대한내과학회지: 제 94 권 제 4 호 2019 https://doi.org/10.3904/kjm.2019.94.4.309

Perspectives

공통 데이터 모델과 분산연구망: 오딧세이 컨소시엄(Observational Health Data Sciences and Informatics, OHDSI) 연구사업

아주대학교 의과대학 의료정보학과

박래웅

The Distributed Research Network, Observational Health Data Sciences and Informatics, and the South Korean Research Network

Rae Woong Park

Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea

개 요

의료 데이터는 데이터 구조, 형식의 이질성, 데이터의 질과 양 등 기술적인 어려움과 기관의 허락, 개인 정보보호 문제 등 법적 문제 그리고 타인에게 제공하는 데이터가 자신에게 불리하게 사용될지 모른다는 두려움 등의 문제로 연구자 간 공유가 쉽지 않다. 현재까지 대부분의 기관 간 공동 연구는 극히 일부의 환자 데이터를 연구 주도 기관과 공유함으로써 진행하였는데, 한 번의 공동 연구를 위하여 막대한 노력과 시간, 자금이 들어가는 현실적인 문제와 개인 정보 공유를 제한하는 법적/제도적 문제들이 있다[1]. 최근 이런 제약을 극복하기 위하여 공통 데이터 모델(common data model, CDM)을 이용한 분산연구망(distributed research network)이 주목받고 있다.

의료 데이터 표준화의 필요성 및 분산연구망의 대두

분산연구망은 원본 데이터의 공유 없이 분산된 형태로 데이터를 관리하면서 기관 내에서 분석한 결과만 공유하는 방식이다. 즉, 각 병원의 환자정보를 표준화 및 가명화한 후 데이터를 병원 폐쇄망 안에 두고 사용자의 요청에 따라서 기관 안에서 R이나 파이썬 등 프로그램/분석코드를 실행하여 분석된 요약 집합정보(평균, 합, 표준편차, 오즈비, 위험도 등)만 수요자에게 회신하는 방식이다. 수요자는 폐쇄망 안에 있는 환자의 개별 정보를 보거나 취득할 수 없지만, 전체 데이터를 모아서 분석한 것과 동일한 분석 결과를 도출할 수 있다(Fig. 1). 다만 이를 위해서는 각 병원의 정형화된 전체 임상 데이터를 CDM을 이용하여 형식과 의미를 표준화하여야 한다.

Received: 2019. 6. 14
Accepted: 2019. 6. 17

Correspondence to Rae Woong Park, M.D., Ph.D.
Department of Biomedical Informatics, Ajou University School of Medicine, 164 World cup-ro, Yeongtong-gu, Suwon 16499, Korea
Tel: +82-31-219-4471, Fax: +82-31-219-4472, E-mail: veritas@ajou.ac.kr

