Real-World Implementation of the Medical Imaging CDM:

An Alzheimer's Disease Use Case

Woo Yeon Park, M.S.[1], Teri Sippel Schmidt, M.S. [1], Blake Dewey, Ph.D. [2], Paul Nagy, Ph.D. [1]
[1]Biomedical Informatics and Data Science, School of Medicine, Johns Hopkins University, Maryland, U.S.A
[2]Department of Neurology, School of Medicine, Johns Hopkins University, Maryland, U.S.A

Background

Medical imaging comprises up to a vast part of healthcare data volume, yet the current OMOP Common Data Model (CDM) captures imaging studies only via procedure codes, which cannot represent the complexity or diversity of real-world imaging data[1]. As a result, observational researchers lack access to valuable imaging-derived information for cohort discovery and downstream analysis[2].

To address this gap, the OHDSI Imaging Workgroup proposed the Medical Imaging Common Data Model (MI-CDM), an extension to OMOP CDM. MI-CDM introduces two new tables and imaging-specific vocabularies that support structured, standardized representation of DICOM metadata without including pixel data[3]. While previous work demonstrated this framework using publicly available datasets, real-world implementation and validation remain limited.

This study presents the first large-scale implementation of MI-CDM at Johns Hopkins Medicine using 1 million DICOM series from an Alzheimer's Disease neuroimaging cohort. We demonstrate the feasibility of extracting and transforming DICOM metadata into OMOP-compliant imaging tables, enabling integration of imaging data with electronic health records for observational research. This work was reviewed and approved by the Johns Hopkins Institutional Review Board #IRB00228485.

Methods

Study Population and Data Source

The Alzheimer's Disease Projection project includes patients aged 50 years or older with at least one clinical encounter within the Johns Hopkins health system. The cohort was already mapped to a standard OMOP CDM instance but lacked imaging metadata.

DICOM Metadata Extraction

DICOM metadata (headers) were extracted from the institution's Vendor Neutral Archive (VNA), excluding pixel arrays and private tags. Pixel data remain in the imaging archive and are retrieved only when specific phenotype criteria are met, reducing storage redundancy. Private tags, which contain vendor- or site-specific fields not standardized across institutions, were excluded to enhance model generalizability. In this study, we sampled 1 million DICOM series from the VNA extract for efficient prototyping[4].

MI-CDM Implementation Workflow (Figure 1):

1. Load DICOM vocabulary into OMOP vocabulary tables
2. Index source DICOM metadata from VNA, excluding pixel data
3. Transform metadata into OMOP-compliant records and load into MI-CDM extension tables
4. Define target cohorts using ATLAS cohort discovery platform
5. Enable phenotype-driven queries using imaging metadata parameters for targeted data retrieval

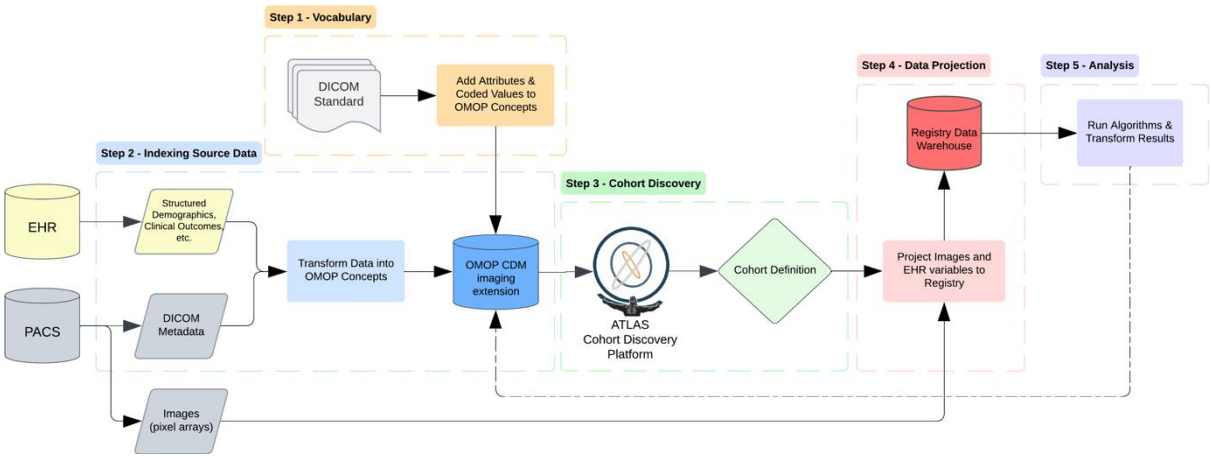

Figure 1. The figure illustrates five steps in dotted-lined and colored boxes containing actions and data elements. The actions are encapsulated in rectangles, each with a color matching the steps. The data elements include flat files (DICOM Standard), databases as cylinders (EHR, PACS, OMOP CDM, Registry), components as rhombuses (demographics, clinical outcomes, metadata, images), and a cohort definition as a diamond. The ATLAS cohort discovery platform is shown as the logo of ATLAS.

