Ⅰ-1. 보건의료 분야 CDM 적용 사례 및 미래 활용 방안

박래웅 교수, 유승찬 연구원
아주대학교 의료정보학과

1. 들어가는 글

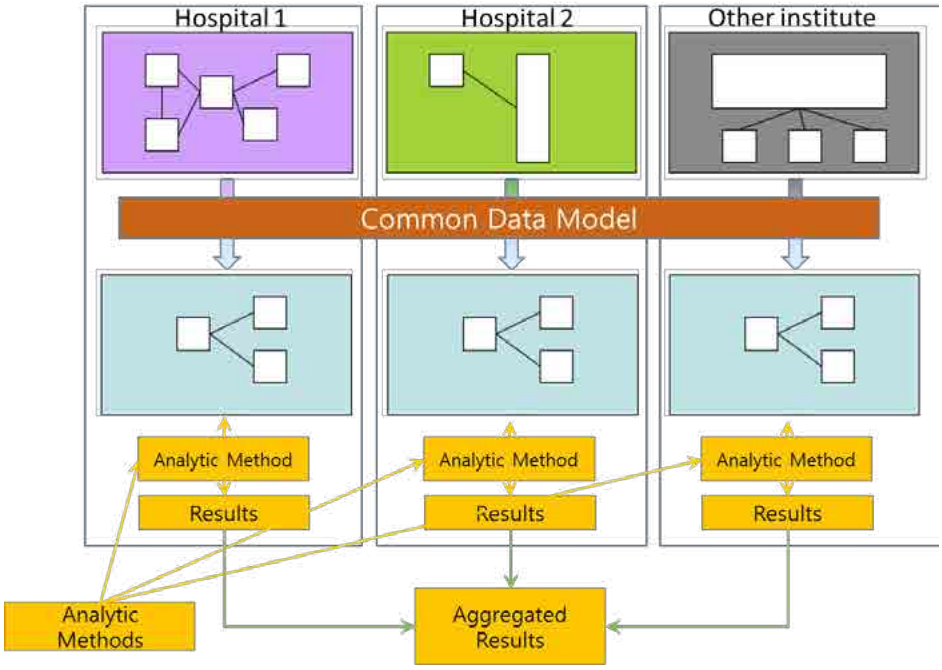
- 최근 전세계적으로 보건의료 빅데이터 활용에 대한 관심이 증가함
  - 보건의료 빅데이터를 이용한 질병 예방에 따른 의료비 절감, 의료기관의 운영비용 절감, 오류에 따른 손실비용 절감 등의 경제적 효과도 기대되어 활용이 증대되고 있음
  - 헬스케어 서비스를 통해 생산되는 건강정보 관리와 활용에 대한 논의가 활발히 진행되고 있으며, 보건의료 빅데이터에 대한 니즈와 관련 데이터, 대응 방향 등이 이슈화되고 있음

- 국내에서는 건강보험심사평가원(이하 심평원)에서 '보건의료빅데이터개방시스템'을 통해 연구자들에게 보건의료 빅데이터를 개방하면서 보건의료 빅데이터의 활용이 본격적으로 대두됨
  - 심평원은 전국민 기반의 실제 임상 데이터 (Real World Data)를 개인 단위의 세부 의료이용 데이터를 포함하여 관리하고 있으며, 이를 모두 표준화된 청구 서식을 기반으로 수집하고 있음
  - 하지만 데이터 수집의 목적이 심사 업무 수행이다보니 의료 연구를 위한 재처리 과정이 필요하고, 외부 데이터(타 기관, 타 국가 등) 연계, 비급여정보 부재 등의 한계점이 존재함

- 국제 공동연구를 위해 국내 대형 병원을 중심으로 용어와 구조가 표준화된 공통데이터모델 (Common Data model, CDM) 도입이 증가하고 있음
  - CDM을 통한 분산연구망 연구는 각 기관에서 같은 의미와 내용으로 연구에 필요한 데이터를 저장한 데이터 웨어하우스를 구축하고, 분석 코드를 공유하여 외부 데이터 연계, 공동 연구가 가능함

