









MI-Common Data Model: Extending Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) for Registering Medical Imaging Metadata and Subsequent Curation Processes

Varvara Kalokyri, PhD¹ ; Haridimos Kondylakis, PhD¹ ; Stelios Sfakianakis, PhD¹ ; Katerina Nikiforaki, PhD¹ ; Ioannis Karatzanis, MS¹ ; Simone Mazzetti, MD^{1,2,3} ; Nikolaos Tachos, PhD^{1,4}; Daniele Regge, MD, PhD^{1,3}; Dimitrios I. Fotiadis, PhD^{1,4}; Konstantinos Marias, PhD¹ ; and Manolis Tsiknakis, PhD¹ 

DOI <https://doi.org/10.1200/CCI.23.00101>

ABSTRACT

PURPOSE The explosion of big data and artificial intelligence has rapidly increased the need for integrated, homogenized, and harmonized health data. Many common data models (CDMs) and standard vocabularies have appeared in an attempt to offer harmonized access to the available information, with Observational Medical Outcomes Partnership (OMOP)-CDM being one of the most prominent ones, allowing the standardization and harmonization of health care information. However, despite its flexibility, still capturing imaging metadata along with the corresponding clinical data continues to pose a challenge. This challenge arises from the absence of a comprehensive standard representation for image-related information and subsequent image curation processes and their interlinkage with the respective clinical information. Successful resolution of this challenge holds the potential to enable imaging and clinical data to become harmonized, quality-checked, annotated, and ready to be used in conjunction, in the development of artificial intelligence models and other data-dependent use cases.

METHODS To address this challenge, we introduce medical imaging (MI)-CDM—an extension of the OMOP-CDM specifically designed for registering medical imaging data and curation-related processes. Our modeling choices were the result of iterative numerous discussions among clinical and AI experts to enable the integration of imaging and clinical data in the context of the ProCancer-I project, for answering a set of clinical questions across the prostate cancer's continuum.

RESULTS Our MI-CDM extension has been successfully implemented for the use case of prostate cancer for integrating imaging and curation metadata along with clinical information by using the OMOP-CDM and its oncology extension.

CONCLUSION By using our proposed terminologies and standardized attributes, we demonstrate how diverse imaging modalities can be seamlessly integrated in the future.

ACCOMPANYING CONTENT

 [Data Supplement](#)

Accepted September 29, 2023
Published December 7, 2023

JCO Clin Cancer Inform
7:e2300101
© 2023 by American Society of
Clinical Oncology

Creative Commons Attribution
Non-Commercial No Derivatives
4.0 License

INTRODUCTION

Observational Medical Outcomes Partnership (OMOP)-Common Data Model (CDM)¹ developed by The Observational Health Data Sciences and Informatics (OHDSI) community is one of the most widely used CDMs on a global scale, enabling conducting research on health care data by using standard vocabularies and open-source tools for reproducible science. OMOP-CDM allows the standardization and harmonization

of health care information both on a structural and semantic level, enabling distributed network research and federated analytics.

Currently, creating cohort populations by using electronic phenotypes has been extensively studied and implemented in the OHDSI community. However, creating cohorts through combined clinical and imaging metadata remains a challenge because of the absence of a complete standard for

CONTEXT

Key Objective

This study presents an extension of the Observational Medical Outcomes Partnership (OMOP)-Common Data Model (CDM) framework to incorporate standardized medical imaging attributes and curation processes. Our work aims to enhance the integration of imaging data into clinical research, enabling more accurate cohort discovery and the development of AI models for improved medical decision making.

Knowledge Generated

The proposed model extends the OMOP-CDM to include standardized imaging attributes, facilitating cohort discovery using DICOM metadata for AI model training and quality assurance. Additionally, the model includes curation processes, linking initial and curated images through standardized vocabularies. This approach is demonstrated using prostate cancer imaging data from the ProCancer-I project.

Relevance (F.P.-Y. Lin)

Incorporation of medical imaging attributes within the extension of the OMOP-CDM framework is relevant for both clinical and research informatics applications in oncology. This extension has the potential to facilitate harmonization of metadata, allowing the retrieval of combined clinical information, thereby supporting cohort discovery and the integrated development of decision models that utilize radiology data.*

*Relevance section written by JCO CCI Associate Editor Frank P.-Y. Lin, MB ChB, PhD, FRACP, FAIDH.

image-related information and subsequent image curation processes within the OMOP-CDM. Such curation processes include data quality processes (eg, motion correction, coregistration etc) or any other important annotation and labeling process (eg, labeling anatomical structures, noting specific pathologies through segmentation masks etc). By using high-quality imaging and curation metadata, health care practitioners (often with the support of artificial intelligence [AI] and image processing algorithms) can accurately identify diagnoses, provide guidance for treatment choices, and enhance surgical interventions. Therefore, standardizing imaging and curation-related data in the realm of medical imaging research is imperative.

Although the Digital Imaging and Communications in Medicine (DICOM) standard² for collecting, storing, and transferring medical imaging data could be used for accessing important image acquisition parameters (eg, how the image was acquired, what the field of view was, what the slice thickness was, etc) for cohort discovery and AI model training as well as for quality assurance purposes, unfortunately, it lacks important information required to identify relevant images because there is information that is not standardized in the DICOM metadata. For example, the fact that a series is a T2-weighted axial series is usually registered in the DICOM tag Series Description (0008,103E), which is free text and highly heterogeneous in clinical institutions.

