Representing and Utilizing Clinical Textual Data for Real World Studies: An OHDSI Approach

Vipina Keloth, Juan Banda, Michael Gurley, Paul Heider, Georgina Kennedy, Hongfang Liu, Feifan Liu, Timothy Miller, Karthik Natarajan, Olga Patterson, Yifan Peng, Ruth Reeves, Masoud Rouhizadeh, Jianlin Shi, Xiaoyan Wang, Yanshan Wang, Wei-Qi Wei, Andrew Williams, Rui Zhang, Rimma Belenkaya, Christian Reich, Clair Blacketer, Patrick Ryan, George Hripcsak, Noémie Elhadad, Hua Xu

Background

Clinical documentation in electronic health records (EHRs) contains crucial narratives and details about patients and their care. Several general clinical natural language processing (NLP) tools, such as MedLEE[1], MetaMap/MetaMap Lite[2,3], cTAKES[4], and CLAMP[5], have been developed and have evolved over the years to contribute to multiple types of real-world studies, including pharmacovigilance, comparative effectiveness research, and drug repurposing. To promote the use of textual information present in EHRs for observational studies, the OHDSI NLP Working Group (WG) was established in 2015 as part of the OHDSI consortium. In this abstract, we describe a framework for representing and utilizing textual data for real-world evidence generation, the workflow and tools that were developed to extract, transform and load (ETL) data from clinical notes into tables in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), as well as current applications and specific use cases of the proposed OHDSI NLP solution at large consortia and individual institutions.

Methods

To enable storing of clinical text and the information extracted by the NLP tools from the text into the OMOP CDM, the NLP WG was engaged to work closely with the CDM and Vocabulary Development Working Group (CDM WG). As a result of this endeavor, two tables namely NOTE and NOTE_NLP, were incorporated into the OMOP CDM. The NOTE table includes the unstructured clinical documentation of patients in EHRs, along with additional meta information such as dates the notes were recorded and types of notes. The NOTE_NLP table encodes all NLP output from the clinical notes.

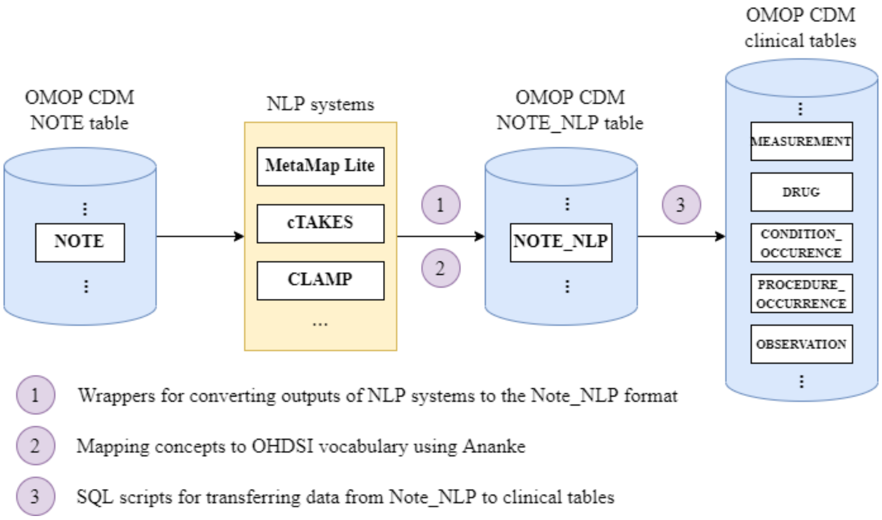
OMOP CDM
NOTE table
⋮
NOTE
⋮
NLP systems
MetaMap Lite
cTAKES
CLAMP
...
1
2
OMOP CDM
NOTE_NLP table
⋮
NOTE_NLP
⋮
3
OMOP CDM
clinical tables
⋮
MEASUREMENT
DRUG
CONDITION_
OCCURENCE
PROCEDURE_
OCCURRENCE
OBSERVATION
⋮
1 Wrappers for converting outputs of NLP systems to the Note_NLP format
2 Mapping concepts to OHDSI vocabulary using Ananke
3 SQL scripts for transferring data from Note_NLP to clinical tables

Figure 1. An overview of the workflow for transforming clinical text in NOTE table.

The ETL workflow for transforming clinical text in the NOTE table is shown in Figure 1. This transformation includes 1) executing the NLP systems to process the textual notes in the NOTE table, 2) converting NLP system output into the NOTE_NLP table, and 3) transferring concepts from NOTE_NLP to individual clinical tables in CDM. More details of the developed tools are described below.

1. Wrappers for converting outputs of NLP systems to the NOTE_NLP format: To ease transforming the output of clinical NLP tools into structured fields, wrappers[6] were implemented in Java to support concept extraction. The wrappers take the clinical text files as input and output the extracted concepts along with other information corresponding to the fields in NOTE_NLP table.

2. Mapping concepts to OHDSI vocabulary using Ananke: The NLP systems (and therefore the wrappers) map the extracted concepts to the UMLS concept unique identifiers (CUIs), not to the concept identifiers within the OHDSI vocabulary. Ananke[7] provides direct mappings between UMLS CUIs and OHDSI concept identifiers.

