

LLM-Assisted Code Conversion for OMOP Implementation of a Pregnancy Algorithm: Similarity-Based Quality Control and Pairing Logic

Michael Ray¹, Carrie Nielson¹, Ram Varma¹, Lauren Liu¹, Nirali Kotowsky¹, Jesse Najarro Cermeño¹, Sonia Hernandez-Diaz², Marc Suchard³

1. Gilead Sciences, Inc.
2. Harvard T.H. Chan School of Public Health
3. UCLA Fielding School of Public Health

Background

Accurate identification of pregnancy episodes in observational healthcare data is foundational for credible real-world evidence (RWE) in maternal safety research. Medical insurance claims data enable large-scale pregnancy ascertainment, but limited clinical detail often complicates characterization of pregnancy timing and pregnancy-related health outcomes. Established claims-based pregnancy algorithms—such as the validated approach described by Sumner et al.¹—support scalable pregnancy episode identification and timing (trimester, gestational age) ascertainment in native coding systems by combining multiple codes and hierarchical logic. However, adapting such algorithms to the OMOP Common Data Model (CDM) introduces challenges, as mapping nonstandard source codes (e.g., ICD, CPT) to OMOP standard concepts can result in loss of clinically meaningful pregnancy phenotype intent and temporal (trimester) intent. To support federated OMOP implementations, scalable methods are needed to (1) quantify mapping similarity, (2) prioritize mappings for human review, and (3) explicitly detect when preservation of intent requires combining multiple OMOP concepts (e.g., condition + timing).

Methods

The large language model (LLM)-assisted mapping and adjudication workflow is summarized in Figure 1. We first converted pregnancy-related code lists from Sumner et al.'s published algorithm (ICD-9-CM, ICD-9 procedures, ICD-10-CM, ICD-10-PCS, CPT-4) to OMOP standard concepts using OMOP vocabulary “maps to” relationships. For each source-to-standard mapping, we then computed term frequency-inverse document frequency (TF-IDF) cosine similarity. TF-IDF cosine similarity represents each concept description as a weighted term vector, assigning less weight to common terms and more weight to distinctive terms, and uses cosine similarity to quantify overlap between the two vectors (source vs target descriptions) as a signal for mapping quality. This similarity score was used to (i) identify exact description matches (cosine = 1.0) that can be accepted without additional (LLM or human) adjudication and (ii) stratify non-exact mappings into quartiles of similarity for subsequent human review. For codes with non-exact mappings (cosine < 1), we then applied an LLM-assisted adjudication step to classify

whether each mapping preserved pregnancy phenotype intent and temporal intent (trimester), and whether preserving intent required pairing multiple OMOP concepts (e.g., a condition concept paired with a trimester concept).

The adjudication output produced structured fields (pregnancy phenotype preserved [Yes/No], trimester preserved [Yes/No], and pairing required [Yes/No]) and, when trimester timing was lost during mapping, suggested standardized OMOP trimester concepts for pairing—4239938 (First trimester pregnancy), 4244438 (Second trimester pregnancy), and 4218813 (Third trimester pregnancy). When pairing was deemed to be required by the LLM, the workflow flagged the mapping for human reviewer input, including specification of combination logic (e.g., AND/OR) and an appropriate claims-based timing window (e.g., same day, $\pm N$ days, or within pregnancy episode). For example, trimester-specific source codes may map to separate OMOP concepts such that a condition concept must be paired with a trimester concept to recover temporal intent (e.g., O46.91 “*Antepartum hemorrhage, first trimester*” mapped to an OMOP hemorrhage concept + 4239938 “*First trimester pregnancy*”). A mapping was also marked “flagged for review” when the LLM indicated either a potential loss of pregnancy phenotype or trimester intent.

