NHIS GLOBAL FORUM 2025

Bridging Innovation and Policy:

Leveraging Emerging Data for Healthcare Decisions

Hyun-Young Park, MD, PhD

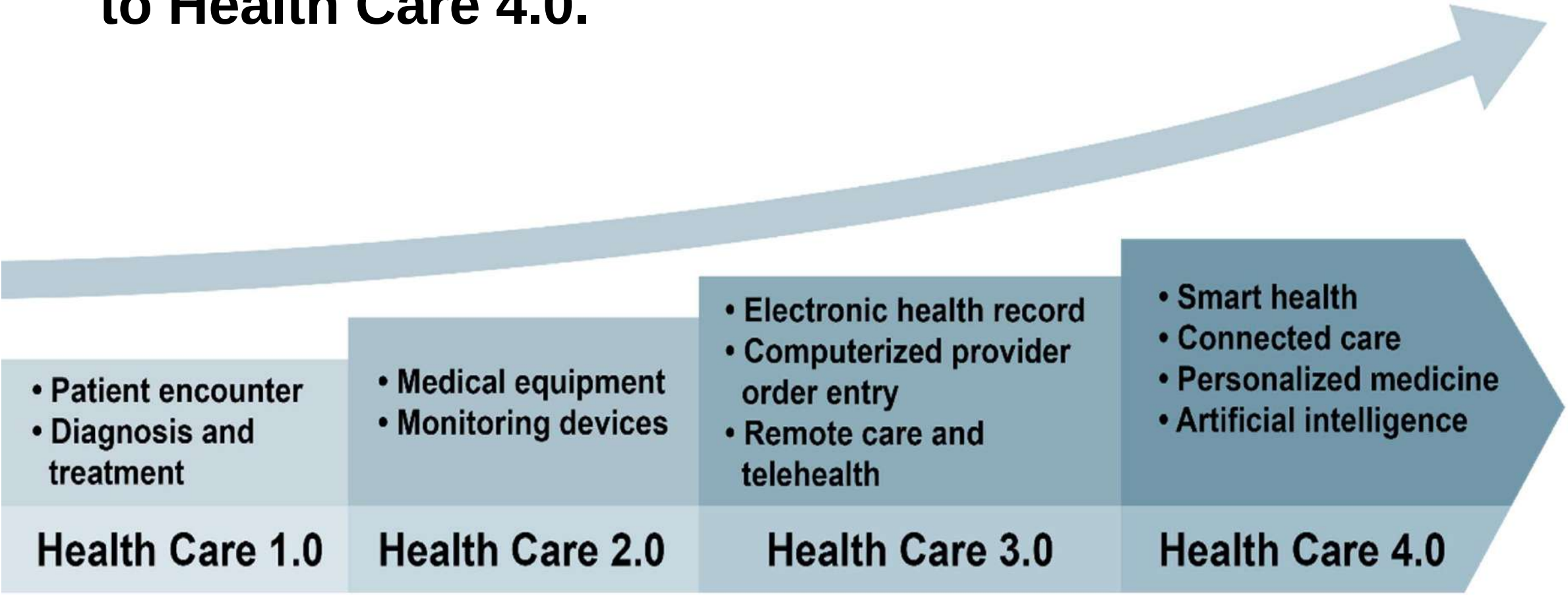
National Institute of Health, Republic of Korea

National Institute of Health
Republic of Korea

Contents

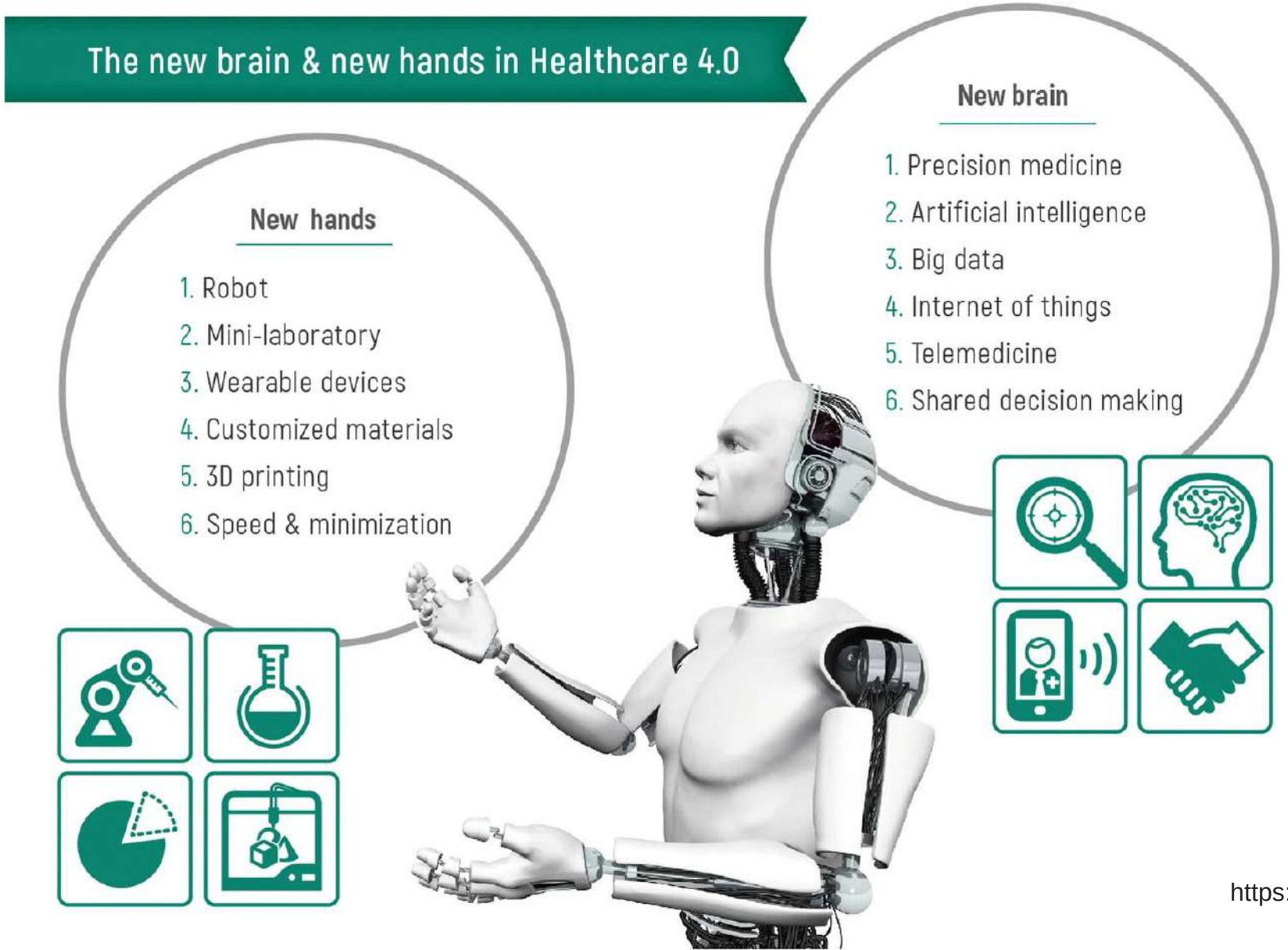
- Healthcare big data
- Role of healthcare big data for outbreak control
- NHIS and EMR data integration
- Genomic data for precision medicine
- NIH’s Bio Big Data Initiative

Historical evolution of health care 1.0 to Health Care 4.0.



https://doi.org/10.1080/24725579.2021.1884627

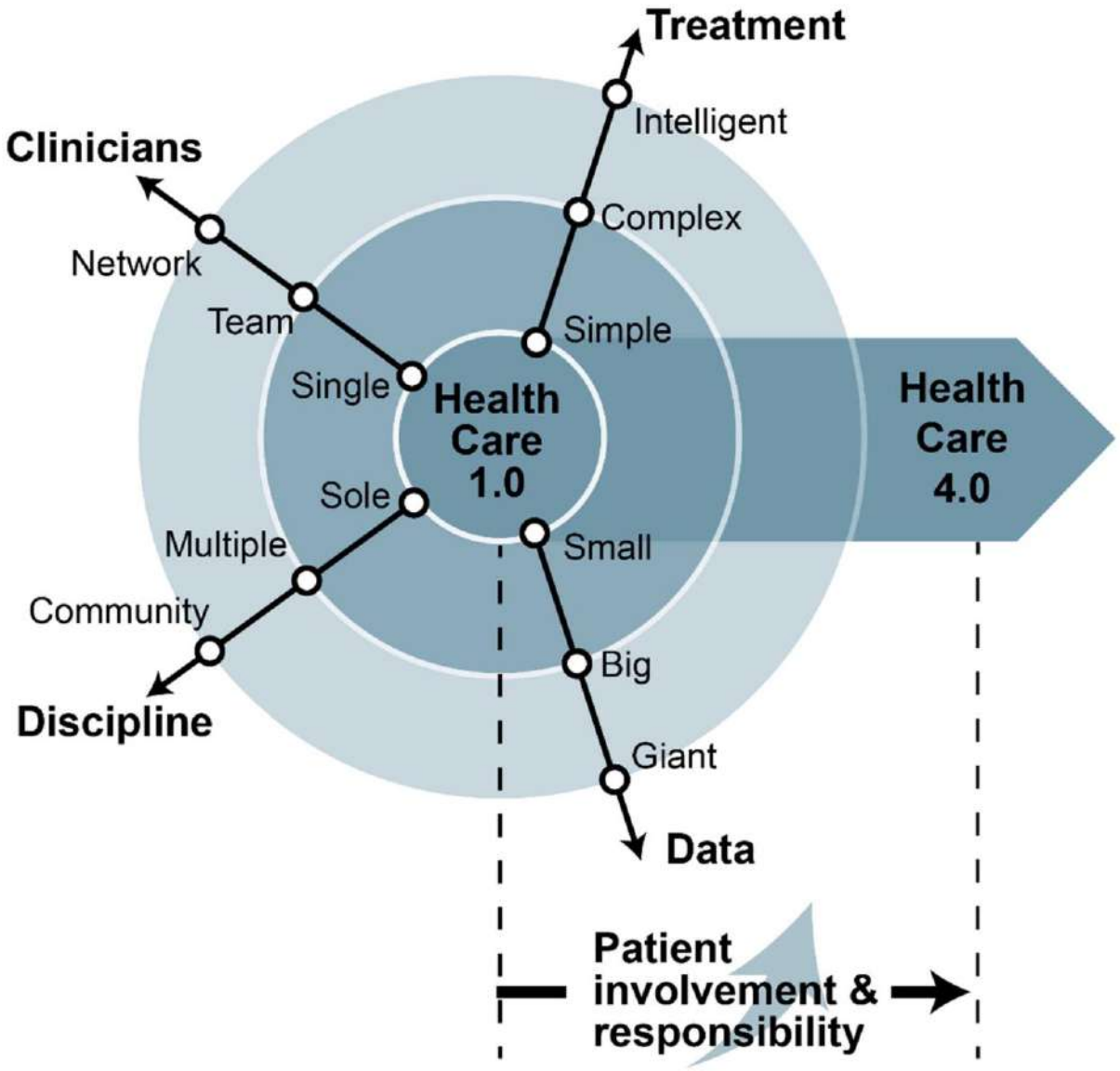
Healthcare 4.0, *Smart Medicine*



https://doi.org/10.1007/s10916-019-1513-0

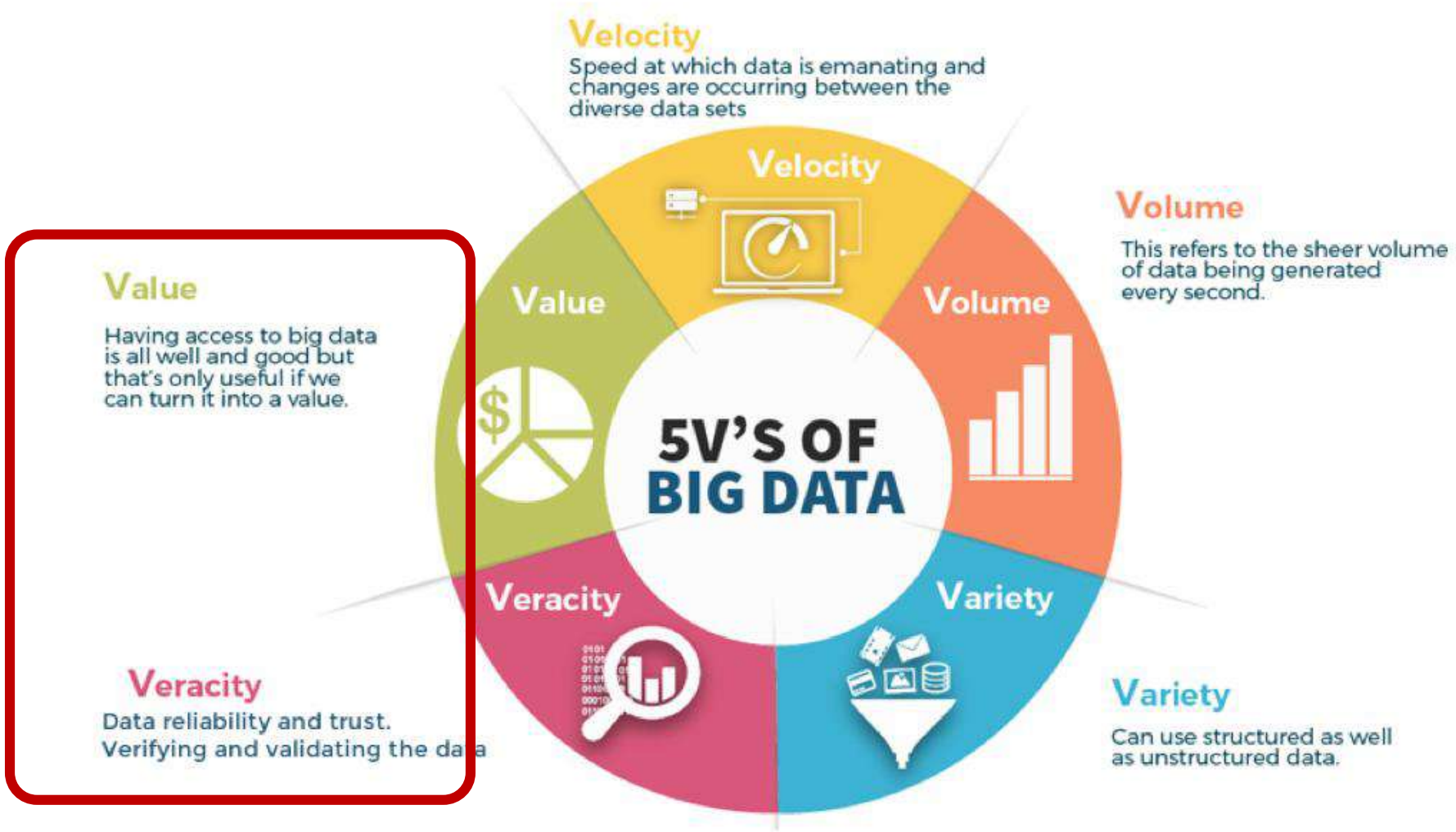
Characteristics

- ***Big and giant data*** streams with large variations in dimension, quality, format, and characteristic
- Patients and clinicians are ***increasingly involved and share responsibilities*** for monitoring their health, reporting symptoms, and participating in shared decision making for treatment and care planning



https://doi.org/10.1080/24725579.2021.1884627

Big Data in Healthcare



https://medium.com/nerd-for-tech/what-is-big-data-6c47407bfbe5

Current Healthcare Big Data

- Electronic Health Records (EHRs): Diagnoses, lab results, prescriptions, vital signs, clinical notes
- Medical Imaging Data
- Claims and Administrative Data: Insurance billing, procedures, medications, healthcare utilization
- Health Screening
- Social Determinants of Health (SDOH):Income, education, employment, environment

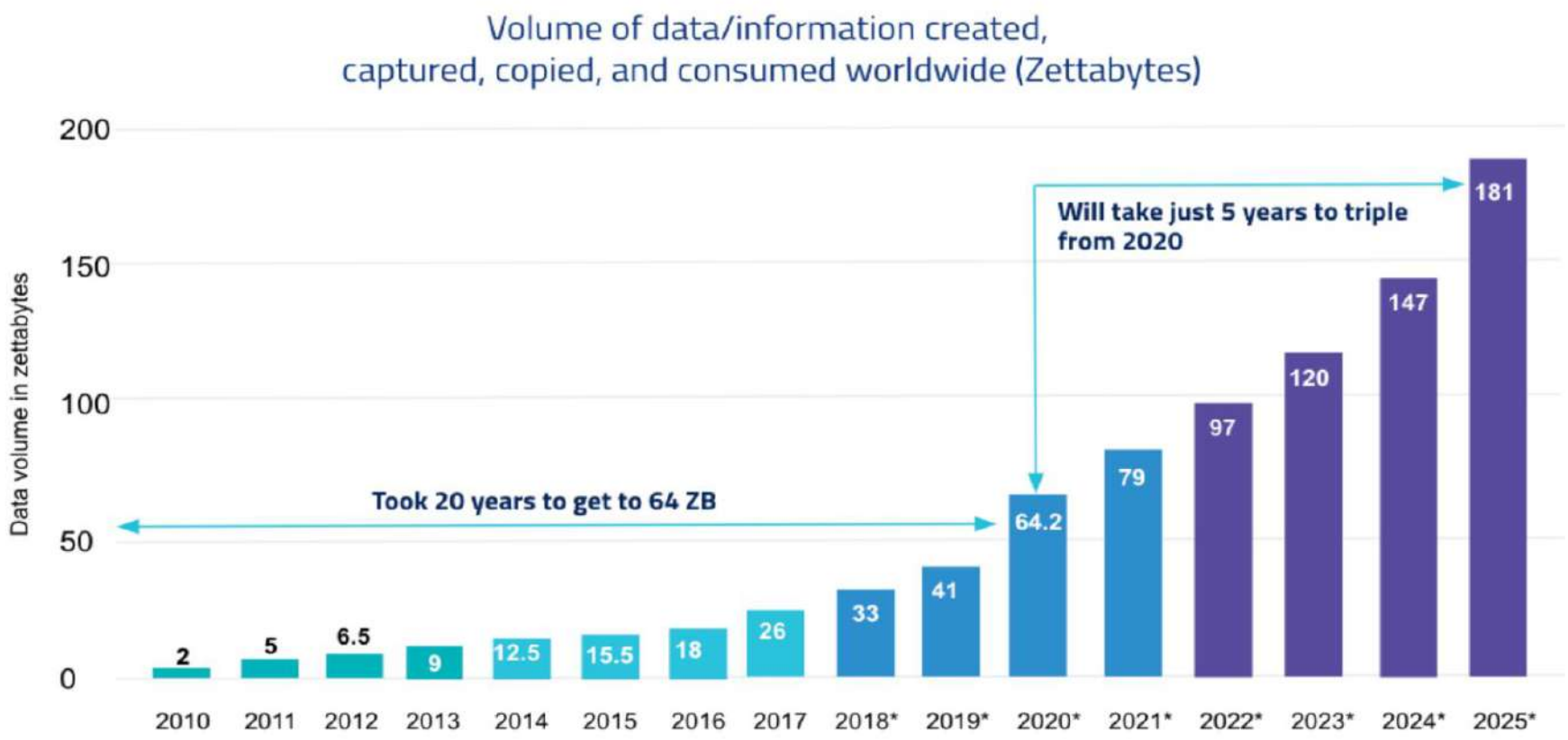
What is Emerging Healthcare Data?