3. SQL scripts for transferring data from NOTE_NLP to clinical tables: The NOTE_NLP table acts as an intermediate storage for the extracted concepts. SQL scripts were developed to transfer the data from the NOTE_NLP table to the corresponding clinical event tables (e.g., CONDITION_OCCURRENCE, PROCEDURE_OCCURRENCE). All developed tools have been made publicly available at OHDSI NLP WG GitHub repository[8].

Results

Since the release of the NOTE/NOTE_NLP tables in OMOP CDM in 2017, researchers have started exploring the use of them for real world research, including several large initiatives and many individual healthcare systems. We highlight some of large initiatives below and summarize a few other in Table 1.

The All of Us Research Program (AoU): The AoU[9] is building a nationwide cohort to support precision medicine research by collecting genomic, clinical (e.g., EHRs), and lifestyle data for more than one million patients in the U.S. A detailed plan for collecting and processing textual data from AoU participating sites has been developed, following the OHDSI NLP workflow, and hopefully will be ready for the research community to use in 2023.

The National COVID Cohort Collaborative (N3C): The N3C[10] is a collaborative initiative to collect COVID-19 clinical data to answer critical research questions related to the pandemic. An NLP workgroup at N3C have developed an ETL process to populate signs and symptoms of COVID-19 into the NOTE_NLP tables using an example NLP engine MedTagger[11], and implemented and evaluated its performance across multiple participating sites.

The Veterans Health Administration (VHA): The VHA is a branch of the U.S. Department of Veterans Affairs (VA) that provides healthcare to over 9 million veterans every year. To facilitate collaboration and reuse of analytic tools, the VA Informatics and Computing Infrastructure (VINCI) resource center mapped all VHA medical records to OMOP CDM. The use of NOTE_NLP table has been evaluated for mapping the output of an NLP system designed to extract left ventricular ejection fraction (LVEF) from echocardiogram reports.

Table 1. A summary of healthcare systems that adopted the OHDSI NLP solution.

Healthcare organization	NLP tools used	Applications and use cases
University of Utah Health (1.5M patients)	EasyCIE	Venous thromboembolism and pulmonary embolism
Columbia University Irving	MedLEE, HealthTermFinder, and	eMERGE phenotypic algorithms,

Medical Center (6.6M patients)	MedTagger	infectious disease surveillance
Weill Cornell Medicine (3M patients)	RadText	Information extraction tasks from radiology reports
University of Minnesota M Health Fairview (4.5M patients)	Locally trained NLP algorithms	COVID-19 sign/symptom extraction, dietary supplements information extraction.
UMass Memorial Health (3.2M patients)	cTAKES	Suicide prediction models by extracting features (e.g., history of self-harm)
University of Pittsburgh Medical Center (5.5M outpatients)	Locally trained NLP algorithms	Extracting lifestyle-related Social Determinants of Health factors such as sleep-related concepts
Sydney Partnership for Health, Research, Education and Enterprise	Luigi library, multiple spaCy and Hugging Face models	Prevalence and impact of variation in clinical cancer care
Sema4 Mount Sinai Genomics Inc. (multiple health systems serving >10M patients)	Locally developed NLP pipelines based on CLAMP	Extracting genetic variants, protein biomarkers, family medical history, diseases and procedures
Biomedical Informatics Center at the Medical University of South Carolina (~1.5M patients)	DECOVRI built on Apache UIMA; custom medspaCy pipelines	Data Extraction for COVID-19 related Information (specifically, symptom monitoring)

Conclusion

Clinical notes in EHRs are important parts of real-world data and NLP enables the use of textual data in real world studies. Although issues still exist, the OHDSI NLP WG has proposed a framework for representing and utilizing textual data in real world evidence generation, as an initial step to advance the field. Future work includes the development of more methods, tools, and applications to enable efficient and accurate use of textual data for real world research.

References/Citations

1. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. Natural language processing in an operational clinical information system. Natural Language Engineering. 1995;1(1):83-108.
2. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010;17(3):229-36.
3. Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. Journal of the American Medical Informatics Association. 2017;24(4):841-4.
4. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.

Journal of the American Medical Informatics Association. 2010;17(5):507-13.

5. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association. 2018;25(3):331-6.
6. OHDSI NLP tools - Wrappers [cited 2022 Jun 24]. Available from: https://github.com/OHDSI/NLPTools/tree/master/Wrappers.
7. OHDSI Ananke - A Tool for Mapping Between OHDSI Concept Identifiers to Unified Medical Language System (UMLS) identifiers [cited 2022 Jun 24]. Available from: https://github.com/thepanacealab/OHDSIananke.
8. OHDSI NLP tools repository [cited 2022 Jun 24]. Available from: https://github.com/OHDSI/NLPTools.
9. Cronin RM, Jerome RN, Mapes B, Andrade R, Johnston R, Ayala J, et al. Development of the initial surveys for the All of Us Research Program. Epidemiology (Cambridge, Mass). 2019;30(4):597.
10. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association. 2021;28(3):427-43.
11. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. AMIA Summits on Translational Science Proceedings. 2013;2013:149.