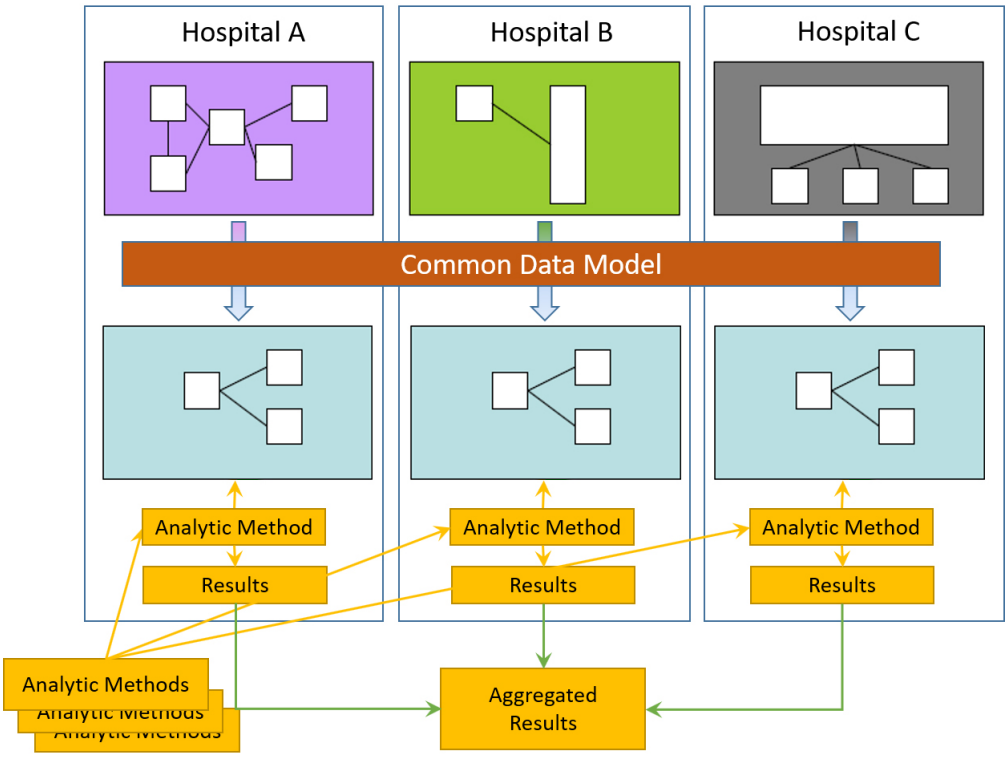

Figure 1. The common data model (CDM) and the distributed research network (DRN). The DRN uses the CDM, enabling analysis of data generated by the various software packages used by the participating organizations, and combines the results obtained from the entire network.

OHDSI

Observational Health Data Sciences and Informatics (또는 "Odyssey"라고 발음되는 OHDSI)는 2008년에 미국 정부의 지원으로 결성된 Observational Medical Outcomes Partnership (OMOP)으로부터 파생된 국제적 협의체이다. 초기에는 관찰 연구 방법론과 데이터를 활용하기 위한 분석 도구 및 시각화 도구, 그리고 각 기관마다 다른 진단, 처방 용어를 통일한 표준 용어를 만들었다. OMOP는 2013년 정부의 지원이 예정대로 종료된 후, 참여 연구자들이 OHDSI라는 이름으로 자발적으로 결성한 비영리 국제 연구단체이며 중앙조정센터는 Columbia University에 위치하고 있다. OMOP-CDM과 표준 용어 정의 등을 이관하여 계속 발전시키고 있다. 특히 OMOP 시절에는 약물 부작용 조사 방법론에 초점을 맞추었지만, OHDSI로 이관한 이후에는 약물의 안전성, 비교효과 연구, 경제성 분석, 의료의 질, 인공지능 기반의 환자 개별 위험도 예측 등 임상 빅데이터 분석으로 진화해 나가고 있다.

OHDSI (https://www.ohdsi.org/)의 모든 솔루션은 Git Hub에 오픈 소스로 제공되며(https://github.com/ohdsi), OHDSI 연구 커뮤니티(http://forums.ohdsi.org/)는 여러 분야(임상 의학, 생물 통계학, 컴퓨터 과학, 역학, 생명 과학)에 걸쳐 연구자들의 적극적인 참여를 가능하게 하고 있고, 다양한 이해 관계자 그룹(예: 연구원, 환자, 제공자, 지불자, 제품 제조업체, 규제기관)을 포괄하고 있다.

OHDSI 프로젝트는 협업 구성원이 주도하고 리더십은 프로젝트별로 결정된다. 데이터 표준화, 안전 감시(safety surveillance), 비교 효과 연구, 인구학적 평가, 개인 맞춤형 위험 예측, 데이터 특성 분석, 품질 개선, 지리정보 등 빅데이터 기반의 거의 모든 관찰 연구 영역에 초점을 맞추고 다양한 연구가 진행되고 있다(https://github.com/OHDSI/StudyProtocolSandbox). OHDSI의 목표는 건강을 향상시키기 위하여 지역 사회가 협력하여 보다 나은 결정과 보살핌을 촉진하는 증거를 생성할 수 있도록 권한을 부여하는 것이다[2].

CDM

OHDSI가 개발한 OMOP-CDM은 지금까지 개발된 CDM 중 임상 정보를 가장 광범위하게 포함할 수 있는 데이터 구조를 가지도록 설계되어 진화하고 있다. OMOP-CDM은 서로 다른 관측 데이터베이스일지라도 체계적인 분석이 가능하도록 설계되었다. 이것이 가능한 이유는 서로 다른 데이터베이스에 포함된 데이터를 표준적인 공통 표현(용어, 어휘, 코드 체계)으로 변환한 다음, 데이터의 구조 또한 공통 형식(데이터 모델)으로 변환하여, 공통 형식을 기반으로 작성된 표준 분석 루틴 라이브러리를 사용하여 체계적인 분석을 수행하기 때문이다.