- Genomic data (e.g., whole-genome sequencing)
- Multi-omics data generated by research
- Real-world evidence (RWE) from clinical practice and patient registries
- Social and behavioral health data
- Wearable/IoT health data
- Patient-Reported Outcomes (PROs)
- Artificial intelligence-generated insights

Big Data Grows Ever Bigger

1 ZETTABYTE =
1 000 000 000
000 000 000 000
BYTES

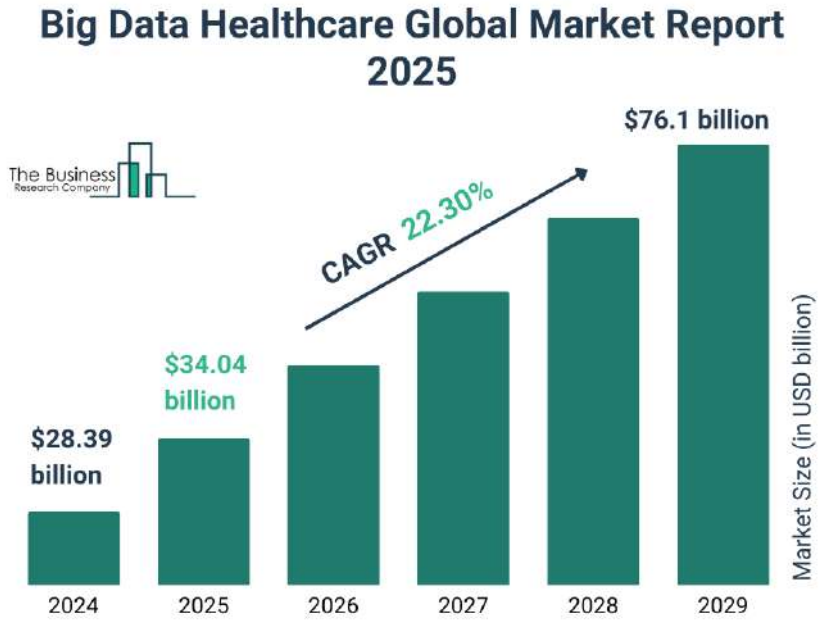
ComputerHope.com

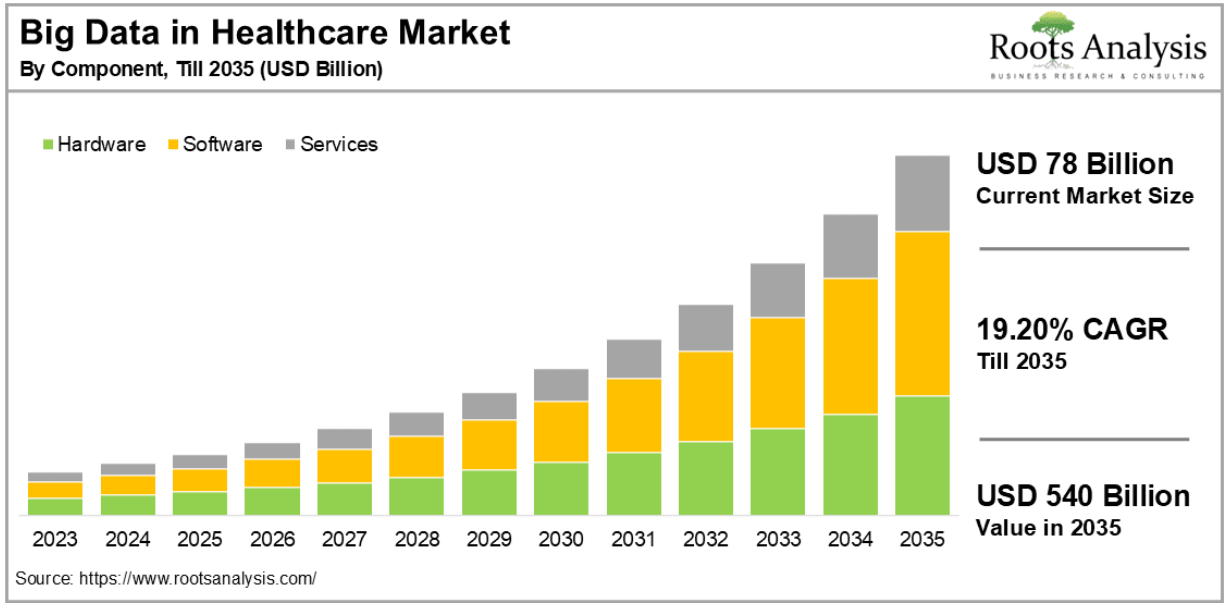


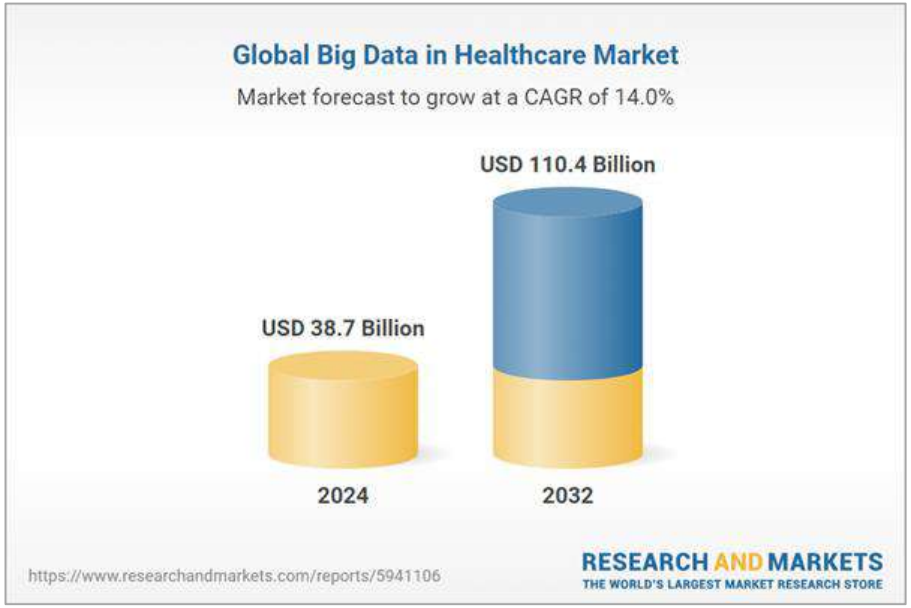
35%
of all data will be life sciences + healthcare by 2025

https://www.signalfire.com/blog/healthcare-gpt

The Era Where Data Becomes Economic Value



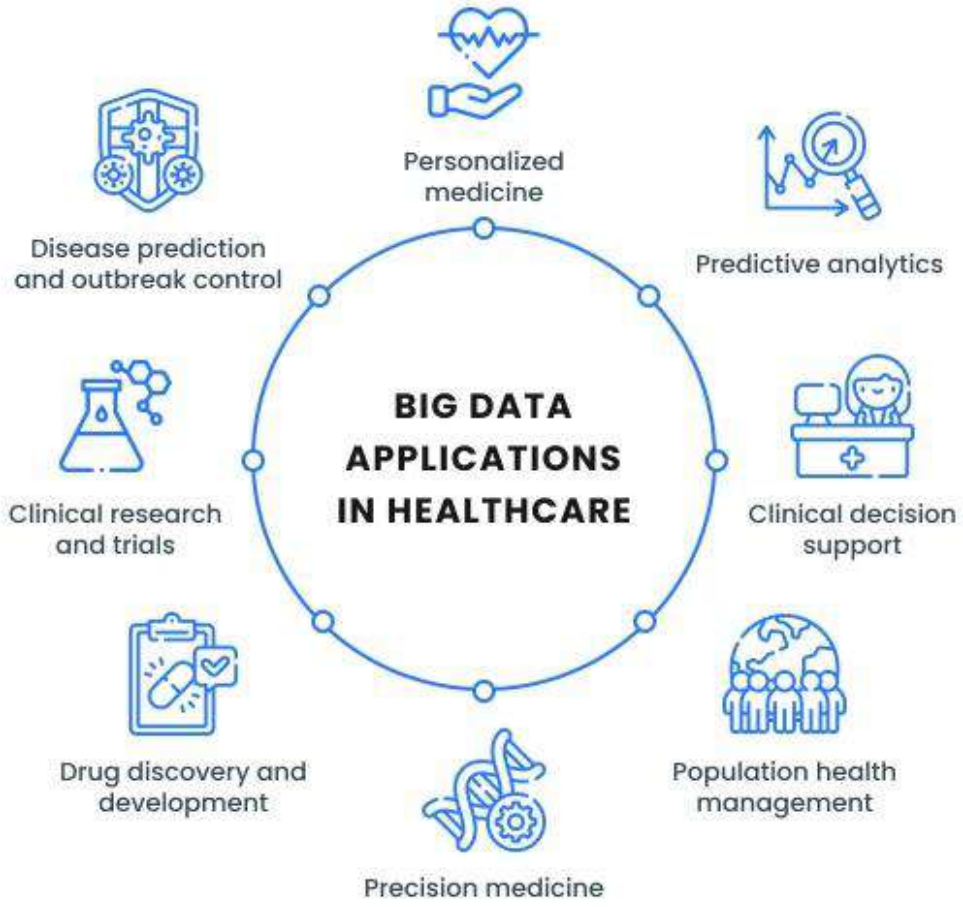




Healthcare Big Data

Opportunities and Impact

- **Clinical decision support**
- **Precision medicine**: Tailored therapies based on genomic and lifestyle data
- **Population health management**: Predictive analytics for early intervention
- **Policy innovation**: Data-driven design of reimbursement, screening, and public health programs
- **Drug discovery and clinical trials**: Accelerated recruitment and targeted trials



https://binariks.com/blog/big-data-applications-in-healthcare/

Role of Healthcare Big Data for Outbreak Control

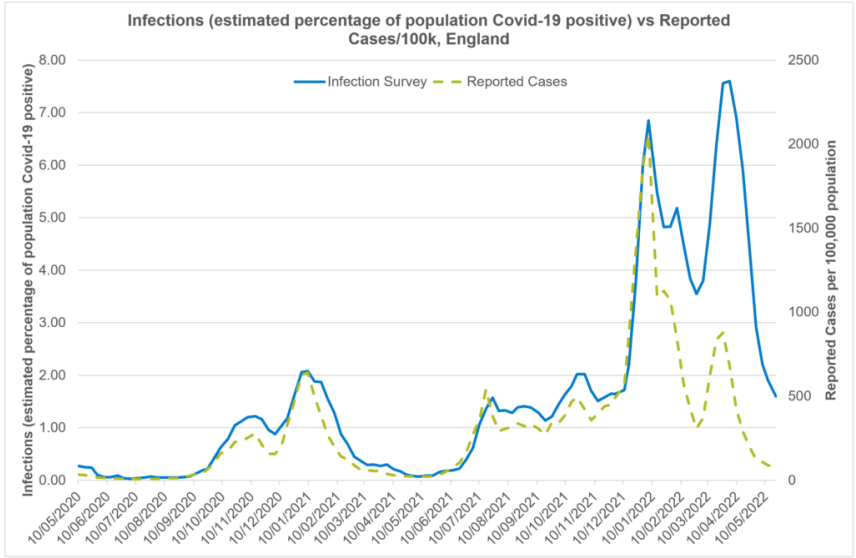
COVID-19 Infection Survey (ONS)

National Health Service (NHS) and Office for National Statistics (ONS) to monitor infections, hospitalizations, and deaths

- Regular testing of random household samples, including asymptomatic individuals
- Used big data analytics to estimate:National and regional infection rates, Infection prevalence by age group, region, and time

Impact:

- Provided real-time community transmission insights
- Guided local lockdown decisions and vaccine rollout
- One of the most accurate surveillance tools globally



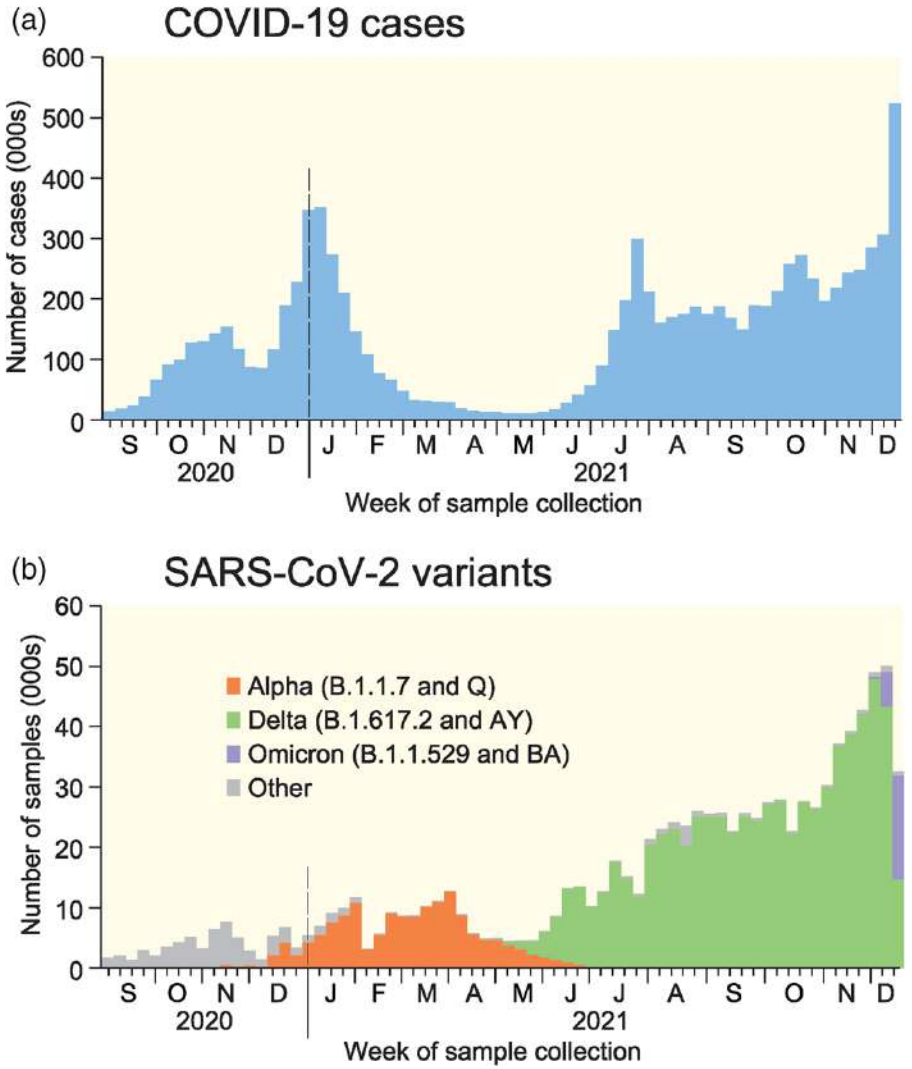
COVID-19 GENOMICS UK CONSORTIUM

COVID-19 Genomics UK (COG-UK) Consortium established in March 2020

- Sequenced over 2 million SARS-CoV-2 genomes
- Linked viral genome data with patient demographics and clinical outcomes

Impact:

- Detected and tracked variants of concern (e.g., Alpha, Delta, Omicron)
- Informed public health strategies and vaccine updates
- A model for integrated pathogen surveillance globally



OpenSAFELY

Supported by NHS England

- Secure, open-source analytics platform created in March 2020
- Designed to enable secure, large-scale analysis of electronic health records (EHRs) for COVID-19 research while maintaining strict privacy protection
- Data: Primary care records (diagnoses, medications, demographics), Linked to hospital data, death registrations, and COVID-19 test results
- Impact
    1) COVID-19 Risk Factors identification
    2) Vaccine Safety and Effectiveness: Monitored vaccine uptake and safety in real time, Evaluated rare adverse events
    3) Health Inequalities: disparities in COVID-19 mortality among ethnic minorities, Access to care and shielding recommendations, Geographic variation in vaccine uptake

Big Data and Public Health Crisis

Integrated big data and genomics can turn a reactive public health response into a proactive and precision-driven strategy.

