다기관 임상연구를 위한 의료 데이터 셋 관리 시스템

이충섭[1], 김승진[1], 김지언[1], 노시형[1], 김태훈[1,3], 윤권하[1,2,3], 정창원[1,3]
[1]원광대학교 의료융합연구센터
[2]원광대학교 의과대학 영상의학과
[3]원광대학교병원 스마트헬스 IT 사업단

e-mail : {cslee99, koch369369, kakasky112, nosij123, tae_hoonkim, khy1646, mediblue}@wku.ac.kr

Medical Dataset Management System for Multi-Center Clinical Research

Chung-Sub lee[1], Seung-Jin Kim[1], Ji-Eon Kim[1], Si-Hyeong No[1], Tae-Hoon Kim[1,3],
Kwon-Ha Yoon[1,2,3], Chang-Won Jeong[1,3]
[1]Medical Convergence Research Center, Wonkwang University
[2]Dept of Radiology, Wonkwang University School of Medicine and Hospital
[3]Smart Health IT Center, Wonkwang University Hospital

요 약

본 논문은 국제표준화인 OHDSI OMOP-CDM 의 확장으로 의료영상 표준기반의 R_CDM 으로 변환하고 그 데이터를 기반으로 다기관 임상연구를 위한 관리시스템에 대해 기술한다. 이를 위해 기존 공통 데이터모델과 연계에 중점을 두어 DICOM 태그정보를 기반으로 의료영상 표준 모델의 스키마와 다기관 연구를 위한 Report 정보를 포함하여 모델링하였다. 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

제 4 차 산업혁명의 핵심 기술인 사물인터넷, 인공지능, 클라우드, 빅데이터는 의료 서비스의 패러다임을 변화시키고 있다[1]. 특히, 임상데이터기반의 인공지능(AI), 빅데이터 분석 관련 기업이 급성장하고 있다[2]. 최근 1 차병원과 2, 3 차병원간 진단과 처방 등 진료기록을 교류하는 시스템을 구축하는 사업을 국가적으로 추진하고 있다. 이러한 시스템 개발에 있어서 가장 중요한 요소는 표준화 연계모듈 고도화를 위해 국제표준화 용어(SNOMED_CT)를 사용한다. 이와 관련하여 OHDSI(Observational Health Data Science and Informatics)에서 제안하는 공통데이터모델(CDM)[3]은 임상데이터기반 연구를 위한 의료정보의 표준화에 대한 대표적인 모델이다. 이를 기반으로 다기관 공동연구 플랫폼으로 분산형 바이오헬스 빅데이터 플랫폼(FEEDER-NET)이 개발되어 국내외 공동연구가 활발하게 진행되고 있다[3]. 그동안 우리는 OMOP-CDM 을 기반으로 의료영상표준인 DICOM 의 태그 정보를 추출하여 메타데이터의 표준화와 의료영상데이터의 관리에 중점을 둔 R_CDM 을 제안하였다. 우리가 제안한 R_CDM 은 머신러닝 연구를 위한 표준화된 의료영상 데이터셋을 검색하여 다양한 형태로 생성할 뿐만 아니라 다기관 공동연구를 위한 표준화된 영상정보를 수집하고, 익명화된 데이터를 공유할 수 있다. 그러나 실제 활용하기 위해서는 수집된 의료영상기반의 데이터셋 뿐만 아니라 의료영상에 대한 설명을 리포트하는 기능이 요구되었다. 그리고 다기관 영상정보를 관리하기 위한 이질성 문제(DICOM 헤더 정보, 파일 확장자 등)를 해결해야 했다.

본 논문에서는 웹 기반으로 다자간 임상연구를 위한 의료영상 데이터 셋 관리 시스템에 대해서 기술한다.

