Review
Others

Diabetes Metab J 2022;46:552-563
https://doi.org/10.4093/dmj.2022.0193
pISSN 2233-6079 · eISSN 2233-6087

dmj
DIABETES & METABOLISM JOURNAL

Current Trends of Big Data Research Using the Korean National Health Information Database

Mee Kyoung Kim[1], Kyungdo Han[2], Seung-Hwan Lee[3,4]

[1]Division of Endocrinology and Metabolism, Department of Internal Medicine, Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul,
[2]Department of Statistics and Actuarial Science, Soongsil University, Seoul,
[3]Division of Endocrinology and Metabolism, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul,
[4]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Korea

Recently, medical research using big data has become very popular, and its value has become increasingly recognized. The Korean National Health Information Database (NHID) is representative of big data that combines information obtained from the National Health Insurance Service collected for claims and reimbursement of health care services and results obtained from general health examinations provided to all Korean adults. This database has several strengths and limitations. Given the large size, various laboratory data, and questionnaires obtained from medical check-ups, their longitudinal nature, and long-term accumulation of data since 2002, carefully designed studies may provide valuable information that is difficult to obtain from other forms of research. However, consideration of possible bias and careful interpretation when defining causal relationships is also important because the data were not collected for research purposes. After the NHID became publicly available, research and publications based on this database have increased explosively, especially in the field of diabetes and metabolism. This article reviews the history, structure, and characteristics of the Korean NHID. Recent trends in big data research using this database, commonly used operational diagnosis, and representative studies have been introduced. We expect further progress and expansion of big data research using the Korean NHID.

Keywords: Database; Diabetes mellitus; Korea; Metabolism; National health programs

INTRODUCTION

In recent years, big data analysis has become one of the mainstream areas of medical research. Since studies using big data have both strengths and limitations, they provide important insights that cannot be achieved by other forms of research. However, some possibilities of bias and caution in interpretation need to be acknowledged. The Korean National Health Information Database (NHID), which contains a nationwide claims database and health examination data, represents the Korean population and has become an attractive source of research in various fields. In this review, we describe the characteristics of the Korean NHID and provide an overview of recent trends in research related to diabetes.

HISTORY OF NATIONAL HEALTH INSURANCE SERVICE AND HEALTH EXAMINATION

The Medical Insurance Act was enacted in 1963 in Korea. At that time, the medical insurance society could be established voluntarily at an industrial establishment with 300 workers or

Corresponding author: Seung-Hwan Lee https://orcid.org/0000-0002-3964-3877
Division of Endocrinology and Metabolism, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Korea
E-mail: hwanx2@catholic.ac.kr

Received: Jun. 6, 2022; Accepted: Jun. 30, 2022

more. Through several amendments, medical security for the entire population was achieved in 1989. The National Health Insurance Act, which was enacted in 1999 and enforced in January 2000, integrated all insurers into a single insurer (National Health Insurance Corporation [NHIC]) and established an independent organization for health care review and evaluation (Health Insurance Review Agency—currently the Health Insurance Review & Assessment Service [HIRA]) [1].

The implementation of health examinations for public officials, faculty and staff of private schools, and policyholders within the medical insurance corporation began in 1980. In the 1990s, screening for specific types of cancer was implemented and expanded to include public officials, faculty and staff of private schools, workplace policyholders, regional policyholders, and their dependents. Through legislation and promulgation of the Framework Act on Health Screening in 2008, the target diseases for general health screening were established, and the list of items included in the screening test was improved in 2009. General health screening for regional household members and dependents was expanded to include people aged 20 years or older in 2019 [2].

The NHID refers to big data combining information obtained from the National Health Insurance Service (NHIS) and health examinations. It includes qualification, insurance rate, medical check-up results, treatment details, elderly long-term nursing insurance data, clinic status, registered information on cancer and rare diseases, etc. This database was established in 2011, and a sample cohort database was established in 2012 [3].

OPERATIONAL STRUCTURE OF THE NATIONAL HEALTH INSURANCE SYSTEM

The NHIS and HIRA are under the supervision of the Ministry of Health and Welfare, which plays a role in the formulation and implementation of policies. The NHIS is a non-profit organization and a single insurer that manages the system in Korea. They are responsible for (1) managing the qualifications of insured individuals and their dependents; (2) imposing and collecting contributions; (3) paying healthcare service costs to healthcare service providers; and (4) purchasing health screening. Health service providers claim reimbursement of corporations' share of healthcare service costs to the NHIS and HIRA and receive copayment from insured individuals. The HIRA evaluates the adequacy of healthcare service costs by reviewing medical billing and claims and announces the review results to the NHIS and healthcare service providers (Fig. 1). The contribution of an em-