In this study, we take a step forward in extending the OMOP-CDM for imaging data. Our proposed model moves

beyond the preliminary Radiology Extension (R-CDM) proposed by Park et al³ by (1) incorporating standardized imaging attributes for the most important DICOM tags used for cohort discovery and (2) extending the model for registering curation processes on images, which allows for the retrieval of the curated images based on standardized vocabularies enabling the connection between the initial and the resulted images as a way to keep provenance of information. We present our modeling choices, and we demonstrate a use case by using prostate cancer imaging data, acquired from the ProCancer-I European project.⁴

METHODS

The goal of our proposed CDM extension is to enable cohort discovery by using a combination of clinical and imaging metadata for addressing a set of clinical scenarios as these were defined in the context of the ProCancer-I project. ProCancer-I aspires to collect the largest data set of anonymized prostate cancer multiparametric magnetic resonance (mpMR) images worldwide, following the European Union General Data Protection Regulation rules. The clinical scenarios defined in the project range from prostate cancer diagnosis and characterization to prediction of treatment response and occurrence of side effects after treatment.

To address these clinical use cases, experts within the ProCancer-I consortium defined all clinical, imaging, pathology, and follow-up data that needed to be collected. In all use cases, mandatory clinical information accompany the images—including prostate-specific antigen levels, biopsy

records, and/or prostatectomy confirmations of prostate cancer. Moreover, mandatory information regarding the medical images was also defined and aligned with the specific use case to facilitate the optimal development of the proposed AI model.

To generate our model, we followed an approach on the basis of the Stanford seven-step approach for building ontologies.⁵ These steps include the following:

1. Domain and scope determination: We first determined the domain of our model, which is the representation of imaging metadata and image-related curation processes. The model is to be used for the creation of homogeneous and quality-assured cohort populations to be used in the development of AI models.
2. Reusing existing approaches: Before starting to construct our model, we first tried to review existing approaches on the domain and OMOP-CDM extensions. For example, the OMOP-CDM oncology extension⁶ was a natural choice for representing the cancer-related clinical information required by the project. In addition, Park et al³ have proposed an initial version of an R-CDM trying to address the aforementioned challenges; however, it is still not complete and it does not take into consideration curation-related processes on images. This information is critical in developing AI models for supporting research on the detection of tumors and tumor characterization, to name a few, thus assisting in answering diagnostic and/or treatment-related challenges along various cancer types. For the annotation model, we also explored the most important concepts from the caBIG Annotation and Image Markup Project.⁷
3. Collection of important terms: The set of concepts included in our model was the result of numerous iterative discussions between clinical experts and AI model experts. Both defined a set of the most important terms (clinical and imaging) to be included in each use case, and they also defined a set of possible queries to be addressed in order for the resulted cohort population to be used as input to the respective AI models for training and evaluation. Some of these terms were deemed important for quality assurance purposes (eg, slice thickness) and others for creating the cohort populations (eg, the sequence type and the manufacturer of MR images). An example of such a query for defining the cohort for addressing use case 1—for developing a vendor-specific AI model for prostate cancer diagnosis—is *retrieve all T2 axial series from a Siemens scanner with confirmed prostate cancer at pathology (either biopsy or prostatectomy) and their subsequent segmentation masks containing labels to the different prostate gland zones as well as to the lesions observed.*
4. Define entities and entity relationships: A natural choice for representing the domain of our model is to have two different high-level entities, one for representing imaging metadata and one representing the curation metadata. However, since in the DICOM world images are described using the concepts of *studies* and *series* and each *series* has a

number of acquisition parameters depending on the *modality*, we included imaging studies, series, modality, and curation entities.

5. Define the attributes of the entities: Given the defined entities and the important terms as these were defined by the domain experts, we defined the attributes for each of the entities. These correspond mostly to important DICOM metadata and user-defined attributes as these were extracted from the functional requirements of the project (eg, the url through which a series can be accessed, an algorithm used in a curation process etc).
6. Define the restrictions on attributes: For each of the attributes defined, we determined the cardinality, the possible value types, and the standard concepts to be used on the basis of the OMOP standardized vocabularies and other related ontologies whenever these were not enough for representing our data.
7. Instance creation: As a validation of our modeling choices, we created instances of the defined model, by using prostate cancer imaging and clinical data collected from the ProCancer-I project.

The proposed OMOP Medical Image (MI) extension is shown in [Figure 1](#).

In the following sections, we present our model in detail, starting with the imaging metadata model, followed by the imaging curation model, concluding with the standardized terminologies and ontologies used for both imaging and curation metadata.

Imaging Metadata Model

Starting with standardizing the structure of the imaging metadata, we chose to represent the most important acquisition parameters of a DICOM study into two main classes, the *Imaging_Study* table and the *Imaging_Series* table, which contain the most important image metadata as extracted from the DICOM instances.

