Original Article
Yonsei Med J 2022 Jan;63 Suppl:S74-83
https://doi.org/10.3349/ymj.2022.63.S74
Yonsei Medical Journal
YMJ
pISSN: 0513-5796 · eISSN: 1976-2437

Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data

ChulHyoung Park[1*], Seng Chan You[2*], Hokyun Jeon[1], Chang Won Jeong[3], Jin Wook Choi[4], and Rae Woong Park[1,5]

[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon;
[2]Department of Preventive Medicine, Yonsei University College of Medicine, Seoul;
[3]Medical Convergence Research Center, Wonkwang University, Iksan;
[4]Department of Radiology, Ajou University Medical Center, Suwon;
[5]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea.

Purpose: Digital Imaging and Communications in Medicine (DICOM), a standard file format for medical imaging data, contains metadata describing each file. However, metadata are often incomplete, and there is no standardized format for recording metadata, leading to inefficiency during the metadata-based data retrieval process. Here, we propose a novel standardization method for DICOM metadata termed the Radiology Common Data Model (R-CDM).

Materials and Methods: R-CDM was designed to be compatible with Health Level Seven International (HL7)/Fast Healthcare Interoperability Resources (FHIR) and linked with the Observational Medical Outcomes Partnership (OMOP)-CDM to achieve a seamless link between clinical data and medical imaging data. The terminology system was standardized using the RadLex playbook, a comprehensive lexicon of radiology. As a proof of concept, the R-CDM conversion process was conducted with 41.7 TB of data from the Ajou University Hospital. The R-CDM database visualizer was developed to visualize the main characteristics of the R-CDM database.

Results: Information from 2801360 cases and 87203226 DICOM files was organized into two tables constituting the R-CDM. Information on imaging device and image resolution was recorded with more than 99.9% accuracy. Furthermore, OMOP-CDM and R-CDM were linked to efficiently extract specific types of images from specific patient cohorts.

Conclusion: R-CDM standardizes the structure and terminology for recording medical imaging data to eliminate incomplete and unstandardized information. Successful standardization was achieved by the extract, transform, and load process and image classifier. We hope that the R-CDM will contribute to deep learning research in the medical imaging field by enabling the securement of large-scale medical imaging data from multinational institutions.

Key Words: Metadata, standardization, radiology information system

Received: September 8, 2021 **Revised:** October 28, 2021
Accepted: October 31, 2021
Corresponding author: Rae Woong Park, MD, PhD, Department of Biomedical Informatics, Ajou University School of Medicine, 164 World cup-ro, Yeongtong-gu, Suwon 16499, Korea.
Tel: 82-31-219-4471, Fax: 82-31-219-4472, E-mail: veritas@ajou.ac.kr

*ChulHyoung Park and Seng Chan You contributed equally to this work.
•The authors have no potential conflicts of interest to disclose.

INTRODUCTION

Together with recent advances in artificial intelligence (AI), current deep learning methods, especially convolutional neural networks, have proven to match or even surpass humans in several specific radiological tasks.[1] The application and verification of deep learning in medicine have been primarily led by the radiology community, because of the relatively large and standardized body of data, to some extent, that is available in this field. Digital Imaging and Communications in Medicine (DICOM) is an internationally standardized file format for the collection, storage, and transmission of medical imaging data.[2] One

of the most important features of the DICOM standard is that it simultaneously contains imaging data showing the captured picture and metadata or a header describing it. The DICOM standard has successfully standardized the basic format of data across the medical imaging industry to improve the interoperability of medical systems.

To date, however, a lack of sufficient available imaging data for patients from diverse geographic areas has been the single most critical challenge in the development of accurate AI with generalizability.[1] To prepare medical imaging data for machine learning, radiology data need to be properly de-identified, accessed, queried, and integrated with ground-truth or electronic phenotypes.[3] In many instances, these procedures are semi-automated or manual processes, because key information required to identify relevant images is usually missing, incorrect, or non-standardized in the metadata of DICOM files.[4] Gueld, et al.[5] found that the DICOM tag "Body Part Examined" (0018,0015) was incorrectly annotated in 15% of cases. Although there is a DICOM attribute called "Series Description" (0008,103E), this description is free text and is hardly standardized even within single institutions.[6] Hence, it is usually challenging to query DICOM files of interest for a specific patient cohort because of a lack of standardization in DICOM metadata, which in turn hinders reproducible science in radiology.[7]

Here, we propose a standardized schema, the Radiology Common Data Model (R-CDM), for essential DICOM metadata. The R-CDM was designed as a radiology module for the Observational Medical Outcomes Partnership (OMOP)-CDM to incorporate radiology data with standardized clinical data and electronic phenotypes.[8] R-CDM holds the potential to facilitate the preparation of scalable image datasets for machine learning across various institutions, which is of paramount importance for the development of robust AI for radiology.

MATERIALS AND METHODS

In order to standardize medical imaging data with the standardized data model based on OMOP-CDM, four tasks were performed as shown in Supplementary Fig. 1 (only online). First, 41.7 TB of deidentified data were transferred from the Ajou University Hospital, a Korean tertiary teaching hospital, to secure large-scale medical imaging data for research purposes. Second, we standardized the terms in the field of radiology using the RadLex glossary.[9] Logical Observation Identifiers Names and Codes (LOINC) was used to map the radiology protocol terminology of RadLex to the OMOP-vocabulary used by the Observational Health Data Sciences and Informatics (OHDSI) community as an international standard. Third, two tables in the R-CDM were designed to contain medical imaging data in a standardized structure and were developed as an extension model of the OMOP-CDM for seamless connection with clinical data. Last, metadata and imaging data from DICOM, an international standard file format for storing medical imaging data, were used in the process of standardizing medical imaging data.

As a proof of concept, we converted the radiology data of Ajou University Hospital into the R-CDM. Moreover, a desired patient cohort was designed using standardized clinical data of OMOP-CDM, and specific types of images were extracted from the patient cohort through linkage with the R-CDM.[10] Lastly, an R-CDM database viewer was applied in order to confirm the characteristics of the standardized medical image database. This study was approved by the Institutional Review Board at Ajou University Hospital of Republic of Korea (IRB approval number: AJIRB-MED-MDB-20-088).

