OMOP CDM 기반 의료 데이터 ETL 툴 개발

한만욱[1], 이푸름[2], 이호웅[3]
[1]호서대학교 컴퓨터공학부 학부생
[2]호서대학교 컴퓨터공학부 학부생
[3]호서대학교 컴퓨터공학부 교수
aksdnr507@gmail.com, lpm0831@naver.com, always14@gmail.com

Development of A HealthcareData ETL Tool Based on OMOP CDM

Man-Uk Han[1], Pureum Lee[2], Ho-Woong Lee[3]
[1]Dept. of Computer Science, Hoseo University
[2]Dept. of Computer Science, Hoseo University
[3]Dept. of Computer Science, Hoseo University

요 약

디지털 헬스케어 서비스 활성화에 따라 디지털 의료 데이터의 양은 매년 급속하게 증가하고 있으며, 의 데이터의 상호 교환과 연동을 위한 다양한 CDM(Common Data Model)이 개발되고 있다. 그러나, 의료 데이터 교류에 대한 요구가 증가하면서, 기존 레거시 시스템의 데이터를 CDM으로 변환하기 위한 추가적인 비용이 소요될 수 밖에 없다. 이에 본 연구에서는 OMOP CDM (Observational Medial Outcomes Partnership Common DataModel) 기반 의료 데이터 ETL (Extract, Transform, Load) 툴을 개발하였다. OMOP CDM ETL 툴은 기존의 레거시 데이터베이스 정보를 CDM으로 변환할 수 있는 효과적인 료인터페이스를 제공함으로써, 디지털 의료 데이터 공유와 관리 및 분석의 효율성을 증대할 수 있을 것이다.

1. 서론

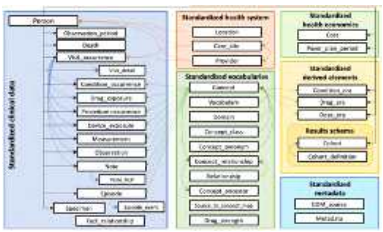
의료 데이터 교류의 중요성이 증가함에 따라, 의료 데이터 상호 교환과 연동을 위한 다양한 CDM이 개발되었다[1]. 그러나, 기존 레거시 시스템 데이터를 CDM 데이터로 변환하는데 추가적인 비용이 소모될 수 밖에 없다. 이에 본 논문에서는 CDM 데이터로 변환하는데 필요한 과정을 효율적으로 간략화하여, OMOP CDM 기반 의료 데이터 ETL 툴을 개발하여 변환 비용을 절감하였다.

2. 관련연구

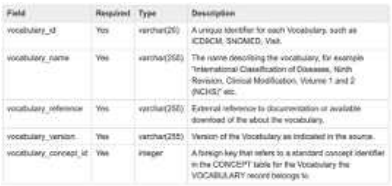
2.1 OMOP CDM

OMOP CDM은 미국 식품의약국(FDA)에서 주관하던 프로젝트로서 OMOP에서 개발되었다[2].

OMOP CDM은 의료 데이터를 여러 도메인으로 구분하고, 각 도메인에서 발생하는 다양한 이벤트 및 관찰을 표현한다. 또한 풍부한 의료 용어집(Vocabulary)을 포함하여 다양한 의료 용어를 표준화된 코드 체계로 매핑하여 의료 시스템이나 데이터 소스에서 발생하는 용어의 다양성을 관리하고 의미론적 일관성, 통일성을 유지한다[3].



(그림 1) OMOP CDM 도메인



(그림 2) OMOP CDM vocabulary

2.2 White Rabbit

White Rabbit은 OMOP CDM 의료 데이터베이스 ETL을 돕기 위한 소프트웨어 도구이며, 소스 데이터를 스캔하여 테이블, 필드 및 값에 대한 자세한 정보를 제공하고 ETL을 설계할 때 참고할 수 있는 보고서를 생성한다[4].

2.3 Rabbitin a Hat

Rabbitin a Hat은 White Rabbit과 함께 제공되며 White Rabbit의 스캔 문서를 읽고 그래픽 사용자 인터페이스(GUI)를 통해 사용자가 소스 데이터를 CDM 내의 테이블 및 열에 연결할 수 있도록 한다 [5].

