인공지능 기반의 임상연구를 위한 의료 데이터 셋 관리 시스템[1]

박민기*, 한성민**, 김승진*, 이충섭***, 김태훈***, 정창원***, 윤권하****
*원광대학교 의과학과
**원광대학교 컴퓨터소프트웨어공학과
***원광대학교 의료융합연구센터
****원광대학교 의과대학 영상의학과

e-mail : {pmg0612, zhsks528, koch369369, cslee99, tae_hoonkim, mediblue, khy1646}@wku.ac.kr

Medical Dataset Management System for Artificial Intelligence-Based Clinical Research

Min-Gi Pak*, Seong-Min Han**, Seung-Jin Kim*, Tae-Hoon Kim***, Chung-Sub lee***, Chang-Won Jeong***, Kwon-Ha Yoon****
*Department of Medical Science, Wonkwang University
**Department of Computer Software Engineering, Wonkwnag University
***Medical Convergence Research Center, Wonkwang University
****Dept of Radiology, Wonkwang University School of Medicine and Hospital

요 약

본 논문은 국제표준화인 OHDSI OMOP-CDM 의 확장으로 의료영상 표준기반으로 한 관리시스템에 대해 기술한다. 이를 위해 기존 공통데이터모델과 연계에 중점을 두어 DICOM 메타태그정보 기반의 의료영상 표준 모델의 스키마를 제시한다. 이를 기반으로 머신러닝 기술개발을 위한 데이터 셋 생성과 관리를 위한 웹 기반 시스템 구조와 기능에 대해서 기술한다. 끝으로 구현된 시스템에서 제공하는 웹 서비스 수행 결과를 보인다.

1. 서론

최근 4 차 산업혁명의 빅데이터 및 인공지능(AI) 기술을 기반으로 의료산업은 변화하고 있다. 특히, 의료영상 데이터와 인공지능 기술을 접목한 데이터 분석에 관한 연구가 활발하게 이루어지고 있다. 그러나, 비정형데이터인 의료영상에 대한 표준화는 미흡한 실정이다. 따라서, 연구에 필요한 데이터 셋을 구축하는데 어려움이 따른다. 이에 대한 대안으로 각 기관의 데이터를 OHDSI(Observational Health Data Science and Informatics)에서 제안하는 공통데이터모델(CDM)로 변환하여 다양한 임상연구들을 진행하는 방법이 제시되고 있다 [1-3]. 의료영상정보는 국제 표준인 DICOM 을 준수하여 생성되지만 세부적인 데이터 형식은 각 기관별로 상이하다. 그렇기 때문에 수집된 영상을 바로 임상연구나 인공지능 학습 연구에 적용하기 위해서는 해결해야 할 문제들이 있다 [4]. 이에 대한 해결책으로 의료영상정보를 담고 있는 DICOM 파일로부터 태그 정보를 추출하여 표준화된 CDM 과 연계한 의료영상표준에 대해 제안한다. 그리고 표준화된 임상데이터를 관리할 뿐만 아니라 데이터 셋을 생성하여 제공할 수 있는 시스템에 대해 기술한다.

2. 관련 연구

2-1. 빅데이터 표준화 연구

OHDSI Research Network 는 분산 네트워크를 통한 CDM 기반의 임상 연구를 지원한다. 상이한 자체 데이터베이스를 가진 의료기관들이 CDM 으로 구축하고, 분석알고리즘을 각각 수행하여 연구결과만을 공유하여 연구하는 방법이다. 주요 도구인 Achilles 는 CDM 을 테이블 별로 시각화하며, Atlas 는 웹 기반 자료 분석 도구로 웹 인터페이스로 코호트 구축, 성향 변수 맞춤, 생존 분석, 상대 위험도 계산 등의 통계 분석을 쉽게 할 수 있다. 최근 이런 분산 네트워크를 활용하여 임상연구가 활발하게 진행되고 있다. 그리고 다양

[1] This study was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute(KHIDI), funded by the Ministry of Health & Welfare(HI18C1216) and the Technology Innovation Program (or Industrial Strategic Technology Development Program(20001234)

