임상데이터웨어하우스 기반 공통데이터모델 구축 및 데이터 추가 경험

고수정¹, 박상준², 장동진³

¹가톨릭대학교 의과대학 의료정보학교실, ²서울대학교 의과대학 분당서울대학교병원 안과, ³가톨릭대학교 의과대학 여의도성모병원 안과

Experience of Converting Clinical Data Warehouse to Common Data Model and Additional Data Loading

Soo Jeong Ko¹, Sang Jun Park², Dong-Jin Chang³

¹Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul; ²Department of Ophthalmology, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam; ³Department of Ophthalmology, Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea

Correspondence to:

Dong-Jin Chang

Department of Ophthalmology, Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 10 63(yuksam)-ro, Yeongdeungpo-gu, Seoul 07345, Korea
Tel: +82-2-3770-1612
Fax: +82-2-761-6869
E-mail: hpalways@catholic.ac.kr

Received: May 11, 2021

Revised: October 10, 2021

Accepted after revision: October 29, 2021

Background: Clinical Data Warehouse (CDW) enables healthcare organizations to allow investigators to utilize their anonymous data securely. The Common Data Model (CDM) enables collaborative research between institutions. The authors describe the experience of converting CDW data into the Observational Medical Outcomes Partnership (OMOP) CDM (OMOP-CDM).

Methods: Data from electronic health records of a medical center was converted into OMOP-CDM ver. 5.3.1 (<https://www.ohdsi.org/data-standardization/the-common-data-model/>). Source codes of the institution were mapped to the codes in the standard vocabularies. Predefined data were extracted, transformed, and loaded into the CDM tables. Ophthalmologic data were digitalized and structured per their data configuration.

Results: The number of total patients with available data in the CDW was 8,532,392. The data from 878,141 patients were converted into CDM structure after the codes were mapped to those in the standard vocabularies. Visual acuity, intraocular pressure, automated visual field test, and optical coherence tomography were digitalized and stored.

Conclusion: Although it requires plenty of resources to convert the data set to CDM, it is worthy of keeping the activity since it provides opportunities for investigators to participate in various studies.

Keywords: Common Data Model; Clinical Data Warehouse; Observational Medical Outcomes Partnership

서론

빅데이터는 통용되는 데이터 수집, 관리 소프트웨어의 한계를 넘어서는 크기의 데이터를 말한다[1]. 디지털과 모바일 커뮤니케이션이 발전하면서 세상은 더 연결되고, 추적 가능하게 되었으며, 그 결과 큰 규모의 데이터를 구축할 수 있게 되었다[1]. 의료 영역에서도 전자의무기록(electronic medical record)의 보급을 통해 전 세계적으로 많은 수의 환자에 대한 데이터가 다양한 형태로 수집, 저장되고 있으며, 이를 의료 빅데이터로 칭하고 있다. 의료데이터는 그 생성의 일차적 목적에 따라 크게 세 가지로 분류할 수 있다[2]. 첫째, 직접적으로 특정 영역의 연구활동을 용이하도록 한 것이다. 그 예로, 설문조사, 레지스트리 정보, 혹은 임상시험자료 등이 있다. 둘째, 진료행위를 지원하기 위한 것이다. 전자의무기록에서 생성되는 의료데이터가 대표적인 예이다. 셋째, 의료비의 지급 관리를 위한 청구데이터이다. 진료 및 청구데이터는 비록 임상연구를 위해 수집되지 않았다 하더라도 임상연구에 충분히 활용할 수 있다. 다만, 상기 세 종류의 데이터는 데이터의 목적과 운용기관에 따라 대개 고유한 형식 및 구조로 이루어져 있다. 따라서 서로 다른 데이터 출처에서 연구결과를 도출하기 위해서는 데이터를 동일한 데이터 표준으로 변환할 필요가 있으며, 이는 공통데이터모델(Common Data Model, CDM)을 구현함으로써 가능해진다.

규제기관의 관점에서는 일반화 가능한 근거를 빨리 생성하고, 노출빈도가 낮은 요인에 대한 임상결과를 용이하게 수집하기 위하여 분산연구망을 구축할 필요가 있었고, 이를 충족시키기 위해 다양한 CDM이 고안되었다. 그 예로 미국 Food and Drug Administration (FDA)의 Sentinel, Patient-Centered Outcomes Research Network (PCORnet), Exploring and

Understanding Adverse Drug Reactions (EU-ADR), European Medical Informatics Framework, Vaccine Safety Datalink 등이 있다[3-5]. 이와 같은 의료 CDM은 형태에 따라 데이터구조화 모델(organizing data model), 데이터 매핑 모델(mapping data model), 적응형 규칙시스템(adaptive rule system)으로 분류할 수 있다[6]. 데이터 구조화 모델은 다수의 원천데이터를 하나의 표준화된 데이터 구조로 구성하는 합의된 규칙을 의미한다. 이는 환자의 의료기관 입원, 외래, 응급 등 의료기관의 의료이용 시의 진단정보와 처치 및 행위, 검사실 검사, 처방정보, 조제 정보 등의 정보를 구조화한다. 데이터 구조화 모델은 원자료의 유형과 순서를 구별하여 정렬하는 방법을 제공하기는 하지만, 기록된 정보를 해석하거나 다른 정보로의 변환 또는 매핑하는 규칙을 적용하지는 않는다. 예시로 FDA Sentinel CDM, i2b2, PCORnet 등이 있다[7-11].

데이터 매핑 모델은 표준화된 데이터 구조와 분석 가능한 변수 집합을 정의하기 위해 다양한 규칙을 적용하는 CDM을 의미한다. 이러한 형태의 CDM은 주어진 원천데이터에서 파생된 포괄적인 데이터 세트로 구성되며, 전체 범위의 의료이용 구조를 포함하고 다양한 변수를 수록한다. 따라서 한번 변환이 되면 데이터는 고정되고, 연구자가 사용할 수 있게 된다. 이러한 모델은 원자료에서 발견되는 용어나 정보 대신 분석을 위해 매핑된 데이터를 사용하도록 설계되어 있어, 연구자가 볼 수 있는 변수명은 늘 동일하다. 따라서 한번 변환이 되면 데이터는 고정되며, 이러한 의미에서 연구자에게는 데이터가 표준화된 것으로 보인다. 원천데이터의 CDM 변환을 위한 알고리즘은 원자료의 구조, 데이터의 양, 완전성에 따라 다양하게 적용해야 한다. 나이가 변환 알고리즘은 변환을 수행하는 조직과 변환 목적에 따라 서로 다를 수도 있다. 그

뿐만 아니라 데이터의 갱신이 있을 때는 데이터 변환을 다시 해야 하고, 원자료 데이터 구조가 바뀔 때는 새로운 방법으로 데이터 변환을 해야 한다. 이러한 CDM의 대표적인 예시로 Observational Medical Outcomes Partnership (OMOP)-CDM이 있다[2].

