Development of Medical Imaging Data Standardization for Imaging Based Observational Research: OMOP Common Data Model Extension

<Woo Yeon Park, M.S.>[1], <Kyulee Jeon, B.S.>[2,3], <Teri Sippel Schmidt, M.S.>[1], <Haridimos Kondylakis, Ph.D.>[4], <Tarik Alkasab, M.D., Ph.D.>[5], <Blake E. Dewey, Ph.D.>[6], <Seng Chan You, M.D., Ph.D.>[2,3], <Paul Nagy, Ph.D.>[1]

[1]Biomedical Informatics and Data Science, Johns Hopkins University, Baltimore, MD, USA. [2]Department of Biomedical Systems Informatics, Yonsei University, Seoul, Korea. [3]Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, Korea. [4]Institute of Computer Science, Foundation of Research & Technology-Hellas (FORTH), Heraklion, Greece. [5]Department of Radiology, Massachusetts General Hospital, Boston, MA, USA. [6]Department of Neurology, Johns Hopkins University, Baltimore, MD, USA.

Background

Imaging researchers focus on obtaining knowledge from medical images, while observational researchers rely on structured information from electronic health records (EHRs). Accessing organ-level measurements is essential to quantify disease progression and treatment efficacy. The emergence of deep learning models provides important disease biomarkers. Our goal is to link image-based measurements to the OMOP CDM to harness deeper phenotypes tracked in the EHR.

The current OMOP CDM, designed for observational healthcare data, has limited potential to fully represent the diverse information stored in medical images that are formatted in Digital Imaging and Communication in Medicine (DICOM) standard. Medical imaging events and the findings requires different data representation and management compared to EHR or claims.

Building upon the work of Park et al. (2022) on the Radiology Common Data Model (R-CDM), the Medical Imaging extension is designed to represent not only radiological studies but all DICOM-based studies from different specialties, such as pathology and ophthalmology. The proposed model provides a methodology to incorporate DICOM into the OMOP CDM, while providing provenance for reproducibility. The proposed model tracks provenance of each feature which can result from structured reports as well as deep learning algorithms.

Methods

The Medical Imaging Working Group for the OHDSI community was formed in 2021, comprised of imaging scientists and observational health researchers familiar with OMOP CDM. The working group evaluated standard vocabularies, defined fields containing key imaging events, and identified limitations of the model. The working group started with the R-CDM in the development of the medical imaging extension. Imaging researchers across the field—radiology, pathology, neurology—were consulted to gather requirements and gain insights into the structure, vocabularies, and usability of the proposed model.

The extension adheres to the OMOP CDM convention to seamlessly integrates with existing tables and OHDSI applications like ATLAS, facilitating access to information from other domains when defining cohorts based on imaging findings (Figure 1). The medical imaging tables focus specifically on imaging information, minimizing duplication, and limited to a specific domain. We leverage the flexible structure of the OMOP CDM to incorporate widely used imaging vocabularies such as DICOM and RadLex, along with SNOMED and LOINC.

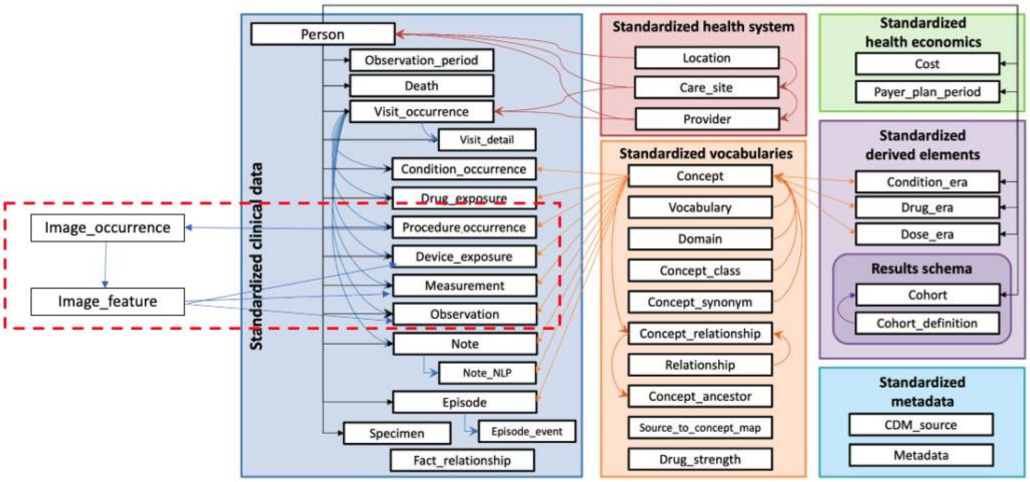

Figure 1. Incorporation of proposed Medical Image Data Model to existing OMOP CDM v5.4

Results

We have developed two tables, Image_occurrence and Image_feature, as an extension to the OMOP CDM for standardized representation of complex medical imaging events and features (Table 1).

