Integrating Clinical Data and Medical Imaging in Lung Cancer: A Feasibility Study Using the OMOP Common Data Model Extension

Sooyoung Yoo, Hyerim Ji, Seok Kim, Leonard Sunwoo, Sowon Jang, Ho-Young Lee

Submitted to: JMIR Medical Informatics
on: April 04, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Integrating Clinical Data and Medical Imaging in Lung Cancer: A Feasibility Study Using the OMOP Common Data Model Extension

Sooyoung Yoo[1] PhD; Hyerim Ji[1, 2] MS; Seok Kim[1] MPH; Leonard Sunwoo[3] MD, PhD; Sowon Jang[3] MD; Ho-Young Lee[1, 4] MD, PhD

[1]Office of eHealth Research and Business Seoul National University Bundang Hospital Seongnam-si KR
[2]Department of Health Science and Technology Graduate School of Convergence Science and Technology Seoul National University Seoul KR
[3]Department of Radiology Seoul National University Bundang Hospital Seongnam-si KR
[4]Department of Nuclear Medicine Seoul National University Bundang Hospital Seongnam-si KR

Corresponding Author:
Sooyoung Yoo PhD
Office of eHealth Research and Business
Seoul National University Bundang Hospital
172, Dolma-ro, Bundang-gu
Seongnam-si
KR

Abstract

Background: Digital transformation, particularly the integration of medical imaging with clinical data, is vital in personalized medicine. The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) standardizes health data. However, integrating medical imaging remains a challenge.

Objective: This study proposes a method for combining medical imaging data with the OMOP CDM to improve multimodal research.

Methods: Our approach included the analysis and selection of Digital Imaging and Communications in Medicine (DICOM) header tags, validation of data formats, and alignment according to the OMOP CDM framework. The FHIR ImagingStudy profile guided our consistency in column naming and definitions. Medical Imaging CDM (MI-CDM), constructed using the entity-attribute-value (EAV) model, facilitates scalable and efficient MI data management. For lung cancer patients diagnosed between 2010 and 2017, we introduced four new tables—IMAGING_STUDY, IMAGING_SERIES, IMAGING_ANNOTATION, and FILEPATH—to standardize various imaging-related data and link to clinical data.

Results: This framework underscores the effectiveness of MI-CDM in enhancing our understanding of lung cancer diagnostics and treatment strategies. The implementation of the MI-CDM tables enabled the structured organization of a comprehensive dataset, including 275,446 IMAGING_STUDY, 5,346,571 IMAGING_SERIES, and 34,449 IMAGING_ANNOTATION records, illustrating the extensive scope and depth of the approach. A scenario-based analysis using actual data from patients with lung cancer underscored the feasibility of our approach. A data quality check applying 44 specific rules confirmed the high integrity of the constructed dataset, with all checks successfully passed, underscoring the reliability of our findings.

Conclusions: These findings indicate that MI-CDM can improve the integration and analysis of medical imaging and clinical data. By addressing the challenges in data standardization and management, our approach contributes toward enhancing diagnostics and treatment strategies. Future research should expand the application of MI-CDM to diverse disease populations and explore its wide-ranging utility for medical conditions.

(JMIR Preprints 04/04/2024:59187)
DOI: https://doi.org/10.2196/preprints.59187

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Abstract

Background: Digital transformation, particularly the integration of medical imaging with clinical data, is vital in personalized medicine. The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) standardizes health data. However, integrating medical imaging remains a challenge.

Objective: This study proposes a method for combining medical imaging data with the OMOP CDM to improve multimodal research.

Methods: Our approach included the analysis and selection of Digital Imaging and Communications in Medicine (DICOM) header tags, validation of data formats, and alignment according to the OMOP CDM framework. The FHIR ImagingStudy profile guided our consistency in column naming and definitions. Imaging CDM (I-CDM), constructed using the entity-attribute-value (EAV) model, facilitates scalable and efficient MI data management. For lung cancer patients diagnosed between 2010 and 2017, we introduced four new tables—*IMAGING_STUDY, IMAGING_SERIES, IMAGING_ANNOTATION, and FILEPATH*—to standardize various imaging-related data and link to clinical data.

Results: This framework underscores the effectiveness of I-CDM in enhancing our understanding of lung cancer diagnostics and treatment strategies. The implementation of the I-CDM tables enabled the structured organization of a comprehensive dataset, including 282,098 *IMAGING_STUDY*, 5,674,425 *IMAGING_SERIES*, and 48,536 *IMAGING_ANNOTATION* records, illustrating the extensive scope and depth of the approach. A scenario-based analysis using actual data from patients with lung cancer underscored the feasibility of our approach. A data quality check applying 44 specific rules confirmed the high integrity of the constructed dataset, with all checks successfully passed, underscoring the reliability of our findings.

Conclusions: These findings indicate that I-CDM can improve the integration and analysis of medical imaging and clinical data. By addressing the challenges in data standardization and management, our approach contributes toward enhancing diagnostics and treatment strategies. Future research should expand the application of I-CDM to diverse disease populations and explore its wide-ranging utility for medical conditions.

Keywords DICOM; OMOP CDM; Lung Cancer; Medical Imaging; Data Integration; Data Quality

Introduction

The accessibility and use of health information in various formats and standards are limited, further limiting the development of advanced data analytics technologies, especially in an era where machine learning and other cutting-edge technologies have become essential for medical research. Integrating these sophisticated analytical tools requires a paradigm shift toward the standardization and harmonization of healthcare data.[1,2] Standardized data structures are not only beneficial but essential for the effective application of machine learning algorithms as they ensure consistent data quality, interoperability, and comprehensive analysis across different healthcare domains. By moving to standardized data formats, we laid the foundation for a more powerful and scalable application of emerging technologies, opening up new possibilities in medical research and patient care.

Several standardization projects and technologies have emerged in response to the demand for integrated approaches.[3,4] Among these, the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) is noteworthy for its advantages in converting diverse sources of data into a consistent format.[5,6] It harmonizes the structure and content of various clinical datasets, facilitating consistent analytical approaches in multi-institutional research. These characteristics ensure high efficiency and accuracy of data interpretation and utilization, thereby enhancing both the quality and pace of research in the rapidly evolving field of medicine.

Digital Imaging and Communications in Medicine (DICOM) is a universal standard for managing, storing, and transmitting medical images, ensuring interoperability and improved exchange of medical image data and associated information between healthcare systems. Recent research has focused on the integrated analysis of DICOM and OMOP CDM to promote accessibility to complex medical imaging data and electronic health records.[7-11] These efforts aim to combine detailed imaging metrics with diverse clinical data to contribute to the development of diagnostic and therapeutic strategies through comprehensive data analysis. However, the data duplication problem caused by constructing an instance-level table and the absence of a table that can store annotation data such as labeling (commonly used in image analysis) are major limitations. These limitations must be addressed to effectively manage the complex characteristics of medical imaging data and perform an integrated analysis with clinical information in the OMOP CDM. As the complexity of medical imaging data and range of DICOM tags increase, effective solutions are required to integrate data seamlessly and consistently.

Lung cancer is the leading cause of cancer-related deaths worldwide and accounts for >20% of all cancer fatalities in South Korea. The etiology, progression, and therapeutic response of lung cancer are intricately linked to a myriad biological and genetic factors. Therefore, a systematic understanding of the characteristics of lung cancer is paramount for its early detection, prevention, and construction of personalized treatment strategies.[12,13] However, this requires an approach that efficiently integrates high-resolution data across various fields.

In this study, we propose a method to integrate medical imaging data with the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), aimed at enhancing multimodal research capabilities. This approach involves converting DICOM metadata and its annotation data to fit within the OMOP CDM framework and subsequently integrating it into a designed Imaging Common Data Model (I-CDM). We applied this integrated framework to a specific cohort of patients with lung cancer and brain metastases to not only test the feasibility and utility of our approach but also to demonstrate its practical application through a series of research scenarios. Additionally, the use of scenarios was intended to showcase use cases that validate the operational functionality of our proposed model within real-world research settings.

Methods

We systematically analyzed and selected the DICOM header tags, verified their data formats, and mapped them to the OMOP CDM framework. To ensure consistency and interoperability in the column naming and definitions, we referenced the FHIR image study profiles, constructed I-CDM table incorporating an entity-attribute-value (EAV) model for scalability, and performed data preprocessing to maintain data integrity. This allowed us to construct and validate a series of scenarios that combined clinical and imaging information from patients with lung cancer using a structured approach while ensuring interoperability. Figure 1 provides a visual overview of the processes employed.