The *Imaging_Study* table contains the same attributes as the *Radiology_Occurrence* table of the R-CDM proposed by Park et al. Conceptually, when a patient undergoes an imaging procedure, there is a DICOM study with multiple DICOM series that get created. In that case, an *Imaging_Study* instance will be created containing the most important information of the general study module of the DICOM standard.⁸ Note that this is also in accordance with the *Imaging_Study* entity introduced by the FHIR HL7 framework,⁹ therefore enabling future interoperability between the two standards. In addition, we propose the addition of the *procedure_occurrence_id* as a foreign key to the instance generated in the standard OMOP-CDM *Procedure_Occurrence* table, when the patient undergoes an imaging study. This attribute addition is important for mainly two reasons: (1) for associating information about the provider and the clinical site that performed this imaging procedure and the visit occurrence information and (2) and, most importantly, through the *procedure_occurrence_id*, which is connected to the

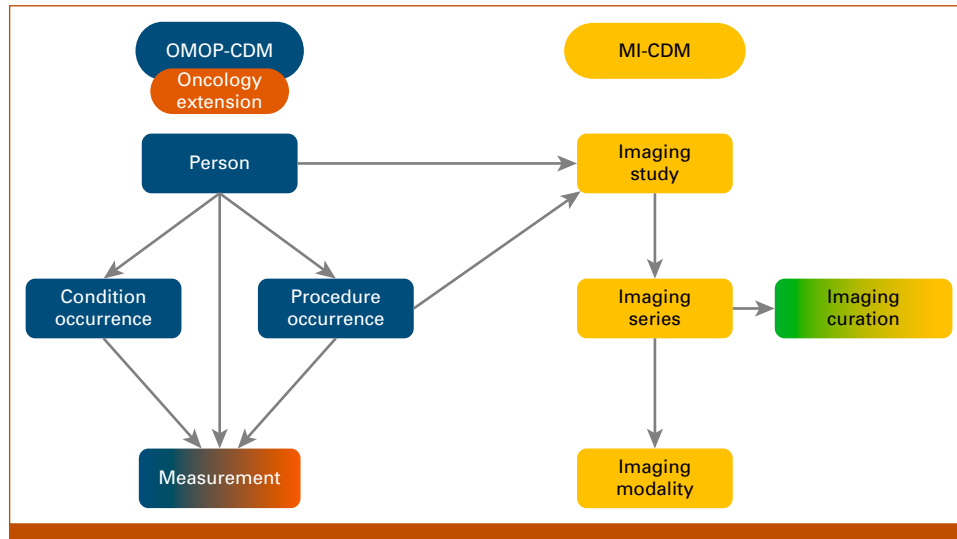


FIG 1. MI-CDM model representation and its connection to the OMOP-CDM and the oncology extension model. CDM, Common Data Model; MI, medical imaging; OMOP, Observational Medical Outcomes Partnership.

measurement_event_id of the OMOP *Measurement* table, information about the cancer modifiers in the Oncology extension can be exploited. Therefore, every cancer modifier, including the clinical TNM staging, size of lesions, location of tumor etc extracted from the medical images, can be inter-linked through the *measurement_event_id*, the *procedure_occurrence_id*, and the *imaging_study_id*.

The definition of the *Imaging_Study* table is shown in Table 1.¹⁰

However, queries corresponding to acquisition parameters of specific DICOM series should also be possible. For example, for the MR series, one should be able to retrieve all the images with slice thickness less than three (eg, for quality purposes), which are T2-weighted axial series. For answering this type of query, it is crucial that some lower-level information (instance-level) about each type of modality is present. However, having a table for registering each DICOM instance/image of a series (as proposed by Park et al³) is redundant as it is not common to develop AI models by querying individual images of series but rather the whole series themselves, we have elevated concepts from the instance level to the series level, for easier discovery of cohort populations.

Therefore, we propose an *Imaging_Series* table for capturing important metadata across all image modalities. Furthermore, we introduce an *Imaging_Modality* table, which stores distinct modality-related information, enabling tailored queries for each modality. For representing this kind of information, we added a set of attributes adhering to the OMOP-CDM conventions.

The full set of attributes of the *Imaging_Series* and the *Imaging_Modality* tables are shown in Table 2.

Image Curation Metadata Model

Besides standardizing imaging metadata, it is also important that a structured representation of the curation processes of images is present for describing and querying medical imaging data. Such a model can enable researchers, clinicians, and other stakeholders to annotate and curate imaging data in a consistent and interoperable manner, which can facilitate data sharing and analysis across different institutions and research projects.

As such, the information to be included in this model relates to image curation processes for enhancing and correcting image-related issues, as well as to image annotation processes such as segmentation of images for labeling specific regions of interest. In the following section, we explore the key components of such a model.

We introduce an extra primary table to our original MI-CDM extension model, named *Imaging_Curation* along with an associated secondary table called *parameter*. Beginning with the *Imaging_Curation* table, its most important attributes include the image curation task (eg, motion correction, coregistration etc), the input series and output series IDs of the curation process (both foreign keys to the *Imaging_Series* table), the curation method, the curation process initiator, Data Supplement metadata relevant to the algorithm used, the software used etc, as a way of keeping provenance of information.

In scenarios involving annotation tasks, apart from the aforementioned metadata, the model incorporates the anatomical site and the image finding/label (eg, an index lesion is being annotated on the peripheral zone of the prostate). Moreover, for the *image_curator_id*, we have included a connection to the *provider* standard OMOP table which designates the radiologist responsible for curating the image.

TABLE 1. Definition of the Imaging_Study Table (logical data model)

Imaging_Study Table			
Attribute	Required	Type	Description
imaging_study_id (PK)	Yes	Integer	Unique ID for each patient study to be created
person_id (FK)	Yes	Integer	Foreign key to the id that identifies the person who underwent the imaging study
procedure_occurrence_id (FK)	Yes	Integer	Foreign key to the id of the imaging procedure that the person underwent
imaging_study_date	Yes	Date	Date when the study was taken
manufacturer_name	No	String	Name of the manufacturing company of the imaging equipment
manufacturer_model	No	String	Model of the imaging equipment
imaging_study_source_uid	No	String	Source identifier of the study (DICOM Study UID)
imaging_study_access_uri	No	String	The access URI of the study, either on a web DICOM server (eg, via the use of the WADO-RS DICOMweb REST API ¹⁰) or on local machine via the path name to the folder containing the study
imaging_study_description_source_value	No	String	Study source description
number_of_images	No	Integer	The number of instances inside all the series within a study
number_of_series	No	Integer	The number of all the series within a study

Abbreviations: DICOM, Digital Imaging and Communications in Medicine; FK, foreign key; PK, primary key; REST API, Representational State Transfer Application Programming Interface; UID, unique identifier; URI, uniform resource identifier; WADO-RS, Web Access to DICOM Objects - Retrieve Studies.