Acquisition of medical imaging data for standardization into R-CDM

The randomly sampled medical imaging data (41.7 TB; 87203226 images from 2801360 cases), about 10% of the whole dataset, from Ajou University Hospital were approved for usage in research purposes. First, two processes were undertaken: deidentification of the data and transfer into a dedicated server. Although the data were not exported beyond the firewall, we tried to avoid ethical problems with infringement of personal information through the deidentification process. Moreover, the data-transfer process was designed to not overburden networks of the Picture Archiving and Communication System server. Thus, the overall data-transfer process was conducted under the supervision of the information management team of Ajou University Hospital, a third party that was not related to the study. An honest blocker deidentified the transferred data to minimize the risk of personal information infringement. To deidentify the DICOM files of Ajou University Hospital, metadata that contained personal information were chosen. Deidentification was achieved by deleting 12 items of metadata containing personal information, such as the patient's name, sex, date of birth, and location of image recording.

Subsequently, we analyzed the modality composition of the 41.7 TB image database for research purposes. Fig. 1 depicts the counts of imaging occurrences and images by modality. X-ray images with modality values of "CR" and "DX" accounted for 1661414 of the cases (i.e., 59.3% of all 2801360 cases). Ultrasonography (291155 cases, 10.4%) and computed tomography (CT) (259775 cases, 9.3%) scans were assessed. Among all 87203226 images, 55184910 (63.3%) were CT scans, followed by 14164055 images (16.2%) of magnetic resonance imaging (MRI). X-ray images, which occupied an overwhelming majority of occurrences among all imaging modalities, accounted for only 2509684 images (2.9%). This is because, unlike CT or MRI, in which dozens to hundreds of images are created in a single procedure, in X-ray, the number of images created in a single imaging procedure is remarkably small.

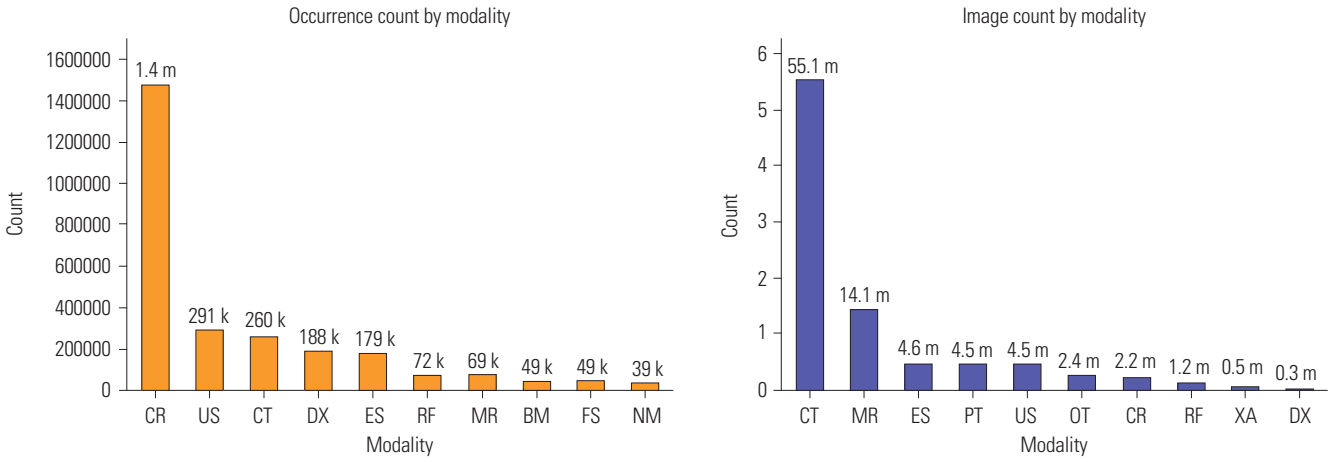


Fig. 1. Occurrence and image count according to modality. CR, computed radiography; US, ultrasound; CT, computed tomography; DX, digital radiography; ES, endoscopy; RF, radio fluoroscopy; MR, magnetic resonance; BM, bone densitometry (X-ray); FS, fundoscopy; NM, nuclear medicine; PT, positron emission tomography; OT, other; XA, X-ray angiography.

Table 1. Structure of Radiology Occurrence Table and Detailed Description of Each Row

Field	Required	Type	Description
radiology_occurrence_id (PK)	Yes	Integer	Unique ID for each image shooting, acting as the primary key of the Radiology study table
person_id (FK)	Yes	Integer	Foreign key that identifies the person who took the image
radiology_occurrence_date	No	Date	Date when the study was taken
radiology_occurrence_datetime	No	Datetime	Date and time when the study was taken
modality	No	Varchar (10)	Value which represents DICOM file type
manufacturer	No	Varchar (50)	Manufacturing company of imaging equipment that carried out image shooting
protocol_concept_id	No	Integer	Value indicating the type of the study
protocol_source_value	No	Varchar (255)	Additional source values describing the study
count_of_series	No	Integer	Count of series generated per imaging study
count_of_images	No	Integer	Count of instances (images) generated per imaging study
radiology_note	No	Varchar (Max)	Recognition findings described by radiology specialists

DICOM, Digital Imaging and Communications in Medicine.

Designing a standard structure for R-CDM

We designed a standardized structure for the R-CDM consisting of two tables linked to each other: the Radiology Occurrence table and the Radiology Image table. To help with understanding the concepts of each table, consider the following example of a patient who underwent brain CT because of a head injury: brain CT imaging without a contrast agent is usually performed for this type of patient, and approximately 40 images are created by the imaging procedure. In this situation, information explaining the imaging of brain CT itself, such as the date and time of the image acquisition and the type of imaging device used for the imaging, is systematically organized in the Radiology Occurrence table. In addition, information describing each of the 40 images generated by one brain CT scan, such as the file path, resolution, and contrast agent administration status of each image, is summarized in the Radiology Image table. Tables 1 and 2 summarize the type, character format, and specific content of each table.

This R-CDM structure is manufactured in a form that is very convenient with the Fast Healthcare Interoperability Resources (FHIR) of Health Level Seven International (HL7), so it is expected that interoperability of standardized image data will be greatly improved in the future. HL7 FHIR stores DICOM metadata in three tables. FHIR's Imaging Study and Series tables are compatible with R-CDM's Radiology Occurrence table, and FHIR's Instance table is compatible with the Radiology Image table.

