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

Hua Xu PhD, FACMI

August 30th 2022

Disclosure

- Founder:
  - Melax Technologies Inc. - Dr. Hua Xu and The University of Texas Health Science Center have research related financial interests at Melax Technologies Inc.

- Consultant:
  - Hebta LLC.
  - More Health INC.
  - Bayer US LLC.

Outline

01 Introduction to EHR, clinical notes, and NLP

02 NLP Working Group at OHDSI: CDM, Tools, Use Cases

03 Challenges and future work

01 Introduction to EHR, clinical notes, and NLP

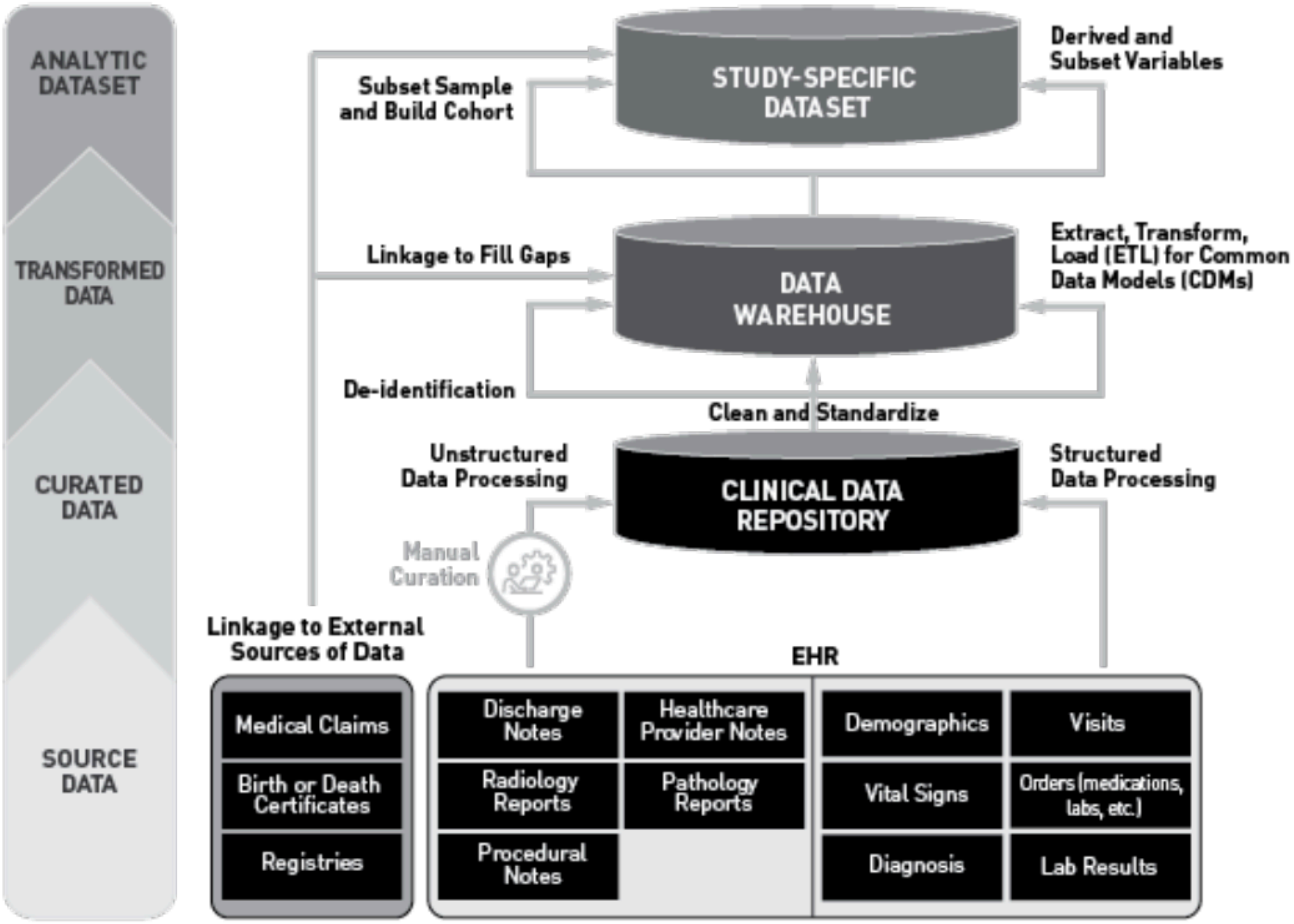
02 NLP Working Group at OHDSI: CDM, Tools, Use Cases

03 Challenges and future work

Electronic Health Records (EHRs) for Real World Evidence (RWE)

- EHRs (and linked data) becomes an enabling resource for RWE

RWD Guidance 2021 *



* Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products – Draft Guidance by FDA, September 2021

Textual Documents in EHRs

Admit 10/23
Medical History: 71 yo woman h/o DM, HTN, Dilated CM/CHF, Afib s/p embolic event, chronic diarrhea, admitted with SOB. CXR pulm edema. Rx'd Lasix.
Social History: PT isolates to self in her apartment.
All: none
Meds Lasix 40mg IVP bid, ASA, Coumadin 5, Prinivil 10, glucophage 850 bid, glipizide 10 bid, immodium prn

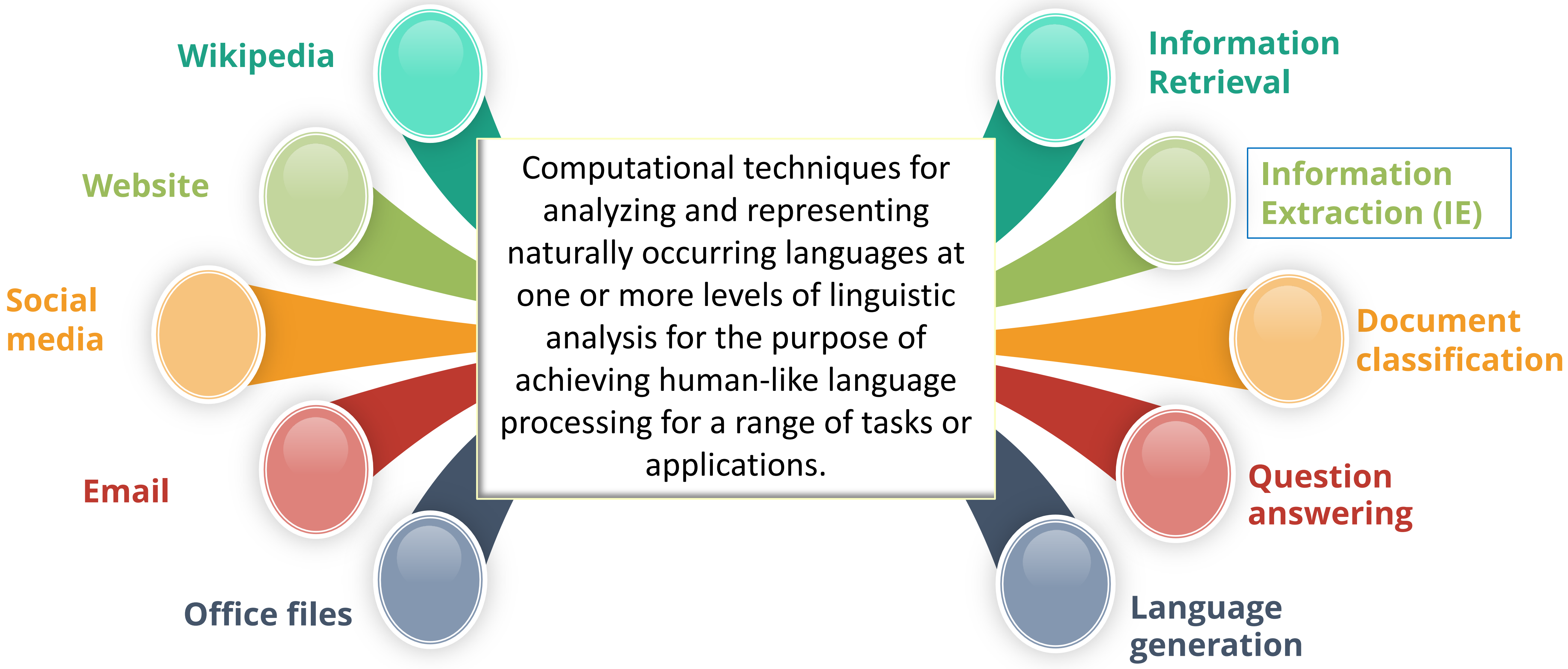
Medical History

Social History

Treatment Response

More details …

Natural Language Processing (NLP)



