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

Vipina K. Keloth[1], Juan M. Banda[2], Michael Gurley[3], Paul M. Heider[4], Georgina Kennedy[5], Hongfang Liu[6], Feifan Liu[7], Timothy Miller[8], Karthik Natarajan[9], Olga V Patterson[10,11,12], Yifan Peng[13], Ruth M. Reeves[14,15], Masoud Rouhizadeh[16,17], Jianlin Shi[10,11,18], Xiaoyan Wang[19], Yanshan Wang[20], Wei-Qi Wei[15], Andrew E. Williams[21], Rui Zhang[22], Rimma Belenkaya[23], Christian Reich[24], Clair Blacketer[25,26], Patrick Ryan[9,25], George Hripcsak[9], Noémie Elhadad[9,*], Hua Xu[1,*]

1. School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA
2. Department of Computer Science, Georgia State University, Atlanta, Georgia, USA
3. Applied Research Informatics Group, Northwestern University, Chicago, Illinois, USA
4. Biomedical Informatics Center, Medical University of South Carolina, Charleston, South Carolina, USA
5. Ingham Institute for Applied Medical Research, Sydney, Australia
6. Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, Minnesota, USA
7. Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, Massachusetts, USA
8. Computational Health Informatics Program, Boston Children's Hospital, and Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA
9. Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, New York, USA
10. VA Informatics and Computing Infrastructure, Department of Veterans Affairs Salt Lake City Health Care System, Salt Lake City, Utah, USA
11. Division of Epidemiology, Department of Internal Medicine, School of Medicine, University of Utah, Salt Lake City, Utah, USA
12. Verily Life Sciences, Mountain View, California, USA
13. Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA
14. TN Valley Healthcare System, U.S. Department of Veterans Affairs, Nashville, Tennessee, USA
15. Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA
16. Department of Pharmaceutical Outcomes & Policy, University of Florida, Gainesville, Florida, USA
17. Biomedical Informatics and Data Science, Johns Hopkins University, Baltimore, Maryland, USA
18. Department of Biomedical Informatics, University of Utah, Salt Lake City, USA
19. Sema4 Mount Sinai Genomics Incorporation, Stamford, Connecticut, USA
20. Department of Health Information Management, Department of Biomedical Informatics, and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
21. School of Medicine, Tufts University, Boston, Massachusetts, USA
22. Institute for Health Informatics, and Department of Pharmaceutical Care & Health Systems, University of Minnesota, Minneapolis, Minnesota, USA
23. Memorial Sloan Kettering Cancer Center, New York, New York, USA
24. Real World Solutions, IQVIA, Durham, North Carolina, USA
25. Janssen Pharmaceutical Research and Development LLC, Titusville, New Jersey, USA
26. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

*Corresponding authors:
hua.xu@uth.tmc.edu (Hua Xu)
noemie.elhadad@columbia.edu (Noémie Elhadad)

Abstract

Clinical documentation in electronic health records contains crucial narratives and details about patients and their care. Natural language processing (NLP) can unlock the information conveyed in clinical notes and reports, and thus plays a critical role in real-world studies. The NLP Working Group at the Observational Health Data Sciences and Informatics (OHDSI) consortium was established to develop methods and tools to promote the use of textual data and NLP in real-world observational studies. In this paper, we describe a framework for representing and utilizing textual data in real-world evidence generation, including representations of information from clinical text in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), the workflow and tools that were developed to extract, transform and load (ETL) data from clinical notes into tables in OMOP CDM, as well as current applications and specific use cases of the proposed OHDSI NLP solution at large consortia and individual institutions with English textual data. Challenges faced and lessons learned during the process are also discussed to provide valuable insights for researchers who are planning to implement NLP solutions in real-world studies.

Keywords: *Real-world study; Natural language processing; Electronic health records*

1. Introduction

The use of real-world data has gained increasing popularity in drug development, drug regulation, clinical trial feasibility, and observational research, especially in cases where clinical trials are too difficult or expensive to conduct [1-4]. The United States Food and Drug Administration (FDA) defines real-world data (RWD) as "the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources." [5] The sources of RWD include electronic health records (EHRs), claims and billing data, disease registries, and patient-generated health data like the ones from electronic devices (wearables), software applications (apps), and social media [6]. While EHRs, claims data, and patient registries are used widely for clinical evidence generation [7-9], the COVID-19 pandemic witnessed an increased use of data from wearables and social media for epidemiological studies [10-12]. As data from these diverse sources can provide new insights and evidence regarding the benefits and risks of medical products and services, RWD is being generated and used by multiple stakeholders such as pharmaceutical companies, researchers, payers, providers, patients, and regulatory agencies. For example, FDA's Real-World Evidence (RWE) program promotes shared learning and provides guidelines to assist developers interested in RWD to develop RWE for regulatory decisions [13].