Internationally standardized terminology system for the R-CDM

We adopted the LOINC/RadLex vocabulary as the standard terminology of the R-CDM for the following reasons: First, RadLex, which was developed by the Radiological Society of North America (RSNA; one of the world's most authoritative groups in the field of radiology), is a glossary that is used for unifying terms in the medical imaging field. The RadLex playbook is the most comprehensive glossary in the field, covering more than 75000 terms.[11] Furthermore, RSNA collaborated with Regen-

https://doi.org/10.3349/ymj.2022.63.S74

Table 2. Structure of Radiology Image Table and Detailed Description of Each Row

Field	Required	Type	Description
radiology_image_id (PK)	Yes	Integer	Unique ID of each image, acting as the primary key of the Radiology image table
radiology_occurrence_id (FK)	Yes	Integer	Unique ID for each image shooting, acting as the primary key of the Radiology study table
radiology_series_id	Yes	Integer	Unique ID of each series
file_path	Yes	Varchar (255)	File path of each image files
body_part_source_value	No	Varchar (20)	Value indicating the photographed body part
laterality_concept_id	No	Varchar (20)	Image shooting direction (anatomical plane)
series_type_concept_id	No	Varchar (20)	Value indicating the type of the series
series_type_source_value	No	Varchar (20)	Additional source values describing the series
series_total_number	No	Integer	Number of images constituting each series
series_serial_number	No	Integer	Order of images within each series
image_resolution_rows	No	Integer	Image resolution (number of horizontal pixels)
image_resolution_columns	No	Integer	Image resolution (number of vertical pixels)
CT_slice_thickness	No	Float	Thickness of CT image slide

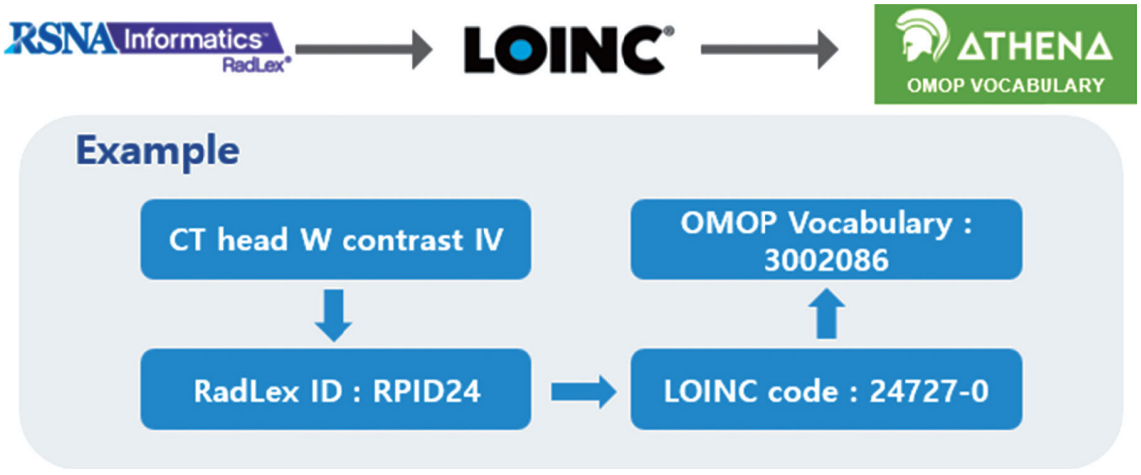


Fig. 2. Mapping process of the RadLex terminology to the Observational Medical Outcomes Partnership (OMOP) vocabulary.

strief and launched a LOINC/RSNA playbook that combines the radiology protocol terminologies of the RadLex playbook and LOINC.[6] Second, the LOINC/RSNA playbook has been demonstrated to cover most of the CT terminology across 40 health information exchange sites in the US.[12] Third, the LOINC/RadLex vocabulary has already been incorporated into the OMOP vocabulary, with LOINC being one of the standard vocabularies in OMOP-CDM.[13]

Fig. 2 shows how the terminologies in the field of radiology were mapped from RadLex to OMOP vocabulary. The figure describes the whole process used to map the protocol of brain CT with a contrast agent to the standard OMOP vocabulary. First, "CT head W contrast IV" with the RadLex ID of "RPID24" in the RadLex glossary was identified. Through the LOINC/RSNA playbook, the LOINC code "24727-0" corresponding to "RPID24" was queried. Finally, the OMOP concept ID "3002086" linked to the LOINC code "24727-0" was searched through the OMOP vocabulary system.[14] Overall, we created a mapping table for 5753 protocol terminologies in RadLex and shared it on GitHub (https://github.com/ABMI/Radiology-CDM).

R-CDM conversion through the DICOM metadata Extract, Transform, and Load process

We identified 16 essential elements from DICOM metadata that contained necessary information to query medical imaging data for machine learning.[3] Through an appropriate extract, transform, and load (ETL) process using the values recorded in the metadata, meaningful information was converted into the format of the R-CDM and loaded into the database. Fig. 3 comprises a diagram showing a part of the ETL process of DICOM metadata. The medical record number in the "Patient ID" DICOM metadata was replaced with "person_id" of OMOP-CDM through deidentification and incorporation into the OMOP-CDM standardized clinical database. Various types of metadata were used to form the "protocol_concept_id" column of the Radiology Occurrence table. Because essential medical information was distributed in several metadata, it was necessary to collect and map them to a single OMOP vocabulary. Further-