Additionally, a new property called *practice_start_date* has been introduced to the *provider* table, for capturing the years of experience of the curator. Recognizing that the seniority of a radiologist executing a curation process can influence the outcomes of AI models, this property is important for transparency and traceability purposes.

However, the curation processes usually contain multiple steps. As these steps, their execution order and the algorithm parameters used in those, affect the result of the curation processes greatly, it is recommended that we keep all the steps being executed as a way to maintain data provenance and enable traceability and transparency of the AI models to be developed. For this reason, in addition to the *image_curation_algorithm*, which stores the curation step name, we have added an *imaging_curation_parent_id* property for maintaining the order of the different curation process steps. Furthermore, we introduce an additional secondary table called *parameter*, for storing all the parameters used in the different curation process steps. For example, for image motion correction processes, we can store the fact that there was a translation registration curation step, with $\sigma = 0.2$, 10 iterations, and scale factor = 2.

The full set of attributes of the *Imaging_Curation* table along with the *property* table are shown in [Table 3](#).

Standardized Terminologies and Ontologies for Imaging and Curation Metadata

Currently, the most widely used ontology for supporting imaging data integration and interoperability is the RadLex lexicon,¹¹ developed by Radiological Society of North America (RSNA) with the support of the National Institutes of Health.

RadLex is organized into a hierarchy of concepts, with each concept assigned a unique RadLex ID, which can be used to identify and reference concepts in a standardized manner. In addition to RadLex, Semantic DICOM¹² is another well-known framework that extends the DICOM standard by incorporating semantic annotations and additional knowledge into DICOM objects. Although Semantic DICOM is flexible in supporting various ontologies, RadLex was chosen as the primary ontology for imaging metadata standardization because of its radiology-focused coverage.

Nonetheless, RadLex vocabulary is still not a part of the OMOP vocabularies, and therefore, there are no standard concept IDs that map directly to the RadLex IDs. This issue has been partly overcome by the RadLex Playbook,¹³ an effort by RSNA and the Regenstrief Institute (ie, LOINC), which provides mappings between the RadLex and *LOINC Part* concepts, facilitating the use of LOINC within OMOP vocabularies. However, to the best of our knowledge, the OMOP vocabularies have not yet included the *LOINC Part* of the LOINC terminology. Therefore, although the imaging procedures can be mapped using the LOINC terminology (which is included in the OMOP vocabularies), all the information from the RadLex Playbook lexicon cannot be yet used. Additionally, not all the concepts from RadLex have been incorporated into the RadLex Playbook, such as the patient-device orientation (RID10461). We believe that incorporating the RadLex Lexicon into the OMOP standard vocabularies will assist in imaging terminology standardization. In our implementation, we have included the RadLex concepts used in the project, inside the OMOP vocabulary table, as shown in the Results section.

Apart from standardizing imaging-related metadata, it is also critical that training data sets contain standardized curation

TABLE 2. Definition of the Imaging_Series and Imaging_Modality Table

Imaging_Series Table			
Attribute	Required	Type	Description
imaging_series_id (PK)	Yes	Integer	Unique ID for each series within a study to be created
person_id (FK)	Yes	Integer	Foreign key to the id that identifies the person who underwent the imaging study
imaging_study_id (FK)	Yes	Integer	Foreign key to the id that identifies the imaging study that this series belongs to
imaging_series_date	Yes	Date	Date when the series was acquired
imaging_modality_concept_id	Yes	Integer	Radiologic procedure modality
laterality_concept_id	No	Integer	Body site laterality, where applicable
body_region_concept_id	Yes	Integer	Refers to the body parts to be imaged
patient_position_concept_id	No	Integer	A generic descriptor of the patient's anatomic configuration
patient_orientation_concept_id	No	Integer	Orientation of patient relative to an imaging device
series_number	No	Integer	Numeric identifier of the series
pixel_data_characteristics	No	Enum	The characteristics of the image taken, if the image pixel values are based on original or source data or they have been derived from pixel values or other images. Possible values: ORIGINAL, DERIVED
patient_exam_characteristics	No	Enum	The characteristics of the image taken relative to the patient examination, ie, if the image was created as a direct result of the patient examination or after the initial patient examination. Possible values: PRIMARY, SECONDARY
imaging_series_source_uid	No	String	Source identifier of the series (DICOM Series UID)
imaging_series_access_uri	No	String	The access URI of the series, either on a web DICOM server (eg, via the use of the WADO-RS DICOMweb REST API) or on local machine via the path name to the folder containing the series instances
imaging_series_source_description	No	String	The source series description of the imaging series
number_of_images	No	Integer	The total number of images/instances in the imaging series
Imaging_Modality Table			
Attribute	Required	Type	Description
imaging_modality_field_id (PK)	Yes	Integer	Unique key to each modality field being instantiated
imaging_series_id (FK)	Yes	Integer	Foreign key to the ID of each series within a study for which important acquisition parameters are being stored
person_id (FK)	Yes	Integer	Foreign key to the id that identifies the person that underwent the imaging study
imaging_study_id (FK)	No	Integer	Foreign key to the id that identifies the imaging study that this series belongs to
imaging_modality_concept_id	Yes	Integer	Radiologic procedure modality
imaging_modality_field_concept_id	No	Integer	The concept ID of the acquisition parameters relevant to the modality of the series (eg, RID10738 for the MR echo type of the MR modality)
imaging_modality_field_value_as_concept_id	No	Integer	The concept ID of the field value (eg, RID10746 for the spin echo value of the MR echo type)
imaging_modality_field_value_as_number	No	Decimal	The numerical value of the modality field (eg, 0 for the gantry tilt angle [RID12343] in case of a CT modality)
imaging_modality_field_unit_concept_id	No	Integer	Unit concept ID of the modality field (eg, 9,484 for the degree unit of the gantry tilt angle field)
imaging_modality_field_source_concept_id	No	String	The source id of the modality field. It is usually the dicom tag
imaging_modality_field_source_value	No	String	The source name of the modality field, in case it cannot be mapped to the imaging_modality_field_concept_id
imaging_modality_field_value_source_value	No	String	The source value of the modality field, in case it cannot be mapped to the imaging_modality_field_value_as_concept_id or imaging_modality_field_value_as_number