2. 관련 연구

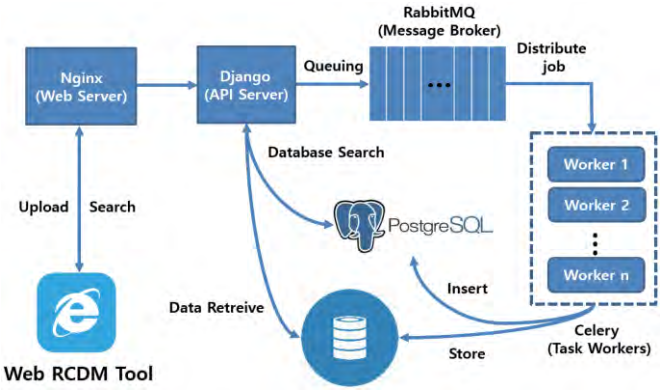
OMOP-CDM 은 정형화된 임상데이터를 표준화하는데 중점을 두고 있다. 그러나 최근 유전체, 영상 그리고 생체신호와 같은 비정형 데이터의 표준화로 확장하고 있다. 특히, 현재 각 병원에서는 PACS 를 사용하여 DICOM 국제 표준을 준수하여 저장하고 있으나 이러한 방대한 양의 데이터가 있더라도 실제 임상연구를 위해서는 각 질환 별로 최적화된 임상 프로토콜에 의한 선별, 핵심적인 의료영상에 저장되는 의료정보까지

[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, HI18C2383) and the Technology Innovation Program (or Industrial Strategic Technology Development Program(20001234).

표준화되어 저장되어야 한다[4]. 이와 관련하여 수행된 연구는 국내외에서도 미흡하며 더욱이 의료기관별 의료영상의 표준화된 정보 없이 인공지능 학습 연구에 적용하기에는 어려움이 있다[5]. 또한 인공지능 학습을 위해서는 방대한 양의 의료영상 데이터가 요구되며, 인공지능 알고리즘의 최적화에 필요한 검증 및 테스트 데이터 수집도 매우 어렵다. 이러한 문제점을 해결하기 위해 의료영상에 대한 표준화의 요구사항을 정리하였고, 기존의 OMOP-CDM 과 연계하여 확장 모델을 제시하였다[6]. 또한 다기관 공동연구를 위한 영상데이터에 대한 리포트 기술에 대한 요구사항에 따라 기존 제시한 R_CDM 관리 시스템을 개선하고자 한다.

3. 제안 시스템

본 논문에서 제안하는 다기관 임상연구를 위한 의료 데이터 셋 관리 시스템은 각 기관에서 수집한 데이터를 R_CDM 기반의 표준화된 데이터로 변환하여 함께 공유하고 해당 영상에 대한 Report 를 작성하여 다기관 공동연구가 가능하도록 개발되었다. 본 시스템의 구조는 다음 그림 1 과 같다.



(그림 1) 다기관 의료 데이터셋 관리 시스템

React UI Library 기반의 Front-End (Web Client)와 Python Django Rest Framework 기반의 Back-End (REST API Server)를 설계하였다. 또한, 각 기관에서 발생되는 대량의 의료 데이터를 수집하기 위해 Nginx 웹 서버와 Message Queue, Task Worker 을 통해 비동기 분산 업로드 방식을 도입하였다. 의료기관의 데이터 관리 시스템은 환자 정보를 비식별화 해야하기 때문에 익명화(Anonymize)를 지원하고 Client 와 Server 간의 통신프로토콜을 암호화하여 전송된 환자 정보 및 데이터에 대한 보안을 유지하도록 SSL:보안 소켓 계층(Secure Sockets Layer) 프로토콜을 지원하고 있다.