Active Development of Clinical IE Systems

General Purpose

- MedLEE
- MetaMap
- cTAKES
- CLAMP

Specific Purpose

- Smoking status
- PHI De-identification
- Social determinants
- Bleeding events
- Cancer metastasis
- ……

Three Main Components for Clinical Information Extraction

Named Entity Recognition - NER

Recognize boundary and type of an entity mention in the text

Relation Extraction - RE

Extract modifiers of main entities, such as negation, subject, conditional, certainty, temporal etc.

Concept Normalization - CN

Link an entity to a concept in an ontology, also called entity linking

NLP Challenge Tasks		Ranking
Named Entity Recognition	2009 i2b2 medication information extraction	#2
	2010 i2b2 problem, treatment, test extraction	#2
	2013 SHARe/CLEF abbreviation recognition	#1
	2016 CEGS N-GRID, De-identification	#2
Relation Extraction	2012 i2b2 Temporal information extraction	#1
	2015 SemEval Disease-modifier extraction	#1
	2015 BioCREATIVE Chemical-induced disease from literature	#1
	2016 SemEvel, temporal information extraction	#1
	2017 TAC ADR extraction from drug labels	#1
	2018 n2c2, medication and associated ADR	#1
Concept Normalization	2014 SemEval, disorder encoding	#1

Named Entity Recognition (NER)

- The 2010 i2b2 Challenge: recognize problem, treatment and test

"Plavix was not recommended, given her recent GI bleeding."

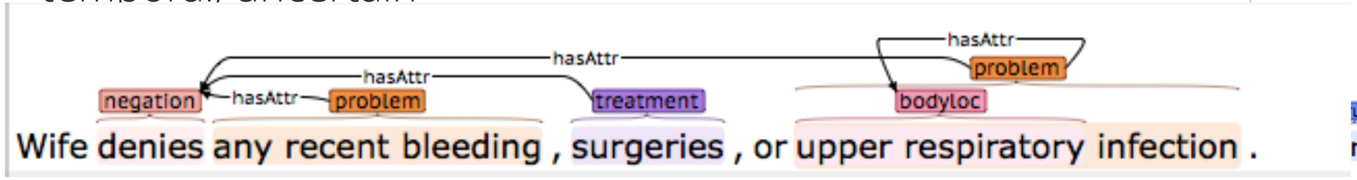
B O O O O B I I I

Algorithms	Feature	F1
CRFs (Jiang et al., 2010) (#2 in challenge)	Bag of words	77.33
	Optimized features	**83.60**
Semi-Markov (deBruijn B, et al., 2010) (#1 in challenge)	Optimized features + Brown clustering	**85.23**
SSVMs (Tang et al., 2014)	Optimized features + Brown clustering + Random indexing	85.82
CNN (Wu et al., 2015)	Word embedding	82.77
Bi-LSTM-CRF (Wu et al., 2017)	Word embedding	85.91
BERT (Si et al., 2020)	Pre-trained language model - BERT, fine tuned on clinical text	**90.25**

Relation Extraction (RE) – Modifiers of Clinical Entities

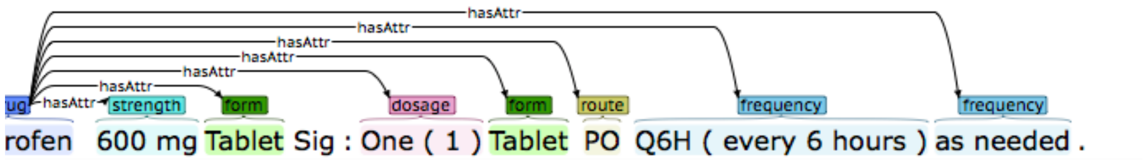
Problem

severity, condition, negation, subject, bodyloc, temporal, uncertain



Drug

duration, dosage, route, strength, form, frequency …



2018 Drug-ADR	SVM	post-processing	CNN-RNN	+ post-processing	biLSTM-CRF	+ post-processing
Strength -> Drug	0.9704	0.9792	0.9760	0.9853	0.9865	0.9916
Dosage -> Drug	0.9637	0.9798	0.9642	0.9818	0.9720	0.9860
Duration -> Drug	0.84	0.8947	0.8519	0.9125	0.8829	0.9292
Frequency -> Drug	0.9525	0.9735	0.9592	0.9810	0.9692	0.9873
Form -> Drug	0.9728	0.9867	0.9713	0.9864	0.9765	0.9890
Route -> Drug	0.9581	0.9742	0.9668	0.9805	0.9736	0.9858
Reason -> Drug	**0.7328**	**0.8364**	**0.7464**	**0.8466**	**0.7579**	**0.8488**
ADE -> Drug	**0.7604**	**0.8221**	**0.7528**	**0.8112**	**0.7946**	**0.8502**
Overall	0.9256	0.9521	0.9304	0.9574	0.9399	0.9630

Concept Normalization (CN)

- Example: “right below - knee amputation”
- Candidates:
  - 1: C2202463 amput below knee leg right
  - 2: C0002692 amput below knee
  - 3: C0002692 amput below bka knee
  - …

Task	Dataset	Method	Accuracy
SNOMED-CT	clinical text 2013 ShARe/CLEF 2014 Semeval	BM25 + Domain knowledge+RankSVM (#1 in challenge) (Zhang, 2014)	0.873
		BM25 + domain Knowledge + CNN (Tang, 2017)	0.903
		BM25 + BERT (Ji, 2019)	**0.911**
MedDRA	drug labels 2018 TAC ADR	BM25 + Translational model + RankSVM (#1 in challenge) (Xu, 2018)	0.911
		BM25 + BERT (Ji, 2019)	**0.932**
MeSH	biomedical literature NCBI	BM25 + domain Knowledge + CNN (Tang, 2017)	0.861
		BM25 + BERT (Ji, 2019)	**0.891**