Abbreviations: CT, computed tomography; DICOM, Digital Imaging and Communications in Medicine; FK, foreign key; MR, magnetic resonance; PK, primary key; REST API, Representational State Transfer Application Programming Interface; UID, unique identifier; URI, uniform resource identifier; WADO-RS, Web Access to DICOM Objects - Retrieve Studies.

metadata for developing AI models, which requires the most human effort. Several standardized vocabularies can be used to support the values of our proposed curation metadata model and more specifically, the attributes that refer to the imaging

finding and the anatomic site of the annotation processes. The DICOM standard itself with the DICOM segmentation image module (DICOM-SEG)¹⁴ defines a number of standardized coding schemes and controlled vocabularies for describing

TABLE 3. Structure of the Imaging_Curation and Parameter Table

Imaging_Curation Table			
Attribute	Required	Type	Description
imaging_curation_id (PK)	Yes	Integer	A unique identifier for the curation task
source_imaging_study_id (FK)	Yes	String	The unique identifier for the imaging study that contains the series that is being curated
source_imaging_series_id (FK)	Yes	String	The unique identifier for the imaging series being curated
source_static_imaging_series_id (FK)	No	Integer	The unique identifier for the imaging series used as a static series for coregistration curation processes or other processes requiring a reference series
derived_imaging_series_id	Yes	String	The unique identifier for the derived curated imaging series
imaging_curation_task	Yes	String	The type of the curation process (eg, motion correction, coregistration, annotation)
imaging_curation_datetime	No	DateTime	The date and time the curation task was performed
imaging_curator_id (FK)	No	Integer	The unique identifier for the person who initiated the task as a foreign key to the "Provider" OMOP table
imaging_curation_status	No	String	The current status of the curation, such as final or pending
anatomic_site_concept_id	No	Integer	The anatomic location being annotated (eg, peripheral zone of the prostate gland)
imaging_finding_concept_id	No	String	The imaging observation that is annotated (eg, lesion of the prostate)
imaging_curation_method	No	String	The method used for the curation process, such as manual, automatic, or semiautomatic
imaging_curation_algorithm	No	String	The name of the curation step/algorithm used, if applicable
imaging_curation_software	No	String	The software name and version of the preprocessing/curation tool used
imaging_curation_review_status	No	String	The status of the review process, if applicable
imaging_curation_reviewer_id (FK)	No	Integer	The unique identifier for the person responsible for reviewing the annotation, if applicable as a FK to the Provider table
imaging_curation_review_datetime	No	DateTime	The date and time the curation process was reviewed, if applicable
imaging_curation_parent_id (FK)	No	Integer	The unique identifier for the parent curation instance in case the curation process requires multiple steps

Parameter Table			
Attribute	Required	Type	Description
parameter_id (PK)	Yes	Integer	A unique identifier for the parameter used
imaging_curation_id (FK)	Yes	Integer	The image_curation_id of the curation process for which the algorithm parameter is saved
parameter_name	Yes	String	The parameter name (eg, sigma)
parameter_type	No	String	Parameter type (eg, decimal, integer, etc)
parameter_value	Yes	String	The actual value of the parameter

Abbreviations: FK, foreign key; OMOP, Observational Medical Outcomes Partnership; PK, primary key.

imaging findings and observations. Other relevant ontologies can also be used to support the proposed curation metadata model, such as SNOMED CT and International Classification of Diseases for Oncology, third edition.

Using direct annotations on medical images supported by RadLex coupled with the RadLex descriptors of imaging series offers substantial benefits for groundbreaking radiomics and novel AI-based image analysis solutions. When RadLex-annotated images leverage the full ontology hierarchy, underlying meanings and connections within existing image data sets emerge, even when not explicitly stated earlier. For example, the recognition of a Prostate Imaging Reporting and Data System (PI-RADS) 4 lesion¹⁵ could be automatically inferred, if lesion

characteristics are encoded via annotations using the *image_finding_concept_id* property. For example, according to the PI-RADS assessment, if the following image findings were encoded through RadLex: (RID49501: T2 hypointensity, RID49495: lenticular, RID6059: homogeneous), the fact that this lesion is PI-RADS 4 could be deduced, even if this information is not present in the clinical data.