3. OMOP CDM 기반 데이터 통합 ETL 툴 개발

3.1 개발 환경

개발을 위한 환경은 다음 <표1>과 같다.

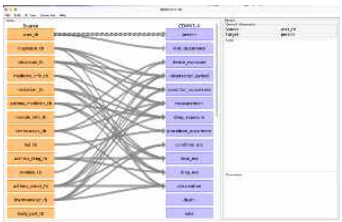
<표 1> 개발 환경

	버전
CDM	OMOP CDM v5.4
Java	Amazon Corretto Version 11.0
IDE	Intellij IDEA 2023.2
데이터베이스	PostgresSQL 14.2, MariaDB 10.9.3
빌드도구	Gradle 7.1
기타 도구 / 프레임워크	Lombok 1.18.26, Log4jdbc 1.16

3.2 ETL 툴 개발

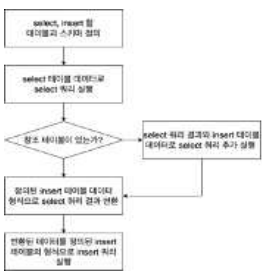
기존의 데이터 소스를그림 3, 그림 4 과 같이 White Rabbit 과 Rabbit in a Hat을 사용하여 시각화 한다.

(그림 3) White Rabbit



(그림 4) Rabbit in a Hat

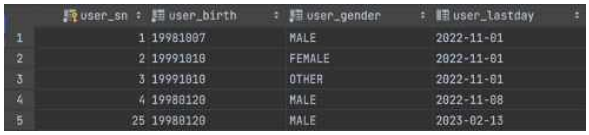
시각화 된 정보를 바탕으로표준화가 필요한 테이블과 스키마를 코드로 작성한다. 작성된 코드로 테이블과 스키마를 정의하고, 이를 해당하는 OMOP CDM 구조와 용어로 변환한다. 변환된 결과를 OMOP CDM 데이터베이스에 적재한다.



(그림 5) ETL 툴 개발 흐름도

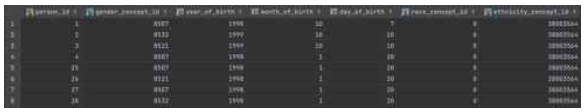
4. 개발 결과

ETL 툴을 통해, 다양한 형식의 의료 데이터를 효율적으로 OMOP CDM 구조로 변환할 수 있었다. 변환 된 데이터는 일관성과 표준화된 용어 사용을 통해 데이터의 정확성과 신뢰성을 높였다.



	user_sn	user_birth	user_gender	user_lastday
1	1	19981007	MALE	2022-11-01
2	2	19991010	FEMALE	2022-11-01
3	3	19991010	OTHER	2022-11-01
4	4	19980120	MALE	2022-11-08
5	25	19980120	MALE	2023-02-13

(그림 6) 기존 데이터 소스



(그림 7) 적재된 OMOP CDM 데이터 소스

5. 결론

본 연구에서 개발한 ETL 툴은 간결한 정의를 기반으로 의료 데이터를 CDM 데이터로 변환하여 효율적인 데이터 통합이 가능했다. 이를 통해 레거시 시스템 데이터을 CDM 데이터로의 변환을 제공함으로써 의료 분야의 효율적인 데이터 공유와 관리 및 분석이 이루어질 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부와 정보통신기획평가원의SW중심대학사업의 연구결과로 수행되었음 (2019-0-01834)

참고문헌

[1] FitzHenry, F., Resnic, F. S., Robbins, S. L., Denton, J., Nookala, L., Meeker, D., ... & Matheny, M. E., Creating a common data model for comparative effectiveness with the observational medical outcomes partnership. *Applied clinical informatics*, *6*(03), 536-547. 2015

[2] Yu, Y., Zong, N., Wen, A., Liu, S., Stone, D. J., Knaack, D.,... & Jiang, G., Developing an ETL tool for converting the PCORnet CDM intothe OMOP CDM to facilitate the COVID-19 data integration., Journal ofBiomedical Informatics, 127, 104002, 2020

[3] https://www.ohdsi.org/data-standardization/

[4] https://www.ohdsi.org/analytic-tools/whiterabbit-for-etl-design/

[5] https://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html