**Introduction**

DICOM (Digital Imaging and Communications in Medicine) 은 의료 영상 데이터를 수집, 저장, 그리고 전송하기 위하여 국제적으로 표준화된 파일 형식이다. 각 DICOM 파일은 촬영한 사진을 보여주는 이미지 데이터와 사진을 설명해주는 메타데이터를 동시에 포함한다. 각 사진은 수십 개에서 많게는 수백 개의 메타데이터를 포함하며 각 메타데이터에는 촬영한 환자의 이름과 성별, 나이, 환자 번호, 촬영 날짜와 시간, 촬영 장소, 촬영 신체 부위 등의 정보가 기술되어 있다. 연구자들은 데이터베이스 내에서 본인이 필요한 의료 영상 데이터를 검색하고 추출하는 데에 메타데이터를 활용한다. 하지만 메타데이터의 값을 기록하기 위해 정해진 표준화된 형식이 없으며, 메타데이터에 중요한 의학 정보가 기록되지 않거나 틀린 값으로 기록된 DICOM 파일들이 있기 때문에 원하는 자료를 검색하는 일에 많은 비효율이 발생한다. 검색한 데이터 세트에 연구자가 원하지 않는 다른 종류의 데이터가 들어있을 수 있고, 필요로 하는 데이터가 데이터베이스에서 빠짐없이 추출되었는지 확신할 수 없다. 이러한 문제를 극복하기 위해 DICOM 파일들에 기록된 메타데이터를 표준화하고, 틀리거나 비어 있는 정보를 올바른 정보로 채워 넣어 주어야 한다. 의료 영상 데이터를 국제적으로 표준화함과 동시에 표준화를 이룬 데이터를 데이터베이스에 적재하여 효율적인 자료 검색을 할 수 있게 돕는 R-CDM을 소개한다. R-CDM은 OHDSI 커뮤니티 내의 다기관 의료 영상 데이터를 표준화하여 통합하고 연구자가 원하는 대규모의 데이터를 효율적으로 확보할 수 있도록 돕는다. 또한 R-CDM은 OMOP-CDM의 확장 모델로써 작동하기 때문에 둘은 연결되어 있다. OMOP-CDM의 구조로 표준화 변환되어있는 병원의 임상 데이터와 R-CDM으로 변환한 의료 영상 데이터를 연계한 연구를 효율적으로 가능케 한다.

**Method**

**가명화된 연구 데이터 준비**

의료 영상 데이터의 표준화를 위해 데이터에 접근하기에 앞서 데이터의 가명화 작업이 이루어져야 한다. 각 이미지 파일의 메타데이터에는 환자의 이름과 환자번호, 생년월일, 성별 등의 개인정보가 포함되어 있다. 따라서 Raw data를 활용한 연구는 개인정보 침해라는 윤리적 문제를 발생시킬 수 있다. 본 연구에서는 의료원의 데이터 제공 절차에 따라 연구와 관련 없는 제3자인 의료 빅데이터 센터로부터 가명화된 데이터를 수령하여 연구를 진행했다. 환자의 개인정보가 포함된 메타데이터 12항목에 기록된 값들을 모두 삭제했으며, 환자번호는 복원이 불가능한 연구 아이디로 변환되었다. 아주대병원에서 1994년에서 2018년 사이에 촬영되었으며, 가명화 작업을 거친 68TB의 의료 영상 데이터를 활용하여 연구를 진행하였다.

