

보건의료 영상정보 데이터의 가명처리 및 데이터의 활용 방안에 대한 연구

송현목, 김경호, 우정훈, 오현우*, 박종욱*

건양대학교, *건양대학교

shm1029@misoinfo.co.kr, ghkim@misoinfo.co.kr, woojh@misoinfo.co.kr,

*hwoh@konyang.ac.kr, *jupark@konyang.ac.kr

De-identification Processing and Data Utilization of Health and Medical Image Information

Song Hyun Mok, Kim Gyeong Ho, Woo Jeong Hoon, Oh Hyun Woo*, Park Jong Uk*

*Konyang Univ.

요약

보건의료 데이터의 활용은 90년대부터 지금까지 보건의료 산업의 중요한 과제이다. 최근 데이터 3법 개정, 보건의료 5대 빅데이터 플랫폼 가동 등으로 보건의료데이터의 활용 기반이 갖춰졌다. 하지만 개인정보 보호법, 생명윤리법 등과 같은 법령 제한에 의해 데이터를 활용하기 위해서는 개인정보를 비식별화하는 데이터 가명화가 필수적으로 동반되어야 한다. 따라서 본 연구에서는 가명화가 필요한 데이터의 가명화 및 활용 방법에 대해 연구를 진행했다. 보건의료 활용 가이드라인에서 제시한 데이터의 속성 중 영상정보 DICOM에 초점을 맞췄다. 가이드라인 기준을 준수해 DICOM을 가명화하는 것과 동시에 DICOM 데이터의 활용을 위해 DICOM 메타데이터 및 DICOM 영상의 텍스트 데이터를 추출했다. 본 연구는 추출한 데이터들은 가명화와 더불어 데이터 활용 방안에 대해 연구를 진행했다.

I. 서론

보건의료 데이터 활용을 위한 노력은 90년대 진료기록 관리, 건강보험 청구 전산화에서 시작되었다. 그 후 15년까지 데이터 교류, 표준화 등 의료기관 정보화를 추진하였고, 건강보험 데이터 개방, 보건의료빅데이터 플랫폼 구축 등 보건의료 데이터 활용을 위한 논의가 본격화 되었다. 그리고 최근에는 데이터 3법 개정(가명정보 활용), 보건의료분야 5대 빅데이터 플랫폼 가동 등 보건의료데이터 활용 여건 및 기반 조성 단계로 진입하였다.[1] 하지만 보건의료데이터 활용 여건이 조성되었다 하더라도 개인정보 보호법, 생명윤리법 등과 같은 법령의 제한에 의해 개인정보가 포함되어 있는 데이터는 가명정보 처리 가이드 라인을 준수해 데이터의 가명화가 필요하다[2] 본 연구에서는 보건의료데이터 활용을 위해서 데이터 가명화 방법과 그 활용 방안에 대해서 연구했다.

II. 본론

본 연구에서는 보건의료 분야에서 의료영상을 표준 규격

에 맞게 저장, 가공, 전송하는 시스템(PASC : Picture Archiving and Communications System)에서 이미지형 데이터들의 저장에 활용되는 DICOM파일을 선정하였으며, 보건의료데이터 활용 가이드라인에서 제시한 가명화 대상 데이터를 DICOM을 기준으로 테스트 및 비식별화를 진행하고, 나아가 영상정보 데이터의 활용도를 높이기 위하여 DICOM 영상정보에서 텍스트를 추출하였다. 영상정보에서 텍스트를 추출하기 위하여 광학 문자 인식(OCR, Optical character recognition) 기술을 활용했다.

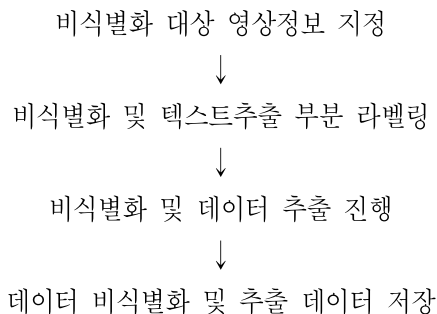
표 1은 DICOM 파일을 생성할 경우 같이 저장되는 데이터의 속성 값이며 속성 값들의 각 정보들 중 영상정보의 비식별화 처리를 위하여 DICOM 파일의 JPG, PNG 등의 파일로 형 변환을 진행하였으며, 형 변환을 진행할 시 기존 속성 값들은 그 과정에서 삭제가 되기 때문에 별도의 조치를 취하지 않았다. 비식별화 필요 정보들은 형변환 과정에서 삭제되며 영상정보 상 개인정보를 비식별화 조치하기 위한 라벨링을 지정하고 같거나 혹은 비슷한 서식지를 모아 라벨링 처리

를 진행한다.