2. 분석연구망과 공통데이터모델

1. 분산연구망(Distributed Research Network, DRN)



[그림 1] 분산연구망의 개념

각 기관의 정형화된 임상정보 전체를 전세계적으로 동일한 구조와 의미를 갖도록 공통데이터모델로 변환한 후, 동일한 분석코드를 참여기관에 보내어 각 기관 안에서 분석하고 분석된 집합정보(평균, 합, 표준편차, 오즈비, 위험도 등)만 기관 외의 연구자에게 회신하며, 기관 외의 연구자는 여러 기관에서 회신된 집합정보를 모아서 결과적으로 환자의 개별정보는 보지 않으면서도 여러 기관의 자료를 모아서 분석한 것과 같은 결과를 낼 수 있음

- 의료 데이터는 기술적인 어려움(데이터 구조, 형식의 이질성 등)과 규제(개인정보보호문제, 기관승인) 등으로 인해 공유가 어려움
  - ▶ 현재까지 대부분 공동연구는 일부 환자 데이터를 연구 주도 기관과 공유하며 진행하였는데 기술적, 윤리적 문제들이 있었음

- 분산연구망은 수요자간 원본데이터 공유없이 분산된 데이터만 관리하고, 분석결과만 공유하여 위와 같은 문제점을 극복할 수 있음

2. 공통데이터모델(Common Data Model, CDM)

- 데이터 표준화는 협업 연구, 대규모 분석 및 정교한 도구 및 방법론 공유를 가능하게 하는 중요한 프로세스임
  - 연구에 사용할 데이터를 공통 형식으로 저장하여 협업 연구, 대규모 분석 및 정교한 도구 및 방법론 공유가 가능하지만 오랜 시간과 비용이 소요됨

- 공통데이터모델(CDM)은 여러 병원들의 데이터를 효율적으로 활용하기 위하여 정의한 표준화된 데이터 구조임
  - 기관별로 상이한 데이터 구조와 의미를 동일한 하나의 구조와 의미를 갖도록 변환하여, 다기관 공동 연구 수행 시, 기관 간 다른 데이터 구조로 인해 다양한 어려움이 따르는 것을 해결할 수 있음
  - 공통데이터모델을 따르기 위해서는 기존 데이터를 공통데이터모델로 변환하는 과정(ETL: Extract, Transform, Load)이 필요하며, 기존의 한계점 등을 고려하여 지속적으로 업데이트 되고 있음

- 대표적인 공통데이터모델로 비영리 국제컨소시엄인 오딧세이(Observational Health Data and Informatics, 이하 OHDSI), 약물부작용 조사를 위한 미국 FDA의 센티넬 공통데이터모델(이하 Sentinel CDM), 미국 국내에서의 비교효과연구를 위한 피코르넷(The National Patient-Centered Clinical Outcomes Research Network, 이하 PCORnet) 등이 있음

- OHDSI는 2008년에 미국정부의 지원으로 결성된 Observational Medical Outcomes Partnership(OMOP)으로부터 파생된 국제적 협의체로 초기에는 관찰연구 방법론과 데이터를 활용하기 위한 분석 및 시각화 도구와각 기관마다 다른 진단, 처방 용어를 통일한 표준용어를 만듬

- OMOP의 업무는 2013년 정부 지원이 종료된 후 OHDSI로 이관되어 계속되고 있으며, 인공지능 기반 환자 개별 위험도 예측 등의 임상 빅데이터 분석으로 진화해 나가고 있음
- OHDSI 프로그램은 대규모 분석을 통해 헬스 데이터의 가치를 창출하는 여러 이해 관계자 간의 학제 간 협력을 끌어내고 있으며, 연구자 및 관찰 건강 데이터베이스(observational health databases)의 국제 네트워크를 구축하였음
- 중앙 조정센터(central coordinating center)는 Columbia University에 위치하고 있으며, OHDSI에 참여하는 국제 협력기관들은 각 대륙에 고루 분포되어 있음