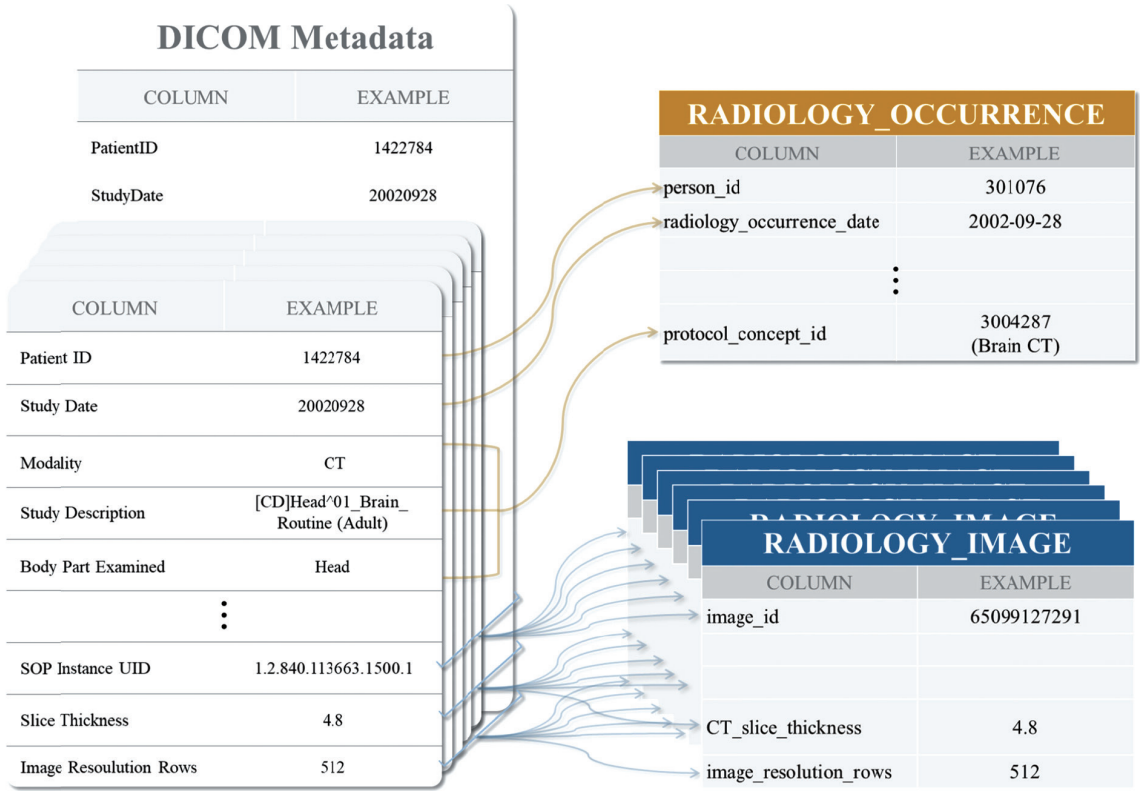


Fig. 3. Metadata extract, transform, and load process.

more, metadata, such as "SOP Instance UID", "Slice Thickness," or "Image Resolution Rows," were entered in a one-to-one correspondence to the appropriate column of Radiology Image table. Values of "SOP Instance UID" were converted for deidentification before being loaded into the "image_id" column.

Interworking of the R-CDM and OMOP-CDM

The R-CDM, which is an extension model of the OMOP-CDM, enables the efficient progress of research by linking patient clinical data to medical imaging data. OMOP-CDM and R-CDM are connected by a common column called person_id, and Fig. 4 shows how the two models are connected. Researchers can easily build a desired patient cohort by utilizing a standardized phenotyping platform called ATLAS from the OMOP-CDM database.[15] The imaging data of interest with specific modality, procedure, and series can be easily retrieved from the R-CDM, which is incorporated into the OMOP-CDM.

As a proof of concept, we attempted to show that the efficiency of the data extraction process can be maximized by combining R-CDM and OMOP-CDM. We extracted an axial view pre-contrast image from primarily acquired brain CT scans of a patient group who visited the emergency room because of cerebral hemorrhage. Supplementary Fig. 2 (only online) depicts the process used for setting a specific patient cohort and extracting only the desired type of image within that cohort using four OMOP-CDM tables and two R-CDM tables. The date of admission to the emergency room of Ajou University Hospital for cerebral hemorrhage was designated as the index date. Only patients who had been enrolled into the database for at least 2 months before the index date were included in the analysis to avoid left censoring. Using patient data from Ajou University Hospital spanning 30 years (from 1998 to 2018) standardized with the OMOP-CDM, the size of the patient cohort was 4685 individuals. Axial view pre-contrast brain CT images acquired on the day of the emergency room visit for the previously set patient group could be extracted from the R-CDM converted database. For the data retrieval process, OMOP vocabulary "3004287," which means brain CT, was searched in the protocol_concept_id column of the Radiology Occurrence table. Subsequently, "28833" and "10579," which mean "pre-contrast" and "axial plane," respectively, were searched in the series_type and anatomical_plane columns of the Radiology Image table.

Development of the R-CDM database viewer

R-CDM database viewer is a tool that visualizes the main characteristics of a standardized medical image database. Using the R-CDM database viewer, researchers can obtain information about what kind of data the database consists of and when the data was captured. Furthermore, researchers can identify the distribution of certain data types of interest in the database. Since the application was developed in the form of a web application to maximize user convenience, the system can be used in any terminal or connection environment.

https://doi.org/10.3349/ymj.2022.63.S74

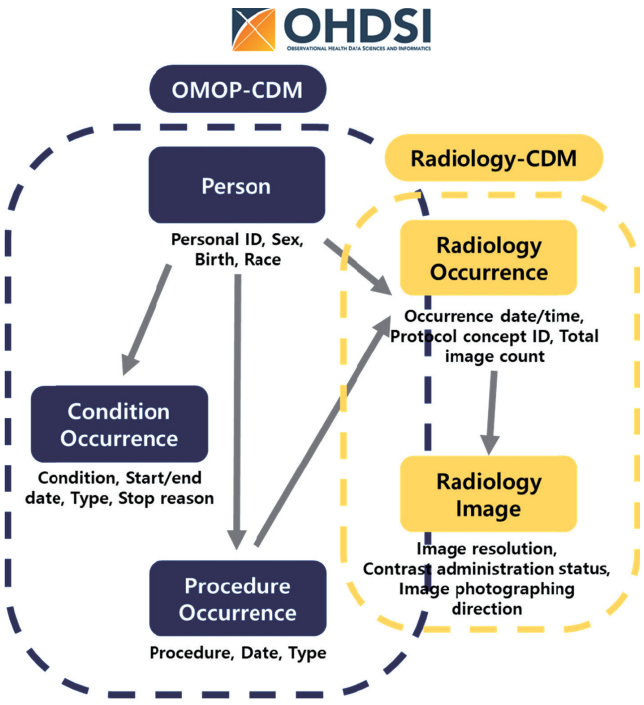


Fig. 4. Interworking of R-CDM and OMOP-CDM. R-CDM, Radiology Common Data Model; OMOP-CDM, Observational Medical Outcomes Partnership CDM.