**아주대학교 의료 영상 데이터의 문제점 파악**

아주대병원 의료영상데이터의 메타데이터가 기록되어 있는 현황을 살펴본 결과 크게 세 가지의 문제점을 발견할 수 있었다. 첫 번째로 메타데이터 값을 기술하는 용어가 표준화되어 있지 않아 동일한 의학 용어도 각기 다른 방식으로 기술되어 있었다. 동일한 종류의 혈관조영영상 뇌CT촬영을 199케이스를 진행하여 발생한 65010장의 사진에서 어떤 종류의 촬영을 진행했는지 기록하는 메타데이터(Study Description)를 분석한 결과 총 14종류의 방식 (e. g. Head^01\_Brain\_angio(Adult), Head^02\_Brain\_Angiography (Adult), C-T Acute Brain Angio c Contrast etc.)으로 기록되어 있었다. 각종 특수기호들이 함께 포함되어 있고, 동일한 용어도 다르게 기술하고 있음을 확인할 수 있는 등 값을 기술할 때에 전혀 표준화를 이루고 있지 않음을 확인할 수 있었다. 이러한 Raw data가 R-CDM을 통한 표준화 작업을 거친다면 OHDSI 네트워크가 사용하고 있는 국제적 표준화 용어 시스템 OMOP vocabulary 36713289 (CT angiography of head)로 매핑될 수 있을 것이다.

두 번째로, 메타데이터에 틀린 정보들이 기록되어 있는 경우가 있다. 메타데이터를 기반으로 한 검색을 통하여 Chest X-ray posteroanterior (PA) view 사진을 추출하여 자세히 들여다보았다. 촬영 기기 종류에 대한 정보가 담기는 ‘Modality’ 메타데이터가 ‘CR’ 혹은 ‘DX’라는 값을 가지며, 촬영한 신체 부위에 관한 정보가 담기는 ‘BodyPartExamined’ 메타데이터가 ‘CHEST-PA’라는 값을 가지는 DICOM 파일을 검색했다. 메타데이터를 통해 검색한 바에 따르면 모두 Chest X-ray posteroanterior (PA) view 사진이어야 하지만, 의사 1명이 수기로 정확한 종류를 분류한 결과 14.3%에 해당하는 113장은 Chest X-ray anteroposterior (AP) view 사진이었으며, 1.7%에 해당하는 13장은 Chest X-ray lateral view 사진, 1.1%에 해당하는 8장은 Chest X-ray lateral decubitus view 사진이었다. 환자가 Chest X-ray 촬영을 한 번 진행하는 경우, 대부분의 환자는 기본적으로 Chest X-ray PA view 사진을 촬영하고, 환자의 상황이나 의심되는 질병에 따라 신체를 여러 방향으로 기울여 여러 장의 사진을 찍게 된다. 이 때 촬영기사는 각 사진에 대한 정확한 설명을 기록하기보단 촬영을 대표할 수 있는 기록값 하나를 선택하고 모든 사진에 동일하게 기입하는 방법을 택하곤 한다. 앞서 살펴본 케이스는 chest X-ray PA view와 다른 방향에서 촬영한 사진을 연이어 찍고, 촬영을 대표할 수 있다고 생각한 ‘CHEST-PA’라는 값을 모두 기록한 경우이다.

세 번째로, 특정 종류의 정보가 원래 기록되어야 할 곳이 아닌 다른 메타데이터에 기록되어 있는 경우가 있다. 앞서 예를 든 Chest X-ray 데이터의 경우만 보아도 BodyPartExamined 메타데이터에 ‘CHEST-PA’라는 값이 기록되어 있음을 확인할 수 있다. 촬영한 신체 부위에 관한 정보가 담기는 항목이기 때문에 ‘Chest’라는 값이 들어가야 하며, 촬영한 방향을 나타내는 정보 ‘PA’는 ‘PatientOrientation’, ‘PatientPosition’ 등의 메타데이터에 들어가는 것이 합당하다. 하지만 총 3673 종류의 메타데이터 중에서 특정 정보가 어느 종류의 메타데이터에 기록되어야 하는지 판단하기는 절대 쉬운 일이 아니다.

마지막으로 중요한 의학 정보가 어디에도 기록되어 있지 않은 데이터도 있다. 총 87,203,228개의 DICOM 파일의 메타데이터를 확인한 결과 촬영한 신체 부위를 포함하여 어떤 종류의 촬영인지에 관한 정보가 전혀 기술되어 있지 않은 1,129,417개의 데이터를 발견할 수 있었다. 더욱이 31,209개의 파일에는 환자 번호가 기록되어 있지 않아서 어떤 환자를 촬영한 의료 영상 데이터인지 확인할 수가 없다.