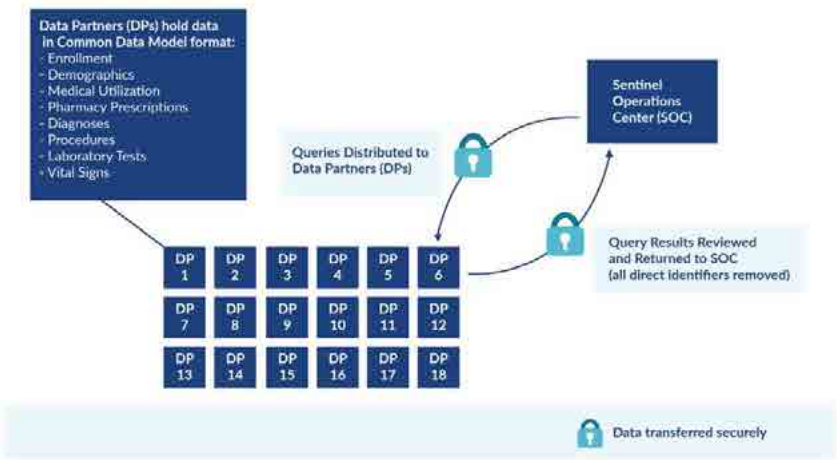
[그림 2] OHDSI Collaborators

- OHDSI의 모든 솔루션은 오픈 소스로 제공되고 있기 때문에, OHDSI 연구 커뮤니티에 여러 분야(임상 의학, 생물 통계학, 컴퓨터 과학, 역학 등) 연구자들의 적극적인 참여가 가능하고, 다양한 이해 관계자(연구자, 환자, 제조업체 등) 그룹을 포괄할 수 있음
- OHDSI 프로젝트는 협업 구성원이 주도하고 리더십은 프로젝트별로 결정되는데, 현재 빅데이터 기반의 대부분 분야(데이터 표준화, 의료제품 안전 감시, 개인 맞춤형 위험 예측, 지리정보 등)에 걸쳐 다양한 연구가 진행되고 있음

- Sentinel은 미국 식품의약국(Food and Drug Administration, 이하 FDA)로부터 시작되었으며, 의료 제품의 안전성 감시를 위한 국가적 전자시스템으로 Sentinel 시스템을 개발하였음

- 이 시스템은 FDA 규제 제품을 사용하여 보고된 이상 반응을 추적하는 기존의 감시 기능을 보완하여, FDA가 이러한 제품의 안전성을 사전에 평가할 수 있도록 함

- Sentinel은 데이터 파트너가 기존 환경에서 전자 데이터에 대한 물리적 및 운영상의 제어를 유지하는 분산 데이터 접근 방식을 사용함
  - 분산된 접근 방식은 Sentinel CDM으로 저장되며, 참여하는 파트너는 자신들이 보유한 데이터를 통일된 Sentinel CDM으로 변환하므로 하나의 동일한 분석 프로그램으로 여러 기관 결과를 동시에 분석할 수 있음
  - 개인정보보호를 위하여 분석 쿼리가 배포되고, 검색 결과가 보안 포털을 통해 반환됨
  - 모든 데이터 파트너들 사이에서 합쳐진 데이터 집합을 Sentinel Distributed Database (SDD)라고 한다.