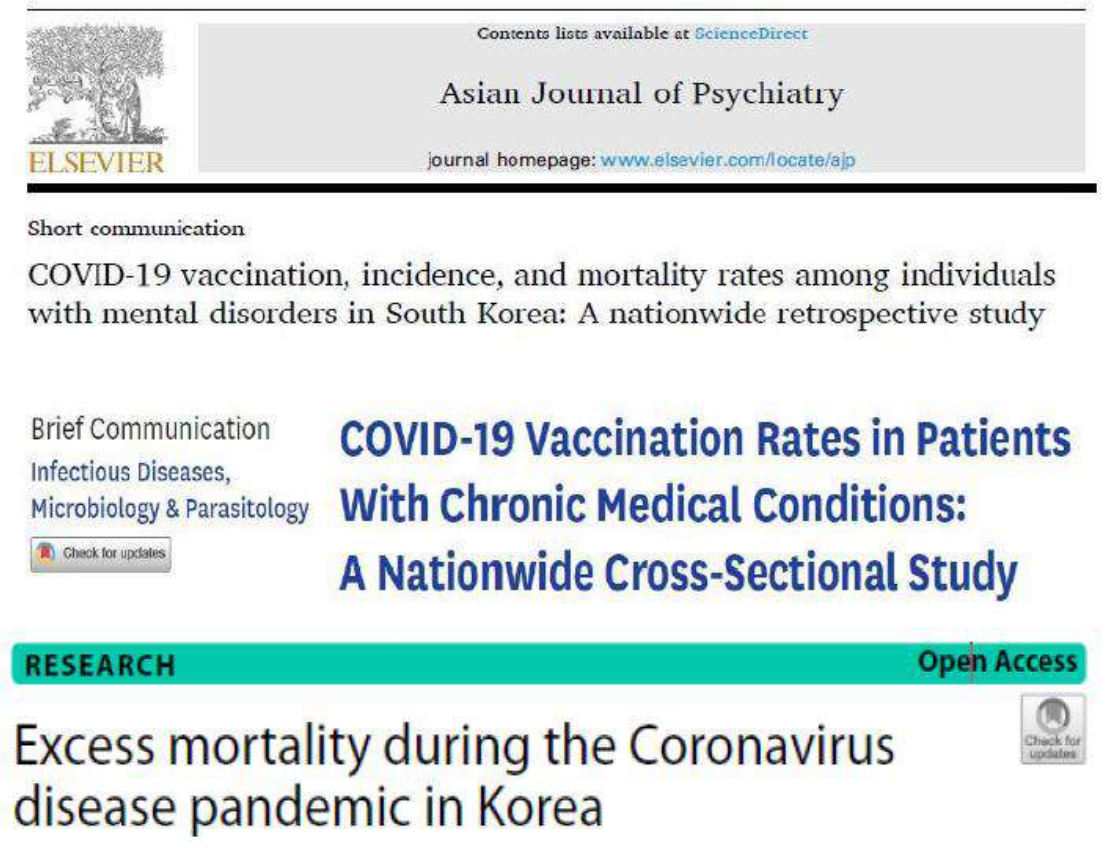
KDCA Covid-19 NHIS cohort

KDCA Covid-19 NHIS cohort

Data sharing

K-COV-N sharing website (http://nhiss.nhis.or.kr)

Key achievements using K-COV-N



- 224 studies were approved, 36 papers published (April 2022~2025)

NHIS Data

Korea’s National Health Insurance Service (NHIS) data serves as a representative example of how large-scale health data can inform and shape healthcare policy decisions.

Overview of NHIS Data

- Operated by the **National Health Insurance Service**
- Covers **over 98%** of the Korean population
- Contains longitudinal data: eligibility, healthcare utilization, prescriptions, health screening, death
- Linked to other datasets: national cancer registry, cause-of-death registry, etc.

Example

Article

Effectiveness of the Korean National Cancer Screening Program in Reducing Colorectal Cancer Mortality

Hyeon Ji Lee [1], Kyeongmin Lee [2], Byung Chang Kim [3], Jae Kwan Jun [1,2], Kui Son Choi [1,2] and Mina Suh [1,2,*]

Cancers 2024, 16, 4278.

- Objective: To assess the impact of the Korean National Cancer Screening Program (KNCSP) on colorectal cancer (CRC) mortality.
- Methodology: A nested case-control study utilizing cohort data from the KNCSP, encompassing 5,944,540 individuals aged ≥50 years as of 2004. The study linked this data with the Korea Central Cancer Registry (KCCR) and death certificate data from Statistics Korea.
- Findings: Individuals who underwent CRC screening using the fecal immunochemical test (FIT) had a **26% lower risk of CRC-specific mortality** compared to those who were never screened (Odds Ratio: 0.74; 95% Confidence Interval: 0.71–0.76). The reduction in mortality was more pronounced with increased frequency of screening.

NHIS Database

Pros and Cons

Strengths

Nationally representative

Longitudinal, large-scale

Real-world utilization and claims data

Useful for RWE and health policy evaluation

Limitations

Limited clinical detail

Potential for miscoding

No patient-reported outcomes or genomics

Access constraints and time lag

*NHIS data is a **powerful tool for public health research** and policymaking, but should be **complemented** with clinical, genomic, and patient-reported data*

Key Obstacles in Linking Healthcare Big Data in Korea

Lack of Interoperability

- Heterogeneous data formats
- Limited adoption of international standards

Data Ownership and Governance Issues

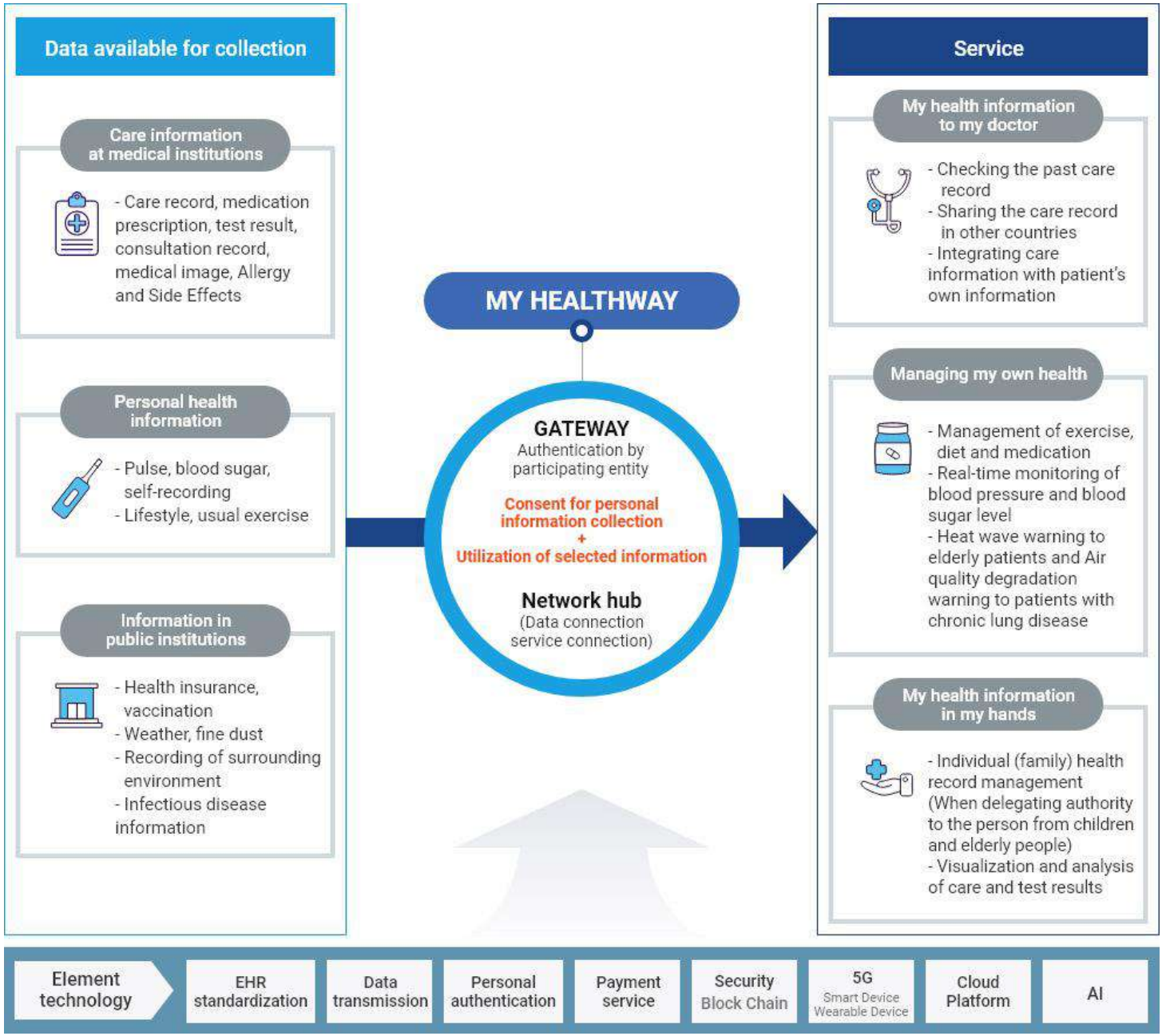
- Unclear data ownership between healthcare providers, public institutions, and patients obstructs coordinated data use.
- Absence of a unified governance framework for managing cross-sector data access and accountability.

Privacy and Public Trust Concerns

- High public sensitivity regarding health data misuse or commercial exploitation undermines support for industry-driven uses.
- Lack of **trusted intermediary institutions** to manage anonymization, consent, and ethical oversight weakens social acceptance.

(재)한국보건의료정보원
Korea Health Information Service

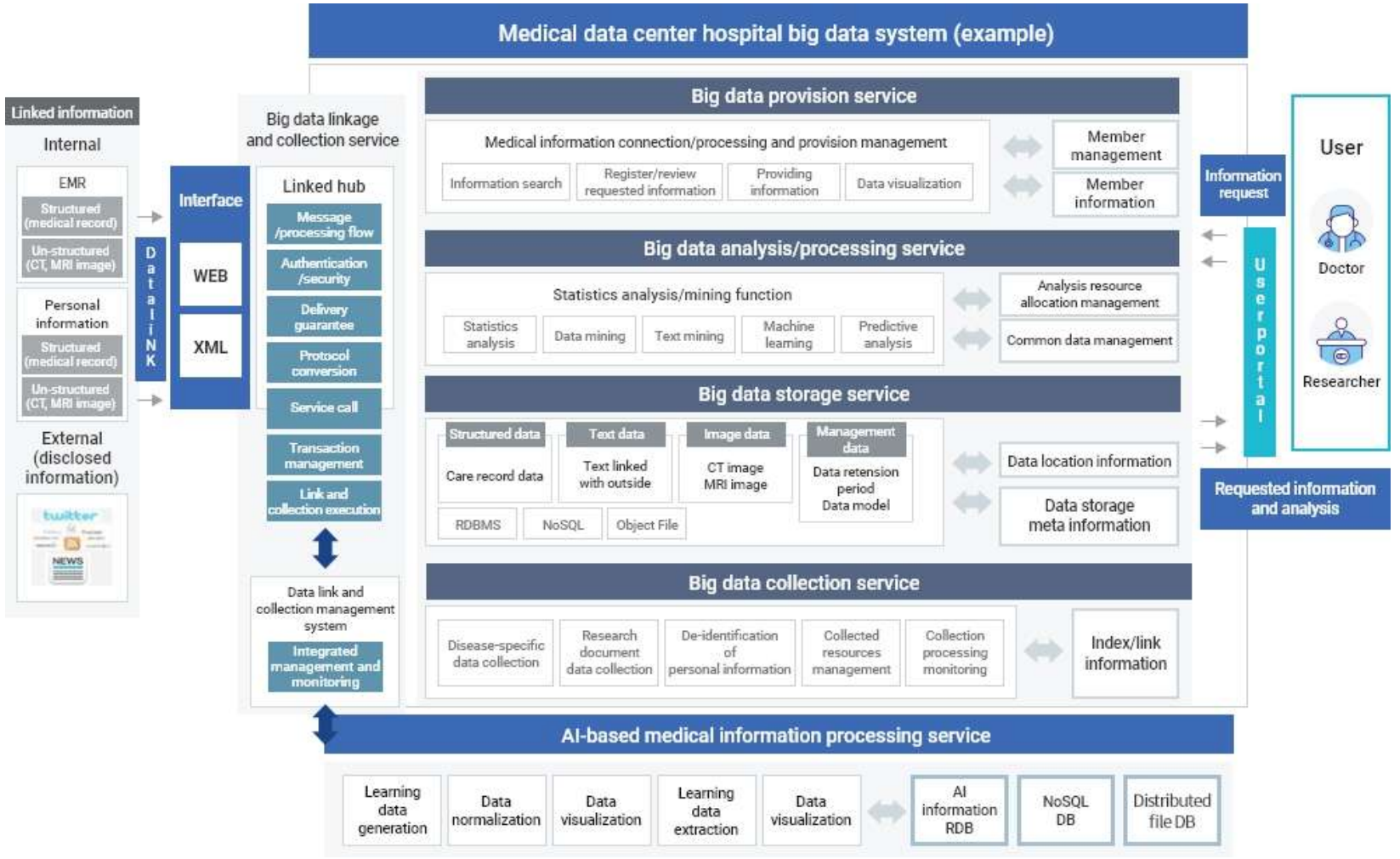
My Healthway



Healthcare data center hospitals

Promoting research on utilization of medical information

- Definition and construction of medical data set for promotion of submitted application scenario research
- Contributing to the development of new medical technologies and promoting research for the development of new drugs, medical devices, and AI by using medical data



Healthcare data center hospitals

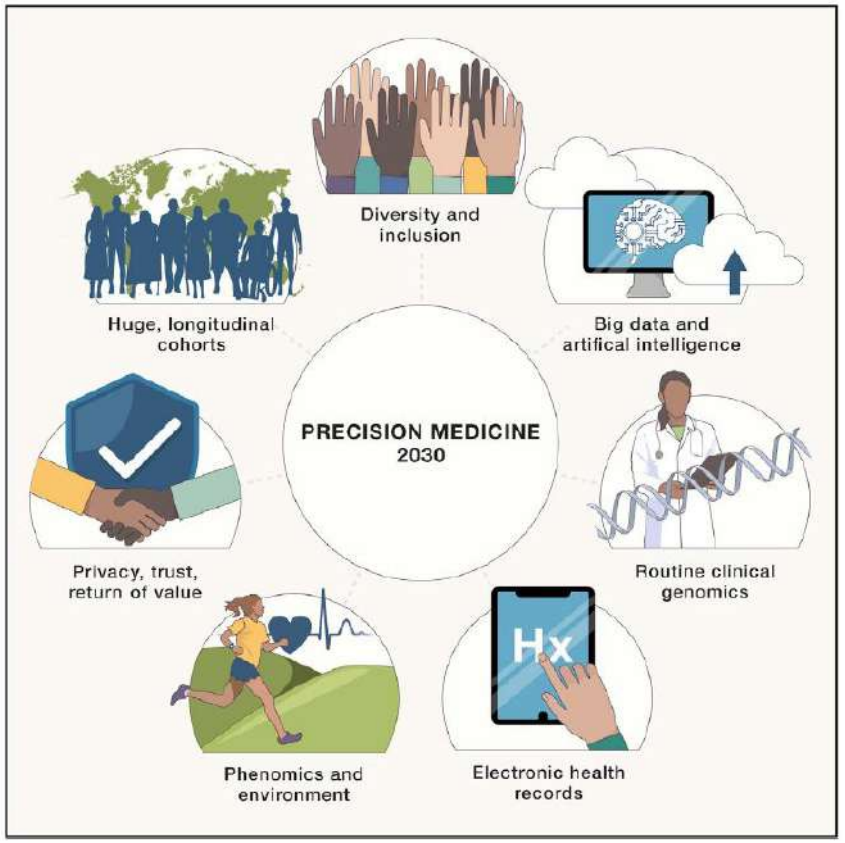
(재)한국보건의료정보원
Korea Health Information Service

Genomic Data
for Precision Medicine

Importance of Genomic Data

- **Precision medicine**: personalized drug selection, pharmacogenomics.
- **Rare diseases**: faster and more accurate diagnoses.
- **Cancer genomics**: tumor profiling, immunotherapy targeting.
- **Public health**: genomic surveillance, population-level risk stratification.

Seven opportunities for precision medicine by 2030

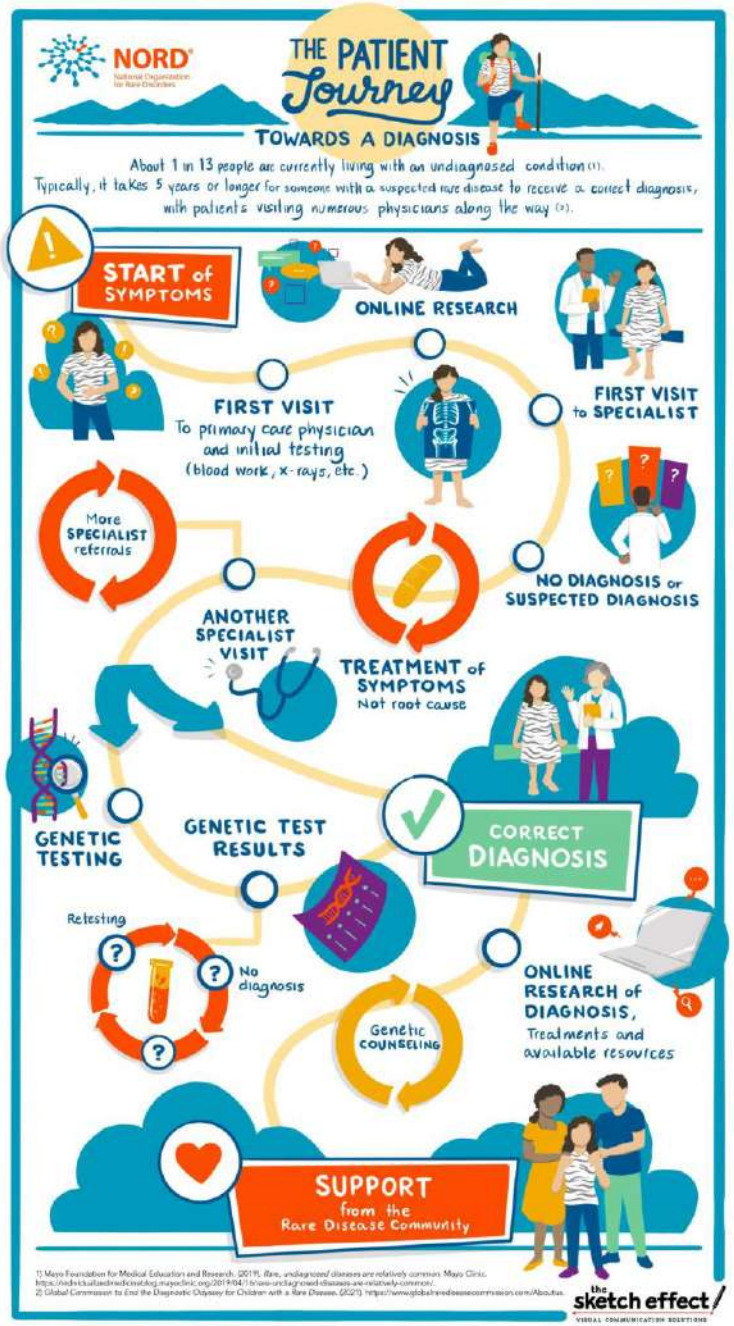


https://doi.org/10.1016/j.cell.2021.01.015

Table 1. Envisioning how precision medicine will affect clinical medicine and research in the next decade

	Where we are today	Where we will be in 2030
Clinical applications		
Genomics for disease	Primarily limited to rare disease and select cancers.	Genomics is routine. Genetic causes and targeted therapies are discovered for many "common" diseases. Microbiome measures are routinely included.
Pharmacogenomics (PGx)	Common in cancer and within select applications of older medications at select sites.	Genome-aware EHRs make PGx easy and automatically update rules from central guidelines. New PGx associations discovered from clinical data.
Genomics for healthy individuals	In research, whole-genome sequencing and search for mutations in one of the ACMG59 genes, present in about 3% of people. Variant interpretation is hard.	ACMG59 grows to > 200, variant interpretation improved by huge, diverse sequenced populations. Cell-free DNA becomes a mainstay of cancer screening
EHRs	Episodic capture from healthcare without robust genomics support. EHR data is essentially not portable.	Genome- and device- enabled. Data can be easily moved between EHRs and to participant apps.
Environmental influences on health	Patient-reported habits and exposures	Geocode-based exposure linkage Real time monitoring of multiple environmental exposures Precision nutrition
Wearable sensors	Ad hoc use of activity monitors	Continuous monitoring of physical activity, sleep, metabolic parameters
Research applications		
Population demographics	>80% European ancestry	>50% non-European ancestry
Routinely available data	Surveys of health conditions, lifestyle, behavior, and diet. GWAS data, lab assays, structured EHR data, and geocoded exposure linkages.	Whole genomes, lab assays, surveys, full EHRs, environmental, genomic and sensor data. Includes imaging, narrative, geocoded, and continuous monitoring approaches to clinical care, activity, precision nutrition, and environment.
Size of cohorts used in analysis	Up to 500K, data downloaded and manually harmonized to sets of several million	>100M using cloud-based federated analyses facilitated by common standards
Largest genomic studies performed on a trait	>1M (GWAS)	>50M (GWAS) >2M (WGS)
Cost of a whole genome	$500	$20*

*Sequencing costs have often fallen faster than Moore's law. Using Moore's law, sequencing costs would be 1/32 of US $500, or $15.63.