아주대병원 의료 영상 데이터의 메타데이터가 불완전하며 표준화되어 있지 않은 상태임을 충분히 확인할 수 있었다. 의료 영상 데이터의 국제적 표준화 변형 툴인 R-CDM은 표준화되지 않은 채 각기 다른 방식으로 기록되어 있던 의학 용어를 국제적 표준화 용어가 대체하며, 내재하고 있는 딥러닝 이미지 분류기를 통해 틀린 정보와 비어 있는 정보를 올바른 값으로 수정할 것이다. 또한 수백 종류의 서로 다른 메타데이터에 분산되어 저장되고 있던 의학 정보를 한 데 모아 표준화된 구조로 맞추어줄 것이다.

**R-CDM 표준화 용어 시스템 설계**

영상의학 분야에서 가장 권위 있는 북미영상의학회(RSNA)에서 관리하고 있는 RadLex playbook을 사용하여 R-CDM의 표준화된 용어 시스템을 구축했다. RadLex playbook은 75000개 이상의 용어를 보유하고 있는 영상 의학 분야 최대 규모의 용어집이다. 최근 RSNA와 Logical Observation Identifiers Names and Codes (LOINC)를 관리하는 Regenstrief가 협업하여 RadLex playbook와 LOINC의 용어 일부를 결합한 LOINC/RSNA playbook을 출시했다. LOINC는 현재 OMOP vocabulary와 결합되어 있는 건강 측정, 관찰 및 문서를 기록하기 위한 국제 표준이다. 따라서 R-CDM이 채택한 용어 시스템은 영상의학 분야 전체를 아우르며, 국제적 표준화를 이루게 되었다. 추가적으로 RadLex, LOINC, OMOP vocabulary를 연결하는 매핑 테이블을 작성해 github에 공유하였다. *(https://github.com/ABMI/Radiology-CDM/tree/master/RadETL)*

**R-CDM 표준화 구조 설계**

Radiology Occurrence table, Radiology Image table의 서로 연결된 2개의 테이블로 구성되는 R-CDM의 표준화된 구조를 설계했다. 보통 환자가 한 번의 촬영을 할 때 적게는 몇 장에서부터 많게는 수백, 수천 장의 이미지가 창출되는데 각 촬영에 대한 정보는 Radiology Occurrence table에, 각 이미지에 대한 정보는 Radiology Image table에 들어가게 된다. Radiology Occurrence table에는 어떠한 종류의 촬영이 언제 일어났는지 등의 촬영 자체를 설명해주는 데이터가 들어가게 되고, Radiology Image table에는 이미지의 해상도, 이미지 촬영 방향 등 1번의 촬영에서 발생한 각각의 이미지들에 대한 데이터가 들어가게 된다.

또한 R-CDM과 OMOP-CDM을 연결하는 외래 키를 구축하여 의료 영상 데이터와 임상 데이터 간의 원활한 연계가 이루어지도록 했다. 연구자는 OMOP-CDM의 형태로 표준화 변환되어 있는 환자의 임상 데이터를 활용하여 특정 환자 코호트를 구축할 수 있으며, R-CDM으로 표준화 변환되어 있는 의료 영상 데이터베이스 내에서 해당 환자 코호트가 촬영한 의료 영상 데이터를 한 번에 추출할 수 있다. 그 중 연구자가 원하는 종류의 의료 영상 데이터를 국제적 표준화를 이룬 용어를 사용하여 검색할 수 있다.

