




# scientific **data**



**OPEN**  
ARTICLE

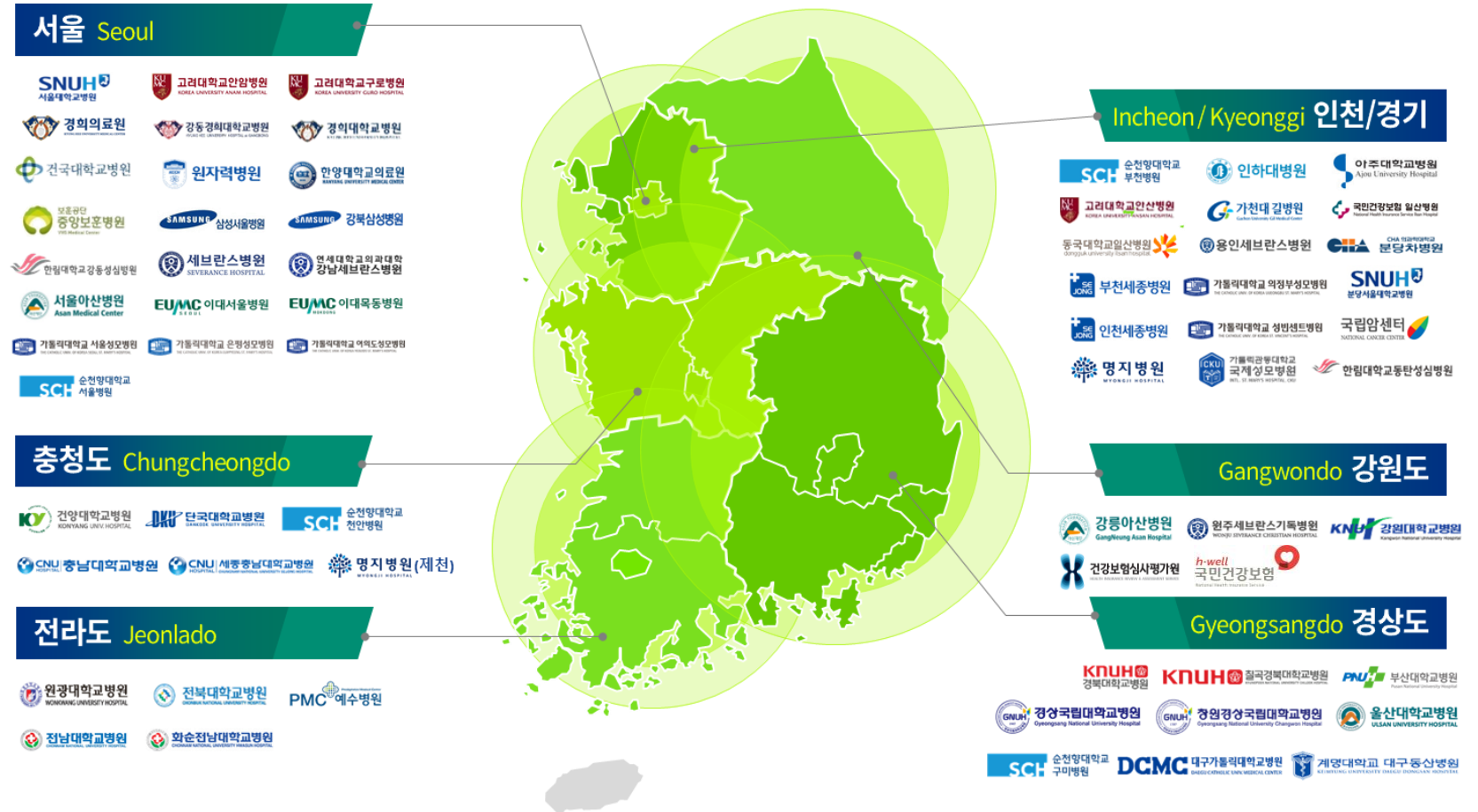
## Scalable Infrastructure Supporting Reproducible Nationwide Healthcare Data Analysis toward FAIR Stewardship

Ji-Woo Kim<sup>1,7</sup>, Chungsoo Kim<sup>2,7</sup> , Kyoung-Hoon Kim<sup>3</sup>, Yujin Lee<sup>3</sup>, Dong Han Yu<sup>1</sup>,  
Jeongwon Yun<sup>1</sup>, Hyeran Baek<sup>1</sup>, Rae Woong Park<sup>2,4,8</sup>  & Seng Chan You<sup>5,6,8</sup> 