Rare Diseases

Diagnostic Odyssey

- Multiple specialists
- Diagnostic uncertainty
- Impact on patient quality of life
- Financial burden

Rare Disease Diagnosis in Korea

- Patients undergo multiple **single-gene or small panel tests**
- Each test targets specific genes, often requiring **dozens to hundreds** of sequential tests
- Covered by National Health Insurance (NHI),
  Patient out-of-pocket cost: ~10%
  For panel test, patient have to pay up to 80%
- Access to WGS is mainly available through **government-supported research or diagnostic aid programs**.

Diagnostic Yield of Genetic Test

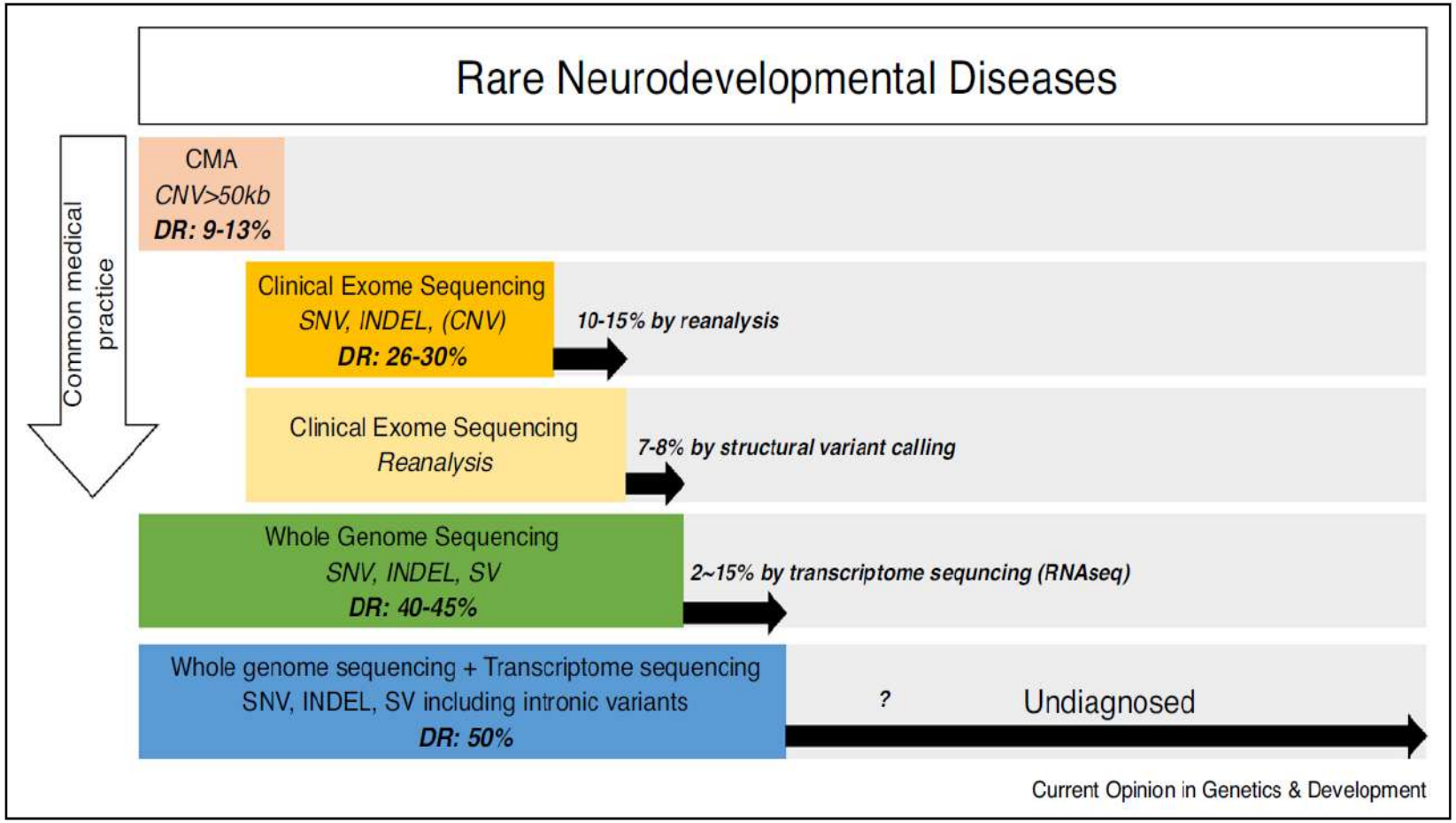
- **Traditional genetic testing (e.g., microarrays, gene panels):**
  Diagnostic yield: **10–25%**

- **Whole Exome Sequencing (WES):**
  Diagnostic yield: **25–40%**

- **Whole Genome Sequencing (WGS):**
  Diagnostic yield: **40–60%**

Diagnostic Yield of Genetic Test

Genomic Tests for Diagnosing Rare NDD



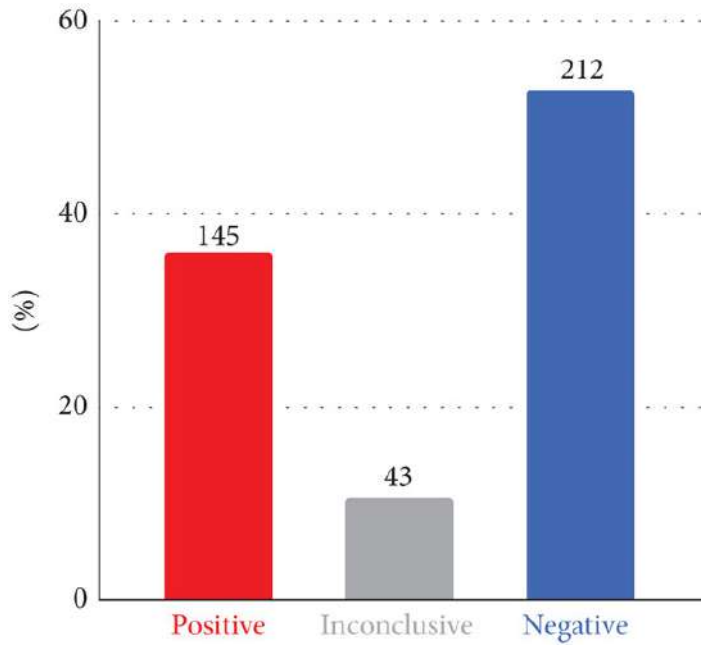
2021_CurOpiGenDev_The frontiers of sequencing in undiagnosed neurodevelopmental diseases

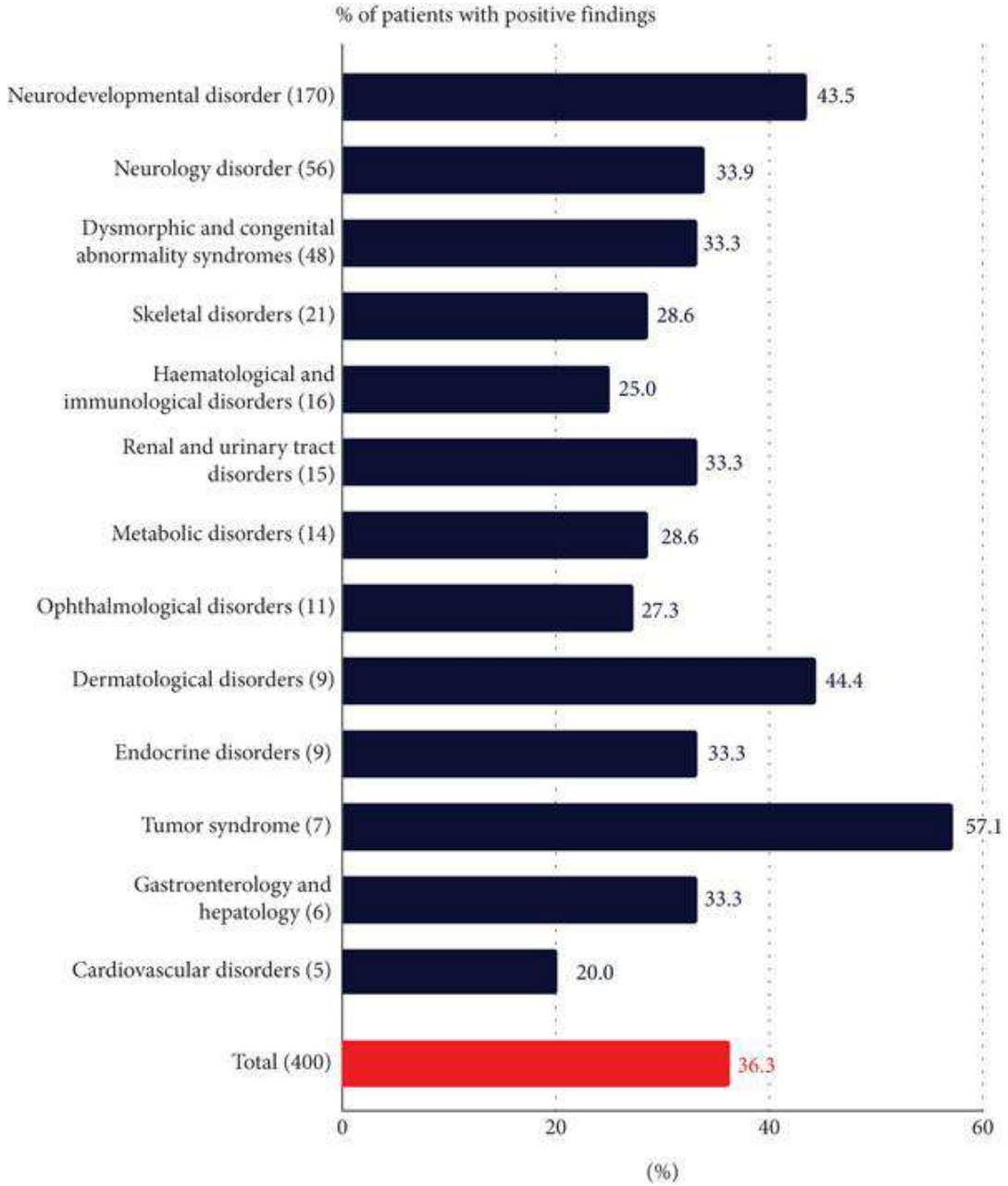
Research Article

Genome Sequencing of Rare Disease Patients Through the Korean Regional Rare Disease Diagnostic Support Program

https://doi.org/10.1155/humu/6096758

overall diagnostic yield was 36.3% (145/400), with 4.8% (7/145) of the diagnosed patients being reported with variants that could not have been identified

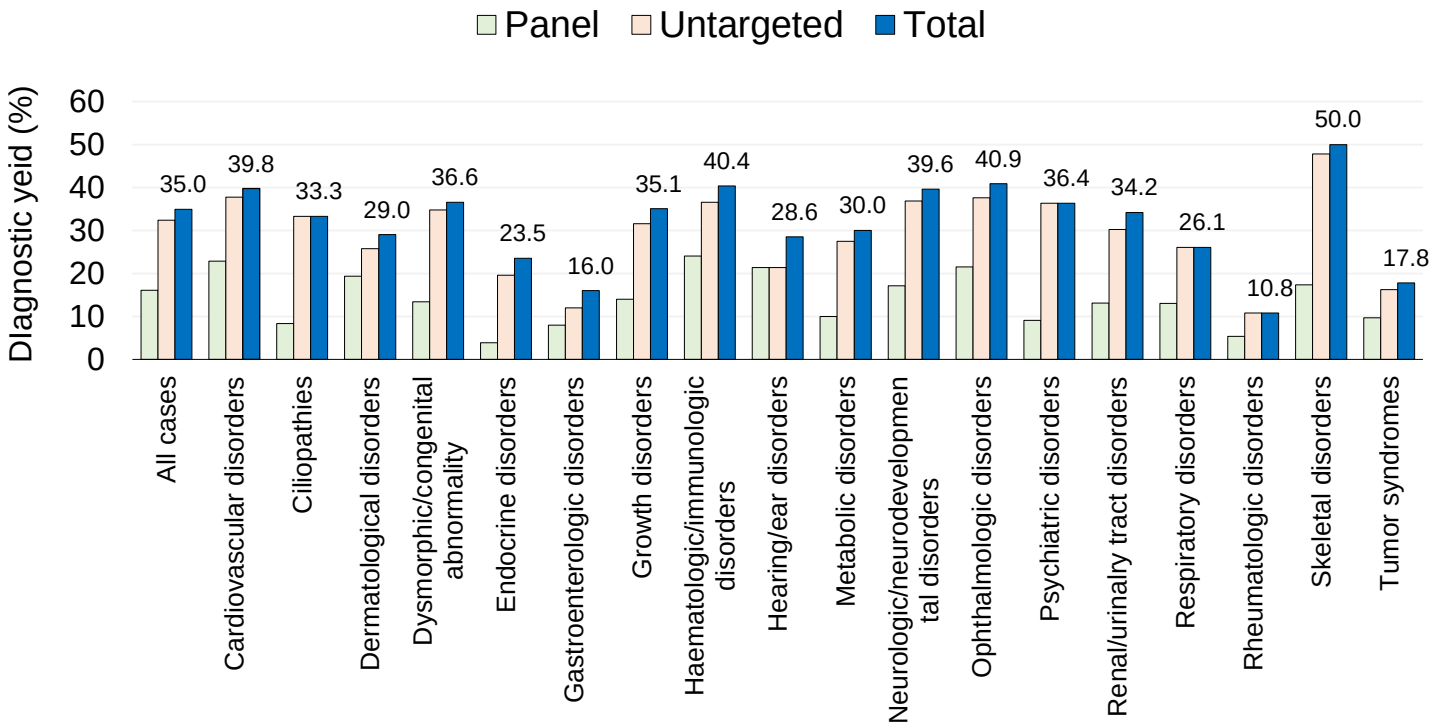


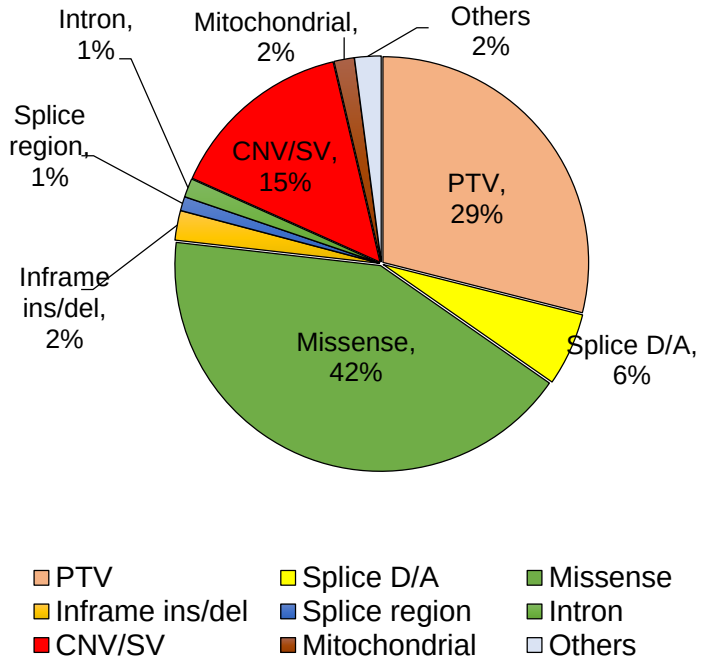


Pilot stage of National Bio Big Data Project

Diagnostic Yield in Rare Diseases

Panel-based analysis vs. WGS





- Total diagnostic yield: 35.0%
- 19 disease categories, 491 diagnostic genes

Challenges in Applying WGS for Rare Disease Diagnosis

1. **Interpretation of variants**
   - Many results fall into Variants of Uncertain Significance (VUS).
   - Difficulty in linking novel or rare variants to clinical symptoms.
   - Requires continuous updates from global genomic databases (e.g., ClinVar, gnomAD).
2. **Lack of standardized guidelines**
   - Interpretation and classification of variants may vary across labs and institutions.
   - Need for consistent application of international standards (e.g., ACMG/AMP guidelines).
3. **Limited clinical actionability**
   - Not all findings are immediately clinically actionable.
   - In many cases, diagnosis may not lead directly to treatment options.
   - Raises ethical questions about disclosure and patient communication.