RESULTS

Results of R-CDM conversion via the DICOM metadata ETL process

Through the ETL process using metadata from DICOM files, 41.7 TB of deidentified medical imaging data were standardized to the structure and terminology system of R-CDM. Information from 2801360 cases was loaded onto the Radiology Occurrence table, and information from 87203226 DICOM files was loaded onto the Radiology Image table. Through sophisticated ETL work using DICOM metadata, most of the columns constituting the R-CDM could be filled with valid values. The conversion results are summarized in Tables 3 and 4. Information pertaining to the dates of recording, replaced patient ID, types of imaging devices, manufacturers of the imaging device, and image resolution could be filled with valid values, with a probability of more than 99.97%. The "CT_slice_thickness" column only contained 81.51% of valid values, as only CT images can have a significant value in the column. To determine the value of the "protocol_concept_id" column, it was necessary to collect meaningful information from several attributes of DICOM metadata and refine them into standardized terms. Detailed information on the image, such as the direction of photography or contrast agent administration status, could not be extracted from the metadata.

Table 3. Results of the ETL Process Using DICOM Metadata in the Radiology Occurrence Table

Field	Unmapped case	Mapped case	Mapping accuracy
radiology_occurrence _id	0	2801360	100
person_id	0	2801360	100
radiology_occurrence_date	22	2801338	99.99
radiology_occurrence _datetime	85	2801275	99.99
modality	0	2801360	100
manufacturer	1930	2799430	99.93
protocol_concept_id	782968	2018392	72.05
protocol_source_value	782968	2018392	72.05
count_of_series	0	2801360	100
count_of_images	0	2801360	100
radiology_note			

ETL, extract, transform, and load; DICOM, Digital Imaging and Communications in Medicine.

Table 4. Results of the ETL Process Using DICOM Metadata in the Radiology Image Table

Field	Unmapped case	Mapped case	Mapping accuracy
radiology_image_id	0	87336478	100
radiology_occurrence_id	0	87336478	100
radiology_series_id	0	87336478	100
file_path	0	87336478	100
body_part_source_value	7765143	79571335	91.11
laterality_source_value			
series_type_source_value			
series_total_number	0	87336478	100
series_serial_number	0	87336478	100
image_resolution_rows	29964	87306514	99.97
image_resolution_columns	29964	87306514	99.97
CT_slice_thickness	16149733	71186745	81.51

ETL, extract, transform, and load; DICOM, Digital Imaging and Communications in Medicine.

Data extraction process by combining R-CDM and OMOP-CDM

Using the imaging data that were available for research purposes, 445 cases and 18275 axial view pre-contrast images could be extracted from the cohort of 4685 patients. Moreover, we extracted images from more detailed patient cohorts. Additional conditions were applied to the cohort to design a group of patients with a good prognosis with a hospitalization period of less than 15 days and a patient group with a poor prognosis who died within a period of 30 days or more and 2 months of admission. Finally, 8136 axial view pre-contrast brain CT images from 198 cases and 4970 images from 121 cases were extracted, respectively.

R-CDM Database viewer

Fig. 5 is the main page of the R-CDM DB viewer, along with a

Fig. 5. Main page of the Radiology Common Data Model (R-CDM) database viewer.

brief explanation of R-CDM and how to use the R-CDM DB viewer. After a simple login process, a user can see a page that visualizes the main features of the R-CDM database (Fig. 6). Barplot and pie charts on the left show the distribution of image data by shooting year and modality. On the right, there is a table that lists the most frequently included protocols in the R-CDM DB and a table that indicates the count of combinations of metadata. Users can easily determine what kind of image data is included the most in the database and the distribution of the desired data using the search function.

DISCUSSION

We identified the limitations of the DICOM international standard and developed a new standardization method to overcome them. We designed the standardized structure and terminology system of R-CDM and applied a deep learning image classifier to improve the quality of the converted data. As a proof of concept, 41.7 TB including 87203226 images obtained for the purpose of the study were converted into the R-CDM. Among them, 10813807 brain CT images were accurately classified by the deep learning image classifier. Furthermore, we showed that, through a combination of the R-CDM and OMOP-CDM, studies that link clinical data with imaging data can be conducted efficiently.

Despite the increasing number of publications about machine learning in radiology, the development and implementation of a standardized infrastructure for the large-scale retrieval of medical images has been a very daunting challenge. Basu, et al.[4] described the challenges associated with the development of The Cancer Imaging Archive, which faces fundamental difficulties with retrospectively harmonizing imaging and non-imaging data for cancer research from multiple institutions, including the inevitable extensive review of DICOM headers. Although content-based image retrieval has been extensively researched over the past decades, it has provided few valid large-scale retrieval systems to date.[16-18] Recently, Pizarro,