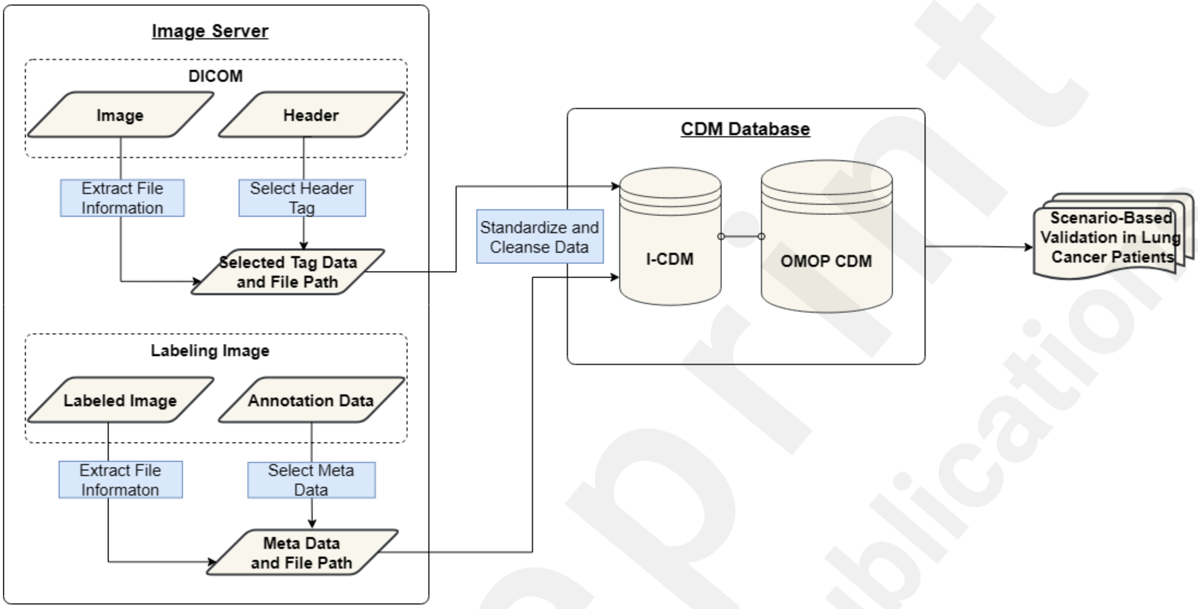


Figure 1. Workflow for I-CDM implementation in lung cancer research

DICOM Header Analysis

We selected preliminary DICOM header tags that are universally applicable across a spectrum of modalities through a systematic procedure. To ensure the relevance and appropriateness of these tags, we sought consultation with radiology specialists.[14-16] While the DICOM standard provides a framework, it does not enforce a uniform data format. This lack of uniformity has led to variations in data formats across medical institutions and modalities. Accordingly, we validated the extractable data formats, ensuring that only data from non-empty DICOM header tags are extracted to maintain data quality and relevance. For data values where no standard concepts were unavailable, we introduced new custom concepts that are used consistently and comprehensively within the OHDSI framework. Concurrently, we examined the FHIR standard's ImagingStudy profile of radiological image data to determine the appropriate table and column names.[17] This approach was adopted to ensure that the selected header tags provided comprehensive coverage with respect to established imaging information standards.

I-CDM Table Modeling

Our database architecture was designed to consolidate a variety of medical imaging data types, including DICOM header tags, study and series-level preprocessing information, and annotation information through labeling. The organization of *IMAGING_STUDY*, *IMAGING_SERIES*, *IMAGING_ANNOTATION*, and *FILEPATH* tables ensured structured and accessible data collection.

The *IMAGING_STUDY* table refers to the CDM *PROCEDURE_OCCURRENCE* table associated with the corresponding imaging-order information.

To understand the significance of series-level analysis in medical imaging with reference to the ImagingStudy profile of the FHIR standard, we modeled the *IMAGING_SERIES* table [18-20] to house values derived from DICOM header tags pertinent to the series. Considering the potential variability in series-specific details owing to distinct imaging equipment or research requirements, we adopted the EAV format. The EAV model provides a data representation framework for the scalable and flexible storage of entities, in which the number and types of attributes (properties and parameters) can vary.[21,22] Using EAV, each attribute-value pair is stored as a separate record, making it easier to populate new data rows without altering the foundational database schema. This model facilitates data expansion without necessitating changes to the foundational table structure.

To observe the emergence and ubiquity of automated labeling tools in radiology, we designed an *IMAGING_ANNOTATION* table [23-26] structured to retain minimal metadata originating from the tools. Similar to the *IMAGING_SERIES* table, we employed an EAV approach to promote extension of the metadata. Finally, in response to the importance of the file size in image-oriented research and AI implementations, we designed a *FILEPATH* table. This database captured fundamental attributes, such as file size, location, and specific format. Figure 2 shows the diagram constructed according to the I-CDM table definition. In APPENDIX.1, we provide detailed definitions for each column in the *IMAGING_STUDY*, *IMAGING_SERIES*, *IMAGING_ANNOTATION*, and *FILEPATH* tables for I-CDM. This appendix elucidates the data format and captures broadly the attributes of the I-CDM framework. In addition, it specifies whether a column is mandatory, ensuring comprehensive documentation and consistency across datasets.

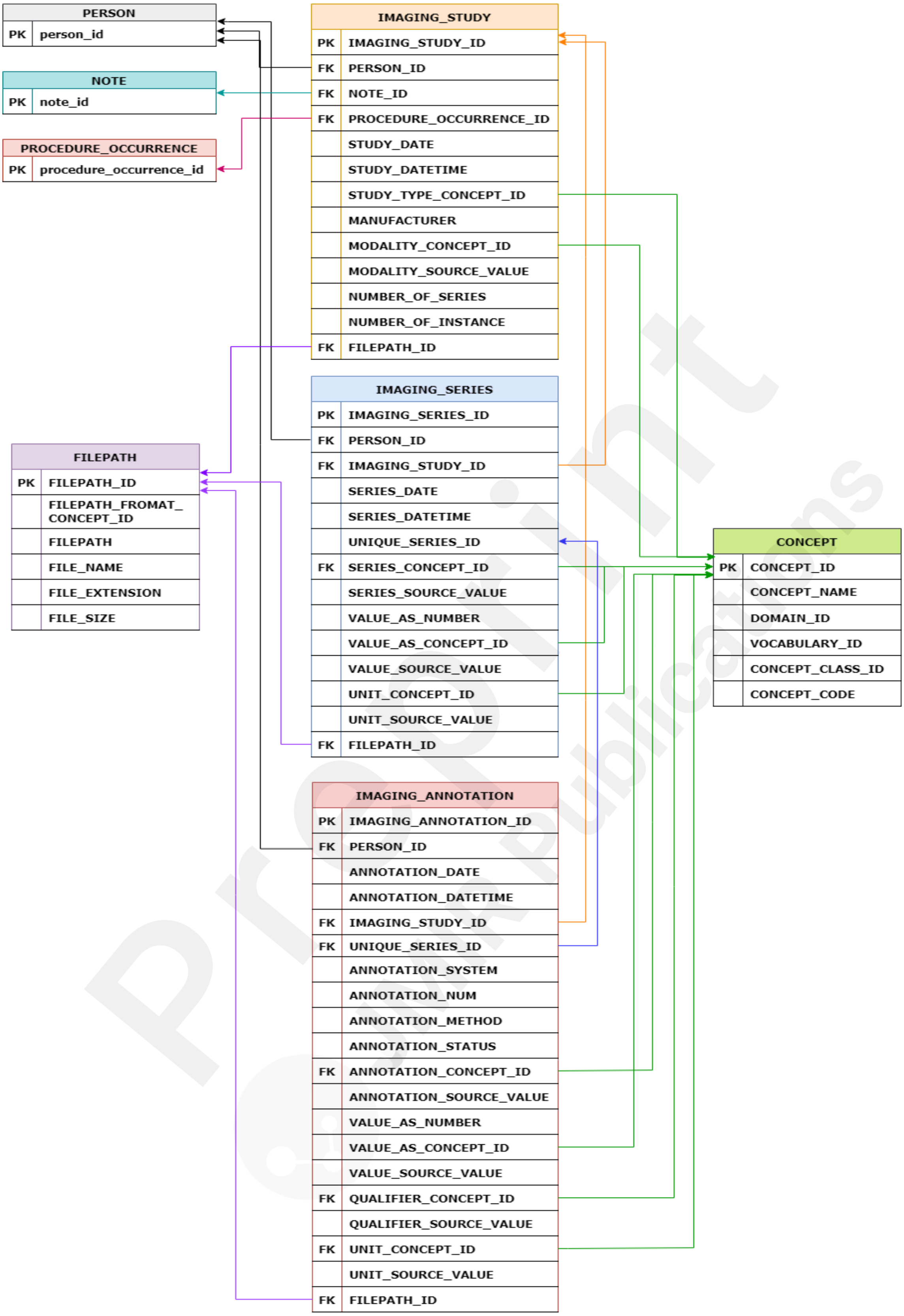


Figure 2. Diagram of I-CDM workflow and attributes

We analyzed and mapped the categorically constructed columns to the corresponding CDM concept IDs where feasible. The modality information was mapped to the concept IDs of the Procedure class in the CDM specifically related to the imaging equipment. This mapping enabled cross-validation using the method attributes of the linked procedure. The body part

category was also aligned with the CDM concept, resulting in terms like "chest" being mapped to the Procedure class as "chest imaging." For the *IMAGING_ANNOTATION* table, attributes such as the labeling plane and area were harmoniously mapped to the standard CDM concept IDs. In instances where mapping to standard CDM codes proved challenging, custom concepts were designated without limiting the classes and domains. Here, the original data extracted from the images were consistently included in the source-value column to ensure data fidelity.
Additionally, in anticipation of RadLex potentially being adopted as a standard vocabulary within the OHDSI framework, we have also suggested additional RadLex mappings for our dataset. RadLex, developed by the Radiological Society of North America (RSNA), is a comprehensive terminology system designed to standardize the names of radiological diagnoses, findings, and procedures. It encompasses a wide range of terms used in medical imaging, making it an invaluable resource for enhancing the comprehensiveness of medical imaging vocabularies within our dataset. Appendix 2 provides a detailed map of the standard terminology used in our I-CDM with their corresponding OMOP CDM concept IDs.

