

Mapping Colombian Pharmaceutical Registry Data to OMOP Standard Concepts Using RxNorm and RxNorm Extension

A structured pipeline evaluated against a large language model baseline

Anabel Barrera García¹, Vaclav Papez²

¹ Sura Colombia, Medellín, Colombia

² University College London (UCL), London, United Kingdom

Background

In Colombia, pharmaceutical data are managed through INVIMA, the national authority for drug marketing authorizations. Despite this regulatory framework, its records show marked semantic heterogeneity: local naming conventions, free-text descriptions, variable units and incomplete attributes.

This variability creates a critical gap for integrating Colombian medication data into the OMOP Common Data Model (CDM), limiting participation in international real-world evidence initiatives and the OHDSI network¹⁻³; similar barriers were reported for national drug terminologies in Austria, Japan and France⁴⁻⁶.

This study develops a two-phase framework for normalizing and mapping INVIMA records to OMOP Concept IDs via RxNorm and RxNorm Extension, and benchmarks it against a large language model (LLM) performing the same task. We hypothesized that the structured pipeline would achieve broad coverage and complement, rather than outperform, the LLM, with residual failures reflecting vocabulary gaps rather than the method.

Methods

Three sources were integrated: the INVIMA sanitary registry (9,757 unique medication records in Spanish free text), RxNorm and RxNorm Extension as the target concept space (standard, non-deprecated concepts only), and the Anatomical Therapeutic Chemical (ATC) classification as an independent semantic anchor for ingredient identification.

Phase 1 — Attribute Normalization: Active ingredients were identified through two reconciled pathways: traversal of ATC-to-RxNorm relationships in the OMOP vocabulary and entity extraction by two LLMs (Gemini 2.5 Flash and GPT-4.1 mini), combined through a confidence cascade rather than majority voting and accepted only if present in the vocabulary. Strength was standardized by unit-conversion rules, including concentrations in mg/mL. Pharmaceutical form was mapped by an LLM to the equivalent RxNorm dose form, returning null when none existed. Brand and manufacturer were extracted by an LLM and accepted only when verified against the RxNorm brand and supplier catalogues.

Phase 2 — Hierarchical Matching: Matching operates on concept identifiers rather than strings, using DRUG_STRENGTH and CONCEPT_RELATIONSHIP. Candidates are first restricted to concepts whose active-ingredient set exactly matches the record; the engine then descends the hierarchy through strength (2% tolerance for salt-to-base rounding), dose form, brand and supplier, assigning each record the deepest class reached: Marketed Product, Branded Drug, Clinical Drug, Clinical Drug Component or Ingredient (Figure 1). Output follows the SOURCE_TO_CONCEPT_MAP convention; records absent from RxNorm are retained with target concept 0 and flagged as vocabulary-extension candidates.

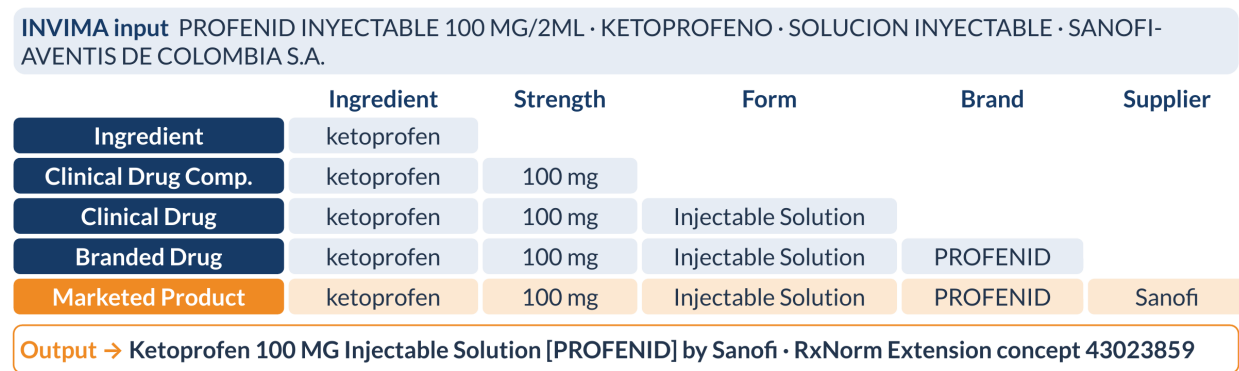


Figure 1. Worked example of the hierarchical descent: an INVIMA record accumulates attributes until it reaches a Marketed Product concept in RxNorm Extension.

