Converting **MIMIC to OMOP CDM** locally on consumer hardware.

Democratizing critical care research: A portable dbt and DuckDB Pipeline for MIMIC-IV to OMOP CDM Conversion

Background: Converting MIMIC-IV, a gold standard dataset of healthcare text analytics has previously been restricted by computation and performance. This GitHub repository can be used to quickly convert the MIMIC-IV into the OMOP CDM (v5.4) format.

Result 1: Comparison of PostgreSQL solution vs our solution dbt/DuckDB Metrics

Metric	PostgreSQL	dbt/DuckDB	Change
Record Counts (millions)			
Person	0.32M	0.36M	+14.6%
Measurement	169.1M	296.7M	+75.4%
Procedure	0.86M	6.56M	+7.6×
Observation	3.52M	14.02M	+4.0×
Note NLP	0	195.6M	New
Concept Mapping Rates			
Condition concepts	100%	100%	—
Drug concepts	86.1%	96.6%	+10.5%
Measurement concepts	0%	95.1%	+95.1%
Unit concepts	N/A	98.6%	New

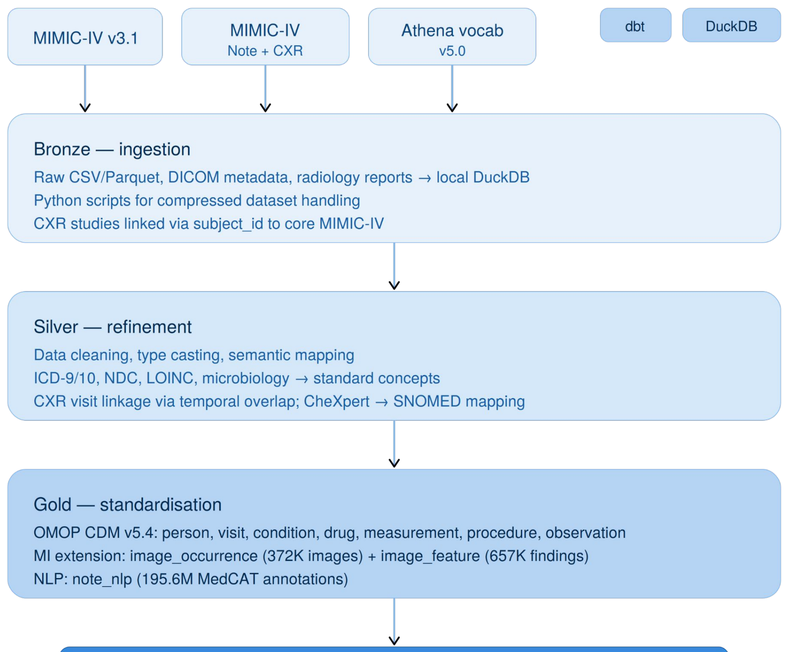
Result 2: Coverage of key OMOP concept ids on selected tables and row counts of various tables.

Table	Concept Column	Coverage
condition_occurrence	condition_concept_id	100%
procedure_occurrence	procedure_concept_id	100%
image_occurrence	modality_concept_id	100%
image_occurrence	anatomic_site_concept_id	100%
image_occurrence	image_type_concept_id	100%
image_feature	feature_concept_id	100%
image_feature	feature_type_concept_id	100%
drug_exposure	drug_concept_id	96.59%
drug_exposure	route_concept_id	97.79%
measurement	measurement_concept_id	95.13%
measurement	unit_concept_id	98.60%
specimen	specimen_concept_id	97.53%
person	gender_concept_id	100%
person	race_concept_id	52.23%
person	ethnicity_concept_id	0% (unmapped)

Table	Count
measurement	296705150
note_nlp	195642167
fact_relationship	74356046
drug_exposure	20351014
observation	14018411
drug_era	10296936
procedure_occurrence	6564445
condition_occurrence	5328942
person	364627
death	38301

Method

- **Data sources:** MIMIC-IV v3.1, MIMIC-IV-Note (331K discharge summaries, 2.3M radiology reports), MIMIC-IV-CXR (377K chest X-rays), CogStack/MedCAT NLP annotations, OHDSI Athena vocabularies v5
- **Architecture:** dbt + DuckDB medallion pipeline, runs entirely on a consumer laptop (<3 hrs, 16GB RAM)
  - **Bronze:** Ingest raw CSVs, Parquet, DICOM metadata, and radiology reports into local DuckDB
  - **Silver:** Data cleaning, type casting; map ICD-9/10, NDC, LOINC, microbiology codes to standard OMOP concepts; link CXR studies to visits via temporal overlap; map CheXpert findings to SNOMED
  - **Gold:** Materialise OMOP CDM v5.4 core tables (person, visit, condition, drug, measurement, procedure, observation) + MI extension tables (image_occurrence, image_feature)
- **NLP integration:** 195.6M structured annotations via note_nlp from MedCAT; pre-transformed table available for download
- **Output:** 343M+ clinical events, 364,627 patients, 372K CXR images (88.8% visit-linked), 656K CheXpert findings (100% concept-mapped), 69 dbt models



Conclusion: This repository can generate all OMOP relevant tables from the MIMIC-IV dataset in less than an hour on light consumer hardware (CPU & 16GB RAM). We also supplement this by supplying data for additional data from MIMIC-CXR (image_occurrence) and CogStackDashboards (note_nlp).

Adam Sutton[1], Niko Möller-Grell[6], Thomas Searle[1], Richard Dobson[1-6]

1 Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, UK 2 Institute for Health Informatics, University College London, UK. 3 NIHR Biomedical Research Centre, University College London Hospitals NHS Foundation Trust, UK. 4 Health Data Research UK London, University College London, UK. 5 NIHR Biomedical Research Centre, South London and Maudsley NHS Foundation Trust and King's College London, UK. 6 DRIVE-health, Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, UK