Challenges in Applying WGS for Rare Disease Diagnosis

4. **Data integration and infrastructure**
   - WGS data must be integrated with clinical records (EHRs) for full value.
   - Requires robust data storage, analysis pipelines, and security infrastructure.
5. **Workforce and expertise gaps**
   - Shortage of trained clinical geneticists, bioinformaticians, and genetic counselors.
   - Need for professional training to interpret WGS results in a clinical context.
6. **Cost and reimbursement issues**
   - In Korea, WGS is not yet reimbursed in standard care settings.
   - Need for policy development on cost-effectiveness and insurance coverage.
7. **Ethical, legal, and social issues** (ELSI)
   - Management of incidental or secondary findings.
   - Ensuring informed consent, data privacy, and patient rights.
   - Long-term data governance for secondary use and research.

ACMG PRACTICE GUIDELINE

Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG)

Kandamurugu Manickam[1,2], Monica R. McClain[3], Laurie A. Demmer[4], Sawona Biswas[5], Hutton M. Kearney[6], Jennifer Malinowski[7], Lauren J. Massingham[8,9], Danny Miller[10], Timothy W. Yu[11,12], Fuki M. Hisama[13] and ACMG Board of Directors[14]*

PURPOSE: To develop an evidence-based clinical practice guideline for the use of exome and genome sequencing (ES/GS) in the care of pediatric patients with one or more congenital anomalies (CA) with onset prior to age 1 year or developmental delay (DD) or intellectual disability (ID) with onset prior to age 18 years.

METHODS: The Pediatric Exome/Genome Sequencing Evidence-Based Guideline Work Group ($n = 10$) used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) evidence to decision (EtD) framework based on the recent American College of Medical Genetics and Genomics (ACMG) systematic review, and an Ontario Health Technology Assessment to develop and present evidence summaries and health-care recommendations. The document underwent extensive internal and external peer review, and public comment, before approval by the ACMG Board of Directors.

RESULTS: The literature supports the clinical utility and desirable effects of ES/GS on active and long-term clinical management of patients with CA/DD/ID, and on family-focused and reproductive outcomes with relatively few harms. Compared with standard genetic testing, ES/GS has a higher diagnostic yield and may be more cost-effective when ordered early in the diagnostic evaluation.

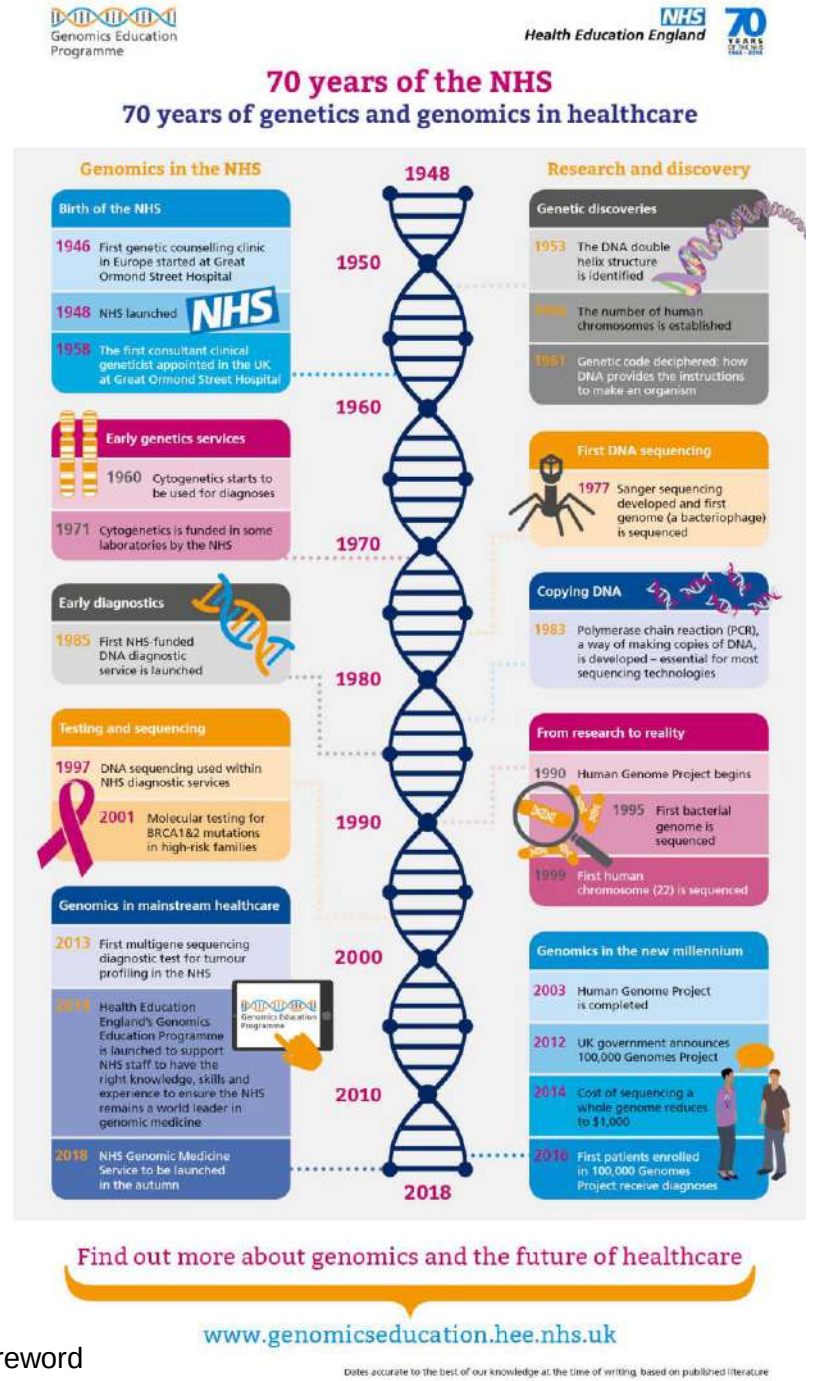
CONCLUSION: We strongly recommend that ES/GS be considered as a first- or second-tier test for patients with CA/DD/ID.

Genetics in Medicine (2021) 23:2029–2037; https://doi.org/10.1038/s41436-021-01242-6

NHS England established the NHS GMS in 2018

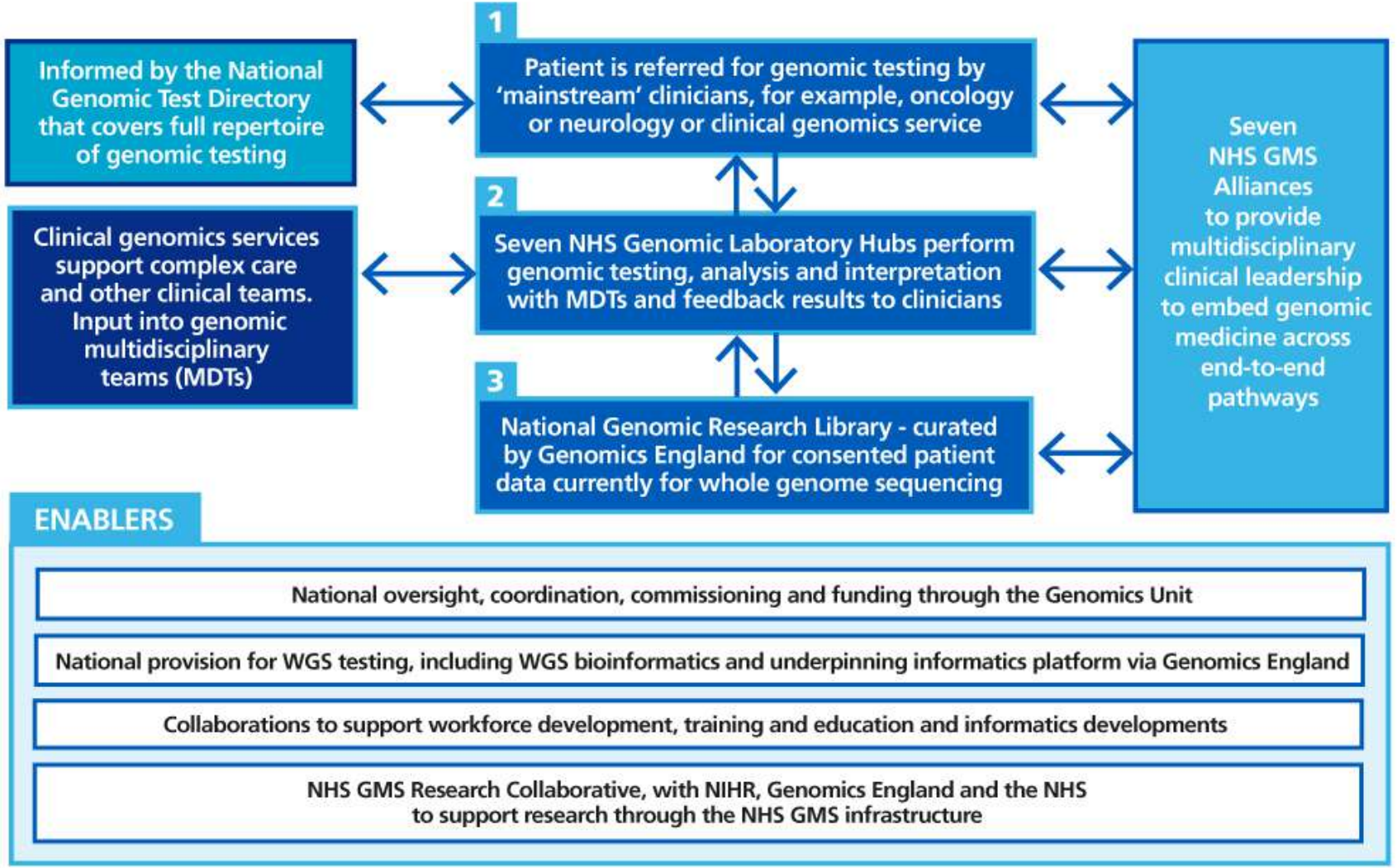
Strategies

- Embedding genomics across the NHS, through a world leading innovative service model from primary and community care through to specialist and tertiary care.
- Delivering equitable genomic testing for improved outcomes in cancer, rare, inherited and common diseases and in enabling precision medicine and reducing adverse drug reactions.
- Enabling genomics to be at the forefront of the data and digital revolution, ensuring genomic data can be interpreted and informed by other diagnostic and clinical data.
- Evolving the service through cutting-edge science**, research and innovation to ensure that patients can benefit from rapid implementation of advances**.



https://www.england.nhs.uk/long-read/accelerating-genomic-medicine-in-the-nhs/#foreword

Infrastructure of NHS Genomic Medicine Service



Therapeutic Application

Zolgensma for SMA Gene Therapy

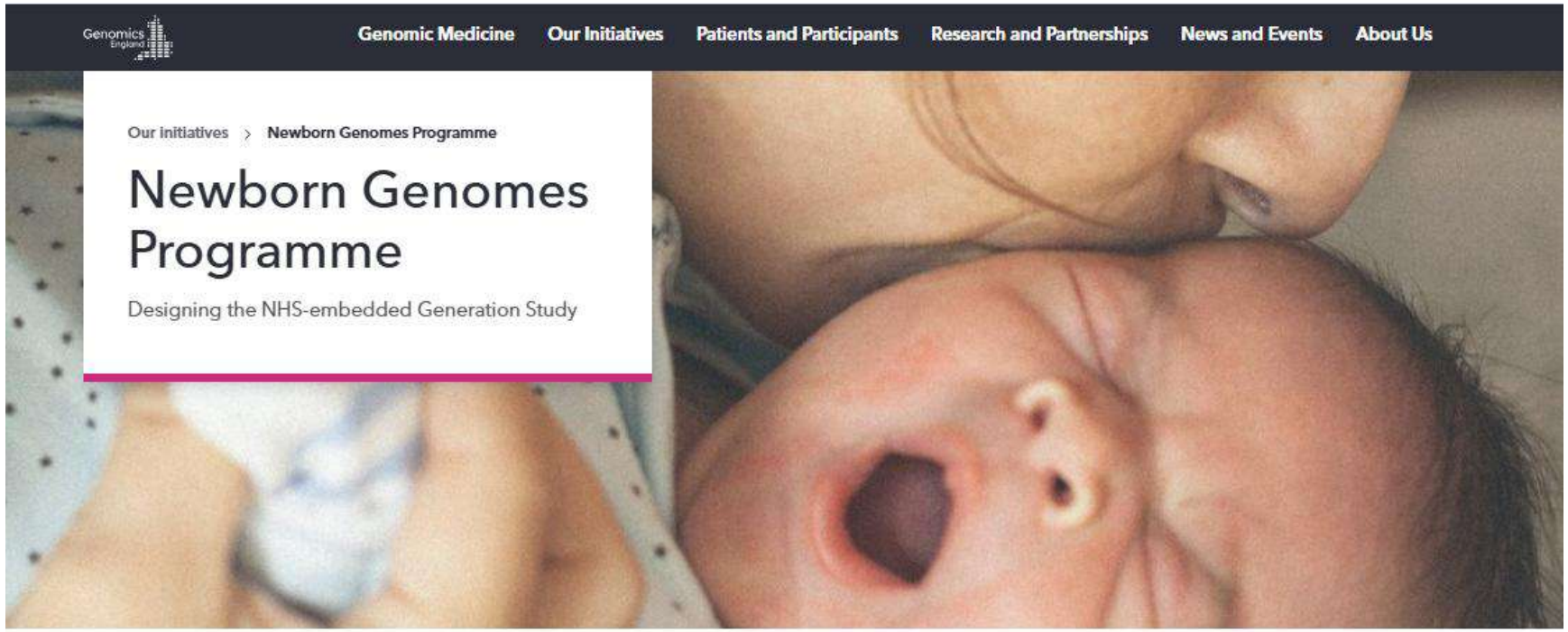
Table 1. *Approved gene therapies for rare diseases in Europe*

Drug name (Company)	Type	Clinical indication	Gene	Inheritance pattern	Year of approval in Europe
Zolgensma (Novartis)	AAV vector-based gene therapy	Pediatric patient (<2 years) spinal muscular atrophy	*SMN1*	Autosomal recessive	2020
Upstaza (PTC Therapeutics International)	AAV vector-based gene therapy	Adults and children (from 18 months) with severe aromatic L-amino acid decarboxylase deficiency	*DDC*	Autosomal recessive	2022
Strimvelis (Orchard Therapeutics)	Gamma-retroviral vehicle for ex-vivo stem-cell therapy	Patients with severe combined immunodeficiency with no available human leukocyte antigen-matched related stem-cell donor	*ADA*	Autosomal recessive	2016
Roctavian (BioMarin International)	AAV vector-based gene therapy	Severe hemophilia A	*F8*	X-linked recessive	2022
Luxturna (Novartis)	AAV vector-based gene therapy	Adults and children with loss of vision due to inherited retinal dystrophy	*RPE65*	Autosomal recessive	2018
Libmeldy (Orchard Therapeutics)	Lentiviral vehicle for ex-vivo stem-cell therapy	Children with metachromatic leukodystrophy	*ARSA*	Autosomal recessive	2020

Abbreviation: AAV, adeno-associated virus.

https://doi.org/10.1111/joim.13655

Newborn Screening



Newborn Genomes Programme | How we work | Engagement | Ethics | How we choose conditions | Evaluation

The Generation Study

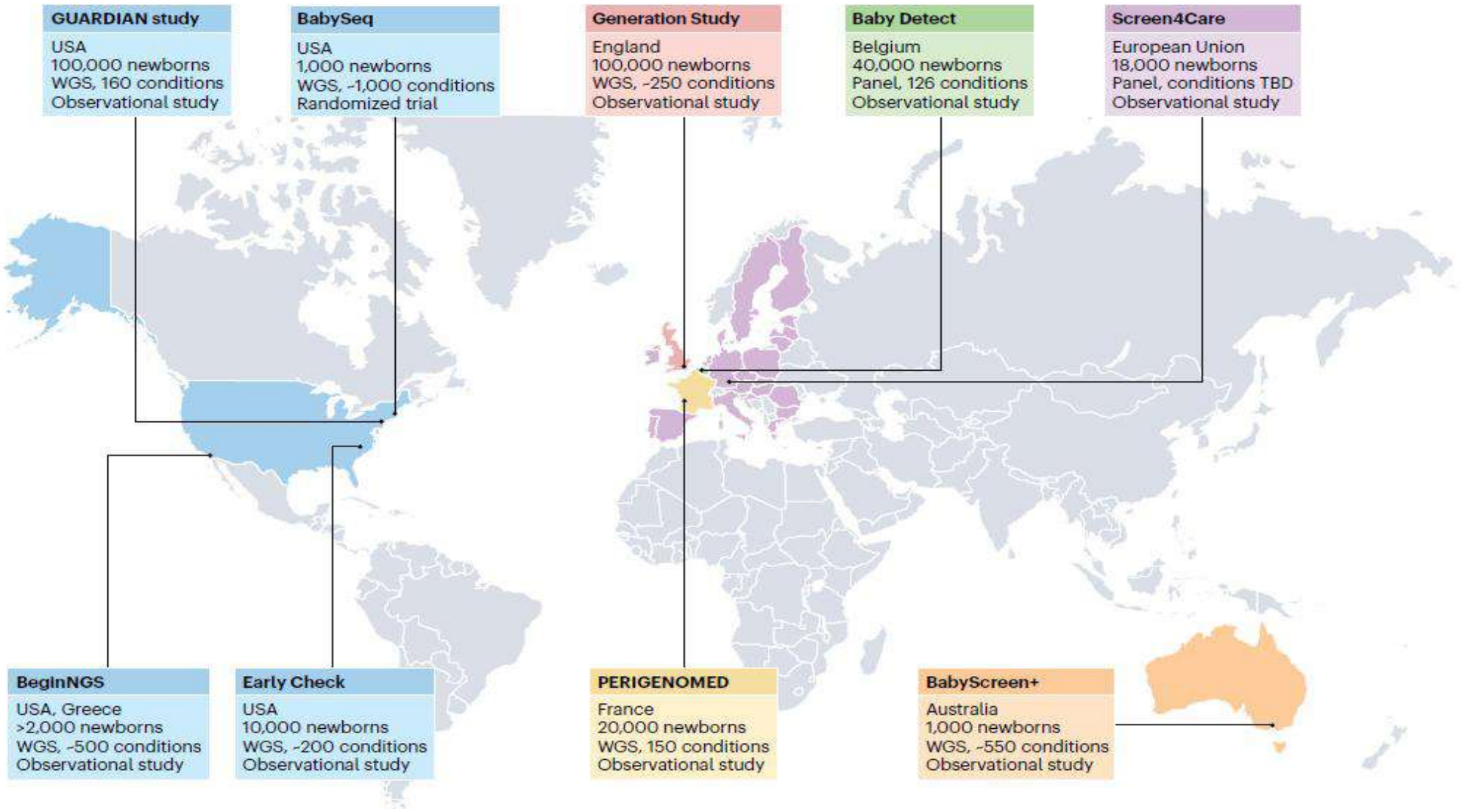
Every year hundreds of babies are born in the UK with rare genetic conditions. Early intervention can enhance the health and quality of life of many of these babies. But these conditions can be hard to diagnose, leading to delays in care.