One of the main challenges in conducting high-quality, reproducible real-world studies is achieving data standardization across different collaborative sites. To achieve that goal, extensive efforts have been

devoted to developing and maintaining common data models (CDM) for real-world clinical data by various initiatives. The Informatics for Integrating Biology and the Bedside (i2b2) [14] data model is one of the earliest models and follows the entity-attribute-value (EAV) approach. The schema has a central "fact" table with each row representing a single observation about a patient. The FDA leads the Sentinel System which coordinates the development of the Sentinel Common Data Model (SCDM) [15]. The SCDM v8.0.0 includes 16 tables with a major focus on using RWD to study medical product safety. The National Patient-Centered Clinical Research Network (PCORnet) CDM [16] is based on the FDA Mini-Sentinel CDM [17] thus leveraging existing analytic tools and expertise, and prioritizing analytic functionality in the CDM design. The Observational Health Data Sciences and Informatics (OHDSI) [18] community has invested tremendous effort in the development and maintenance of the Observational Medical Outcomes Partnership (OMOP) CDM and Standardized Vocabularies [19]. At present, OMOP CDM has been applied to over 300 sites, containing data on over 800 million unique patients, resulting in over 300 publications [20].

One of the features that sets OMOP CDM apart from other CDMs is the incorporation of representations and tools that can deal with clinical textual data. Clinical notes contain abundant information about patients' prior medical history, psychosocial and family history, disease course and progress, as well as information regarding the healthcare process (e.g., tests, treatments, and procedures). Clinical Natural Language Processing (NLP), while being an active area of research, has played a crucial role in extracting relevant patient information embedded in clinical narratives [21-23]. Several general clinical NLP tools, such as MedLEE [24], MetaMap/MetaMap Lite [25, 26], cTAKES [27], and CLAMP [28], 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 [29] was established in 2015 as part of the OHDSI consortium. The major focus includes defining representations of textual data, developing NLP and ETL (Extract, Transform and Load) tools, and facilitating real world studies incorporating evidence in clinical documents.

In this paper, we summarize the development of representations for storing clinical text and its NLP outputs in the OMOP CDM and its current applications to real-world studies with multiple use cases of English textual data. A summary of the ETL tools developed for aiding this process is also provided. Furthermore, we discuss the lessons learned during this process and future development plans.

2. Methods

The NLP framework of the OMOP CDM has been developed in close collaboration with the 'CDM and Vocabulary Development Working Group' (CDM WG) and the 'NLP Working Group' (NLP WG) with

active input from the OHDSI community. The CDM WG at OHDSI is responsible for the development, maintenance, and promotion of the OMOP CDM. To enable the 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 WG. NLP WG scheduled multiple meetings to gather opinions from team members regarding the tables to be added to the OMOP CDM. Based on information collected during the conference calls and the open discussions on OHDSI forums, a proposal was submitted to the CDM WG. This was followed by further discussions between CDM WG and NLP WG. The initial proposal was updated after every discussion. The process was iterated upon until the proposal's approval. Following the design principles of OMOP CDM (specified in the Book of OHDSI [30]), a lean standardized representation of two tables: NOTE and NOTE_NLP, was incorporated into the OMOP CDM. These tables store the clinical text and the output of clinical NLP tools (**Figure 1**).

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	A standard Concept Id representing the HL7 LOINC Document Type Vocabulary 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	If the Note record is related to another record in the database, this field is the primary key of the linked record.
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.

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.

Figure 1: Schema of NOTE (left) and NOTE_NLP (right) tables in OMOP CDM 5.4 [31]

The NLP tools can extract a wide variety of information including concepts (e.g., cardiovascular disease), values (e.g., body temperature, blood pressure), and modifiers (e.g., severity, certainty). To determine the fields in the NOTE and NOTE_NLP tables, our approach was aligned with the principles of OMOP CDM, focusing on fields that demonstrated their usefulness in real-world applications. Extensive search and review of use cases from the literature, use cases from networks such as eMERGE [32], and use cases

recorded by several OHDSI working groups for various projects were collected. Trade-offs were made to decide the fields to be included (see Discussion).

2.1 Representing clinical text data in OMOP CDM

The two NOTE and NOTE_NLP tables store the note information and output from clinical NLP tools respectively. 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. Each row represents a clinical concept and includes information on the concept (i.e., name and concept id), its commonly used modifiers (e.g., existence, temporal), and other relevant contextual information such as the section it belongs to, a snippet of surrounding words, and offsets in the note. Such context information could be useful for NLP-based manual chart review, (e.g., to highlight extracted information) as well as for retraining NLP systems. Data provenance information like the name and version of the NLP system used for processing the clinical text is also included. As NLP systems may output different types and values for modifiers, a field called "*term_modifiers*" was proposed to store a list of modifier names and values.

2.2 Workflow and tools for generating NOTE_NLP table

As shown in **Figure 2**, the ETL workflow for textual data in the CDM consists of the following steps: (1) executing the NLP systems to process the textual notes in the NOTE table. Users can use their preferred NLP systems such as cTAKES, MetaMap Lite, or CLAMP for this task; 2) converting NLP system output into the NOTE_NLP table. As different NLP systems have different output formats and use different coding terminologies, we developed additional tools to facilitate the conversion (see sections below); and (3) transferring concepts from NOTE_NLP to individual clinical tables in CDM. SQL scripts were developed for this purpose. More details of the developed tools are described below.