We expanded our methodological framework by developing a Python-based tool, publicly available on GitHub, for automatic conversion and integration of DICOM files into our I-CDM.[27] This tool was designed to work in conjunction with PostgreSQL to effectively create and populate essential tables such as *IMAGING_STUDY*, *IMAGING_SERIES*, and *FILEPATH* directly from specified DICOM file directories. Notably, the tool includes an algorithm to map the extracted DICOM header data systematically to the corresponding CDM concept IDs. This functionality ensures that the medical image data are not only accurately integrated into the I-CDM but also align with the standardized terminologies and classifications of the OMOP CDM.
For practical applications, we chose the NSCLC-Radiomics open dataset from The Cancer Imaging Archive. The NSCLC-Radiomics dataset was utilized solely and only for the purpose of testing our DICOM files to PostgreSQL conversion tool, confirming its functionality with generic DICOM files, and providing a publicly shareable example of the processed output. After reading DICOM files from the NSCLC-Radiomics dataset, our tool methodically constructs I-CDM tables within PostgreSQL, thereby streamlining the data integration process.

Data preprocessing for I-CDM table construction

Before data pre-processing, all personal identifiers were removed to maintain patient confidentiality and protect personal information. Our first step was to ID and categorize the images into a series, ensuring that each series-specific folder contained only pertinent images, thereby maintaining a hierarchical directory structure. The Series Description in a DICOM header often comprises terms and abbreviations that describe the image characteristics. We performed a detailed analysis of the Series Descriptions of selected images to discern imaging attributes, such as the image plane, the presence of enhanced contrast, and designations such as low-dose, T1, or T2, among others. Based on a combination of these criteria, we designed rule-based naming conventions for folders, aiming for descriptive and meaningful names.
Furthermore, we extracted information on the presence of “Black Blood” imaging in MRI scans using DICOM header data, which assisted in preparing data for the construction of the imaging series table.

To construct the imaging annotation table, two radiology experts identified and labeled the lesions on the chest CT and brain MRI scans. Subsequently, we extracted metadata related to the labeled regions, such as area dimensions and characteristics.

Validation of I-CDM for Lung Cancer Studies

This study sought to define a research cohort comprising patients aged 18 years and above with primary lung cancer and structure the metadata of all chest X-ray, chest CT, and brain MRI images using I-CDM. Using the structured data, we aimed to elucidate the unique and major characteristics of patients with lung cancer through analyzing various scenarios.

Scenario 1: Association of Hypertension on Imaging Frequency in Patients with EGFR Mutation-Positive Lung Cancer Receiving Osimertinib

We investigated the association of hypertension on imaging frequency in patients with lung cancer who were prescribed osimertinib, an EGFR tyrosine kinase inhibitor. By comparing the frequency of CT imaging between groups with and without hypertension, this study aimed to determine whether the presence of hypertension affects imaging frequency in patients undergoing osimertinib treatment.

Scenario 2: Correlation Between Ground-Glass Nodules and Solid Tumor Volume in Lung Cancer

Utilizing annotated data from chest CT scans to compare tumor volumes in lung cancer patients with ground-glass nodules (GGN) with those with solid nodules, we aimed to explore the relationship between ground-glass opacity nodules and tumor volume in lung cancer.

Scenario 3: Utilization of Low-Dose CT in Diagnostic Imaging for Lung Cancer

This scenario investigates the number of CT series, each consisting of more than 150 image instances, in patients who have undergone low-dose CT for lung cancer diagnosis. This study aimed to evaluate the adequacy of low-dose CT for providing diagnostic information while minimizing radiation exposure in patients.

Scenario 4: Number of Enhanced T1-Weighted MRI Images with a Slice Thickness of <1 mm in Lung Cancer Patients Diagnosed Under 60 Years of Age

This scenario targets lung cancer patients diagnosed at <60 years of age and involves quantifying the number of enhanced T1-weighted MRI images with a slice thickness of <1 mm. The collected data highlighted the volume of images available for subsequent annotation, and facilitated an in-depth radiological analysis of patient demographics.

Ethics Statement

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of Seoul National University Bundang Hospital (Approval Number: B-2202-738-004, Date: 2022.04.14).

Statistical Analysis

Statistical analyses were performed using R software, version 4.2.2 (The R Foundation, Vienna, Austria). Descriptive statistics were used to summarize the data, including the calculation of means for continuous variables and frequency counts for categorical variables. The data were categorized as necessary to facilitate further analysis.

Results

Conversion of DICOM Data to the OMOP CDM Format: Realization and Integration in the I-CDM Framework

To systematically organize and efficiently manage the extensive collection of imaging data for the cohort of lung cancer patients diagnosed between 2010 and 2017, sourced at Seoul

National University Bundang Hospital, we structured the I-CDM into four basic tables: *IMAGING_STUDY*, *IMAGING_SERIES, IMAGING_ANNOTATION*, and *FILEPATH* (Table 1). The dataset included imaging data from follow-ups in 2003–2021.

Table 1. I-CDM Data Summary for Lung Cancer Cohort (Number of Tables and Records for Data from 2003 to 2021 for Patients Diagnosed with Lung Cancer from 2010 to 2017)

I-CDM Table/Column Name	Record Count (N)	Records with Data (N & %)	Unique Values
IMAGING_STUDY	282,098		
Modality	282,098	282,028 (99.9)	3
Manufacturer	282,098	281,940 (99.9)	265
Number of Series	282,098	282,098 (100)	35
Number of Instance	282,098	282,098 (100)	1,825
IMAGING_SERIES	5,674,425		
Body part Examined	382,517	382,517 (100)	25
Laterality	85,118	85,118 (100)	3
Slice Thickness	411,351	411,351 (100)	1,099
Series Description	654,247	635,208 (100)	12,535
Window Center	685,526	685,526 (100)	29,815
Window Width	685,848	685,848 (100)	31,393
Patient Position	458,770	444,770 (100)	11
Columns	717,154	717,154 (100)	2,020
Rows	717,154	717,154 (100)	2,066
Number of instance	717,169	717,169 (100)	626
BB/NonBB	12,943	12,943 (100)	2
IMAGING_ANNOTATION	48,536		
Annotation System	48,536	48,536 (100)	5
Annotation Text	3,013	2,689 (100)	69
Volume	11,153	11,153 (100)	3,298
Long Axis	31,353	31,333 (100)	159
Surface	3,009	3,009 (100)	2,492
FILEPATH	1,000,361		
File Path	1,000,361	1,000,361 (100)	998,844
File Size	1,000,361	1,000,361 (100)	681,026

The I-CDM categorized 282,098 *IMAGING_STUDY* records, which were systematically linked to OMOP. This link provides an extensive overview of patient imaging trajectories and clinical data. The database contains 5,674,425 image series encompassing 47,381,027 individual image instances. Of the 282,098 records in the dataset, 282,028 records contained information across various modalities, each contained to a corresponding "number of series and instance" columns within the database.

The IMAGING_SERIES data represents a testament to the scale and complexity of the dataset, with the 5,674,425 series illustrating the vast range of radiological examinations included in this framework. A total of 382,517 records detailing the examined body parts underscored the targeted nature of radiological diagnostics. The dataset was also characterized by an array of parameters, with 685,526 records for window centers and 685,848 for window widths. The structural resolution was meticulously captured with 717,154 data points for both the columns and rows, reflecting the intricate images. Each series was contextualized with descriptions recording the purpose and context of the imaging sequence in 654,247 data records. In total, 12,943 images were classified into BB/NonBB categories to indicate the presence or absence of blood–brain barrier contrast, highlighting the usefulness of specialized imaging sequences for detailed neurovascular assessments.