# OMOP in Korea

## 65 Data partners

- 63 Hospitals  
(70% of tertiary hospitals)
- 2 National Institutions
- Coordinating center  
: Ajou University



# National Health Insurance Data in Korea

## National Claims

- From Clinics, general/tertiary hospital, pharmacy
- Including demographics, income level, diagnosis, prescription, measurement (no value of measures), procedure, device, etc
- Cost
- National Health checkup data (NHIS only)

## Data Linkage

- Death registry (Statistics Korea)
- Cancer registry (National Cancer Center)
- Vaccine registry (KCDC)

## Health Insurance Review and Assessment Service (HIRA)



# HIRA CDM projects

## 2018 - 2020 Pilot CDM ETL Study published in JAMA (Chan et al)

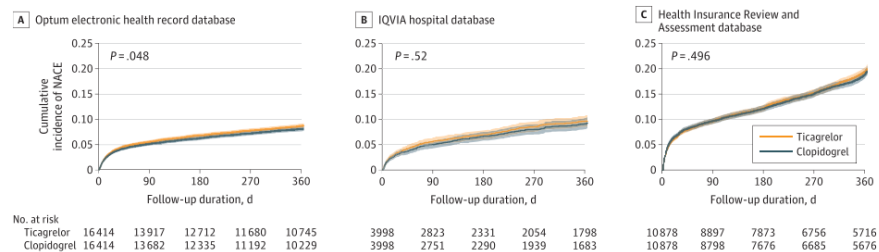
Research

JAMA | Original Investigation

### Association of Ticagrelor vs Clopidogrel With Net Adverse Clinical Events in Patients With Acute Coronary Syndrome Undergoing Percutaneous Coronary Intervention

Seng Chan You, MD, MS; Yeunsook Rho, PhD; Behnood Bikdeli, MD, MS; Jiwoo Kim, MS; Anastasios Sianos, MSc; James Weaver, MSc; Ajit Londhe, MPH; Jaehyeong Cho, BS; Jimyung Park, BS; Martijn Schuemie, PhD; Marc A. Suchard, MD, PhD; David Madigan, PhD; George Hripcsak, MD, MS; Aakriti Gupta, MD, MS; Christian G. Reich, MD; Patrick B. Ryan, PhD; Rae Woong Park, MD, PhD; Harlan M. Krumholz, MD, SM

Figure 2. Kaplan-Meier Plots for the Risks of the Primary Outcome (Net Adverse Clinical Events) Associated With Ticagrelor and Clopidogrel



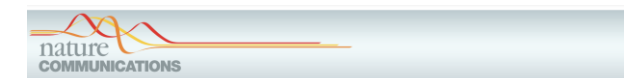
## 2020 OpenData4Covid project World first data open for research (Published in many journals)



ARTICLE

<https://doi.org/10.1038/s41467-022-29160-4> OPEN

DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models



ARTICLE

<https://doi.org/10.1038/s41467-020-18849-z> OPEN

Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study

Edward Burn et al.



A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data

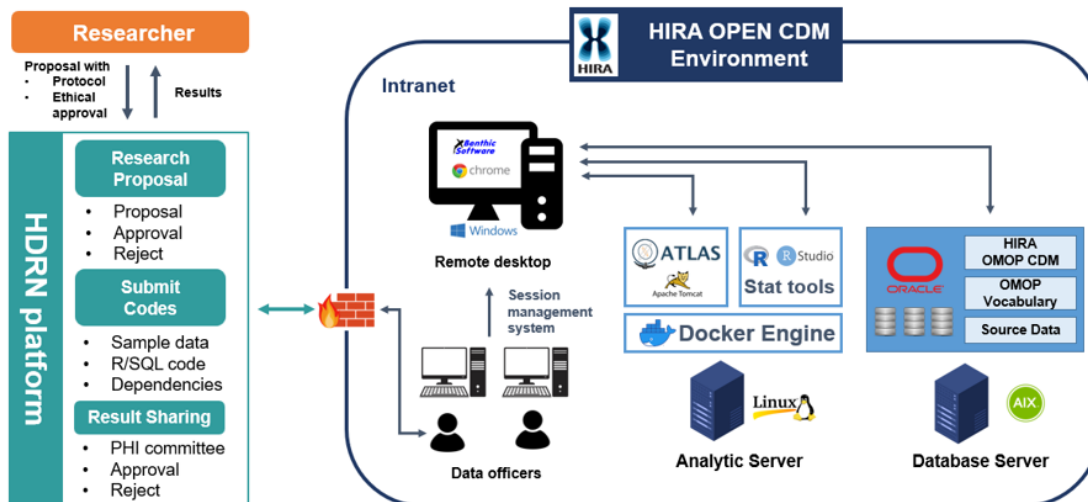
Sara Khalid<sup>1,2</sup>, Cynthia Yang<sup>1,2</sup>, Clair Blacketer<sup>3</sup>, Talita Duarte-Salles<sup>4</sup>, Sergio Fernández-Bertolin<sup>5</sup>, Chungsoo Kim<sup>6</sup>, Rae Woong Park<sup>6,7</sup>, Jimyung Park<sup>8</sup>, Martijn J. Schuemie<sup>9</sup>, Anthony G. Sena<sup>10</sup>, Marc A. Suchard<sup>8</sup>, Seng Chan You<sup>1</sup>, Peter R. Rijnbeek<sup>11</sup>, Jenna M. Reps<sup>12,13</sup>

