NHIS GLOBAL FORUM 2025

보건의료 빅데이터와 AI 활용: 장애물, 돌파구, 의료기관의 역할

Healthcare Big Data and AI: Barriers, Breakthroughs, and Roles of Medical Institutions

Hyeon Chang Kim, MD, PhD, FAHA

Professor, Department of Preventive Medicine, Yonsei University College of Medicine

Director, Institute for Innovation in Digital Healthcare, Yonsei University Health System

Editor-in-Chief, Epidemiology and Health, Korean Society of Epidemiology

Big Data

Secondary Data

Administrative Data

Claims Data

Routinely-Collected Data

Real-World Data

https://scet.berkeley.edu/big-data-solutions-for-small-firms/word-cloud-big-data/

Most Available Healthcare Big Data Source in Korea

Category	Description	Representative Examples
Structured EHR Data	Electronic health records used in retrospective cohort studies, often in CDW or CDM format	• Clinical Data Warehouse (CDW) • Common Data Model (CDM) • Hospital-based cohort DBs
Unstructured Health Data	Data from clinical notes, free-text reports, and non-coded documents	• Physician's notes • Discharge summaries • Operative reports
Medical Imaging Data	Radiology and imaging data for diagnosis and AI model training	• CT/MRI/PET scans • PACS (Picture Archiving and Communication System) • Radiology reports
Genetic Data	Genomic epidemiology and hospital-based genome sequencing programs	• Korea Biobank Project • Hospital-based Whole Genome Studies • KoGES Genetic DB
Epidemiology & Behavioral	Data from longitudinal and national health surveys related to lifestyle and behavior	• Korean National Health & Nutrition Examination Survey (KNHANES) • Korean Genome and Epidemiology Study (KoGES) cohort • BioBigData.Korea (Korean PMI chort)
Health Screening	Mass screening data from national and private sources	• National Health Screening Program • National Cancer Screening Program • Corporate health checkups
National Health Insurance	Administrative data for health service eligibility, claims, and reimbursement	• Eligibility and enrollment data • Diagnosis/procedure codes • Medication claims
Disease & Death Registries	Nationally curated registries for disease incidence and mortality tracking	• National Cancer Registry • Rare & Severe Disease Registry • Cause-of-death data

Private

Public

Most Commonly Used Healthcare Big Data Source in Korea

Category	Description	Representative Examples
Structured EHR Data	Electronic health records used in retrospective cohort studies, often in CDW or CDM format	• Clinical Data Warehouse (CDW) • Common Data Model (CDM) • Hospital-based cohort DBs
Unstructured Health Data	Data from clinical notes, free-text reports, and non-coded documents	• Physician's notes • Discharge summaries • Operative reports
Medical Imaging Data	Radiology and imaging data for diagnosis and AI model training	• CT/MRI/PET scans • PACS (Picture Archiving and Communication System) • Radiology reports
Genetic Data	Genomic epidemiology and hospital-based genome sequencing programs	• Korea Biobank Project • Hospital-based Whole Genome Studies • KoGES Genetic DB
Epidemiology & Behavioral	Data from longitudinal and national health surveys related to lifestyle and behavior	• Korean National Health & Nutrition Examination Survey (KNHANES) • Korean Genome and Epidemiology Study (KoGES) cohort • BioBigData.Korea (Korean PMI chort)
Health Screening	Mass screening data from national and private sources	• National Health Screening Program • National Cancer Screening Program • Corporate health checkups
National Health Insurance	Administrative data for health service eligibility, claims, and reimbursement	• Eligibility and enrollment data • Diagnosis/procedure codes • Medication claims
Disease & Death Registries	Nationally curated registries for disease incidence and mortality tracking	• National Cancer Registry • Rare & Severe Disease Registry • Cause-of-death data

Private

Public

건강보험 빅데이터 연구의 장단점
Strengths and Limitations of Korean NHI Big Data

Strengths	Limitations
전국민 단일보험 + 요양기관 강제지정 Universal coverage & unified insurance system	다른 인구집단에 대한 일반화 한계 Limited generalizability to non-Korean populations
매우 큰 N수: 안정성, 세분성 Very large sample size → stable & granular data	매우 큰 N수 → 통계적 오류 우려 High power → trivial differences become statistically significant
비교적 긴 추적기간 Relatively long follow-up period	후향적·이차자료 연구 Mostly retrospective and secondary data-based research
행위별 수가에 기반한 상세 진료내역 포함 Detailed billing-level clinical information	진료 목적·결과의 부재 Lack of indication and clinical outcomes in claim records
반복된 청구자료 및 검진자료 Repeated claims and screening data over time	청구자료·검진자료의 정확도 한계 Validity and reliability concerns with administrative data