마지막으로 적응형 규칙시스템은 시간에 따라 배열된 종단 데이터를 해석하기 위한 규칙을 적용하여, 새로운 변수를 구성할 수 있는 논리 쿼리 언어로 구성되어 있다. 이 시스템은 사용자가 새로운 변수를 필요에 따라 결합하고 변경할 수 있으면서도 높은 투명성을 유지하는 특징이 있다. Sentinel CIDA (Cohort Identification and Descriptive Analysis)와 EU-ADR은 적응형 규칙시스템으로 분류된다[12,13].

국내 대학병원과 공공기관은 주로 데이터 매핑 모델의 일종인 OMOP-CDM을 채택하여 사용하고 있다. OMOP-CDM을 이용한 CDM이 많이 보급되면서 OMOP-CDM의 표준화된 분석방법과 분석도구를 적용할 경우 국내·외의 여러 기관의 데이터를 쉽게 분석할 수 있어 빠르게 대규모 다기관 연구를 진행할 수 있는 환경이 조성되고 있다. 다만, OMOP-CDM은 정해진 데이터 구조를 사용하여야 하므로, 전자의무기록시스템에 적재된 자료의 형태를 모두 수용할 수 없으며, 표준화된 용어를 사용하여야 하므로 전자의무기록 등에 기록된 자료에 대해 표준용어로의 매핑 작업을 선행하여 수행하여야 한다는 한계가 있다. 이 때문에 일부 의료기관은 임상데이터 웨어하우스(Clinical Data Warehouse, CDW)를 독자적으로 구축하여 사용하기도 하였다. 본 논문에서는 의료기관 CDW 유래의 CDM을 구축 경험과 안과 영역에서 해당 시스템을 개선하기 위한 사례를 소개함으로써 CDM을 이용하여 의료기관의 연구역량을 재고하고 그 과정에서 특정 진료과 혹은 연구영역의 연구요구를 충족시키기 위한 노력을 기술하고자 한다.

방 법

1. 원천데이터의 구성

원천데이터를 구성하기 위해 1997년 4월부터 2021년 3월까지 서울성모병원, 여의도성모병원, 은평성모병원, 의정부성모병원, 부천성모병원, 성빈센트병원, 인천성모병원 등 7개 병원에서 진료를 받은 환자의 기록을 익명화하였다. 익명화 데이터는 방문 기록, 진단정보, 수술정보, 처방정보, 투약정보, 진단 검사, 영상검사, 병리검사, 내시경검사, 기능검사, 인체유래물은행, 혈액투석, 서식기록, 수술기록, 안과 검사 테이블로 재구성되었다. 개별 연구자를 위해 연구자가 수행하는 연구의 자격조건(eligible criteria)에 맞는 코호트를 정의하고, 해당 코호트에 해당하는 환자를 검색하고, 검색된 환자들의 분포를 시각화할 수 있는 웹페이지를 구성하였으며, 연구자가 소속기관의 기관생명윤리위원회(institutional review board, IRB)와 데이터 활용위원회에서 연구계획서 승인을 받았을 때 데이터를 추출하는 등의 기능도 구현하였다. CDW의 구축은 전문개발업체에 의뢰하여 수행하였다.

2. 용어 매핑과 추출, 변환, 적재(extraction, transformation, load, ETL)

병원에서 사용하는 진단명과 약물, 처치, 수술, 검사 용어를 표준용어체계로 매핑하였다. 원내 용어 마스터를 ATHENA에서 제공하는 Korean Standard Classification of Diseases7 (KCD7), ICD-10 (10th revision of the International Statistical Classification of Diseases and Related Health Problems), RxNorm, Anatomical Therapeutic Chemical (ATC) classification system, SNOMED Clinical Terms (SNOMED-CT), Logical Observation Identifiers

Names and Codes (LOINC) 등 표준용어집을 기반으로 하고, Usagi를 이용하여 일차 매핑을 수행한 뒤, 그 결과에 대해 의료전문가의 리뷰를 시행하여 재확인하였다[14]. Usagi를 이용하여 용어를 찾기 어려운 경우에는 ATHENA 사이트에서 직접 검색을 통해 매핑하였다[15]. 의약품은 한국의약품안전관리원에서 발행한 공통데이터모델 구축을 위한 약물 용어 매핑가이드를 준수하였다[16]. OMOP-CDM ver. 5.3.1 (<https://www.ohdsi.org/data-standardization/the-common-data-model/>)을 변환대상으로 선정하였다. OMOP-CDM의 테이블과 각 필드, 그리고 필드의 내용 등의 세부 사항을 확인하여 이미 구축된 CDW 원천데이터의 테이블, 필드, 내용에 매칭하였다. 이 과정에서 7개 병원에서 사용하는 원내 용어집의 코드를 상술한 표준용어집의 개념 아이디(concept_id)에 매칭하였다. 이에 원내 용어집의 코드는 매핑된 표준용어집의 개념 아이디의 형태로 OMOP-CDM의 테이블과 필드에 재구성되었다. 데

이터 추출은 전문개발업체에 위탁하여 수행하였다.

3. 안과 데이터의 적재 형태 분석 및 데이터 추출

먼저 전자의무기록시스템에 적재된 안과 데이터의 저장 형태와 양을 분석하여, 안질환 연구를 위한 기본 데이터로 시력, 안압 측정이 포함된 안과기본검사와, 시야검사, 빛간섭단층촬영(optical coherence tomography, OCT)검사를 선정하였다. 시력은 전자 의무기록시스템의 시력 서식지를 통해 기록된 자료를 CDW에 적재하여 사용하였다. 안압은 진료기록 내부의 자유 텍스트에서 추출하기 위해 정규표현식을 작성하여 추출하였다. 시야검사와 빛간섭단층촬영검사는 두 단계로 검사결과를 추출하였다. 먼저, 이미지 내부 개체 검출을 위한 Faster R-CNN 알고리즘을 이용하여 결과의 이미지 내부에서의 위치를 자동 검출한 뒤 광학문자인식(optical character recognition) 알고리즘인 Tesseract를 이용한 코드를 작성하여 검사결과를 추출하였다.

그림 1. Clinical Data Warehouse 데이터 적재 현황. (A) 누적 방문 수. (B) 수술 처방 수. (C) 진단서. (D) 처방 수.

표 1. Clinical Data Warehouse 주요 테이블 적재 데이터 현황

구분	적재 기간(연도)	적재량(백만)
방문(입원, 외래)	1997-2021	118
진단	1997-2021	243
수술	2005-2021	4
처방	2005-2021	1,777
투약	2005-2021	517
진단검사	2005-2021	1,169
영상검사	2005-2021	37
병리검사	2006-2021	5
내시경검사	2005-2021	2
기능검사	2005-2021	14
인체유래물은행	2007-2021	4
혈액투석	2002-2021	6
서식기록	2008-2021	291
수술기록	2008-2021	1
안과검사	2008-2021	3

4. 안과 용어의 매핑

안과에서 사용하는 진단명, 검사명, 행위, 의약품 등 매핑된 용어 검토를 시행하였다. 안과 데이터 추출 항목 각각에 대한 용어를 ATHENA에서 검색하여 매핑하고, 일치하는 concept_id를 찾지 못한 경우 비표준 개념 아이디(non-standard concept_id)를 부여하였다.