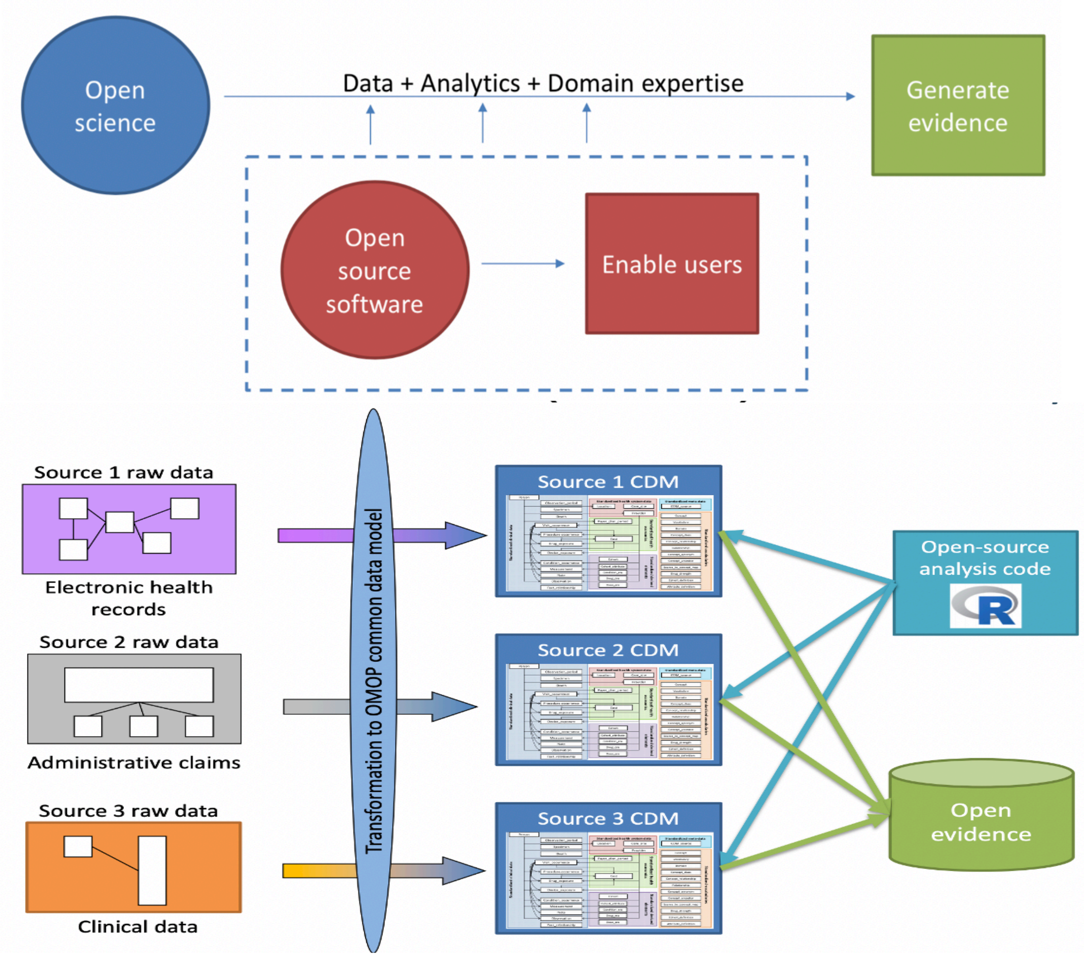
01 Introduction to EHR, clinical notes, and NLP

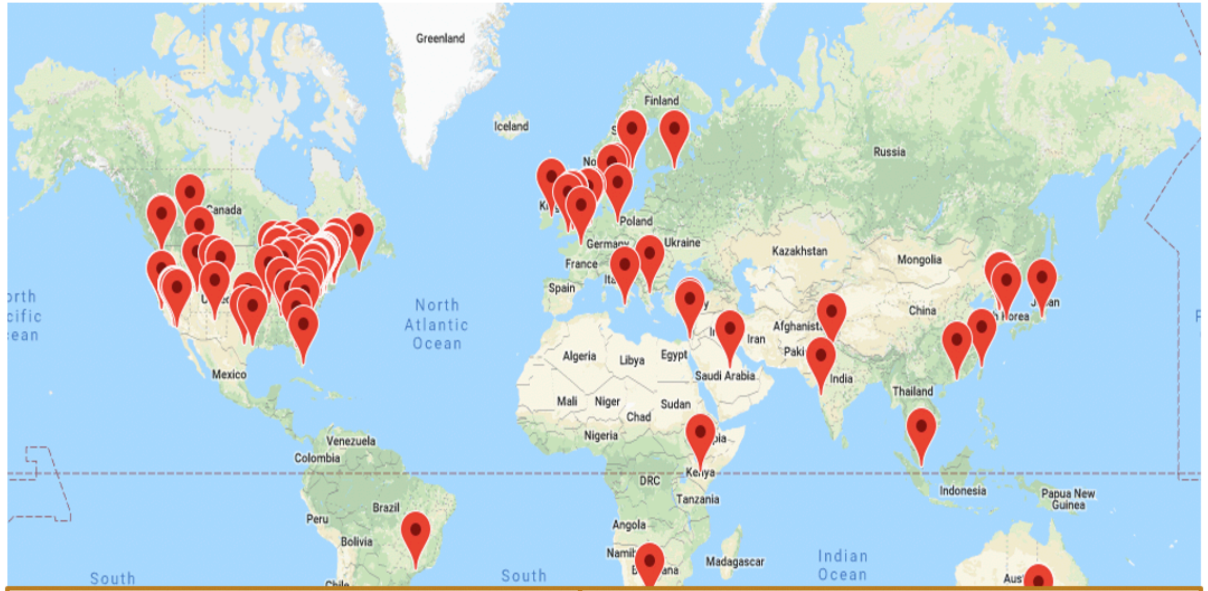
02 NLP Working Group at OHDSI: CDM, Tools, Use Cases

03 Challenges and future work

The Observational Health Data Sciences and Informatics (OHDSI) Consortium

- A multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics





OHDSI Collaborators:
- >2,770 researchers in academia, industry and government
- >21 countries

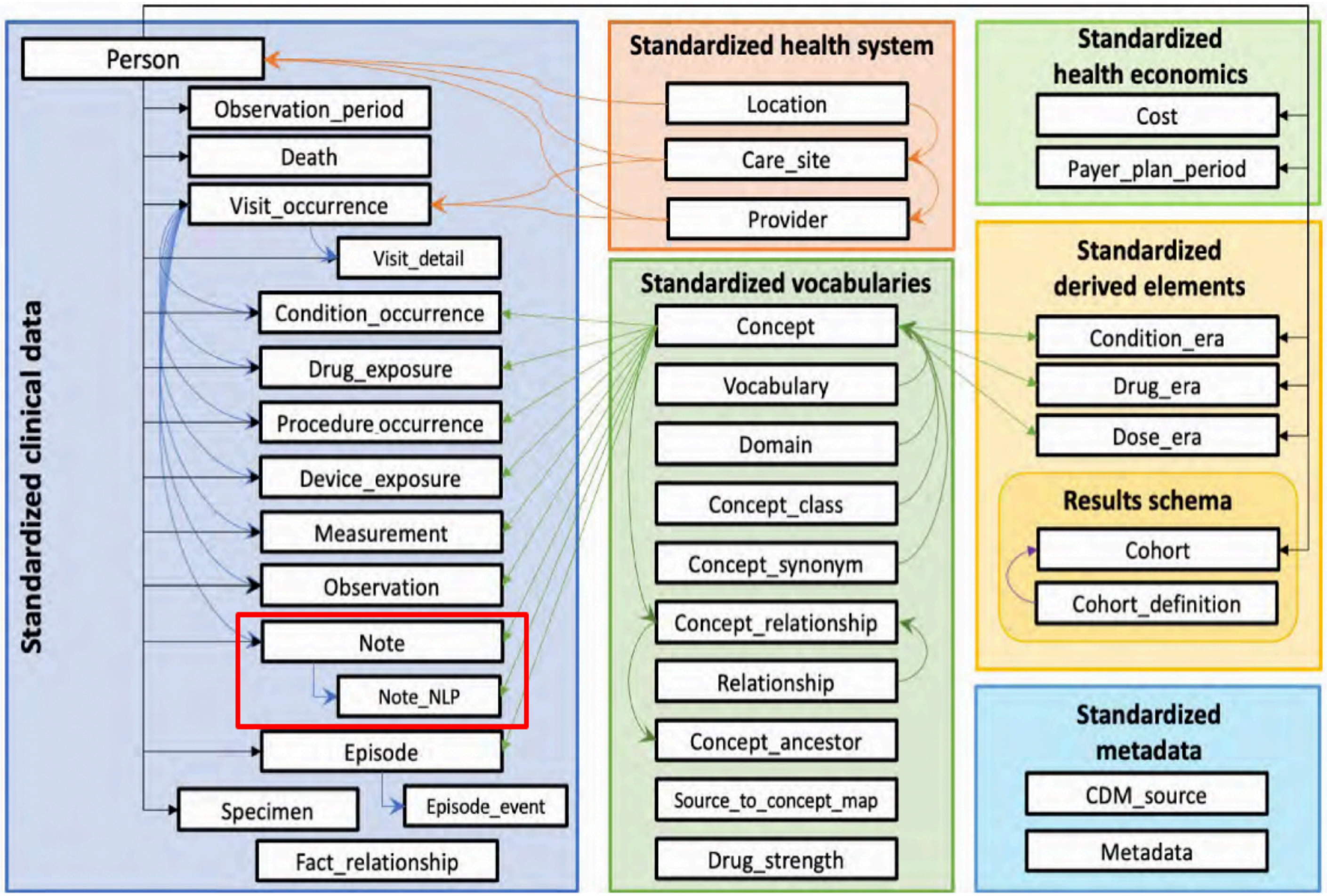
OHDSI Data Network:
- >133 databases from 18 countries
- 1.9 billion patient records (duplicates)
- ~369 million non-US patients

