SPECIAL ARTICLE
Medical Informatics

JKMS

https://doi.org/10.3346/jkms.2017.32.5.718 • J Korean Med Sci 2017; 32: 718-728

Towards Actualizing the Value Potential of Korea Health Insurance Review and Assessment (HIRA) Data as a Resource for Health Research: Strengths, Limitations, Applications, and Strategies for Optimal Use of HIRA Data

Jee-Ae Kim,[1] Seokjun Yoon,[2] Log-Young Kim,[1] and Dong-Sook Kim[1]

[1]Pharmaceutical Policy Research Team, Health Insurance Review & Assessment Service, Wonju, Korea; [2]Department of Preventive Medicine, College of Medicine, Korea University, Seoul, Korea

Received: 4 September 2016
Accepted: 28 January 2017

Address for Correspondence:
Dong-Sook Kim, PhD
Department of Research, Health Insurance Review & Assessment Service, 60 Hyeoksin-ro, Wonju 26465, Republic of Korea
E-mail: sttone@hira.or.kr

Health Insurance and Review Assessment (HIRA) in South Korea, also called National Health Insurance (NHI) data, is a repository of claims data collected in the process of reimbursing healthcare providers. Under the universal coverage system, having fee-for-services covering all citizens in South Korea, HIRA contains comprehensive and rich information pertaining to healthcare services such as treatments, pharmaceuticals, procedures, and diagnoses for almost 50 million beneficiaries. This corpus of HIRA data, which constitutes a large repository of data in the healthcare sector, has enormous potential to create value in several ways: enhancing the efficiency of the healthcare delivery system without compromising quality of care; adding supporting evidence for a given intervention; and providing the information needed to prevent (or monitor) adverse events. In order to actualize this potential, HIRA data need to actively be utilized for research. Thus understanding this data would greatly enhance this potential. We introduce HIRA data as an important source for health research and provide guidelines for researchers who are currently utilizing HIRA, or interested in doing so, to answer their research questions. We present the characteristics and structure of HIRA data. We discuss strengths and limitations that should be considered in conducting research with HIRA data and suggest strategies for optimal utilization of HIRA data by reviewing published research using HIRA data.

Keywords: National Health Insurance; Claims Data; Health Research; Healthcare Services; Health Insurance Review and Assessment Service; HIRA Data; Korea

INTRODUCTION

Medical practices today strive to be evidenced-based. Randomized clinical trials (RCTs) are widely recognized as the type of studies that generally have the greatest level of quality with respect to generating evidence for medical decisions. However, due to some limitations of RTCs, quasi-experimental studies and observational studies are also important sources for providing evidence. One drawback of RCTs is that they usually have sample sizes that are often not large enough to capture rare events. RCTs also tend to under-represent vulnerable populations such as the elderly and those associated with "low income." Moreover, they usually require significant resources — such as time, labor, and funds — the needs for which tend to increase as target sample sizes increase. Despite their advantages in minimizing the risk of bias due to confounding, their generalizability vis-à-vis the general population may be limited because RCTs are usually conducted in highly controlled environments that are different from those in routine practice (1-3).

Observational studies with secondary health data can be an alternative to RCTs by addressing these shortcomings of RCTs, and in recent years, they have been grown (4). One kind of important secondary data is health insurance claims data, which are routinely collected for payment of healthcare services. Such data offer details on medical procedures, treatments including pharmaceuticals, and socio-demographic characteristics of beneficiaries such as gender, residence, and income status. By contrast, claims data have variations depending on the type and scope of program. They are representative and comprehensive for large patient populations, encompassing elderly, children, the very poor and the institutionalized — i.e., all who exist at the furthest margins of our society and thus easily excluded or under-represented in clinical trials. Since claims data are already collected, they are less expensive to procure than an RCT and tend to be free of ethical obstacles — for instance a clinical equipoise that assumes no better intervention exists than designing an RCT (5); the necessity of sufficient belief that the intervention under investigation is safe; and the risk of ad-

pISSN 1011-8934
eISSN 1598-6357

verse health effects as a consequence of an RCT.

Several forms of claims data have been utilized in North America, Europe, and Asia. The United States have multiple databases operated by private insurers such as Kaiser Permanente and those belonging to public programs such as Medicare and Medicaid. In Europe, the General Practice Research Database (GPRD) in the United Kingdom is most widely used (6). The GPRD consists of longitudinal medical records from primary care and offers accurate information on diagnoses, prescriptions and other healthcare services because the database is generated from medical records from general practitioners. Taiwan is the leading country in utilizing claims data for healthcare service research in Asia. Health insurance in Taiwan is based on fee-for-services and has features of health policy similar to that of Korea such as the single-payer, universal, and compulsory healthcare insurance model. This system covers 99% of the residents in Taiwan. The Taiwan National Health Research Institute (NHRI) is entrusted with continuously and systematically collecting relevant registration and claims data present in the insurance system. The National Health Insurance Research Database (NHIRD) contains comprehensive computerized National Health Insurance (NHI) records of the entire population in Taiwan. As a reliable data source for health research, it has been utilized for academic research in various scientific disciplines. As a result, the number of research articles pertaining to such data have rapidly increased (7).