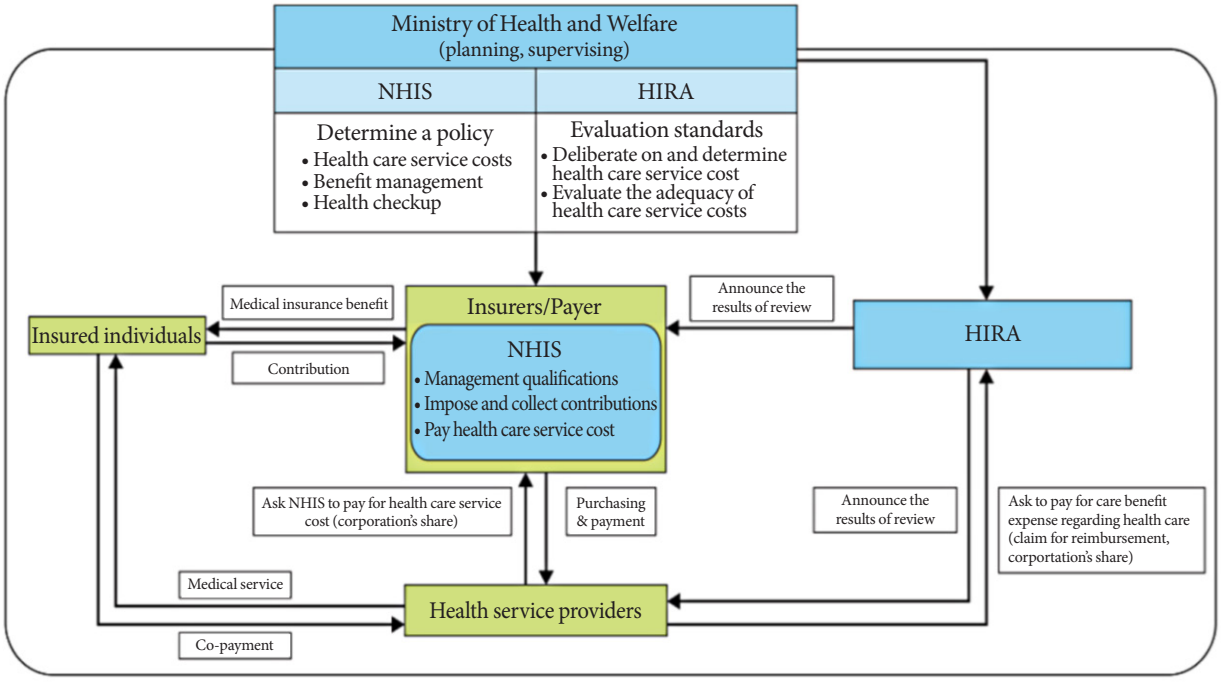

Fig. 1. Operational structure of National Health Insurance System (NHIS). Reproduced from Kim et al. [4]. HIRA, Health Insurance Review & Assessment Service.

ployee to the NHIS is determined based on wages, and that of a self-employed person is calculated from age, gender, household income, property, and owned vehicles. National Health Insurance, Medical Aid, and Long-term Care Insurance are the main health care programs that universally cover the Korean population. Approximately 97% of the population is enrolled in the National Health Insurance program, and 3% of the population is covered by medical aid programs [1,4,5].

HEALTH SCREENING POLICY

Health screening is performed to improve the health of citizens and reduce health care costs through the prevention of cardiovascular and cerebrovascular diseases and early detection of major cancers. Heads and members of regional policyholders aged 20 years or older are recommended to undergo health screening once every 2 years. All employees engaged in office work and employee dependents are also recommended to undergo health screening once every 2 years. Employment-based policyholders engaged in non-office work must undergo health screening annually [2]. Cancer screening includes tests for stomach, liver, colorectal, breast, cervical, and lung cancers. The starting age of screening and test intervals is different for different types of cancers. All fees for general health screenings were charged to the NHIC. For cancer screening, 90% was charged to the NHIC, and 10% was co-paid by the examinee. However, all fees for medical aid beneficiaries are charged to national or local governments [2]. The number of participants who underwent health screening in the last 10 years is shown in Table 1. The number of eligible individuals and actual examinees has increased gradually, and approximately 15 million people participate in health examinations every year. The rate of general health screening was approximately 75% in the last 10 years but 67.8% in 2020 [6], possibly due to the coronavirus disease 2019 (COVID-19) pandemic.

TYPES AND VARIABLES COMPRISING NHID

There are two types of research databases. A customized database refers to health information data collected, managed, and

Table 1. Number of eligible individuals and actual examinees of health examination in recent 10 years

Variable	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
No. of eligible individuals	15,249,528	15,673,188	15,775,891	16,456,214	17,356,727	17,633,406	17,818,302	19,593,149	21,716,582	21,446,220
No. of actual examinees										
Total no. (%)	11,070,569 (72.6)	11,419,350 (72.9)	11,381,295 (72.1)	12,301,581 (74.8)	13,213,329 (76.1)	13,709,413 (77.7)	13,987,129 (78.5)	15,076,899 (76.9)	16,098,417 (74.1)	14,544,980 (67.8)
Sex										
Men	6,117,787	6,277,362	6,258,804	6,716,277	7,152,110	7,360,929	7,470,196	8,106,914	8,395,046	7,659,607
Women	4,952,782	5,141,988	5,122,491	5,585,304	6,061,219	6,348,484	6,516,933	6,969,985	7,703,371	6,885,373
Age, yr										
≤19	22,066	25,852	30,395	28,855	27,898	27,698	25,498	21,548	16,162	13,126
20–24	292,806	289,877	310,544	320,157	331,153	348,864	340,926	337,873	544,396	525,980
25–29	959,981	861,405	879,338	886,824	906,928	974,937	972,343	1,008,398	1,144,773	1,095,797
30–34	1,161,993	1,181,946	1,206,389	1,232,766	1,271,907	1,203,259	1,166,903	1,195,162	1,340,699	1,235,064
35–39	1,070,355	1,083,236	1,020,708	1,139,037	1,193,888	1,231,963	1,267,513	1,335,464	1,385,978	1,209,388
40–44	1,238,902	1,274,646	1,304,791	1,330,964	1,446,585	1,426,743	1,411,857	1,839,238	1,919,130	1,656,855
45–49	1,329,572	1,361,423	1,371,396	1,512,407	1,653,299	1,729,097	1,751,848	1,732,167	1,767,840	1,520,351
50–54	1,661,191	1,759,631	1,647,344	1,801,231	1,885,250	1,895,002	1,907,258	1,965,960	2,055,588	1,848,045
55–59	1,062,443	1,152,283	1,204,758	1,337,416	1,492,845	1,586,881	1,644,551	1,662,173	1,648,391	1,497,048
60–64	972,055	1,053,108	1,004,503	1,162,690	1,285,409	1,456,209	1,551,359	1,597,421	1,780,520	1,614,717
65–69	333,237	322,477	318,096	362,290	451,578	455,019	490,695	868,891	860,339	900,574
70–74	594,159	648,373	638,440	684,102	687,162	756,759	764,036	778,593	850,860	757,803
75–79	226,827	245,203	267,378	293,277	334,352	343,885	387,835	416,163	414,459	372,947
80–84	114,366	128,812	139,344	165,798	193,357	217,859	241,803	253,020	292,102	234,836
≥85	30,616	31,078	37,871	43,767	51,718	55,238	62,704	64,828	77,180	62,449

maintained by the NHIC to be modified as requested for policy and academic research. The sample research database refers to the data standardized by extracting the sample to improve the limited access and use by investigators owing to the large size and personal, identifiable information issues. The sample cohort, medical check-up, elderly cohort, working women cohort, and infant medical check-up databases are available as sample research databases and allow long-term observation of the same individuals as a cohort [3]. The most recent sample cohort database includes one million people sampled based on data from 2006, which is approximately 2% of the total population (48,222,537). Stratified random sampling was used from 2,142 (2×17×21×3) strata constructed by sex (male and female: two groups), age (5-year age groups between 1 and 79 and 80 years and above: 17 groups), eligibility and contribution (deciles of regional policyholders, deciles of employment-based policyholders, and medical aid beneficiaries: 21 groups), and region (big city, middle or small cities, and rural areas: three groups) [7].

