Fast, Accurate, and Local Conversion of MIMIC-IV to OMOP with DBT

Adam Sutton[1], Niko Möller-Grell[1], Thomas Searle[1], Richard Dobson[1,2,3]
[1] Department of Biostatistics and Health Informatics
Institute of Psychiatry, Psychology and Neuroscience King's College London, London, UK
[2] Institute of Health Informatics, University College London, London, United Kingdom
[3] NIHR Mausley Biomedical Research Centre, London, United Kingdom
Correspondence: adam.sutton@kcl.ac.uk

Abstract

dbt_mimic_omop is a free, open-source resource that converts the MIMIC-IV dataset to the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) format on consumer level hardware. CDM approaches are increasingly adopted in both industry and academia due to the need for interoperability and reproducibility, including in clinical NLP tasks such as cohort selection, information extraction, and retrieval-augmented generation. The MIMIC-IV database is among the most widely used critical care research datasets, yet existing pipelines to transform it to OMOP depend on enterprise database infrastructure and complex orchestration, limiting accessibility for practitioners and resource-constrained researchers. We further integrate free-text clinical notes (195.6M clinical annotations) and chest radiographs into the OMOP note_nlp and imaging extension tables, making all MIMIC-IV modalities (structured data, free-text, and imaging) accessible through a common data model. This resource generates a more comprehensive dataset than existing alternatives and is intended to be used to aid in system development, testing, and evaluation.

1 Introduction

A growing number of clinical datasets are becoming available to the research community (All of Us Research Program Investigators, 2019; Clark et al., 2013; Pollard et al., 2018; Johnson et al., 2023b). However, each resource adopts its own bespoke schema, limiting interoperability and increasing development overhead of multi-source studies.

Common Data Models (CDMs) address this by enabling consistent data representation across institutions and applications through a standardised framework of schemas, attributes, entities, and relationships, thus facilitating collaboration, benchmarking, and reproducible analytics (Finster et al., 2025; Brown et al., 2022; PCORnet, 2025; Garza et al., 2016).

These data models have proven to be useful in emerging fields within healthcare and computational linguistics, such as information extraction and retrieval augmented generation (Keloth et al., 2023; Möller-Grell et al., 2026).

The Medical Information Mart for Intensive Care (MIMIC-IV) database is a cornerstone of critical care research, providing granular clinical data for thousands of admissions to the ICU (Johnson et al., 2023c). The MIMIC-IV data collection covers a wide range of data modalities, from structured data and textual data to images.

However, the joint analysis of these disparate data sources can be resource intensive, requires custom adaptations, and is not easily extensible to additional data sources. Converting these datasets into a CDM such as the OMOP CDM provides a unified extensible representation to cross-institutional data and a consistent foundation for systematic annotation of clinical text via Natural Language Processing (NLP).(Observational Health Data Sciences and Informatics, 2026).

Existing MIMIC ETL pipelines rely on enterprise database infrastructure (BigQuery) and complex orchestration (Paris et al., 2021; Sciences and Informatics, 2020), creating barriers for those who are resource-constrained.

dbt_mimic_omop' [1] is a lightweight, modular and portable ETL pipeline that uses dbt-Core (data build tool) ((dbt Labs, Inc., 2026)) and DuckDB ((Raasveldt and Mühleisen, 2019)) to transform MIMIC-IV data into the OMOP CDM (v5.4).

In addition to the core MIMIC-IV dataset, we also offer conversion for supplementary MIMIC-IV datasets. MIMIC-IV-Note contains more than 2 million de-identified free text notes, MIMIC-IV-CXR is an imaging database with more than

[1] link removed for reivew

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 992–996

300,000 medical images. Both datasets are linked to the same patients as the core dataset. Further to this, we provide a resource with clinical annotations for the free text data in MIMIC-IV-Note.

2 Data Sources

2.1 MIMIC-IV

MIMIC-IV is well known in multiple fields of research, providing access to de-identified healthcare data. MIMIC-IV contains data for more than 65,000 patients admitted to the ICU and more than 200,000 patients admitted to the emergency department (Johnson et al., 2023a). The dataset does not adhere to any common data model.