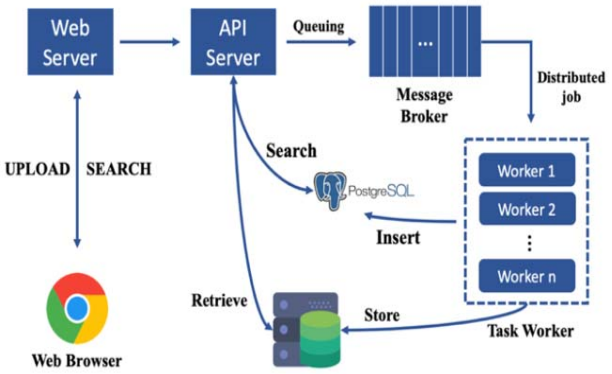
한 Open Source Software [4]를 개발하여 연구자들에게 제공하고 있다.

2-2. Radiology CDM 연구

OMOP-CDM 은 정형화된 임상데이터를 표준화하는데 중점을 두고 있다. 그러나 최근 영상, 이미지, 생체신호 등 비정형 데이터에 관련된 임상데이터의 활용을 위해 모델을 확장하고 있다. 특히, 현재 각 병원에서는 의료영상을 DICOM 국제 표준을 준수하여 저장하고 있으나 각 질환 별로 최적화된 임상 프로토콜에 대한 선별, 핵심적인 의료영상에 저장되는 의료정보까지 표준화되어 저장되어야 한다[5]. 이와 관련하여 수행된 연구는 국내외에서도 미흡한 실정이며 더욱이 의료기관별 의료영상의 표준화된 정보 없이 인공지능 학습 연구에 적용하기에는 어려움이 있다 [6, 7]. 또한 인공지능 학습을 위해서는 방대한 양의 의료영상 데이터로 학습을 시켜야 정확도가 높은 인공지능 알고리즘을 개발할 수 있지만 해당 연구에 대한 케이스의 수집도 매우 어려운 일이다. 이러한 문제점을 해결하기 위해 의료영상에 대한 표준화의 요구사항을 정리하였고, 기존의 OMOP-CDM 과 연계하여 확장형 모델을 제시하였다[8, 9].

3. 제안 시스템

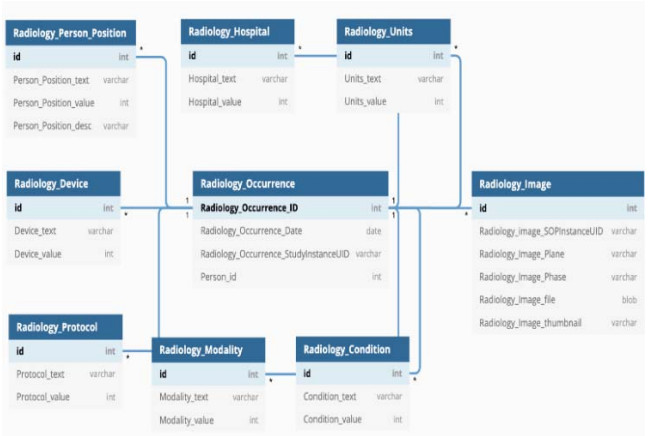
본 논문에서 제안하는 의료영상의 표준화와 데이터셋의 생성 및 검색, 인공지능 학습을 위한 데이터 셋을 제공하는 웹 기반의 관리 시스템 구조는 다음 그림 1 과 같다. React UI Library 기반의 Web Client 와 Python Django Rest Framework 기반의 API 서버를 설계하였다. 또한, 각 기관에서 발생되는 대량의 의료 데이터를 수집하기위해 Nginx 웹 서버와 Message Queue, Task Worker 을 통해 비동기 분산 업로드 방식을 도입하였다.