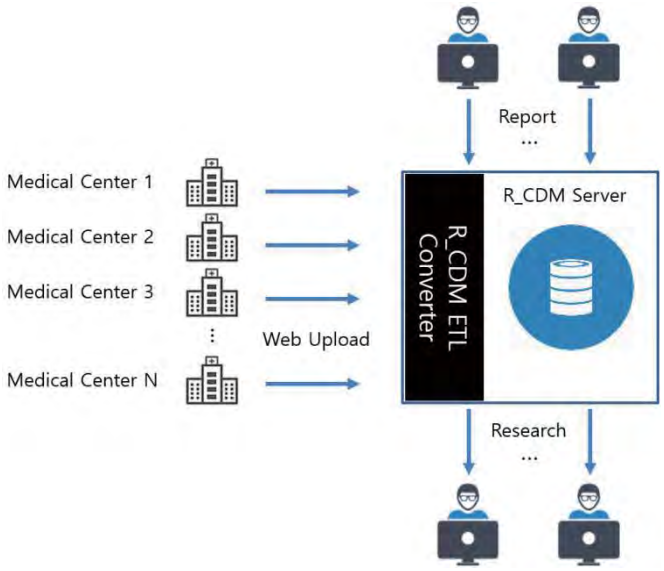
3-1. 의료영상정보 표준화를 위한 데이터베이스 설계

본 논문에서 제안한 다기관 임상연구를 위한 의료 데이터 셋 관리 시스템의 DB 설계는 다음 그림 2 와 같다. 데이터베이스는 크게 Radiology CDM 기반의 의료영상 표준화를 위해 DICOM 태그 정보로부터 추출되는 데이터 셋의 촬영 정보를 저장하기 위한 Radiology Occurrence 테이블과 각 데이터 셋에 포함된 이미지들에 대한 정보를 저장하는 Radiology Image 테이블로 설계하였다.

(그림 2) R_CDM 을 위한 데이터베이스 스키마

또한, 각 데이터 셋의 정보를 표준화하기 위해서 병원 별 촬영 조건이 담긴 Radiology Protocol, 어떤 질환에 대한 영상인지를 판단할 수 있는 Radiology Condition, 어떤 자세로 촬영된 지 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, 해당 영상의 임상적 의견을 관리하는 Radiology Report 등 임상연구에 필요한 정보를 저장하기 위한 테이블로 설계하였다

3-2. 다기관 의료영상 데이터 수집



(그림 3) R_CDM Workflow

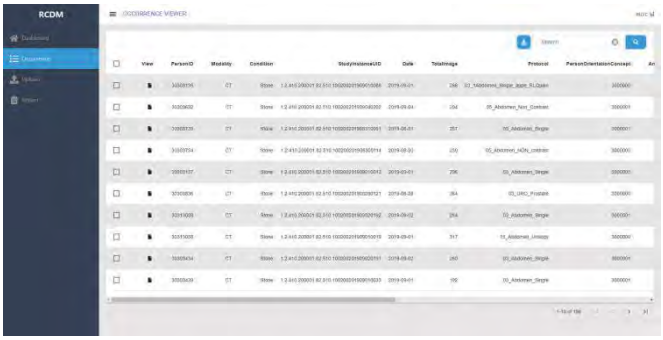
웹 기반으로 구축된 시스템은 다 기관으로부터 의료영상 데이터를 그림 3 과 같이 R_CDM 으로 변경하여 데이터를 수집할 수 있다. R_CDM 으로 표준화하여 저장된 데이터는 판독의로부터 해당 영상의 Report 를 작성하여 소견을 작성할 수 있다. 이러한 소견을 바탕으로 특정 질환군에 세부적인 검색 조건이 되어 연구자들에 의해서 필요한 데이터셋을 생성하여 연구에 사용될 것으로 전망된다

표 1 과 같이 Radiology Report 테이블은 Radiology Occurrence Table 과의 연결성을 같기 위해 Study Instance UID 를 Key 로 관리되고 있고 Report 생성일, Report 결과에 대한 컬럼을 포함한다.

<표 1> Report 를 위한 임상 정보

Table Name	Column	Remarks
Report Table	Study Instance UID	Occurrence Table
Report Table	Modality	Occurrence Table
Report Table	Study Date	Occurrence Table
Report Table	Report Create Date	Create
Report Table	Report Approval Date	Create
Report Table	Report Text	Create