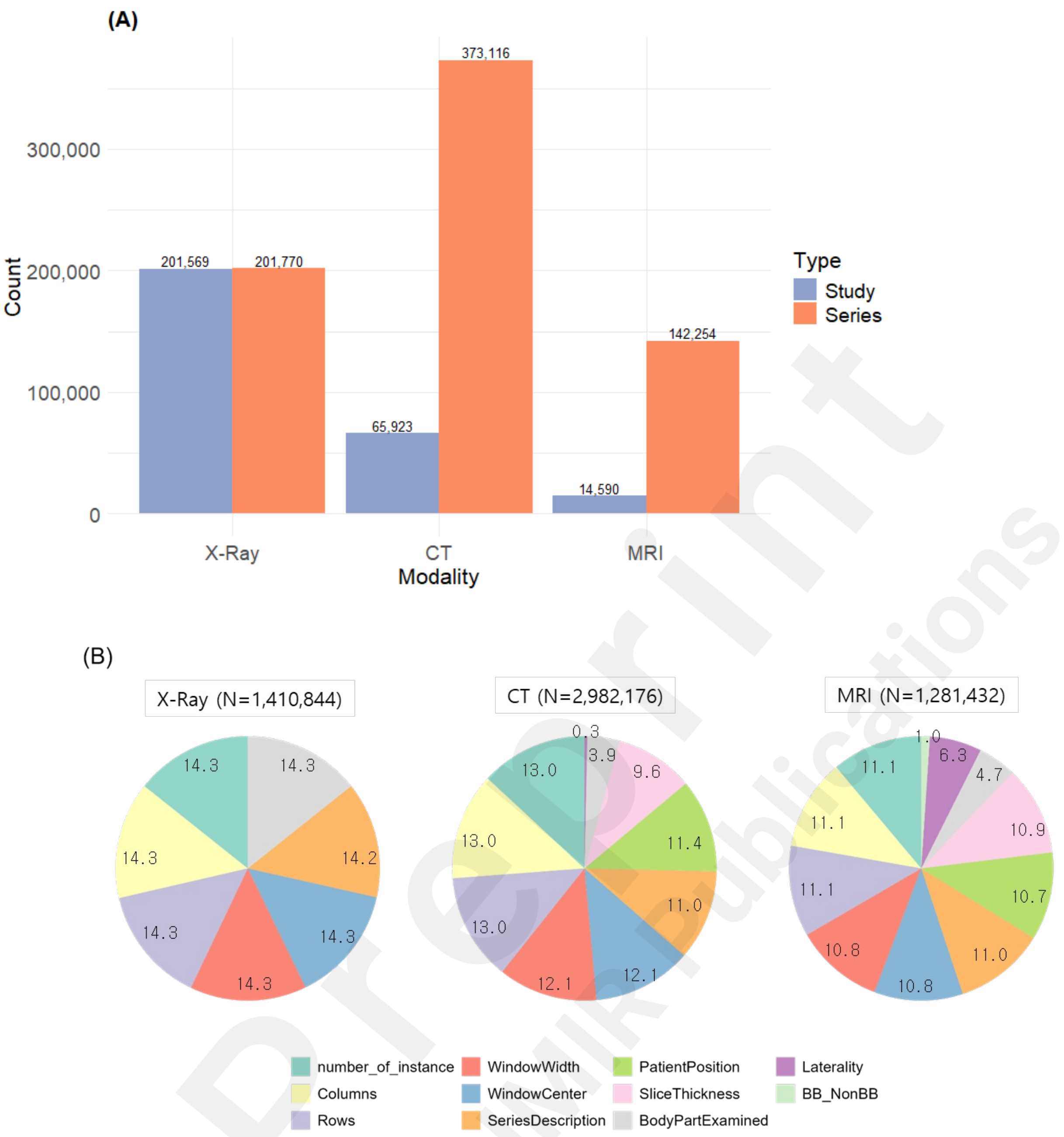


Figure 3. Imaging study and series data distribution: (A) Count data based on modality and type, (B) Percentage data distribution according to modality

Figure 3(A) displays the distribution of DICOM data across different modalities for patients with lung cancer, indicating the number of studies and series for each modality. X-ray examinations usually consist of a single series, whereas MRI and CT scans frequently include multiple series per study to accommodate a variety of imaging sequences. This highlights the detailed and complex nature of lung cancer diagnosis and monitoring. Figure 3(B) displays the structured categorization within the *IMAGING_SERIES* table utilizing an EAV model, where the MRI data comprise 11 categories, including the BB/NonBB distinction. This implies a detailed classification of the MRI data, unlike the X-ray data, which are classified into seven categories representing the variability of the desired parameters for each imaging modality.

The *IMAGING_ANNOTATION* table had 48,536 annotations predominantly sourced from CT and MRI scans. These annotations offer a detailed exploration of the examined regions, with volumetric and long-axis measurements documented. This granular level of detail is critical for the precise characterization and monitoring of diseases supported by a comprehensive understanding of the annotated regions. The FILETATH records within the I-CDM table (totaling 1,000,361) served as a bridge between the CDM tables and actual image file paths, spanning approximately 9.6 terabytes of image data. This illustrates not only the substantial volume of image data, but also the expansive nature of our image repository within the I-CDM framework. In addition, DICOM folders were organized in series, and common imaging characteristics were identified through series descriptions to assign meaningful folder names, further streamlining the data structure for efficient management and retrieval.

Validation of I-CDM Scenarios for Enhanced Imaging and Treatment Classification in Lung Cancer Patients

Scenario 1: Hypertension and Imaging Frequency in Patients Treated with Osimertinib

Among the total cohort of 7,842 patients with lung cancer, 176 (2.24% of the total) prescribed osimertinib were diagnosed with hypertension. In the osimertinib arm, 28 patients (0.36%) had hypertension. The average number of CT scans in patients with and without hypertension was 19.5 and 20.3, respectively, indicating that hypertension did not significantly affect the frequency of CT imaging in lung cancer patients receiving osimertinib treatment. Figure 4 shows this comparison of CT scan frequency over hypertension status among patients treated with osimertinib.

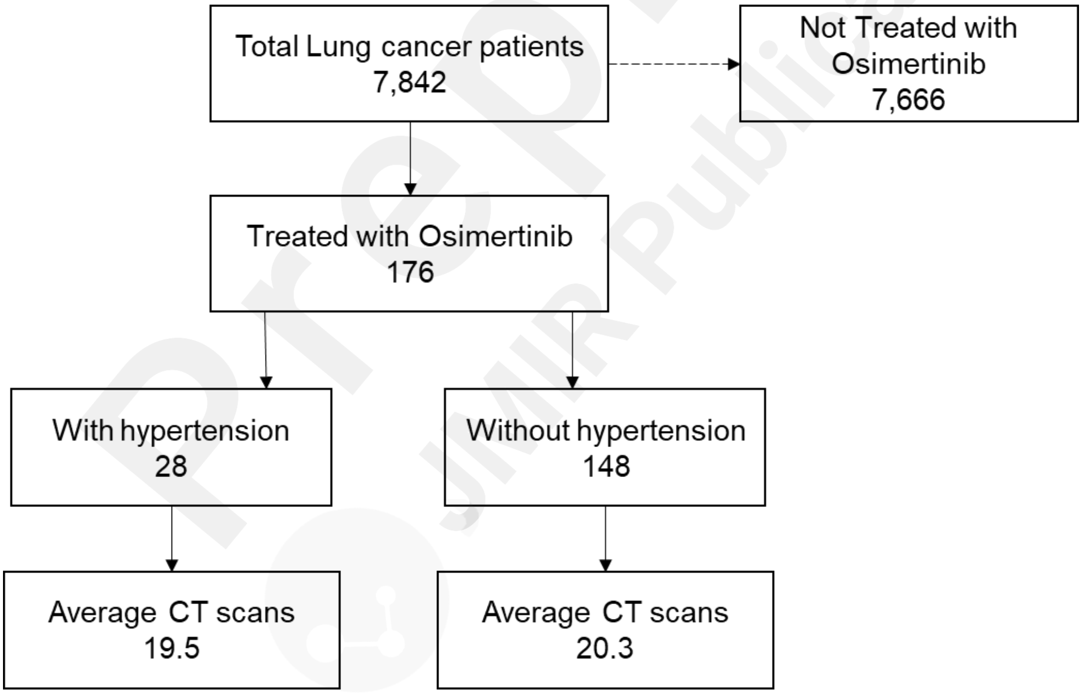


Figure 4. Diagram of CT scan frequency comparison according to hypertension status among osimertinib-treated lung cancer patients.

Scenario 2: Nodule Characterization and Volume Measurement in CT Imaging

We evaluated 1,947 annotated CT scans from 1,929 patients and observed that a significant number of GGNs contained solid components, necessitating labeling of both components within the GGNs. Our comparative analysis focused on the mean volume of solid nodules within GGNs compared with those without GGNs. In total, 673 GGNs were identified in 626 patients, of which 649 were classified as having solid nodules. The average volume of the GGN was 8,135.616 mm^3, whereas the volume of solid nodules within the GGN was 2,578.006 mm^3. In contrast, 1,343 solid nodules without GGN were found in 1,319 patients, with an average volume of 34,712.58 mm^3. This corroborates the findings of previous studies indicating that solid nodules, especially those not associated with GGNs, tended to be more abundant.[27] Our results provide additional evidence supporting these observations on the nodal nature of lung pathologies.

Scenario 3: Utilization of Low-Dose CT and Instance Range

Among the 63,446 CT studies conducted in the lung cancer patient cohort, which included 2,725,899 series, 48,587 were identified as low-dose imaging studies. Of these, 41,336 included >150 instances. The instance count in the low-dose CT images varied significantly, with the smallest and largest series consisting of three and 633 instances, respectively. This highlights the need for low-dose imaging to capture extensive data while minimizing radiation exposure.

Scenario 4: MRI Imaging with 1-mm-Thick Slice and T1 Enhancement

In Scenario 4, which focused on MRIs of 1-mm-thick slices under T1 enhancement, our analysis of 137,566 MRI series identified 31,851 series employing T1-weighting at the specified slice thickness, 5,235 of which were associated with patients aged ≤60 years. This scenario allows the preemptive examination of images to be labeled, integrating image patterns and nodule characteristics with clinical data. This approach not only facilitates identification of the scale of target images to be annotated but also enables precise quantification of the images that meet the specified criteria. Figure 5 shows the categorization of the MRI series using I-CDM, providing a visual summary of the data refined by slice thickness and T1 enhancement and further filtered to include patients aged ≤60 years within the studied patient cohort.

