Improving VSAC to OMOP Mapping Using LLM Assisted Curation

Robert B Barrett[a], Star Liu[a], Kyle Zollo-Venecek[b], Benjamin Riesser[c], Benjamin Martin[a]

[a]Biomedical Informatics and Data Science, Johns Hopkins University, Baltimore, MD, USA

[b]CTSI, Tufts University School of Medicine, Boston, MA

[c]Improving Health Outcomes, American Medical Association, Greenville, SC

Background

The National Library of Medicine (NLM) Value Set Authority Center (VSAC) provides a crowd-sourced repository of standardized representations of clinical concepts, commonly used in quality measures, decision support tools, and for interoperability. Yet, VSAC value sets are not without error and have demonstrated significant variation in quality measure outcomes when addressed[3]. These errors often manifest in the form of missing, extraneous, or misconstrued codes. Moreover, cross-terminology evaluations have compared ICD-10-CM to SNOMED CT mappings, showing divergence in several domains[2]. These issues assume critical importance when used in standardized data models like the Observational Medical Outcomes Partnership (OMOP) Common Data Model, where concept mappings between vocabularies play a pivotal role.

The Observational Health Data Sciences and Informatics (OHDSI) community reviews and maintains relationships between clinical concepts, mapping concepts from various source vocabularies to a central "standard" concept. The designated standard concepts and defined mappings between existing vocabularies is maintained as a bi-annually updated public resource: the OHDSI Standardized Vocabularies.[6] The default deterministic route for deriving these standard concepts, walking "Maps to" links in OMOP's *concept_relationship* table, matches source concepts to the corresponding OMOP standard concept, but it does not guarantee semantic fidelity. These semantic relationships may involve the concept's clinical meaning, relationship with the value set, and the greater, often nuanced, intention of the tool it is used in.

Recent studies demonstrated that large language models (LLMs) can detect or correct such mapping errors with high validity. GPT-4-based methods achieved a 96% valid-term match rate when screening local terminology against SNOMED CT[4], and an additional study using LLMs for data-exchange tasks highlighted their utility for harmonizing structured clinical data[5]. These findings suggest that an LLM can serve as an intelligent filter, flagging overly broad, overly specific, or misaligned OMOP concepts that deterministic rules alone cannot detect. This study used OpenAI's GPT-4o to map 6 commonly used VSAC value sets and evaluated the overall fitness to value sets' descriptions and the precision of concept mappings.

Methods

To translate VSAC value sets into a validated set of OMOP concepts, we executed a three-phase pipeline.

1. Value set identifiers were resolved using the NLM VSAC REST API. This process expanded each value set into its full list of constituent codes, descriptions, and source vocabularies (e.g., SNOMED CT, ICD-10, CPT).

2. The expanded source codes were cross walked to standard concepts within the OHDSI Standardized Vocabularies, using OHDSI Vocabulary v5.0 31-AUG-23. This produced a list of OMOP concept IDs for each value set, which served as the input for the final validation stage.
3. We performed LLM-assisted semantic filtering to ensure the lexical and contextual accuracy of the mappings, with relation to the mapped concepts and to the value set intentions. For this, we prompted OpenAI's GPT-4o with the value set's narrative description (including its clinical focus and inclusion/exclusion criteria), value set concept details, and OMOP concept details.

The filtering LLM agent was instructed to classify the mappings into two categories:

Concept-to-concept relationship:

- Equivalent: The concepts are clinically identical. Minor wording differences or updates in coding definitions are allowed if the core meaning is preserved.
- Narrower: The OMOP concept is a more specific, stricter subtype of the VSAC concept.
- Broader: The OMOP concept is more general than the VSAC concept. –
- Unrelated: The OMOP and VSAC concepts are clinically different.

Relevance to value set:

- In scope: The mapping is consistent with the Value Set's purpose.
- Out of scope: The mapping is not related to the Value Set's purpose.

Four informaticists independently annotated 283 VSAC-to-OMOP mappings, classifying concept relationships and relevance. Ground truth was derived via majority voting. Performance was evaluated using precision, recall, F1-score, accuracy, and Cohen's κ to compare GPT-4o classifications against the established human consensus.

Results

GPT-4o was assessed on 283 mappings across 6 VSAC value sets, covering cardiovascular, endocrine, renal, and administrative domains. For concept relationship classification, GPT-4o achieved high performance (F1-score = 0.968, Cohen's κ = 0.918), with precision of 0.972 and recall of 0.965 (**TP**: 273, **FP**: 8, **TN**: N/A, **FN**: 10). Value set relevance performance was lower (F1-score = 0.500, κ = 0.493), primarily due to the rarity of unrelated (n=4) concepts (**TP**: 277, **TN**: 2, **FP**: 2, **FN**: 2).