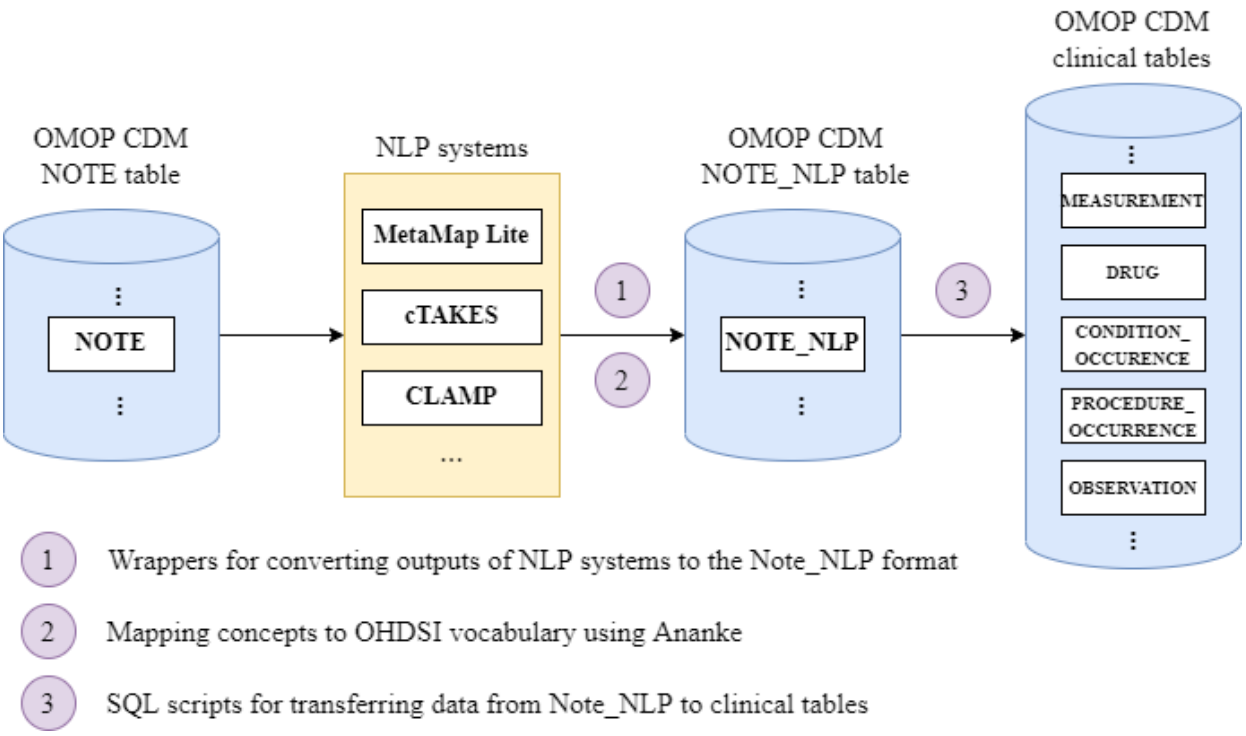

Figure 2: An overview of the workflow for transforming clinical text in the NOTE table

Wrappers for converting outputs of NLP systems to the NOTE_NLP format: To reduce the complexity of transforming the output of clinical NLP tools into structured fields, wrappers were implemented in Java to support concept extraction [33]. Currently, wrappers are available for cTAKES, MetaMap Lite, and CLAMP [33]. The wrappers take the clinical text files as input and output the extracted concepts along with text snippets, offset, the corresponding UMLS Concept Unique Identifiers (CUIs), date, and other temporal and existential modifiers as .txt or .xmi files.

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. To bridge this gap, Ananke [34] was built as a tool that provides direct mappings between UMLS CUIs and OHDSI concept identifiers. It leverages the OHDSI vocabulary and the UMLS synonyms and concept relationships. As a single mapping file, this can be plugged in at the time of moving NLP output into the NOTE_NLP table to standardize extracted concepts.

SQL scripts for transferring data from NOTE_NLP to clinical tables: The terms extracted and normalized from the clinical text data may correspond to conditions, procedures, measurements, etc. The NOTE_NLP table acts as an intermediate storage for these concepts. We developed SQL scripts to transfer

the data from the NOTE_NLP table to the corresponding clinical event tables such as CONDITION_OCCURRENCE, PROCEDURE_OCCURRENCE, MEASUREMENT, DRUG_EXPOSURE, and OBSERVATION. All developed tools have been made publicly available at the OHDSI NLP WG GitHub repository [35].

3. Implementation Status and Lessons Learned

3.1 Implementation Status

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

The All of Us Research Program (AoU): The AoU [36] 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. The Data and Research Center at Vanderbilt University Medical Center is following the OMOP CDM to standardize clinical data from EHRs, including textual documents. The ETL workflow developed by the OHDSI NLP WG has been implemented and tested using a small collection of synthetic textual data and the CLAMP NLP tool. A more 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 [37] is a collaborative initiative to collect COVID-19 clinical data to answer critical research questions related to the pandemic. A Data Enclave was developed to serve as a secure platform used by contributing members to store clinical data. An NLP workgroup at N3C has been formed to promote the use of textual documents for COVID-19 research. They have developed an ETL process to populate signs and symptoms of COVID-19 into the NOTE_NLP tables using an example NLP engine MedTagger [38], and implemented and evaluated its performance across multiple participating sites [39].