OHDSI NLP Working Group

- Established in 2015, with the goal to promote the use of textual data in electronic health records (EHRs) for observational studies under the OHDSI umbrella
- Three objectives:
  - Develop standard representations for clinical text and NLP output data
  - Build methods and tools to facilitate textual data processing
  - Conduct cross-institutional studies and disseminate best practice of using textual data for real world evidence generation
- Available at https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:nlp-wg

Representing Clinical Texts and NLP Outputs 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 - store select NLP outputs from clinical notes (e.g., name and concept id, modifiers)



Note Table

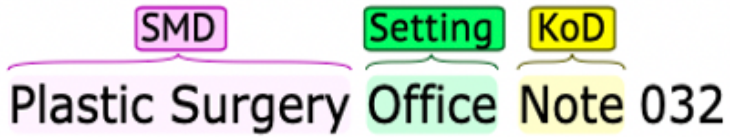
Field	Required	Type	Description
note_id	Yes	integer	A unique identifier for each note.
person_id	Yes	integer	A foreign key identifier to the Person about whom the note was recorded.
note_date	Yes	date	The date the note was recorded.
note_datetime	No	datetime	The date and time the note was recorded.
note_type_concept_id	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

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.

Extract Note Table from EHRs – Note Type Standardization

- Can we extract note type information from note titles only?
  - Convert into an NER task
  - Develop ML/DL methods



18,075 clinical document titles from five institutions

- Boston Children's Hospital (7,400)
- Vanderbilt University Medical Center (3,434)
- Stanford University Medical School (3,128)
- The University of Texas Health Science Center at Houston (3,232)
- Columbia University Medical Center (881)

LOINC Document Ontology (DO) Axis:

- ***Type of Service (ToS)***: the kind of healthcare services provided to patients. e.g., Consultation, Evaluation and Management, Procedure
- ***Kind of Document (KoD)***: the type of clinical documents based on its structure. e.g., Note, Report, Checklist
- ***Setting***: the location or channel where clinical care is provided. e.g., Ambulance, Birthing Center, Intensive Care Unit
- ***Role***: people and their occupations involved in the service or authors who created the clinical note. e.g., Physicians, Nurse, Pharmacist
- ***Subject Matter Domain (SMD)***: clinical specialty relevant to the document or the main purpose of creating the document. e.g., Anesthesiology, Urology, Cardiovascular Disease

Dataset Statistics

- Annotated 4,000 note titles from 5 institutions

Institution	Criteria	ToS	KoD	Setting	Role	SMD
BCH	Exact Match	47%	87%	90%	93%	42%
	Fuzzy Match	51%	13%	10%	7%	55%
	Not Covered	2%	-	-	-	3%
Columbia	Exact Match	67%	81%	86%	95%	41%
	Fuzzy Match	30%	19%	14%	4%	55%
	Not Covered	3%	-	-	1%	4%
UT Health	Exact Match	91%	95%	94%	92%	87%
	Fuzzy Match	9%	5%	6%	7%	13%
	Not Covered	1%	-	-	1%	-
Stanford	Exact Match	53%	83%	72%	92%	48%
	Fuzzy Match	44%	17%	28%	8%	48%
	Not Covered	3%	2%	-	-	4%
Vanderbilt	Exact Match	89%	86%	90%	95%	87%
	Fuzzy Match	10%	14%	9%	4%	12%
	Not Covered	1%	-	1%	1%	1%

NER Results

LOINC DO Axis	Precision		Recall		F1	
	BERT	CRF	BERT	CRF	BERT	CRF
ToS	0.7187	0.7880	0.7848	0.7270	**0.7494**	0.7120
KoD	0.9076	0.9110	0.9286	0.8930	**0.9179**	0.9020
Setting	0.8911	0.9190	0.9226	0.8940	0.9058	**0.9060**
Role	0.8810	0.9210	0.8837	0.8610	0.8811	**0.8900**
SMD	0.8153	0.8139	0.8434	0.7880	**0.8290**	0.8000

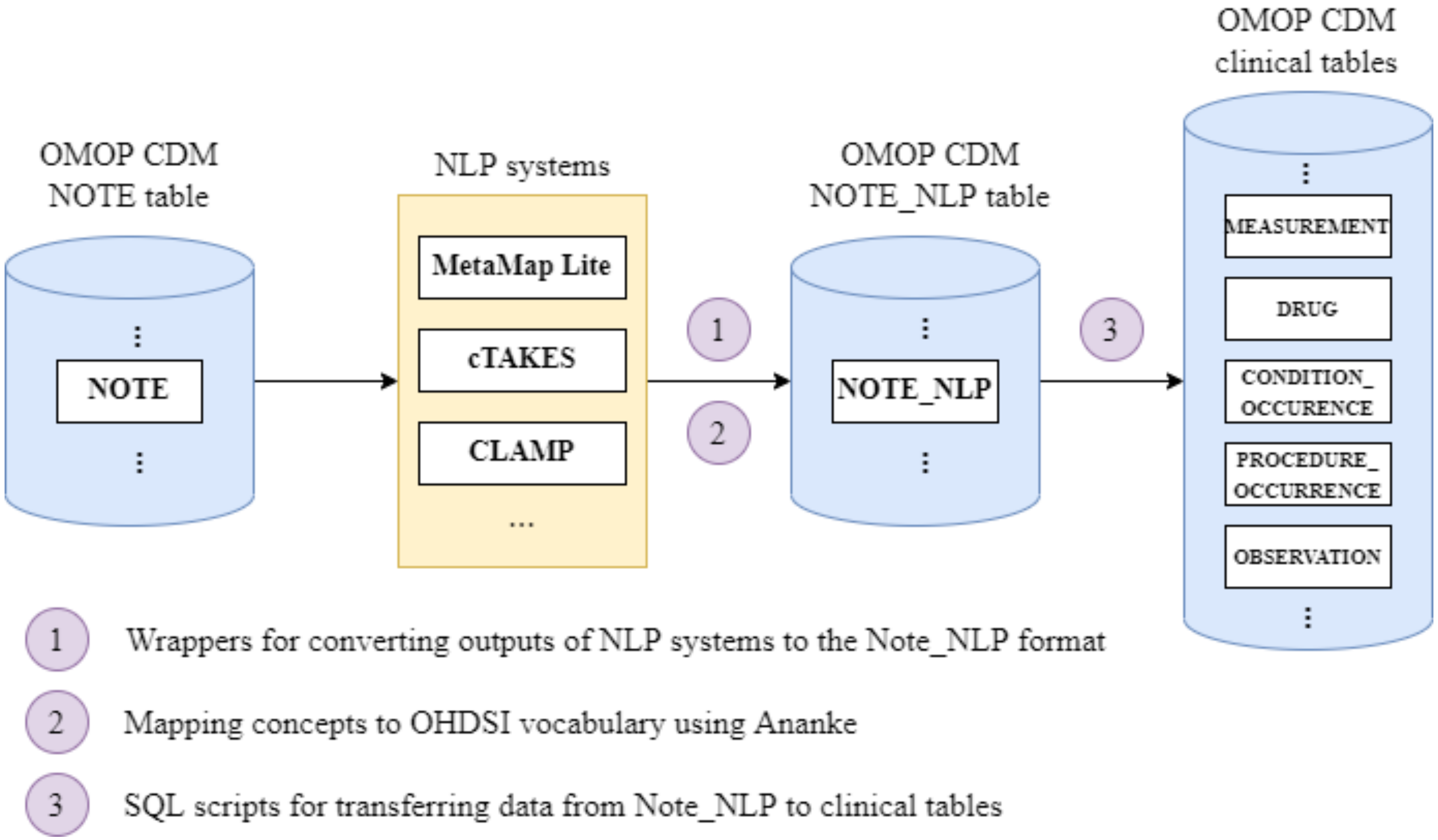
Institution	ToS		KoD		Setting		Role		SMD	
	BERT	CRF	BERT	CRF	BERT	CRF	BERT	CRF	BERT	CRF
BCH	0.4567	0.5030	0.8418	0.7010	0.8862	0.8480	0.7592	0.5780	0.7290	0.6470
Columbia	0.6533	0.6250	0.8860	0.8600	0.8823	0.9160	0.6957	0.6670	0.7234	0.6630
UT Health	0.8185	0.8130	0.9317	0.9340	0.9284	0.8000	0.9431	0.9420	0.9397	0.9240
Stanford	0.5657	0.6440	0.8983	0.8520	0.8326	0.7940	0.8256	0.8500	0.7284	0.6400
Vanderbilt	0.9165	0.9190	0.9679	0.9440	0.9544	0.9730	0.9450	0.9590	0.9487	0.9260