(그림 1) 웹 기반 의료영상정보 관리 시스템

3-1. 의료영상정보 표준화를 위한 데이터베이스 설계

본 논문에서 제안한 의료영상정보 표준화를 위한 DB 의 설계는 다음 그림 2 와 같다. OHDSI CDM 기반의 의료영상 표준화를 위해 DICOM 메타태그정보로부터 추출되는 데이터 셋의 촬영 정보를 저장하기 위한 Radiology Occurrence 테이블과 각 데이터 셋에 포함된 이미지들에 대한 정보를 저장하는 Radiology Image 테이블로 설계하였다. 또한, 각 데이터 셋의 정보를 표준화하기 위해서 병원 별 촬영 조건이 담긴 Radiology Protocol, 영상이 어떤 질환에 대한 영상인지를 판단할 수 있는 Radiology Condition, 현재 영상이 환자의 어떤 자세로 촬영된 건지 판단할 수 있는 Radiology Person Position, 촬영된 Modality 를 판단할 수 있는 Radiology Modality, 의료영상의 각종 단위를 표시하는 Radiology Units, 영상에 촬영한 장비를 표시하는 Radiology Device, 영상이 촬영된 병원을 표시하는 Radiology Hospital 정보 등 임상연구에 필요한 정보를 저장하기 위한 테이블로 설계하였다.

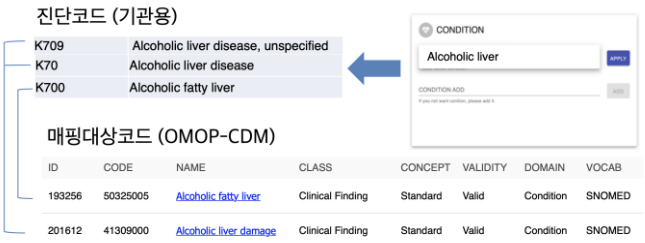


(그림 2) R-CDM 을 위한 데이터베이스 스키마

4. 결과

4-1. 의료영상 데이터의 수집

구축된 시스템의 의료영상 데이터의 진단명은 그림 3 과 같이 사용자에 의해 입력된 값을 표준화 코드(snomed)로 매핑하여 DB 상에 Condition ID 로 입력된다. 표준화 코드와의 매핑을 통해 OMOP-CDM 의 표준화 데이터와 수집된 의료영상 데이터를 함께 사용하여 연구에 사용할 수 있다 [10].



(그림 3) 진단명과 CDM 표준화 코드의 매핑

4-2. 표준화를 위한 DICOM 메타태그정보 추출

본 논문에서 제안한 의료영상정보의 표준화를 위해 DICOM 파일의 메타태그정보를 사용한다 [11]. 표준화를 위해 필요한 메타태그정보는 다음 Radiology Occurrence 을 위한 표 1 과 Radiology Image 객체를 생성하기 위한 표 2 와 같다. Occurrence 는 환자정보, 기관, 프로토콜과 같은 데이터 셋을 구분하기 위한 태그 정보로 구성되어 있다. Image 는 각 DICOM 파일의

의료영상정보를 위한 메타태그정보로 구성이 되어있다.

<표 1> Occurrence 을 위한 DICOM 메타태그정보

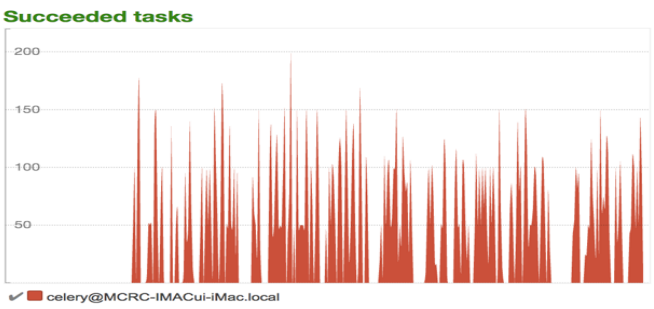
DICOM Tag Number	DICOM Tag Name
(0010, 0010)	Patient Name
(0010, 0020)	Patient ID
(0008, 0030)	Study Time
(0008, 0020)	Study Date
(0010, 1010)	Patient Age
(0010, 0040)	Patient Sex
(0008, 0033)	Content Time
(0018, 5101)	View Position
(0018, 0087)	Magnetic Field Strength
(0008, 1010)	Station Name
(0008, 1030)	Protocol Name
(0018, 0060)	KVP
(0008, 0060)	Modality
(0018, 1150)	Exposure Time
(0010, 4000)	Patient Comments
(0020, 000D)	Study Instance UID