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. The Corporate Data Warehouse (CDW) is a national data repository comprising data from several VHA clinical and administrative systems. 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 [40]. The data gets updated regularly and ongoing effort is dedicated to maintaining and expanding the dataset with additional data

sources within VHA. 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 [41].

Table 1 shows a list of individual healthcare systems which have adopted the OHDSI NLP solution, together with their use cases.

Table 1. Summary of sites that adopted the OHDSI NLP solution.

Healthcare organization	**NLP tools used**	**Applications and use cases**
University of Utah Health (1.5 million patients)	A generic rule-based NLP system, EasyCIE [42]	Two NLP pipelines to identify and classify venous thromboembolism (VTE) and pulmonary embolism (PE) patients [43]
Columbia University Irving Medical Center (6.6 million patients)	Multiple locally trained tools including MedLEE [24], HealthTermFinder [44], and MedTagger [45] for N3C.	Cohort identification, characterization studies, and predictive analytics tasks, for instance, eMERGE phenotypic algorithms [46], infectious disease surveillance [47]
Weill Cornell Medicine (3 million patients)	An open-source radiology text analysis system, RadText [48]	Information extraction tasks from radiology reports.
University of Minnesota M Health Fairview (4.5 million patients)	Locally trained NLP algorithms [49, 50]	Two applications: (1) COVID-19 sign/symptom extraction from clinical notes; and (2) dietary supplements (DS) information extraction.
UMass Memorial Health (the largest health care system in Central Massachusetts, serving 3.2 million patients)	cTAKES [27]	A pilot study to develop suicide prediction models by extracting features (e.g., history of self-harm) from clinical notes.

University of Pittsburgh Medical Center (the largest health care system in west Pennsylvania, with over 5.5 million outpatient visits every year)	Locally trained NLP algorithms [51]	Extracting lifestyle-related factors such as sleep-related concepts
Sydney Partnership for Health, Research, Education and Enterprise (includes data from multiple local health districts in New South Wales, Australia)	A broad data engineering system based on the Luigi library [52], which supports multiple spaCy and Hugging Face [53, 54] models trained on local data.	Study the prevalence and impact of variation in clinical cancer care.
Sema4 - Mount Sinai Genomics Inc. (the Centrellis health ecosystem including multiple health systems serving >10 million patients)	Locally developed NLP pipelines based on CLAMP [28]	Five NLP pipelines for extracting genetic variants, protein biomarkers, family medical history, diseases, and procedures.
Medical University of South Carolina (the primary Research Data Warehouse includes ~1.5 million patients)	DECOVRI [55] built on Apache UIMA [56]; custom medspaCy pipelines [57, 58]	Data Extraction for COVID-19 Related Information (specifically, symptom monitoring); custom pipelines used by the NLP Core, the unstructured data request service center for research at MUSC

3.2 Lessons Learned

Based on discussions with sites that have implemented the OHDSI NLP solution, several common issues have been identified and we summarize the lessons learned as follows:

Gaps in standardization of concepts extracted by NLP: As clinical documents often contain deeper and broader information about patients, customized NLP systems may extract concepts that are not included in the current OMOP standard vocabularies. Several sites reported this issue, including the University of

Pittsburgh Medical Center - UPMC (i.e., sleep-related concepts), Sydney Partnership for Health, Research, Education and Enterprise (i.e., cancer-related information), Sema4 (i.e., genetic variants), and Medical University of South Carolina - MUSC (i.e., social determinants of health). Potential solutions to address this challenge often rely on two approaches: (1) leveraging existing extensions to OMOP CDM, e.g., OMOP oncology extension for cancer information and the genomic common data model (G-CDM) for genetic variants [59]; and (2) developing custom CDM extensions locally (e.g., the sleep project at UPMC).

Challenges regarding the use of NLP systems: As shown in Table 1, different NLP systems have been used by different healthcare systems to support diverse applications, which makes it difficult to develop a unified NLP solution based on a single NLP tool. The main reason is that off-the-shelf clinical NLP systems often cannot achieve optimal performance on local data, often requiring customization by a local team. Furthermore, as reported by several sites (e.g., Columbia University Irving Medical Center – CUIMC and MUSC), sometimes multiple NLP systems are deployed at one site, to support different applications. According to CUIMC, as each NLP system outputs large amounts of data into the NOTE_NLP table, it adds a data storage burden to their infrastructure. MUSC also expressed the challenge of comparing extractions between NLP systems/versions, as the current design of the NLP workflow does not differentiate incremental vs. repeated processing of textual data, which we plan to address in our future work.

Implementation issues to meet local application needs: All sites agree that full implementation of the NLP solution to include all textual documents in an EHR is a complex task, requiring significant amounts of resources and effort. A practical solution is to start with pilot projects that are focused on specific applications (e.g., disease-specific studies in Table 1) to assess the feasibility and required resources in a local setting, before rolling out for full implementation. Some sites also reported customized solutions to meet their local needs. For example, the University of Utah Health 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 and is used to save the snippet-level NLP output. Furthermore, they use two additional tables to save document-level and patient-level NLP output, so that the identified patients with corresponding evidence in notes can be easily searched using join queries.

