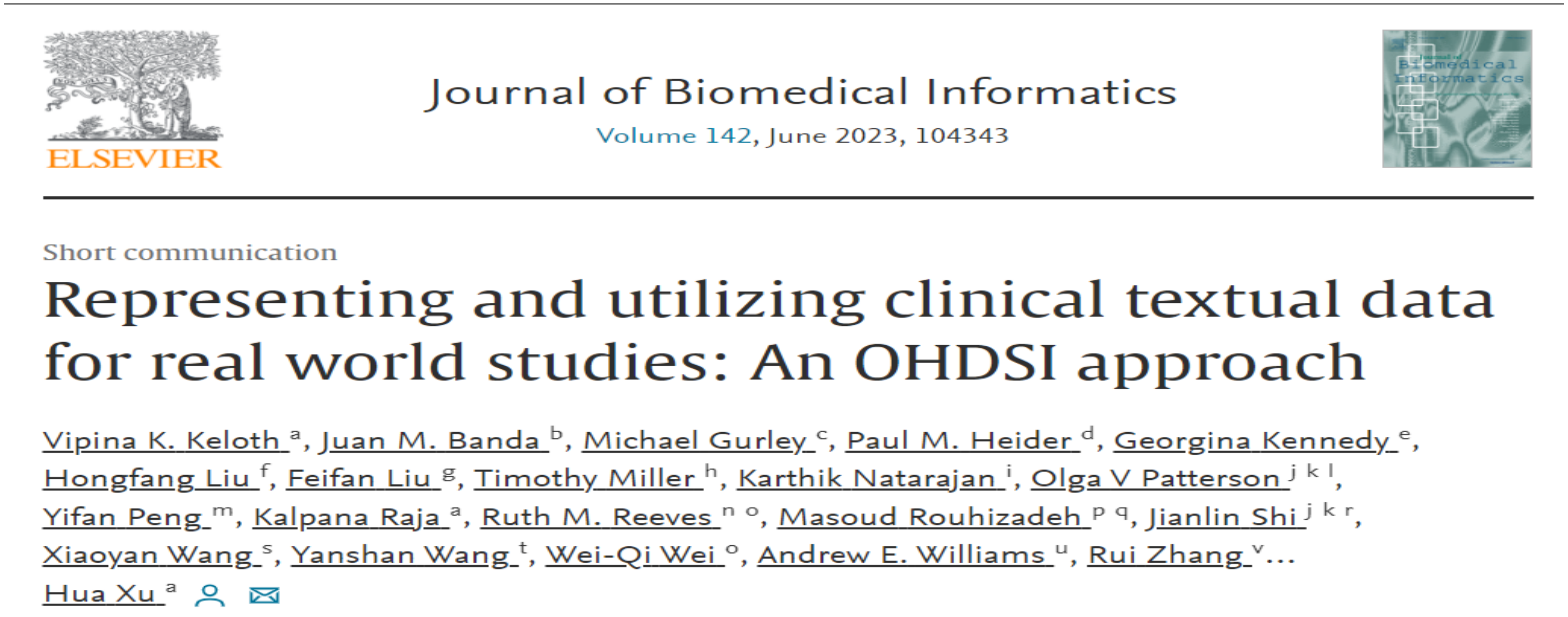
Representing and Utilizing Clinical Textual Data for Real World Studies: An OHDSI Approach

OHDSI Natural Language Processing Workgroup

Vipina K. Keloth

Co-authors: Juan Banda, Michael Gurley, Paul Heider, Georgina Kennedy, Hongfang Liu, Feifan Liu, Timothy Miller, Karthik Natarajan, Olga Patterson, Yifan Peng, Kalpana Raja, Ruth M. 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



ELSEVIER

Journal of Biomedical Informatics
Volume 142, June 2023, 104343

Short communication

Representing and utilizing clinical textual data for real world studies: An OHDSI approach

Vipina K. Keloth [a], Juan M. Banda [b], Michael Gurley [c], Paul M. Heider [d], Georgina Kennedy [e], Hongfang Liu [f], Feifan Liu [g], Timothy Miller [h], Karthik Natarajan [i], Olga V Patterson [j k l], Yifan Peng [m], Kalpana Raja [a], Ruth M. Reeves [n o], Masoud Rouhizadeh [p q], Jianlin Shi [j k r], Xiaoyan Wang [s], Yanshan Wang [t], Wei-Qi Wei [o], Andrew E. Williams [u], Rui Zhang [v]…
Hua Xu [a]

Paper Overview

- Representing textual data in OMOP CDM
- ETL workflow – from unstructured clinical notes to concepts
- Use cases
- Lessons learned/challenges
- Future work