<표 2> Image 을 위한 DICOM 메타태그정보

DICOM Tag Number	DICOM Tag Name
(0028, 0010)	Rows
(0028, 0011)	Columns
(0008, 0008)	Image Type
(0028, 1050)	Window Center
(0018, 0050)	Slice Thickness
(0008, 0031)	Series Time
(0020, 0011)	Series Number
(0008, 0032)	Acquisition Time
(0020, 0012)	Acquisition Number
(0008, 103E)	Series Description
(0020, 0037)	Image Orientation (Patient)
(0020, 0013)	Instance Number
(0008, 0018)	SOP Instance UID

4-3. 표준화 작업

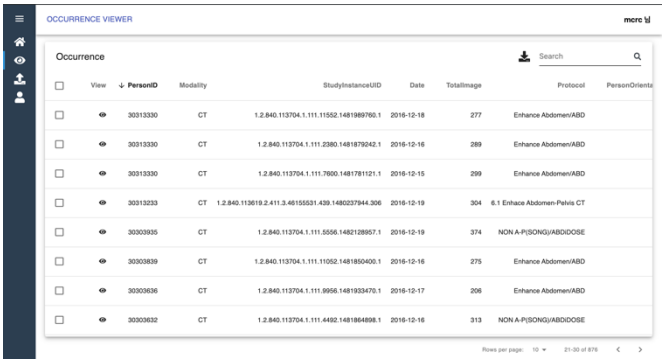
논문에서 제시하는 표준화 작업은 DICOM 메타태그정보를 추출 및 Radiology Occurrence, Radiology Image Object 을 생성하는 작업과 Pixel Data(7FE0, 0010) 메타태그정보를 통해 PNG 파일생성작업을 포함한다. 표준화 작업의 성능을 측정한 결과 초당 100~150 개의 처리량을 그림 4 와 같이 보이고 있다.



(그림 4) 표준화 작업 처리량

4-4. Radiology Occurrence Viewer

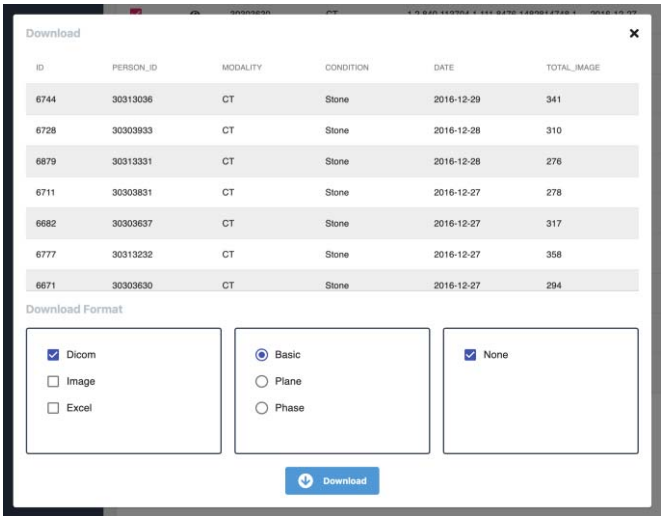
표준화된 데이터의 검색을 위한 웹 기반의 관리시스템에서 제공하는 사용자인터페이스는 다음 그림 5 와 같다. Dashboard 형태의 Radiology Occurrence Viewer 는 각 기관으로부터 수집된 데이터를 CDM 기반의 표준화된 Radiology Occurrence 데이터 셋을 나타내고, 사용자가 원하는 조건(질환별, 디바이스, 모달리티 등)으로 데이터를 검색할 수 있다.



