





Short communication

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

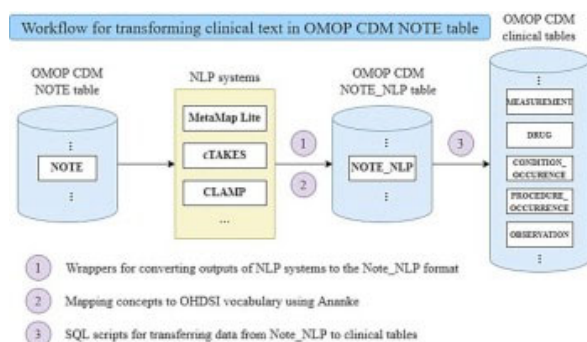
[Vipina K. Keloth](#)^a, [Juan M. Banda](#)^b, [Michael Gurley](#)^c, [Paul M. Heider](#)^d, [Georgina Kennedy](#)^e, [Hongfang Liu](#)^f, [Feifan Liu](#)^g, [Timothy Miller](#)^h, [Karthik Natarajan](#)ⁱ, [Olga V Patterson](#)^{j k l}, [Yifan Peng](#)^m, [Kalpana Raja](#)^a, [Ruth M. Reeves](#)^{n o}, [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  

[Show more](#)  Share  Cite<https://doi.org/10.1016/j.jbi.2023.104343> [Get rights and content](#) 

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.

Graphical abstract



[Download : Download high-res image \(78KB\)](#)

[Download : Download full-size image](#)

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], [2], [3], [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], [8], [9], the COVID-19 pandemic witnessed an increased use of data from wearables and social media for epidemiological studies [10], [11], [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 through 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], [22], [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.

Section snippets

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....

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 [35] 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...

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 Section 3, 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], [61], [62], [63]. For example, one study described the experiences in transforming the notes from MIMIC into OMOP CDM and evaluated the difficulty and analyzed ...

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 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....

Declaration of Competing 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....

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...

[Recommended articles](#)

References (77)

E. Skovlund *et al.*

[The use of real-world data in cancer drug development](#)

Eur. J. Cancer (2018)

S. Velupillai *et al.*

[Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances](#)

J. Biomed. Inform. (2018)

S.A. Johnson *et al.*

[A comparison of natural language processing to ICD-10 codes for identification and characterization of pulmonary embolism](#)

Thromb. Res. (2021)

N. Shang *et al.*

[Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network](#)

J. Biomed. Inform. (2019)

P. Zachariah *et al.*

[Using the “Who, What, and When” of free text documentation to improve hospital infectious disease surveillance](#)

Am. J. Infect. Control (2020)

Y. Huang *et al.*

[ELII: A novel inverted index for fast temporal query, with application to a large Covid-19 EHR dataset](#)

J. Biomed. Inform. (2021)

A. Stubbs *et al.*

[Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1](#)

J. Biomed. Inform. (2015)

J. Corrigan-Curay *et al.*

Real-world evidence and real-world data for evaluating drug safety and effectiveness

JAMA (2018)

E. Baumfeld Andre *et al.*

Trial designs using real-world data: the changing landscape of the regulatory approval process

Pharmacoepidemiol. Drug Saf. (2020)

M. Trojano *et al.*

Treatment decisions in multiple sclerosis—insights from real-world observational studies

Nat. Rev. Neurol. (2017)



View more references

Cited by (0)

[View full text](#)

© 2023 Elsevier Inc. All rights reserved.



All content on this site: Copyright © 2023 Elsevier B.V., its licensors, and contributors. All rights are reserved, including those for text and data mining, AI training, and similar technologies. For all open access content, the Creative Commons licensing terms apply.