[그림 3] Sentinel's distributed data approach

- PCORnet은 2013년에 The Patient-Centered Outcomes Research Institute (PCORI)에서 설립한 프로젝트로 환자 전자건강기록(Electronic health records, EHR)을 이용하여, 비교효과연구(Comparative effectiveness research, CER)를 수행하기 위한 목적으로 시작되었음

- 50개 주에 걸쳐 11개의 임상 데이터 연구 네트워크(Clinical data research networks, CDRNs)와 18개의 환자 참여 연구 네트워크(Patient - powered research networks, PPRNs)를 설립
- PCORnet이 구축하고 있는 연구 플랫폼의 핵심은 환자 중심의 접근 방식(patient-centered approach)이며 데이터는 중추 역할을 함

3. 학술연구 진행 및 소프트웨어 개발 현황

1. 분산연구망 기반 학술 연구 진행 현황

- 분산 연구망기반 학술 연구는 기본적으로 비실험적 혹은 후향적 관찰 연구(non-experimental or retrospective observational study)로 분류됨
  - 기존의 다기관 공동 연구 방식과 달리 하나의 연구기관이 다른 기관들의 데이터를 수집하여 통일화하지 않고, 약속된 공통데이터모델의 형식에 맞춘 데이터를 기관들이 각각 소유하며 진행됨
  - 현재까지의 분산 연구망 기반 학술 연구 진행 방식은 크게 협력센터(Coordinating center)를 거쳐 진행되는 방식, 중앙형 분산 연구망 연구(Centralized research system through DRN)방식, 개별 연구 기관의 연구자들간 협력을 통해 진행되는 방식, 탈중앙형 분산 연구망 연구(Decentralizd reserach system through DRN)방식이 있음

- Sentinel initiative, PCORNet은 협력센터 또는 그에 상응하는 협력 연구 집단(Collaborative research group)이 존재하여 중앙식 분산 연구망 연구를 진행하며, 모든 분석 및 연구는 협력센터 또는 협력 연구 집단을 통해 시작됨
  - 협력 연구 집단, 개인 연구자가 원하는 연구 디자인을 협력 센터에 공동 연구를 신청하면 Sentinel initiative에서는 미국 FDA가 자발적 부작용 보고 등을 기반으로 협력센터에 의약품 안정성 검사를 요청하는 형식으로 시작됨
  - 협력 센터는 제안된 연구를 검토한 후 승인할 수 있는 권한을 가지며, 이후 승인된 연구에 대해 분석 코드를 작성하여 각 데이터 파트너들에게 분석을 요청함

- 데이터 파트너들은 분석 결과를 협력 센터에 전송하고, 협력 센터는 분석 결과를 검토하고 취합하는 역할을 함

- OHDSI는 탈중앙형 분산 연구망 연구를 채택하고 있으며, 개별 연구자들은 누구나 본인이 원하는 연구에 대한 프로토콜과 분석 코드를 작성하여 홈페이지 등에 게시할 수 있음
  - CDM 의료 데이터를 보유하고 있는 기관의 연구자들은 이를 개별적으로 검토한 후 연구에 대한 참여의사를 밝히고, 분석 코드를 수행하여 분석 결과를 연구자에게 전송함
  - 연구자는 이 분석 결과를 취합하여 발표할 수 있으며, 논문 작성 과정도 클라우드 서비스 등을 이용하여 공동으로 수행할 수 있음