표 2. 진단명 매핑 현황

구분	총진단 수	매핑 수(%)	상위 개념 매핑 수(%)
KCD6	60,610	55,743 (91.97)	35,834 (64.28)
KCD7	62,058	54,508 (87.83)	35,237 (64.65)
Other	138,611	132,650 (95.70)	108,755 (81.99)

KCD, Korean Standard Classification of Diseases.

결 과

1. 원천데이터의 구성

1990년 이후 출생 환자를 기준으로 하여, 7개 병원의 전자의무기록에서 CDW로 추출된 총환자의 수는 8,532,392명이었다. 전체 환자의 외래와 입원을 포함한 방문 건수는 일억천팔백만 건, 진단은 이억사천삼백만 건이었다. 방문과 진단정보는 1990년부터 누적된 정보였으나, 수술, 처방, 투약, 진단검사 등 해당 방문 때 이루어진 처방과 검사에 대한 정보는 전자의무기록시스템의 본격적인 도입이 이루어진 2007년부터 데이터의 누적이 나타나기 시작하였다(그림 1). 표 1은 CDW 내 주요 테이블의 데이터 적재 건수를 보여준다.

2. 용어 매핑과 ETL

CDW를 구성하는 원천진단 코드의 수는 268,090개였다. 내부 관리를 위한 6,811개의 코드를 제외하면, KCD6 유래 진단 60,610건, KCD7 유래 진단 62,058건, 기타 138,611건이었다. 의료기관의 원천진단명은 각 용어집의 진단명과 n:1 관계를 맺고 있었다. 예를 들어 KCD7의 ‘상세 불명의 노인성 백내장(H259)’은 ‘상세 불명의 노인백내장(H259.44G.000)’, ‘노인성백내장(상세불명의; H259.888.00)’ 등으로 존재하였다. KCD6 유래 진단명 중 91.97% (55,743개), KCD7 유래 진단명 중 87.83% (54,508개), 기타

95.69% (132,650개)의 진단명이 SNOMED-CT 진단명으로 매핑되었다. 정확히 일치하는 용어가 아닌 상위 개념의 진단명으로 매핑된 경우는 KCD6, KCD7, 기타 순으로 각각 35,834 (64.28%), 35,237 (64.65%), 108,755 (81.99%)이었다(표 2).

2020년 9월 1일 현재 “약제 급여목록 및 급여상한 금액표”에 등재된 약제는 총 24,414건이었고, 동일성분 동일제형 약제를 제외하면, 고유한 성분과 제형을 가진 약제는 4,527종이었다. CDW 구성 7개 기관은 성분과 제형이 동일한 의약품을 중복하여 원내에 도입하는 것을 지양하고 있으며, 작용기전이 같은 계열의 의약품은 최소한으로 도입한다는 원칙을 가지고 있다. 7개 기관의 의약품 코드는 총 6,674건이었고, 임상시험용 의약품 코드 2,040개를 제외하면 4,632개였다. 이 중 2,809개(60.6%)가 RxNorm (National Library of Medicine, Bethesda, MD, USA), Rx-Norm Extension (National Library of Medicine), ATC로 매핑되었으며, 해당 코드가 총 처방빈도의 91.5%를 담당하였다. 수술, 처치 코드는 총 4,525건이었고 3,018건(66.7%)이 CPT4 코드로 매핑되었으며, 해당 코드는 총 처방빈도의 88.4%를 차지하였다. 진단검사의 경우 원천 진단검사 오더 코드의 총수는 1,010개였으며, 검체 채취 부위는 195곳이었다. 따라서 이론적으로 196,950개의 진단검사 조합이 가능하였다. 실제로 오더가 발생한 오더-채취부위 조합수는 1,439건이었다. 모든 조합에 대하여 LOINC로 매핑하였고, 이 중 1,021건(70.95%)은 일치하는 코드, 나머지 418건(29.05%)은 상위개념의 코드로 매핑되었다.

CDW로부터 매핑을 반영하여 여의도성모병원 등록환자 2,279,292명 중 2009년부터 2020년 12월 31일까지 여의도성모병원을 방문한 731,846명의 데이터를 OMOP-CDM ver. 5.3.1 테이블로 추출, 적재

하였다. 데이터와 환자 수의 구성은 표 3과 같다.

3. 안과 데이터의 적재 형태 분석 및 데이터 추출

안과 데이터를 기존의 CDW, CDM에 적재하기 위하여 전자의무기록시스템 내에 저장되어 있는 안과 검사정보의 패턴과 저장 형태를 분석한 결과 이미지(image), 연속이미지(serial image), 문자형(text), 이미지와 텍스트의 조합(image+text), 이미지 조합(image set), 좌표변환 가능 이미지(coordination)로 구분할 수 있었다(표 4).

표 3. 여의도성모병원 OMOP-CDM 데이터 현황

(2021년 7월 1일 현재)

OMOP-CDM 테이블	구분데이터 건수 (행수)	환자 수 (명)
Observation_period	873,463	873,463
Person	878,141	878,141
Specimen	685,451	177,448
Condition_occurrence	23,946,055	738,475
Visit_occurrence	10,432,288	877,204
Device_exposure	9,552,905	384,433
Visit_detail	10,574,504	877,204
Condition_era	11,039,962	738,475
Note	1,212,322	223,904
Measurement	139,939,726	540,870
Observation	5,155,624	162,981
Death	6,655	6,655
Drug_exposure	59,499,428	588,768
Drug_era	12,542,373	570,752
Procedure_occurrence	116,463,727	799,553
Location	250	-
Care_site	1	-
Provider	17,266	-

OMOP-CDM, Observational Medical Outcomes Partnership-Clinical Data Warehouse.

표 4. 안과 검사 정보의 패턴과 저장 형태(일부)

분류	예시
Image
Serial image
Text
Image+text
Image set
Coordination

표 5. 안과 주요 검사 환자 수(중복 제외)

검사명	검사 인원(명)	기록형태
기본안저촬영	180,673	이미지
시신경유두입체검사	2,397	이미지
전안부촬영	61,922	이미지
시신경섬유층 사진	42,802	이미지
광각 안저촬영	67,673	이미지
자가형광안저촬영	8,758	이미지+텍스트
광간섭단층 혈관영상	10,625	이미지+텍스트
패턴 망막전위도검사	-	이미지+텍스트
정밀시야검사	700	이미지+텍스트
자동시야검사	50,542	이미지+텍스트
Swap 자동시야검사	8,064	이미지+텍스트
Threshold 자동시야검사	3,512	이미지+텍스트
굴절및조절검사	244,177	이미지, 텍스트
안압 측정(기타)	325,158	텍스트
안근기능검사 및 폭주검사 (랑카스터검사)	750	이미지
각막곡률반경 측정	45,063	이미지
각막내피세포검사	59,744	이미지
전산화 각막형태검사 (두께 포함; corneal topography; 편안)	25,635	이미지
ICG 맥락막 촬영	8,135	이미지
안구광학단층촬영(OCT) -망막(황반부)	34,739	이미지+텍스트
안구광학단층촬영(OCT) -녹내장(시신경)	29,555	이미지+텍스트
전산화 각막형태검사 (wave front scan)	2,145	이미지+텍스트
망막혈류계측	-	
시신경유두 측정	15,066	이미지
전안부정밀촬영	17,431	이미지
눈의계측검사, 레이저간섭계 이용(IOL)	35,106	텍스트
안구광학단층촬영(OCT) -전안부(전방영상)	1,330	이미지
간섭에 의한 눈물 지질층 두께 측정	-	이미지+텍스트
안구광학단층촬영	83,229	이미지+텍스트
대비감도검사	252	이미지+텍스트

ICG, indocyanine green; OCT, optical coherence tomography; IOL, intraocular lens.

