건강보험심사평가원
HEALTH INSURANCE REVIEW & ASSESSMENT SERVICE

vol.33 (2024-3)
ISSN 3022-8204 (online)

HIRA ISSUE

건강보험심사평가원 공통데이터모델 (Common Data Model) 구축과 개방

박소정 주임연구원
건강보험심사평가원 빅데이터실 빅데이터결합부

| 키워드 | 공통데이터모델, Common Data Model, CDM

1. 들어가며

건강보험심사평가원(심평원)은 공공데이터법[1]에 의거하여 심평원이 보유한 보건의료 데이터를 누구나 활용할 수 있도록 HIRA빅데이터개방포털 및 공공데이터포털을 통해 개방하고 있다. 또한 학술 · 공공 · 산업계를 대상으로 맞춤형 연구자료, 환자 표본자료, 결합자료 등을 원격 분석서버 또는 별도의 공간에서 분석하고 결과값을 반출할 수 있도록 지원하고 있다[1]. 이러한 보건의료 데이터 개방과 활용 지원은 점차 확대되고 있으나, 데이터 연계 및 활용에는 법적, 구조적 제약사항이 존재한다. 데이터를 보유하고 있는 각 기관의 데이터 구조와 용어가 상이하며, 개인정보보호법 등 법적 규제가 강화되고 있기 때문이다. 심평원은 위와 같은 제한점을 극복하고자 보유하고 있는 건강보험 청구자료(청구자료)를 공통데이터모델(Common Data Model, CDM)로 변환 · 구축하였다. 이 글에서는 심평원의 CDM (HIRA CDM) 구축 및 개방에 대하여 소개하고자 한다.

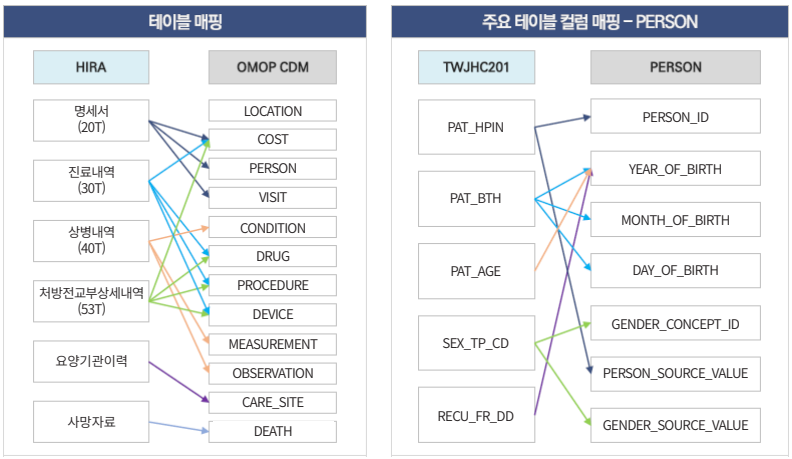
2. HIRA CDM 개요

가. 개념과 구조

CDM이란 데이터의 구조와 용어를 표준화한 데이터 모델로[2], CDM 데이터를 보유한 기관은 기관 간 데이터를 연계하여 분석할 수 있다. 또한, CDM을 통해 데이터 자체가 아닌 분석 결과만을 공유하는 분산연구망(distributed research network, DRN) 실현이 가능하므로[3], 개인정보유출의 위험으로부터 안전하다.

1) 공공데이터의 제공 및 이용 활성화에 관한 법률 제1조(목적), 제3조(기본원칙)

HIRA CDM은 다양한 CDM 중 임상 정보를 가장 광범위하게 포함하는 데이터 구조로 설계된 OMOP-CDM (Observational Medical Outcomes Partnership-Common Data Model) 형태[3]로 구축되었다. HIRA CDM DB는 OMOP-CDM 5.3.1 버전을 기준으로 청구자료를 변환하였으며, 임상 데이터 테이블(a), 보건의료 시스템 데이터 테이블(b), 보건의료 경제 데이터 테이블(c), 기타 의료정보 테이블(d), 의료용어 정보 테이블(e)로 구성된다[4].