The NHID includes qualification, treatment, medical check-up, and clinic tables. Variables included in the qualification table are age, sex, location, type of subscription, and socioeconomic statuses, such as income rank, disability, and death. The cause of death was determined upon request in the sample cohort database. The treatment table is composed of a database including statements (T20), details of treatment (T30), type of disease (T40), and details of prescription (T60) on the data from medical institutions, dental, oriental, and pharmacy [7,8]. The details of the variables included in each table of the sample cohort database are presented in Table 2.

The variables included in the health examination and questionnaire were changed over time (Table 3). Fifty-one variables were included in 2002–2008, 57 variables in 2009–2017, and 108 variables in 2018–2019. Currently, the parameters measured using blood tests include fasting blood glucose (FBG), total cholesterol, triglyceride, high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), hemoglobin, creatinine, aspartate aminotransferase, alanine aminotransferase, and gamma-glutamyl transferase. The increase in the number of variables in 2018 to 2019 was mainly due to the detailed questionnaire on smoking and alcohol consumption habits [7].

STRENGTHS AND LIMITATIONS OF THE NHID

There are several strengths and limitations that should be carefully considered when using the NHID. The most powerful

Table 2. Variables included in the Korean National Health Information sample cohort database

Qualification table	Year of construction, individual unique number, age, sex, location, type of subscription, deciles of contribution, type of disability, severity of disability, eligibility of medical check-up, sample type
Birth and death table	Year of birth, date of death, cause of death
Treatment table	
Statement (T20)	Start date of medical care, medical subject code, principal diagnosis, additional diagnosis, first date of hospitalization, route of hospitalization, official injury, operation (yes/no), days of medical care, days of hospital visit, days of total prescription, result of medical care, medical expenses (cost paid by insurer, cost paid by beneficiaries)
Treatment details (T30)	Start date of medical care, classification and item of specification, code of medical care classification, dosage and frequency of medication or procedure, type of medical expense, unit price, total cost, drug classification
Type of disease (T40)	Start date of medical care, medical subject code, principal diagnosis, additional diagnosis, ruled-out diagnosis
Prescription details (T60)	Start date of medical care, code of medication, drug classification, dosage, total days of administration, cost of medication
Medical check-up table	Anthropometry, blood pressure, vision, hearing ability, blood test (fasting glucose, lipid levels, hemoglobin, creatinine, estimated glomerular filtration rate, aspartate aminotransferase, alanine aminotransferase, gamma-glutamyl transferase), chest radiography, electrocardiogram, past medical history, family history, questionnaires (smoking, alcohol consumption, exercise)
Clinic table	Institution classification code, address of institution, subject type, numbers of doctors, nurses, beds for admission, beds for operation, and beds for emergency room
Elderly long-term nursing table	General information and rating result of application, claim specification, status of long-term nursing facility

Table 3. Variables and questionnaires included in the health examination database

Classification	Variable	Year of health examination		
		2002–2008	2009–2017	2018–2019
Health examination				
Obesity	Height	○	○	○
	Weight	○	○	○
	Body mass index	○	○	○
	Waist circumference		○[a]	○
Hypertension	Systolic blood pressure	○	○	○
	Diastolic blood pressure	○	○	○
Sensory	Vision	○	○	○
	Hearing ability	○	○	○
Diabetes	Fasting glucose	○	○	○
Hypertension, dyslipidemia, atherosclerosis	Total cholesterol	○	○	○
	Triglyceride		○	○
	HDL-cholesterol		○	○
	LDL-cholesterol		○	○
Anemia	Hemoglobin	○	○	○
Kidney disease	Urine glucose	○		
	Urine occult blood	○		
	Urine pH	○		
	Urine protein	○	○	○
Chronic kidney disease	Serum creatinine		○	○
	Estimated glomerular filtration rate		○[b]	○
Liver disease	Aspartate aminotransferase	○	○	○
	Alanine aminotransferase	○	○	○
	Gamma-glutamyl transferase	○	○	○
Pulmonary disease	Chest radiography	○	○	○
Cardiac disease	Electrocardiogram	○		
Questionnaire				
Past medical history		○[c]	○[d]	○[d]
Family history		○[e]	○[f]	○[f]
Smoking	Smoking status	○	○	○
	Daily smoking amount	○		
	Average daily smoking amount (ex-smoker)		○	○
	Average daily smoking amount (current smoker)		○	○
	Smoking duration	○		
	Smoking duration (ex-smoker)		○	○
	Smoking duration (current smoker)		○	○
Alcohol consumption	Drinking frequency	○		○
	Days of drinking per week		○	
	Amount of drinking per time	○		
	Amount of drinking per day		○	○
	Type of alcohol			○
	Maximum amount of drinking per day			○
Exercise	Exercise frequency per week	○		
	Days of strenuous exercise per week		○	○
	Time of strenuous exercise per day			○
	Days of moderate intensity exercise per week		○	○
	Time of moderate intensity exercise per day			○
	Days of walking exercise per week		○	
	Days of strength training per week			○
Hepatitis B	Hepatitis B		○	○