2009년부터 2020년까지 주요 검사 시행 환자 수와 기록 형태는 표 5와 같으며, 시력의 손상 연구에 우선순위를 주어, 시력, 안압, 자동시야검사, 빛간섭단층촬영검사를 변환하여 추출하기로 하였다. 시력과 안압을 제외한 안과 기록은 이미지 파일 형태로 전자 의무기록시스템에 저장되어 있었다. 시력은 전자의무기록시스템 내부에 서식지로 기록되어 별도의 데이터베이스에 저장되고 있었다. 따라서 데이터 매핑 후 CDW에 적재하는 작업을 수행하였다. 안압은 전자의무기록시스템 내부에 자유 양식의 의무기록으로 기록되어 있었다. 2020년 12월 한 달간 서울성모병원에서 기록된 안과 초진 및 경과 기록지는 12,965건이었으며, 분석결과 26가지 패턴으로 안압이 기록되어 있었다. 자동시야검사 결과는 텍스트와 그림으로 구성되었으며, 기기의 종류 및 세대에 따라 상이한 양식의 결과지로 저장되어 있었다. 이 중 Humphery Visual Field 결과지 내에서 25가지의 변수를 정의하고, 각 변수의 위치를 표시한 뒤 Faster R-CNN을 이용하여 각 변수의 위치를 학습시켰다. 총 301,487건의 검사지에서 인식된 위치의 화상 정보를 추출하여 광학 문자인식을 통해 데이터를 추출하고, 추출된 데이터를 식명화하여 CDW에 적재하였다. 빛간섭단층촬영의 경우 총 34종의 결과 보고서로 분류할 수 있었다. 처방량이 가장 많은 두 종류의 검사용지에 대하여 상기 방법으로 데이터를 추출, 식명화하여 총 312,609건의 검사결과를 CDW에 적재하였다(그림 2).

4. 안과 용어의 매핑

안질환은 우안, 좌안, 양안에 발생할 수 있다. KCD에는 편측성을 나타내는 코드가 있다. KCD7에는 919개의 안과 질환 진단명이 있으며, 중복을 제외한 코드가 581개 수록되어 있었다. 그 중 188개(47종)의 병명의 코드만 오른쪽, 왼쪽, 양쪽, 상세부위

불명 등의 위치정보를 포함하고 있는데, 해당 코드를 검토한 결과 실제로는 백내장과 녹내장 계열의 진단만이 편측성 여부를 표기하고 있었고, 중복 코드를 제거하면 128개(33종)의 정보가 편측성 정보를 포함하고 있었다(표 6). 해당 진단명은 백내장 또는 녹내장의 하위 개념이었다. 2020년 서울성모병원에서 백내장 진단이 새롭게 내려진 건수는 20,751건이었고, 그 중 1,683건(8.10%)에서 편측성 정보가 입력되었다.

녹내장의 경우는 편측성 정보가 입력된 사례가 없었다. 건강보험요양급여비용에 등재된 안과 수술(자511-자554)은 모두 어느 눈에 수술을 수행하는지 표현할 수 없다. 시기능검사(나666-너798)의 경우 편측검사의 비용을 기록해 두기는 하지만, 해당 검사가 어느 쪽 눈에 시행하였는지에 대한 정보를 기록하도록 하고 있지 않았다. CDW에서는 수술실에서 이루어지는 행위인 경우, 그 행위뿐 아니라 수술 전 진단

그림 2. (A) Faster R-CNN을 통한 자동시야검사 결과 내 관심 영역 추출. (B) 자동시야검사 변수. (C) Faster R-CNN을 통한 빛간섭단층촬영 결과 내 관심 영역 추출. (D) 빛간섭단층촬영 변수.

표 6. 편측성 정보를 가진 안과 진단(오른쪽만 발체함)

상병기호	한글 명칭	영문 명칭
H2500	노년성 초기 백내장, 오른쪽	Senile incipient cataract, right
H2500	관상 노년성 백내장, 오른쪽	Coronary senile cataract, right
H2500	피질 노년성 백내장, 오른쪽	Cortical senile cataract, right
H2500	점상 노년성 백내장, 오른쪽	Punctate senile cataract, right
H2500	수정체낭하 극성 노년성 백내장(전, 후), 오른쪽	Subcapsular polar senile cataract (anterior, posterior), right
H2500	수극, 오른쪽	Water clefts, right
H2510	노년성 핵백내장, 오른쪽	Senile nuclear cataract, right
H2510	갈색백내장, 오른쪽	Cataracta brunescens, right
H2510	핵경화백내장, 오른쪽	Nuclear sclerosis cataract, right
H2520	모르가니형 노년성 백내장, 오른쪽	Senile cataract, morgagnian type, right
H2520	노년성 과숙백내장, 오른쪽	Senile hypermature cataract, right
H2580	기타 노년성 백내장, 오른쪽	Other senile cataract, right
H2580	노년성 백내장의 복합형, 오른쪽	Combined forms of senile cataract, right
H2590	상세불명의 노년성 백내장, 오른쪽	Senile cataract, unspecified, right
H2600	영아, 연소 및 초로 백내장, 오른쪽	Infantile, juvenile and presenile cataract, right
H2610	외상성 백내장, 오른쪽	Traumatic cataract, right
H26200	안구내 수술에 의한 이차성 백내장, 오른쪽	Cataract secondary to intraocular surgery, right
H26280	기타 합병백내장, 오른쪽	Other complicated cataract, right
H26280	눈 장애에 의한 이차성 백내장, 오른쪽	Cataract secondary to ocular disorders, right
H26280	녹내장성 수정체 혼탁(낭하), 오른쪽	Glaucomatous flecks (subcapsular), right
H26280	만성 홍채섬모체염에서의 백내장, 오른쪽	Cataract in chronic iridocyclitis, right
H2630	약물-유발 백내장, 오른쪽	Drug-induced cataract, right
H2640	후발 백내장, 오른쪽	After-cataract, right
H2640	이차성 백내장, 오른쪽	Secondary cataract, right
H2640	쉴메링고리, 오른쪽	Soemmerring's ring, right
H2680	기타 명시된 백내장, 오른쪽	Other specified cataract, right
H2690	상세불명의 백내장, 오른쪽	Cataract, unspecified, right
H4000	녹내장 의심, 오른쪽	Glaucoma suspect, right
H4000	고안압, 오른쪽	Ocular hypertension, right
H40100	저안압 녹내장, 오른쪽	Low-tension glaucoma, right
H40110	색소성 녹내장, 오른쪽	Pigmentary glaucoma, right
H40120	수정체의 거짓낙설을 동반한 수정체낭성 녹내장, 오른쪽	Capsular glaucoma with pseudoexfoliation of lens, right
H40130	개방우각녹내장 잔류기, 오른쪽	Residual stage of open-angle glaucoma, right