The Health Insurance and Review Assessment data are health insurance claims data which is also called NHI data, as they are generated in the process of reimbursing claims for healthcare services under the NHI system in Korea. Availability of HIRA data for research was limited to a few research projects that were mainly commissioned by the government. HIRA data become publically available for research in 2009 and the use of the data has increased since then. The availability of the data has been accelerated by a government initiative promoting "Opening and Sharing Big Data for Value Creation" in 2013.

Despite the large potential value of HIRA data in answering a wide spectrum of research questions in health research encompassing outcomes, public health, epidemiology, biostatistics, health informatics and health economics, research with HIRA data was not actively conducted in Korea compared to the US and other European countries. This may be partly due to lack of awareness of its existence and its potential and barriers to accessing the data arising from its complicated structures and numerous variables that require significant efforts to understand before extracting any of the necessary information.

To transform the potential of HIRA data into actual value, the data need to be precisely utilized for research. Thus understanding the characteristics of the data is essential. To this purpose, we provide guidelines on the use of HIRA data for researchers who are interested in using the data in answering their research questions. The paper consists of 5 sections. The first section presents the establishment and roles of HIRA and describes the general features of HIRA. The second section examines advantages and limitations of HIRA data that can compromise the validity of studies. In the third section, we present application of HIRA data by examining previous studies that have used this data, to suggest areas of research where HIRA data may be applicable and strategies to overcome limitations of the data. In this section, we also present ways to access the data. The next section contains the discussion and implications for policy and the final section is conclusion.

THE GENERAL FEATURES OF HIRA DATA

Establishment of HIRA

Having introduced a medical insurance program for companies with 500 or more employees in 1977, Korea gradually expanded the coverage and achieved universal coverage in 1989. In order to drive the health insurance system more efficiently and effectively — through implementing measures such as reducing administrative cost, enhancing management efficiency, simplifying systems for claims and reimbursement, bolstering financial stability, and pooling national risk — integration of multiple-insurers into single-insurers was deliberated. In 1998, the health insurance system was reformed to integrate the fragmented regional medical insurance societies for the self-employed, farmers and fishermen with the then-prevalent medical insurance service provided for government employees and private school employees. In order to complete this reform, the NHI Act was enacted in 1999. Through the enactment of this legislation, the full integration of all medical insurance societies, including those that had served company employees, was attained through the formation of the unified NHI system in 2000. The HIRA and the National Health Insurance Service (NHIS) were thereby established.

The HIRA was founded as an independent single agency distinct from the insurer, providers and other interested parties. The HIRA has 2 main roles: reviewing medical fees for reimbursement decisions; and assessing quality of healthcare services provided to beneficiaries. The HIRA assures the appropriate healthcare provisions through the fair and objective review and assessment in the partnership with NHIS. In addition to these routine procedures, the HIRA has the research department perform research to improve reviews and assessment and to provide the government with the policy-making resources based on its research. The HIRA develops data and information concerning clinical, social, and economic implications of health care. The NHIS reviews the eligibility of insured policy holders; imposes and collects contributions; negotiates the medical fee schedule with healthcare service providers; and reimburses healthcare services provided in accordance with the HIRA's re-

Table 1. NHI program from 2010

Parameters	Reported numbers by year				
	2010	2011	2012	2013	2014
Total population, No. (unit: 1,000)	49,410	49,779	50,004	50,220	50,424
Beneficiaries, No. (unit: 1,000)	50,581	50,909	51,169	51,448	51,757
Health insurance	48,907	49,299	49,662	49,990	50,316
Medical aid	1,674	1,609	1,507	1,459	1,441
Coverage rate, % (beneficiaries/total population)	102.3	102.3	102.3	102.4	102.6
No. of claims (unit: 1,000)	1,307,823	1,327,233	1,420,857	1,418,710	1,453,776
Inpatient	12,491	13,201	14,338	15,512	17,491
Outpatient	1,295,332	1,314,032	1,406,519	1,403,199	1,436,285
No. of providers	81,681	82,948	83,811	84,971	86,629
Tertiary hospitals	44	44	44	43	43
General hospitals	274	275	278	281	287
Hospitals	2,182	2,363	2,524	2,683	2,811
Clinics	27,469	27,837	28,033	28,328	28,883
Community health centers	3,515	3,508	3,502	3,504	3,516
Oriental clinics	12,229	12,585	12,906	13,312	13,654
Dental clinics	14,872	15,257	15,566	15,930	16,377
Pharmaceuticals	21,096	21,079	20,958	20,890	21,058

NHI = National Health Insurance.