Representing Clinical Textual Data in OMOP CDM

- To enable the storing of clinical text and the information extracted by the NLP tools from the text into the OMOP CDM
  - Note table
    - includes the unstructured clinical documentation of patients in EHRs, along with additional meta information (e.g., dates the notes were recorded, types of notes)
  - Note_NLP table
    - encodes all NLP output (info. extracted by NLP tools) from clinical notes (e.g., concept id, offset, modifiers (temporal, existential))

Note Table

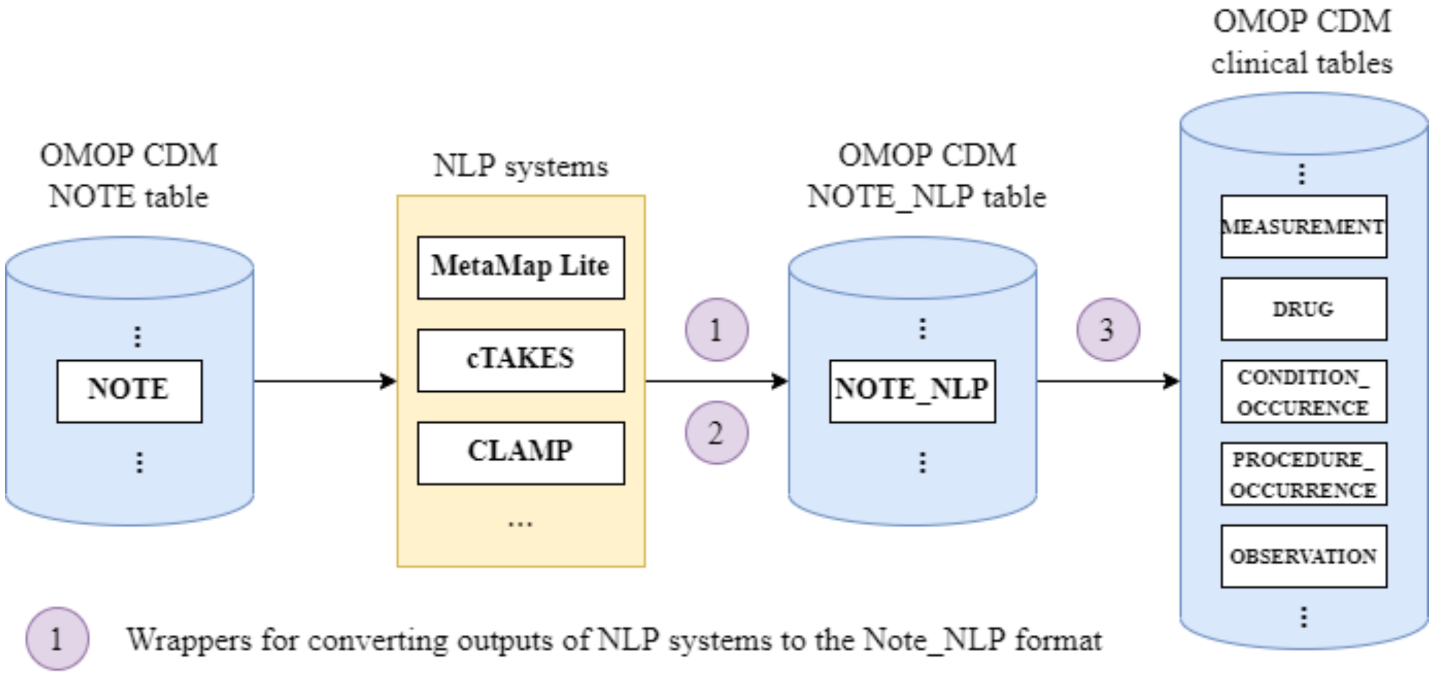
Field	Required	Type	Description
note_id	Yes	integer	A unique identifier for each note.
person_id	Yes	integer	A foreign key identifier to the Person about whom the note was recorded.
note_date	Yes	date	The date the note was recorded.
note_datetime	No	datetime	The date and time the note was recorded.
note_type_concept_id	Yes	integer	The provenance of the note.
note_class_concept_id	Yes	integer	Std. Concept id repr. the HL7 LOINC Doc. Type Vocab. classification of the note.
note_title	No	varchar(250)	The title of the note.
note_text	Yes	varchar(MAX)	The content of the note.
encoding_concept_id	Yes	integer	This is the Concept representing the character encoding type.
language_concept_id	Yes	integer	The language of the note.
provider_id	No	integer	The Provider who wrote the note.
visit_occurrence_id	No	integer	The Visit during which the note was taken.
visit_detail_id	No	integer	The Visit Detail during which the note was written.
note_source_value	No	varchar(50)	The source value mapped to the NOTE_CLASS_CONCEPT_ID
note_event_id	No	integer	primary key of the linked record if the Note record is related to another record in the database
note_event_field_concept_id	No	Integer	If the Note record is related to another record in the database, this field is the CONCEPT_ID that identifies which table the primary key of the linked record came from.