명, 수술 후 진단명 등을 기록하고 있어 수술이 이루어지는 환자에 한하여 편측성 정보를 추출할 수 있었다. OMOP-CDM ver. 5.3.1의 PROCEDURE_OCCURRENCE 테이블에는 해당 행위의 편측성 정보를 기록하는 변수로 modifier_concept_id가 있다. modifier_concept_id는 행위와 관련한 추가적인 정보이며, CPT4 Modifier나 HCPCS Modifier에 해당하는 표준용어를 사용하도록 규정되어 있다. CPT4 Modifier와 HCPCS Modifier의 right/left에 대응하는 concept_id를 modifier_concept_id에 기록함으로써 이를 편측성을 표현할 수 있으며, 연구진은 CPT4 Modifier를 채용하였다(표 7).

새롭게 추출한 시력, 안압, 자동시야검사, 빛간섭단층촬영의 검사결과를 CDM에 적재하기 위하여 LOINC 용어집을 검색하였다. 시력의 경우 국내에서 사용하는 시력 단위와 일치하는 용어는 발견할 수 없었다. 안압의 경우 좌안과 우안의 안압에 해당하는 용어가 각 한 개씩 존재하였다. 각 용어는 안압 측정방법을 정의하지 않았다. 자동시야검사의 변수와 일치하는 용어는 발견할 수 없었으며, 빛간섭단층촬영의 경우 11개의 변수에 해당하는 용어가 있었으나 좌, 우측을 구분하지 않고 있었다(표 8).

LOINC는 검체, 성격, 검체채취 간격, 채취 부위, 단위, 검사방법 등 6가지 항목을 맞추어 매핑하도록

규정하고 있다. 대부분의 안과 검사는 이에 해당하는 항목이 존재하지 않아 임시로 20억 번째 정수를 가진 임의의 concept_id를 비표준 개념(non-standard concept)으로 생성하여 임시 아이디를 부여하였다. 새로 생성한 concept_id는 안압 12개, OCT 117개, 자동시야 44개였다.

고 찰

본 논문은 의료데이터의 구축 경험과 임상 전공과 중의 하나인 안과 연구를 위하여 데이터를 확장, 보강하려는 노력을 기술함으로써, 의료데이터 활용에 대한 기대와 현황 사이의 간극을 고찰하고 그 차이를 메우는 방법을 제시하고자 하였다. 의료데이터를 활용하면 다양한 연구를 큰 어려움 없이 시행할 수 있을 것이라는 기대감이 크다[17]. 규제의 관점에서도 실사용근거(real world evidence, RWE)를 데이터를 통해 추출하고 활용할 수 있을 것으로 기대하고 있다. 2016년 12월 발효된 21세기 치료법(21st Century Cures Acts)은 미국의 의료혁신 역량을 강화하며 국민이 신속하고 의료혁신 결과의 혜택을 받을 수 있도록 하는 다양한 내용을 담고 있다[18]. 주목할 점은 미국 FDA의 인허가 절차에 관한 내용을 포함하고 있다는 것이다. FDA는 규제의 관점에서 RWE를 이용

표 7. 좌측과 우측을 나타내는 modifier_concept_id

Concept_id	VOCAB	Concept_code	Name	Domain	Class
2720905	HCPCS	RT	Right side (used to identify procedures performed on the right side of the body)	Observation	HCPCS Modifier
45888271	CPT4	RT	Right side (used to identify procedures performed on the right side of the body)	Observation	CPT4 Modifier
2720489	HCPCS	LT	Left side (used to identify procedures performed on the left side of the body)	Observation	HCPCS Modifier
45888270	CPT4	LT	Left side (used to identify procedures performed on the left side of the body)	Observation	CPT4 Modifier

표 8. LOINC에 수록된 시력, 안압, 시야검사, 빛간섭단층촬영 용어

LOINC_number	Component	Property	Time_aspect	System	Scale_type	Method_type
79876-9	Visual acuity^best corrected	LogLenRto	Pt	Eye.right	Qn	ETDRS eye chart
79877-7	Visual acuity^best corrected	LogLenRto	Pt	Eye.left	Qn	ETDRS eye chart
79878-5	Visual acuity^uncorrected	LogLenRto	Pt	Eye.right	Qn	ETDRS eye chart
79879-3	Visual acuity^uncorrected	LogLenRto	Pt	Eye.left	Qn	ETDRS eye chart
79880-1	Visual acuity^best corrected	LenRto	Pt	Eye.right	Qn	Snellen eye chart
79881-9	Visual acuity^best corrected	LenRto	Pt	Eye.left	Qn	Snellen eye chart
79882-7	Visual acuity^uncorrected	LenRto	Pt	Eye.right	Qn	Snellen eye chart
79883-5	Visual acuity^uncorrected	LenRto	Pt	Eye.left	Qn	Snellen eye chart
79892-6	Intraocular pressure	Pres	Pt	Eye.right	Qn	Tonometry
79893-4	Intraocular pressure	Pres	Pt	Eye.left	Qn	Tonometry
57110-9	Macular grid.inner superior subfield thickness	Len	Pt	Eye	Qn	OCT
57114-1	Macular grid.outer superior subfield thickness	Len	Pt	Eye	Qn	OCT
57108-3	Macular grid.center point thickness	Len	Pt	Eye	Qn	OCT
57109-1	Macular grid.center subfield thickness	Len	Pt	Eye	Qn	OCT
57110-9	Macular grid.inner superior subfield thickness	Len	Pt	Eye	Qn	OCT
57111-7	Macular grid.inner nasal subfield thickness	Len	Pt	Eye	Qn	OCT
57112-5	Macular grid.inner inferior subfield thickness	Len	Pt	Eye	Qn	OCT
57113-3	Macular grid.inner temporal subfield thickness	Len	Pt	Eye	Qn	OCT
57114-1	Macular grid.outer superior subfield thickness	Len	Pt	Eye	Qn	OCT
57115-8	Macular grid.outer nasal subfield thickness	Len	Pt	Eye	Qn	OCT

OCT, optical coherence tomography.

하여 허가된 약제의 적응증 추가, 의약품의 안전성 자료의 수집에 관한 제도와 절차를 만들기 위한 노력을 지속적으로 수행하겠다고 천명하였다[19].