Note Type Normalization Discussion

- Findings from this study:
  - LOINC DO has a relatively high coverage over document titles
  - BERT model works better than CRF in general
  - Note titles alone are not sufficient to provide note type information
- Practical solution
  - Extract metadata of notes from EHRs to provide additional LOINC DO information
  - Currently OHDSI and PCORNet (GPC) are working together to develop queries to derive LONIC DO axes from Epic and Cerner
  - We may also rely on document content to decide note types

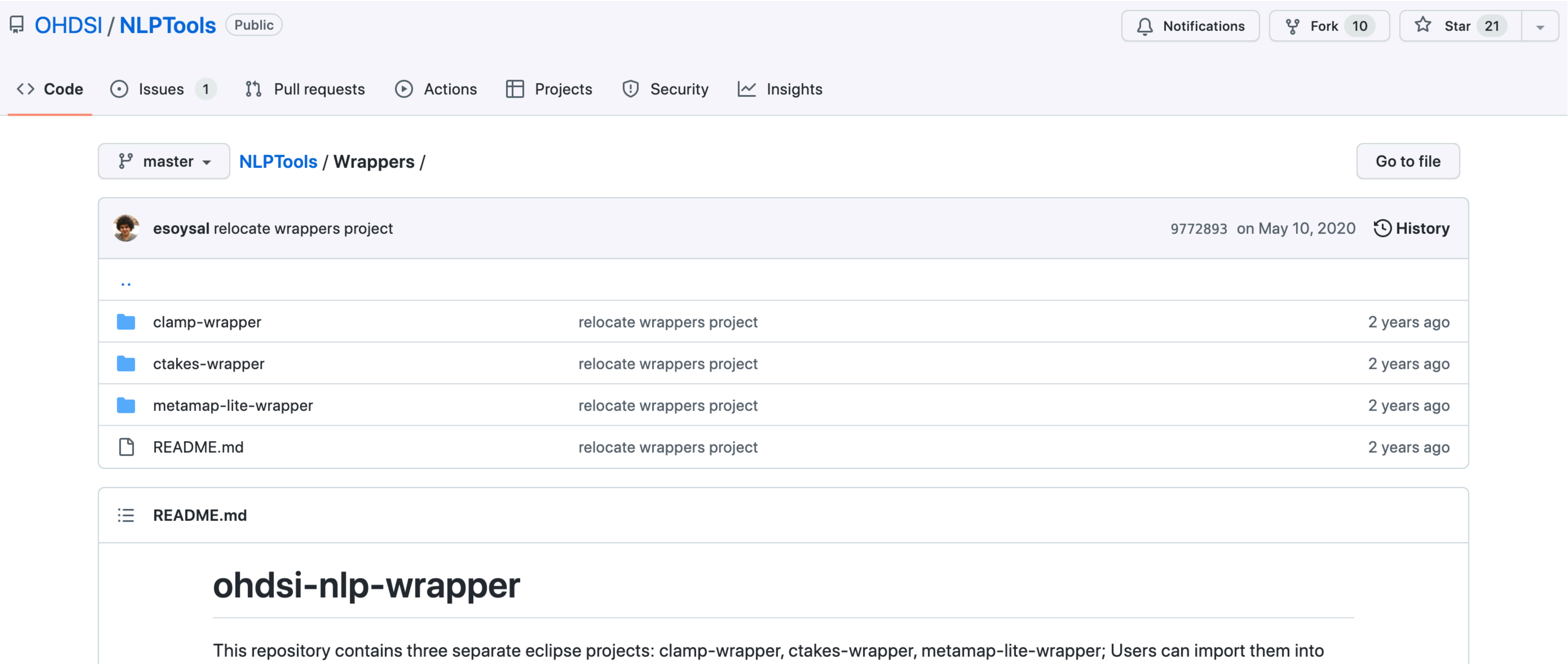
NLP Workflow for Textual Data in CDM

- Run NLP systems to process textual notes in NOTE table
- Convert NLP system output into NOTE_NLP table
- Transfer concepts from NOTE_NLP to clinical tables in CDM