2. CDM 기반 분석 소프트웨어 및 인공지능 소프트웨어 개발 현황

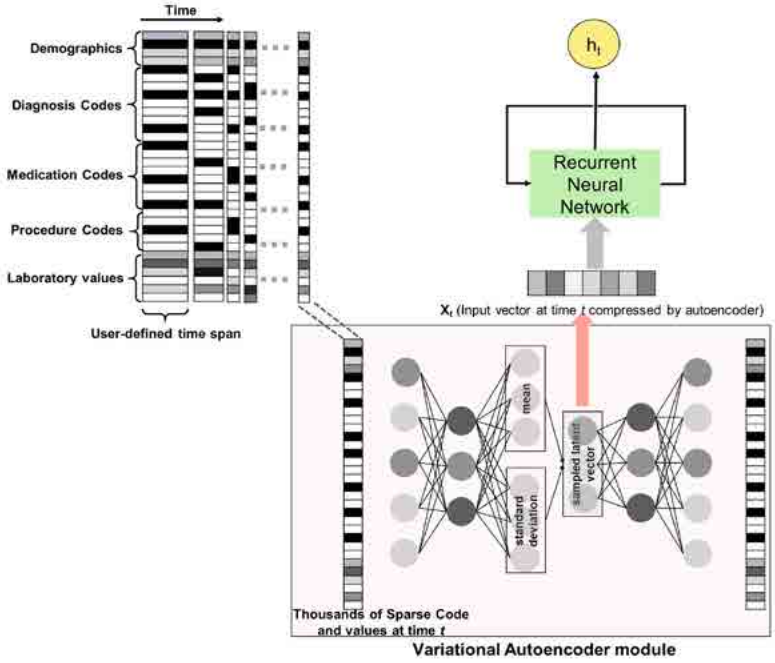
- 2017년 10월 구글팀은 EHR의 전송 표준인 FHIR을 기반으로 원내 사망 및 30일 이상 장기 입원을 예측하는 딥러닝 알고리즘을 개발 · 발표
  - 예측 및 분석 알고리즘 개발 시, 알고리즘 구성 자체가 아닌 데이터 전처리, 융합, 가공 등에 80% 이상의 노력이 사용되고 있다고 지적
  - 표준에 기반한 알고리즘을 개발하여, 해당 노력을 반복하지 않고도 확장 가능한 알고리즘 개발한 것을 장점으로 부각시키고 있음

- 소프트웨어 개발에 있어서 CDM의 장점은 다양한 의료기관 데이터에 적용할 수 있는 확장성임
  - 국내에서 여러 병원들이 다양한 기업과 협업하여 인공지능 시스템을 개발하고 있지만 개별 병원에서만 작동하며, 타 병원에서 작동할 수 있도록 설치하기 위해서는 막대한 데이터 전처리 작업이 필요함
  - 하지만 CDM 기반으로 개발한 소프트웨어는 한 번의 개발로 전 세계의 다양한 의료 기관 데이터에 적용이 가능하다는 장점이 있음

- OHDSI 내 소프트웨어 개발 환경의 또다른 장점은 다양한 오픈소스 프로그램에 기반한

에코시스템으로 OHDSI github(github.com/ohdsi)에 100개 이상의 다양한 프로그램들이 등록되어 무료로 사용 가능함

- OHDSI methods library내의 핵심 프로그램들은 서로와 밀접한 연관을 맺고 발전하고 있으며, 새로운 프로그램 개발 시 이러한 핵심 프로그램들을 다양하게 활용하여 개발 시간을 현격하게 단축할 수 있음
- 현재 OHDSI에는 연구자가 원하는 환자에서 원하는 이벤트를 예측할 수 있는 딥러닝 코드를 구성해주는 패키지(CIReNN, 그림4)가 공개되어 있음



[그림 4] OHDSI 딥러닝 모듈(CIReNN)

4. 심평원 빅데이터의 CDM 변환 경험

- 심평원 빅데이터의 CDM 변환 가능 및 활용 여부를 확인하기 위하여 2017년 하반기부터 심평원 빅데이터의 일부를 CDM으로 변환하여 연구를 수행하는 프로젝트가 진행됨
  - ▸ 최근 의료이용이 증가하고 있는 경피관상동맥시술(Percutaneous Coronary Intervention, PCI)을 받은 환자 대상(2007 ~2016년)
  - ▸ 변환 후 명세서 건수 및 환자수를 비교하여, 변환 시 누락데이터가 없음을 확인

[성별에 따른 연도별 명세서 건수 및 환자수 비교: 청구데이터 vs. CDM 데이터]