OMOP CDM은 v6.0 기준 총 41개 테이블로 이루어져 있으며, 임상 정보를 담당하는 clinical data, 임상 정보로부터 파생되는 derived elements, 기관의 정보를 담고 있는 health system data, 비용 관련 데이터를 담당하는 health economics, 표준 용어를 통칭하는 vocabulary, CDM의 메타 정보를 담고 있는 meta-data로 구분하여 정의하고 있다(https://github.com/OHDSI/CommonDataModel/wiki). OHDSI의 연구자들은 OMOP CDM의 발전과 CDM으로써의 적용을 위하여 노력하고 있는데, 특히 OHDSI의 여러 활동적인 커뮤니티들은 CDM 변환, 유지 보수 등을 위하여 서로 협력하고 있다.

표준 용어

CDM은 데이터 구조는 물론 의미의 표준화를 추구한다. 현재 OMOP 표준 용어에서는 약 60여 종류 이상의 국제표준 용어들이 300만 개 이상 포함되어 있으며 각 표준 용어들은 의미론적으로 상·하위 관계를 이루며 구성되어 있다. 하지만 국제표준 용어가 존재하더라도 각 의료기관이 그 국제표준 용어를 사용하지 않거나 부분적으로만 사용하기 때문에, 각 병원의 데이터를 조사하여 비표준 용어를 표준 용어로 변환하기 위한 용어 매핑 작업이 필요하다. OHDSI에서는 지속적인 OMOP 용어의 업데이트를 진행하고 있으며, 업데이트에 따라 OMOP 표준 용어가 비표준 용어로 변경되거나 용어들 간의 새로운 관계가 정의되기도 하기 때문에 업데이트 내용을 파악하여 용어들의 매핑 작업을 주기적으로 수행 및 관리하여야 한다.

분석툴

OHDSI는 다양한 데이터 세트를 CDM으로 변환할 수 있는 자원들을 제공할 뿐만 아니라 변환된 CDM의 각종 특징이나 통계값을 보여주는 Achilles, 표준 용어 검색툴인 Athena, 코호트 추출, 유병률 계산, 성향점수매칭, 기술 분석, 단변량/다변량 분석, 기계학습기반 예측모형 구축 등 통합 분석툴인 Atlas, 기관 간 공동 연구를 위한 Arachne라는 툴 등 100종 이상의 다양한 도구들을 제공한다. OHDSI의 오픈 소스 소프트웨어는 Git Hub 저장소에서 무료로 사용할 수 있다(https://GitHub.com/OHDSI/). 국내에서는 아주대학교 의료정보학과에서 CDM 기반 약 30종 이상의 오픈 소스 프로그램과 연구 프로토콜을 개발하여 공개하고 있다(https://GitHub.com/abmi).

연구 재현성

OHDSI의 가장 큰 특징 중 하나는 CDM 기반의 재현 가능한 연구(reproducible research)를 추구한다는 점이다. 최근 빅데이터 유행에 영합하여 많은 임상 연구들이 위양성 증거(false-positive evidence)를 양산하고 있는 것이 아닌가 하는 우려가 잇따르고 있다. 실제로 PLOS Medicine에 실린 논문에 따르면 연구 대상자 수가 충분치 않은 역학 연구의 경우 1/10 경우만이 믿을 수 있고, 논문을 위한 논문(discovery-oriented exploratory research with massive testing)의 경우 1,000개 중 1개만이 믿을만하다고 한다[3]. 연구 재현성을 가로막는 것으로는 다음의 4가지가 주요 요인으로 꼽힌다[4]: 충분치 않은 연구 대상자 수 또는 낮은 위험도 비(low statistical power), 출판 편향(publication bias), *p*값 해킹(*p*-value hacking), 결과를 알고 난 후 가설 재수립 HARKing (hypothesizing after results are known).