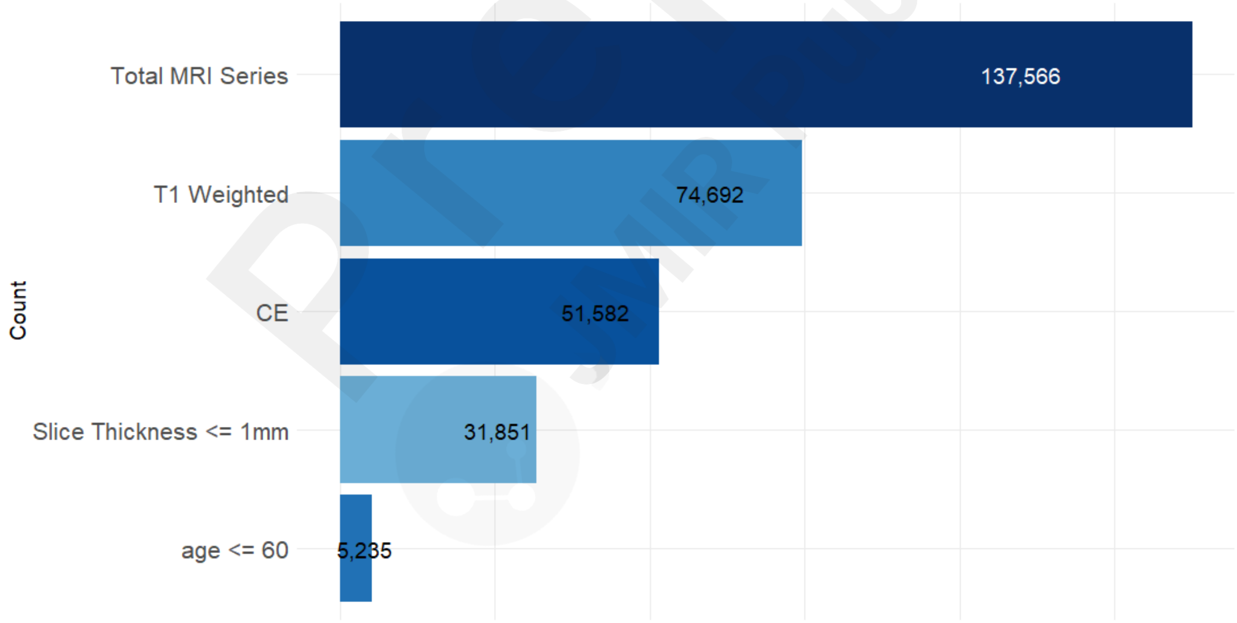


Figure 5. MRI series analysis using I-CDM

I-CDM Data Quality Check

We ensured data quality based on a set of 44 comprehensive data quality (DQ) rules (outlined in Appendix 3), focusing on the Radiation CDM quality assurance framework. These rules encompassed a broad spectrum of checks, including the evaluation of DICOM series and instance counts, data type consistency, and the accuracy of the linkage between the dataset and existing clinical tables within the CDM. All data entries successfully met these criteria, indicating compliance with quality standards. Among these rules, those pertaining to inter-data relationships and outlier detection were instrumental in validating the integrity of the dataset. A selection of these quality checks and their outcomes are presented in Table 2, highlighting the importance of ensuring data quality and integrity within I-CDM.

Table 2. Selected Data Quality Assurance Rules and Outlier Analysis Results from Appendix 3

No	CDM_ TABLE	CONTCEPT_ NAME	Check description	Threshold	Result %, (Error N)
1	IMAGING _STUDY	NUMBER_ OF_SERIES	The NUMBER_OF_SERIES must be equal to the number of series in the IMAGING_SERIES table with the same IMAGING_STUDY_ID. This ensures that the number of series recorded in the IMAGING_STUDY matches the actual series entries in the related table	At least 95% match	PASS 99.9, (202)
2		NUMBER_ OF_ INSTANCE	The NUMBER_OF_INSTANCE must equal the sum of VALUE_AS_NUMBER for entries in the IMAGING_SERIES table where SERIES_CONCEPT_ID equals NUMBER_OF_INSTANCE, under the condition that they are mapped between the two tables. This is to verify that the number of instances (images) reported in the IMAGING_STUDY corresponds to the aggregated count of instances from the series data	At least 95% match	PASS 99.9 (109)
3		NUMBER_ OF_SERIES, NUMBER_ OF_ INSTANCE	The presence of a Rule of NUMBER_OF_SERIES necessitates the presence of a Rule of NUMBER_OF_INSTANCE.	At least 95% match	PASS 99.9, (1)
4	IMAGING _SERIES	SERIES_CONCEPT_ID = SliceThickness	VALUE_AS_NUMBER must exist and be a numeric value for at least 99% of the records	At least 99% of records must have a non-missing	PASS 99.8, (496)
5		SERIES_CONCEPT_ID = Rows	VALUE_AS_NUMBER must exist and be a numeric value for at least 99% of the records	At least 99% of records must have a non-missing	PASS 100
6			Outliers, defined as values beyond the 1st and 99th percentiles, should be reviewed	Outliers should be under 5%	PASS 1.2, (8,952)
7		SERIES_CONCEPT_ID = Columns	VALUE_AS_NUMBER must exist and be a numeric value for at least 99% of the records	At least 99% of records must have a non-missing	PASS :100
8			Outliers, defined as values beyond the 1st and 99th percentiles, should be reviewed	Outliers should be under 5%	PASS 1.0, (7,251)
9		SERIES_CONCEPT_ID = BB/NonBB	Values must be exclusively 'Positive' or 'Negative', ensuring they represent these specific states without including the concept IDs	100% of records must have as one of the specified valid IDs	PASS 100

			45884084 and 45878583		
10	IMAGING_ANNOTATION	ANNOTATION_CONCEPT_ID = Long axis	VALUE_AS_NUMBER must exist and be a numeric value	No missing values for VALUE_AS_NUMBER	PASS 100
			Outliers, defined as values beyond the 1st and 99th percentiles, should be identified and reviewed to ensure they accurately reflect the intended measurements.	Outliers should be under 5%	PASS 0.2, (77)
11		ANNOTATION_CONCEPT_ID = Volume	VALUE_AS_NUMBER must exist and be a numeric value	100% of records must be numeric and non-null	PASS 100
12		ANNOTATION_CONCEPT_ID = annotation_text	VALUE_SOURCE_VALUE must contain a non-empty text value	100% of records must be numeric and non-null	PASS 100
13		ANNOTATION_CONCEPT_ID = surface area	VALUE_AS_NUMBER must exist and be a numeric value	100% of records must be numeric and non-null	PASS 100

Discussion

Principal Results

This study proposes a method for integrating clinical and imaging data using I-CDM. By converting DICOM data into the OMOP CDM format and integrating it into the I-CDM framework, we implemented a systematic approach to efficiently manage medical imaging data. This approach enabled the connection and analysis of clinical and imaging data in different contexts.

Limitations

Limitation of this study is its lack of consideration for the resources required for processing DICOM images and integrating annotation information. To customize and add data according to researcher needs using the EAV model, comprehensive knowledge and expertise on DICOM standards and tags are required [31]. Furthermore, integrating image annotation data within the I-CDM framework not only demands sufficient resources [32-34], but also requires advanced data management strategies to expand the integration and harmonization of datasets from various imaging modalities beyond chest CT, X-ray, and brain MRI [35,36]. Moreover, focusing exclusively on a cohort of patients with lung cancer narrowed the scope of the study. Additionally, in our study, the scenarios were designed to validate the functionality of the proposed model using actual medical data. These scenarios were deliberately simplified to ensure effective management within the capabilities of the implemented I-CDM framework. Future research will benefit from the incorporation of expanded annotation data, enabling more complex analyses, such as longitudinal comparisons of tumor sizes pre- and post-treatment in individual patients.

Comparison with Prior Work

This study diverging from previous Radiology-CDM research, we refined the integration of imaging examination data by linking them with the *PROCEDURE_OCCURRENCE* table, which enables a more efficient analysis through improved data connectivity. Moreover, unlike previous research that relied on RadLex for standard terminology, this study directly mapped

DICOM terms to OMOP CDM standard terminologies. This direct mapping simplifies the process and enables the use of custom codes, thus facilitating a deeper analytical integration of clinical and imaging data. And this study takes a distinct approach compared with recent I-CDM studies [29,30]. Our method enhances the analytical scope by facilitating the storage and management of annotation information. This ensures that imaging-related data, including annotations, can be comprehensively managed within the I-CDM framework. By utilizing the EAV model, our study introduced flexibility in managing various data types and structures, rendering our approach adaptable to evolving research needs and data characteristics. Consequently, it exhibits good flexibility and adaptability, especially in research requiring integrated analysis of clinical and imaging data. Considering the file sizes associated with imaging data, effective file management is essential. Our study used the *FILEPATH* table to connect I-CDM with the original imaging data, including file extension and size information, to ensure quick access to file details and facilitate efficient management.