**아주대병원 의료 영상 데이터의 R-CDM 변환**

아주대병원 의료 영상 데이터를 R-CDM의 표준화된 구조와 용어 시스템으로 변환하기 위해 메타데이터의 ETL과 딥러닝 이미지 분류기 적용의 2가지 단계를 거치도록 했다. 일차적으로 의료 영상 데이터에 기록된 메타데이터로부터 최대한의 정보를 추출해 R-CDM의 표준화된 형태로 담아낸다. 하지만 앞서 살펴보았다시피 메타데이터는 불완전하며, 틀린 값을 보유하고 있는 경우가 흔하며, 이렇게 원자료가 문제가 있는 경우 표준화 변환을 마친 데이터도 불완전하거나 틀린 값을 가지게 될 수밖에 없다. 이러한 문제를 해결하기 위해 딥러닝 이미지 분류기가 R-CDM의 형태로 표준화 변환된 데이터를 한 번씩 스크리닝하여 틀린 데이터나 결측치는 올바른 값으로 수정하며, 메타데이터로부터 추출하기 힘들었던 세부적인 정보들을 추출해 추가적으로 기입하도록 하였다.

먼저 수백 개의 DICOM 파일을 수기로 확인하여 각 데이터가 보유하고 있는 수십~수백 종류의 메타데이터 중 유의미한 의학 정보가 공통적으로 기록되어 있는 16개 항목의 메타데이터를 결정했다. 해당 항목에 기록된 값들은 적절한 Extract, Transform, and Load (ETL) 과정을 거쳐 R-CDM의 용어와 구조 시스템으로 표준화된 다음 데이터베이스에 적재되었다.

메타데이터에 기록된 값만으로도 R-CDM 대부분의 컬럼을 올바른 값으로 채울 수 있었지만, 촬영 신체 부위에 관한 정보는 표준화를 이루지 않은 용어들로 기록되어 있었고, 다양한 종류의 메타데이터에 분산되어 저장되어 있었기 때문에 이를 한 데 모아 표준화된 용어로 정제하는 과정이 필요했다. 더욱이 각 이미지들의 촬영방향, 조영제 투여 현황과 같은 세부적인 정보는 채워 넣을 수 없었다.

메타데이터로부터 추출하기 힘들었던 세부적인 정보들을 정확하게 추출함과 동시에 오류치, 결측치를 수정하기 위해 딥러닝 이미지 분류기를 설계했다. 본 연구에서 설계한 딥러닝 이미지 분류기는 메타데이터 ETL과정을 거쳐 Brain CT (OMOP-vocabulary: 3004287로 매핑)를 촬영했다는 정보를 가지게 된 DICOM 파일을 촬영 방향과 조영제 투여 현황에 따른 총 8개의 클래스로 세부적인 분류를 해준다. Brain CT이미지를 axial view pre contrast image, coronal view pre contrast image, axial view post contrast image, axial view Maximum intensity projection (MIP) image, coronal view MIP image, sagittal view MIP image, 이미지 촬영 범위를 알려주는 topogram 이미지, 3D reconstruction 이미지의 8개의 class로 분류를 해준다.

첫 번째로 딥러닝 모델의 학습 및 평가에 이용될 이미지 데이터가 균질한 형태를 가지도록 하는 resizing, reshaping, pixel regularization, 밝기 조절 등의 데이터 전처리 과정이 이루어진다. 전처리를 마친 데이터는 이미지 분류에 높은 성능을 지니고 있다고 알려진 Resnet50 딥러닝 모델에 의해 학습된다. 마지막으로 각 이미지 데이터의 분류 결과는 데이터베이스에 새롭게 업데이트된다. 이러한 데이터 전처리, 데이터 multiclass classification, 데이터베이스에 적재하는 3가지 과정을 자동화하는 파이프라인을 구축했다.