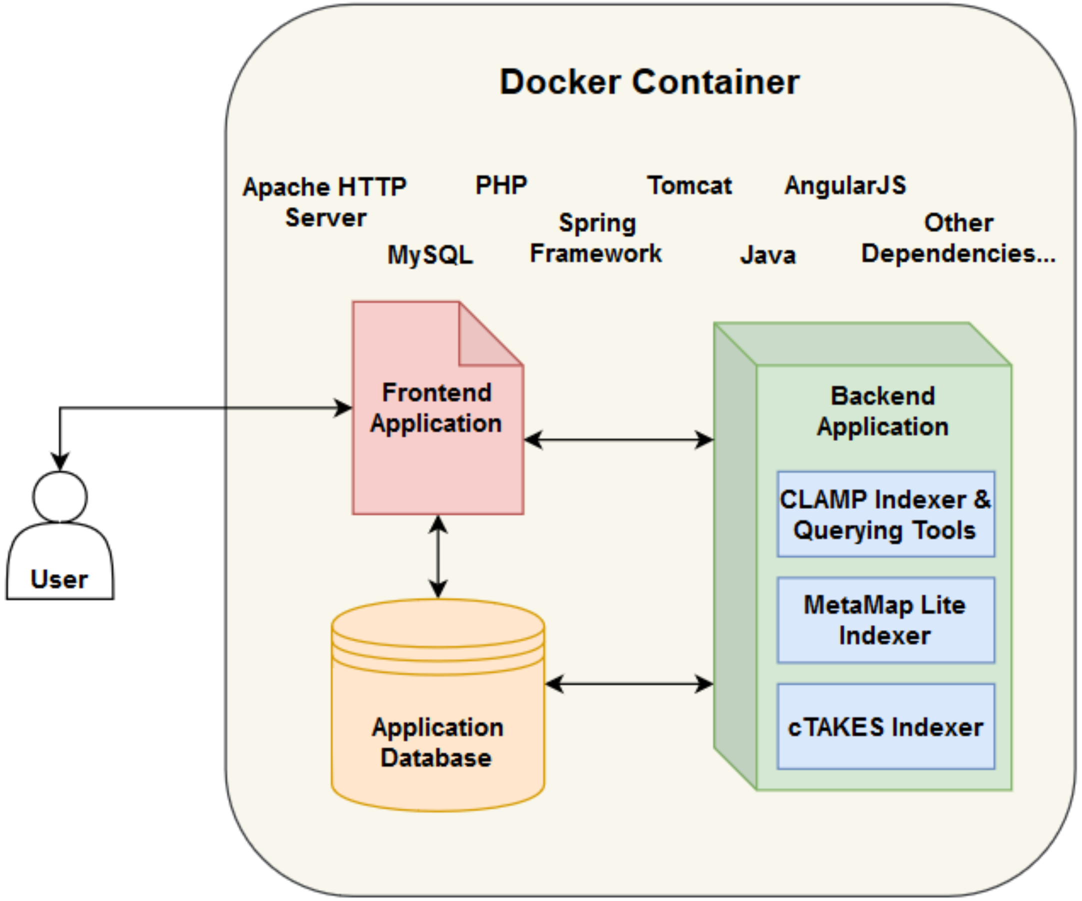
NLP Wrappers – Convert CLAMP / cTAKES / Metamap to Note_NLP

- https://github.com/OHDSI/NLPTools/tree/master/Wrappers



THEIA – A Web Application to Process and Visualize Textual Data

- Select own NLP tools (i.e., cTAKES, MetaMap, and CLAMP)
- Process selected clinical documents
- Convert different NLP systems' outputs into standard OMOP CDM tables
- Query and visualize their results
- Configurable access among multiple users



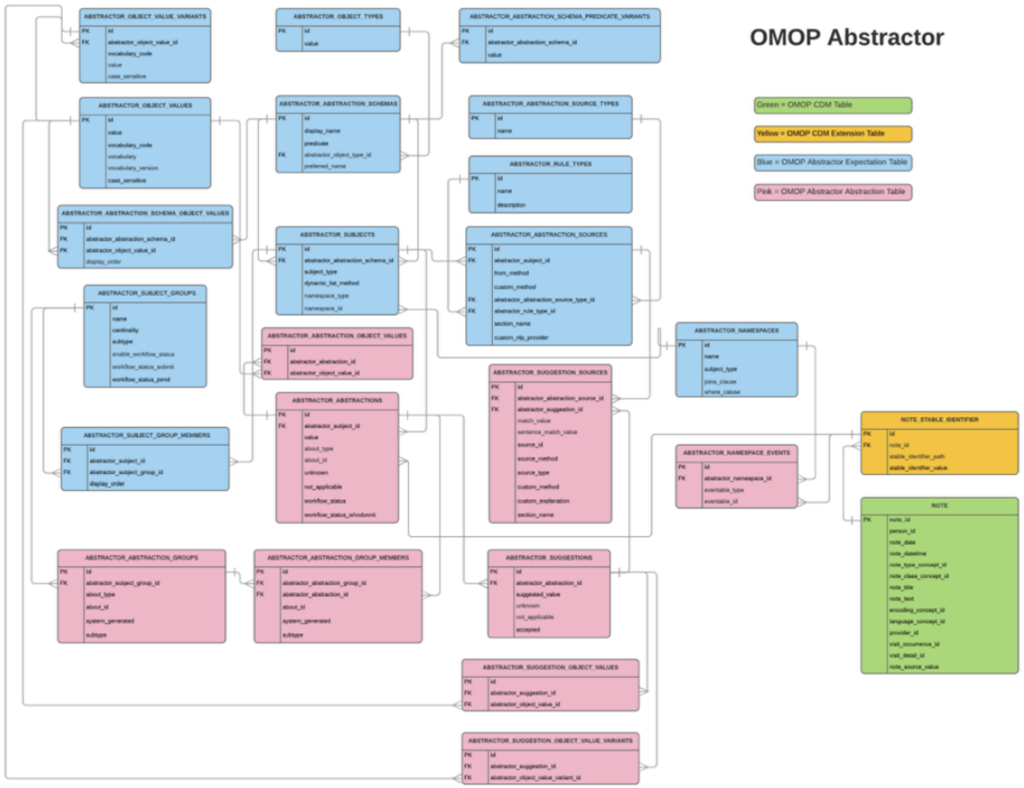
Ananke – Convert UMLS CUIs to OMOP Concept IDs

- Most NLP tools map concepts to UMLS Metathesaurus Concept Unique Identifiers (CUIs)
- UMLS Metathesaurus contains over three million concepts and over 130 English vocabularies
- OHDSI vocabulary on the other hand covers over 70 vocabularies with many of them overlapping

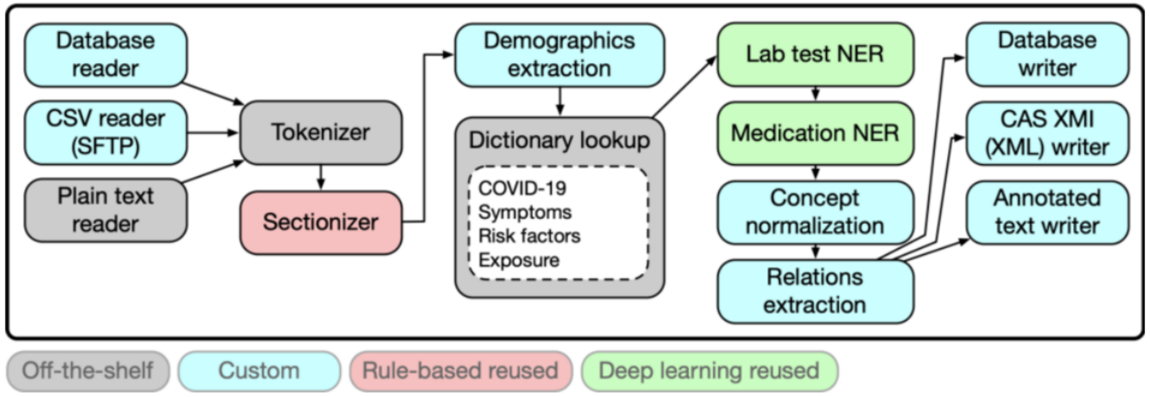
UMLS CUIs → OMOP Concept IDs

Other NLP tools

- OMOP Abstractor
  - NLP-aided assisted chart abstraction platform built upon the OMOP CDM



- DECOVRI (Data Extraction for COVID-19 Related Information)
  - Free and open source tool to convert unstructured notes into structured data within an OMOP CDM-based ecosystem
  - Built on the Apache UIMA framework



Heider PM, Pipaliya RM, Meystre SM. A Natural Language Processing Tool Offering Data Extraction for COVID-19 Related Information (DECOVRI). Stud Health Technol Inform. 2022 Jun 6;290:1062-1063