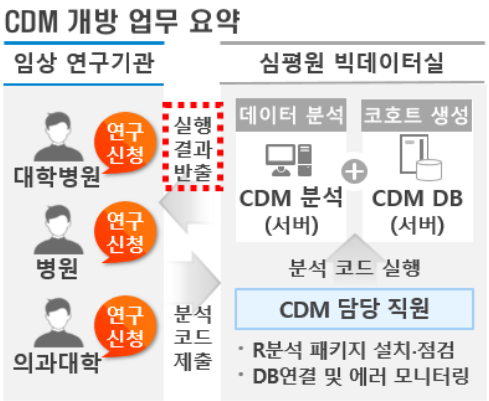
[그림 1] 심평원 청구자료 테이블과 CDM 테이블 매핑

출처: 심평원 내부자료

(a) PERSON, OBSERVATION_PERIOD, VISIT_OCCURRENCE 등; (b) CARE_SITE 등; (c) PAYER_PLAN_PERIOD, COST 등; (d) CONDITION_ERA, DRUG_ERA 등; (e) VOCABULARY 등

나. 장점

HIRA CDM 분석은 연구자가 심평원에 분석코드를 제출하면 담당 직원이 내부 폐쇄망에 위치한 익명화된 CDM에 접근하여 분석코드를 실행하는 방식으로 이루어진다. 이러한 과정에서 개인정보유출이 없는 것은 물론이며, 심평원 담당 직원이 직접 R분석 패키지를 설치 · 점검하고 오류 발생 시 연구진과 공유하며 분석을 지원하는 등 연구자 친화적인 분석 서비스를 제공하고 있다. 이는 연구가 사전 정의된 프로토콜에 의해 진행되며, 연구자가 연구 결과를 원하는 방향으로 도출하기 위하여 분석 방법을 임의로 변경하는 것이 불가능하여 객관적인 연구 결과 도출이 가능함을 의미한다.



[그림 2] HIRA CDM 분석 과정

출처: 심평원 내부자료

또한, HIRA CDM은 높은 품질 수준을 갖추고 있다고 자랑할 수 있다. 2024년 4월 심평원은 HIRA CDM의 품질을 CDM 공식 품질평가 도구[2]로, 완전성(Completeness), 순응성(Conformance), 타당성(Plausibility)을 평가하여 검증한 결과, 표준화율 97%, 완전성과 순응성은 각 99%, 타당성은 97%[3]라는 높은 품질 수준을 보여 신뢰성을 확보하였다.

HIRA_CDM_MFRN

DataQualityDashboard Version: 1.4.1
Results generated at 2024-03-04 07:48:53 in 16 days

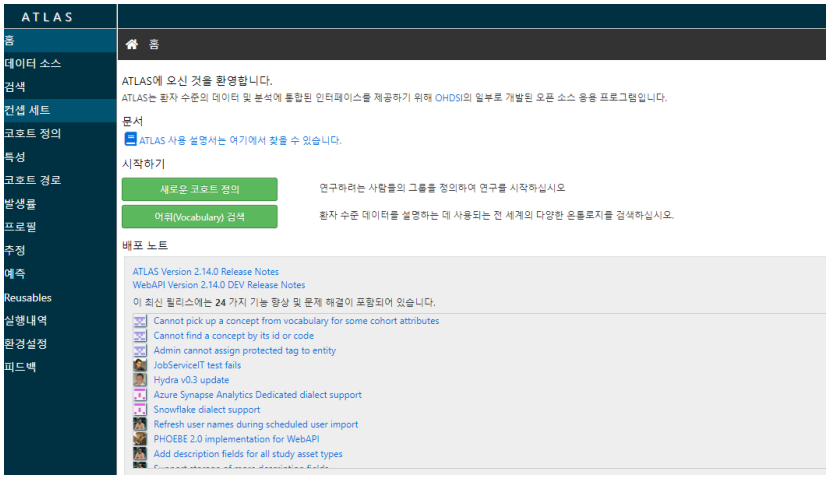
	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1899	75	1974	96%	287	0	287	100%	2186	75	2261	97%
Conformance	145	2	147	99%	44	0	44	100%	189	2	191	99%
Completeness	132	1	133	99%	10	1	11	91%	142	2	144	99%
Total	2176	78	2254	97%	341	1	342	100%	2517	79	2596	97%

[그림 3] 심평원 CDM DB의 품질평가 결과
출처: 심평원 내부자료

다. 이용방법

HIRA CDM 이용 신청은 HIRA빅데이터개방포털(https://opendata.hira.or.kr)에서 할 수 있다. 사전 공지를 통해 안내하며 접수된 과제는 심의를 통해 선정하여 분석을 지원한다.

연구자는 웹기반의 CDM 분석 도구인 'ATLAS'를 활용하여 데이터 추출 · 분석 코드를 작성할 수 있다. 'ATLAS'에서는 환자 코호트 정의, 연구 대상자 특성 분석, 분석코드 생성 등을 지원하며 누구나 무료로 사용할 수 있다[5].