HDL, high-density lipoprotein; LDL, low-density lipoprotein.
[a]Waist circumference measurement was started in 2008, [b]Estimated glomerular filtration rate measurement was not performed in 2010 to 2011, [c]Past medical history, development year, cured or not on pulmonary tuberculosis, hepatitis, liver disease, hypertension, cardiac disease, stroke, diabetes, cancer, and other disease, [d]Past medical history and medical treatment of stroke, cardiac disease, hypertension, diabetes, dyslipidemia, pulmonary tuberculosis, cancer, and other disease, [e]Family history of liver disease, hypertension, stroke, cardiac disease, diabetes, and cancer, [f]Family history of hypertension, stroke, cardiac disease, diabetes, cancer, and other disease.

strength is the number of individuals included in the database. Since the NHIC is a single insurer that manages the National Health Insurance System in Korea, virtually all Koreans (approximately 50 million) are enrolled in this program; therefore, the NHID could represent the entire Korean population. Furthermore, the health screening policy described above enables the accumulation of medical check-up data, including anthropometric measurements, past medical history, family history, laboratory data, and detailed questionnaires on lifestyle factors. Combining the claims database with health examination data makes the Korean NHID unique. Mortality data from Statistics Korea or other databases can be linked using resident registration numbers for wider application of the database [9]. Given the large size of the database, it can be utilized to study rare diseases or rare complications of treatment and to study specific populations such as the elderly group [10]. One example is an observational study on acromegaly and cardiovascular outcomes, which included 1,874 patients with acromegaly [11]. It is also appropriate for long-term follow-up owing to the longitudinal nature of the database.

Since the data were not collected for research purposes, it is difficult to define causal relationships when performing outcome studies. The main purpose of establishing this database was to record claims and reimbursements; therefore, data on medications or procedures not covered by the NHIS are not available. Also, information on the severity of medical conditions are lacking and it is hard to reflect the health behaviors of beneficiaries. In addition, discrepancies may exist between the diagnosis encoded to claim medical bills and the actual disease. Therefore, setting an appropriate operational definition and validation may be crucial. As shown in Table 1, not all eligible individuals undergo health check-ups, which may impose a possibility of selection bias. Importantly, the linkage between the NHID and the electronic medical records of each hospital is very limited due to legal and privacy issues. Solving this problem might lead to a new leap forward in research using the NHID [12].

To request data from the National Health Insurance Sharing Service (http://nhiss.nhis.or.kr), researchers must obtain approval of the study protocol from the institutional review board and the data provision review committee at the NHIC. Access to and analysis of the NHID can be performed only in designated places, and the raw data cannot be retrieved from the server. Only the analyzed data can be obtained after approval. However, remote access and analysis are available for sample research databases. Recently, the process of requesting and reviewing data applications has taken a long time owing to the great increase in the number of researchers interested in the NHID.

TRENDS OF RESEARCH USING NHID

Since the establishment of the NHID, research and publications based on this database have increased significantly (Fig. 2). We searched PubMed using the keyword 'NHIS' or 'National Health Insurance System' or 'NHID' and 'Korea' or 'Korean.' A total of 1,692 published articles were identified. Among these, 595 articles (35.2%) were on diabetes, metabolism, metabolic

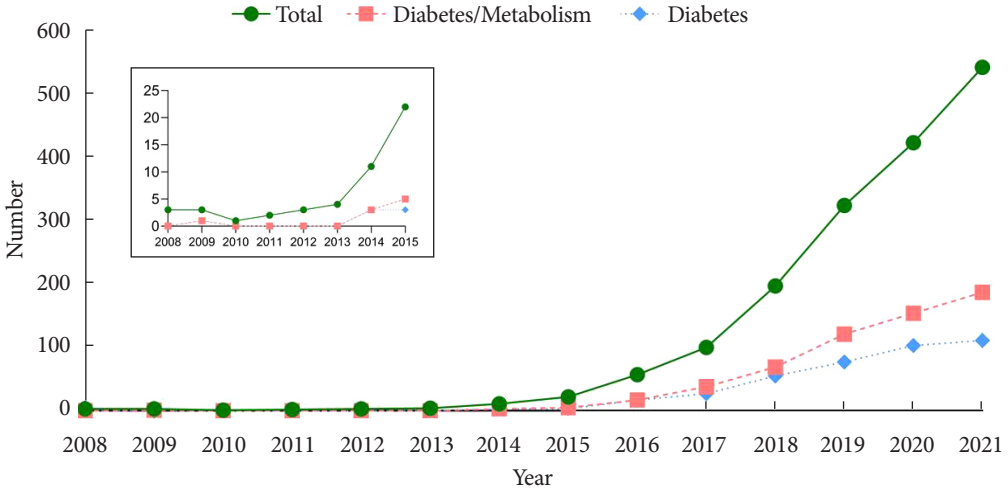

Fig. 2. The number of publications using National Health Information database from 2008 to 2021.

syndrome (MetS), obesity, lipids, and cholesterol. A total of 397 articles were identified using the keyword 'diabetes.'

THE OPERATIONAL DEFINITION OF OUTCOMES IN DIABETES AND METABOLISM RESEARCH

Type 2 diabetes mellitus

In research using the NHID, the operational definition of diabetes was applied considering the characteristics of the database. The proportion of patients with diabetes was 13.2% according to the International Classification of Disease, 10th revision (ICD-10) codes (E11-14) alone, and 8.7% based on prescription data alone in 2013 [13]. The Taskforce Team of Diabetes Fact Sheet of the Korean Diabetes Association concluded the operational definition of diabetes as either (1) patients who had both recordings of diagnosis (ICD-10 codes E11-14 for diabetes as either principal diagnosis or 1st to 4th additional diagnosis at least once a year) and prescription of anti-diabetic drugs; or (2) patients whose FBG levels from health check-up data were ≥126 mg/dL (Table 4) [13]. According to this definition, the prevalence of diabetes was 11.4% in 2013.