Table 1. Image_occurrence (upper) and Image_feature (lower) tables

Image_Occurrence Table			
Field	Required	Data type	Description
image_occurrence_id (PK)	Yes	integer	The unique key is given to an imaging study record (often referred to as the accession number or imaging order number)
person_id (FK)	Yes	integer	The person_id of the Person for whom the procedure is recorded. This may be a system-generated code
procedure_occurrence_id (FK)	Yes	integer	The unique key is given to a procedure record for a person. Link to the Procedure_occurrence table

visit_occurrence_id (FK)	No	integer	The unique key is given to the visit record for a person. Link to the Visit_occurrence table
anatomic_site_concept_id (FK)	No	integer	Anatomical location of the imaging procedure by the medical acquisition device (gross anatomy). It maps the ANATOMIC_SITE_SOURCE_VALUE to a Standard Concept in the Spec Anatomic Site domain. This should be coded at the lowest level of granularity
wadors_uri	No	varchar (max)	A Web Access to DICOM Objects via Restful Web Services Uniform Resource Identifier on study level
local_path	Yes	varchar (max)	Universal Naming Convention (UNC) path to the folder containing the image object file access via a storage block access protocol (e.g., \\Server\Directory)
image_occurrence_date	Yes	date	The date the imaging procedure occurred
image_study_UID	Yes	varchar (250)	DICOM Study UID
image_series_UID	Yes	varchar (250)	DICOM Series UID
modality	Yes	varchar (250)	DICOM-defined value (e.g., US, CT, MR, PT, DR, CR, NM)

Image_feature Table

Field	Required	Data type	Description
image_feature_id (PK)	Yes	integer	The unique key is given to an imaging feature
person_id (FK)	Yes	integer	The person_id of the Person table for whom the the procedure is recorded. This may be a system-generated code
image_occurrence_id (FK)	Yes	integer	The unique key of the Image_occurrence table

table_concept_id	Yes	integer	The concept_id of the domain table that feature is stored in Measurement, Observation, etc. This concept should be used with the table_row_id
table_row_id	Yes	integer	The row_id of the domain table that feature is stored
image_feature_concept_id	Yes	integer	Concept_id of standard vocabulary—often a LOINC or RadLex of image features
image_feature_type_concept_id	Yes	integer	This field can be used to determine the provenance of the imaging features (e.g., DICOM SR, algorithms used on images)
image_finding_concept_id	No	integer	RadLex or other terms of the groupings of image feature (e.g., nodule)
image_finding_id	No	integer	Integer for linking related image features. It should not be interpreted as an order of clinical relevance
anatomic_site_concept_id	No	integer	This is the site on the body where the feature was found. It maps the ANATOMIC_SITE_SOURCE_VALUE to a Standard Concept in the Spec Anatomic Site domain
alg_system	No	varchar (max)	URI of the algorithm that extracted features, including version information
alg_datetime	No	datetime	The date and time of the algorithm processing

The Image_occurrence table includes information about imaging acquisition and other image-specific procedure extracted from the DICOM metadata, which is beyond the scope of the Procedure_occurrence table. Researchers can retrieve pixel data and other DICOM attributes at the study or series level, using study_UID and series_UID through DICOMweb and wadors_uri queries. Certain DICOM metadata elements, such as modality and anatomic site location, have been standardized and structured for imaging study. It also serves as provenance for the Image_feature table.

The Image_feature table captures the characteristics and provenance of features derived from medical images and links them to the clinical domain tables such as the Measurement or Observation tables. It enables researchers to identify and reconstruct the algorithms and parameters used. If multiple imaging features are closely related, such as descriptions of the same nodule or acquisition parameters of an image, they can be grouped by an image_finding_id. The image_feature_concept_id represents features, such as a “round mass”. The image_finding_concept_id provides a clinical interpretation of the finding, such as a “nodule”.