[그림 4] ATLAS 접속화면
출처: OHISI ATLAS 웹페이지[5]

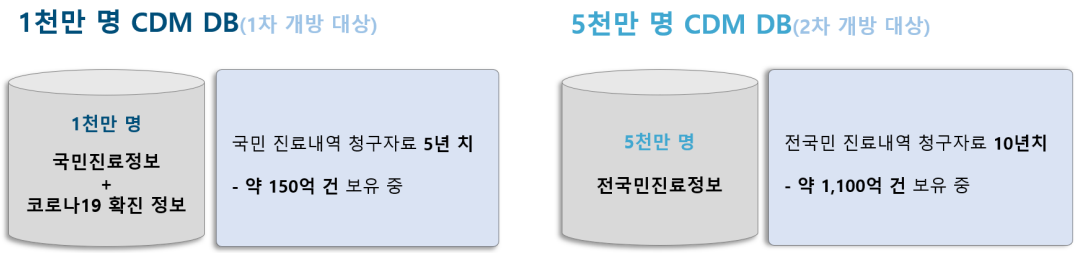
2) OHDSI (Observational Health Data Sciences and Information)의 OMOP-CDM 공식 품질평가 도구인 DQD (Data Quality Dashboard)를 활용함
3) 타당성 97% (청구자료에는 시간 정보가 없기 때문에 DATATIME(일시)을 검사한 부분에서 탈락(Fail) 75건 이 발생함. 실제 타당성은 99% 수준)

3. HIRA CDM 데이터베이스 구축 및 개방

가. 진행경과

심평원은 보건복지부 주관으로 진행된「기관 간 분석자료 공유 · 활용 네트워크 운영」사업에 2018년부터 참여하여 분석서버 도입, 표준용어 매핑사전 구현 등 청구자료를 CDM으로 변환 · 구축하기 위한 노력을 기울여왔다. 그러던 중 2020년 코로나19 감염 환자가 급증하자 국가 단위의 임상 근거를 마련하기 위해, 심평원은 보건복지부 중앙사고수습본부와 공동으로 국내 코로나19 의심 · 확진자의 의료이용 정보(확진여부, 기저질환 등)를 비식별화하여 CDM으로 구축 · 개방하였다[6].

이어 2022년에 심평원은 전국민의 20%에 해당하는 1천만 명[4]의 5년간(2018.1.1.~2022.4.30.) 청구자료를 CDM으로 구축하였으며, 해당 DB는 코로나19 확진 정보를 포함한 국민 진료내역 약 150억 건을 보유하고 있다[7]. 또한 2023년에는 CDM 개방 범위 확대 요구에 대응하여 전국민 5천 6백만 명의 10년간(2013.1.1.~2022.12.31.) 청구자료 약 1,100억 건을 CDM 데이터로 구축하였다[8].



[그림 5] 심평원에서 개방한 CDM DB 2종
출처: 심평원 내부자료

나. 활용성과

심평원은 2020년 세계 최초로 코로나19 CDM 데이터를 개방한 이후, 개방 범위를 확대하여 2022년과 2023년 두 차례 데이터 확대 개방을 실시하였다. 심평원은 2020년 3월부터 8월까지 코로나19 CDM 데이터를 국내외 연구자들에게 개방하여 국제협력 연구를 지원하였다. 총 32건의 연구과제가 수행되고[6], 일부는 The American Journal of Geriatric Psychiatry 등 주요 저널에 게재되어 코로나19와 관련된 다수의 근거를 생성하는 기반이 되었다.

2022년 9월, 심평원은 1천만 명 CDM을 개방하고 이용 신청을 받았다. 22개 기관에서 총 42건의 연구 과제를 신청하는 등 CDM에 대한 수요가 상당했다. 기관별로는 대학교에서 12건, 의료기관에서 30건을 신청하였으며, 주제별로는

4) 2021년 한 해 동안 국내 의료서비스를 이용한 전체 환자의 약 20%를 층화 표본 추출함

코로나19 관련 31건, 기타 질환 및 약물의 연관성 비교연구 등이 11건이었다. 2023년 9월에 5천만 명 CDM을 개방하자 해당 데이터 활용을 희망하는 17개 기관에서 총 40건(대학교 18건, 의료기관 22건)의 연구 과제를 접수하였다. 그 중 15개 과제를 선정하여 현재 분석 지원 중이다. HIRA CDM은 2020년 최초 개방이 이루어진 이래로 참여 연구과제 중 6편이 SCI(E)급 학술지에 발표되는 등 학술적 가치를 인정받았고[9]. 앞으로의 성과가 더욱 기대된다.