(그림 5) Radiology Occurrence Viewer

4-5. 인공지능 기반의 임상 연구를 위한 데이터 셋 다운로드

기존에는 인공지능 연구를 위한 데이터 셋을 구축하기 위해서는 HIS 또는 임상연구목적의 CDW 등을 활용하여 수집하였다. 최근 임상시험을 위한 e-CRF 시스템이 자동화 파이프라인을 가진 시스템으로 대안이 되었으나 중앙집중식 시스템으로 의료정보 관리의 문제점이 도출되었다. 또한 연구자가 원하는 형태의 데이터 포맷을 구축하는데 어려움이 있었다. 데이터의 규모가 증가함에 따라서 사용자의 요구에 따른 데이터 셋을 자동으로 구축해주는 기능이 필요함에 따라서 그림 6 과 같이 Phase, Plane 형태에 따라 데이터 셋 다운로드 기능을 설계하였다. 분류 기준은 의료영상의 해부학적 평면에 따라서 Plane Mode 기능을 설계하였고, 의료영상의 촬영 시간에 따라서 Phase Mode 을 설계하였다. 데이터 셋 구축 기능을 제공함으로써 인공지능 연구를 수행하기 위한 기존의 데이터 셋 구축 문제점을 해결하였다.



(그림 6) 커스텀 데이터 셋 다운로드

5. 결론

본 논문에서는 의료영상정보의 표준화 및 인공지능 기반의 임상연구를 위한 데이터 셋의 수집 및 커스텀 데이터 셋을 제공하는 웹 기반의 관리시스템을 제안한다. 구축된 웹 기반 관리시스템을 통해 인공지능 기반의 임상 연구에 적용하기 위한 학습 또는 검증 그리고 테스트 데이터를 위한 데이터 셋을 제공함을 보였다. 제안한 시스템은 기존 CDM 과 연계하여 다기관 임상연구를 수행할 수 있을 것으로 기대된다. 향후 연구내용으로는 표준화 작업을 통해 변환된 각 의료영상 이미지를 웹 기반 관리시스템 상에서 다양한 이미지 툴들을 지원하여 분석 연구를 위한 이미지 뷰어 개발을 진행할 예정이다. 또한, 수집된 데이터를 활용하여 웹 기반 관리시스템 상에서 다양한 인공지능 학습 모델을 기반으로 데이터 셋을 학습시키고 분석 도구 등을 지원할 계획이다.

참고문헌

[1] G Hripcsak et al., “Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers”, in Stud Health Technology Information, Vol. 216, pp. 574-578, Aug 2015.

[2] FitzHenry F, Resnic FS, Robbins SL, et al., “Creating a common data model for comparative effectiveness with the observational medical outcomes partnership”, Appl Clin Inform, Vol. 6, No.3, pp. 536-547, Aug 2015.

[3] Voss EA, Makadia R, et al., “Feasibility and utility of applications of the common data model to multiple, disparate observational health

[4] Huser V1, Kahn MG, Brown JS, Gouripeddi R. “Methods for examining data quality in healthcare integrated data repositories,” Pac Symp Biocomput, Vol. 23, pp. 628-633, Apr 2018.

[5] W.Dean Bidgood, Jr., MD, MS, Steven C. Horii, MD, Fred W. Prior, PhD, and Donald E. Van Syckle “Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging,” Vol. 4, No. 3, pp. 199-212, May-Jun 1997.

[6] Adrian V. Dalca, Katherine L. Bouman, William T. Freeman, Natalia S. Rost, Mert R. Sabuncu, Polina Golland, “Medical Image Imputation From Image Collections,” IEEE transactions on medical imaging, Vol. 38, No. 2, pp. 504-514, Feb 2019.

[7] J Zhang, Y Xie, Q Wu, Y Xia, “Medical Image Classification Using Synergic Deep Learning,” Medical image analysis, Vol. 54, pp. 10-19, May 2019.

[8] OHDSI Forum, https://forums.ohdsi.org/t/oncology-radiology-imaging-integration-into-cdm/2018/7

[9] OHDSI/Radiology-CDM, https://github.com/OHDSI/Radiology-CDM

[10] Noh. Sihyeong, et al., “Medical Image Dataset for Machine Learning Based on OMOP CDM”, EMBC18

[11] Kohli, Marc D., Ronald M. Summers, and J. Raymond Geis. "Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session." Journal of digital imaging 30.4 (2017): 392-399.