Data Transformation Process

All transformations were conducted using SQL scripts and DICOM vocabulary files from the OHDSI Imaging Working Group GitHub repository (https://github.com/OHDSI/ImageWG). Core imaging attributes (patient ID, study UID, series UID, modality, body part examined, and study date) were extracted to populate the image_occurrence table. All remaining DICOM tags were systematically transformed and loaded into image_feature and measurement tables, with multi-part values, empty fields, and long strings such as study description excluded during processing.

Clinical Integration and Phenotyping

Following DICOM metadata integration, imaging parameters can be queried alongside clinical variables to enable granular phenotype definitions. Then we created a phenotype leveraging both features: patients who had T1-weighted brain MRIs (Table 1). This phenotype criteria will be used to populate necessary images and patients for an imaging registry to run a brain segmentation algorithm.

Table 1. Computable Phenotype Inclusion Criteria

Variables	**Ranges/ Concept IDs**	**Remarks**
Age	>= 50 years old	
Procedure (Brain MRI)	3037128, 3024397	Terminology: LOINC
DICOM Acquisition Parameters*	2128000816 (Repetition Time)	between 1500 ms and 2500 ms OR < 10 ms
	2128000817 (Echo Time)	< 10 ms

	2128000818 (Inversion Time)	between 800 ms and 1100 ms OR < 500 ms

* Ranges were determined by an imaging expert based on the most popular manufacturer (Siemens, GE, Philips)

Results

We successfully extracted and transformed DICOM metadata from 1 million series representing 108,250 patients into MI-CDM tables. The dataset contained 73,965,730 individual DICOM metadata elements (average 73 elements per series), demonstrating the rich information available beyond traditional procedure codes.

Table 1. The high-level counts on the DICOM image sample extract.

	Counts
Patients	108,250
DICOM Studies	311,810
DICOM Series	999,972
Average Series per Patient	9
DICOM metadata elements	73,965,730
Average elements per Series	73
Most popular modalities	Magnetic Resonance (MR), Computed Tomography (CT), Structured Report (SR)

The transformation populated image_occurrence tables with core imaging attributes and image_feature and measurement tables with technical parameters. This integration enables phenotype-driven queries combining clinical and imaging criteria. We selected T1-weighted MRI sequences for patients, which was required to run brain segmentation algorithm. We found 13,924 patients with 47,977 DICOM series, meeting the criteria.

Conclusion

This study demonstrates a real-world implementation of MI-CDM model in a large academic medical center. While this use case focused on Alzheimer's Disease research, the MI-CDM is domain agnostic and extensible to other imaging types (e.g., CT, ultrasound, ophthalmologic imaging). By enabling standardized and scalable representation of imaging metadata within OMOP CDM, MI-CDM supports reproducible, phenotype-driven use of imaging data across diverse research and clinical domains.

Reference

1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014 Feb 7;2:3. doi: 10.1186/2047-2501-2-3. PMID: 25825667; PMCID: PMC4341817.
2. OHDSI Collaborative. (2019). The Book of OHDSI: Observational Health Data Sciences and Informatics. Available at: https://ohdsi.github.io/TheBookOfOhdsi/
3. Park WY, Jeon K, Schmidt TS, Kondylakis H, Alkasab T, Dewey BE, You SC, Nagy P. Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension. J Imaging Inform Med. 2024 Apr;37(2):899-908. doi: 10.1007/s10278-024-00982-6. Epub 2024 Feb 5. PMID: 38315345; PMCID: PMC11031512
4. Park WY, Sippel Schmidt T, Salvador G, et al. Breaking data silos: incorporating the DICOM imaging standard into the OMOP CDM to enable multimodal research. J Am Med Inform Assoc. 2025;32(10):1533-1541. doi:10.1093/jamia/ocaf091