Note_NLP Table

Field	Required	Type	Description
note_nlp_id	Yes	integer	A unique identifier for the NLP record.
note_id	Yes	integer	This is the NOTE_ID for the NOTE record the NLP record is associated to.
section_concept_id	No	integer	The SECTION_CONCEPT_ID should be used to represent the note section contained in the NOTE_NLP record.
snippet	No	varchar(250)	A small window of text surrounding the term.
offset	No	varchar(50)	Character offset of the extracted term in the input note.
lexical_variant	Yes	varchar(250)	Raw text extracted from the NLP tool.
note_nlp_concept_id	No	integer	Foreign key to Concept table. Represents the normalized concept for extracted term.
note_nlp_source_concept_id	No	integer	A foreign key to a Concept that refers to the code in the source vocabulary used by the NLP system.
nlp_system	No	varchar(250)	Name and version of the NLP system that extracted the term.
nlp_date	Yes	date	The date of the note processing.
nlp_date_time	No	datetime	The date and time of the note processing.
term_exists	No	varchar(1)	Term_exists is defined as a flag that indicates if the patient actually has or had the condition.
term_temporal	No	varchar(50)	Term_temporal is to indicate if a condition is “present” or just in the “past”.
term_modifiers	No	varchar(2000)	Term_modifiers will concatenate all modifiers for different types of entities (conditions, drugs, labs, etc.) into one string. Lab values will be saved as one of the modifiers.

ETL workflow for textual data in the CDM

1. Execute NLP systems to process textual notes in NOTE table
2. Convert NLP system output into NOTE_NLP table
3. Transfer concepts from NOTE_NLP to clinical tables in CDM



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

Use Cases of Note/Note_NLP tables

- The All of Us Research Program (AoU)
  - plan for collecting and processing textual data from AoU participating sites developed following the OHDSI NLP workflow (available by 2023)
- The Veterans Health Administration (VHA)
  - use of NOTE_NLP table evaluated for mapping the output of an NLP system designed to extract left ventricular ejection fraction (LVEF) from echocardiogram reports
- The National COVID Cohort Collaborative (N3C)
  - populated signs and symptoms of COVID-19 into the NOTE_NLP tables using MedTagger and implemented and evaluated its performance across multiple participating sites

Use cases - Individual Healthcare Systems

Healthcare organization	NLP tools	Applications
University of Utah Health (1.5 million patients)	A generic rule-based NLP system, EasyCIE	Venous thromboembolism (VTE) and pulmonary embolism (PE)
Columbia University Irving Medical Center (6.6 million patients)	MedLEE, HealthTermFinder, and MedTagger	eMERGE phenotypic algorithms, infectious disease surveillance
Weill Cornell Medicine (3 million patients)	Radiology text analysis system, RadText	Information extraction tasks from radiology reports.
University of Minnesota M Health Fairview (4.5 million patients)	Locally trained NLP algorithms	COVID-19 sign/symptom and dietary supplements.
UMass Memorial Health (3.2 million patients)	cTAKES	Suicide prediction

Healthcare organization	NLP tools	Applications
University of Pittsburgh Medical Center (over 5.5 million outpatient visits every year)	Locally trained NLP algorithms	Social Determinants of Health (SDoH) factors
Sydney Partnership for Health, Research, Education and Enterprise	Luigi library, spaCy and Hugging Face models	Prevalence and impact of variation in clinical cancer care
Sema4 Mount Sinai Genomics Inc. (serving >10 million patients)	Locally developed NLP pipelines based on CLAMP	Five NLP pipelines for extracting genetic variants, protein biomarkers, family medical history, diseases and procedures
Medical University of South Carolina (~1.5 million patients)	DECOVRI built on Apache UIMA; custom medspaCy pipelines	Data Extraction for COVID-19 symptom monitoring

Challenges/Lessons Learned

1. Gaps in standardization of concepts extracted by NLP
   - Some concepts extracted by NLP systems are not present in OMOP vocabularies (e.g., social determinants of health)
2. Challenges regarding the use of NLP systems
   - Different NLP systems are used by different organizations making it difficult to develop a unified NLP solution based on a single NLP tool
3. Implementation issues to meet local application needs
   - Significant amount of resources

Future Work

1. Proposal for modifications in the Note_NLP table
   - adding polymorphic foreign keys to the NOTE_NLP table to link it to the clinical event tables

2. Representing modifiers and more…

References

- Wrappers - https://github.com/OHDSI/NLPTools/tree/master/Wrappers
- Ananke - https://github.com/thepanacealab/OHDSIananke
- OHDSI NLP WG GitHub repository - https://github.com/OHDSI/NLPTools