임상현장에서도 데이터 자원을 활용한 연구에 대한 요구가 증가하였다. 이에 의료기관은 연구활동을

지원하면서도 개인정보의 유출을 막고, 의료데이터의 관리를 위한 제도를 마련하여 그 필요를 충족시키고자 하였다. 가톨릭중앙의료원에서는 연구지원을 위하여 독자적인 CDW인 CMC nU CDW를 구축하였다. CMC nU CDW는 개인정보 익명화와, 접근 보안을

구현하였다. 소정의 활용교육과 보안교육을 이수한 연구자는, 웹 플랫폼을 통하여 인구학적 조건, 수진 정보, 진단정보, 투약정보, 진단검사정보, 수술처방, 처방정보, 병리검사, 영상검사, 내시경검사, 기능검사, 수술기록, 인체유래물은행 등록, NGS (next-generation sequencing) 정보 등의 조건을 검색하여 코호트를 정의할 수 있다. 연구자는 정의된 코호트에 속한 환자의 분포를 확인한 후, 연구를 진행하기로 결정하면, IRB의 연구계획서 승인을 득하고, 정보활용 협의체의 데이터활용 승인을 받아 익명화된 데이터를 수령할 수 있게 된다. CMC nU CDW는 서울성모병원, 여의도성모병원, 은평성모병원, 의정부성모병원, 부천성모병원, 성빈센트병원, 인천성모병원 등 총 7개 기관의 데이터를 아우르고 있어, 가톨릭중앙의료원 내 다기관 연구가 가능하도록 구성되어 있다. 2019년 11월부터 2021년 1월까지 총 104건의 활용승인이 이루어져 매주 평균 약 2건의 연구가 새롭게 이루어지고 있다. 비록 CDW가 연구자의 데이터 활용을 촉진하고 있으나, CDW 활용 연구에는 몇 가지 한계가 있다. 첫째, 코호트를 이루는 환자가 대학병원에 방문한 환자라는 점이다. 데이터베이스 내 특정 환자에서 처음 나타나는 진단명이 해당 환자에 있어 처음으로 진단받은 것인지, 다른 의료기관에서 해당 진단을 받고 전원한 환자인지 알 수 없다. 따라서 대학병원의 질환 분포가 대한민국 또는 의료기관이 설치된 지역의 질환 분포를 대변한다고 판단할 수 없다. 둘째, 다른 의료기관과의 공동연구가 어렵다. 셋째, 데이터 분석과정이 온전히 연구자의 역량과 분석기술에 의존한다는 점이다. 현재 CMD nU CDW는 분석 기능을 제공하지 않는다. 코호트와 데이터 추출필터가 정의되면, 그에 맞추어 각 정보 테이블별로 Excel file 또는 CSV 형태의 파일을 생성하여 연구자에게 전달한다. 이후 분석은 모두 연구자가 처리하게 된

다. 따라서 분석과정 중에 연구자에게서 발생하는 오류를 배제할 수 없게 된다. 이 같은 한계를 극복하기 위하여 CDM의 도입을 고려할 수 있다. 예를 들어, CDM으로 구축된 공공데이터를 사용하면 특정 환자의 진단이 최초로 발생한 것인지 판단할 수 있다. 나아가, 다기관에서 동일한 방법을 통해 얻은 연구결과를 비교함으로써 특정 기관의 데이터가 편향되었는지 판단할 수도 있으며, 표준화된 분석방법론을 사용하여 분석과정에서 발생하는 오류를 최소화할 수 있다.

CDM 구축의 과정은 순탄하지만은 않다. 우선 의료데이터의 CDM 적재에 대한 행정적 결정과 이를 추진하기 위한 인적자원을 배정하는 일이 선행되어야 한다. 이후 원천데이터를 이해하고, CDM 데이터 구조에 대한 파악이 필요하다. OMOP-CDM의 경우 데이터 구조에 관한 설명이 공개되어 있어 누구나 접근할 수 있다. 그러나 전자의무기록시스템이나 원천데이터에 대한 이해는 평소 해당 시스템을 다루지 않는 사람에게는 매우 어려운 일이다. 따라서 첫째, 데이터 전문가와 CDM 전문가가 함께 ETL 설계할 것, 둘째, 의학 지식이 있는 사람이 용어 매핑을 할 것, 셋째, 엔지니어가 ETL을 수행할 것, 마지막으로, 모든 사람이 질 관리에 참여할 것을 ELT 진행의 모범사례로 제시하고 있다[20]. CMC nU CDW를 데이터 원천으로 하여 OMOP-CDM을 구축하는 과정에는 원천데이터를 잘 이해하고 있는 개발팀, 저자를 포함한 의학지식이 있는 연구자들이 참여하는 등 모범사례를 따르기 위해 노력했다.

데이터를 적재하기 전에, 구축하려는 CDM이 요청하는 사양의 표준용어로의 매핑이 필수적이다. OMOP-CDM에서는 표준 개념(standard concept)에 매핑할 것을 요청하고 있다. 대개 기관이 보유하고 있는 많은 용어를 단시간에 매핑하기는 쉽지 않다. 일례로 본 연구에서 원천데이터 소스에 등재된 진단 용어

는 총 26만여 항목이었다. 본 연구에 사용된 CDM의 경우, (1) 원내코드-원천코드 관계에서 원천코드가 표준코드에 매핑되어 있는 경우 이를 원내코드에 기계적으로 우선 적용, (2) 원내코드 중 명칭이 동일하면서 코드만 다른 경우, 매핑 일괄 적용, (3) 의무기록사, 간호사, 약사 등 다양한 직종의 수개월 간 검토, (4) 전문의의 검증 등의 과정을 거쳤다. 그럼에도 불구하고 완벽한 매핑은 불가능함을 인정하여, 실제 분석 전에 해당 연구에 필요한 진단명의 매핑 적절성을 검토하고 연구를 수행하는 등의 조치를 하고 있다. 아울러, 최근 기관 표준화 팀을 신설하여 매핑 전면 검토를 수행하는 등 지속적인 매핑 업데이트를 시행하고 있다. 매핑 작업 시 모든 용어가 그 개념에 일치하는 standard concept을 가지고 있지는 않았다. 진단명을 매핑하는 경우, 가장 가까운 의미 진단명에 매핑하거나 상위 개념의 용어에 매핑하는 경우가 발생한다. 이때 임상적으로 중요한 일부 정보가 사라지기도 한다. 예를 들어, KCD7 코드 H2520 “모르가니형 노년성 백내장, 오른쪽”의 경우 정확히 일치하는 표준 용어는 존재하지 않는다. 이 경우 “모르가니형 백내장” 또는 “노년성 백내장, 오른쪽” 상위 개념으로 매핑하기도 한다. 그러나 이러한 방식은 편측성 정보가 사라지거나 백내장 형태 정보가 사라지게 되는 결과를 가져온다. Observational Health Data Sciences and Informatics (OHDSI)에서는 이러한 경우 1:n 방식으로 매핑하여 하나의 원천코드를 두 개 이상의 코드로 매핑하도록 하여 임상정보를 유지하도록 하고 있다. 다만, 매핑 실무를 담당하는 사람의 의학지식의 정도와 판단에 따라 기관별로 결과가 달라질 수 있다는 점은 염두에 두어야 한다. 의약품은 주로 RxNorm과 RxNorm Extension으로 매핑하였다. 성분명, 용량, 제형 등 다양한 정보수준이 있는데, 어떤 항목을 매핑하느냐에 따라 매핑 정보가 달라질 수 있다. 때로