Dyslipidemia

The presence of dyslipidemia was defined as the presence of at least one claim per year under ICD-10 code E78 and at least one claim per year for the prescription of lipid-lowering agents or total cholesterol ≥240 mg/dL (Table 4) [14-16]. Lipid-lowering drugs include statins, ezetimibe, and fibrates. In addition to this operational definition, dyslipidemia was defined using health check-up data and prescriptions of lipid-lowering drugs. Hypercholesterolemia was defined as a total cholesterol level ≥240 mg/dL or the use of a lipid-lowering drug. Hyper-LDL cholesterolemia was defined as a serum LDL-C level ≥160 mg/dL or the use of a lipid-lowering drug. Hypo-HDL cholesterolemia was defined as a serum HDL-C level <40 mg/dL. Hypertriglyceridemia was defined as a serum triglyceride ≥200 mg/dL. In the Korea Dyslipidemia Fact Sheet 2020, dyslipidemia was defined as satisfying one of the definitions for LDL-C, HDL-C, or triglyceride as stated above [17].

Hypertension

The presence of hypertension was defined as the presence of at least one claim per year under ICD-10 codes I10 or I11 and at least one claim per year for the prescription of antihypertensive agents or systolic blood pressure (BP) ≥140 mm Hg or diastolic BP ≥90 mm Hg (Table 4) [18,19]. Other studies have used a different operational definition of hypertension: ICD-10 codes I10-I13 or I15 for the hypertensive disease usually recorded twice in the outpatient clinic, or once during hospitalization, and a history of prescription of antihypertensive drugs [20,21]. This definition includes hypertensive end-organ damage, such as hypertensive renal disease (I12), hypertensive heart and renal disease (I13), and secondary hypertension

Table 4. The operational definitions of commonly used outcomes and covariates in the field of diabetes and metabolism research

	ICD-10 codes and additional definitions		General health check-up results
Type 2 diabetes mellitus	E11-14	Recording as either principal diagnosis or 1st to 4th additional diagnosis at least once a year and prescription of anti-diabetic drugs	Fasting blood glucose ≥126 mg/dL
Dyslipidemia	E78	Recording at least once a year and prescription of lipid-lowering agents (statin, ezetimibe, fenofibrate)	Total cholesterol ≥240 mg/dL
Hypertension	I10-I11	Recording at least once a year and prescription of antihypertensive agents	Systolic blood pressure ≥140 mm Hg or diastolic blood pressure ≥90 mm Hg
Myocardial infarction	I21, I22	Recording at admission ≥1	
Ischemic stroke	I63, I64	Recording at admission ≥1 with claims for the imaging studies (brain CT or MRI)	
Heart failure	I50	Recording at admission or outpatient clinic ≥1	
Chronic kidney disease	N18, N19	Recording at admission ≥1 or outpatient clinic ≥2	eGFR <60 mL/min/1.73 m^2
End-stage renal disease	N18–N19, Z49, Z94.0, Z99.2	Dialysis ≥30 days or kidney transplantation during hospitalization	

ICD-10, International Classification of Disease, 10th revision; CT, computed tomography; MRI, magnetic resonance imaging; eGFR, estimated glomerular filtration rate.

(I15). In the Korea Hypertension Fact Sheet 2020, hypertension is defined as at least one health insurance claim for the diagnosis of essential hypertension (I10) each year [22].

Myocardial infarction and stroke

Myocardial infarction (MI) was defined according to ICD-10 codes I21 or I22 recorded during hospitalization [23,24]. Stroke was defined using principal diagnosis codes from I60 to I64 with the enforcement of brain computerized tomography or magnetic resonance imaging at the emergency center or outpatient clinic, or during hospitalization [25]. Ischemic stroke was defined as a recording of ICD-10 codes I63 or I64 during hospitalization with a claim for brain magnetic resonance imaging or brain computerized tomography (Table 4) [23,24]. This definition has been widely adopted in previous studies using claims databases [26,27]. According to the validation of diagnostic codes of clinical outcomes in the NHID, the primary discharge diagnostic codes for MI (ICD-10 codes I21 and I22) showed favorable reliability, with a positive predictive value (PPV) of 92% [28]. In stroke and intracranial hemorrhage (ICH), in addition to the primary discharge diagnostic codes, consideration of relevant clinical information, such as hospitalization duration, imaging studies, and prescription of antithrombotic agents, could improve the accuracy of diagnosis. For ischemic stroke (ICD-10 codes I63 and I64) and ICH (ICD-10 I60-62), the combination of primary diagnostic codes during hospitalization and brain imaging studies showed a PPV and sensitivity of 92.2% and 91.2%, respectively [28]. For ICH, the combination of primary diagnostic codes with hospitalization and brain imaging studies showed a PPV and sensitivity of 81.4% and 95.1%, respectively [28].

Heart failure

Heart failure was defined using ICD-10 code I50 with more than one diagnosis during hospitalization or in an outpatient clinic (Table 4) [29]. Another study defined heart failure as ICD-10 I50 during hospitalization [30].

Chronic kidney disease and end-stage renal disease

Chronic kidney disease (CKD) was defined using the ICD-10 codes N18 or N19 and an estimated glomerular filtration rate of <60 mL/min/1.73 m^2 was calculated using the CKD Epidemiology Collaboration Equation on more than two occasions during the medical check-up [31,32]. End-stage renal disease was defined using a combination of ICD-10 codes (N18–N19, Z49, Z94.0, Z99.2) and initiation of renal replacement therapy for 30 days or more, and/or kidney transplantation during hospitalization (Table 4) [33].

REPRESENTATIVE STUDIES RELATED TO DIABETES AND METABOLISM USING THE KOREAN NHID

Diabetes Fact Sheet in Korea 2021

The representative national estimates of diabetes in Korea can be analyzed using the Korea National Health and Nutrition Examination Survey (KNHANES) and the Korea NHID [34]. Among Korean adults aged ≥30 years, the estimated prevalence of diabetes was 16.7% in 2020. The proportion of adults with diabetes who achieved a glycosylated hemoglobin target of <6.5% was 24.5%. The prescription patterns of anti-diabetic drugs were analyzed. It was reported that 86.0% of adults with previously diagnosed diabetes were taking oral glucose-lowering medications without insulin, and 7.5% were treated with insulin. Sulfonylurea was the most commonly used drug, followed by metformin in 2002. During the past decade, the use of metformin has increased steadily to 86% of total antidiabetic drug prescriptions and metformin was the most frequently prescribed antidiabetic agent in Korea in 2018. The use of dipeptidyl peptidase-4 (DPP-4) inhibitors increased markedly after their release in 2008 and dramatically increased to 62.0% in 2018. There was a steady decrease in the use of sulfonylureas/glinides, from 84% in 2002 to 43% in 2018. The use of insulin and thiazolidinediones remained stable from 2002 to 2018 [34].