Evaluation: a stratified sample of 436 records was annotated by a domain expert, blinded and attribute by attribute (ingredient, strength, form, brand). The pipeline was benchmarked against an LLM baseline that named the active ingredients, retrieved candidates by exact ingredient set and selected one real concept or none. We report exact concept-identifier precision, per-attribute precision, hierarchical depth and pairwise dominance, with stratum-weighted estimates.

Results

After a targeted correction motivated by the evaluation itself, which recovered 563 records without altering any prior assignment, 8,834 of 9,757 records (90.5%) were anchored to a standard RxNorm or RxNorm Extension concept at ingredient level or deeper. The remaining 923 comprised 919 combination products absent from RxNorm, represented by ingredient decomposition and flagged as extension candidates, and four biologicals with no identifiable ingredient. Clinical Drug was the most frequent class reached (4,724; 48.4%), followed by Clinical Drug Component (2,502; 25.6%), Ingredient (1,662; 17.0%), Marketed Product (621; 6.4%) and Branded Drug (244; 2.5%).

The pipeline and the LLM baseline showed complementary profiles (Table 1). The pipeline was more precise on ingredient and brand and near-perfect on branded products (exact precision 0.986 vs 0.493), whereas the LLM was more accurate on strength and form. Overall exact concept-identifier precision was 0.224 vs 0.296, but this metric is structurally zero for records correctly resolved to Component or Ingredient, which carry

no form; per-attribute precision is therefore the fairer comparison. Dominance was balanced (110 vs 111 records; 215 equivalent).

Table 1. Precision over applicable records, pipeline versus LLM baseline (n = 436).

Measure	Pipeline	LLM baseline
Ingredient	0.853	0.764
Strength	0.580	0.761
Form	0.596	0.750
Brand (all applicable)	0.316	0.222
Brand (present in RxNorm)	0.516	0.388
Exact concept ID, Branded Drug stratum	0.986	0.493
Exact concept ID, overall (weighted)	0.224	0.296

Bold marks the higher value. "Present in RxNorm" excludes Colombian brands absent from the vocabulary, which no system can assign.

The LLM omitted a brand that existed in RxNorm in 95 records versus 55 for the pipeline. The dominant limitation was the absence of local brands and combination products from the vocabularies.

Conclusions

- 8,834 of 9,757 INVIMA records (90.5%) were anchored to a standard concept, but only 6.4% reached Marketed Product, reflecting limited coverage of the Colombian market in international vocabularies.
- Structured descent and LLM parsing are complementary: the pipeline is more precise on ingredient and brand, the LLM more complete on strength and form.
- Per-attribute evaluation is a fairer basis than exact concept-identifier precision for hierarchical drug normalization.
- Vocabulary gaps, not methodological limitations, are the main barrier; the 919 absent combinations and local brands are catalogued as candidates for OHDSI vocabulary extension.

References

1. OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics. 2019.
2. Reich C, et al. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. *J Am Med Inform Assoc.* 2024;31:583–590.
3. Ostropelets A, Zhuk A, Korchmar E, Ryan P, Reich C. Developing RxNorm Extension: a step toward global drug data harmonization in observational drug research. *AMIA Annu Symp Proc.* 2024;2024:969–978.
4. Katsch F, Mészáros Á, Héja T, Hussein R, Duftschmid G. Semiautomatic mapping of a national drug terminology to standardised OMOP drug concepts using publicly available supplementary information. *BMC Med Res Methodol.* 2025;25:213.
5. Kimura E, Kawakami Y, Inoue S, Okajima A. A dataset for mapping the Japanese drugs to RxNorm standard concepts. *Data Brief.* 2025;59:111418.
6. Delange B, et al. A hybrid pipeline for mapping French UCD drug codes to RxNorm with dosage preservation. 2026. doi:10.3233/SHTI260385.