Scalability and applicability of the I-CDM

The lung cancer cohort in this study was initially utilized to validate the functionality of the proposed I-CDM tables using actual medical image data. In future studies, the model is not only adaptable to lung cancer but also designed to accommodate a wider spectrum of medical conditions, including various tumor types and cardiovascular diseases. By leveraging OMOP CDM's standard vocabulary for 'modality_concept_id' and 'body part examined' ('value_as_concept_id'), the model can be broadly adapted to accommodate various diseases or different settings beyond lung cancer. This adaptability ensures that any additional data items users might require can be seamlessly integrated by aligning with OMOP CDM standard vocabulary, underlining the framework's potential for broad application across diverse medical data and settings. "Furthermore, while RadLex is extensively used in medical imaging vocabularies, it is not yet included as a standard vocabulary in the OHDSI framework. Even if RadLex were incorporated, it would not cover all concepts related to imaging. Therefore, we had to consider various vocabularies to ensure comprehensive coverage. Recognizing this, we aimed to build the I-CDM by maximizing the use of existing OHDSI vocabularies according to OHDSI principles, rather than proposing new vocabularies. We have proactively suggested mapping terms compatible with RadLex within our study wherever possible. In the new scenario, the principle for term selection involves mapping the standard concept to the granularity level of the source data. This is achieved by selecting the term from the standard vocabulary that most accurately represents the clinical meaning. In addition to the features we have currently mapped, our study focused on lung cancer, but for other diseases, there are important concepts in imaging that should be considered. For instance:

- Ultrasound Image Tags: Commonly used tags include "Transducer Frequency," "Gain," and "Depth of Field," which are critical for analyzing the quality and characteristics of ultrasound images.
- Spine X-Ray Tags: Relevant tags such as "KVP" (Kilovoltage Peak), "Exposure Time," and "Focal Spot Size" are essential for understanding the technical parameters that affect image quality.
- Body Part Imaging Concepts: Terms like "Entire Thorax," "Entire Liver," and "Entire Pelvis" are crucial for precisely describing the anatomical region being imaged, which can vary significantly depending on the disease or condition being studied.

These examples ensure that the I-CDM framework is adaptable, capable of integrating a wide range of imaging data characteristics and supporting diverse medical conditions and research scenarios."

Conclusions

This study implemented a systematic approach for the efficient management of medical imaging data, achieving integration of clinical and imaging data through the development of the I-CDM framework and the conversion of DICOM data into the OMOP CDM format. Future efforts should strive to broaden the application of the I-CDM framework to encompass various disease populations and include diverse imaging techniques for different body parts, such as abdominal CT, spine MRI, and liver MRI, thereby enhancing its applicability. Expanding its scope to incorporate these imaging modalities is crucial for conducting more comprehensive investigations into the utility of merging clinical and imaging data across different health conditions.

Acknowledgements

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI22C0471).

Conflicts of Interest

None declared

Abbreviations

OMOP CDM: Observational Medical Outcomes Partnership Common Data Model
DICOM: Digital Imaging and Communications in Medicine
I-CDM: Imaging CDM
EAV: entity-attribute-value
DQ: data quality

Reference

1. Esteva A, Kuprel B, Novoa RA, et al. A guide to deep learning in healthcare. Nature medicine; 2019(25):24-29. PMID:1234567
2. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature biomedical engineering; 2018(2):719-731. doi:10.1136/bmj.331.7529.1391
3. Rho MJ, Kim HS, Chung K, et al. Common data model for decision support system of adverse drug reaction to extract knowledge from multi-center database. Information Technology and Management; 2016(17):57-66. PMID:12345678
4. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. Journal of the American Medical Informatics Association; 2015(22):553-564. doi:10.1136/bmj.331.7529.1392
5. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics; 2015(216):574. PMID:12345679
6. Lim JE, Kim D, Lee JE, et al. Association between dyslipidemia and asthma in children: a systematic review and multicenter cohort study using a common data model. Clinical and Experimental Pediatrics; 2023(66):357. PMID:12345680

7. Park CH, Kim TH, Seo JB, et al. Development and validation of the radiology common data model (R-CDM) for the international standardization of medical imaging data. Yonsei Medical Journal; 2022(63) Suppl:S74. PMID:12345681
8. Kim TH, Lee YG, Kang DK, et al. Development and validation of a management system and dataset quality assessment tool for the Radiology Common Data Model (R_CDM): A case study in liver disease. International Journal of Medical Informatics; 2022(162):104759. PMID:12345682
9. Malik B, Kuo PH, Lo YC, et al. Development of the Medical Imaging Extension for OMOP-CDM. ohdsi.org; 2022 OMOP CDM symposium. [Medline]
10. Praeta J, Scherer M, Smeets D, et al. Application of the R-CDM extension to capture metadata and features extracted from quantitative brain MRI and CT data. 2023 OMOP CDM symposium. [Medline]
11. Kwon EY, Lee JH, Kim YJ, et al. Development of common data module extension for radiology data (R-CDM): A pilot study to predict outcome of liver cirrhosis with using portal phase abdominal computed tomography data. European Congress of Radiology-ECR; 2019. [Medline]
12. Tsui DC, Camidge DR, Rusthoven CG. Managing central nervous system spread of lung cancer: the state of the art. Journal of Clinical Oncology; 2022(40):642-660. PMID:12345683
13. Eaton KD. Lung cancer: Translational and emerging therapies. 2008. ISBN:0195176332
14. Mildenberger P, Eichelberg M, Martin E. Introduction to the DICOM standard; European radiology; 2002(12):920-927.
15. Bidgood Jr WD, Horii SC, Prior FW, et al. Understanding and using DICOM, the data interchange standard for biomedical imaging; Journal of the American Medical Informatics Association; 1997(4):199-212.
16. Haripriya P, Porkodi R. A survey paper on data mining techniques and challenges in distributed DICOM; International Journal of Advanced Research in Computer and Communication Engineering; 2016(5):741-747.
17. Kamel PI, Nagy PG. Patient-centered radiology with FHIR: an introduction to the use of FHIR to offer radiology a clinically integrated platform; Journal of digital imaging; 2018(31):327-333.
18. Tournier J-Donald, Smith RE, Raffelt D, et al. MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation; Neuroimage; 2019(202):116137.
19. Aiello M, Cavaliere C, D'Albore A, et al. How does DICOM support big data management? Investigating its use in medical imaging community; Insights into Imaging; 2021(12):1-21.
20. Onken M, Riesmeier J, Bennett A, et al. Digital imaging and communications in medicine; In: Biomedical Image Processing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 427-454.
21. Lelong R, Léon C, Nicol P, et al. Querying EHRs with a semantic and entity-oriented query language; Stud Health Technol Inform; 2017(235):121-125.
22. Nadkarni PM, Marenco L, Chen R, et al. QAV: querying entity-attribute-value metadata in a biomedical database; Computer methods and programs in biomedicine; 1997(53):93-103.
23. Morozov SP, Pekar V, Gubern-Merida A, et al. A simplified cluster model and a tool adapted for collaborative labeling of lung cancer CT scans; Computer Methods and Programs in Biomedicine; 2021(206):106111.
24. Diaz-Pinto A, Ravikumar N, Attar R, et al. Monai label: A framework for ai-assisted interactive labeling of 3d medical images; arXiv preprint arXiv:2203.12362; 2022.

25. Philbrick KA, Weston AD, Akkus Z, et al. RIL-contour: a medical imaging dataset annotation tool for and with deep learning; Journal of digital imaging; 2019(32):571-581.
26. Lösle PD, Schwier M, Skibbe H, et al. Introducing Biomedisa as an open-source online platform for biomedical image segmentation; Nature communications; 2020(11):5577.
27. RadiologyCDM. Available from: https://github.com/HIRC-SNUBH/ImagingCDM
28. Chu ZG, Kim JH, Yoon SH, et al. Primary solid lung cancerous nodules with different sizes: computed tomography features and their variations; BMC cancer; 2019(19):1-8.
29. Kalokyri V, Tsiknakis M, Marias K, et al. MI-Common Data Model: Extending Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) for Registering Medical Imaging Metadata and Subsequent Curation Processes; JCO Clinical Cancer Informatics; 2023(7):e2300101.
30. Park WY, Lee KH, Kim JY, et al. Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension; Journal of Imaging Informatics in Medicine; 2024(1):1-10.
31. Arvanitis TN, Demetriou G, Kolios P, et al. A hybrid EAV-Relational model for consistent and scalable capture of clinical research data; Integrating Information Technology and Management for Quality of Care; 2014(202):32.
32. Gu Y, Lu H, Niu Y, et al. Reliable label-efficient learning for biomedical image recognition; IEEE Transactions on Biomedical Engineering; 2018(66):2423-2432.
33. Wu Y, Zhou Z, Wu W, et al. OneSeg: Self-learning and One-shot Learning based Single-slice Annotation for 3D Medical Image Segmentation; arXiv preprint arXiv:2309.13671; 2023.
34. Dimitrovski I, Kocev D, Kitanovski I, et al. Hierarchical annotation of medical images; Pattern Recognition; 2011(44):2436-2449.
35. Martinez-Garcia M, Hernández-Lemus E. Data integration challenges for machine learning in precision medicine; Frontiers in medicine; 2022(8):784455.
36. Parciak M, Kleine-Brueggeney H, Raschke F, et al. FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital; BMC Medical Informatics and Decision Making; 2023(23):94.

Supplementary Files

Figures

Workflow for MI-CDM implementation in lung cancer research.