# HIRA CDM projects

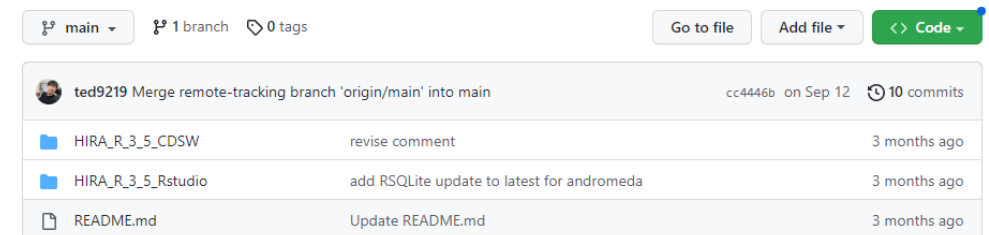
## HIRA CDM analytic environment

- ABMI supports HIRA to manage their own analytic environment for future research.
- It was not easy for implementing R into the off-line environment, but we resolved using Docker.

## On-premise environment (manually developed)



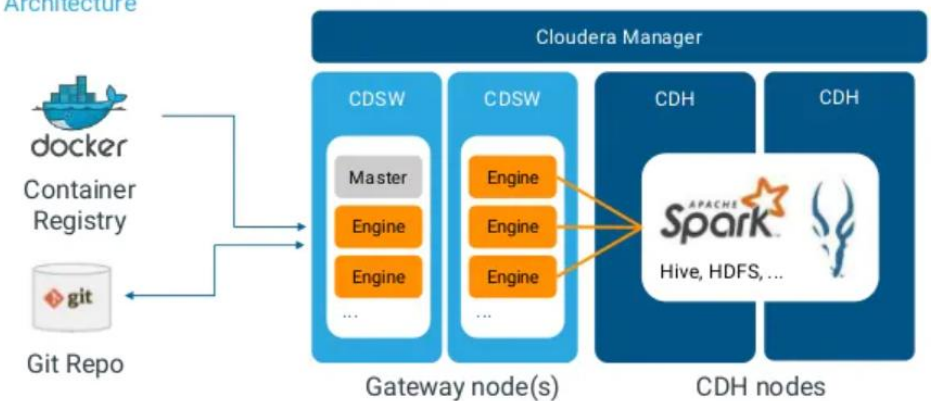
## Off-the-shelf platform + Docker (flexible)



HiraRDocker

Cloudera Data Science Workbench

Architecture



# Data ETL

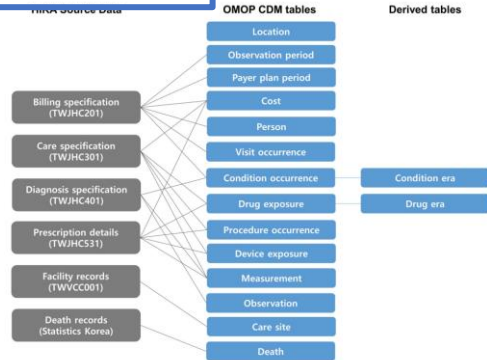


Fig. 3 Data mapping to OMOOP-CDM from HIRA source claims database. OMOOP: Observational Medical Outcome Partnership; CDM: common data model; HIRA: Health Insurance Review and Assessment Service.