구분		청구데이터		CDM 데이터	
년도	성별	청구건수	환자수	청구건수	환자수
2007	여	4,892,538	144,474	4,892,538	144,474
	남	6,298,763	292,318	6,298,763	292,318
2008	여	6,259,273	144,630	6,259,273	144,630
	남	8,076,416	294,645	8,076,416	294,645
2009	여	6,518,688	143,839	6,518,688	143,839
	남	8,623,031	296,083	8,623,031	296,083
2010	여	6,511,313	142,649	6,511,313	142,649
	남	8,940,677	295,925	8,940,677	295,925
2011	여	6,530,394	140,685	6,530,394	140,685
	남	9,286,745	294,737	9,286,745	294,737

구분		청구데이터		CDM 데이터	
년도	성별	청구건수	환자수	청구건수	환자수
2012	여	6,814,840	138,254	6,814,840	138,254
	남	10,304,676	292,415	10,304,676	292,415
2013	여	6,663,341	135,086	6,663,341	135,086
	남	10,396,879	288,629	10,396,879	288,629
2014	여	6,493,095	131,526	6,493,095	131,526
	남	10,403,274	284,264	10,403,274	284,264
2015	여	6,125,843	127,269	6,125,843	127,269
	남	10,256,063	278,781	10,256,063	278,781
2016	여	5,864,972	122,555	5,864,972	122,555
	남	10,156,918	272,197	10,156,918	272,197

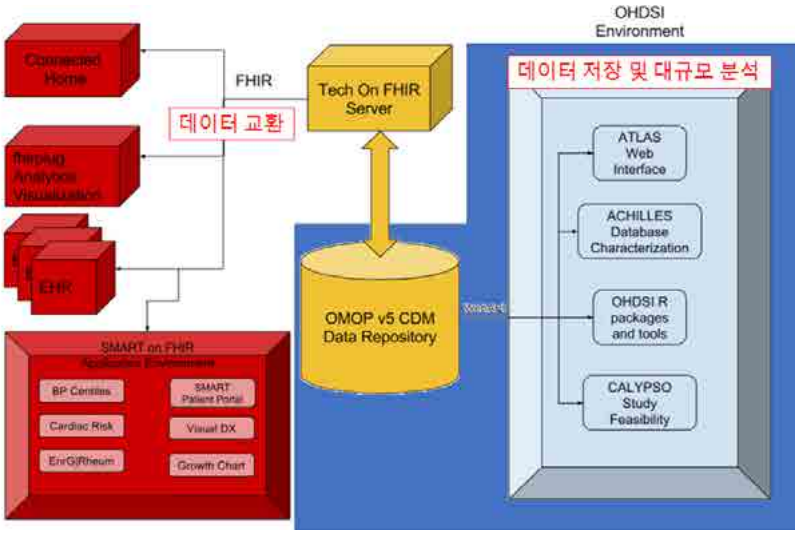
[그림 5] 데이터 비교

- 현재 해당 데이터를 이용하여 아스피린 클로피도그렐 병합 요법과 이스피린 티가그렐러 병합 요법의 주요 합병증(관상동맥재시술, 심근경색, 출혈 부작용)을 비교하는 국제 공동 연구를 진행
  - ▸ 본 연구를 통하여 서구 데이터를 중심으로 효과가 입증된 신약의 시판 후 조사(post-marketing surveilliance)를 CDM 및 분산 연구망을 이용하여 효율적으로 수행할 수 있는지 확인
  - ▸ 같은 연구 프로토콜의 결과를 해외의 다양한 데이터베이스에서 확인하여 국내 인구의 보건학적 특성을 다른 나라와의 비교할 수 있을 것으로 기대됨

5. 분산연구망을 이용한 CDM의 한계

- 기존의 CDM은 EMR, EHR 혹은 보험 청구자료 등의 임상데이터는 포함할 수 있으나 생체신호, 라이프로그, 유전체 정보, 영상정보 등의 대규모 비정형자료를 포함할 수 없었음
  - 또한 분석에만 초점이 맞추어져 있기 때문에 쌓인 과거 데이터를 한꺼번에 변환하여 이용하는 방식을 취하고 있으므로 실시간으로 쌓이는 데이터를 변환하지는 않음
  - 따라서 타 기관 자료를 통해 병원에서 실시간 임상의사 결정 지원시스템에 활용하기 위해서는 데이터의 실시간 변환기능이 필요함

