OMOP mapping and automated metadata harmonization for UK Biobank

Rohan Karthikeyan*, Param Shah*, Harshavarthan PK*, Jaya Singh, Radhakrishna Bettadapura, Badri Padhukasahasram
Strand Life Sciences, Bangalore, India *Contributed equally

VISIT OUR WEBSITE

strand

Contact
Badri Padhukasahasram
Vice President - Data Science
91-9591670961 badri.p@strandls.com

Introduction

- UK Biobank holds rich primary care (GP) data for ~500,000 participants, but its native format limits interoperability with the global network of OMOP-standardized observational health databases.
- Mapping UKB data into the OMOP Common Data Model unlocks the ability to apply validated OHDSI analytic tools, reuse established phenotype definitions, and run federated analyses alongside hundreds of other OMOP datasets worldwide.
- For translational work this standardization can be especially critical — it enables reproducible longitudinal modeling of drug exposures against clinical and laboratory outcomes, using a vocabulary framework (SNOMED, RxNorm, LOINC) that ensures semantic consistency and downstream portability.

Methods

Baseline ETL mapping

- Framework to map UK Biobank to OMOP framework
- Align CDM version & vocabulary releases
- Generate baseline OMOP instance from GP data
- Produce domain-level coverage statistics

Targeted Harmonization

- Map prescriptions Rx → Norm
- Map disease outcomes → SNOMED / ICD-10
- Map lab measurements → LOINC + unit normalization
- Resolve ambiguous / free-text entries via rule-based + LLM-assisted mapping

Ontology Mapping Pipeline

Raw GP Records → Rule-Based Normalization

LLM-Assisted Semantic Mapping (metadata only, no PII) → Expert Validation

→ OMOP Standard Concepts

Quality control and OMOP compliance

Ensuring Data Integrity Across All Transformed Domains

Domain Validation
- Row counts by OMOP domain.
- Sparsity & missingness metrics.
- Consistency across person, observation_period, drug_exposure, condition_occurrence, measurement tables.

Mapping QC
- Manual spot validation of concept mappings.
- Documentation of all transformation rules.
- Reproducibility verification of mapping logic

OHDSI Best Practices
- Vocabulary alignment checks (SNOMED, RxNorm, LOINC, ICD-10)
- Schema consistency with current UKB release.
- CDM version compatibility validation.

UK Biobank: OMOP mapping summary

- We executed mapping of 61 meta-data fields pertaining to a neurological condition from UK Biobank using an automated pipeline
- Table and piechart below denote the distribution of mapping as well as OMOP domains respectively
- Observations is the most common domain (~59%) followed by measurements (~19%), covering the vitals (blood pressure) and biometric data (BMI, sleep duration). Clinical domains (CONDITION_OCCURRENCE, DRUG_EXPOSURE) and administrative tables (VISIT_OCCURRENCE, PERSON, CARE_SITE) account for the remaining~22%

Mapping Category	Description	Percentage
Fully Mapped	These have clear OMOP standard concept matches (e.g., BMI → LOINC 39156-5, systolic/diastolic BP → LOINC codes, smoking status → SNOMED, ethnicity → PERSON table)	68.9%
Partial Mapping	A reasonable OMOP concept exists but the mapping requires qualifiers, loses OCP/HRT start/stop, comparative alcohol intake over time, and household income (social determinant concepts are still maturing in OMOP)	23.0%
Unmapped	No suitable OMOP standard concept. These are mostly UK Biobank-specific survey items like "variation in diet," "major dietary changes in last 5 years," "getting up in morning," "alcohol intake versus 10 years previously".	8.2%

5,829 Total prescription rows

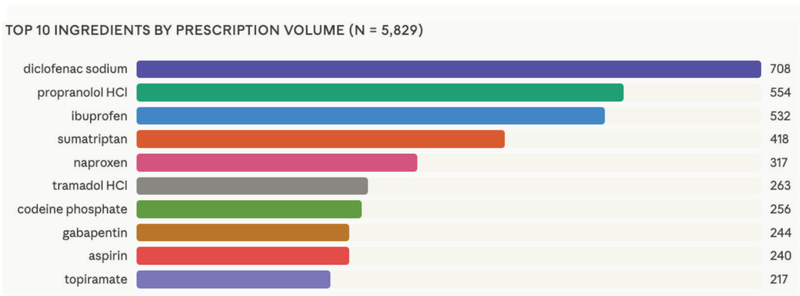
912 Unique source descriptions

319 Standardized RxNorm names

82 Unique RxNorm ingredients

HARMONIZATION PIPELINE

Raw input 912 heterogeneous UK drug descriptions → **Parse & extract** Ingredient, strength, dose form → **Normalize** Map to RxNorm terminology → **SCD output** 319 semantic clinical drug names



Ontology harmonization pipeline mapping heterogeneous UK prescription drug descriptions to RxNorm Semantic Clinical Drug (SCD) names. Normalizations include UK → US naming conventions (paracetamol → acetaminophen, salbutamol → albuterol), salt form resolution, dose form standardization to RxNorm term types, combination drug decomposition, and brand-to-generic resolution.

Summary and Achievements

- Mapping of UK Biobank metadata fields to OMOP framework.
- Quality control and validation of mapping by expert review.
- Use of LLM system to speedup harmonization of longitudinal prescription records and clinical events

Conclusions

- A pipeline for mapping UK Biobank metadata to OMOP
- Automated ontology harmonization for metadata fields along with QC checks and manual review.
- Summarization of variables information for longitudinal data after cleanup and standardization steps

Quality control and data readiness for longitudinal modeling

Longitudinal Readiness Assessment Evaluating GP Lab Data for Time-Series Translational Modeling

Longitudinal Density Metrics

Time-series continuity
Gap analysis across observation windows

Median observation intervals
Typical time between repeated measures

Per-patient measurement frequency
Distribution of measurement counts per individual

Temporal coverage span
Duration of follow-up per lab variable

Readiness Report Outputs

1. Data sufficiency per lab variable
2. Eligible cohort sizes for modeling
3. Potential modeling constraints identified
4. Data quality limitations & recommendations

Target Deliverable

Analysis-ready OMOP datasets with validated prescriptions, disease outcomes, and lab measurements suitable for downstream longitudinal and translational modeling.

Illustration: Prescription records for a neurological disease

Summary Metric	Measured
Patient Coverage	69.01%
Median Record per patient	179
Median Span of Coverage	18.47 years
Missingness	20.89%
Median inter-prescription interval	31 days
Percentage of clinical events descriptions that could be normailzed to ICD10 codes by LLM	84.1%