RESULTS

We assessed the utility of the medical imaging extension, for the use case of prostate cancer in the context of the ProCancer-I project. Through the ProCancer-I infrastructure, we have currently collected 9,822 distinct patients,

corresponding to 69,420 DICOM series and 6,071,355 total DICOM instances. Each patient’s corresponding clinical data have been converted into the OMOP-CDM along with the OMOP oncology extension. All the DICOM metadata from all the series collected have also been converted into the MI-CDM extension. The MI-CDM ddl files for a Postgres database are accessible in the GitHub website.¹⁶ Note that all the patient IDs were inserted into the database after they were fully anonymized per project’s guidelines and best practices. Figure 2 shows an example instance of the

Imaging_Study, the Imaging_Series and the Imaging_Modality tables, whereas Figure 3 shows how imaging curation processes are registered into the Imaging_Curation and Imaging_Series tables.

DISCUSSION

Our MI-CDM extension addresses a significant challenge in health care research—integrating imaging data into the OMOP-CDM. By introducing standardized imaging and

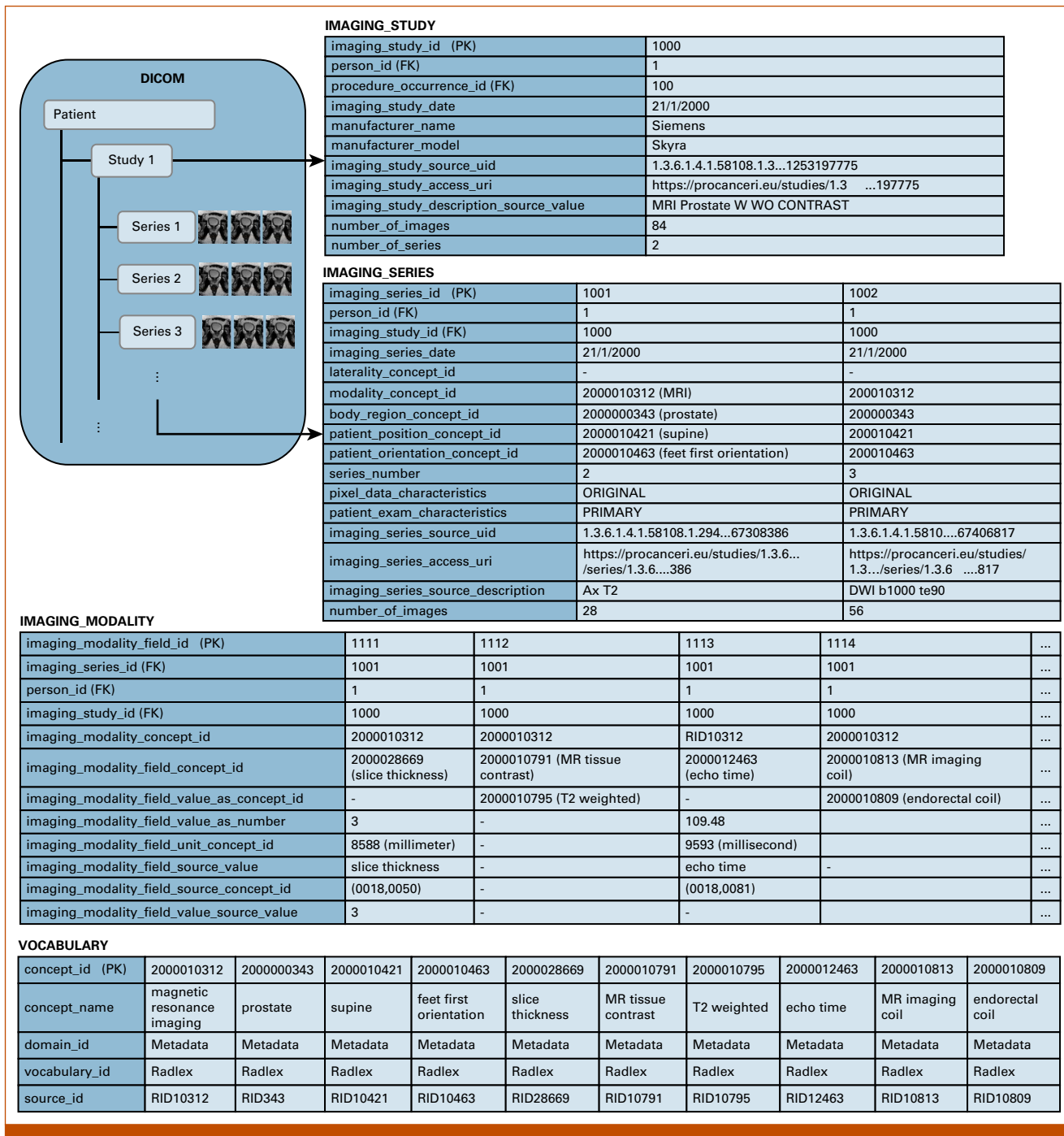


FIG 2. MI-CDM imaging metadata model instantiation for the use case of a patient with an imaging study with two imaging series (T2 axial and DWI) along with standardized metadata for the T2 axial series. CDM, Common Data Model; DWI, diffusion weighted imaging; MI, medical imaging.