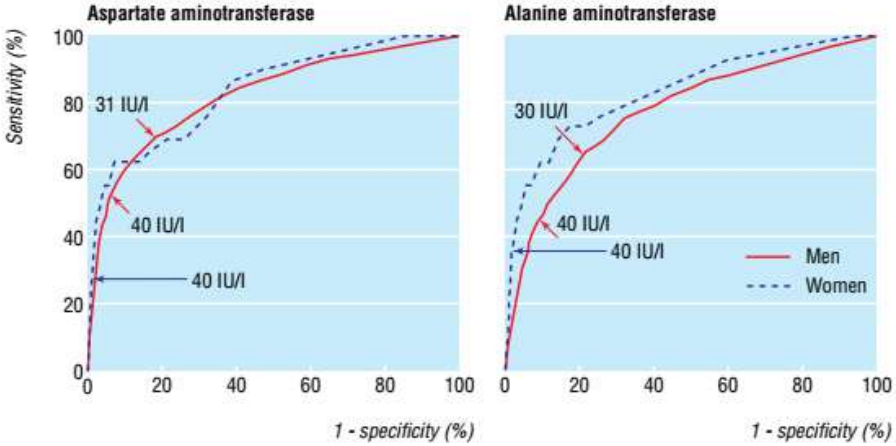
NHI Big-Data Research in 2000s

Cite this article as: BMJ, doi:10.1136/bmj.38050.593634.63 (published 17 March 2004)

Papers

Normal serum aminotransferase concentration and risk of mortality from liver diseases: prospective cohort study

Hyeon Chang Kim, Chung Mo Nam, Sun Ha Jee, Kwang Hyub Han, Dae Kyu Oh, Il Suh



Kim HC, et al. ***BMJ.*** 2004;328(7446):983.

Stroke
Volume 36, Issue 8, 1 August 2005; Pages 1642-1647
https://doi.org/10.1161/01.STR.0000173404.37692.9b

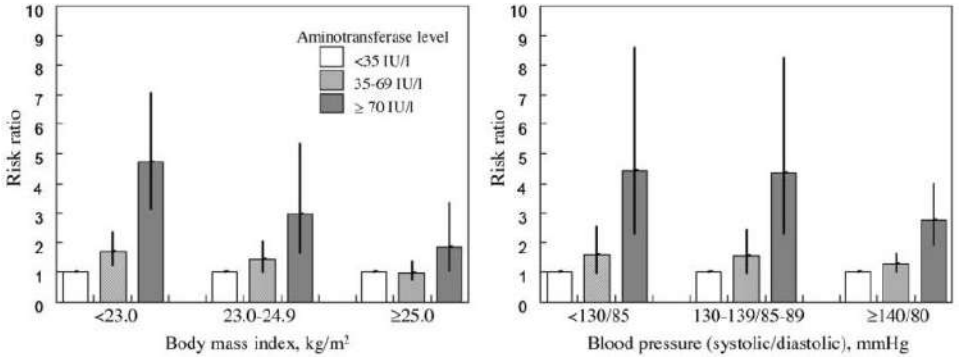
American Heart Association.

ORIGINAL CONTRIBUTIONS

Elevated Serum Aminotransferase Level as a Predictor of Intracerebral Hemorrhage

Korea Medical Insurance Corporation Study

Hyeon Chang Kim, MD, PhD, Dae Ryong Kang, PhD, Chung Mo Nam, PhD, Nam Wook Hur, PhD, Jee Seon Shim, MPH, Sun Ha Jee, PhD, and Il Suh, MD, PhD