https://doi.org/10.3349/ymj.2022.63.S74

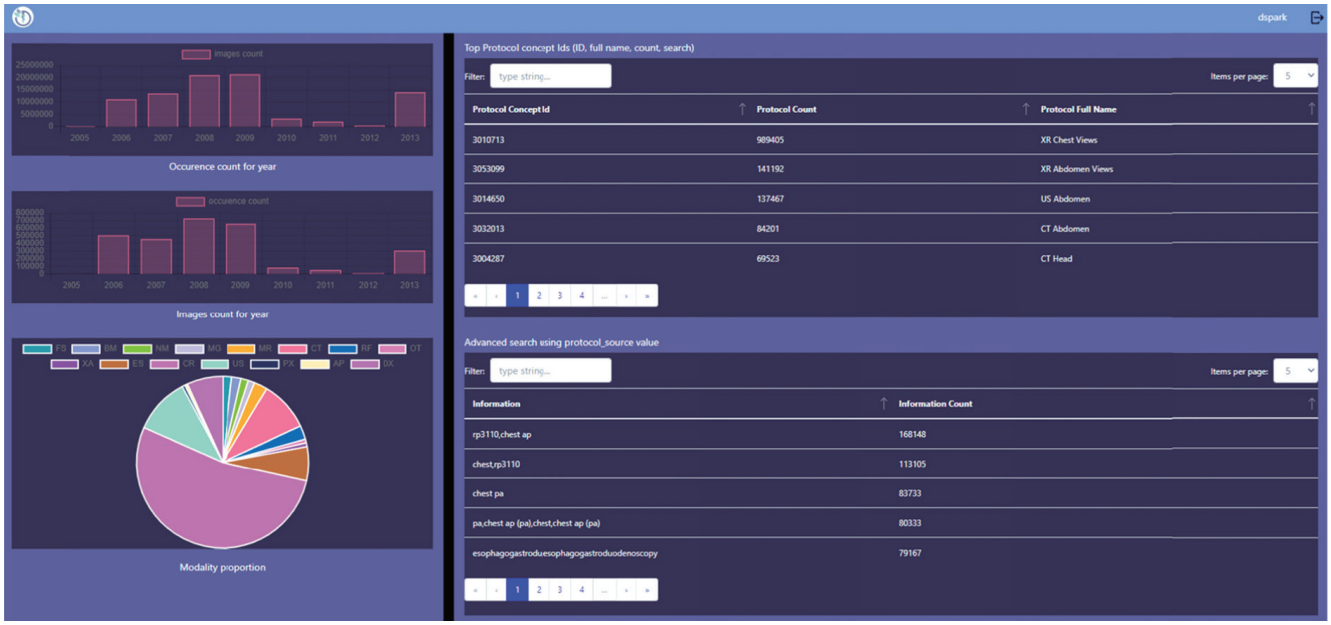


Fig. 6. Analysis results on the visualization page of the Radiology Common Data Model database viewer.

et al.[19] proposed a deep learning algorithm to automatically categorize the series of brain MRI using images, while Gauriau, et al.[6] developed an algorithm for a similar task using DICOM metadata. The ultimate goal of these efforts was to build infrastructure for the large-scale retrieval of medical images to prepare datasets for machine learning.

The reproducibility of this new technique in the medical field is associated with unique challenges and obstacles, which should be carefully considered in the context of its validity, safety, and effectiveness.[20] These challenges can mainly be overcome by the development of standards and methods for data curation, distribution, sharing, and management.[21] OHDSI is an international collaborative community of researchers that maintains the OMOP-CDM. A standardized framework to generate and evaluate machine learning models has been implemented across the distributed research network of CDM databases.[22] As a response to the current coronavirus disease 2019 (COVID-19) pandemic, the OHDSI built distributed COVID-19 cohorts, evaluated a proposed algorithm to identify vulnerable patients, and developed a more reliable and reproducible algorithm for the same task using heterogeneous but standardized databases across the world.[23-25] As the proposed R-CDM is fully compatible with the clinical data of the OMOP-CDM, the implementation of the R-CDM can facilitate the development and evaluation of AI for radiology using standardized electronic phenotyping via collaborative and reproducible research.[26]

The use of the CDM approach, which is a federated standardized data network, to standardize medical imaging data afforded several advantages: mainly bringing the algorithm to the data, rather than data to the algorithm.[27] As data owners can host, build, and evaluate their own algorithm or externally develop an algorithm on their data inside their firewall, a lower level of concern exists regarding privacy, governing, and intellectual property issues, compared with a conventional approach with aggregation of data.[27] Moreover, there have been proposals for the integration of genomic or oncology data into the CDM.[28,29] The implementation of standardized data infrastructure combining imaging biomarkers with genomic and clinical phenotype information based on standardized data may facilitate precision medical research.[26]

Our proposal of the integration of electronic structured clinical data with radiology images had several limitations. Radiology Common Data Elements were not included in our proposal, which standardizes reading data recorded by radiology specialists, an essential piece in the development of AI.[30] The extraction, standardization, and integration of radiology notes into the R-CDM, which can be supported by previous natural language processing, should be investigated in a future study.[31,32] Another vital piece of the imaging data for machine learning would be annotation on images. The previously proposed standards, such as the Annotation and Image Markup, should be considered in a future study.[33] Second, the radiology protocol terminology alone in the LOINC/RSNA playbook has been mapped to the OMOP-vocabulary to date. However, the RadLex playbook itself includes a much greater number of concepts, in addition to the protocol terminologies. For instance, "10579," indicating the direction of photography, and "28833," indicating the administration status of a contrast agent, are terminologies that are not included in the LOINC/RSNA playbook; therefore, they are not currently mapped to the OMOP vocabulary. To construct a complete standard terminology system for the R-CDM, it is necessary to create a mapping table that maps other types of radiology terminologies to the OMOP vocabulary.

Third, we applied the proposed system to the radiology data of a single center. The interoperability of the system should be further validated in future research.

In conclusion, the R-CDM was developed to standardize the structure and terminology of incomplete and unstandardized medical imaging data. As a proof of concept, an ETL process was performed on the metadata of Ajou University Hospital DICOM files in accordance with the terminology and structure of R-CDM. Furthermore, through linkage of R-CDM and OMOP-CDM, it was possible to efficiently link medical imaging data with clinical data. We hope that R-CDM will contribute to the development of deep learning in medical imaging by enabling the securement of large-scale medical imaging data from multinational institutions in the OHDSI community and by linking clinical data with the OMOP-CDM and medical imaging data.

ACKNOWLEDGEMENTS

This research was funded by the Bio Industrial Strategic Technology Development Program (20003883, 20005021, 20001234) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea), and grants from the Korea Health Technology R&D Project, MD-Phd/Medical Scientist Trair Program through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health &Welfare, Republic of Korea (grant number: HR16C0001).

AUTHOR CONTRIBUTIONS

Conceptualization: all authors. Data curation: ChulHyoung Park, Seng Chan You, Chang Won Jeong, and Rae Woong Park. Formal analysis: ChulHyoung Park, Seng Chan You, and Rae Woong Park. Funding acquisition: ChulHyoung Park, Seng Chan You, and Rae Woong Park. Investigation: ChulHyoung Park, Seng Chan You, and Rae Woong Park. Methodology: ChulHyoung Park, Seng Chan You, and Rae Woong Park. Project administration: ChulHyoung Park, Seng Chan You, Chang Won Jeong, and Rae Woong Park. Resources: ChulHyoung Park, Seng Chan You, Hokyun Jeon, and Rae Woong Park. Software: ChulHyoung Park, Seng Chan You, and Rae Woong Park. Supervision: ChulHyoung Park, Seng Chan You, Hokyun Jeon, and Rae Woong Park. Validation: ChulHyoung Park, Seng Chan You, and Rae Woong Park. Visualization: ChulHyoung Park, Seng Chan You, Hokyun Jeon, and Rae Woong Park. Writing—original draft: ChulHyoung Park, Seng Chan You, and Rae Woong Park. Writing—review & editing: ChulHyoung Park, Seng Chan You, Hokyun Jeon, and Rae Woong Park. Approval of final manuscript: all authors.