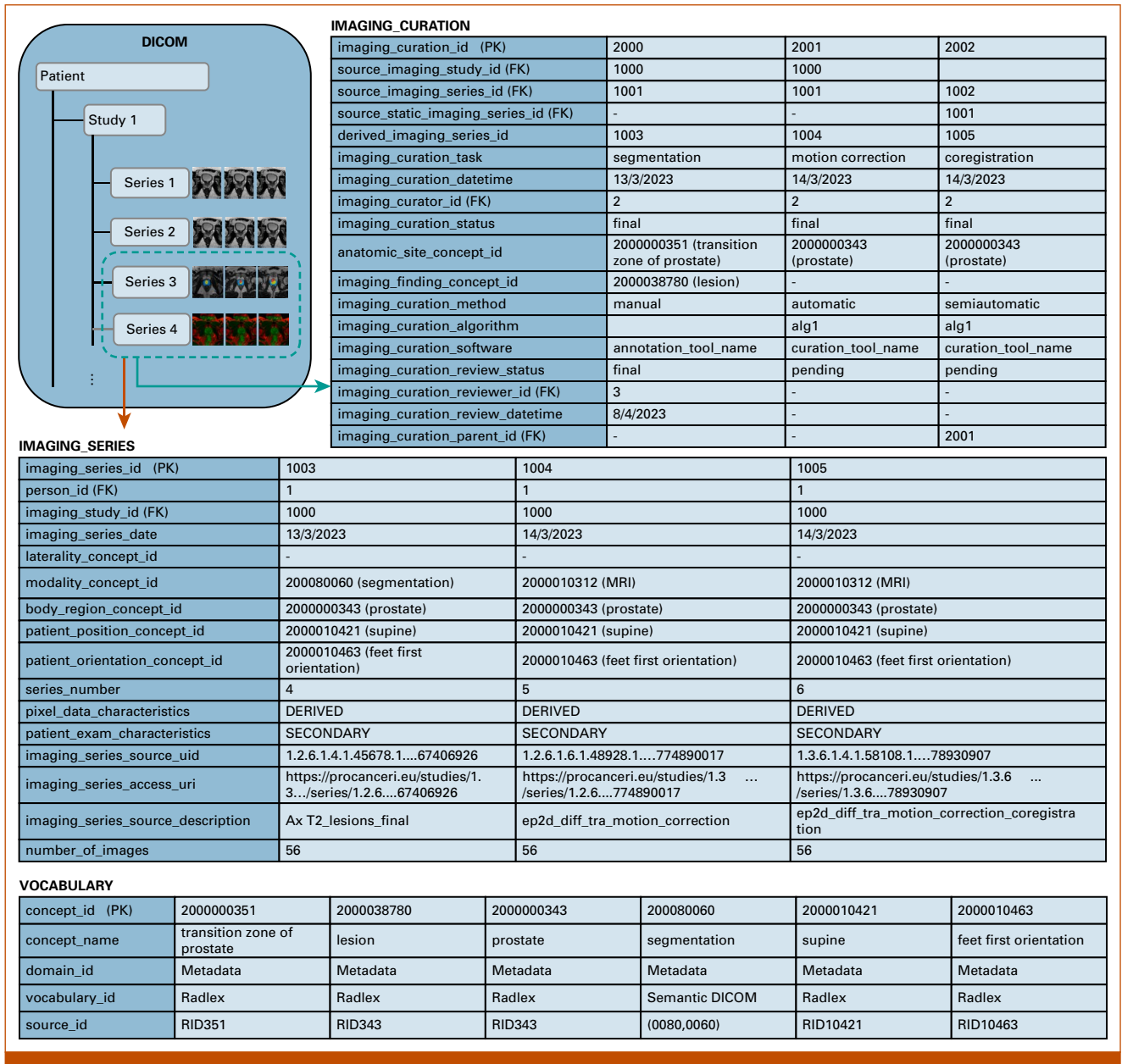


FIG 3. MI-CDM curation metadata model instantiation for the use case of a series that has undergone motion correction and coregistration processes producing two new Imaging_Series instances and a lesion segmentation on the transitional zone, producing a new Imaging_Series instance (SEG modality). CDM, Common Data Model; MI, medical imaging; SEG, segmentation.

curation-related attributes, our model facilitates cohort discovery and AI model advancement. RadLex as a reference ontology assists in harmonizing diverse imaging methods and metadata, leading to a more complete understanding of patient health, improving diagnostic precision and personalized treatment strategies.

However, although the MI-CDM extension offers a significant advancement in the integration of imaging data, a few limitations still exist. First, the extension relies on the existing DICOM standard, which, despite its widespread use, lacks standardized information, necessitating careful mappings, and integration efforts, which may vary across

institutions and settings. Second, our extension was primarily guided by the demands of prostate cancer imaging within the ProCancer-I project. Extending its use to other medical areas and would require validation, potentially introducing complexities in standardizing specific attributes and metadata.

Finally, the MI-CDM's success depends on the adoption and standardization of its terminologies within the broader health care and research community. Collaborative efforts are essential to ensure consistent implementation, harmonization with existing standards, and the continued evolution of the MI-CDM framework.

In conclusion, this study presents MI-CDM, an extension of the OMOP-CDM for imaging data. The extension capitalizes state-of-the-art models in capturing imaging metadata, standardized terminologies, and ontologies and proposes a model able to capture not only information regarding

the images but also information regarding their curation processes. Using the aforementioned model, data for more than 9,800 patients have been acquired, integrated, and homogenized comprising the largest European data set currently available on prostate cancer.