- CDM은 데이터 저장표준으로서 데이터 전송 기능이 없지만 데이터의 전송을 목적으로 하는 HL7의 FHIR 표준과 병행하여 사용하면 해결할 수 있음
  - OHDSI내에서 FHIR와 OMOP-CDM을 결합하여 사용하기 위한 워크그룹이 결성되어 활동 중이며, 국내에서는 아주대학교를 중심으로 42개 병원이 한국 OHDSI 컨소시엄을 결성하여 활발하게 활동 중 임



[그림 6] HL7 FHIR 및 OMOP-CDM 협업 모델

6. 향후 발전 방안

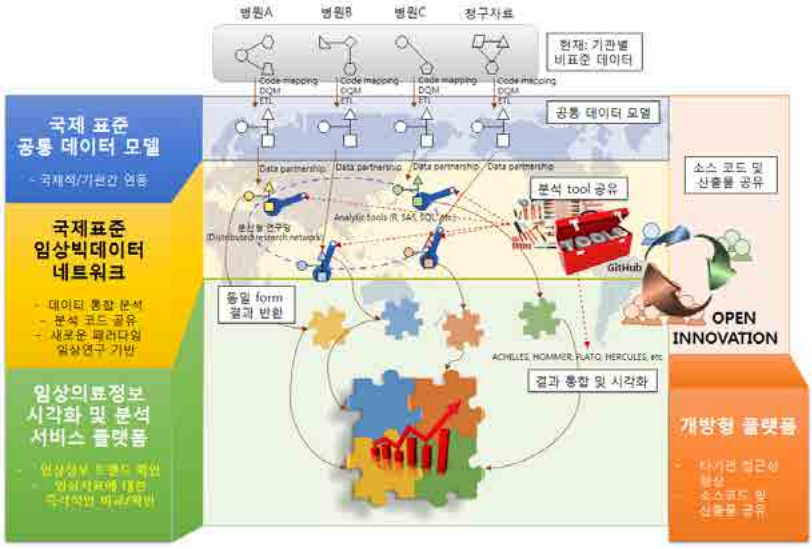
- 현재까지 의료정보 공유를 위한 많은 노력과 연구가 있었지만 특정기관에서 만든 시스템을 다른 기관에서 사용하기는 거의 불가능했음
  - ▶ 데이터 공유에 대한 동기와 보상이 없고, 개인정보보호법과 생명윤리 및 안전에 관한 법률 등에 의해 여러 가지 제약이 있었음

- 최근 분산연구망 이용한 공통데이터모델을 활용한 연구가 시작되고 있으며, 국내에 도입된 OHDSI 국제컨소시엄은 임상자료, 병원 자료, 원무자료, 메타자료, 추출요소, 표준용어 등 총 6개의 대분류 하에 총 36개의 테이블로 구성되어 있음
  - ▶ 이 기준에 따르면 국가, 언어, 기관에 상관없이 모든 데이터가 같은 구조와 의미를 갖고 데이터는 익명화됨
  - ▶ 이에 연구자들은 여러 기관의 자료를 바탕으로 연구 분석이 가능함

- 최근 개발하려는 시도가 있는 한국형 CDM은 쉽고 간단하게 만들 수 있는 장점이 있지만 국외 기관과의 공동연구나 분석의 가능성이 아직 확인되지 않았고, 활용 가능한 임상자료원, 분석플랫폼이나 분석툴이 국내 사용자에게만 국한되는 단점이 존재함
  - ▶ 결과적으로 빅데이터화 하거나 다양하고 혁신적인 응용소프트웨어 개발 및 활용과 국제적 리더쉽 확보가 어려워 지속적으로 성장 · 발전하는 플랫폼이 되기 어려울 것으로 판단됨