2.2 MIMIC-IV-Note

MIMIC-IV-Note contains 331,794 de-identified discharge summaries from 145,915 patients and 2,321,355 radiology reports and for 237,427 patients, from the same sample of patients from the core MIMIC-IV dataset.

2.3 MIMIC-IV-CXR

The MIMIC Chest X-ray Database (MIMIC-CXR) is a large publicly available dataset of chest radiographs in DICOM format with free-text radiology reports (Johnson et al., 2019). The dataset contains 377,110 images corresponding to 227,835 radiographic studies. This dataset uses the CheXpert labelling system (Irvin et al., 2019), 14 distinct radiographic diagnostic observations through rule-based extraction are mapped to SNOMED concept IDs via custom vocabulary mappings.

2.4 CogStack Dashboards

CogStack Dashboards is an OpenSearch cluster that hosts the MIMIC-IV and MIMIC-IV-Note dataset[2]. In addition to the MIMIC dataset, annotations have been generated, provided via MedCAT (Kraljevic et al., 2021; Sutton et al., 2026). These generated annotations use the SNOMED CT ontology (SNOMED International, 2025) to provide coding for all detected entities. They also provide additional information about the entity, such as meta information and character positioning within the text. This model has been trained on various open-source datasets as well as privately trained data from King's College Hospital and Guy's and St Thomas' Hospital. Annotations such as these are useful in information extraction and retrieval augmented generation tasks (Keloth et al., 2023; Möller-Grell et al., 2026).

[2]https://cogstackdashboards.sites.er.kcl.ac.uk/

2.5 OHDSI Athena

OHDSI Athena is the repository for OHDSI standardised vocabularies used in the OMOP CDM. It allows users to search, browse, and download standardised mapped medical codes (concepts) for clinical, drug, and procedure data, allowing harmonisation and interoperability (Observational Health Data Sciences and Informatics, 2025).

3 Implementation

The DBT project's medallion architecture to transform raw MIMIC-IV to OMOP is as follows:

- **Bronze Layer (Ingestion):** Raw MIMIC-IV CSV files (v3.1), MIMIC-IV-CXR metadata and radiology reports, and OHDSI Athena vocabularies (v5) are ingested into a local DuckDB database. Python scripts facilitate efficient loading of large compressed datasets directly from Parquet/CSV sources. For MIMIC-IV-CXR, DICOM metadata is extracted and loaded alongside associated free-text radiology reports, which are linked to the patient and study identifiers shared with the core MIMIC-IV dataset.
- **Silver Layer (Refinement):** Intermediate views handle data cleaning, type casting, and semantic mapping. Specific transformations include cleaning visit details from various source tables (such as admissions/transfers) and mapping source codes (ICD-9/10, NDC, LOINC, local microbiology codes) to standard OMOP concepts using the OHDSI Athena vocabulary lookup tables. CXR radiology reports are linked to their corresponding MIMIC-IV hospital admission and processed for downstream annotation. CXR studies are linked to OMOP visit occurrence records by temporal overlap, matching subject ids and cross-referencing visit windows.
- **Gold Layer (Standardisation):** The final OMOP CDM v5.4 tables are compiled, covering key OMOP tables such as person, visit occurrence, condition occurrence, drug exposure, measurement, procedure occurrence, and observation. Tables generated from supplementary data do not contribute to the core

Domain	Field	PostgreSQL	dbt/DuckDB
condition_occurrence	condition_concept_id	100%	100%
procedure_occurrence	procedure_concept_id	100%	100%
drug_exposure	drug_concept_id	86.10%	96.59%
drug_exposure	route_concept_id	0%	97.79%
measurement	measurement_concept_id	0%	95.13%
measurement	unit_concept_id	0%	98.60%
specimen	specimen_concept_id	0%	97.53%
person	gender_concept_id	100%	100%
person	race_concept_id	0%	52.23%
person	ethnicity_concept_id	0%	0% (unmapped)

Table 1: Key OMOP Concept Coverages - each percentage represents the number of rows in each corresponding table that has appropriate concept ID in the specified field. These fields are required by the OMOP CDM standard.