# Mapping (EDI to SNOMED)

요약 [기타] 의료행위 분류 용어(처치 및 수술 등) EDI-SNOMED CT 매핑 결과 공개  
담당부서: 빅데이터연구부 작성일: 2021-05-31 조회수: 6,442

안녕하십니까?  
건강보험심사평가원 빅데이터실 빅데이터연구부입니다.  
건강보험심사평가원은 건강보험행정규약어목(EDI)을 개발 및 관리하는 주제로서, 국제표준체계와 연계하여 국내 의료행위 용어 표준화를 위한 기틀을 마련하고자 '건강보험행정규약어목', '처치 및 수술 등' 28개 관리진료과 2,628개 EDI코드에 대하여 SNOMED CT와 매핑 테이블을 생성하였습니다.

이와 관련, 결과 보고서와 매핑테이블을 첨부하였으니, 일무에 참고하여 주시기 바랍니다.  
아울러 정기적 매핑 테이블 업데이트를 통한 매핑 결과 유지보수를 위해 용어 표준화 개선의견 창구가 마련되었습니다. (\*심사평가원 보건정보빅데이터개발소통: opendata.hira.or.kr ->고객문의->개선의견 -용어 표준화 개선의견) 의료현장에서 수용성 높은 매핑 테이블로 발전할 수 있도록 많은 관심과 의견 부탁드립니다.

기타 문의사항은 건강보험심사평가원 빅데이터실 빅데이터연구부 Te(033-739-1049, 1069로 연락 주십시오. 감사합니다.

첨부파일

- [심사평가원]EDI-SNOMED CT 매핑테이블(처치/수술/분류)\_202105.xlsx (536.2 KB)
- [심사평가원]매핑결과보고서(처치)\_202105.pdf (88.2 MB)

# Open to public

# Data catalog (ETL rule and sample data)

# Quality check

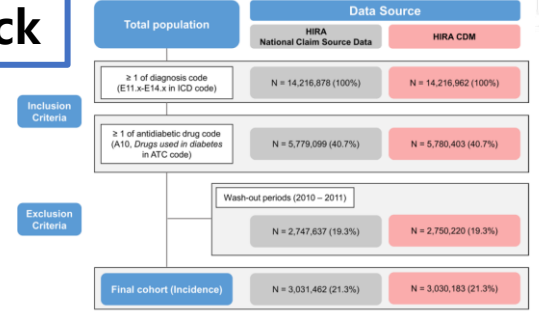


Fig. 1 Flow chart of type 2 diabetes mellitus phenotype and comparison of incidences from the source and converted CDM databases. CDM: common data model.

# Analytic infra

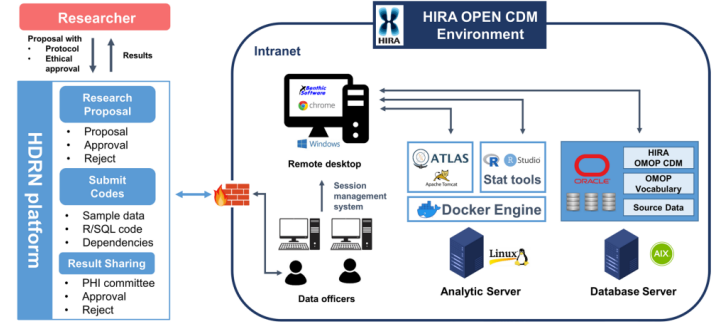


Fig. 2 HIRA CDM analytic environment and data open process. Researchers can request the use of the HIRA CDM through the HDRN platform, which is an open public healthcare data platform. HDRN, Healthcare Distributed Research Network; PHI, personal health information; HIRA, Health Insurance Review and Assessment Service; CDM, common data model; OMOP, Observational Medical Outcome Partnership.

scientific data

OPEN ARTICLE

## Scalable Infrastructure Supporting Reproducible Nationwide Healthcare Data Analysis toward FAIR Stewardship

Ji-Woo Kim<sup>1,2</sup>, Chungsoo Kim<sup>3,7</sup>, Kyoung-Hoon Kim<sup>3</sup>, Yujin Lee<sup>3</sup>, Dong Han Yu<sup>1</sup>, Jeongwon Yun<sup>1</sup>, Hyeran Baek<sup>1</sup>, Rae Woong Park<sup>3,4,5,6</sup> & Sang Chan You<sup>3,4,5,6</sup>

Transparent and FAIR disclosure of meta-information about healthcare data and infrastructure is essential but has not been well published. In this paper, we provide a transparent disclosure of the process of standardizing a common data model and developing a national data infrastructure using national claims data. We established an Observational Medical Outcome Partnership (OMOP) common data model database for national claims data of the Health Insurance Review and Assessment Service of South Korea. To introduce a data openness policy, we built a distributed data analysis environment and released metadata based on the FAIR principle. A total of 10,096,730,243 claims and 56,579,726 patients' data were converted as OMOP common data model. We also built an analytics environment for distributed research and made the metadata publicly available. Disclosure of this infrastructure to researchers will help to eliminate information inequality and contribute to the generation of high-quality medical evidence.