는 모든 조건을 만족하는 표준코드를 찾을 수 없는 경우가 발생하고 이때도 매핑 실무자의 의견에 따라 매핑할 수밖에 없게 된다. 진단검사 코드는 의무기록시스템에서 처방을 내릴 때 사용하는 오더 코드와 결과를 조회할 때 사용하는 결과 코드로 존재한다. 연구에서 사용하고자 하는 코드는 주로 결과 코드인데, 결과 코드는 검체, 성격, 검체채취 간격, 채취부위, 단위, 검사방법 등 6가지 항목 모두 일치하여야 정확한 분석이 가능하다. 따라서 이를 매핑하는 작업에 상당한 노력이 들어가게 된다. 이러한 노력을 최소화하면서 매핑 실무자의 차이에서 발생하는 기관 간의 매핑 차이를 극복하기 위해, 이미 공개된 매핑 정보를 활용할 필요가 있다. ATHENA에는 KCD7과 electronic data interchange (EDI) 코드가 공개되어 있다[15]. 비록 현재 KCD7과 EDI 코드는 비표준 코드(non-standard code)로 분류되어 있기는 하지만 향후 추가적인 매핑을 통해 standard code로 편입되면, 기존 용어와의 위계가 연결될 것이다. 이와 별도로 한국보건 의료정보원에서도 KCD, EDI, 보건의료용어표준을 SNOMED-CT로 매핑한 자료를 공개하였다[21]. 공개된 자료 덕에 진단, 의약품, 진단검사의 매핑에 들어가는 수고는 줄어들 것이고, 기관 간의 매핑의 차이도 점차 좁아질 것이다.

CDM의 MEASUREMENT 테이블에 적재된 정보는 주로 진단검사항목이다. 각 임상 과정에서 이루어지고 있는 기능검사는 그 처방 여부를 알 수 있을지언정, 그 결과는 CDW, CDM에 적재되어 있지 않은 경우가 많다. 가장 큰 요인은 CDW, CDM을 구축할 때 의료정보시스템의 데이터베이스에서 정규화된 데이터를 추출하여 변환, 적재하기 때문이다. 따라서 원천 데이터가 정제되어 있지 않은 때에는 적재가 어렵다. 이 경우 연구의 필요에 따라 새롭게 데이터를 마련하여 적재하는 방법을 사용해야 한다. 안과 연구를 수행

하기 위해 필요한 정보를 파악해본 결과, 시력을 제외한 모든 정보는 정제되지 않은 상태로 존재하고 있었다. 기록지의 내용에 섞여 있는 정보를 캐내기 위해 정규표현식을 사용하기는 하였으나, 다른 기관의 안과 의료진이 같은 패턴으로 기록을 하고 있을 가능성은 적다. 따라서 향후 기계학습을 통한 자연어처리를 이용한 추출 연구가 필요하다. 시야와 OCT는 딥러닝 기법을 이용하여 관심 영역을 찾아내고 광학문자인식을 하여 데이터를 추출하는 방법을 사용하였다. Faster R-CNN을 이용하여 관심 영역을 찾아내는 것은 1,000여 장의 학습데이터를 사용하여 비교적 쉽게 수행할 수 있었으나, 광학문자인식작업에서는 정확도를 높이기 위한 이미지 전처리 작업을 수행하여야 하였다. 문자가 위치한 곳의 배경 이미지나 색이 서로 다른 경우 문자 인식률이 떨어지는 현상이 나타나서 수차례의 코딩 수정을 해야 했다. 고성능 GPU (graphics processing unit) 서버를 사용하였음에도 불구하고 시야검사 30만 장을 처리하는 데 걸리는 시간은 약 30일 정도였고, 빛간섭단층촬영 30만 장의 처리에도 비슷한 시간이 걸렸다. 시야검사와 빛간섭단층촬영 이미지별로 문자인식을 달리해야 하듯, 새로운 데이터가 추가될 때마다 유사한 작업을 반복하여 수행할 필요가 있을 것으로 보인다. 이러한 과정을 거쳤더라도 본 연구의 대부분의 안과 검사는 이에 해당하는 항목이 존재하지 않아 임시로 20억 번대 정수를 가진 임의의 concept_id를 비표준 개념(non-standard concept)으로 생성하여 임시 아이디를 부여하였다. 이는 CDM의 안과 영역의 확대 구축을 이루고 시스템에 적재하기 위한 선택이었으며, 용어의 생성이 필요한 경우 OHDSI에 요청하여 새로운 concept_id를 생성 부여받는 절차가 있으므로 추후 이 절차를 활용하여 생성한 임시 아이디를 갱신할 예정이다.

CDM을 활용하여 진단명, 의약품, 진단검사 결과

를 가지고 정의할 수 있는 코호트와 결과가 있다면 비교적 수월하게 특정 치료의 치료 경로나 특정 현상의 발생률을 관찰할 수 있다[22]. 다양한 코호트를 생성한다면 코호트 간의 결과를 비교해볼 수도 있다. 적재된 데이터에서 벗어나는 연구를 수행하기에는 매우 까다로운 작업을 거쳐야 한다. 이를 극복하기 위한 노력은 끊임없이 이루어지고 있다. Shin 등[23]은 유전체 데이터를 OMOP-CDM에 확장한 G-CDM (Genomic CDM)을 선보이고 그 효용을 증명하였다. Michael 등[24]은 백혈병 연구를 위하여 검체 정보와 임상정보를 연계하도록 OMOP-CDM을 확장하는 시도를 하였다. OMOP-CDM에서 부족한 영상검사의 처리를 위한 시도와 종양학 정보를 탑재하기 위한 시도도 ODSHI 커뮤니티 내에서 이루어지고 있다[25].

매핑과 ETL, 자료의 추가에 대해서 생각하면 CDM을 활용한 안과 영역에서의 공동연구의 참여가 요원한 것으로 보일 수 있다. 그러나 완벽하게 구축된 시스템이라는 이상을 버리고 한 번에 하나씩 접근한다면, 적어도 연구를 시작할 수는 있다. 예를 들어, 특정 의약품의 이상반응에 대한 공동연구를 기획하는 경우, 해당 의약품의 코드, 관찰하여야 할 진단검사의 코드와 진단명의 코드가 전 기관에 걸쳐 동일하게 매핑되어 있는지 사전에 조율하고, 관심 의약품이 실제로 처방된 건수가 어느 정도인지, 실제 관심 있는 검사의 결과가 적재되어 있는지를 미리 파악하면 큰 도움이 된다.