Table	Count
measurement	296705150
note_nlp	**195642167**
fact_relationship	74356046
drug_exposure	20351014
observation	14018411
drug_era	10296936
procedure_occurrence	6564445
condition_occurrence	5328942
note	**2652887**
image_feature	**656697**
image_occurrence	**372209**
person	364627
death	38301

Table 2: Key OMOP Table Counts

OMOP CDM tables. MIMIC-IV-Note and annotations from CogstackDashboards are inserted into the note and note_nlp tables respectively. Tables sourced from MIMIC-CXR are generated into the OMOP CDM MI working group extension tables: image_occurrence with visit linkage, view position, modality, local and google cloud file paths; and image_feature, with one row per CheXpert finding, certainty labels mapped per custom SNOMED mapping.

The modular dbt project structure separates logic into 69 distinct models at transformation stage (e.g., separate models for admissions, pharmacy, labevents), facilitating maintenance and community contributions.

4 Results

The pipeline transforms MIMIC-IV v3.1 to OMOP CDM v5.4, generating more than 343 million standardised clinical events in 364,627 patients. On a consumer laptop (16gb of RAM, 8 core cpu), the full ETL process can be completed in less than three hours.

Tab. 2 shows the record counts of key OMOP tables and the tables generated from the supplementary data source. Tab. 1 shows the mapping rates for key concepts within this implementation.

Our proposed pipeline recovers previously unmapped domains, most notably measurement concept ids ($0\% \rightarrow 95.13\%$) and specimen concept ids ($0\% \rightarrow 96.59\%$), and introduces full coverage for all imaging fields. Race concepts improve from 0% to 52.23%, while ethnicity concept ids remains unmapped at 0% in both implementations. These lesser performances are due to the vocabularies lacking standard translations from MIMIC-IV recording of ethnicity and race to OMOP standardized codes.

We also directly compare the coverage between our solution and Legrand et al. (2025). Tab. 3 shows raw record counts and concept mapping rates for both solutions.

Free text clinical notes are also included, sourced from MIMIC-IV-Note. These notes are imported into the note table. The pipeline also incorporates full NLP annotation processing via the note_nlp table, adding 195.6M structured annotations from previously unprocessed clinical notes—a capability absent from prior implementations. However, due to the significant processing cost, we also offer the table already transformed ‘note_nlp‘ as a download,

Table 3: Comparison of PostgreSQL solution (Legrand et al., 2025) vs 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×
Concept Mapping Rates			
Condition concepts	100%	100%	—
Drug concepts	86.1%	96.6%	+10.5%
Measurement concepts	0	95.1%	+95.1%
Unit concepts	0%	98.6%	New

with instructions provided in the repository.

In addition to the core OMOP CDM tables, the pipeline populates the OMOP Medical Imaging (MI) working group extension tables. The table image_occurrence contains 372,209 chest X-ray images, and image_feature captures 656,697 CheXpert findings with 100% concept mapping across all imaging fields (modality, anatomic site, image type concept and feature type). Of these images, 330,401 (88.8%) were successfully linked to visit_occurrence via temporal mapping.

5 Conclusion

The ‘dbt_mimic_omop‘ pipeline is freely available on GitHub [3]. This open-source solution significantly reduces the technical barrier for researchers wishing to conduct OHDSI-compliant studies using MIMIC-IV data. By decoupling the ETL process from enterprise-grade infrastructure, we facilitate broader adoption in educational settings, reproducibility in research, and rapid prototyping of analytical pipelines.

This resource is intended to be used as an aid for those who are developing NLP systems that employ the OMOP Common Data Model. The free text that is available via ‘mimic-iv-note‘ is available and in OMOP format for various tasks such as entity recognition and linking. The generated annotations we have provided can be used as baseline comparison. The dataset could also find use as a test dataset for more complex pipelines such as Retrieval Augmented Generation.

By combining dbt’s software engineering best practices with DuckDB’s embedded analytical capabilities, and extending the CDM with note_nlp and imaging tables, we aim to lower the barrier to clinical NLP, multimodal system development, and reproducible observational research over MIMIC-IV.

Future work includes improvements to this ETL pipeline. Such as additional versions of the MIMIC datasets (Johnson et al., 2016), and other publicly available datasets. Future versions of the repository can also include additional supplementary datasets (Moody et al., 2022). Such pipelines can also speed up the benchmarking process. Various datasets and benchmarking challenges are available, but often require development time for that specific challenge (DrivenData, 2026). Work can be done making such challenges follow a common data model, if applicable.