- 임상정보에 국한된 CDM모델을 확장하여 생체신호, 라이프로그, 유전체 정보, 레지스트리, 영상정보를 포괄할 수 있는 형태로 확장시킬 경우 진정한 의료 빅데이터 시대로 한걸음 더 다가설 것임

※ 미리 보는 이상적인 임상의료 빅데이터 오픈 플랫폼



- 세계의 모든 기관이 국제 표준의 공통데이터 모델 형태의 데이터 보유 (병원/청구자료 포함)
- 공통데이터 모델을 갖고 있는 기관들이 임상빅데이터 네트워크를 형성
- 네트워크 내 모든 기관이 분석 툴(R, SAS, SQL 등)을 공유하여 동일 형태의 결과 반환
- 모든 요소 기술은 공개되고, 사용자들이 직접 수정 및 개발하여 open innovation 실현

참고문헌

[1] Ruping S. [Big data in medicine and healthcare]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2015;58(8):794-8.

[2] Maynard AD. Navigating the fourth industrial revolution. Nat Nanotechnol. 2015;10(12):1005-6.

[3] Yoon D, Ahn EK, Park MY, Cho SY, Ryan P, Schuemie MJ, et al. Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research. Healthc Inform Res. 2016;22(1):54-8.

[4] OHDSI. Welcome to OHDSI! 2017 [cited 2017. Available from: https://www.ohdsi.org/.

[5] Hripcsak G DJ, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Studies in health technology and informatics. 2015;2017(216):574-8.

[6] OHDSI. Areas of Focus 2017 [cited 2017. Available from: https://www.ohdsi.org/who-we-are/areas-of-focus/.

[7] OHDSI. Mission-Vision-Values 2017 [cited 2017. Available from: https://www.ohdsi.org/who-we-are/mission-vision-values/.

[8] Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative--A comprehensive approach to medical product surveillance. Clin Pharmacol Ther. 2016;99(3):265-8.

[9] Sentinel. Distributed Database and Common Data Model [cited 2017. Available from: https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model.

[10] pcornet. PCORnet Common Data Model (CDM) [cited 2017. Available from: http://www.pcornet.org/ pcornet-common-data-model/.

[11] Seongwon Lee PD, Soo-Yeon Cho, R.N., MPH., Seng Chan You, M.D., M.S., Hojun Park, BS., Sungjae Jung, BE., Rae Woong Park, M.D., Ph.D,Yunyoung Bae, MPH., Hangil Lee, MPH., Jahyun Cho, MPH., Keunhui Park, MS. . Conversion of National

Health Insurance Service (NHIS) Data of Korea to the Observational Medical Outcomes Partnership (OMOP) Common Data Model [cited 2017. Available from: http://www.ohdsi.org/web/wiki/lib/exe/ fetch.php?media=resources:2016_ohdsi_symposium_ajou.pdf.

[12] OHDSI. Data Standardization [Available from: https://www.ohdsi.org/data-standardization/.

[13] Sentinel. Overview and Description of the Common Data Model v6.01 [cited 2017. Available from: https://www.sentinelinitiative.org/sites/default/files/data/DistributedDatabase/Sentinel_Common- Data-Model_v6.01.xlsx.

[14] pcornet. PCORnet Common Data Model v3.0 Specification [Available from: http://www.pcornet.org/ wp-content/uploads/2014/07/2015-07-29-PCORnet-Common-Data-Model-v3dot0-RELEASE.pdf.

[15] i2b2. Informatics for Integrating Biology and the Bedside (i2b2) Overview [cited 2017. Available from: https://www.i2b2.org/about/index.html.

[16] Paolino AR, McGlynn EA, Lieu T, Nelson AF, Prausnitz S, Horberg MA, et al. Building a Governance Strategy for CER: The Patient Outcomes Research to Advance Learning (PORTAL) Network Experience. EGEMS (Wash DC). 2016;4(2):1216.