Gestational diabetes mellitus in Koreans

The clinical characteristics of gestational diabetes mellitus (GDM) in Korea have been reported using a large-scale population dataset from the NHID [35]. The prevalence of GDM in Korean women between 2011 and 2015 was 12.7 %. The operational definition of GDM was as follows: visited the outpatient clinic more than twice with GDM codes and no previous history of diabetes; did not have a claim for diabetes based on ICD-10 codes E10-14 or oral antidiabetic drug or insulin before pregnancy; did not have an FBG level ≥126 mg/dL before pregnancy. The incidence rate of GDM increases with advancing age, pre-pregnancy body mass index, waist circumference, and FBG level [35].

Cholesterol and BP levels and development of cardiovascular disease in Koreans with type 2 diabetes mellitus

In recent guidelines, cholesterol targets are based on several primary- and secondary-prevention statin trials that have shown improved outcomes with more intensive LDL-C lowering. In addition to randomized controlled trials (RCTs), the optimal lipid or BP levels to prevent cardiovascular disease (CVD) could be investigated through big data analysis. Patients with type 2 diabetes mellitus over 40 years of age without CVD were divided into statin users and non-users, and the relationship between LDL-C levels and the risk of CVD was analyzed [36]. There was an increased risk of CVD in individuals with an LDL-C level ≥130 mg/dL among those with type 2 diabetes mellitus who did not take statins. The risk of CVD was significantly higher in those taking statins with an LDL-C level of ≥70 mg/dL. The researchers recommended statin therapy for the primary prevention of CVD, with a target LDL-C level of <70 mg/dL [36].

The relationship between BP and CVD risk in patients with type 2 diabetes mellitus without CVD was analyzed. Systolic BP 130 to 139 mm Hg was associated with a significant increase in the incidence of stroke (hazard ratio [HR], 1.15; 95% confidence interval [CI], 1.12 to 1.18) and MI (HR, 1.05; 95% CI, 1.02 to 1.09) compared to systolic BP 110 to 119 mm Hg [18]. Subjects with a diastolic BP of 80 to 84 mm Hg had a higher risk of CVD than those with a diastolic BP of 75 to 79 mm Hg. The overall relationship between BP and CVD risk was positive, with greater strength observed in the younger age groups. The optimal cutoff for Korean patients with type 2 diabetes mellitus associated with a lower CVD risk may be 130 mm Hg for systolic BP or 80 mm Hg for diastolic BP [18]. Another study examined the association of BP categories before age of 40 years with the risk of CVD later in life. In both young men and women, stage 1 hypertension (systolic BP 130 to 139 mm Hg; diastolic BP 80 to 89 mm Hg) and stage 2 hypertension (systolic BP ≥140 mm Hg; diastolic BP ≥90 mm Hg) were associated with increased risk of CVD, coronary heart disease, and stroke [37].

Risk of cardiovascular events and death associated with the initiation of sodium-glucose co-transporter-2 inhibitors compared with DPP-4 inhibitors: CVD-REAL 2 multinational cohort study

This study utilized data sourced from de-identified health records in 13 different countries located in four geographical regions, which could be linked to CVD outcomes and mortality data [38]. Information from the Korean NHID was used. All initial episodes of new initiation of either sodium-glucose co-transporter-2 (SGLT2) inhibitors or DPP-4 inhibitors were selected. The use of a new SGLT2 inhibitor was associated with a substantially lower risk of hospitalization for heart failure (HR, 0.69; 95% CI, 0.61 to 0.77) and death (HR, 0.59; 95% CI, 0.52 to 0.67). The risks of MI and stroke were also significantly lower with SGLT2 inhibitors than with DPP-4 inhibitors [38]. A large number of patients, the consistency of the findings across 13 countries with different healthcare systems, the inclusion of different SGLT2 inhibitors and DPP-4 inhibitors, and the exclusion of anyone who had been on a DPP-4 inhibitor or SGLT2 inhibitor for at least a year before follow-up started all contribute to the robustness and credibility of these findings [39]. In contrast to clinical trials conducted in highly selected populations, real-world evidence (RWE) can be generalized to so-called average patients with type 2 diabetes mellitus.

Use of fenofibrate on cardiovascular outcomes in statin users with MetS

Recently, RWE analysis has been conducted using a large-scale population-based cohort. The value of RWE begins with the limitations of RCTs. RCTs provide the highest level of evidence in medical science but the inevitable limitations of RCTs include limited patient populations and the trial environment, which is difficult to reproduce in the real world [5,40]. The potential role of fenofibrate in cardiovascular risk reduction was analyzed using the Korean NHID [41]. Early clinical trials on fibrates are promising, but their role in CVD risk management has gradually diminished in the statin era. Using the Korean National Health Insurance Service-Health Screening Cohort, researchers attempted to demonstrate the additional benefits of fenofibrate add-on to statins [41]. Patients with MetS were included in the study. Propensity score matching was performed for those treated with fenofibrate plus statins and those treated with statins only. The risk of composite CVD, including coronary heart disease, ischemic stroke, and cardiovascular mortality, was significantly reduced in the combined treatment group compared with the statin-only group (adjusted HR, 0.74; 95% CI, 0.58 to 0.93; $P=0.01$). In particular, the HRs of composite CVD were lower in those with high triglyceride or low HDL-C (adjusted HR, 0.64; 95% CI, 0.47 to 0.87; $P=0.005$) compared with those with low triglyceride and high HDL-C. This study

may influence treatment guidelines for the benefit of fenofibrate in improving residual cardiovascular risk in patients with dyslipidemia during statin use.