연구를 수행할 때마다 데이터를 정비하고 필요한 경우 ETL 과정을 다시 거치며 지속해서 질 관리하기 쉽지 않다. 초기 구축뿐 아니라 데이터의 추가 적재, 변경된 용어 표준의 반영, 발견된 오류의 해결 등의 작업이 반복적으로 발생하기 때문이다. 그러나 이 같은 일은 품질 높은 RWE를 생산하기 위해 피할 수 없는 과정이다. 최근 FeederNet (Federated E-health

Big Data for Evidence Renovation Network), MOA (Medical record Observation and Assessment for drug safety)-net과 같이 CDM을 기반으로 국내 다기관이 공동으로 참여하여 분산연구를 수행하는 플랫폼들이 등장하고, 플랫폼 내에서 CDM의 구축과 관리, 데이터 추가 등의 작업들이 자동으로 이루어지게 하는 등의 노력이 이어지고 있다. 이를 통하여 전체 연구 네트워크의 데이터 질이 지속적으로 향상될 것으로 기대된다. 이와 더불어 본 연구처럼 각 임상과별로 자료의 종류를 하나하나 추가하는 데 대한 연구와 지원도 지속적으로 이루어져야 할 것이다. 종합적으로 잘 모인 데이터는 결국 환자를 위하여 활용될 것이기 때문이다.

감사의 글

본 연구는 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술연구개발사업 지원으로 이루어졌다(과제고유번호: HI19C0373).

참고문헌

1. Rainie H, Wellman B. Networked: the new social operating system. Cambridge (MA): MIT Press; 2012.
2. Blacketer C. The common data model. In: Observational Health Data Sciences and Informatics, editor. The book of OHDSI: observational health data sciences and informatics [Internet]. [place unknown]: OHDSI; 2021 [cited 2021 Jul 8]. [about 36 screens]. Available from: <https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html>.
3. Toh S, Reichman ME, Graham DJ, Hampp C, Zhang R, Butler MG, et al. Prospective postmarketing surveillance of acute myocardial infarction in new users of saxagliptin: a population-based study. *Diabetes Care*. 2018;41(1):39–48. DOI: <https://doi.org/10.2337/dc17-0476>.
4. Trifiro G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *J Intern Med*. 2014;275(6):551–61. DOI: <https://doi.org/10.1111/joim.12159>.
5. Yih WK, Kulldorff M, Fireman BH, Shui IM, Lewis EM, Klein NP, et al. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink project. *Pediatrics*. 2011;127 Suppl 1:S54–64. DOI: <https://doi.org/10.1542/peds.2010-1722I>.
6. Schneeweiss S, Brown JS, Bate A, Trifiro G, Bartels DB. Choosing among common data models for real-world data analyses fit for making decisions about the effectiveness of medical products. *Clin Pharmacol Ther*. 2020;107(4):827–33. DOI: <https://doi.org/10.1002/cpt.1577>.
7. Toh S, Rasmussen-Torvik LJ, Harmata EE, Pardee R, Saizan R, Malanga E, et al. The National Patient-Centered Clinical Research Network (PCORnet) Bariatric Study Cohort: rationale, methods, and baseline characteristics. *JMIR Res Protoc*. 2017;6(12):e222. DOI: <https://doi.org/10.2196/resprot.8323>.
8. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. *J Am Med Inform Assoc*. 2016;23(5):909–15. DOI: <https://doi.org/10.1093/jamia/ocv188>.
9. Raebel MA, Haynes K, Woodworth TS, Saylor G, Cavagnaro E, Coughlin KO, et al. Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. *Pharmacoepidemiol Drug Saf*. 2014;23(6):609–18. DOI: <https://doi.org/10.1002/pds.3580>.
10. Curtis LH, Weiner MG, Boudreau DM, Cooper WO,

- Daniel GW, Nair VP, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf*. 2012;21 Suppl 1:23–31. DOI: <https://doi.org/10.1002/pds.2336>.
11. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010; 17(2):124–30. DOI: <https://doi.org/10.1136/jamia.2009.000893>.
12. Sentinel, Methods, data, & tools [Internet]. Silver Spring (MD): Sentinel; [date unknown] [cited 2021 Jul 8]. Available from: <https://www.sentinelinitiative.org/methods-data-tools>.
13. European Medicines Agency. A common data model for Europe?: why? which? how? Amsterdam: European Medicines Agency; 2018.
14. Observational Health Data Sciences and Informatics Team. OHDSI/Usagi [Internet]. San Francisco (CA): GitHub Inc.; c2021 [cited 2021 Jul 8]. Available from: <https://github.com/OHDSI/Usagi>.
15. ATHENA [Internet]. Cambridge (MA): Odysseus Data Services Inc.; c2021 [cited 2021 Jul 8]. Available from: <https://athena.ohdsi.org/search-terms/start>.
16. 약물역학빅데이터분석팀. 공통데이터모델 구축을 위한 약물 용어 매핑 가이드[Internet]. 안양: 한국의약품안전관리원; 2020 [cited 2021 Jul 8]. Available from: <https://moa.drugsafe.or.kr/pat/fileRoom/view/50>.
17. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff (Millwood)*. 2014;33(7):1163–70. DOI: <https://doi.org/10.1377/hlthaff.2014.0053>.
18. Hudson KL, Collins FS. The 21st Century Cures Act: a view from the NIH. *N Engl J Med*. 2017;376(2): 111–3. DOI: <https://doi.org/10.1056/NEJMp1615745>.
19. Xia AD, Schaefer CP, Szende A, Jahn E, Hirst MJ. RWE framework: an interactive visual tool to support a real-world evidence study design. *Drugs Real World Outcomes*. 2019;6(4):193–203. DOI: <https://doi.org/10.1007/s40801-019-00167-6>.
20. Blacketer C, Voss E. Extract transform load. In: *Observational Health Data Sciences and Informatics*, editor. The book of OHDSI: observational health data sciences and informatics [Internet]. [place unknown]: OHDSI; 2021 [cited 2021 Jul 8]. [about 32 screens]. Available from: <https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html>.
21. 보건의료정보표준. 매핑 테이블 소개: 표준용어체계란? [Internet]. 서울: 보건의료정보표준; c2021 [cited 2021 Jul 8]. Available from: <https://www.hins.or.kr/menu.es?mid=a11301010000>.
22. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A*. 2016;113(27):7329–36. DOI: <https://doi.org/10.1073/pnas.1510502113>.
23. Shin SJ, You SC, Park YR, Roh J, Kim JH, Haam S, et al. Genomic common data model for seamless interoperation of biomedical data in clinical practice: retrospective study. *J Med Internet Res*. 2019; 21(3):e13249. DOI: <https://doi.org/10.2196/13249>.
24. Michael CL, Sholle ET, Wulff RT, Roboz GJ, Champion TR Jr. Mapping local biospecimen records to the OMOP Common Data Model. *AMIA Jt Summits Transl Sci Proc*. 2020;2020:422–9.
25. OMOP Common Data Model. Oncology extension [Internet]. [place unknown]: OMOP Common Data Model; c2021 [cited 2021 Jul 8]. Available from: <https://ohdsi.github.io/CommonDataModel/oncology.html>.