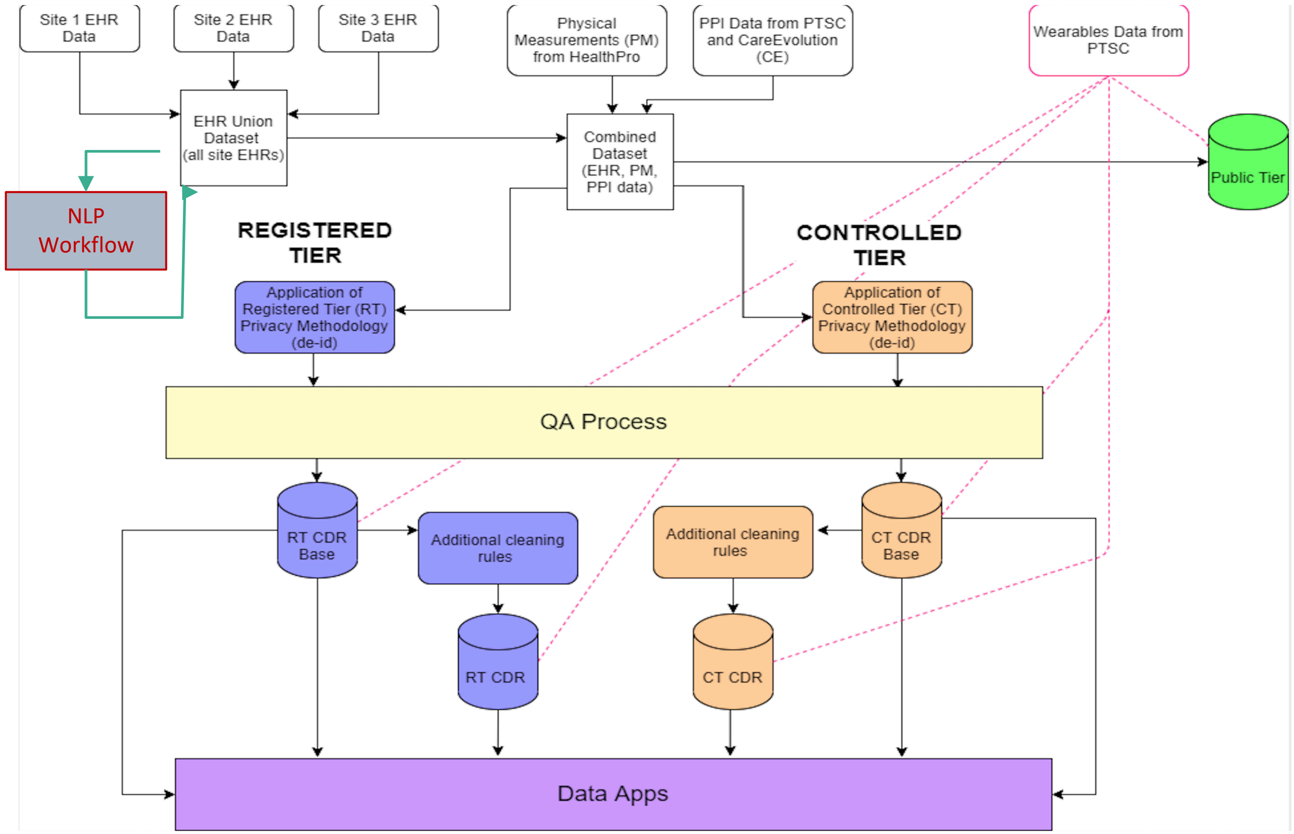
Links to OHDSI NLP Tools

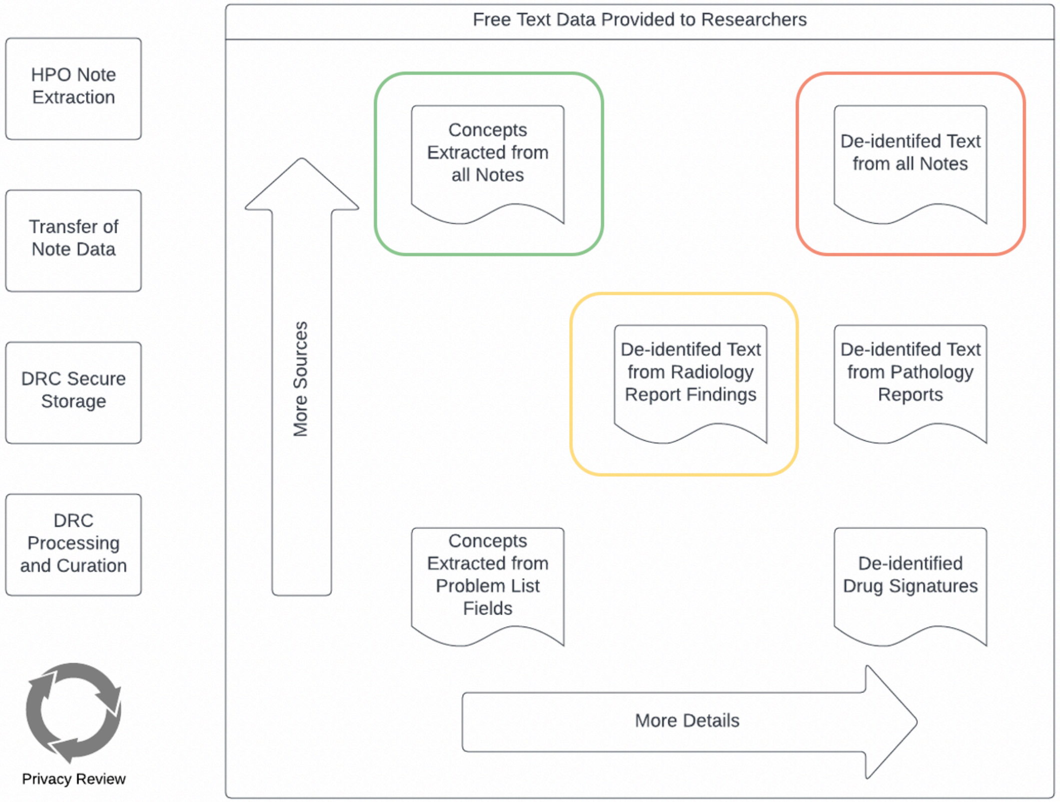
- All software tools are open source, most of them available at OHDSI NLP tools Github: https://github.com/OHDSI/NLPTools

- NLP Wrappers: https://github.com/OHDSI/NLPTools/tree/master/Wrappers
- Ananke: https://github.com/thepanacealab/OHDSIananke
- THEIA: https://github.com/OHDSI/NLPTools/tree/master/THEIA
- COVID-19 TestNorm: https://github.com/UTHealth-CCB/covid19_testnorm