딥러닝 툴 개발에 필요한 데이터의 확보를 위해 DICOM 메타데이터를 기반으로 Brain CT의 촬영 방향 별, 조영제 투여 현황 별 사진을 검색했다. 많은 DICOM 파일에 촬영 방향이나 조영제 투여 현황까지의 세부적인 정보를 기록해 두지 않았지만, 일부 잘 기록된 데이터들도 존재한다. 예를 들어 총 ~개의 3004287로 매핑된 데이터 중에서 axial view pre contrast brain CT image를 검색하기 위해서 촬영 장비를 검색할 때에는 Modality 메타데이터에서 ‘CR’ 혹은 ‘DX’를 검색하면 되지만, 촬영 신체 부위, 촬영 방향에 관한 정보는 5종류의 메타데이터에 표준화를 이루지 않은 형식으로 분산되어 저장되어 있다. 따라서 5종류의 메타데이터 내에 대소문자를 불문하고 chest라는 단어가 포함되어 있는 데이터를 검색함으로써 촬영 신체 부위에 대한 검색했고, 각 방향에 대한 검색도 유사한 방식으로 진행했다. 전체 ~개의 DICOM 파일에 대하여 총 ~장의 chest X-ray를 검색할 수 있었고, 그 중 각 방향에 대한 X-ray ~장, ~장, ~장, ~장, ~장, ~장을 추출할 수 있었다. 앞서 살펴보았듯이 메타데이터에는 틀린 값이 기록되어 있는 경우가 있을 수 있으니 의사 1명이 수기로 6개의 클래스에 대한 라벨링을 진행하였고, 총 ~장의 train, validation dataset을 준비할 수 있었다.

**Results**

**메타데이터 ETL의 정확도**

~장 데이터의 촬영 날짜, 가명화된 환자 번호, 촬영 기기 종류, 촬영 기기 제조 회사, 이미지 해상도에 관한 정보는 99.99%의 정확도로 매핑되어 데이터베이스에 적재되었다. 5종류의 메타데이터에 분산되어 저장되어 있던 촬영 신체 부위에 관한 정보는 ETL되어 80.7%의 정확도로 매핑되었다. 하지만 각 이미지의 조영제 투여 상황 혹은 촬영 방향 등의 보다 구체적이고 세부적인 정보에 대한 매핑은 불가능했다.

**딥러닝 이미지 분류기의 분류 정확도**

개발된 이미지 분류기는 20000장을 training set, 4000장을 validation set, 4000장을 test set을 활용해 총 6개의 클래스에 대한 multiclass classification을 수행하여 chest X-ray 이미지 각각의 촬영 방향에 대한 정보를 알려준다. 이미지 분류기는 150회의 epoch으로 99.87의 AUROC와 97%의 accuracy의 학습결과를 나타내었고 분류 결과는 Radiology Image table의 Radiology Orientation Concept ID에 기록되었다.

**OMOP-CDM과 R-CDM 연동 예시**

R-CDM과 OMOP-CDM를 연결하는 외래 키를 구축하여 의료 영상 데이터와 임상 데이터의 원활한 연계가 이루어졌고, 이를 검증하기 위해 응급실로 내원하여 당일 chest X-ray를 촬영한 환자들의 chest X-ray PA view의 사진을 활용해 재원일수를 예측하는 연구를 진행했다. OMOP-CDM의 구조와 용어로 표준화 변환된 1994년부터 2017년까지의 아주대병원 임상 데이터를 이용하여 응급실로 내원해 당일 chest X-ray를 촬영한 환자 462,932명의 나이와 성별, 그리고 재원일수를 파악했다. 더욱이 R-CDM 변환된 데이터베이스 내에서 해당 환자가 응급실 내원 당일 촬영한 chest X-ray 중 chest X-ray PA view 사진의 파일경로 ~장을 추출할 수 있었다.

환자의 chest X-ray PA view 이미지, 나이, 성별을 학습하여, 예측 재원일수를 3일 이하, 3일 초과 7일 이하, 7일 초과 30일 이하, 30일 이상의 4가지 클래스로 분류해주는 multi input 딥러닝 모델을 설계했다. chest X-ray PA view 이미지는 resnet50으로 학습하여 2048개의 feature로 차원이 축소되며, 여기에 age, sex의 정보를 추가하여 2050개의 feature를 이용해 4가지 클래스로 분류하게 된다.

input으로 들어가는 영상 데이터의 전처리 과정으로써 224\*224의 크기로 resizing, 채널 개수 3개로 늘리기, histogram equalization, 픽셀 정규화 과정을 진행했다.

~장의 데이터를 ~케이스의 training set, ~케이스의 validation set, ~케이스의 test로 나누었고, 총 150회의 epoch으로 99.87의 AUROC와 97%의 accuracy의 학습결과를 나타내었다.

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.