The Generation Study is a groundbreaking research study which will sequence the genomes of 100,000 newborn babies. We are running our study in partnership with the NHS to understand whether we can improve our ability to diagnose and treat genetic conditions.

Learn more about how the Generation Study works by watching this video.

Genomics england

International Consortium on Newborn Sequencing



National Institute of Health
Republic of Korea

Pilot Project for Genomic Newborn Screening in Korea

1. **Target conditions**
   - Identify early-onset, actionable genetic conditions that may benefit from prompt intervention
   - Define which genes and conditions will be screened
   - Examine the impact of returning secondary findings (e.g., adult-onset disorders, carrier status)
2. **Develop protocols for informed consent and clinical decision support**
3. **Data privacy and long-term storage**
   - WGS data is sensitive and must be protected throughout the individual's life
   - Requires long-term secure storage infrastructure and policies on secondary use
   - Consider re-consent models as the child matures
4. **Assess integration with national health records and long-term follow-up systems**

Rapid genome sequencing for critically ill neonates

Genetic diagnoses were achieved in 10 patients (total: 20 trio, diagnostic yield: 50%)

Sample #	Phenotype	TAT	Gene	Variant	Inheritance	Variant type
rWGS-03	polycystic kidney dysplasia	5day 4h	*AGT*	NM_001384479.1:c.292G>A	Homozygous	Missense
rWGS-04	(unconjugated) hyperbilirubinemia * maternal hemolytic anemia	5day 2h	*ANK1*	NM_000037.4:c.2394_2397 del	Heterozygous (Maternal)	Frameshift
rWGS-09	Fetal thrombotic vasculopathy	5day 14h	*NSD1*	NM_022455.5:c.5885T>C	De novo	Missense
rWGS-10	Renal vein thrombosis, both., IVC thrombus	5day 1h	*SERPINC1*	NM_000488.4:c.235C>T	Homozygous	Missense
rWGS-11	Asymptomatic hyperammonemia	7day	*OTC*	NM_000531.6:c.513G>C	X-linked	Missense
rWGS-12	Hyperglycemia, Type I diabetes mellitus	5day 1h	*INS*	NM_000207.3:c.265C>T	De novo	Missense
rWGS-14	ARPKD, Potter syndrome	4day 13h	*PKHD1*	NM_138694.4:c.6840G>A NM_138694.4:c.6602T>A	Compound heterozygous	Stop gained, stop gained
rWGS-17	thrombocytopenia with growth retardation, feeding difficulties	3day	*SAMD9*	NM_017654.4:c.2414A>G	De novo	Missense
rWGS-18	Hyperammonemia, lactic acidosis	4.5day	*OTC*	NM_000531.6:c.841T>G	X-linked	Missense
rWGS-19	Hydrops fetalis, Congenital thrombocytopenia, Lung hypoplasia	3.5day	*LZTR1*	NM_006767.4:c.742G>A	De novo	Missense

Submitted

Pharmacogenetics

Strong guideline frameworks

- CPIC (Clinical Pharmacogenetics Implementation Consortium)
- DPWG (Dutch Pharmacogenetics Working Group)
- FDA labeling includes PGx info for over 400 drugs.

Implementation Models

- Preemptive Testing: Genetic data obtained before prescribing, stored in EHR (e.g., St. Jude, Mayo Clinic)
- Reactive Testing: Ordered at the time of prescribing based on need
- Integration into Clinical Decision Support (CDS) is key for usability.

Genomic medicine initiatives

- USA: Widespread implementation in academic health centers ; PGx panels offered by commercial labs
- Europe: National strategies (e.g., Netherlands, UK) embedding PGx into care pathways
- Asia: Japan and Singapore have PGx in routine drug safety

Implementation of Genomic Medicine in Clinical Practice

Current limitations:

No reimbursement pathway

Limited scalability and clinical feedback loop

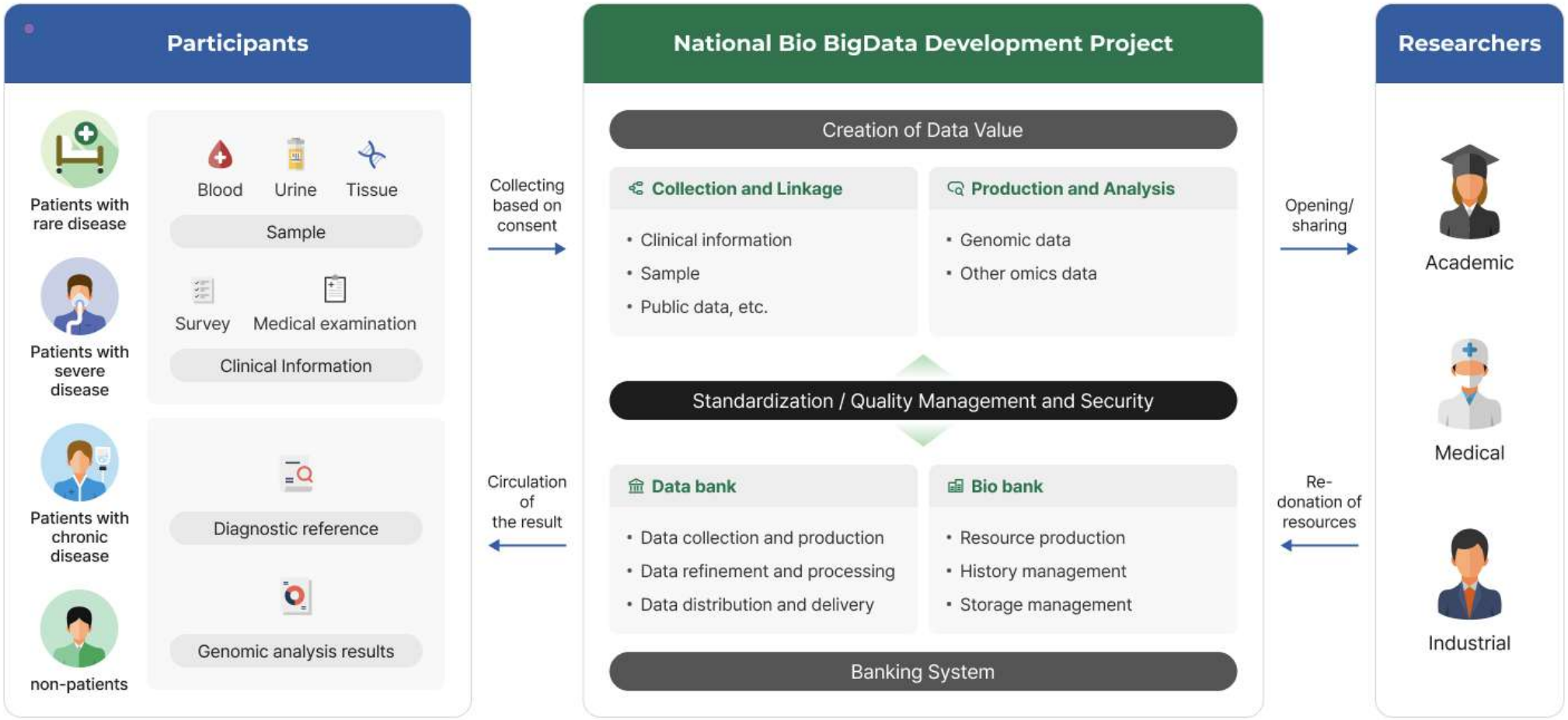
Policy opportunity:

Transition WGS from research-only to **standardized clinical application**, supported by ethical, regulatory, and reimbursement frameworks.

National Bio Big Data Project

Integrated Bio Big Data of 1,000,000 people

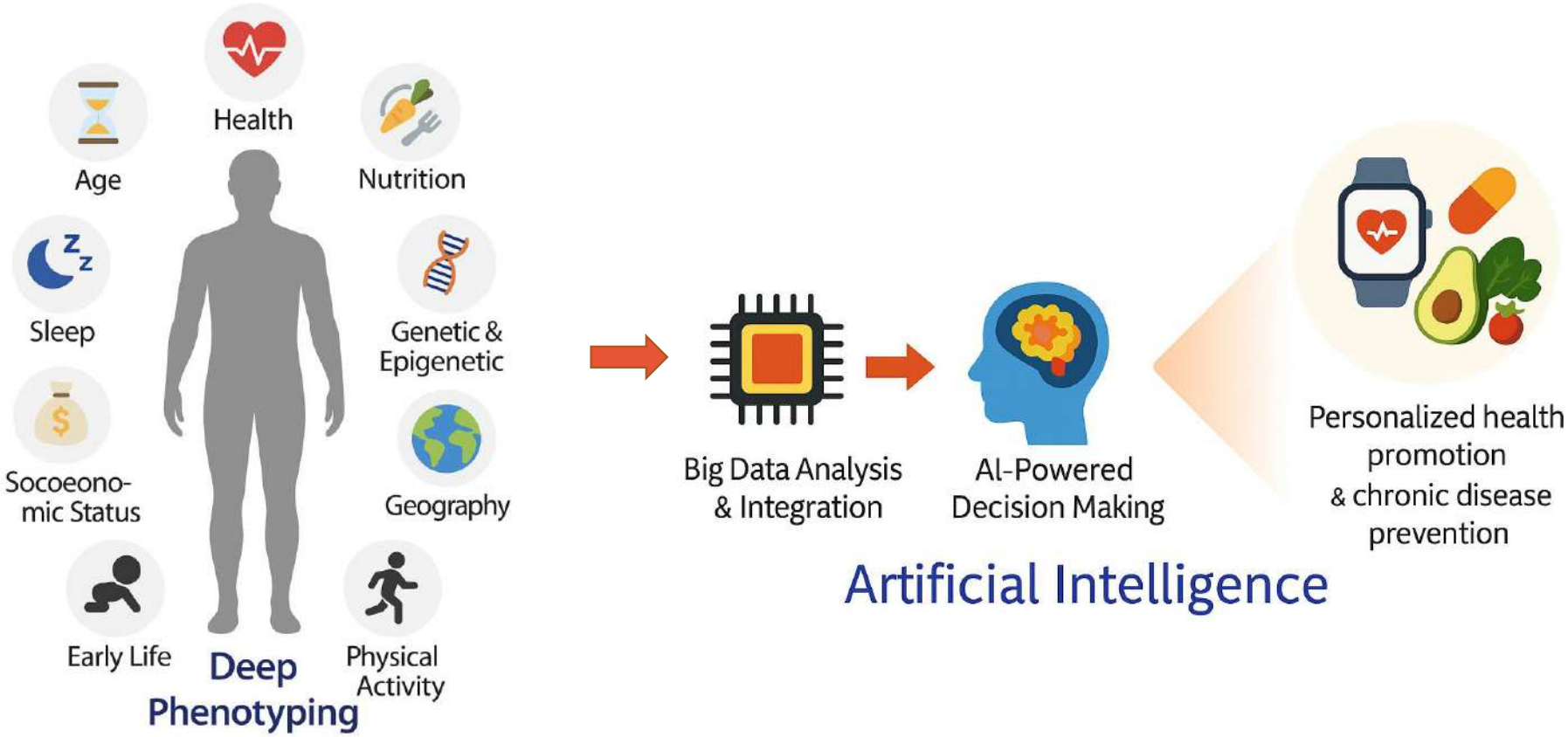
BIKO BioBigData.Korea 국가통합바이오 빅데이터구축사업단



Biomedical Research Data beyond Genomics

KNIH’s Bio Big Data Initiative

Deep Phenotyping is essential for Precision Health



Deep phenotyping and artificial intelligence for health promotion and chronic disease prevention.

modified from the figure. *J Transl Med* **18,** 472 (2020). https://doi.org/10.1186/s12967-020-02658-5

KNIH’s Bio Big Data Initiative

Large-scale
Cohort Studies

Korea Biobank
Project

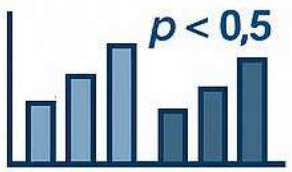
Korean Genome
Analysis Projects

Biomedical data
beyond the genome

Data sharing
CODA

National Institute of Health
Republic of Korea

Large-Scale Cohorts as Key Drivers of Precision Health Research



Capturing Genetic and Environmental Diversity

- Enables analysis of diverse populations across age, ethnicicty, lifestyle, and regions
- Helps uncover gene-environment interactions and rare variant effects

Powering Statistical Validity

- Large sample sizes Improve statistical power for discovering disease biomarkers and risk factors
- Facilitates robust subgroop analysis (e.g. sex-specific, age-specific effects

Longitudinal Insights for Disease Progression

- Tracks health trajectones over time to understand natural history of diseases
- Supports early detection and predictive modeling of chronic diseases

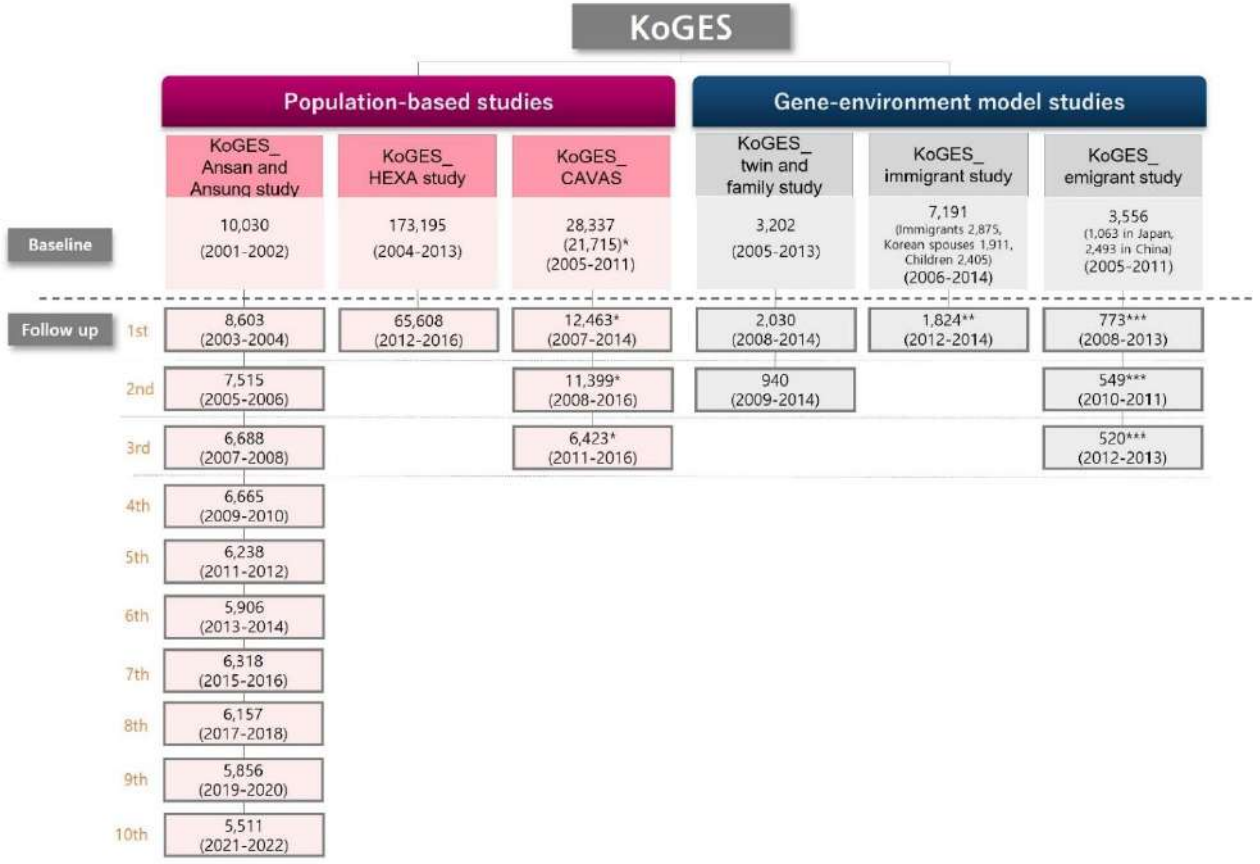
Foundation for Data Integration

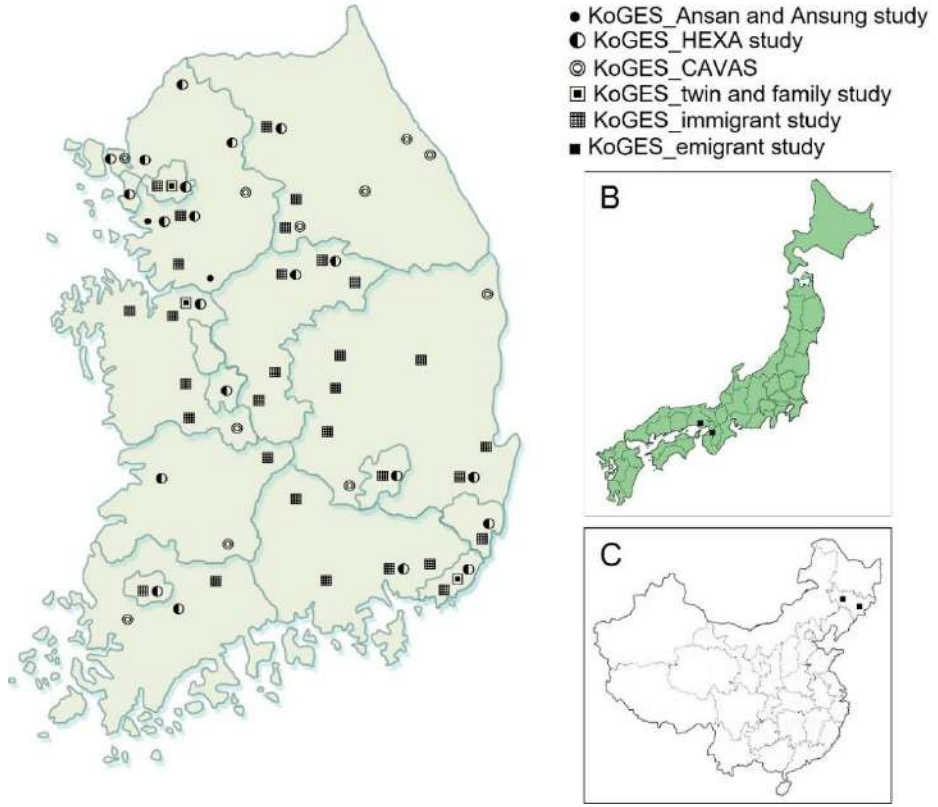
- Allows linkage with genomic, clinical, behavioral, and digital health data
- Supports AI-driven precision diagnostics and personalized treatment pathways