4. Discussion

Following the proposed OMOP CDM representations for clinical textual data, researchers have actively worked on this topic. In addition to the efforts described in the Results section, a few studies were published that utilized the NOTE and NOTE_NLP tables along with the ETL tools for transforming textual data for

use in observational studies [60-63]. For example, one study described the experiences of transforming the notes from MIMIC into OMOP CDM and evaluated the difficulty and analyzed the benefits of this transformation [60]. We are aware that requirements for textual data in real-world research can change over time, which requires us to update the NOTE_NLP table accordingly by taking into consideration the feedback from multi-institutional network studies and other research projects. A process for updating the OMOP CDM NLP framework is in place, by interactively engaging the community, the CDM WG, and the NLP WG. For example, a proposal from the NLP WG to address minor issues such as adding polymorphic foreign keys to the NOTE_NLP table to link it to the clinical event tables [19], which is proposed by end users, is under consideration by the CDM WG.

During the development and refinement of the OHDSI NLP framework for clinical textual data, several issues have been identified and actively discussed by researchers in this area. We summarize them as the following.

4.1 Representations of clinical concepts, modifiers, and more

The need for a lean standardized solution to store NLP output is still an ongoing challenge. The NOTE_NLP table provides an initial version of a standardized representation for clinical concepts extracted by NLP systems, following OHDSI vocabularies. One of the main challenges while developing the first version of the NOTE_NLP table was to strike a balance between the complexity of NLP outputs and the simplicity of the OMOP CDM. NLP systems could generate diverse types of information at different linguistic levels (e.g., syntactic and semantic information), all of which could be useful for downstream text analysis. However, the OMOP CDM WG requires that each field in the table should demonstrate its utility in clinical studies (e.g., frequently used in different studies). Therefore, it has been an iterative process to define the fields in the NOTE_NLP. Each field was extensively discussed and decided based on its importance for NLP and its utility for clinical studies (e.g., use cases in the eMERGE network).

Take modifiers associated with clinical concepts as an example, two potential representations were discussed: (1) store all modifiers in a new table, with linkage to concepts in the NOTE_NLP table; and (2) store a few selected modifiers that have demonstrated clinical utility as columns in the NOTE_NLP table. After discussion, only two modifiers: "*term_exists*" (i.e., to indicate whether a patient has a condition, but is more complex than the simple affirmed/negated dichotomy) and "*term_temporal*" (i.e., presenting a present vs. past condition) are specified in the NOTE_NLP table, while all other modifiers (e.g., values of lab tests and signature information of medications) are concatenated and represented by a single field ("*term_modifiers*"). The decision was made due to several reasons: (1) an additional modifier table will increase the complexity of OMOP CDM, but its utility has not been fully demonstrated yet; (2) modifiers

and their value sets generated by different NLP systems are often diverse and hence it is challenging to define a commonly accepted standard for all modifiers; and (3) "*term_exists*" and "*term_temporal*" have been used in multiple phenotyping tasks in eMERGE studies; and (4) storing all other modifiers in one field makes it simple; but it still allows users to develop more use cases for additional modifiers, to inform additional changes in future. For example, in an effort to map LVEF extraction output to OMOP CDM 5.2 format, researchers revealed that a certain level of data loss has to be accepted when the dataset is large and multiple modifiers need to be recorded [41]. While utilization of this field is simple for data storage and transmission, it may become a considerable bottleneck when the processed dataset grows in size (see 4.2 Efficiency of retrieving NLP derived data). These findings provide valuable evidence for representing modifiers in the next version of NOTE_NLP and other related tables.

Several other representation challenges are also discussed at OHDSI NLP WG meetings. For example, standardizing note types and sections within clinical notes remains challenging. One ongoing project at OHDSI NLP WG is to standardize note types based on the LOINC Document Ontology [64, 65]. More recently, NLP artifacts from novel deep learning architectures such as word embeddings have also been brought up by researchers, and how to represent such more NLP focused data (i.e., to support NLP method development) is also under discussion.

4.2 Efficiency of retrieving NLP derived data

While the current solution (and proposed improvements) allows for data provenance and meta-data storage, the NOTE_NLP table in institutions with millions of patients often surpasses several billion rows. This makes any join operations on any highly indexed and shared RDBMS environments quite computationally expensive. On the storage side, having redundant fields leads to high storage costs, particularly for institutions that rely on cloud platforms as their datastore. It is vital to evaluate the trade-offs between performance, storage, and usability to be able to properly accommodate NLP derived elements into the OMOP CDM. Our current approach moves NLP-extracted concepts to their respective clinical domain tables (CONDITION_OCCURRENCE, OBSERVATION, PROCEDURE_OCCURRENCE, etc.) so that other OHDSI tools (e.g., ATLAS) can be easily applied to analyze textual data without major modifications to the tool's codebase. However, the NOTE and NOTE_NLP tables can be directly applied to support clinical studies, e.g., researchers may conduct a keyword search to identify patients of interest or review patients' charts. In such a scenario, it is critical to efficiently search through massive textual documents and their NLP outputs. Filtering rows by key/value pairs in the "*term_modifiers*" column, which would potentially represent a substring of the column, is computationally expensive for traditional SQL-style databases or expensive in terms of storage because of the additional indexing required on the column. Therefore, schemaless databases such as MongoDB and Elasticsearch have been discussed as potential

solutions. These architectures naturally support faceted field queries, which would mitigate many of the difficulties in representing an open class of valid term_modifiers with an open class set of modifier values. Nevertheless, it will require a substantial amount of effort to implement such scalable solutions for big data.