OHDSI는 *p*-value hacking 및 HARKing을 방지하기 위하여 투명성(transparency), 사전 정의(prespecify), 분석 검증(validaiton of anlaysis)의 3가지 연구 일반 원칙을 권고하고 있다. 투명성과 사전 정의를 위하여 모든 연구 프로토콜 및 분석 코드들은 Git Hub를 통하여 사전에 공개하는 것을 원칙으로 한다. 연구를 진행하기 위한 전체 파이프라인을 먼저 설계하고 공개함으로써, 재현성을 높이고 투명성을 제고할 수 있다. 공개된 코드들은 전 세계 연구자들에 의하여 신뢰성을 검토받을 수 있다. 또한 미리 지정한 100여 가지의 음성 대조군(negative control)을 연구자가 정의한 분석 프로토콜로 동시

에 분석하는 위증 검증(falsification test)을 수행하기를 권하고 있다. 이는 독립변수와 전혀 상관없는 결과 간의 상관관계가 정말로 유의미하게 나오지 않는지를 검증하는 것으로, 예를 들어 항혈소판제제의 효과 분석 연구에서 미리 정의한 음성대조군(예: 내향성 손톱의 발생 등)이 상관관계가 존재하는 것으로 결과가 나온다면 이는 전체 분석 과정에서 systemic error나 unmeasured confounder가 존재하거나 혹은 우연에 의한 결과가 존재함을 알 수 있다.

국가별 동향

미국은 OMOP-CDM을 개발하고 OHDSI를 주도하는 국가로써 약 200여 개 기관이 참여하여 총 19억 명분의 전자의무기록(20%) 및 보험청구자료(80%)가 CDM으로 변환되어 있다. National Institutes of Health, Food and Drug Administration (FDA), National Cancer Institute, eMERGE에서 OMOP-CDM을 공식 데이터 모델 중 하나로 채택하고 있다. 100만 명분의 임상 정보와 유전체 정보를 모으는 All of Us 연구 프로그램에서도 OMOP-CDM을 공식 데이터 모델로 채택하고 있다.

유럽연합의 혁신의학이니셔티브(Innovative Medicines Initiative, IMI)는 총 33억 유로의 예산을 보유한 유럽연합 및 유럽제약산업협회와의 공공-민간 파트너십으로 다양한 연구 프로젝트에 연구비를 지원하고 있다. IMI2는 2018년 12월에 European Health Data & Evidence Network (EHDEN) 프로젝트를 출범하였으며 향후 5년간 총 2,900만 유로(372억 원)의 연구비를 지원할 예정이다. EHDEN 프로젝트는 유럽 12개국 22개 이상의 기관들이 보유한 의료데이터를 CDM으로 변환하는 프로젝트로써 옥스퍼드 대학, Odysseus Data Services, 유럽환자연합 등과 같은 기관에서 인증된 의료 데이터를 통합하여 1억 명분 이상의 환자 데이터를 익명화하여 공통 데이터모델로 표준화하는 것을 목표로 하고 있다. 확보된 실제 임상 데이터를 통하여 임상 근거를 제공하여 건강, 질병, 치료법, 예후 및 새로운 치료법에 대한 연구를 수행할 계획이다. 또한 EHDEN 프로젝트는 미국이 주도하는 OHDSI 사업 및 한국이 주도하는 분산형 바이오헬스 빅데이터 플랫폼(FEEDER-NET)과 긴밀히 협업 중이다.

한국의 경우, OHDSI Korea (http://www.ohdsikorea.org)가 결성되어 활동 중이며, 산업통상자원부는 2018년부터 헬스케어 산업의 발전과 글로벌 경쟁력 확보를 위하여 CDM 기반 분산형 바이오헬스 빅데이터 플랫폼 구축 사업을 추진하고 있다(Fig. 2). 본 사업에서는 병원들의 전자의무기록 데이터를 CDM으로 변환하여 바이오헬스 융합 빅데이터망을 구축하고 유전체, 의료영상, 생체신호 등 다양한 비정형 의료데이터를 포괄할 수 있는 CDM 확장 모델을 개발하고 있다. 데이터 수요자와 공급자 연계 및 다기관 분석 프로세스를 코디네이팅하는 FEEDER-NET이라 명명된 중계 플랫폼을 개발하고 있으며, 보건의료 연구자들과 관련 기업들이 이 플랫폼을 이용하여 혁신적인 서비스를 개발하도록 돕고 있다.