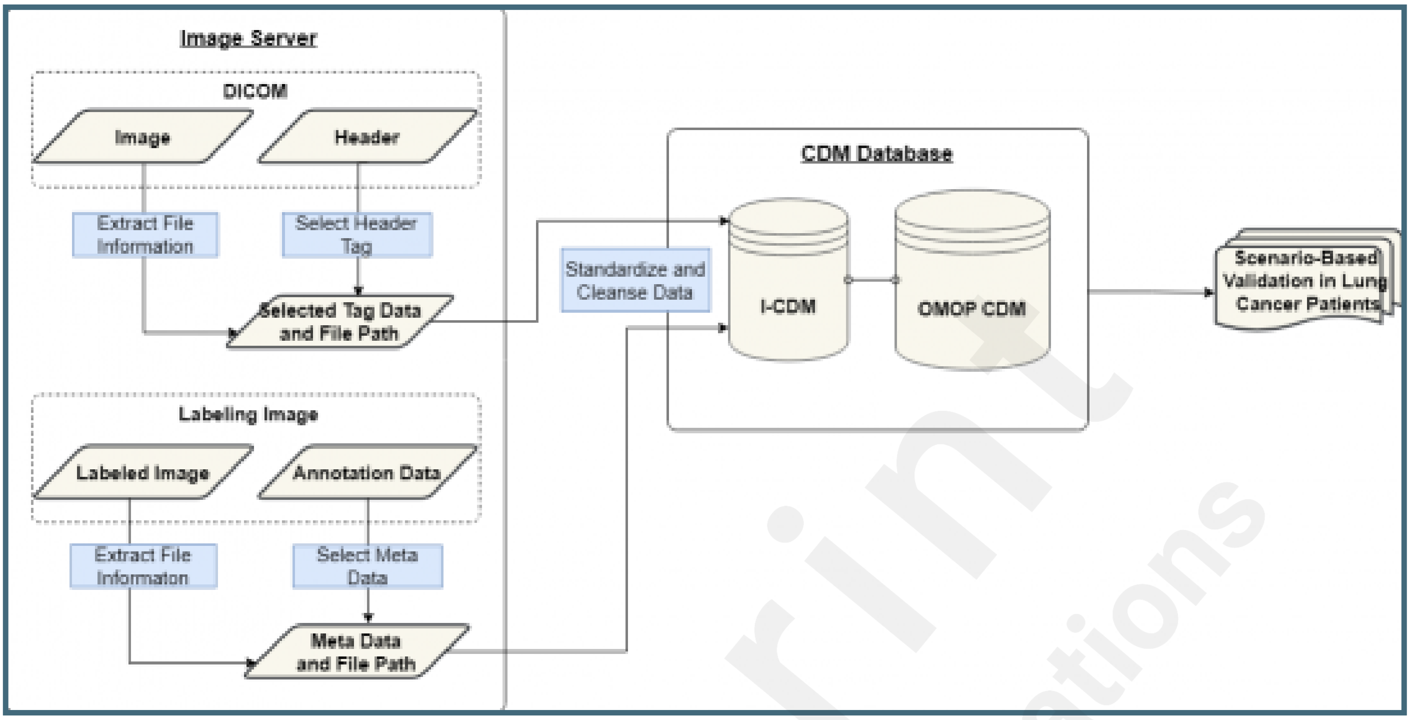
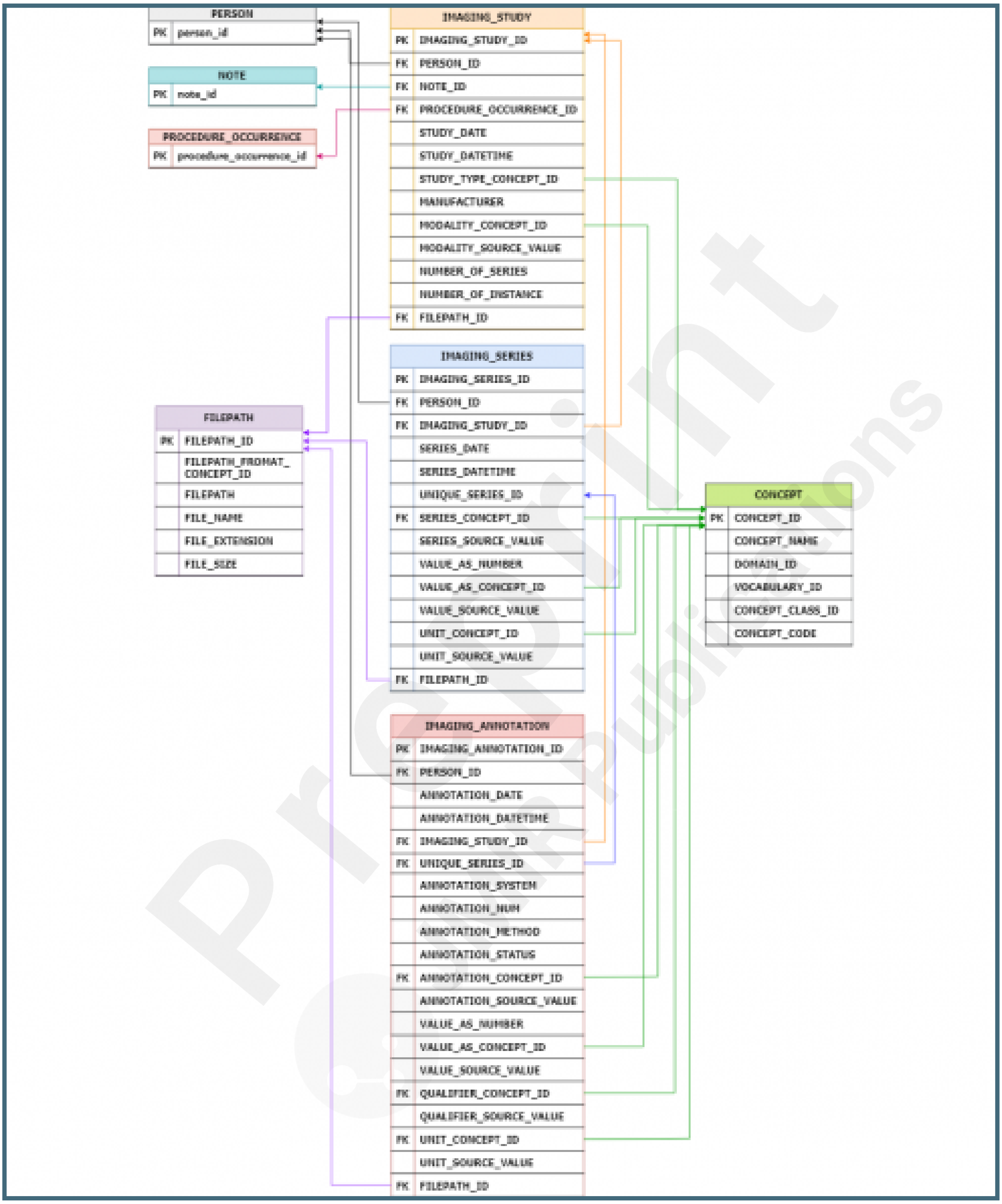


Diagram of MI-CDM workflow and attributes.



Imaging study and series data distribution: (A) Count data based on modality and type, (B) Percentage data distribution according to modality.