[3] https://github.com/CogStack/dbt_mimic_omop

References

All of Us Research Program Investigators. 2019. The “all of us” research program. *New England Journal of Medicine*, 381(7):668–676.

Jeffrey S Brown, Aaron B Mendelsohn, Young Hee Nam, and 1 others. 2022. The us food and drug administration sentinel system: a national resource for a learning health system. *Journal of the American Medical Informatics Association*, 29(12):2191–2200.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. 2013. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057.

dbt Labs, Inc. 2026. dbt (Data Build Tool). Technical report, dbt Labs, Inc. Accessed: 2026-02-20.

DrivenData. 2026. Snomed ct entity linking benchmark. https://www.drivendata.org/benchmarks/

310/benchmark-snomed-ct/. Accessed: 2026-02-25.

Melissa Finster, Markus Wenzel, and Elham Taghizadeh. 2025. Common data models and data standards for tabular health data: a systematic review. *BMC Medical Informatics and Decision Making*, 25(1):422.

Maryam Garza, Guilherme Del Fiol, Jessica Tenenbaum, Anita Walden, and Meredith Nahm Zozus. 2016. Evaluating common data models for use with a longitudinal community registry. *Journal of biomedical informatics*, 64:333–341.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 590–597.

Alistair Johnson, Tom Pollard, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. MIMIC-IV-Note: Deidentified free-text clinical notes.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023b. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023c. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(160035).

Vipina K Keloth, Juan M Banda, Michael Gurley, Paul M Heider, Georgina Kennedy, Hongfang Liu, Feifan Liu, Timothy Miller, Karthik Natarajan, Olga V Patterson, and 1 others. 2023. Representing and utilizing clinical textual data for real world studies: An ohdsi approach. *Journal of Biomedical Informatics*, 142:104343.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, and 1 others. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

Paul Legrand, Kawsar Noor, Satyam Bhagwanani, and Richard Dobson. 2025. An open-source text-to-sql pipeline for omop-formatted electronic health records. Technical report, King's College London. Available from https://github.com/red1dwarf/text-to-omop/tree/main/ETL (accessed 2026-02-20).

Niko Möller-Grell, Linglong Qian, and Shihao Shenzhang. 2026. Multimodal ards detection through agentic medgemma workflows. https://www.kaggle.com/competitions/med-gemma-impact-challenge/writeups/new-writeup-1770377682785. Kaggle MedGemma Impact Challenge write-up. Accessed: 2026-02-25.

Benjamin Moody, Sicheng Hao, Brian Gow, Tom Pollard, Wei Zong, and Roger Mark. 2022. MIMIC-IV Waveform Database. *PhysioNet*. Version 0.1.0.

Observational Health Data Sciences and Informatics. 2025. ATHENA – OHDSI standardized vocabularies.

Observational Health Data Sciences and Informatics. 2026. OMOP Common Data Model. Technical report, OHDSI. Accessed 2026-02-20.

Nicolas Paris, Antoine Lamer, and Adrien Parrot. 2021. Transformation and evaluation of the mimic database in the omop common data model: development and usability study. *JMIR Medical Informatics*, 9(12):e30970.

PCORnet. 2025. Pcornet common data model v7.0 specification. Technical report, Patient-Centered Outcomes Research Network (PCORnet).

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178.

Mark Raasveldt and Hannes Mühleisen. 2019. Duckdb: an embeddable analytical database. In *Proceedings of the 2019 international conference on management of data*, pages 1981–1984.

Observational Health Data Sciences and Informatics. 2020. Ohdsi/mimic: Etl for transforming mimic-iv to the omop common data model. Technical report, OHDSI. GitHub repository. Accessed 2026-02-20.

SNOMED International. 2025. SNOMED CT. International Health Terminology Standards Development Organisation (IHTSDO) / SNOMED International, International Edition.

Adam Sutton, Vlad Dinu, Thomas Searle, and Richard Dobson. 2026. Cogstack dashboards. Accessed: 2026-02-20.