또한 산업통상자원부는 2019년에 바이오헬스 빅데이터망을 미참여 상급종합병원과 종합병원까지 확대하는 사업을 시작하였다. 이에 따라 기존 사업에 더하여 총 61개 병원들이 이 사업에 참여하게 되었다. 신규 사업에서는 기존 FEEDER-NET을 고도화할 분석지원 도구 및 데이터 질/보안 기술 개발 프로젝트는 물론 플랫폼을 활용한 혁신 비즈니스 모델 개발 과제 4개를 선정하였다. 비즈니스 모델 개발 과제의 주제는 다음과 같다: 임상시험 설계 지원 서비스; 병원 간 진료 지원 시스템; 지능형 진단/처방 조회 서비스; 디지털 바이오마커 개발 과제.

산업통상자원부뿐 아니라, 보건복지부에서도 과제 당 3년간 약 5억 원을 지원하는 CDM 활용을 통한 의학·정책·의료기술 등 공공목적의 연구과제 12개를 다음과 같이 선정하였다: 병리, 심전도, 심초음파 검사 결과 비정형데이터 확장모델 개발; 임상시험 확장모델; 자해 및 타해 행동 예측을 위한 정량화 모델 개발; 위암과 대장암 예방의 한국형 맞춤 모형 개발; 내분비 희귀 질환 개방형 플랫폼 구축; 감염성 질환 데이터 확장모델 개발; 스테로이드 약물 부작용의 분석 플랫폼 개발; 응급·중환자 대상 정밀의료 플랫폼 구현; 안질환 CDM 확장 표준 기반 난치성 실명질환의 임상 양상 및 비용 분석 네트워크 연구; 의료방사선 노출이 2차 암 발생에 미치는 영향 연구; 당뇨병 관리 효과 및 비용 분석과 개선방안. 보건복지부는 또한 공익적 목적의 CDM 활용을 위한 제도 및 정보보호기술 연구를 목표로 3년간 약 5억 원을 지원하는 10개 과제를 선정하였다.

도전

정형 임상자료 CDM 표준화에 비하여 비정형 임상자료의 활용 및 정형 자료와의 융합은 아직 미비하다. 비정형 데이터란 미리 정의된 데이터 모델이 없거나 정리되지 않은 정보를 뜻한다. 병원에서 생성되는 비정형 데이터로는 의사가 자유