Kim HC, et al. ***Stroke.*** 2005;36(8):1642

NHI Big-Data Research in 2010s

epiH
Epidemiology and Health

ORIGINAL ARTICLE

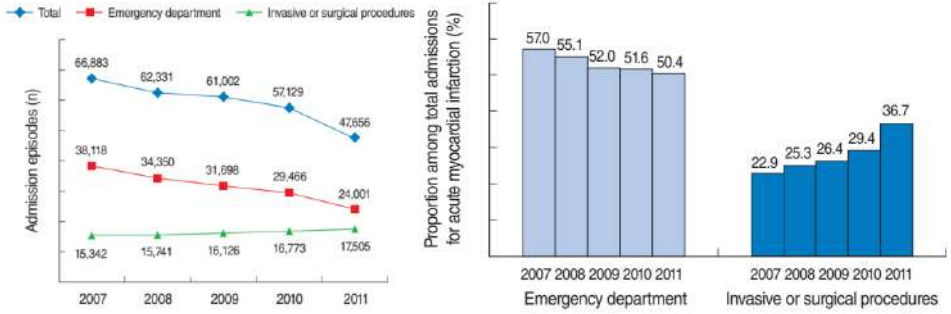
Epidemiology and Health
Volume: 37, Article ID: e2015022, 7 pages
http://dx.doi.org/10.4178/epih/e2015022

Open Access

Admission route and use of invasive procedures during hospitalization for acute myocardial infarction: analysis of 2007-2011 National Health Insurance database

Hyungseon Yeom[1], Dae Ryong Kang[2], Seong Kyung Cho[3], Seung Won Lee[4,6], Dong-Ho Shin[5], Hyeon Chang Kim[1,6]

[1]Department of Preventive Medicine, Yonsei University College of Medicine, Seoul; [2]Research Affairs, Yonsei University College of Medicine, Seoul; [3]Division of Environmental Health, Environmental Policy Research Group, Seoul; [4]Department of Public Health, Yonsei University Graduate School, Seoul; [5]Severance Cardiovascular Hospital, Yonsei University College of Medicine, Seoul; [6]Cardiovascular and Metabolic Diseases Etiology Research Center, Yonsei University College of Medicine, Seoul, Korea



Yeom H, et al. ***Epidemiol Health.*** 2015;37:e2015022

DIABETIC Medicine

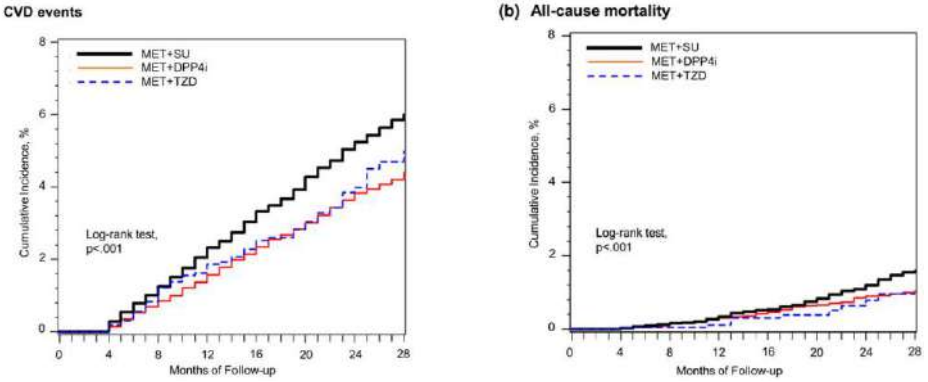
DiABETES UK
KNOW DIABETES. FIGHT DIABETES.

Research: Complications

Cardiovascular events associated with second-line anti-diabetes treatments: analysis of real-world Korean data

K. H. Ha, B. Kim, H. Choi, D. J. Kim, H. C. Kim

First published: 19 May 2017 | https://doi.org/10.1111/dme.13384 | Citations: 11



Ha KH, et al. ***Diabet Med.*** 2017;34(9):1235

NHIS-2016-1-007

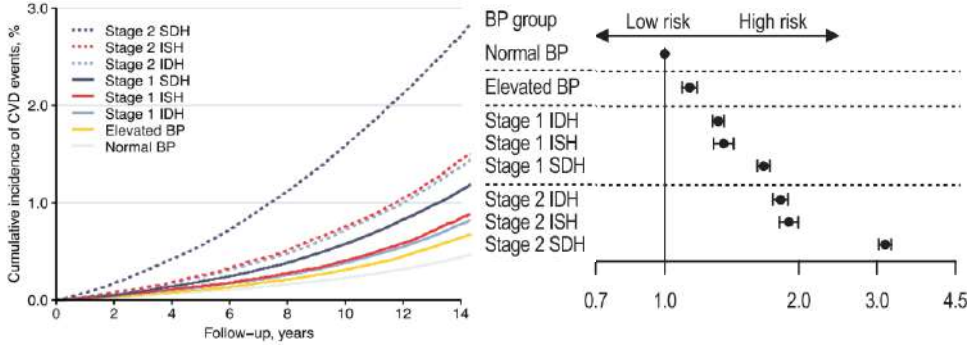
NHI Big-Data Research in 2020s

Circulation
Volume 141, Issue 22, 2 June 2020; Pages 1778-1786
https://doi.org/10.1161/CIRCULATIONAHA.119.044838