All of Us (AoU) Research Program

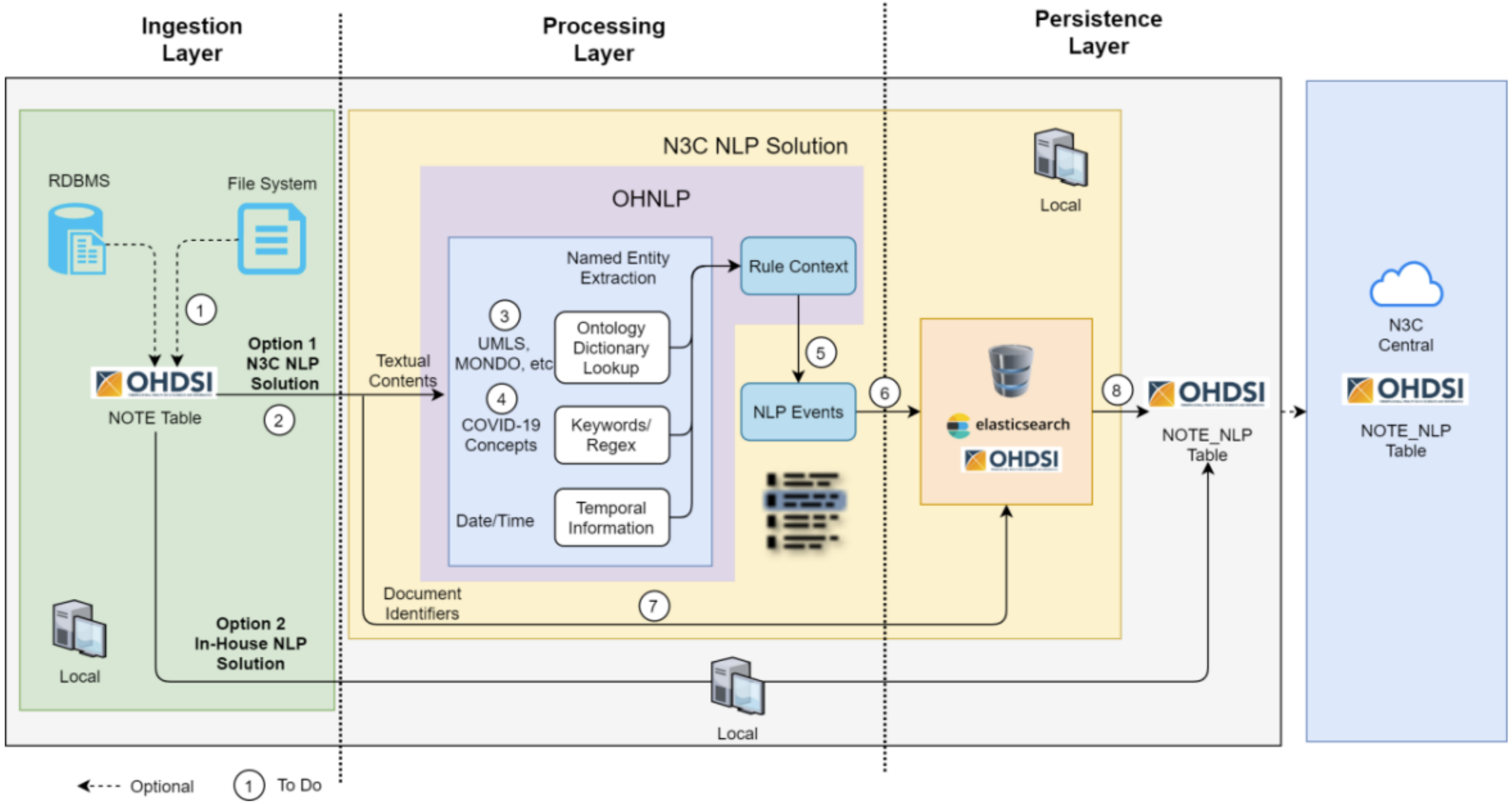
- AoU Data and Research Center is working on collecting and processing textual data from participating sites by following the OHDSI NLP workflow for textual data





The National COVID Cohort Collaborative (N3C) – NLP WG

- N3C NLP WG has populated signs and symptoms of COVID-19 into the NOTE_NLP tables using MedTagger and implemented and evaluated its performance across multiple participating sites



Liu S, et al. An Open Natural Language Processing Development Framework for EHR-based Clinical Research: A case demonstration using the National COVID Cohort Collaborative (N3C). arXiv preprint arXiv:211010780. 2021.

The Veterans Health Administration (VHA)

- The 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 LVEF NLP note findings and source notes were transformed and stored in Note and Note_NLP tables

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

FitzHenry F, Patterson OV, Denton J, Brannen J, Reeves RM, DuVall SL, et al., editors. OMOP CDM for Natural Language Processing: Piloting a VA NLP Data Set. OHDSI Conference; 2017.

Use Cases at Individual Healthcare Systems

Healthcare organization	NLP tools	Applications	Comments
University of Utah Health (1.5 million patients)	A generic rule-based NLP system, EasyCIE	Two NLP pipelines to identify and classify the venous thromboembolism (VTE) and pulmonary embolism (PE) patients	Does not maintain a full OMOP CDM. Instead, a view is created using a schema similar to the NOTE table and the NOTE_NLP table is used to save the snippet-level NLP output.
Columbia University Irving Medical Center (6.6 million patients)	Multiple locally trained tools including MedLEE, HealthTermFinder, and MedTagger for N3C.	Cohort identification, characterization studies, and predictive analytics tasks, for instance, eMERGE phenotypic algorithms, infectious disease surveillance	
Weill Cornell Medicine (3 million patients)	Radiology text analysis system, RadText	Information extraction tasks from radiology reports.	RadText supports a tool to convert from NOTE table and standardizes the output into NOTE_NLP
University of Minnesota M Health Fairview (4.5 million patients)	Locally trained NLP algorithms	COVID-19 sign/symptom extraction from clinical notes; and dietary supplements information extraction.	The COVID-19 related data in the NOTE_NLP table with corresponding CDM data is regularly contributed to the N3C.
UMass Memorial Health (3.2 million patients)	cTAKES	Suicide prediction models by extracting features (e.g., history of self-harm) from clinical notes.	Two OMOP CDM instances built to contribute data to the N3C and TrinetX network

Use Cases at Individual Healthcare Systems

Healthcare organization	NLP tools	Applications	Comments
University of Pittsburgh Medical Center (over 5.5 million outpatient visits every year)	Locally trained NLP algorithms	Extracting lifestyle-related Social Determinants of Health (SDoH) factors such as sleep-related concepts	The extracted SDoH factors could be stored in the NOTE_NLP table. However, due to the lack of standardized SDoH ontology and terminology, it is not trivial to be transferred to OMOP clinical tables.
Sydney Partnership for Health, Research, Education and Enterprise (includes data from multiple local health districts in New South Wales, Australia)	Luigi library, which supports multiple spaCy and Hugging Face models trained on local data	Study the prevalence and impact of variation in clinical cancer care	NOTE_NLP table used to store numerous classes of named entities extracted from clinical notes. Current targets include ECOG performance status, oral chemotherapy agents and smoking history, with the aim of expanding these targets over time.
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	The genomic common data model (G-CDM) [55], an extension of OMOP CDM, was used to map the extracted genetic variants.
Medical University of South Carolina (~1.5 million patients)	DECOVRI built on Apache UIMA; custom medspaCy pipelines	Data Extraction for COVID-19 symptom monitoring	ePhenotyping extractions in the NOTE_NLP table can be difficult for concepts without standard coded forms (e.g., SDoH, section header types).

Ongoing Studies at OHDSI NLP WG

- Post-acute sequelae of SARS-CoV-2 infection (PASC)
  - Led by UTHealth Houston, with 4 participating sites
  - characterize the incidence of PASC, or related symptoms and diagnoses, for COVID-19 patients
- A Delirium Study
  - Led by Mayo Clinic
  - Objective 1: ascertainment of delirium status using natural language processing from EHRs
  - Objective 2: assemble a delirium cohort for a multi-site observational study

01 Introduction to EHR, clinical notes, and NLP

02 NLP Working Group at OHDSI: CDM, Tools, Use Cases

03 Challenges and future work

Remaining Challenges and Future Work

Representations of concepts, modifiers, and more

Scalability of processing textual data using NLP

Efficient retrieval and visualization of textual/NLP data

Security and privacy concerns

Performance and traceability of NLP solutions

Integrating text with structured data for quick analysis

Join OHDSI NLP WG!

- Join us for our monthly meetings:

Second Wednesday of every month @ 2 PM – 3 PM ET

Microsoft Teams meeting (link on our wiki page below)

Wiki page: https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:nlp-wg
(or Google OHDSI NLP WG wiki)
GitHub repository: https://github.com/OHDSI/NLPTools

Hua.Xu@uth.tmc.edu

Upcoming – September 14

2022 OHDSI Symposium

October 14 - 16 Bethesda North Marriott Hotel & Conference Center

NLP WG meeting – Oct 15, 3PM – 5 PM

Acknowledgement

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

Thank you!

Questions?

hua.xu@uth.tmc.edu