데이터 속성값	비식별화 방안
측정수치 정보	▪ 별도의 조치 불요
의료인의 관찰·입력 정보	▪ 별도의 조치 불요
알고리즘이 생산한 건강정보	▪ 별도의 조치 불요
영상정보	<ul style="list-style-type: none"> ▪ 영상정보 상 환자번호·성명·생년월일 등 개인정보 일부를 삭제하거나 전부를 대체 ▪ DICOM 헤더 등 메타데이터 상의 개인 정보 일부를 삭제하거나 일부 또는 전부를 대체 ▪ 영상정보 이미지 상의 특정 개인을 식별할 수 있는 신체적 특징, 문신, 보형물 등은 삭제 또는 모자이크, 표면 가장자리 삭제 등의 방법으로 처리
음성정보	▪ 비식별화 가능여부 유보

표 1 보건의료데이터 활용 가이드라인, 데이터 유형별 비식별화 방법 중 일부

1. DICOM 비식별화 및 데이터 추출 프로세스



비식별화 데이터를 추출하기 위한 프로세스를 정립하고 정립된 프로세스를 활용하여 비식별화 처리를 진행하였으며, 비식별화 지정 및 설명은 아래 2. 비식별화 대상 처리를 참조한다.

2. 비식별화 대상 처리

비식별화를 위한 DICOM 이미지는 환자의 폐 흉부 사진을 JPG 또는 PNG 로 형변환 시킨 데이터를 선정하였으며, 영상의 좌측 우측 상·하단의 라벨링을 진행할 수 있는 이미지를 선정하였다.

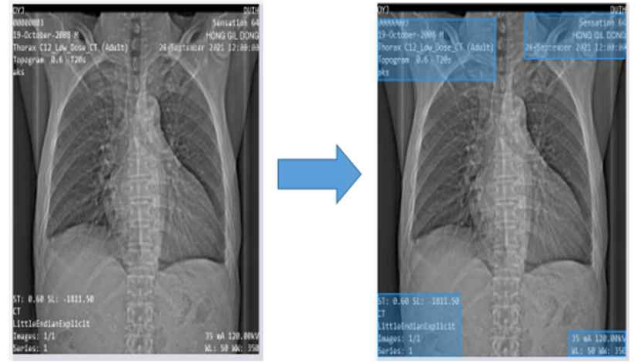


그림 1 비식별화 처리를 위한 라벨링 작업

3. 데이터 비식별화 및 추출 데이터 저장

라벨링 이후 데이터 비식별화 및 텍스트 추출을 진행했다. 텍스트 추출을 먼저 진행해 해당 위치의 텍스트를 추출했으며, 그 이후 데이터 비식별화를 진행해 비식별화 라벨링 부분을 모자이크 처리했다.

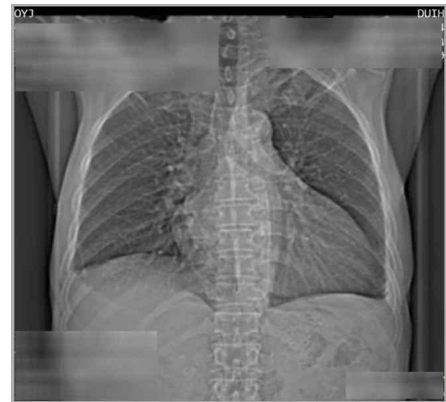


그림 2 비식별화 처리 결과

III. 결론

본 연구에서는 데이터 3법이 개정됨에 따라 그동안 개인 정보로 인하여 내부 데이터베이스에서만 사용되던 이미지형 데이터들을 비식별화 작업을 거쳐 외부에서 사용할 수 있도록 하였다.

비식별화(가명화처리)를 진행하면 환자의 개인정보를 확인할 수 없기 때문에, 개인정보 보호를 수행하며, 아울러 타 병원과의 협업 및 데이터 취합, 활용 등 지금까지 연구할 수 있던 국한된 자료를 더욱 폭넓게 연구 및 활용할 수 있으며, 더 나아가서는 모든 병원의 데이터를 취합하여 더욱 정확한 통계자료 및 연구자료로 활용하고 의학의 발전 가능성을 확인하였다.

비식별화 후 재식별화를 통하여 추후 인가된 사용자는 어떤 환자의 자료인지도 확인하여 더욱 정확한 진료를 가능하게 하며, 기초 데이터를 쌓아 추후에는 인공지능(AI)도입으

로 자동으로 비식별화를 진행할 수 있는 발판을 마련할 수 있을 것으로 생각된다.

참고문헌

- [1] 보건복지부, 보건의료 데이터·인공지능 혁신전략
- [2] 보건복지부 및 개인정보보호위원회, 보건의료데이터 활용 가이드라인
- [3] 백종일, 송경택, 최원균, 유기근, 이필우, 인한진, 김철중, 여광수, and 김순석. "의료 이미지 데이터의 비식별화 방안에 관한 연구." 예술인문사회융합멀티미디어논문지 6.6 (2016): 103-110.
- [4] 이종석, 정민수, 조재찬, 낭종호. (2022). 개인 정보 비식별화를 위한 딥러닝 기반 이미지 및 이미지 특징 벡터 생성 방법. 한국정보과학회 학술발표논문집, (), 1005-1007.
- [5] 조영탁, and 안기옥. "수술 동영상의 비식별화를 위한 개인식별정보 자동 검출 시스템 설계 및 구현." 융합보안 논문지 19.5 (2019): 75-84.