4.3 Tools to support processing and analyzing textual data

Although tools have been developed to facilitate clinical text processing, challenges still exist when users apply such tools to their local data and specific use cases. As shown in Table 1, diverse NLP systems have been used at different institutions for text processing, which increases the complexity of developing common ETL tools that can work with different NLP systems. Our current strategy is to accommodate general purpose, widely used clinical NLP systems such as cTAKES, MetaMap, and CLAMP first when developing ETL tools. The next step is to engage the NLP community to implement an OHDSI NLP output component within each individual NLP system. For example, MedTagger and DECOVRI can now generate outputs that are in compliance with the NOTE_NLP table [39, 55]. To further streamline the NLP ETL process, we have also worked on a toolkit called THEIA [66], which supports indexing and visualization of clinical text stored in OMOP CDM tables using NLP tools such as CLAMP, MetaMap Lite, and cTAKES. Moreover, tools to facilitate the use of NLP outputs have also been developed. For example, Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE) [67] is an R package built for the OHDSI community that implements two methodologies to build probabilistic phenotype definitions: XPRESS [68] and Anchor Learning [69]. It is one of the few OHDSI tools that support using data found in the NOTE_NLP table to identify cohorts of patients and build machine learning models using term mentions in the NOTE_NLP table as features. It has been shown that using NLP-extracted terms as features in probabilistic phenotype models improves accuracy, recall, and positive predictive value by around 10% on average [67], making the utility of having them available in the OMOP CDM crucial for phenotyping efforts. In the future, we expect more tools will be developed to make it more convenient for end users to leverage textual data and NLP for real world studies.

4.4 Other issues for using NLP in real-world studies

There are limitations when applying NLP to real-world evidence generation. First, none of the existing NLP systems extracts information from clinical documents without errors. There are concerns about how such errors from NLP may impact the downstream analyses, which is understandable. However, in general, combining NLP results with coded data often provides a more accurate status of patients, e.g., phenotyping [70]. This is why unstructured data and NLP are mentioned in the FDA guidance for using real-world data to support drug regulatory decision making [71]. Nevertheless, it is very important to carefully evaluate the

workflow of textual data processing, as well as the quality of derived data from the workflow, to ensure the reproducibility of clinical research using NLP. Currently, most studies would include an evaluation of the performance of the NLP system used (e.g., DECOVRI [55]). However, few studies have reported the quality of workflows and data derived from the workflows. Fortunately, researchers have started exploring this important issue, e.g., Digan et al. [72] proposed an approach to document the reproducibility of clinical NLP frameworks using workflow management systems. The OHDSI NLP WG would highly recommend continuing such efforts on quality assessments of textual data, workflows, as well as user experience.

It is also important that we can trace back to the original sources when evidence generated by NLP solutions is questioned and reviewed. Many of the clinical data tables containing coded data provide an explicit column for indicating provenance, which includes "NLP" (with concept code OMOP4976931) as one of the standardized "Accepted Concepts". For instance, the PROCEDURE_OCCURRENCE table includes the column "*procedure_type_concept_id*" to "determine the provenance of the Procedure record, as in whether the procedure was from an EHR system, insurance claim, registry, or other sources".

Another consideration of using clinical documents is privacy and security concerns, as textual data may contain sensitive patient identifier information. For example, the "snippet" field in the NOTE_NLP table records text around a concept, which may contain sensitive information. One potential solution is to run de-identification programs to remove protected health information [73] before extracting snippets. Of course, none of the de-identification programs achieve 100% sensitivity and/or accuracy. In extreme cases, institutions can decide not to populate "snippet", as it is an optional field in NOTE_NLP. For multi-site studies, the OHDSI consortium currently takes a distributed approach, where each participating site shares analysis results from their local datasets, instead of patient-level data. However, the OMOP CDM has been used to standardize data in centralized repositories as well, e.g., N3C and AoU, which often rely on traditional approaches (e.g., multi-site IRB) to ensure privacy protection. In the setting of distributed data networks, federated learning that allows data analysis across sites without sharing individual patient data has received great attention and has shown promise in multi-site studies [74].

5. Conclusion

In summary, text documents 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 advancing 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.

Acknowledgment

Dr. Hua Xu and Dr. Hongfang Liu were supported in part by NCATS 1U01TR002062. Dr. Yifan Peng was supported in part by the National Library of Medicine under Award No. 4R00LM013001. Dr. Rui Zhang was supported in part by NCCIH R01AT009457 and NCATS UL1TR002494. Dr. Paul M. Heider was partially supported through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-2018C3-14549) and the SmartState Endowment for Translational Biomedical Informatics. Dr. Juan M. Banda was supported in part by a grant from the National Institute on Aging (3P30AG059307-02S1). Dr. George Hripcsak was supported by the National Library of Medicine award R01 LM006910.