ORIGINAL RESEARCH ARTICLE

Cardiovascular Risk of Isolated Systolic or Diastolic Hypertension in Young Adults

Hokyou Lee, MD, Yuichiro Yano, MD, PhD, So Mi Jemma Cho, BA, Jong Heon Park, MD, PhD, Sungha Park, MD, PhD, Donald M. Lloyd-Jones, MD, ScM, and Hyeon Chang Kim, MD, PhD



Lee H, et al. ***Circulation.*** 2020;141(22):1778

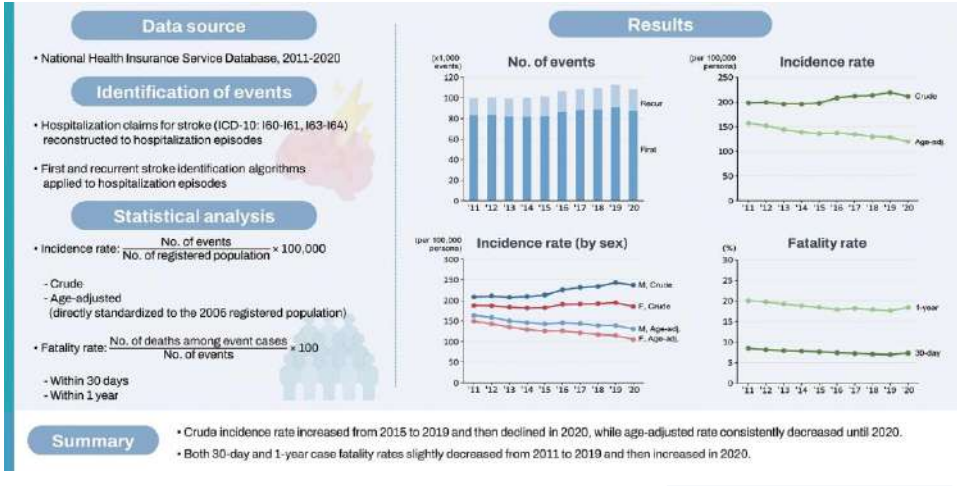
NHIS-2020-1-015

Open Access
Volume: 46, Article ID: e2024003, 9 pages
https://doi.org/10.4178/epih.e2024003

SPECIAL ARTICLE

Incidence and case fatality of stroke in Korea, 2011-2020

Jenny Moon[1*], Yeeun Seo[1*], Hyeok-Hee Lee[2,3,4], Hokyou Lee[2,3,4], Fumie Kaneko[1], Sojung Shin[1], Eunji Kim[2], Kyu Sun Yum[5], Young Dae Kim[6], Jang-Hyun Baek[7], Hyeon Chang Kim[2,4]



Moon J, et al. ***Epidemiol Health.*** 2024;46:e2024003

NHIS-2022-1-679

건강보험 빅데이터 활용 연구의 대표 사례
Typical Applications of Korean NHI Big Data

Until Now	From Now
청구자료 분석: 질병 유병, 의료이용 규모 변화 **Claims:** Disease prevalence, healthcare utilization trends	정밀의료, Exposome 연구의 핵심 자료로 활용 Use in precision medicine and exposome-based research
검진자료 + 청구자료: 주요 위험요인과 질병의 관련성 **Screening + Claims:** Major risk factors and disease risk	질병 예측을 넘어, 건강행태·의료이용 예측으로 확장 From disease prediction to behavior & service utilization forecasting
청구자료 + 공공자료: 환경 위해요인 노출과 건강위해성 **Claims + Public data:** Environmental exposure and health outcomes	새로운 건강이슈의 조기 탐지 Early detection of emerging health issues
청구자료 + 민간자료: 기존 코호트에서 질병 위험도 평가 **Claims + Private data:** Disease risk prediction in existing cohorts	**AI 학습용 데이터 및 헬스 AI 에이전트 개발에 활용** **Training AI models and developing healthcare AI agents**