AFFILIATIONS

¹Institute of Computer Science, Foundation of Research and Technology Hellas, Heraklion, Greece

²Department of Surgical Sciences, University of Turin, Turin, Italy

³Radiology Unit, Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Italy

⁴Biomedical Research Institute, Foundation of Research and Technology Hellas, University Campus of Ioannina, Ioannina, Greece

CORRESPONDING AUTHOR

Varvara Kalokyri, PhD, Computational BioMedicine Laboratory, Institute of Computer Science, Foundation of Research and Technology Hellas, Nikolaou Plastira 100, Heraklion 70013, Greece; e-mail: vkalokyri@ics.forth.gr.

SUPPORT

Supported in part by the ProCancer-I European Union's H2020 program under Grant Agreement No. 952159.

AUTHOR CONTRIBUTIONS

Conception and design: Varvara Kalokyri, Haridimos Kondylakis, Katerina Nikiforaki, Ioannis Karatzanis, Nikolaos Tachos, Daniele Regge, Dimitrios I. Fotiadis, Konstantinos Marias, Manolis Tsiknakis

Financial support: Manolis Tsiknakis

Collection and assembly of data: Varvara Kalokyri, Stelios Sfakianakis

Data analysis and interpretation: Varvara Kalokyri, Stelios Sfakianakis, Ioannis Karatzanis, Simone Mazzetti

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated

unless otherwise noted. Relationships are self-held unless noted.

I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Simone Mazzetti

Patents, Royalties, Other Intellectual Property: System for the detection of tumoral masses based on magnetic resonance imaging. The computer aided diagnosis (CAD) system allows automatic detection of suspicious areas of prostate cancer starting from magnetic resonance images. For each suspected area the CAD also provides an indication of the PI-RADS and a probability index related to tumor aggressiveness, creating a structured report

Daniele Regge

Honoraria: Radmetrix (\$10,000 USD or above in a single calendar year)
Consulting or Advisory Role: Health Triage (less than \$10,000 USD in a single calendar year)

Patents, Royalties, Other Intellectual Property: Patent as the inventor of a tool for prostate cancer diagnosis with MRI. The property of the Patent is of the Candiolo Cancer Institute and of the University of Torino. I will receive compensation following the exploitation of the patent by Health Triage (less than \$10,000 USD in a single calendar year)

Travel, Accommodations, Expenses: Bracco Diagnostics (less than \$10,000 USD in a single calendar year)

Dimitrios I. Fotiadis

Employment: PD Neurotechnology Ltd

Leadership: PD Neurotechnology Ltd

Stock and Other Ownership Interests: PD Neurotechnology Ltd

Research Funding: Pfizer

No other potential conflicts of interest were reported.

REFERENCES

- Hripsak G, Duke JD, Shah NH, et al: Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. *Stud Health Technol Inform* 216:574-578, 2015
- Mildenberger P, Eichelberg M, Martin E: Introduction to the DICOM standard. *Eur Radiol* 12:920-927, 2002
- Park C, You SC, Jeon H, et al: Development and validation of the radiology common data model (R-CDM) for the international standardization of medical imaging data. *Yonsei Med J* 63:S74-S83, 2022
- ProCancer-I Project: ProCancer-I: An AI platform integrating imaging data and models, supporting precision care through prostate cancer's continuum. <https://www.proccancer-i.eu/>
- Noy NF, McGuinness DL: Ontology Development 101: A Guide to Creating Your First Ontology. Technical Report. Stanford, CA, Stanford University, 2001
- Belenkaya R, Gurley MJ, Golozar A, et al: Extending the OMOP common data model and standardized vocabularies to support observational cancer research. *JCO Clin Cancer Inform* 5:12-20, 2021
- Channin DS, Mongkolwat P, Kleper V, et al: The caBIG annotation and image Markup project. *J Digit Imaging* 23:217-225, 2010
- Digital Imaging and Communications in Medicine/National Electrical Manufacturers Association: DICOM PS3.3 2023b—Information Object Definitions - C.7.2.1 general study module. https://dicom.nema.org/medical/dicom/current/output/cthtml/part03/sect_C.7.2.html
- Bender D, Sartipi K: HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, Porto, Portugal, 2013, pp 326-331
- Genereaux BW, Dennison DK, Ho K, et al: DICOMweb™: Background and application of the web standard for medical imaging. *J Digit Imaging* 31:321-326, 2018
- Langlotz CP: RadLex: A new method for indexing online educational materials. *Radiographics* 26:1595-1597, 2006
- Van Soest J, Lustberg T, Grittner D, et al: Towards a semantic PACS: Using Semantic Web technology to represent imaging data. *Stud Health Technol Inform* 205:166-170, 2014
- Vreeman DJ, Abhyankar S, Wang KC, et al: The LOINC RSNA radiology playbook—A unified terminology for radiology procedures. *J Am Med Inform Assoc* 25:885-893, 2018

14. Digital Imaging and Communications in Medicine/National Electrical Manufacturers Association: NEMA PS3/ISO 12052, Digital Imaging and Communications in Medicine (DICOM) Standard, PS3.3 2020b Information Object Definitions - C.8.20.2 segmentation image module. https://dicom.nema.org/medical/dicom/2020b/output/chtml/part03/sect_C.8.20.2.html
 15. Puryso AS, Rosenkrantz AB, Turkbey IB, et al: RadioGraphics update: PI-RADS version 2.1—a pictorial update. *Radiographics* 40:E33-E37, 2020
 16. Kalokyri V: MI-CDM. <https://github.com/vkalokyri/MI-CDM>
-