KoGES

Korean Genome and Epidemiology Study

- Started in 2001, **6 sub-cohorts** representing diverse population groups.
- **KoGES-Ansan & Ansung study** is **still ongoing with over 20 years of F/U.**
- **240,000** participants have been recruited by the end of 2014
- **Major target diseases :** T2DM, metabolic syndrome, hypertension, cardiovascular disease, osteoporosis, cancer





NIH's Cohort or Registry

Currently operating more than 30 cohorts or registries

General Population / Special Group Cohorts

Cohort Name	Start year	Number of Participants
KoGES Rural (Ansung)	2001	5,018
KoGES Urban (Ansan)	2001	5,012
KoGES Rural (CAVAS)	2004	28,337
Korea Nurses Health Study (KNHS)	2013	20,613
Women's Health Study	2014	4,684
Korea Neonatal Network (KNN)	2013	22,651
Korea Frailty and Aging Cohort Study (KFACS)	2016	3,011
Korea Urban and Rural Elderly Study (KURE)	2011	3,517
Cardiovascular & Metabolic Etiology Research Cohort (CMERC)	2013	11,375
Korean Transplant Registry (KOTRY)	2014	45,396

Disease-focused Cohorts

Cohort Name	Start year	Number of Participants
Korean Stroke Registry, KSR	2017	25,515
Korean childhood Asthma Study	2016	961
Korea COPD Subgroup Study	2016	2,952
Resistant hypertension cohort	2018	1,432
Community-based Dementia Cohort	2019	4,789
Parkinson Diseases Registry	2021	749
Hospital-based Dementia Cohort	2021	646
Early-onset Dementia Cohort	2021	406
Korean Severe Asthma Registry	2022	591
Korea HIV/AIDS Cohort	2006	1,644
Korea HCV Cohort	2007	3,858
Korea HBV Cohort	2014	3,021
Tuberculosis Cohort	2019	2,001

Cohorts for Aging Research

KoGES
Korean Genome and Epidemiology Study

Started in 2001
Study population: aged 40~69
Collaboration: Ajou University, Korea University, Hanyang University

KURE
Korean Urban and Rural Elderly cohort

Started in 2012
Study population: aged 65 and older
Collaboration : Yonsei University

KFACS
Korean Frailty and Aging Cohort Study

Started in 2016
Study population: aged 70~84
Collaboration : Kyung Hee University

Korean Centenarian Study
(2025~)

Starts in 2025
Study population: aged 90 and older
Focusing on healthy aging and longevity

Biomedical Data Generation

Omics data

- Genome: K-chip(DNA microarray chip) & imputed data, WGS
- Epigenome: DNA methylation profiles using Illumina arrays (e.g., 850K, 935K)
- Metabolome & Proteome

Brain MRI

- ~5,700 participants in KoGES- Ansan and CAVAS
- Structural & functional analysis of brain imaging(T1, T2, Flair, DTI, fMRI)
- Regional brain volume & morphometric measures

Microbiome

- A total of 800 participants in KoGES-Ansung & KFACS
- including:
  - Fecal DNA samples
  - Microbiome sequencing data
  - Related clinical and epidemiological information

Air pollution

- Cohort data have been linked to air pollution exposure data
  - PM10, PM2.5
  - SO2, NO2, CO, O3
- Meteorological data: Humidity, Wind speed, Precipitation, Cloud cover, Solar radiation, etc

Link with Health Data

- Death data(National statistics)
- Cancer registry (National Cancer Center)
- National Health Insurance data

Korea Biobank project

National Biobank of Korea

Korea Biobank Network

Korea Biobank Network = NBK + 47 regional biobanks

Central Biobank
(**N**ational **B**iobank of **K**orea)

Population-based

Cohort studies
- KoGES
- CMERC

National Survey
- KNHANES
- BICWALZS

Biospecimen/Data

Funding
Informatics Systems
Standardization

Regional Biobanks
(University Hospitals)

Disease-based

- In- & out-Patients
- Health examinees

Biospecimen/Data

Access & Sharing Platform

Biobank Portal (biobank.nih.go.kr)

KBN Portal (koreabiobank.re.kr)

국립중앙인체자원은행
NATIONAL BIOBANK OF KOREA

ARS 1661-9070
Distribution for research

KBN 한국인체자원은행네트워크 Korea Biobank Network

Intramural | University | Industry | Others

Korea Biobank Project

국립중앙인체자원은행 NATIONAL BIOBANK OF KOREA

0.47 million participants

- **Population-based cohorts or Surveys**
  - Korean Genome Epidemiology Study (KoGES, n=249K+)
  - Korea National Health & Examination Survey (KNHANES, n=108K+)
- **Infection related studies**
  - National immunity surveys, etc. (n=48K+)
- **Other biobanking studies**
  - CMERC, KOTRY, BICWALZS etc. (n=52K+)

blood DNA Urine — 10.5 mil vials stored in a central biobank (NBK)

Epidemiological & clinical data — 380K participants

Multi-omics data (K-chip, WES, WGS, RNA-seq, methyl, absolute IDQ, etc) — 170K participants

+

KBN 한국인체자원은행네트워크 Korea Biobank Network

0.76 million participants

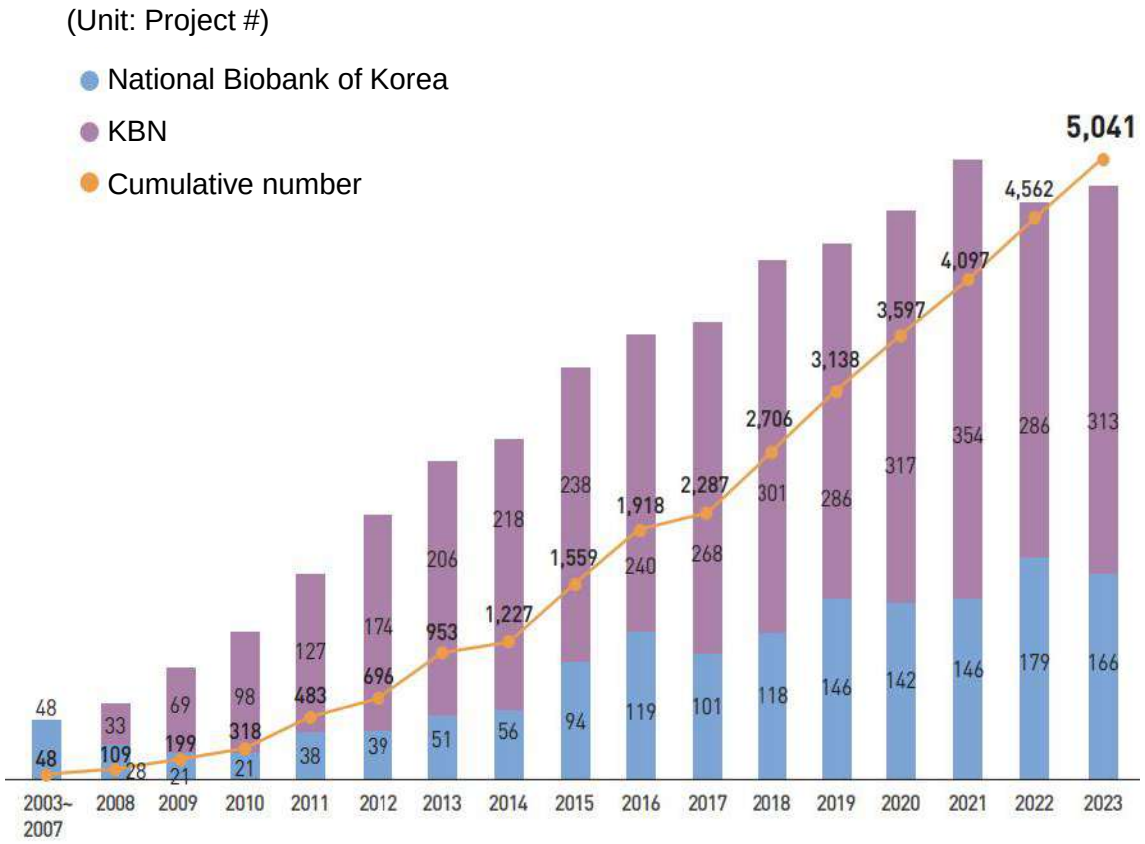
- **Specialized Biobank Subnetwork**
  - 10 subnetwork (10 Hub Biobanks + 20 collaborative Biobanks)
  - Establishment and operation of human biobank specialization subnetwork
- **Innovative Biobanking Consortium**
  - 4 consortium (Chronic cerebrovascular diseases & Alzheimer's diseases, Sarcoma, developmental disorder)
  - Establishment of clinician-led biobanks and healthcare R&D Ecosystem

Tissue blood DNA — 9.5 mil vials stored in across the regional biobanks (n=47)

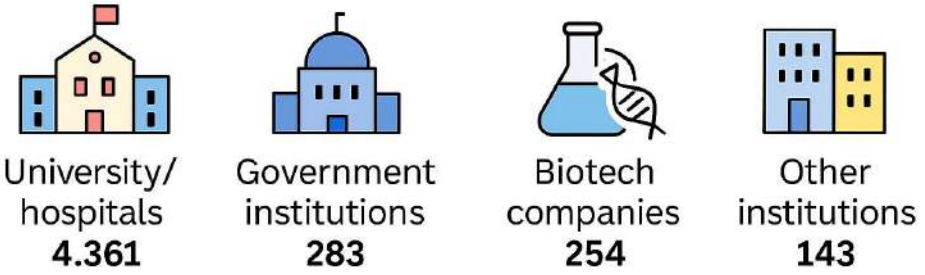
Clinical data

Total 0.76 million (annotated with 11 clinical variables)

Utilization of Human Biological Materials

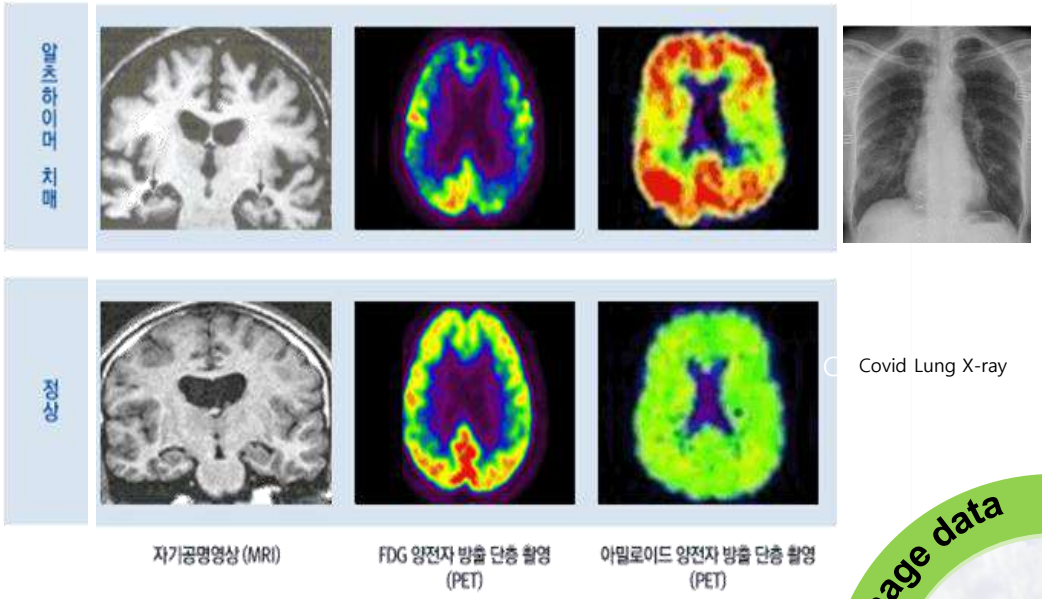


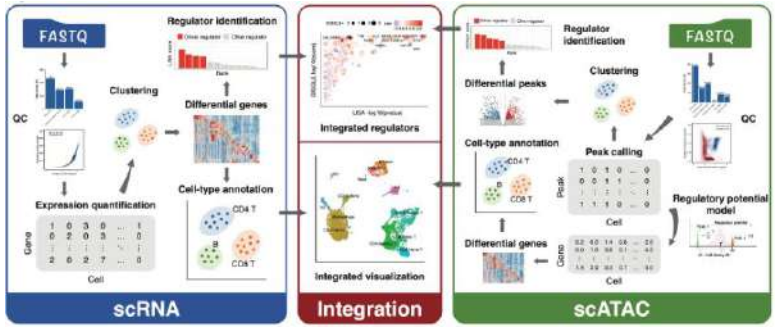
Human biological materials have been used in over 5,000 studies

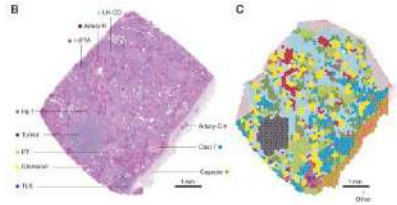


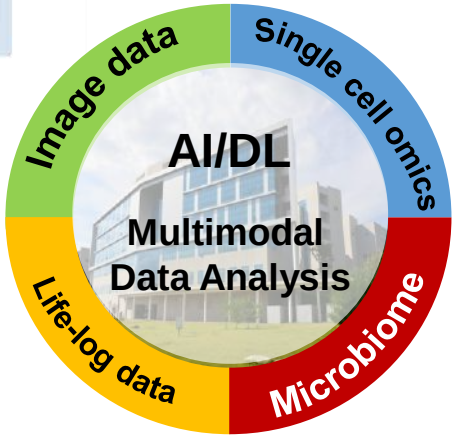
2053 outcomes have been derived from human resources

