC code. For the pilot, we did a manual review of standard titles for matches in the general (TIU) notes to the titles coded in the LOINC document ontology. Just based on this manual review, we were able to assign LOINC codes to about 23% of the general notes. For the echocardiology and radiology notes, we coded the entire source to the same LOINC standard title: 75425-9 [Diagnostic study for cardiovascular disease] and 68604-8 [Radiology diagnostic study note], respectively.

**Conclusions:** The LVEF NLP note findings and source notes were successfully transformed. Numerous design decisions in implementation were required for our use case, including how to map and store the base note data, how often to update the NLP products, how to handle date/time stamping of NLP results, and these lessons learned can be used to help inform future implementations of NLP products in OMOP.

**Financial Support:** This study's work supported with resources and the use of facilities at the TVHS and Salt Lake VA, and is funded by VA HSR&D VINCI and PCORI contract CDRN-1306-04819. Related prior work was supported by AHRQ grant R01HS019913 and NIA grant 1RC4AG039115-01 as part of the American Recovery & Reinvestment Act, National Center for Biomedical Computing (Grant U54 HL108460), and VA HSR&D IIR11-292.

**References**

1. Elhadad N, Belenkaya R, Hua X. Version 5.1 - OHDSI NLP Working Group Recommendations for clinical textual data and NLP output storage and representation schema. 2017 [8-8-2017]; Available from: https://www.ohdsi.org/wp-content/uploads/2016/09/NLPrepresentationschemaforOMOP.docx.pdf.
2. Blacketer C, Elhadad N, Belenkaya R, Hua X. Version 5.2 - Addition of NOTE NLP table and new fields in NOTE table. 2017 [8-9-17]; Available from: https://github.com/OHDSI/CommonDataModel.
3. Patterson OV, DuVall SL. For the Common Good: Sharing data extracted from text. In: *AMIA Annu Symp Proc.*; 2017.