Altered risk for cardiovascular events with changes in the MetS status

The KNHANES data, a nationally representative sample of Korea, is limited in that longitudinal follow-up data for the same subjects cannot be obtained. In contrast, the NHID contains serial data of the same individuals who undergo regular health examinations. In this regard, noteworthy studies have utilized serial data from the Korea NHID to examine the cumulative effect, variability, or changes in metabolic parameters [14-16,19,23,26,33]. An example is a study that showed an altered risk of cardiovascular events with changes in the MetS status [42]. Among those who had undergone three or more health examinations, 72.7%, 15.6%, 6.1%, and 5.6% were in the MetS-free, MetS-chronic, MetS-developed, and MetS-recovery groups, respectively. At a median follow-up of 3.5 years, the MetS-recovery group had a significantly lower major adverse cardiovascular event (MACE) risk than the MetS-chronic group (adjusted incidence rate ratio [IRR], 0.85; 95% CI, 0.83 to 0.87). The MetS-developed group had a significantly higher MACE risk than the MetS-free group (adjusted IRRs, 1.36; 95% CI, 1.33 to 1.39). Among the MetS criteria, the development of the elevated BP criterion was related to the largest increase in MACE. Healthcare providers may consider these results when planning a public health strategy to alleviate the burden of MACE.

CONCLUSIONS

In this review, we have summarized the history, structure, and characteristics of the Korean NHID. Recent trends in big data research using this database and representative studies have been introduced. Due to the purpose and nature of this database, some limitations exist. However, several strengths also highlight the value of this database. A careful study design and analysis of real-world big data may produce valuable information that can complement other forms of research. In the future, institutional support for the linkage between the NHID and other forms of databases would be crucial for the expansion of usability.

CONFLICTS OF INTEREST

Seung-Hwan Lee has been associate editors of the *Diabetes & Metabolism Journal* since 2022. He was not involved in the review process of this review. Otherwise, there was no conflict of interest.

ORCID

Mee Kyoung Kim *https://orcid.org/0000-0003-3205-9114*
Seung-Hwan Lee *https://orcid.org/0000-0002-3964-3877*

FUNDING

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (Grant Number: HI18-C0275).

ACKNOWLEDGMENTS

None

REFERENCES

1. National Health Insurance Service: 2020 National Health Insurance statistical yearbook. Available from: https://www.nhis.or.kr/nhis/together/wbhaec06300m01.do?mode=view&articleNo=10812384&article.offset=0&articleLimit=10 (updated 2021 Nov 5).
2. National Health Insurance Service: 2020 National Health Screening statistical yearbook. Available from: https://www.nhis.or.kr/nhis/together/wbhaec07000m01.do?mode=view&articleNo=10813922&article.offset=0&articleLimit=10 (updated 2021 Dec 30).
3. National Health Insurance Sharing Service: Introduction. Available from: https://nhiss.nhis.or.kr/bd/ab/bdaba012eng.do (cited 2022 Jul 5).
4. Kim HK, Song SO, Noh J, Jeong IK, Lee BW. Data configuration and publication trends for the Korean National Health Insurance and Health Insurance Review & Assessment Database. Diabetes Metab J 2020;44:671-8.
5. Choi EK. Cardiovascular research using the Korean National Health Information Database. Korean Circ J 2020;50:754-72.

6. Korean Statistical Information Service: Health examination statistics. Available from: https://kosis.kr/statisticsList/statisticsListIndex.do?menuId=M_01_01&vwcd=MT_ZTITLE&parmTabId=M_01_01&outLink=Y&entrType=#content-group (cited 2022 Jul 5).
7. National Health Insurance Sharing Service: Sample cohort 2.2 database user manual. Available from: https://nhiss.nhis.or.kr/bd/ab/bdaba002cv.do (cited 2022 Jul 5).
8. Lee J, Lee JS, Park SH, Shin SA, Kim K. Cohort profile: The National Health Insurance Service-National Sample Cohort (NHIS-NSC), South Korea. Int J Epidemiol 2017;46:e15.
9. Bahk J, Kim YY, Kang HY, Lee J, Kim I, Lee J, et al. Using the National Health Information Database of the National Health Insurance Service in Korea for monitoring mortality and life expectancy at national and local levels. J Korean Med Sci 2017;32:1764-70.
10. Kim YI, Kim YY, Yoon JL, Won CW, Ha S, Cho KD, et al. Cohort profile: National Health Insurance Service-Senior (NHIS-Senior) cohort in Korea. BMJ Open 2019;9:e024344.
11. Hong S, Kim KS, Han K, Park CY. Acromegaly and cardiovascular outcomes: a cohort study. Eur Heart J 2022;43:1491-9.
12. Kyoung DS, Kim HS. Understanding and utilizing claim data from the Korean National Health Insurance Service (NHIS) and Health Insurance Review & Assessment (HIRA) Database for research. J Lipid Atheroscler 2022;11:103-10.
13. Lee YH, Han K, Ko SH, Ko KS, Lee KU; Taskforce Team of Diabetes Fact Sheet of the Korean Diabetes Association. Data analytic process of a nationwide population-based study using National Health Information Database established by National Health Insurance Service. Diabetes Metab J 2016;40:79-82.
14. Lee EY, Han K, Kim DH, Park YM, Kwon HS, Yoon KH, et al. Exposure-weighted scoring for metabolic syndrome and the risk of myocardial infarction and stroke: a nationwide population-based study. Cardiovasc Diabetol 2020;19:153.
15. Kim MK, Han K, Kim HS, Park YM, Kwon HS, Yoon KH, et al. Cholesterol variability and the risk of mortality, myocardial infarction, and stroke: a nationwide population-based study. Eur Heart J 2017;38:3560-6.
16. Lee HJ, Choi EK, Han KD, Lee E, Moon I, Lee SR, et al. Bodyweight fluctuation is associated with increased risk of incident atrial fibrillation. Heart Rhythm 2020;17:365-71.
17. Cho SM, Lee H, Lee HH, Baek J, Heo JE, Joo HJ, et al. Dyslipidemia fact sheets in Korea 2020: an analysis of nationwide population-based data. J Lipid Atheroscler 2021;10:202-9.
18. Kim MK, Han K, Koh ES, Kim ES, Lee MK, Nam GE, et al. Blood pressure and development of cardiovascular disease in Koreans with type 2 diabetes mellitus. Hypertension 2019;73:319-26.
19. Cho Y, Han K, Kim DH, Park YM, Yoon KH, Kim MK, et al. Cumulative exposure to metabolic syndrome components and the risk of dementia: a nationwide population-based study. Endocrinol Metab (Seoul) 2021;36:424-35.
20. Lee SR, Choi EK, Kwon S, Jung JH, Han KD, Cha MJ, et al. Oral anticoagulation in Asian patients with atrial fibrillation and a history of intracranial hemorrhage. Stroke 2020;51:416-23.
21. Lee E, Choi EK, Han KD, Lee H, Choe WS, Lee SR, et al. Mortality and causes of death in patients with atrial fibrillation: a nationwide population-based study. PLoS One 2018;13:e0209687.
22. Kim HC, Lee H, Lee HH, Seo E, Kim E, Han J, et al. Korea hypertension fact sheet 2021: analysis of nationwide population-based data with special focus on hypertension in women. Clin Hypertens 2022;28:1.
23. Lee SH, Han K, Kwon HS, Kim MK. Frequency of exposure to impaired fasting glucose and risk of mortality and cardiovascular outcomes. Endocrinol Metab (Seoul) 2021;36:1007-15.
24. Lee SH, Han K, Kim HS, Cho JH, Yoon KH, Kim MK. Predicting the development of myocardial infarction in middle-aged adults with type 2 diabetes: a risk model generated from a nationwide population-based cohort study in Korea. Endocrinol Metab (Seoul) 2020;35:636-46.
25. Kim JY, Kang K, Kang J, Koo J, Kim DH, Kim BJ, et al. Executive summary of stroke statistics in Korea 2018: a report from the Epidemiology Research Council of the Korean Stroke Society. J Stroke 2019;21:42-59.
26. Kim MK, Han K, Park YM, Kwon HS, Kang G, Yoon KH, et al. Associations of variability in blood pressure, glucose and cholesterol concentrations, and body mass index with mortality and cardiovascular outcomes in the general population. Circulation 2018;138:2627-37.
27. Kim MK, Han K, Cho JH, Kwon HS, Yoon KH, Lee SH. A model to predict risk of stroke in middle-aged adults with type 2 diabetes generated from a nationwide population-based cohort study in Korea. Diabetes Res Clin Pract 2020;163:108157.
28. Park J, Kwon S, Choi EK, Choi YJ, Lee E, Choe W, et al. Validation of diagnostic codes of major clinical outcomes in a National Health Insurance database. Int J Arrhythm 2019;20:5.
29. Lee HJ, Kim HK, Han KD, Lee KN, Park JB, Lee H, et al. Age-dependent associations of body mass index with myocardial infarction, heart failure, and mortality in over 9 million Koreans.