We would like to thank all the members of the OHDSI community, especially the NLP and CDM working groups for participating in the discussions and providing valuable feedback.

Conflict of Interest

Dr. Hua Xu and The University of Texas Health Science Center at Houston have research related financial interests at Melax Technologies Inc. Dr. Xiaoyan Wang has related financial interests at Sema4 Mount Sinai Genomics Inc.

References

1. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. Jama. 2018;320(9):867-8.
2. Baumfeld Andre E, Reynolds R, Caubel P, Azoulay L, Dreyer NA. Trial designs using real-world data: the changing landscape of the regulatory approval process. Pharmacoepidemiology and Drug Safety. 2020;29(10):1201-12.
3. Skovlund E, Leufkens H, Smyth J. The use of real-world data in cancer drug development. European Journal of Cancer. 2018;101:69-76.
4. Trojano M, Tintore M, Montalban X, Hillert J, Kalincik T, Iaffaldano P, et al. Treatment decisions in multiple sclerosis—insights from real-world observational studies. Nature Reviews Neurology. 2017;13(2):105-18.
5. U.S. Food and Drug Administration - Real-World Evidence [cited 2022 Jan 30]. Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.
6. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence—what is it and what can it tell us? New England Journal of Medicine. 2016;375(23):2293-7.
7. Patorno E, Gopalakrishnan C, Franklin JM, Brodovicz KG, Masso-Gonzalez E, Bartels DB, et al. Claims-based studies of oral glucose-lowering medications can achieve balance in critical clinical variables only observed in electronic health records. Diabetes, Obesity and Metabolism. 2018;20(4):974-84.
8. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. Journal of the American Medical Informatics Association. 2013;20(e2):e226-e31.

9. Khozin S, Blumenthal GM, Pazdur R. Real-world data for clinical evidence generation in oncology. JNCI: Journal of the National Cancer Institute. 2017;109(11):djx187.
10. Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 social media infodemic. Scientific reports. 2020;10(1):1-10.
11. Ates HC, Yetisen AK, Güder F, Dincer C. Wearable devices for the detection of COVID-19. Nature Electronics. 2021;4(1):13-4.
12. Jeon J, Baruah G, Sarabadani S, Palanica A. Identification of risk factors and symptoms of COVID-19: Analysis of biomedical literature and social media data. Journal of medical Internet research. 2020;22(10):e20509.
13. U.S. Food and Drug Administration - Framework for FDA's Real-World Evidence Program [cited 2022 Jan 30]. Available from: https://www.fda.gov/media/120060/download.
14. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. Journal of the American Medical Informatics Association. 2016;23(5):909-15.
15. Sentinel Common Data Model [cited 2022 Jan 30]. Available from: https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model.
16. Toh S, Rasmussen-Torvik LJ, Harmata EE, Pardee R, Saizan R, Malanga E, et al. The National Patient-Centered Clinical Research Network (PCORnet) bariatric study cohort: rationale, methods, and baseline characteristics. JMIR research protocols. 2017;6(12):e8323.
17. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The US Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiology and drug safety. 2012;21:1-8.
18. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics. 2015;216:574.
19. OMOP Common Data Model [cited 2022 Jan 30]. Available from: https://ohdsi.github.io/CommonDataModel/.
20. Sachson C. Our Journey: Where the OHDSI Community Has Been, and Where We Are Going. 2021 [cited 2022 Jan 30]. Available from: https://www.ohdsi.org/wp-content/uploads/2021/09/OHDSI-OurJourney2021-Final.pdf.
21. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. Journal of biomedical informatics. 2018;88:11-9.
22. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. Journal of the American Medical Informatics Association. 2019;26(11):1297-304.
23. Savova GK, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer research. 2019;79(21):5463-70.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.