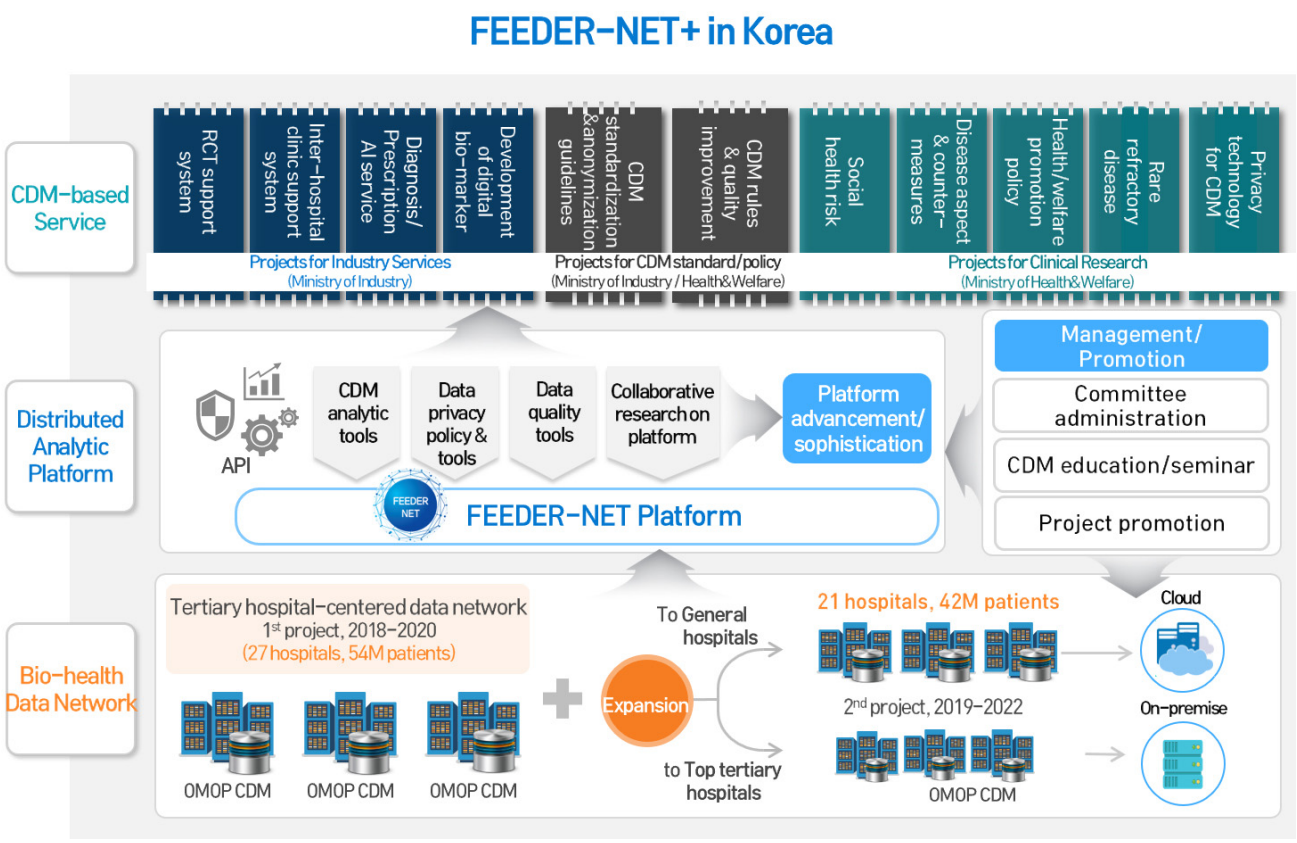

Figure 2. CDM projects funded by the government of South Korea. The Ministry of Industry and Trade supports adoption of the CDM by the 61 largest South Korean hospitals, as well as the development of application services and platform infrastructure. The Ministry of Health and Welfare is currently funding twelve clinical research projects and ten privacy protection and security projects that use the converted CDM data. CDM, common data model; RCT, random clinical trial; API, application programming interface; OMOP, Observational Medical Outcomes Partnership.

롭게 기입하는 자유진술문 형태의 각종 진료 기록이나 검사/시술 기록지, 병리나 방사선 영상, 내시경 사진 및 각종 동영상, 환자 감시 장치에서 생성되는 실시간 생체 신호, 유전자 검사 결과 등을 들 수 있다. 이들 비정형 데이터는 의료기관에서 발생하는 데이터의 80% 이상을 차지하지만 대부분 제대로 수집하지 않거나 관리되지 않아 버려지고 있다. 최근 한국을 중심으로 유전체 데이터를 위한 Genomic CDM [5], 방사선영상 데이터를 위한 Radiology CDM, 이들 데이터를 다룰 수 있는 어플리케이션 등 CDM 확장 모델과 각종 응용 소프트웨어를 활발히 개발하고 있는 점은 매우 고무적이라 할 수 있다.

동일한 데이터 구조에 대하여 하나의 분석 코드를 이용하여 분석을 진행할 경우, 각 기관별 분석에 걸리는 시간 자체는 비교적 짧은 편이다. 하지만 기관별로 윤리심의위원회의 심의를 요청하고 심의를 받는 과정은 다국적·다기관 공동 연구의 가장 큰 장애물이다. Sentinel Initiative의 경우 FDA가 요청에 의한 분석의 경우 Institutional Review Board (IRB) 심의 또는 IRB 면제심의 일체가 필요 없음을 정부에서 공표하였다. Patient-Centered Outcomes Research Institute (PCORI)의 후원으로 세워진 Patient Outcome Research to Advance Learning (PORTAL) 네트워크의 경우 연구별로 연구를 주도한 기관의 IRB 심의를 통과하면, 나머지 기관은 주도 기관의 IRB에 감독권을 이양하는 방식을 채택한다. 국내에서는 분산 연구망에 대한 이해 및 논의가 아직 부족한 실정으로 명확한 가이드라인이 없다. 현재까지는 모든 개별 연구에 대하여 각 기관별 IRB 심의를 받고 있어, 연구가 지연되는 경우가 많다. 생명윤리 및 안전에 관한 법률 제15조(인간대상연구의 심의)에서는 "① 인간대상연구를 하려는 자는 인간대상연구를 하기 전에 연구계획서를 작성하여 기관위원회의 심의를 받아야 한다."로 규정하고 있으며, 시행규칙 제2조