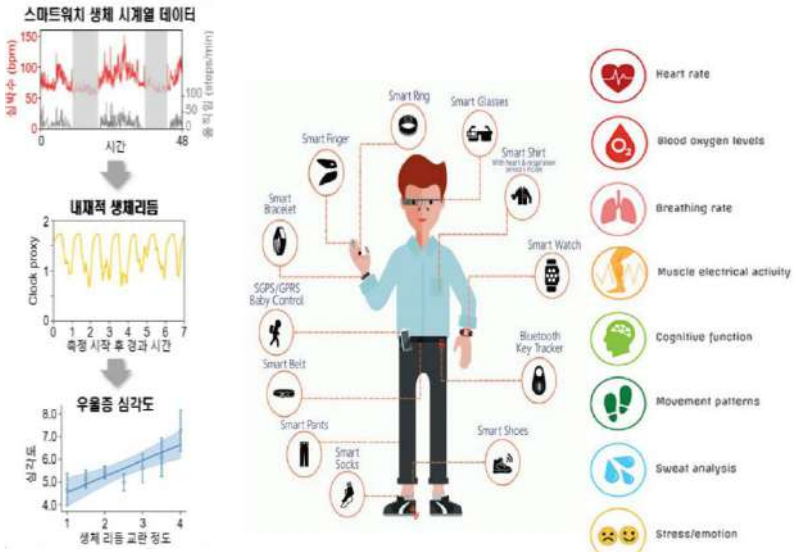
As of the end of 2023





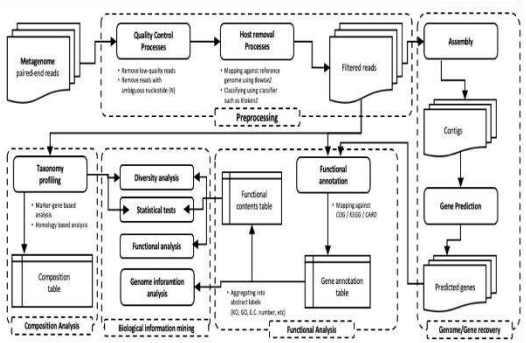


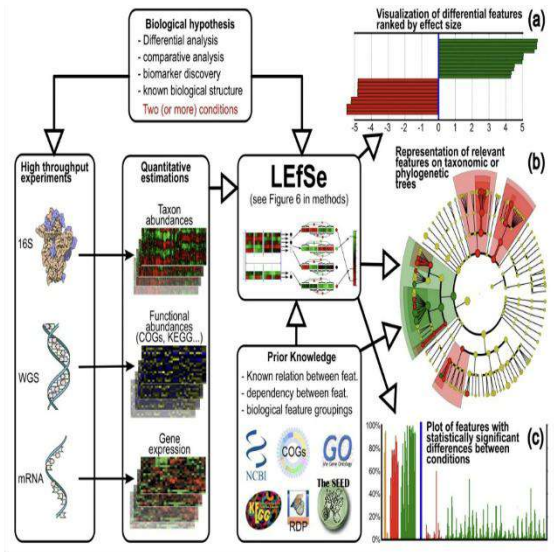




16S rRNA full length

Shotgun

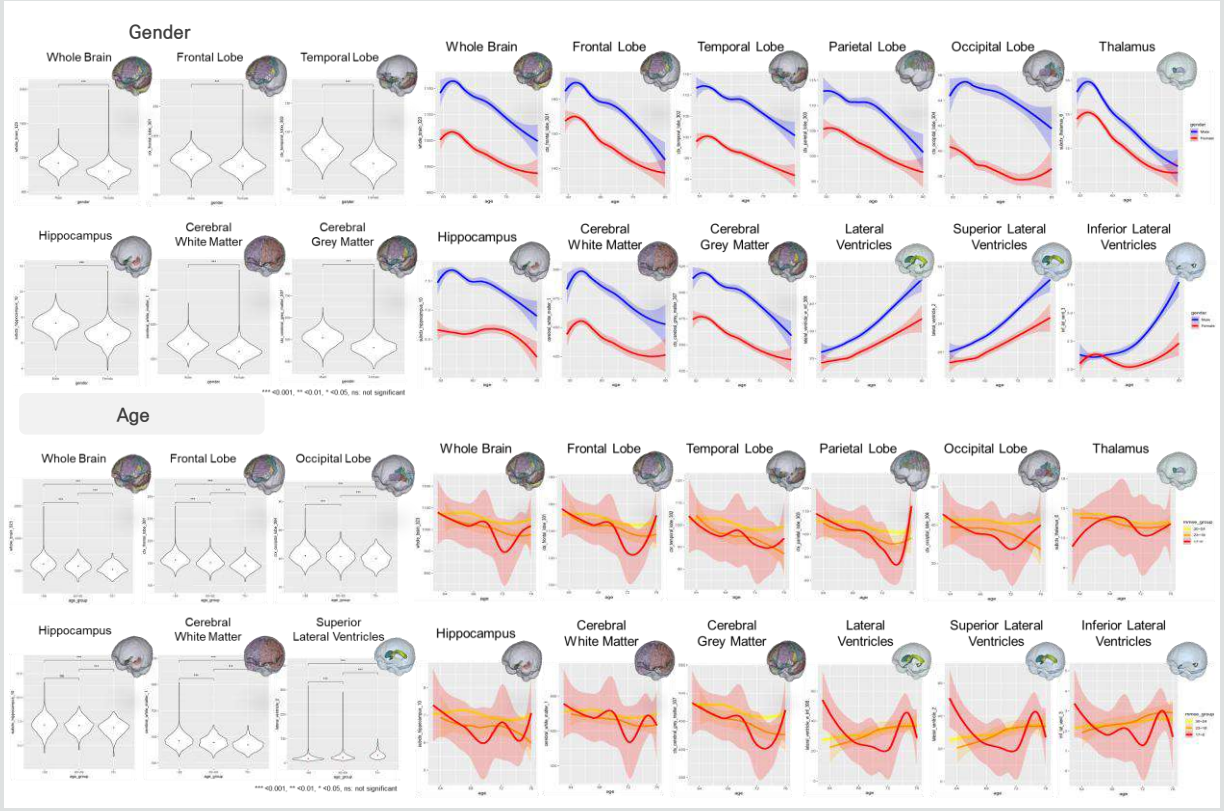




Curation and database development of neuroimaging data

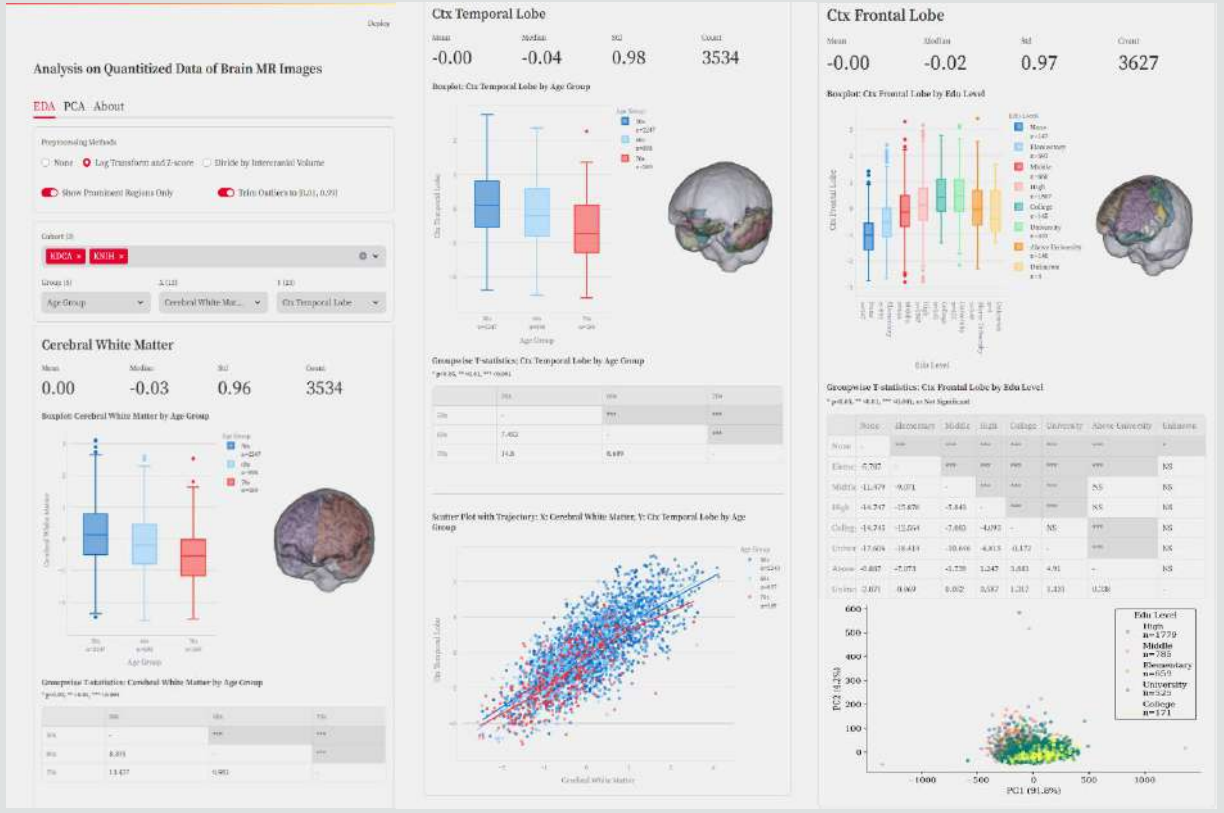
Establishment of a Korean Brain MRI Reference Database

Establishment of a reference dataset of brain regional volume values based on sex, age group, cognitive function, and disease presence/severity (approx. 5,000 individuals)



Development of an App for Visualizing Brain Imaging Data

Development of an interactive web-based visualization tool for multi-center cohort-based brain MRI quantification results and clinical/epidemiological data sets



Central Data Repository

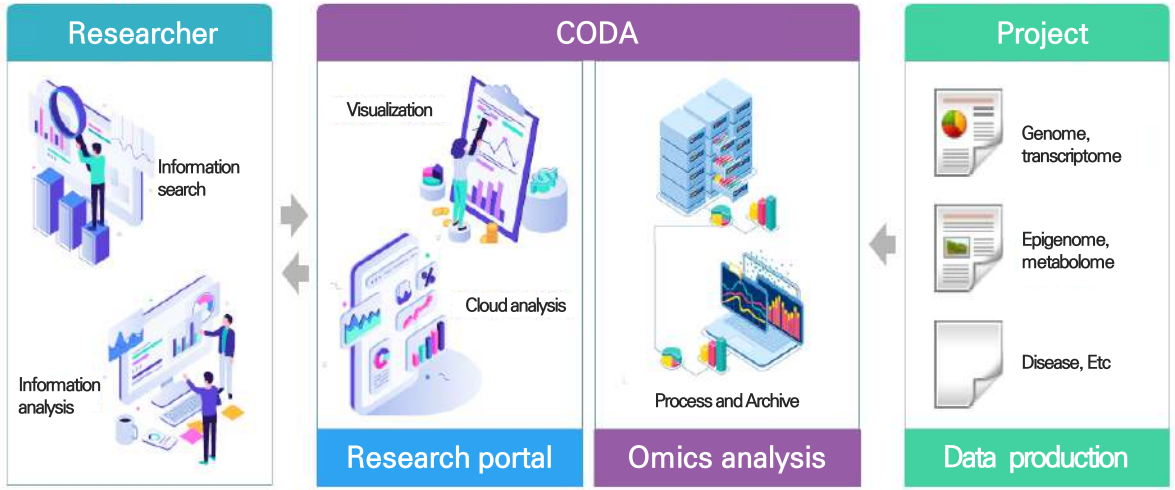
National Institute of Health
Republic of Korea

CODA

Clinical & Omics Data Archive

- It collects **clinical/epidemiological** and **omics data** (microarray, whole-exome, whole-genome, transcriptome, metabolome)
- Collect, share, and utilize biomedical data (resource) derived from national R&D projects or voluntary data deposit
- Established in 2016 at NIH

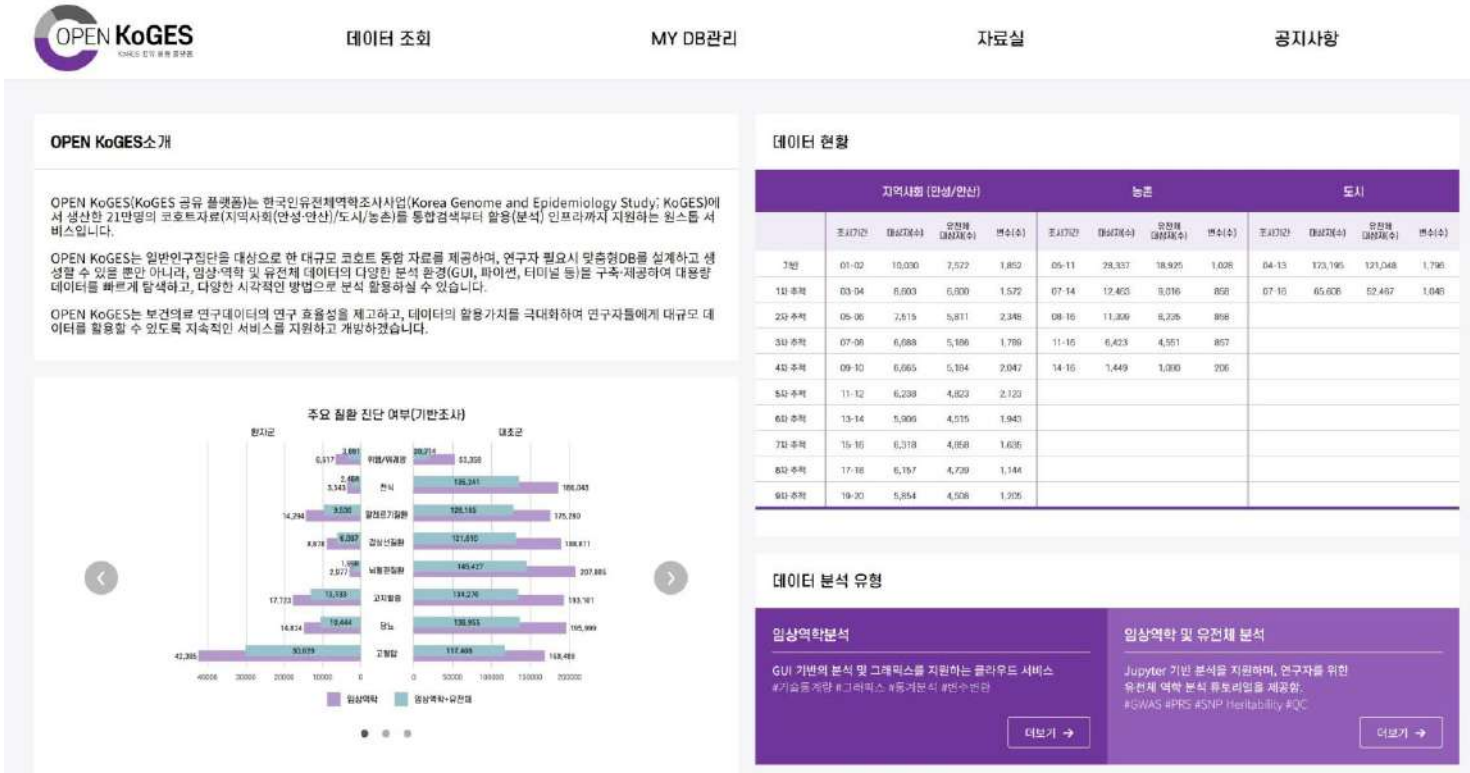
  in accordance with Article 8(2)(3) of the Enforcement Decree of the Bio-Research Resources Act.

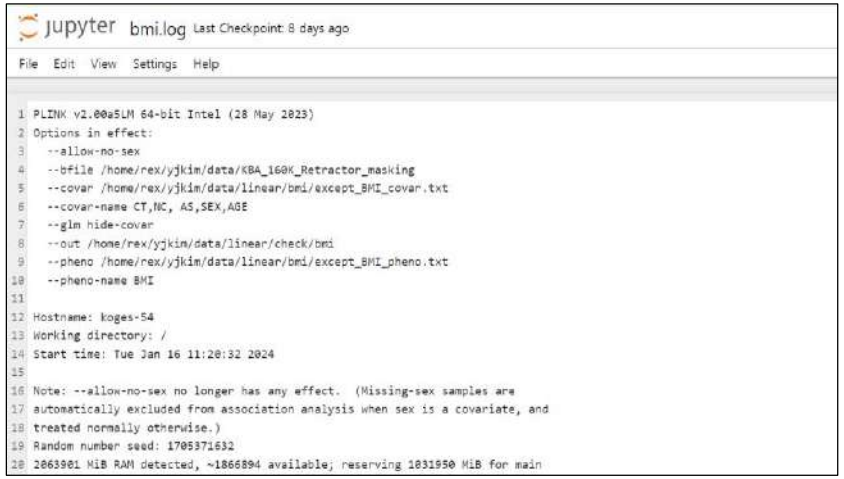


- The system enables researchers to search for and request necessary data, and to perform analyses using a built-in analysis pipeline within the cloud infrastructure.

Open data platform

- ✓ Improve user accessibility
- ✓ Provision of analytical infrastructure via cloud-based operations





GWAS 수행 로그화면

The Need for Integrated Data Ecosystems

- To enable comprehensive and accurate decision-making
- To improve patient outcomes
- To accelerate innovation and research
- To support health system efficiency
- To respond effectively in public health crises

The Need for Integrated Data Ecosystems

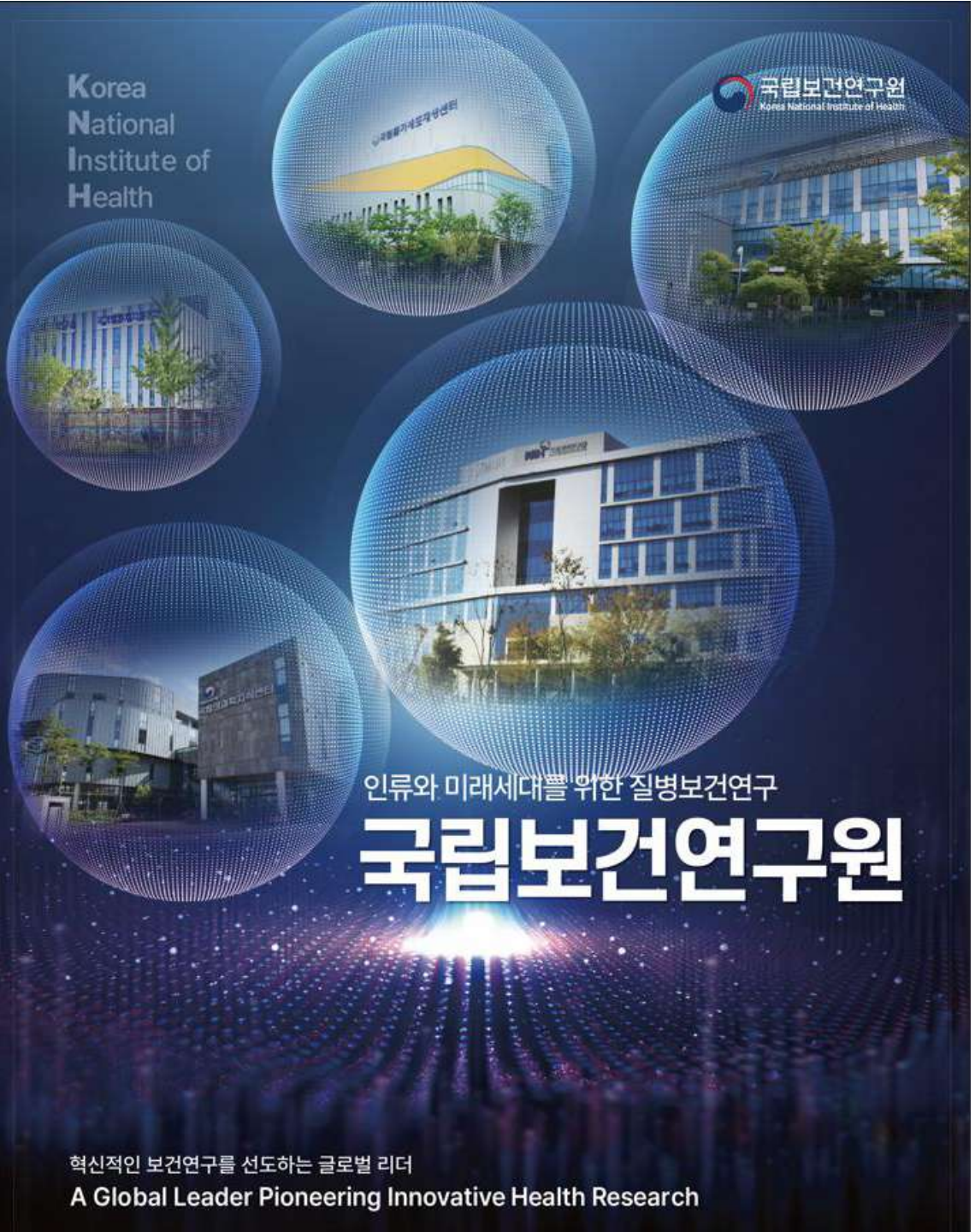
Data	Description	Data source
Clinical Data	Real-time medical data from healthcare providers	EMRs, lab results, prescriptions
Public Data	National-level health, claims, and surveillance	NHIS, HIRA, KDCA, mortality data
Research Data	Discovery-driven data from cohort & omics studies	Biobanks, Cohort/Registry, National Research Initiatives

Barriers to Overcome

- **Technical**: Lack of standards, interoperability gaps
- **Legal**: Regulatory silos between health and research data
- **Ethical**: Consent frameworks for secondary use
- **Institutional**: Siloed data ownership, weak incentives for sharing

"No single dataset is sufficient. Smarter healthcare requires smarter data integration."

By bringing together clinical, public, and research data, Korea can lead the way in evidence-based, personalized, and equitable healthcare decisions.



Thank you

National Institute of Health
Republic of Korea