Eur J Prev Cardiol 2022 May 17 [Epub]. https://doi.org/10.1093/eurjpc/zwac094.

30. Ahn HJ, Lee SR, Choi EK, Han KD, Jung JH, Lim JH, et al. Association between exercise habits and stroke, heart failure, and mortality in Korean patients with incident atrial fibrillation: a nationwide population-based cohort study. PLoS Med 2021;18:e1003659.
31. Han SJ, Ha KH, Lee N, Kim DJ. Effectiveness and safety of sodium-glucose co-transporter-2 inhibitors compared with dipeptidyl peptidase-4 inhibitors in older adults with type 2 diabetes: a nationwide population-based study. Diabetes Obes Metab 2021;23:682-91.
32. Bae EH, Lim SY, Jung JH, Oh TR, Choi HS, Kim CS, et al. Chronic kidney disease risk of isolated systolic or diastolic hypertension in young adults: a nationwide sample based-cohort study. J Am Heart Assoc 2021;10:e019764.
33. Koh ES, Han KD, Kim MK, Kim ES, Lee MK, Nam GE, et al. Changes in metabolic syndrome status affect the incidence of end-stage renal disease in the general population: a nationwide cohort study. Sci Rep 2021;11:1957.
34. Bae JH, Han KD, Ko SH, Yang YS, Choi JH, Choi KM, et al. Diabetes fact sheet in Korea 2021. Diabetes Metab J 2022;46:417-26.
35. Kim KS, Hong S, Han K, Park CY. The clinical characteristics of gestational diabetes mellitus in Korea: a National Health Information Database Study. Endocrinol Metab (Seoul) 2021;36:628-36.
36. Kim MK, Han K, Joung HN, Baek KH, Song KH, Kwon HS. Cholesterol levels and development of cardiovascular disease in Koreans with type 2 diabetes mellitus and without pre-existing cardiovascular disease. Cardiovasc Diabetol 2019;18:139.
37. Son JS, Choi S, Kim K, Kim SM, Choi D, Lee G, et al. Association of blood pressure classification in Korean young adults according to the 2017 American College of Cardiology/American Heart Association guidelines with subsequent cardiovascular disease events. JAMA 2018;320:1783-92.
38. Kohsaka S, Lam CS, Kim DJ, Cavender MA, Norhammar A, Jorgensen ME, et al. Risk of cardiovascular events and death associated with initiation of SGLT2 inhibitors compared with DPP-4 inhibitors: an analysis from the CVD-REAL 2 multinational cohort study. Lancet Diabetes Endocrinol 2020;8:606-15.
39. Gerstein HC. Patient data from routinely collected medical records complement evidence from SGLT2 inhibitor outcome trials. Lancet Diabetes Endocrinol 2020;8:557-8.
40. Kim NH, Kim SG. Fibrates revisited: potential role in cardiovascular risk reduction. Diabetes Metab J 2020;44:213-21.
41. Kim NH, Han KH, Choi J, Lee J, Kim SG. Use of fenofibrate on cardiovascular outcomes in statin users with metabolic syndrome: propensity matched cohort study. BMJ 2019;366:l5125.
42. Park S, Lee S, Kim Y, Lee Y, Kang MW, Han K, et al. Altered risk for cardiovascular events with changes in the metabolic syndrome status: a nationwide population-based study of approximately 10 million persons. Ann Intern Med 2019;171:875-84.