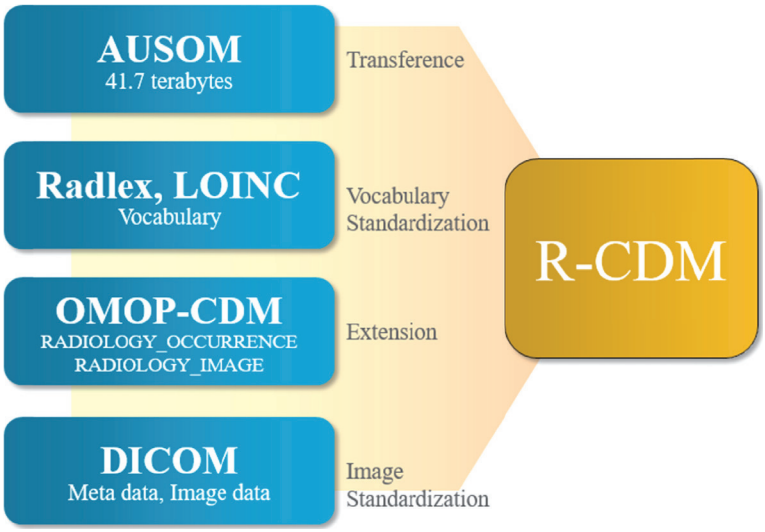
ORCID iDs

ChulHyoung Park https://orcid.org/0000-0003-0531-9144
Seng Chan You https://orcid.org/0000-0002-5052-6399
Hokyun Jeon https://orcid.org/0000-0002-6220-4207
Chang Won Jeong https://orcid.org/0000-0002-9305-4686
Jin Wook Choi https://orcid.org/0000-0002-2396-4705
Rae Woong Park https://orcid.org/0000-0003-4989-3287