Results

The original algorithm utilizes 5,366 distinct source codes, including 3,132 ICD-10-CM diagnoses, 1,552 ICD-9-CM diagnoses, 416 ICD-10-PCS procedures, 85 ICD-9 procedures, and 181 CPT codes. Of these, 793 codes (15%) had cosine similarity = 1.0 (exact description match to an OMOP concept name) and were directly accepted into the OMOP-standard code list. The remaining 4,573 non-exact codes mapped to 2,176 unique OMOP standard concepts. Table 1 summarizes mapping structure by code type (1:1, 1:M, M:1), highlighting heterogeneous mapping and multiplicity across vocabularies. Similarity score distributions by code type are shown in Figure 2a, with procedure code mappings (CPT and ICD-10-PCS) demonstrating a higher concentration of high-similarity matches compared to diagnosis codes, which exhibit broader distributions and more frequent low-similarity mappings. These differences support the use of similarity scoring as a practical triage signal for prioritizing human review.

Stratification of mappings by similarity quartile (Figure 2b) demonstrates that LLM-adjudicated preservation of pregnancy outcome intent increases with similarity, whereas preservation of trimester intent does not follow the same pattern. Trimester preservation peaks at intermediate similarity levels and declines in higher-similarity strata, reflecting cases in which source codes encoding both condition and gestational timing are mapped to OMOP concepts that capture only the condition component. These findings illustrate that high lexical similarity does not guarantee preservation of temporal intent, motivating the need for explicit pairing logic in OMOP implementations. In 40% of mappings, the LLM-assisted adjudication step identified cases where a single source code encoded both condition and gestational timing but mapped to separate OMOP concepts, necessitating explicit pairing logic (e.g., condition concept + trimester concept) to preserve trimester intent.

Conclusions

Relatively few diagnosis source codes in this previously validated pregnancy algorithm had exact matches to OMOP concepts; however, an LLM-assisted workflow allowed for large-scale determination of whether the intention was preserved. Human expert review of the LLM-based suggestions for OMOP concept selection and logical combination is ongoing and will be presented at the symposium.

These results support a reproducible workflow for triaging code-mapping results for expert clinical review. Because LLM outputs can vary across model versions, documenting model configuration and versioning is important for reproducibility of LLM-assisted adjudication workflows. Collectively, this approach provides a scalable blueprint for intent-preserving translation of clinical phenotype code sets to OMOP-based pregnancy research. The workflow may be extended and evaluated in other algorithms requiring temporal and logical combinations of OMOP concepts to preserve the intent of validated claims-based algorithms. These findings demonstrate that while similarity scoring provides an efficient and reproducible mechanism for triaging mapping quality, preservation of temporal intent requires additional structural logic beyond string-based similarity, reinforcing the importance of LLM-assisted adjudication and concept pairing strategies in OMOP-based phenotype implementation.

Reference

1. Sumner, K. M., Ehlinger, A., Georgiou, M. E., & Wurst, K. E. (2021). Development and evaluation of standardized pregnancy identification and trimester distribution algorithms in US IBM MarketScan® Commercial and Medicaid data. *Birth Defects Research*, 113(19), 1357-1367.

Figure 1. LLM-assisted mapping and adjudication framework. The figure below shows the addition of steps 2 through 4 to the OMOP mapping process.

Input ICD-9/10 and CPT source codes are from the published pregnancy algorithm (Sumner et al.), with each code's algorithm utility specifically defined (e.g., code defines a pregnancy phenotype category and/or code defines a specific trimester/pregnancy timing).
1. OMOP Mapping Map each source code to OMOP standard concept(s) using vocabulary "maps to" relationships (concept_relationship table stored in OMOP vocabulary).
2. Similarity Scoring Compute TF-IDF cosine similarity between the source description and the mapped OMOP concept name (text2vec package in R); stratify non-exact (cosine < 1.0) matches into similarity quartiles (Q1–Q4) to prioritize review.
3. LLM Adjudication (Microsoft Copilot Agent) Generate structured labels for each mapping: phenotype preserved (Y/N), trimester preserved (Y/N), pairing required (Y/N); suggest trimester concept IDs when trimester timing is not retained in the mapped concept.
4. Pairing Logic (if needed) If pairing is required, reviewers specify combination logic (AND/OR) and a claims-appropriate timing window (same day, $\pm N$ days, or within pregnancy episode) for condition + trimester concepts.
Output Outputs include similarity score/quartile, LLM adjudication results, suggested additions (e.g., trimester concepts), and a "flagged for review" indicator for triage.
5. Targeted Review (future step) Human review focuses on flagged and/or low-similarity mappings; final mapping decisions and pairing rules are documented for reproducibility; subsequently evaluate human vs LLM adjudication concordance by similarity quartile.

