Harmonization of OMOP Drug and Device source concepts using ChatGPT-4o

David Davila-Garcia[1,2], Adam Wilcox, PhD[2]
1 Columbia University Department of Biomedical Informatics, 2 Washington University in St. Louis School of Medicine

Background

The adoption of the OMOP Common Data Model (CDM) has greatly facilitated the standardization and interoperability of healthcare data across institutions (1,2). However, when it is implemented, a significant proportion of source concepts from individual hospital EHR systems can remain unmapped to standardized OMOP concepts, hindering the efficient exchange of this data for research and clinical purposes (3). Corrections typically require manual mapping of concepts which can be tedious and expensive for large datasets. Large language models (LLMs), such as OpenAI's ChatGPT, have demonstrated impressive performance in various natural language tasks, including those related to biomedical literature (4,5). This study seeks to evaluate the feasibility of ChatGPT-4o, OpenAI's flagship foundation model, to retrieve the most relevant concept from a list of k=5 candidate standardized concepts across the drug and device domains, thereby automating the process of mapping free-text source concept descriptions to OMOP standardized concepts.

Methods

The Washington University in St. Louis OMOP EHR instance contains records for over 2.3 million patients across 14 hospitals within the BJC HealthCare system, encompassing over 100 million drug and device exposures between 2019-2024. Unmapped drug and device exposure source concepts were identified (2.9% of all records), and those without valid target concept IDs were identified using the source_to_concept_map table (25.4% of unmapped records). The source code descriptions from the top 20 most frequent unmapped concepts (15.0% of unmapped records) and a random sample of 20 unmapped concepts (0.1% of unmapped records) were obtained. Free text source code descriptions were searched on ATHENA to retrieve the top 5 unique OMOP candidate concepts. Subsequently, ChatGPT-4o with was prompted to identify the single most relevant concept from the top 5 candidates identified by ATHENA. The model temperature and seed were set to zero to improve reproducibility of predictions. Model inputs included the original free-text source code description, in addition to the concept name, validity, concept type, and domain for the top 5 candidates. Token log probabilities were obtained to measure model confidence in the output. ChatGPT prediction accuracy, area under the receiver operator characteristics curve (AUROC), and area under the precision-recall curve (AUPRC) were calculated using manually annotated concept maps performed by DD as a reference standard. The overall project workflow is shown in Figure 1.

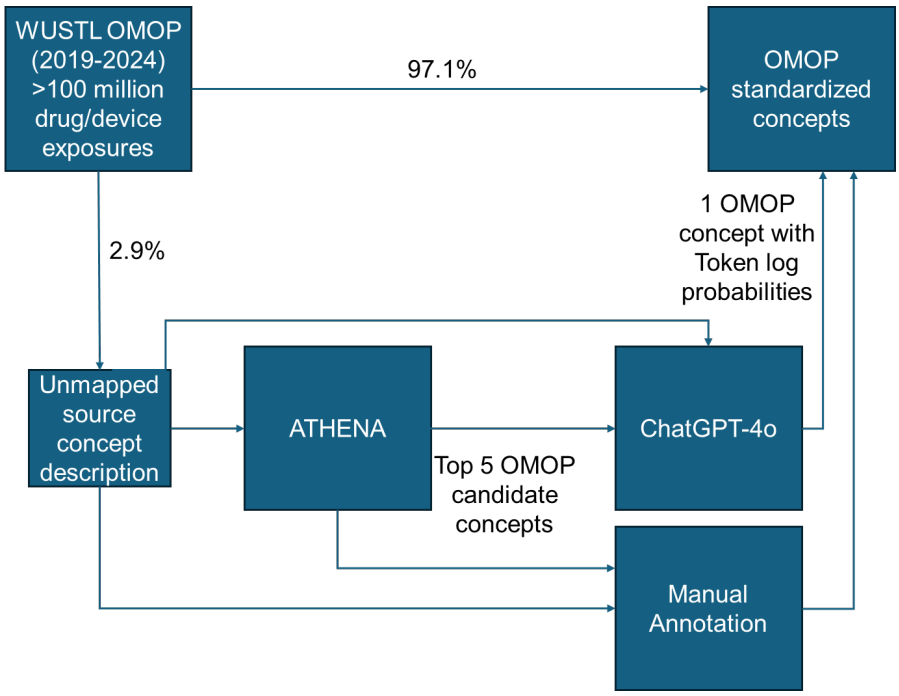

Figure 1. Project workflow for converting unmapped source concepts to OMOP standardized concepts using the ATHENA tool and the ChatGPT-4o language model, with manual annotation serving as a reference standard for performance evaluation.

Results

ChatGPT-4o demonstrated promising performance in harmonizing unmapped source concepts within the Drug and Device domains (Table 1). Overall accuracy reached 68%, with an AUROC of 0.6 and AUPRC of 0.75. Performance was comparable between the Drug and Device domains. Notably, the model exhibited higher AUPRC for the top-20 most frequent unmapped source concepts (0.92) compared to the random sample of 20 concepts (0.64), suggesting enhanced performance for more commonly encountered concepts (Table 2).

Table 1: Performance Metrics by Concept Domain

Domain	# Concepts	Accuracy	AUROC	AUPRC
Drug	22	0.68	0.6	0.73
Device	18	0.67	0.7	0.82
Overall	40	0.68	0.6	0.75

Table 2: Performance Metrics by Group

	% of Unmapped Concepts	Accuracy	AUROC	AUPRC
Top-20 Concepts	15.0	0.65	0.82	0.92
Random 20 Concepts	0.1	0.7	0.42	0.64

Conclusion

This study highlights the potential of state-of-the-art LLMs, such as ChatGPT-4o, to automate data

harmonization of unmapped source concepts to standardized OMOP concepts. The model's strong performance, particularly for frequently encountered concepts, suggests its utility in streamlining the mapping process from institutional source vocabularies to the OMOP CDM Standardized Vocabulary, thus improving the interoperability of healthcare data. Further research is necessary to assess the feasibility of employing LLMs to automate data harmonization across other concept domains. The integration of LLMs into the OMOP CDM harmonization pipeline shows promise in enhancing the efficiency and accuracy of source to concept mapping, ultimately promoting the broader adoption of standardized healthcare data for research and clinical applications.

References

1. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc JAMIA [Internet]. 2012 [cited 2024 Jun 21];19(1):54–60. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3240764/

2. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, et al. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. J Am Med Inform Assoc JAMIA [Internet]. 2024 Jan 4 [cited 2024 Jun 21];31(3):583–90. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10873827/

3. Yoon D, Han C, Kim DW, Kim S, Bae S, Ryu JA, et al. Redefining Health Care Data Interoperability: Empirical Exploration of Large Language Models in Information Exchange. J Med Internet Res [Internet]. 2024 May 31 [cited 2024 Jun 20];26(1):e56614. Available from: https://www.jmir.org/2024/1/e56614

4. Bhagat N, Mackey O, Wilcox A. Large Language Models for Efficient Medical Information Extraction. AMIA Summits Transl Sci Proc [Internet]. 2024 May 31 [cited 2024 Jun 20];2024:509–14. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11141860/

5. Wang A, Liu C, Yang J, Weng C. Fine-tuning large language models for rare disease concept normalization. J Am Med Inform Assoc [Internet]. 2024 Jun 3 [cited 2024 Jun 20];ocae133. Available from: https://doi.org/10.1093/jamia/ocae133