(인간대상연구의 범위)에서는 "인간대상연구"의 정의로써 사람을 대상으로 물리적으로 개입하는 연구, 상호작용을 통하여 수행하는 연구, 개인을 식별할 수 있는 정보를 이용하는 연구로 규정하고 있으므로, 연구자가 가명화된 CDM 데이터에 직접 접근이 불가하고 익명의 분산망을 통하여 간접적으로 분석 결과만 얻는 과정이 "인간대상연구"의 기준에 해당하는지, 따라서 IRB 심의나 심의면제 절차 자체가 필요한지에 대하여 의문이 제기되고 있다. 일부 참여 기관에서는 CDM 기반 다기관 연구가 IRB 심의나 심의면제 요건에 해당하지 않는 것으로 판단하여, 타 기관의 주 연구자가 IRB 심의를 받은 경우로써 단순히 분석코드를 실행만 해주는 경우에는 추가적인 IRB 심의나 IRB 면제심의를 요구하지 않고 있다.

결 론

OHDSI 컨소시엄에 임상 데이터를 가진 많은 기관들이 참여하고 있으며, 다양한 분야의 연구자가 자발적으로 참여하여 많은 응용 소프트웨어와 확장 모델, 다양한 분석 연구를 수행하는 등, 전 세계적으로 그 규모가 빠르게 확장되고 있다. 국내에서도 범정부차원의 지원으로 CDM 중심의 다양한 기초 및 응용 연구 프로젝트가 진행되고 있다.

CDM 기반의 분산연구망은 데이터의 소유와 분석을 분리함으로써 데이터를 소유하지 않으면서도 투명하고 재현 가능한 증거를 범세계적 규모로 쉽게 얻을 수 있는 연구망이다. 기술 분석이나 인과성 평가와 같은 통계적 기법뿐 아니라 기계학습 기반의 인공지능 알고리듬을 이용한 예측 모형도 쉽게 만들 수 있다. R이나 파이썬과 같은 프로그램을 이용하여 데이터 정제에서부터 변환/분석/시각화까지 일괄처리가 가능하여 많은 수의 대규모 분석을 동시에 진행할 수 있다. 일개 기관과 국가를 넘어서 전 세계 많은 기관과의 연계가 가능하므로 전 세계적 규모의 기관 및 국가 간 비교-통합 분석이 가능한 장점이 있다.

하지만 아직 비정형 데이터에 대한 표준화 기술은 더 많은 참여와 개발이 필요한 상황이다. CDM을 활용한 연구방법론에 낯선 많은 임상 연구가들에 대한 교육 프로그램도 시급히 마련되어야 한다. CDM으로 변환한 데이터에 대한 질 평가 방법론과 질 개선 방법론이 지속적으로 개발되어야 한다[6]. 기존의 전통적인 IRB 인증체계에서 요구하는 불필요한 절차 또한 개선되어야 할 필요성이 있다. 이러한 기여와 편익의 조화가 양순환을 이루기 위해서는 분산연구망 참여자들 간에 자유롭게 의견이 제시되고 민주적으로 결과를 도출하는 거버넌스 체계가 구축되어야 하며, 기여한 만큼 혜택을 받는 상호 호혜 기반의 정량 및 정성적 인센티브 제도가 마련되어야 할 것이다.

OHDSI 컨소시엄은 지역별, 주제별, 기술별 작은 단위의 커뮤니티를 지원하고 있으며, 이러한 커뮤니티는 OHDSI 생태계를 지속하는 힘이 되고 있다. 각 커뮤니티들이 협력을 통하여 분산연구망을 활성화시키고 발전시키는 시너지를 발휘한다면 OHDSI 생태계의 기반은 더욱 견고해질 것이며 의학적 발전과 인류 보건에 기여할 수 있게 될 것이다.

REFERENCES

1. Park RW. Sharing clinical big data while protecting confidentiality and security: observational health data sciences and informatics. Healthc Inform Res 2017;23:1-3.
2. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578.
3. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
4. Bishop D. Rein in the four horsemen of irreproducibility. Nature 2019;568:435.
5. Shin SJ, You SC, Park YR, et al. Genomic common data model for seamless interoperation of biomedical data in clinical practice: retrospective study. J Med Internet Res 2019;21:e13249.
6. Huser V, DeFalco FJ, Schuemie M, et al. Multisite evaluation of a data quality tool for patient-level clinical data sets. EGEMS (Wash DC) 2016;4:1239.