As part of this project, we propose mapping RadLex codes and DICOM value sets to OMOP concept_ids. Every image is a "DICOM object" comprising pixel data and a header with attributes identifying information from the modality to acquisition parameters. The DICOM standard has "attribute number" as the key and "Value Set" as the value. If DICOM has a defined Value Set for the attribute, the value is described in DICOM Part 16 "Context Groups". DICOM value sets can refer to SNOMED and LOINC.

The Image_occurrence table uses DICOM attributes and Value Sets as the standard vocabulary. The Image_feature table focuses on imaging-specific characteristics primarily defined in DICOM "Context Groups" and RadLex vocabularies. It should be noted that there is an ACR and RSNA joint initiative to create radiology "Common Data Elements" (CDEs), which will play a critical role in coding the "key-value" pair of radiology imaging findings.

Both tables include unique identifiers from the Person and Visit_occurrence tables. The concept for the imaging study performed is referred to by the procedure_concept_id through the procedure_occurrence_id in the Procedure_occurrence table. The Image_occurrence table has a many-to-many relationship with the Procedure_occurrence table (Figure 2). The Image_feature table has a one-to-many relationship with the Image_occurrence table. The Image_feature table contains fields that explain what the findings are, how the findings are discovered, and a grouper value that aggregates related findings. Each value of imaging findings is recorded in the relevant clinical data table. Thus, the Image_feature table has a one-to-one relationship with clinical data tables.

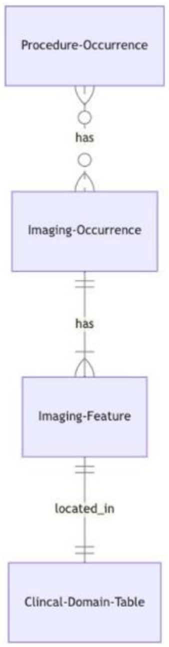

Figure 2. Integration DICOM and OMOP clinical data tables

The proposed imaging tables enable researchers to define computational phenotypes using imaging features. For example, a researcher would like to define cohorts with chest CT scans with a slice thickness of 1 mm and 150 kVp for patients who were ultimately diagnosed with lung cancer. The Image_occurrence table holds information about anatomical location, the date of procedures, and the modality. The

Image_feature table captures features from the image, including image acquisition parameters like a 1mm slice thickness.

Conclusion

In this study, we standardized the representation of medical image events and features within the OMOP CDM framework. We introduced two new tables seamlessly integrating imaging data into the existing CDM structure. This enables efficient storage and retrieval of medical images and facilitates cross-study comparisons and collaboration across different institutions, such as federated evaluation of medical image AI models. Moreover, including imaging features within the OMOP CDM broadens the scope of observational research, allowing for more comprehensive investigations into the associations between imaging findings and various clinical outcomes. The next step in this work is to seek feedback and develop reference implementations to be conducted by the OHDSI Medical Imaging workgroup.

References

1. Park C, You SC, Jeon H, Jeong CW, Choi JW, Park RW. Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. Yonsei Med J. 2022;63(Suppl):S74-S83. doi:10.3349/ymj.2022.63.S74
2. Seong Y, You SC, Ostropolets A, et al. Incorporation of Korean Electronic Data Interchange Vocabulary into Observational Medical Outcomes Partnership Vocabulary. Healthc Inform Res. 2021;27(1):29-38. doi:10.4258/hir.2021.27.1.29
3. Observational Health Data Science and Informatics. OMOP Common Data Model Conventions. Accessed May 29, 2023. http://ohdsi.github.io/CommonDataModel/dataModelConventions.html#Data_Model_Conventions
4. Mildenberger P, Eichelberg M, Martin E. Introduction to the DICOM standard. Eur Radiol. 2002;12(4):920-927. doi:10.1007/s003300101100
5. DICOM Standard. DICOM Part 6. Accessed May 31, 2023. https://dicom.nema.org/medical/dicom/current/output/html/part06.html
6. DICOM Standard. DICOM Part 16. Accessed May 31, 2023.

   https://dicom.nema.org/medical/dicom/current/output/html/part16.html
7. Shore MW, Rubin DL, Kahn CE. Integration of Imaging Signs into RadLex. J Digit Imaging. 2012;25(1):50-55. doi:10.1007/s10278-011-9386-x