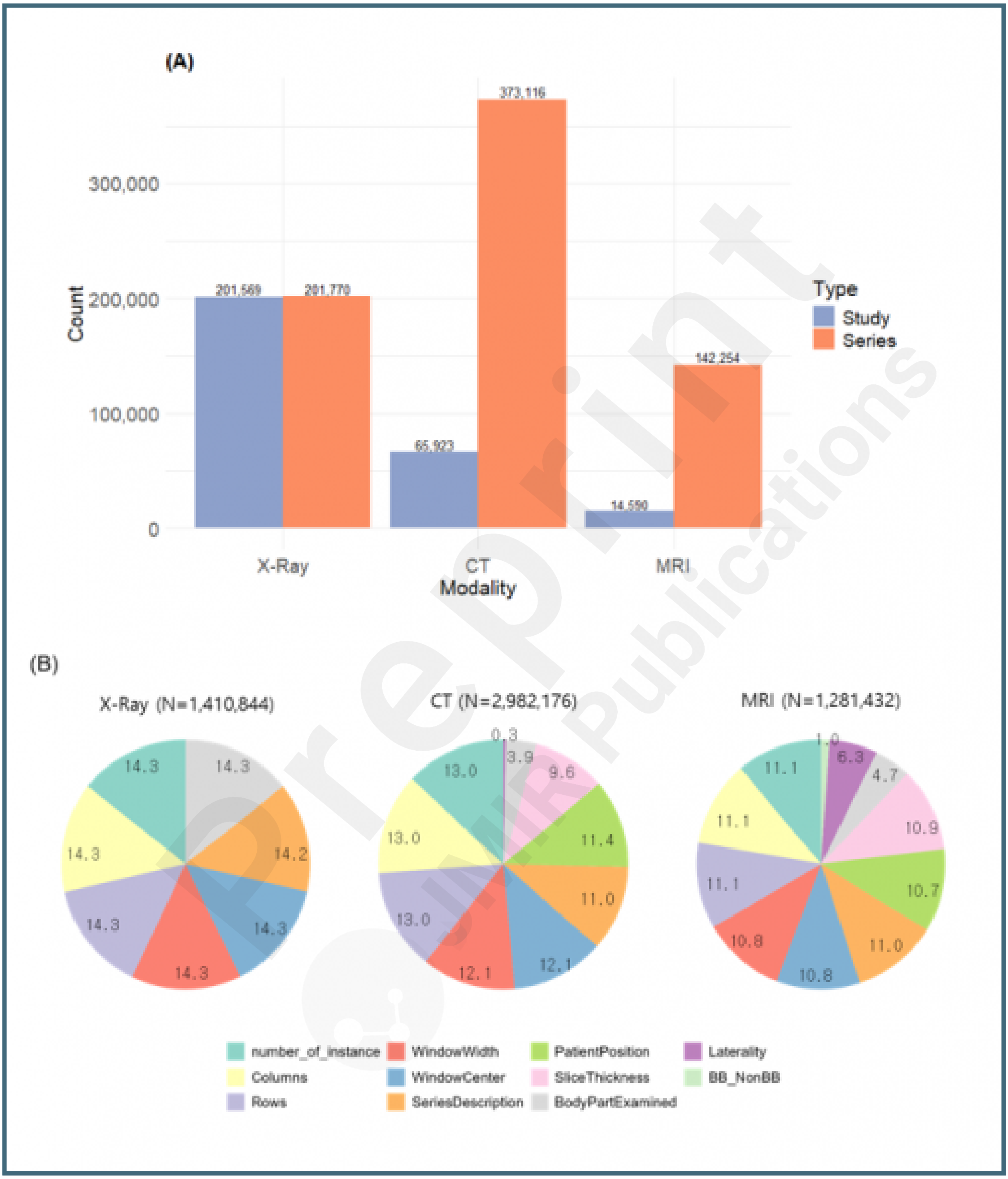
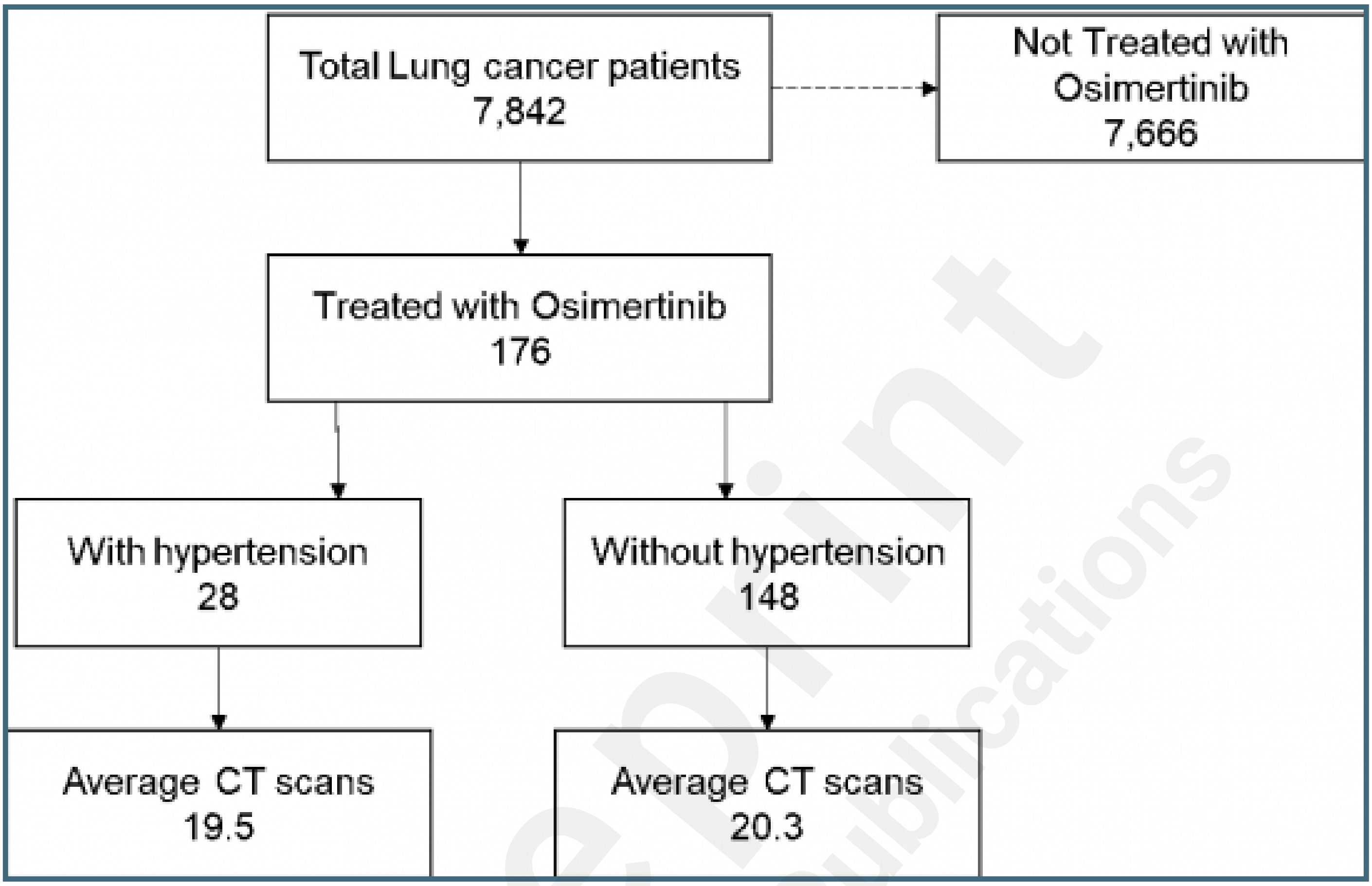
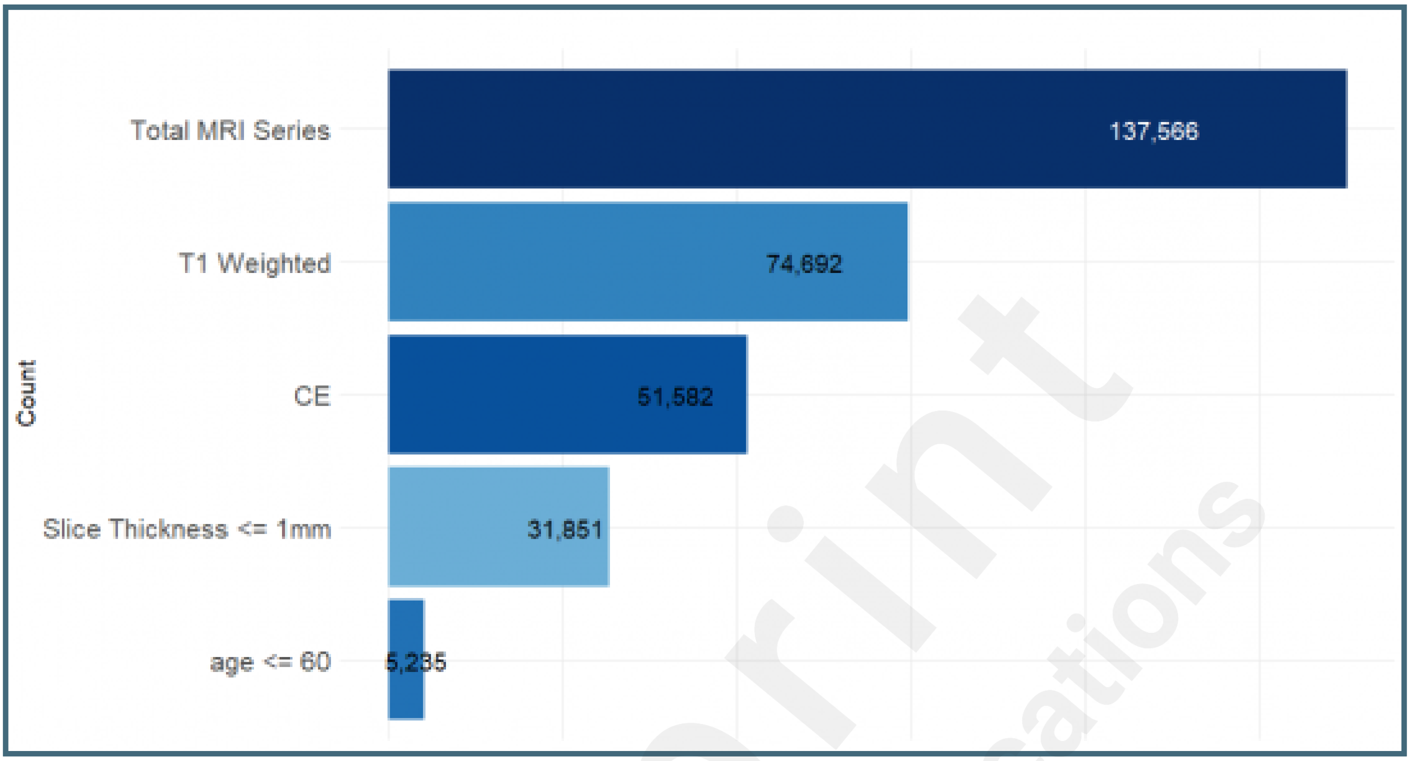


Diagram of CT scan frequency comparison according to hypertension status among osimertinib-treated lung cancer patients.



MRI series analysis using MI-CDM.



Multimedia Appendixes

Radiology CDM Table.
URL: http://asset.jmir.pub/assets/7b8c40db5a9da26a8c291bd26aa2497a.docx

Table. I-CDM mapping OMOP CDM Concept ID.
URL: http://asset.jmir.pub/assets/5acd4fa41c7edadd56b4ab696895834c.docx

Data Quality Check Rule and Result.
URL: http://asset.jmir.pub/assets/87842af737d6db5b075c09fdfc1c738e.docx

Use Case for Integrating I-CDM:Detailed Schema Overview.
URL: http://asset.jmir.pub/assets/606d17e0564ced3cc6149d2733ab132e.png

Comparison of MI-CDM and I-CDM.
URL: http://asset.jmir.pub/assets/37d33609d456d900bea96f4aa8d4e2c5.docx