Introduction  
Numerous studies using routinely collected large healthcare data have provided invaluable evidence representing routine clinical practice<sup>1-4</sup>. Administrative data representing the nationwide population have been used for secondary analysis in healthcare research for various purposes, including consecutive monitoring of disease and medical expenditure, comparative effectiveness of medical interventions, and even machine learning<sup>5-7</sup>. The Korean National Health Insurance system is a single public insurance system for all citizens, and all medical institutions are applied as mandatory designation systems. The Health Insurance Review and Assessment Service (HIRA) establishes health insurance reimbursement criteria and reviews all medical claims for reimbursement. Therefore, the HIRA has accumulated a vast amount of claims data at the national level, and it can be used as a secondary data source for high-quality real-world evidence<sup>8</sup>. For example, statistics from the HIRA database are used in OECD statistics as representative statistics for Korea.  
Administrative data, despite being a commonly used source for research, has drawn significant criticism predominantly due to concerns over the validity of its coded information. For instance, coding practices like "upcoding" can lead to inaccuracies; this is where providers code for a more severe illness than the patient actually has to receive higher reimbursement<sup>9,10</sup>. While the debate on coded information's validity continues, less attention is being directed towards the stewardship of this extensive healthcare data. Chief among these are issues including: 1. Non-scalability and non-interoperability. 2. Ignored reproducibility; and 3. Protection of

<sup>1</sup>Big Data Department, Health Insurance Review and Assessment Service, Wonju, Republic of Korea. <sup>2</sup>Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea. <sup>3</sup>Review and Assessment Research Department, Health Insurance Review and Assessment Service, Wonju, Republic of Korea. <sup>4</sup>Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea. <sup>5</sup>Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea. <sup>6</sup>Institution for Innovation in Digital Healthcare, Yonsei University, Seoul, Republic of Korea. <sup>7</sup>These authors contributed equally: Ji-Woo Kim, Chungsoo Kim. <sup>8</sup>These authors jointly supervised this work: Rae Woong Park, Sang Chan You. \*e-mail: veritas@ajou.ac.kr; chandryou@yuhs.ac

SCIENTIFIC DATA | (2023) 10:674 | https://doi.org/10.1038/s41597-023-02580-7

# HIRA CDM projects

## HIRA CDM Open Project

### 1<sup>st</sup> pilot opening with 10M data (2022)

### 2<sup>nd</sup> pilot opening with 50M data (2023)

#### 심평원, 국제 표준 연구 가능한 보건의료 CDM 개방

정년기사 | A 김은영 기자 | © 입력 2022.05.25 16:30 | [홈](#) | [우측](#)

**데이터 교류 없이 분석 값 제공 '개인정보 유출 위험 No'**

건강보험심사평가원이 국제 표준 연구 가능한 보건의료 데이터 모델을 개방한다.

심평원은 25일 국기와 국가, 국가와 특정 기관과의 비교연구가 가능한 국제표준 공통데이터모델(Common data model, CDM)을 공개한다고 밝혔다. CDM은 각 기관이 보유한 데이터의 구조와 용어를 표준화한 모델이다.

연구자들은 CDM을 통해 심평원의 국가단위 진료정보와 요양기관 등의 자료를 비교해 데이터 분석을 파악할 수 있다. 특히 이 모델은 데이터 자체가 아닌 분석결과 값이 공유된다는 점에서 민감한 개인정보 유출 위험이 없다는 게 특징이다.

심평원은 연구 수요가 높은 신종 코로나바이러스 감염증(코로나19) 환자 데이터를 우선 개방한다. 이 데이터는 전체 국민의 20% 수준인 5년치(2018~2022년) 국민인료정보로 구성된다.



건강보험심사평가원 전경

Health Map 의료이용지도 | 프로그램 | 지원사업

보건의료빅데이터개방시스템

공공데이터 | 의료빅데이터 | 의료동계정보 | 고객지원 | 시스템소개

Home | 고객센터 | 관리자

공지사항 | 보건의료빅데이터개방시스템에서 공지사항을 알려 드립니다.

