**Introduction**

Digital Imaging and Communications in Medicine (DICOM) is the international standard for sharing and storing medical imaging data. It simultaneously includes image data showing the photographs taken and metadata describing each photo. Each photo contains dozens to as many as hundreds of metadata, and each metadata includes information such as the patient's name and patient number, the date and time of the photographing, the location of the photographing, and the body part photographed. Researchers use metadata to search for and extract medical image data they need from a database. However, there is no standardized format for recording the value of metadata, and sometimes some important medical information is not recorded or recorded as wrong value, resulting in a lot of inefficiency in searching for the desired data. Moreover, researchers cannot be sure whether the extracted dataset is clean or that all of the necessary data has been extracted from the database. There is an urgent need to find better ways to collect and reuse medical imaging data. Therefore, we are proposing a new standardization method for medical image data named R-CDM. R-CDM is a tool that helps to efficiently secure large-scale data necessary for research by integrating multi-center medical image data in the OHDSI community. In addition, since it is built as an extension model of Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM), it enables research that links patient data and medical image data.

**Method & Materials**

Before standardizing medical imaging data of Ajou University Hospital, we tried to access the database to check how the data is being collected and stored. However, working with raw data creates an ethical problem of invasion of personal information because it contains personal information such as the patient's name or date of birth due to the nature of the metadata that describes each data. Therefore, before accessing and using raw data for research, pseudonymization was required. 12 metadata including patient's personal information was determined and all values ​​recorded in the corresponding item were deleted. All work was done by using pydicom package from python, and 68TB of medical imaging data from 1994 to 2009 has been pseudonymized.

After completing the pseudonymization process, we identified the status of the metadata being recorded. Two main problems were found, and the first was that the terminology recorded as the value of metadata was not standardized. We analyzed the metadata ‘StudyDescription’ of 65010 images from the same type of brain CT angiography. ‘StudyDescription’ is a metadata which includes information about what kind of photography was taken. Although it can be standardized to the OMOP vocabulary of ‘CT angiography of head (Concept ID: 36713289)’, values of the metadata were currently recorded in 14 types (e. g. Head^01\_Brain\_angio (Adult), Head^02\_Brain\_Angiography (Adult), C-T Acute Brain Angio c Contrast etc.).

Secondly, there was a case where incorrect information was recorded in metadata. After extracting 788 images, one doctor manually classified each data into precise types. Each images were DICOM files which had a value of 'CR' or 'DX' for the metadata item 'Modality' which contains information about the type of photographing device, and 'CHEST-PA' for the 'BodyPartExamined' metadata which contains information about the body part that was taken. Although they all have to be Chest X-ray posteroanterior view images, 14.3% was Chest X-ray anteroposterior view images, 1.7% was Chest X-ray lateral view images, and 1.1% was Chest X-ray lateral decubitus view images.

Moreover, the 'BodyPartExamined' metadata should have a value of 'Chest' due to its characteristics, and it is reasonable to put information indicating the direction of the shooting in the metadata of ‘Patient Orientation’, ‘Patient Position’ or ‘Image Orientation’. However, it is not easy to determine where specific information should be recorded among 3673 types of metadata.

We have identified that the metadata of medical imaging data of Ajou University Hospital is also incomplete and has not been standardized. Through the development and application of R-CDM, we intended to transform medical imaging data into standardized terminologies and structures without loss of information. R-CDM having a Radiology Occurrence table and a Radiology Image table was designed to organize medical image data into a standardized structure. Information about the shooting itself, such as a shooting date or an equipment used for the shooting, is included in the Radiology Occurrence table, and information about each image obtained through the shooting, such as resolution of each image is included in the Radiology Image table. Foreign key was set up so that the R-CDM table can be linked with OMOP-CDM, in order to pursue a seamless link between patient clinical data and medical imaging data. Terminology system was standardized by using RadLex playbook, a comprehensive lexicon of radiology terms containing more than 75,000 terms. Moreover, by using LOINC / RSNA playbook, which combines RadLex playbook and Logical Observation Identifiers Names and Codes (LOINC), RadLex terminology can be mapped to OMOP vocabulary via LOINC. LOINC is the international standard for identifying health measurements, observations, and documents. Finally RadLex, which has been independent of OMOP vocabulary until now, can integrate, and standardize medical imaging terminologies within the OHDSI community. A mapping table that maps a total of 5753 RadLex imaging medical procedure terms to OMOP vocabulary was created and uploaded to github.

Two steps were taken to transform the data into the standardized structure and terminology system of R-CDM. Firstly, ETL process is conducted with 16 types of DICOM metadata that were determined to contain significant information. Secondly, detailed information that was difficult to extract from the metadata is extracted from the image data using the CNN deep learning image classifier.

Results

Accuracy of ETL process with DICOM metadata

Information about photographing date, patient ID, resolution, type of photographing device, and manufacturing company of 68TB medical imaging data was mapped to an accuracy of 99.99% and loaded into the database. Information of photographed body part which were distributed in 5 types of metadata (ProtocolName, PerformedProcedureStepDescription, BodyPartExamined, SeriesDescription, StudyDescription), was ETL processed and mapped with an accuracy of 80.7%. However, it was not possible to map more specific and detailed information such as the contrast administration status or photographing direction of each image.

Classification accuracy of deep learning image classifier

We divided 53000 images into a training set, 7000 images into a validation set, and 7000 images into a test set to train the model to distinguish 20 classes. The image classifier showed learning results of 99.87 AUROC and 97% accuracy with 150 epochs, and the classification results were recorded in the Radiology Series Concept ID and Radiology Orientation Concept ID of the Radiology Image table.