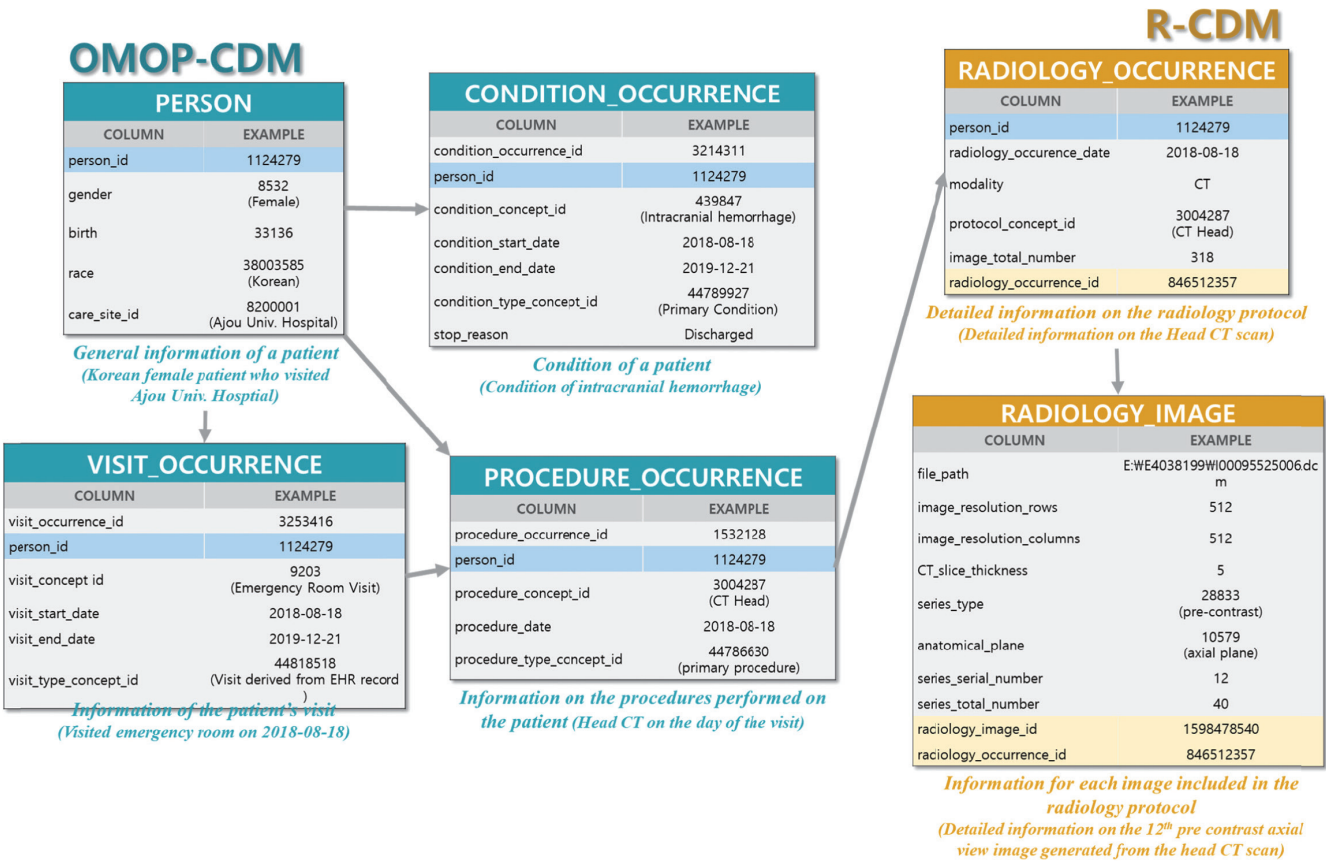
REFERENCES

1. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. Radiology 2019;290:590-606.
2. Mildenberger P, Eichelberg M, Martin E. Introduction to the DICOM standard. Eur Radiol 2002;12:920-7.
3. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. Radiology 2020;295:4-15.
4. Basu A, Warzel D, Eftekhari A, Kirby JS, Freymann J, Knable J, et al. Call for data standardization: lessons learned and recommendations in an imaging study. JCO Clin Cancer Inform 2019;3:1-11.
5. Gueld MO, Kohnen M, Keysers D, Schubert H, Wein BB, Bredno J, et al. Quality of DICOM header information for image categorization. Medical Imaging 2002;4685:280-7.
6. Gauriau R, Bridge C, Chen L, Kitamura F, Tenenholtz NA, Kirsch JE, et al. Using DICOM metadata for radiological image series categorization: a feasibility study on large clinical brain MRI datasets. J Digit Imaging 2020;33:747-62.
7. Blackledge J, Al-Rawi A, Tobin P. Stegacryption of DICOM metadata. Proceedings of the 25th IET Irish Signals & Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies (ISSC 2014/CIICT 2014); 2014 Jun 26-27; Limerick, Ireland: ISSC; 2014. p.304-9.
8. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-8.
9. Vreeman DJ, Abhyankar S, Wang KC, Carr C, Collins B, Rubin DL, et al. The LOINC RSNA radiology playbook-a unified terminology for radiology procedures. J Am Med Inform Assoc 2018;25:885-93.
10. Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, Benoit B, et al. Facilitating phenotype transfer using a common data model. J Biomed Inform 2019;96:103253.
11. Hong Y, Zhang J, Heilbrun ME, Kahn CE Jr. Analysis of RadLex coverage and term co-occurrence in radiology reporting templates. J Digit Imaging 2012;25:56-62.
12. Peng P, Beitia AO, Vreeman DJ, Loo GT, Delman BN, Thum F, et al. Mapping of HIE CT terms to LOINC®: analysis of content-dependent coverage and coverage improvement through new term creation. J Am Med Inform Assoc 2019;26:19-27.
13. Matcho A, Ryan P, Fife D, Reich C. Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. Drug Saf 2014;37:945-59.
14. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform 2012;45:689-96.
15. Lima DM, Rodrigues-Jr JF, Traina AJM, Pires FA, Gutierrez MA. Transforming two decades of ePR data to OMOP CDM for clinical research. Stud Health Technol Inform 2019;264:233-7.
16. Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int J Med Inform 2004;73:1-23.
17. Akgül CB, Rubin DL, Napel S, Beaulieu CF, Greenspan H, Acar B. Content-based image retrieval in radiology: current status and future directions. J Digit Imaging 2011;24:208-22.
18. Li Z, Zhang X, Müller H, Zhang S. Large-scale retrieval for medical image analytics: a comprehensive review. Med Image Anal 2018;43:66-84.
19. Pizarro R, Assemlal HE, De Nigris D, Elliott C, Antel S, Arnold D, et

al. Using deep learning algorithms to automatically identify the brain MRI contrast: implications for managing large databases. Neuroinformatics 2019;17:115-30.
20. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA 2020;323:305-6.
21. Allen B Jr, Seltzer SE, Langlotz CP, Dreyer KP, Summers RM, Petrick N, et al. A road map for translational research on artificial intelligence in medical imaging: from the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop. J Am Coll Radiol 2019;16:1179-89.
22. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 2018;25:969-75.
23. Burn E, You SC, Sena AG, Kostka K, Abedtash H, Abrahão MTF, et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. Nat Commun 2020;11:5009.
24. Reps J, Kim C, Williams R, Markus A, Yang C, Salles TD, et al. Can we trust the prediction model? Demonstrating the importance of external validation by investigating the COVID-19 vulnerability (C-19) index across an international network of observational healthcare datasets. medRxiv [Preprint]. 2020 [accessed on 2021 November 1]. Available at: https://doi.org/10.1101/2020.06.15.20130328.
25. Williams RD, Markus AF, Yang C, Salles TD, Falconer T, Jonnagaddala J, et al. Seek COVER: development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. MedRxiv [Preprint]. 2020 [accessed on 2021 November 1]. Available at: https://doi.org/10.1101/2020.05.26.20112649.
26. Langer SG, Shih G, Nagy P, Landman BA. Collaborative and reproducible research: goals, challenges, and strategies. J Digit Imaging 2018;31:275-82.
27. Kohli MD, Summers RM, Geis JR. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session. J Digit Imaging 2017;30:392-9.
28. Shin SJ, You SC, Park YR, Roh J, Kim JH, Haam S, et al. Genomic common data model for seamless interoperation of biomedical data in clinical practice: retrospective study. J Med Internet Res 2019;21:e13249.
29. Belenkaya R, Gurley MJ, Golozar A, Dymshyts D, Miller RT, Williams AE, et al. Extending the OMOP common data model and standardized vocabularies to support observational cancer research. JCO Clin Cancer Inform 2021;5:12-20.
30. Rubin DL, Kahn CE Jr. Common data elements in radiology. Radiology 2017;283:837-44.
31. Sharma H, Mao C, Zhang Y, Vatani H, Yao L, Zhong Y, et al. Developing a portable natural language processing based phenotyping system. BMC Med Inform Decis Mak 2019;19:78.
32. Park J, You SC, Jeong E, Weng C, Park D, Roh J, et al. A framework (SOCRATex) for hierarchical annotation of unstructured electronic health records and integration into a standardized medical database: development and usability study. JMIR Med Inform 2021;9:e23983.
33. Mongkolwat P, Kleper V, Talbot S, Rubin D. The national cancer informatics program (NCIP) annotation and image markup (AIM) foundation model. J Digit Imaging 2014;27:692-701.



Supplementary Fig. 1. Diagram summarizing the concepts that were required for the standardization of medical image data using R-CDM. LOINC, Logical Observation Identifiers Names and Codes; R-CDM, Radiology Common Data Model; OMOP-CDM, Observational Medical Outcomes Partnership CDM; DICOM, Digital Imaging and Communications in Medicine.



Supplementary Fig. 2. Process used for retrieving pre-contrast axial view images of brain CT acquired on the day of emergency room visitation through the linkage between the OMOP-CDM and R-CDM. CT, computed tomography; R-CDM, Radiology Common Data Model; OMOP-CDM, Observational Medical Outcomes Partnership CDM.