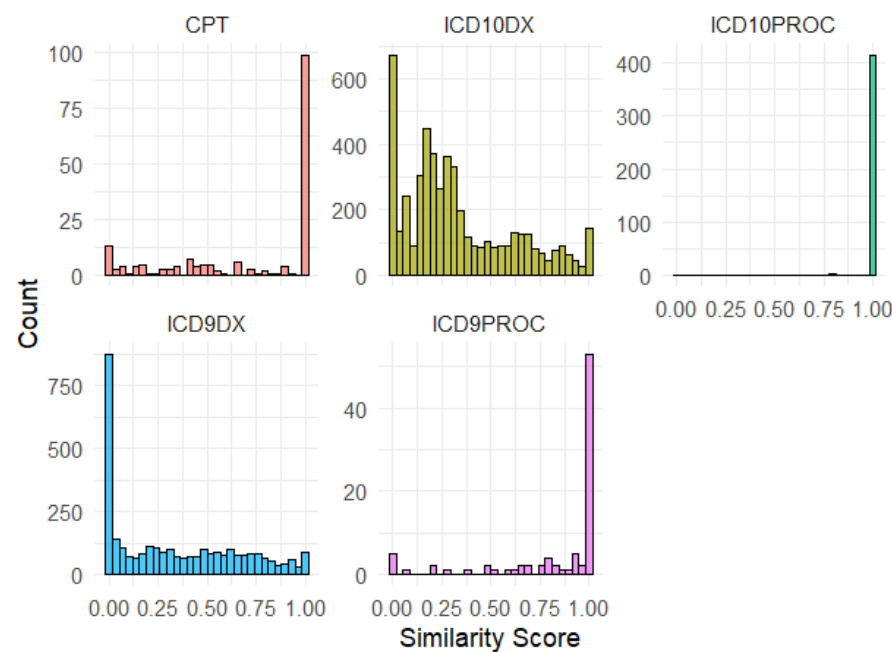
Table 1. Mapping structure by code type for pregnancy algorithm code conversion to OMOP. One-to-many mappings include both (i) semantic expansion where a source code maps to multiple OMOP concepts representing distinct facets of meaning and (ii) decomposition of condition + timing into separate OMOP concepts, where intent is recovered only when concepts are paired (e.g., condition + trimester concept).

Code type	Unique source codes (n)	Unique target concepts (n)	Mapping type (n)		
			1 to 1	1 to many	Many to 1
CPT	181	175	167	6	11
ICD10DX	3132	749	340	3654	1114
ICD10PROC	416	416	416	0	0
ICD9DX	1927	751	207	2131	807
ICD9PROC	86	85	84	0	2

Figure 2a. Distribution of TF-IDF cosine similarity scores by code type. Lower similarity scores indicate mappings more likely to involve semantic mismatch or decomposition of condition and timing

Figure 2b. LLM-adjudicated preservation of outcome and trimester intent by similarity score quartile. Outcome preservation increases with similarity, whereas trimester preservation declines at higher similarity, indicating loss of temporal information in mapping.

2a. Distribution of Similarity Scores by Code Type



2b. LLM-Adjudicated Intent Preservation by Similarity Score Quartile