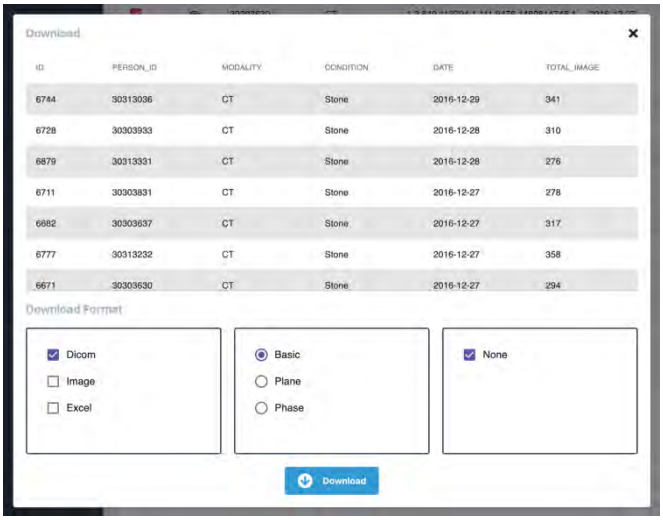
Radiology Occurrence List 는 각 기관으로부터 수집된 데이터의 전체를 그림 4 와 같이 보인다. 또한 사용자가 원하는 조건(질환별, 디바이스, 모달리티 등)으로 검색할 수 있다. 좀더 확장된 검색 기능으로 특정 키워드의 일부분만 입력해도 해당 Occurrence 를 찾고 멀티 키워드를 입력해도 해당 키워드에 맞는 Occurrence 를 검색할 수 있는 기능을 제공한다. 또한 해당 Radiology Occurrence List 에서 검색한 결과를 데이터셋으로 생성하여 다운로드 받을 수 있다.



(그림 4) Radiology Occurrence List

3-3. 데이터 셋 생성 및 다운로드

의료영상을 기반으로 인공지능 연구를 위해서는 해당 연구목적에 맞는 데이터 셋을 확보하여 반복적인 학습을 통해 알고리즘을 개발한다. 그리고 개발된 알고리즘에 대해서 Internal/External 검증을 통해 마무리한다. 최근 임상시험을 위한 e-CRF 시스템이 자동화 파이프라인을 가진 시스템으로 대안이 되었으나 매번 연구 종료 함께 데이터의 재활용이 불가능하다. 또한 연구자가 원하는 형태의 데이터 포맷으로 생성하기에는 어려움이 있다. 데이터의 규모가 증가함에 따라서 사용자의 요구에 따른 데이터 셋을 자동으로 생성하는 기능이 필요하다. 그림 5 와 같이 Phase, Plane 형태에 따라 데이터 셋 다운로드 기능을 설계하였다. 분류 기준은 의료영상의 해부학적 포지션에 따라서 Plane Mode 기능을 설계하였고, 의료영상의 촬영 시간에 따라서 Phase Mode 를 설계하였다. 데이터 셋 생성 기능을 제공함으로써 인공지능 연구를 수행하기 위한 필요한 데이터 셋을 제안한 시스템을 통해서 해결할 수 있다.



(그림 5) 커스텀 데이터 셋 다운로드

3-4. 다기관 연구를 위한 Report 관리

논문에서는 다기관에서 수집된 표준화된 의료영상에 임상적 의미를 부여하기 위해서 그림 6 과 같이 Report 입력을 제공하고 있다. 또한 해당 영상을 공유한 연구자들은 해당 영상의 Report 를 확인할 수 있다.

(그림 6) Report 입력 및 뷰 다이얼로그

4. 결론

본 논문에서는 다기관 연구를 위한 의료영상정보의 표준화와 인공지능 기반의 임상연구를 위한 데이터 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 제공할 수 있음을 보였다. 그리고 기존 CDM 과 연계하여 다 기관 임상연구를 수행할 수 있는 Report 입력을 보였다. 향후 연구내용으로는 표준화 작업을 통해 변환된 각 의료영상 이미지를 다기관 연구를 위한 각 기관별 통계를 보여주고 웹 기반 관리시스템 상에서 다양한 정량화 분석 툴들을 지원하여 다기관 분석 연구를 위한 이미지 뷰어 개발을 진행할 예정이다. 또한, 수집된 데이터를 활용하여 웹 기반 관리시스템 상에서 다양한 인공지능 학습 모델에 생성된 데이터 셋을 학습시키고 최적의 알고리즘 개발을 지원하는 실증 연구를 수행할 계획이다.

참고문헌

[1] 4 차 산업혁명 대정부 권고안, https://www.4th-ir.go.kr/

[2] 박성욱, “빅데이터 기법을 활용한 Data Technology 의 키워드 분석”, 기술혁신학회지, 제 22 권, 2 호 pp. 265~281.

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

[4] 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.

[5] 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.

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