Table 1. Percent Agreement for Concept Mapping and Relevance Judgments Between Human Raters and LLM Across VSAC Value Sets

Concept Set	***H vs LLM Concept Rel.***	***H vs LLM Value Set Rel.***	***Inter-Human Concept Rel.***	***Inter-Human Value Set Rel.***
Diabetes	91.1%	95.7%	93.1%	96.6%
Beta Blocker	100.0%	100.0%	100.0%	100.0%
Dialysis Services	100.0%	98.0%	100.0%	100.0%
Hypertension	98.5%	100.0%	97.1%	100.0%
Kidney Transplant	95.0%	100.0%	96.7%	100.0%
Office Visit	85.7%	100.0%	71.4%	100.0%

Overall (mean)	**95.2%**	**99.0%**	**93.0%**	**99.4%**

Table 2. Summary of LLM–Human Expert Disagreement Cases on OMOP Concept Mappings

Value Set	***VSAC Concept***	***OMOP Concept***	***Disagreement Type***	***LLM Perspective***
Dialysis Services	Hemoperfusion (eg, with activated charcoal or resin)	Hemoperfusion	Technical vs Clinical Scope	Exclude (out-of-scope): "not a dialysis service"
Diabetes	Type 1 diabetes mellitus with periodontal disease	Periodontal disease	Composite-Component Granularity	Narrower relationship; exclude: missing diabetes qualifier
Diabetes	Other specified diabetes mellitus with periodontal disease	Periodontal disease	Composite-Component Granularity	Narrower relationship; exclude: missing diabetes qualifier
Diabetes	Diabetes mellitus in mother complicating pregnancy	Diabetes mellitus in mother complicating pregnancy	Exclusion Criteria Interpretation	Exclude despite equivalence (85% confidence)
Diabetes	Type 1 diabetes mellitus with diabetic nephropathy	Renal disorder due to type 1 diabetes mellitus	Semantic Hierarchy	Equivalent (LLM) vs Broader (Humans)
Diabetes	Uncontrolled type 1 diabetes mellitus	Type 1 diabetes mellitus	Semantic Specificity	Broader (LLM) vs Equivalent (Humans)

Conclusion

Implementing a consensus-driven annotation process alongside an LLM-assisted semantic filtering pipeline enabled robust validation of VSAC-to-OMOP concept mappings. The model demonstrated excellent accuracy for concept relationships (F1 = 0.968, κ 0.918), with reduced performance for value set relevance classifications due to class imbalance. Analysis of disagreement cases highlighted key challenges - including technical versus clinical scope, semantic granularity, and interpretation of exclusion criteria - underscoring the need for clearer value set definitions and tailored LLM prompting strategies. Moreover, the results demonstrated the challenges in human agreement for value set relevance, and in relationships between clinical concepts.

Limitations of this study include the use of OHDSI Vocabulary v5.0 31-AUG-23, excluding expansions, revisions, and corrections observed in more current versions. With the purpose of this study being to evaluate accuracy of mappings and relevance to value set intention, this limitation does not impact the evaluation. Additionally, this evaluation was conducted by a group of informaticists – where clinical review and judgement is likely required for accurate determination of relevance and accuracy of concept relationships. Further, we will refine the strategies used, specifically to used current Vocabulary versions, and include clinician evaluation across broader concept reviews. Finally, it is of interest to explore value set variation and the impact on measure development and result generation for electronic clinical quality measures (eCQMs) results.

Acknowledgments

This research was supported by the NLM, part of the NIH, under award number T15LM013979]

References

1. DuVall SL, Parker CG, Shields AR, et al. Toward Real-World Reproducibility: Verifying Value Sets for Clinical Research. *Stud Health Technol Inform*. 2024;310:164-168.
2. Gold S, Batch A, McClure R, et al. Clinical Concept Value Sets and Interoperability in Health Data Analytics. *AMIA Annu Symp Proc*. 2018;2018:480-489.
3. Zahn LA, Ahmad H, Sittig DF, et al. The Fault in Our Sets: A Mixed Methods Analysis of Clinical Value Set Errors. *medRxiv*. Preprint posted 2025-02-27; doi:10.1101/2025.02.27.25323054.
4. Huh S. Comparative Analysis of ChatGPT-4 for Automated Mapping of Local Medical Terminologies to SNOMED CT. *Stud Health Technol Inform*. 2025;327:813-817.
5. Yoon D, Han C, Kim DW, et al. Redefining Health Care Data Interoperability: Empirical Exploration of Large Language Models in Information Exchange. *J Med Internet Res*. 2024;26:e56614.
6. Reich C, Ostropolets A, Ryan P, et al. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. Journal of the American Medical Informatics Association. 2024;31(3):583-590. doi:10.1093/jamia/ocad247