**심사평가원 공통데이터모델(CDM) 개방(2단계) 이용 신청 안내**

담당부서	빅데이터전략부	작성일	2022-05-07	조회수	145 / 4
------	---------	-----	------------	-----	---------

건강보험심사평가원은 국제표준 공통데이터모델(CDM) 데이터를 개방하여 코로나19 관련 후유증, 이상행동 등의 연구를 지원하고자 합니다.

○ (개방 데이터) 2021년 한 해 동안 의료서비스를 이용한 전체 환자에서 총병 20%의 1천만 명) 기본 주술, 대상 환자의 2018년 1월부터 2022년 4월까지의 항구적 데이터를 공통데이터모델(CDM)로 변환한 데이터

○ (이용대상) 심평원 CDM 데이터 활용을 원하고 소속기관(학교, 의료기관 등)에 CDM을 보유하고 있는 국내 연구자

○ (신청기간) 2022.7.4(월) ~ 7.29(금) 16:00까지, 26일간

○ (제출서류) 첨부파일의 별지 양식 다운로드  
 ① 연구 책임자 소속기관 공문  
 ② 별지 제1호서식 이용개요서  
 ③ 별지 제2호서식 보안유지 및 준수사항 서약서(연구책임자, 이용자 각 1부)  
 ④ 별지 제3호서식 개인정보 수집 이용 동의서(연구책임자, 이용자 각 1부)  
 ⑤ 기관생명윤리위원회(IRB) 심의 결과 통지서  
 ⑥ 기관생명윤리위원회(IRB) 승인된 연구계획서  
 ⑦ 이용자 소속 증명 서류(재직증명서, 재직증명서, 근로계약서 등 1부)

○ (제출방법) 전자우편(cdm@hira.or.kr)으로 송부

자세한 내용은 첨부파일을 확인하여 주시기 바랍니다. (문의: 빅데이터전략부, 033-736-1041, 1098)

**중요** [보건의료빅데이터] 공통데이터모델(CDM) 개방(2단계) 이용 신청 안내

담당부서 | 빅데이터전략부 | 작성일 | 2023-07-31 | 조회수 | 1,098

- 건강보험심사평가원은 국제표준 공통데이터모델(CDM) 데이터를 개방하여 근거 기반 연구를 지원하고자 합니다.
- (이용대상) 심평원 CDM 데이터 활용을 원하고 소속기관(학교, 의료기관 등)에 CDM을 보유하고 있는 국내 연구자
  - (신청기간) 2023.8.1(화) ~ 8.31(목), 31일간
  - (이용 데이터) 전 국민DB와 표본DB 중 하나를 선택하여 신청
    - 전 국민DB
      - 전 국민의 2012년 1월부터 2021년 12월까지의(10년 분) 청구자료를 공통데이터모델(CDM)로 변환한 자료
    - 표본DB
      - 전 국민 20%인 약 천만 명의 2018년 1월부터 2022년 4월까지의(5년 분) 청구자료를 공통데이터모델(CDM)로 변환한 자료
  - (제출서류) HIRA빅데이터개방포털(opendata.hira.or.kr) > 서비스 소개 > CDM > "CDM 이용 관련 서식"에서 다운로드
    - 연구 책임자 소속기관 공문
    - 별지 제1호서식 이용개요서
    - 별지 제2호서식 보안유지 및 준수사항 서약서(연구책임자, 이용자 각 1부)
    - 별지 제3호서식 개인정보 수집 이용 동의서(연구책임자, 이용자 각 1부)
    - 기관생명윤리위원회(IRB) 심의 결과 통지서
    - 기관생명윤리위원회(IRB) 승인된 연구계획서
    - 이용자 소속 증명 서류(재직증명서, 재직증명서, 근로계약서 등 1부)
    - 분석실행 가능성을 확인할 수 있는 분석코드, 분석에상결과(소속기관에서 실행한 분석결과 제시)
  - (제출방법) 전자우편(cdm@hira.or.kr)으로 송부
    - \* 기간 이후 제출은 불가하며, 신청 다음 날까지 심평원 담당자가 발송하는 접수 메일로 신청되었음을 확인

About 40 studies were done using this data

About 15 studies are running using this data

Currently, HIRA is accepting applications 1-2 times a year, but they are preparing to accept applications more frequently.



아주대학교  
AJOU UNIVERSITY



Thank you



Big data department

Ji-Woo Kim  
Kyoung-Hoon Kim  
Yujin Lee  
Dong Han Yu  
Jeongwon Yun  
Hyeran Baek



Rae Woong Park  
Seng Chan You