4. 나가며

디지털 뉴노멀 시대를 맞아 공공 분야의 빅데이터 개방, 민관 간 데이터 연계 및 활용에 대한 요구가 크게 증가하고 있다. 특히, 민감한 건강 정보를 포함한 보건의료 데이터를 활용함에 있어 개인정보의 유출 없이 기관 간 데이터를 연계 · 분석할 수 있는 CDM에 국내외의 관심이 집중되고 있다.

국내에서는 보건복지부와 산업통상자원부가 주도하여 CDM 플랫폼을 구축함으로써 보건의료 데이터의 공적 연구 활성화 기반을 다지고 있다. 산업통상자원부는 3년간 40억 원을 투자하여 민간 의료기관 60개소 이상이 보유하고 있는 의료데이터를 CDM으로 표준화하였으며, 보건복지부는 보건의료 빅데이터 플랫폼 사업을 통해 산하기관 5개 기관[5]의 데이터를 CDM으로 변환하고 활용 플랫폼을 구축하였다[10].

국외에서는 미국 외에도 유럽, 호주, 중국, 일본, 싱가포르 등에서 CDM 변환이 확장되고 있다. 특히 유럽에서는 EHDEN (European Health Data and Evidence Network) 프로젝트를 시행하여 유럽 29개국 187개 데이터 파트너로 구성된 네트워크를 구축하였으며, 8억 6천만 건 이상의 보건의료 데이터를 OMOP-CDM으로 변환하였다[11].

향후 국내외 산업계, 학계, 연구기관에서의 CDM 활용 수요는 더욱 늘어날 것으로 전망된다. 국내에서도 유럽과 같이 다국가간 컨소시엄을 구성하여 국내외 공동연구 과제 발굴 및 지원이 필요하다. 이는 과거 코로나19 국제협력 연구 사례에서 보았듯, 향후 유사 상황에 대비하여 신속한 데이터 기반 의사결정 지원과 국제 비교연구 활성화의 기반이 될 것이다.

5) 건강보험심사평가원, 국민건강보험공단, 국립암센터, 질병관리청, 국립중앙의료원

참고문헌

[1] HIRA빅데이터개방포털[Internet]. [cited 2024 July 1], Available from: https://opendata.hira.or.kr/op/opb/selectHelhMedDataInfoView.do#none

[2] Observational Health Data Sciences and Information 웹사이트[Internet]. [cited 2024 July 1], Available from: https://ohdsi.github.io/CommonDataModel/

[3] 박래웅. 공통 데이터 모델과 분산연구망: 오딧세이 컨소시엄(Observational Health Data Sciences and Informatics, OHDSI) 연구사업. 대한내과학회지. 2019;94(4):309-314. DOI: https://doi.org/10.3904/kjm.2019.94.4.309

[4] Observational Health Data Sciences and Information. OMOP Common Data Model Specifications. OHDSI. 2018.

[5] OHISI ATLAS 웹페이지[Internet]. [cited 2024 July 2], Available from: https://atlas-demo.ohdsi.org/#/home

[6] Rho YS, Cho DY, Son YJ, Lee YJ, Kim JW, Lee HJ, You SC, Park RW, Lee JY. COVID-19 International Collaborative Research by the Health Insurance Review and Assessment Service Using Its Nationwide Real-world Data: Database, Outcomes, and Implications. Journal of Preventive Medicine & Public Health. 2021;54:8-16. DOI: https://doi.org/10.3961/jpmph.20.616

[7] Kim CS, Yu DH, Baek HR, Cho JH, You SC, Park RW. Data Resource Profile: Health Insurance Review and Assessment Service Covid-19 Observational Medical Outcomes Partnership (HIRA Covid-19 OMOP) database in South Korea. International Journal of Epidemiology. 2024;53(3):1-6. DOI: https://doi.org/10.1093/ije/dyae062

[8] 벌크럼. 국민 진료정보의 공통데이터모델(CDM) 데이터 변환 용역사업 결과보고서. 건강보험심사평가원. 2023.

[9] 건강보험심사평가원 보도자료. 심사평가원, 국제 표준 공통데이터모델 확대 개방한다. 2023.7.26.

[10] 보건의료 빅데이터 통합 플랫폼 [cited 2024 July 3], Available from: https://hcdl.mohw.go.kr/static/cat/drnCatalogList

[11] 유럽연합 공공데이터 포털 [cited 2024 July 3], Available from: https://data.europa.eu/en/news-events/news/european-health-data-and-evidence-network-ehden-shaping-future-health-data-europe

HIRA ISSUE

발행일 2024. 7. 31.
발행처 건강보험심사평가원 심사평가정책연구소
발행인 함명일

강원특별자치도 원주시 혁신로 60(반곡동)
Tel. 033-739-0915, 0916 | www.hira.or.kr