Major Sources for Public Health Big-Data

건강보험심사평가원
HEALTH INSURANCE REVIEW & ASSESSMENT SERVICE
보건의료빅데이터개방시스템

h-well 국민건강보험
소통, 배려 국민과함께
건강보험 빅데이터 플랫폼
NHIS-Big Data Platform

국립보건연구원
National Institute of Health
보건의료연구자원정보센터
CODA

한국보건의료정보원
Korea Health Information Service

국립암센터
NATIONAL CANCER CENTER

한국인체자원은행사업
KBN 한국인체자원은행네트워크
Korea Biobank Network

Healthcare Data Center Hospitals

실제진료데이터의 전략적 가치
Real-World Data (RWD): strategic importance in the AI era

임상시험 한계를 보완하는 현실 기반 데이터
Provides real-world evidence beyond the limits of clinical trials

AI 기반 예측·조기진단·맞춤형 정책 설계의 핵심 자원
Enables AI-driven prediction, early detection, and personalized policy design

청구·EMR·검진 등 이기종 데이터 통합으로 모델 성능 향상
Boosts model performance by integrating claims, EMR, and screening data

AI 학습·검증에 필요한 다양하고 반복적인 시계열 데이터 제공
Offers diverse and longitudinal data essential for AI training and validation

의료기관은 RWD의 생산자이자 해석자, 그리고 책임 있는 활용 주체
Hospitals are not only producers but also interpreters and responsible users of RWD

보건의료 빅데이터 및 AI 활용의 주요 장애물
Barriers to Using Healthcare Big Data and AI

기관 간 데이터 파편화 및 표준화 부족
Data fragmentation and lack of standardization across institutions

EMR 및 실제 진료데이터의 상호운용성 미흡
Limited interoperability of EMR systems and real-world data

공공데이터 접근 절차의 복잡성과 제한
Complex and restrictive access procedures for national data

개인정보보호, 윤리, 법적 책임에 대한 우려
Concerns about data privacy, ethics, and legal liability

돌파구와 진전 사례
Breakthroughs and Progress

건강보험공단 및 심사평가원의 폐쇄분석실과 클라우드 기반 분석 플랫폼 운영
Operation of secure on-site data labs and cloud-based analysis platforms by NHIS and HIRA

보건의료 빅데이터 통합 플랫폼 구축 및 운영
Korea Health Big Data Integration Platform: inter-agency data linkage among public institutions

데이터 중심병원 프로젝트: 7개 컨소시엄, 43개 병원 참여
Data-Center Hospital Project involving 7 consortia and 43 participating hospitals

K-CURE 프로젝트: 5개 공공기관과 18개 병원의 협력 체계 구축
K-CURE Project: Joint initiative by 5 public agencies and 18 hospitals

민간기업과 병원 간 CDW 기반 공동 분석 협력체 확산
CDW-based collaborative analytics consortia between private companies and hospitals

의료기관의 변화된 역할
The Evolving Role of Medical Institutions

단순한 데이터 제공자에서 혁신 공동 창출자로
From data providers to co-creators of innovation

임상 맥락, 전문가 통찰, 윤리적 감시 제공
Provide clinical context, expert insight, and ethical oversight

현장 기반 모델 검증과 시범 적용의 중심지 역할
Host real-world model validation and pilot implementation

신뢰 가능하고 설명 가능한 AI 구축의 핵심 주체
Critical to building trustworthy, explainable AI

이러한 역할을 수행할 수 있는 인적 역량, 물적 인프라, 임상 경험, 책임의식을 갖춘 기관
*Medical institutions (teaching hospitals) in Korea are equipped with the **expertise, infrastructure, clinical experience, and a strong sense of responsibility** to lead these roles*

결론 및 정책 제언
Conclusion and Policy Suggestions

데이터 거버넌스에서 개방성과 책임성의 균형 필요
Balance openness and accountability in data governance

다양한 데이터 유형과 접근 방식의 병행 필요
Promote diversity in data types and access mechanisms

의료기관 등 민간 주체의 역량을 강화하고 권한을 부여해야 함
Empower private actors, especially hospitals, with greater roles and authority

정부·공공기관의 지원을 넘어 민간과 의료기관의 자율적 역할 확대
Shift from government-centered support to enabling autonomous leadership by private actors

공공과 민간이 공동 책임을 지고 혁신을 주도하는 파트너십
Foster partnerships where both public and private sectors share responsibility and lead innovation together

Thank you for your attention.