Extraction of desired imaging data through interworking with OMOP-CDM and R-CDM

By establishing a foreign key connecting R-CDM and OMOP-CDM, it was confirmed that medical image data, which has been converted to R-CDM and achieved international standardization, is connected to clinical data contained in the structure of OMOP-CDM. Ajou University Hospital clinical data from 1994 to 2017, which were standardized and converted to the structure and terminology of OMOP-CDM, was used. We have designed a cohort of 1208 people through OMOP-CDM who satisfy the condition by searching for children under 3 years of age who have taken a skull X-ray series and head CT by visiting the emergency department with head trauma since 2010.

**Conclusion**

R-CDM has been developed to internationally standardize structure and terminology system of incomplete and unstandardized medical imaging data. In order to standardize medical imaging data, ETL was primarily done with DICOM metadata, and additional DICOM image data deep learning classifier was applied for more detailed and accurate ETL. Furthermore, through the connection of R-CDM and OMOP-CDM, it is possible to efficiently link medical image data and clinical data. We hope R-CDM to contribute to the development of the field of medical imaging deep learning by enabling the securement of large-scale medical imaging data of multinational institutions in the OHDSI community and the linkage between clinical data of OMOP-CDM and medical image data.

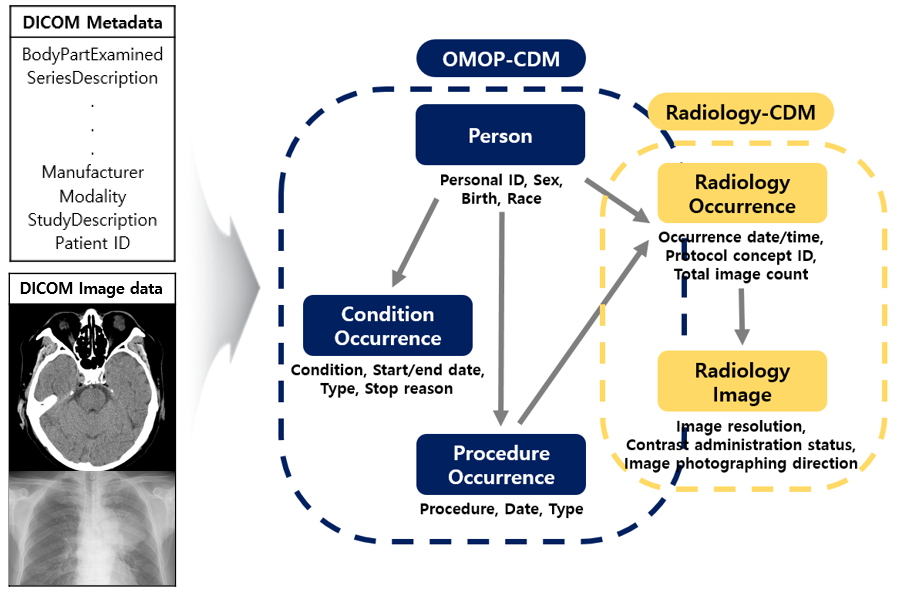
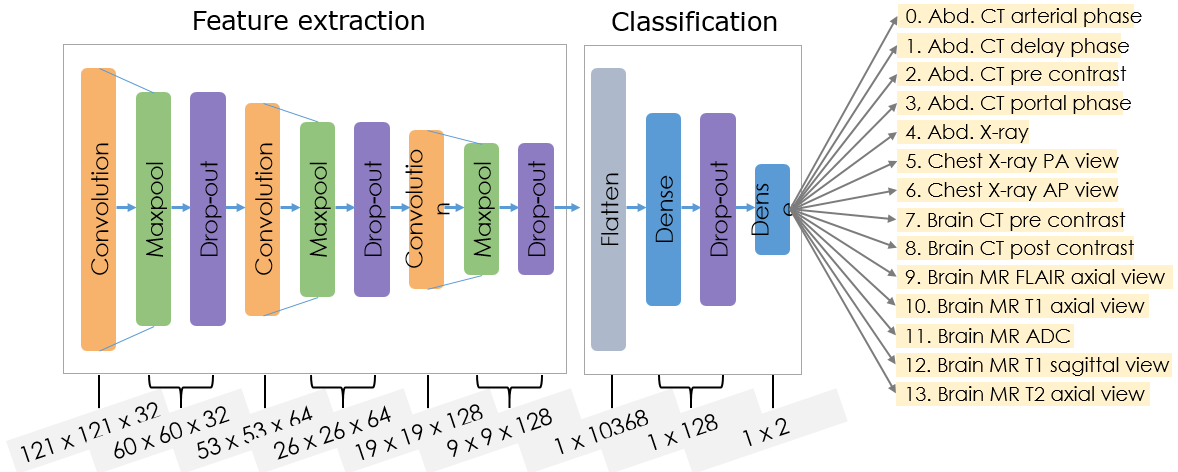
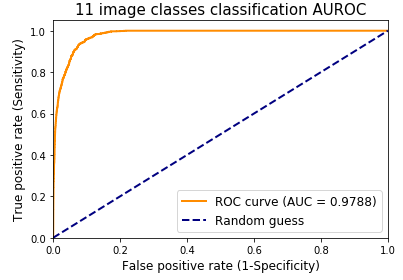


Figure 1.

Figure 2. DICOM 메타데이터 ETL, 이미지데이터 전처리 모식도





14

Figure 3.