29. OHDSI Natural Language Processing Working Group [cited 2022 Jan 31]. Available from: https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:nlp-wg.
30. Observational Health Data Sciences and Informatics: The Book of OHDSI [cited 2022 Sep 13]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/.
31. OMOP CDM 5.4 [cited 2022 May 19]. Available from: http://ohdsi.github.io/CommonDataModel/cdm54.html.
32. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics. 2011;4(1):1-11.
33. OHDSI NLP tools - Wrappers [cited 2022 Jan 31]. Available from: https://github.com/OHDSI/NLPTools/tree/master/Wrappers.
34. OHDSI Ananke - A Tool for Mapping Between OHDSI Concept Identifiers to Unified Medical Language System (UMLS) identifiers [cited 2022 Jan 31]. Available from: https://github.com/thepanacealab/OHDSIananke.
35. OHDSI NLP tools repository [cited 2022 Jan 31]. Available from: https://github.com/OHDSI/NLPTools.
36. 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.
37. 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.
38. 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.
39. Liu S, Wen A, Wang L, He H, Fu S, Miller R, 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.
40. Lynch KE, Deppen SA, DuVall SL, Viernes B, Cao A, Park D, et al. Incrementally transforming electronic medical records into the observational medical outcomes partnership common data model: a multidimensional quality assurance approach. Applied clinical informatics. 2019;10(05):794-803.
41. FitzHenry F, Patterson OV, Denton J, Brannen J, Reeves RM, DuVall SL, et al. OMOP CDM for Natural Language Processing: Piloting a VA NLP Data Set. OHDSI Conference; 2017.
42. Shi J, Mowery D, Zhang M, Sanders J, Chapman W, Gawron L. Extracting intrauterine device usage from clinical texts using natural language processing. 2017 IEEE International Conference on Healthcare Informatics (ICHI); 2017: IEEE.
43. Johnson SA, Signor EA, Lappe KL, Shi J, Jenkins SL, Wikstrom SW, et al. A comparison of natural language processing to ICD-10 codes for identification and characterization of pulmonary embolism. Thrombosis Research. 2021;203:190-5.
44. Hirsch JS, Tanenbaum JS, Lipsky Gorman S, Liu C, Schmitz E, Hashorva D, et al. HARVEST, a longitudinal patient record summarizer. Journal of the American Medical Informatics Association. 2015;22(2):263-74.
45. MedTagger [cited 2022 May 13]. Available from: https://github.com/OHNLP/MedTagger.
46. Shang N, Liu C, Rasmussen LV, Ta CN, Caroll RJ, Benoit B, et al. Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network. Journal of biomedical informatics. 2019;99:103293.
47. Zachariah P, Hill-Ricciuti A, Saiman L, Natarajan K. Using the "Who, What, and When" of free text documentation to improve hospital infectious disease surveillance. American Journal of Infection Control. 2020;48(10):1261-3.
48. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings. 2018;2018:188.

49. Fan Y, Zhang R. Using natural language processing methods to classify use status of dietary supplements in clinical notes. BMC medical informatics and decision making. 2018;18(2):15-22.
50. Fan Y, Zhou S, Li Y, Zhang R. Deep learning approaches for extracting adverse events and indications of dietary supplements from clinical text. Journal of the American Medical Informatics Association. 2021;28(3):569-77.
51. Mohammad HA, Sivarajkumar S, Viggiano S, Oniani D, Visweswaran S, Wang Y. Extraction of Sleep Information from Clinical Notes of Alzheimer's Disease Patients Using Natural Language Processing. medRxiv. 2022.
52. Luigi [cited 2022 May 10]. Available from: https://github.com/spotify/luigi.
53. Honnibal M, Johnson M. An improved non-monotonic transition system for dependency parsing. Proceedings of the 2015 conference on empirical methods in natural language processing; 2015.
54. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations; 2020.
55. Heider PM, Pipaliya RM, Meystre SM. A Natural Language Processing Tool Offering Data Extraction for COVID-19 Related Information (DECOVRI). MEDINFO 2021: The 18th World Congress on Medical and Health Informatics; 2021.
56. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering. 2004;10(3-4):327-48.
57. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. arXiv preprint arXiv:210607799. 2021.
58. Off-the-Shelf Post-hoc Ensemble Generation Algorithms [cited 2022 May 13]. Available from: https://github.com/MUSC-TBIC/ots-ensemble-systems.
59. Genomic-CDM [cited 2022 May 19]. Available from: https://github.com/OHDSI/Genomic-CDM.
60. Paris N, Parrot A. MIMIC in the OMOP Common Data Model. medRxiv. 2020.
61. Ryu B, Yoon E, Kim S, Lee S, Baek H, Yi S, et al. Transformation of pathology reports into the common data model with oncology module: use case for colon cancer. Journal of medical Internet research. 2020;22(12):e18526.
62. Sharma H, Mao C, Zhang Y, Vatani H, Yao L, Zhong Y, et al. Developing a portable natural language processing based phenotyping system. BMC Medical Informatics and Decision Making. 2019;19(3):79-87.
63. Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj D, et al. A new paradigm for accelerating clinical data science at Stanford Medicine. arXiv preprint arXiv:200310534. 2020.
64. Zuo X, Li J, Zhao B, Zhou Y, Dong X, Duke J, et al. Normalizing Clinical Document Titles to LOINC Document Ontology: An Initial Study. AMIA Annual Symposium Proceedings; 2020: American Medical Informatics Association.
65. LOINC Document Ontology [cited 2022 May 13]. Available from: https://loinc.org/document-ontology/.
66. THEIA [cited 2022 May 19]. Available from: https://github.com/OHDSI/NLPTools/tree/master/THEIA.
67. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Summits on Translational Science Proceedings. 2017;2017:48.
68. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. Journal of the American Medical Informatics Association. 2016;23(6):1166-73.
69. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. Journal of the American Medical Informatics Association. 2016;23(4):731-40.
70. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. IEEE/ACM transactions on computational biology and bioinformatics. 2018;16(1):139-53.

71. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products [cited 2022 May 12]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory.
72. Digan W, Névéol A, Neuraz A, Wack M, Baudoin D, Burgun A, et al. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. Journal of the American Medical Informatics Association. 2021;28(3):504-15.
73. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. Journal of biomedical informatics. 2015;58:S11-S9.
74. Luo C, Islam M, Sheils NE, Buresh J, Reps J, Schuemie MJ, et al. DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. Nature communications. 2022;13(1):1-10.