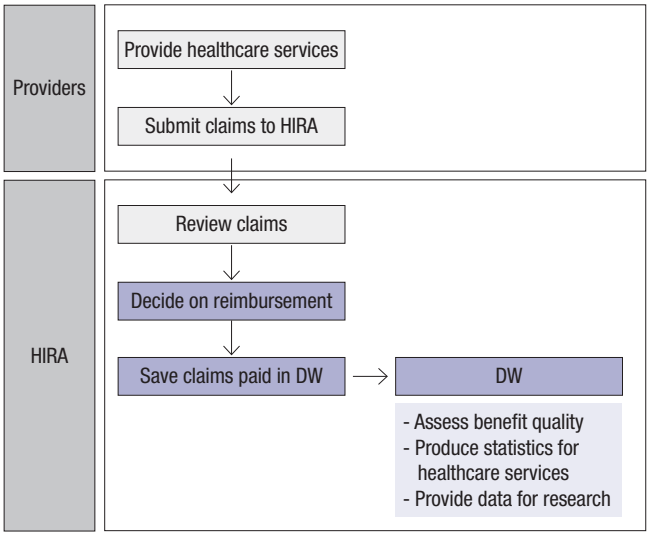

Fig. 1. Flow of data generation of HIRA data.
HIRA = Health Insurance and Review Assessment, DW = data warehouse.

imbursement decisions.

NHI has covered almost 98% of the total population, which numbered approximately 50 million as of 2014 in Korea (Table 1). Under the terms of universal health coverage, all healthcare providers — numbering approximately 80,000 — and all citizens are required to be covered under NHI, which is based on fee-for-services except in cases involving seven types of conditions (Table 1).

Data generation process

HIRA data is generated in the process of reimbursing providers under NHI and contain specific yet comprehensive information on the relevant healthcare services such as procedures, surgeries, examinations, and treatment, including prescriptions as well as the socio-demographic characteristics of patients.

Claims are electronically submitted by providers to HIRA which reviews claims and makes reimbursement decisions. Billing statements for claims for which reimbursement has been completed are stored in the data warehouse (DW) as a record within a database consisting of multiple datasets (Fig. 1). Datasets in the DW become sources of generating statistics on healthcare services, developing indices on quality for each respective type of care, and for health research.

Structure of HIRA data

HIRA research data consists of 6 files (Table 2): 1) the general information file; 2) the healthcare services file, including inpatient prescriptions; 3) the diagnoses file; 4) the outpatient prescriptions file; 5) the drug master file; and 6) the provider information file.

The general information file is a common denominator file and is essential in identifying the study population of interest, as it contains variables representing such fields as the beneficiary's socio-economic characteristics (age and gender), type of insurance (health insurance and medical aid), and 2 diagnoses of which treatment require the most intensive resources (primary and secondary). The file includes variables related to dates pertaining to such events as patient-provider encounters, admission to hospitals and discharge, and lengths of stay for inpatients. Cost information is also included (patient out-of-pocket costs and payer costs) along with payer information (e.g. type and size of practice). The file includes information on whether an operation for a primary diagnosis was performed and the date it was performed. Beneficiary identification (ID) and provider ID are all stored in an encrypted format in order to protect

https://doi.org/10.3346/jkms.2017.32.5.718

Table 2. List of files with variables (selective)

Files	Variables	Common variables
General information	- Beneficiary ID, age, gender, insurance number, type of insurance, date of review, provider ID, indicators for inpatients/outpatients, indicators for types of providers - Operation related to primary diagnosis - Specialty - Dates of treatment, dates of dispensation - Primary diagnosis, secondary diagnosis, surgery, area of provider's practice - No. of days undergoing care, first visit to a physician, dates of encounter, date of admission, date of discharge - No. of days of supply for prescriptions, quantity of prescriptions, special codes for different out-of-pocket costs	Billing statement code, date & year of receipts
Healthcare services	- Procedures, inpatient prescriptions, diagnostic tests, treatments - Operation, injection, and examination - Unit price, quantity per day, days of supply, etc	
Diagnosis	All diagnoses	
Outpatient prescription	Quantity per time, quantity per day, days of supply, drug code, unit price, amount, date of prescription	
Drug master	Drug code, date of starting (and terminating) coverage, drug name, unit, manufacturer, channel of administration, coverage, unit price, etc	
Providers information	Provider ID, location, zip code, name of providers, types of provider, address, date of open, no. of business, no. of beds, etc	

ID = identification.

personal information.

The healthcare services file houses specific and detailed information for healthcare services provided to beneficiaries such as procedures, diagnostic tests, treatments, and inpatient prescriptions.

The diagnosis file includes records of all diagnoses that a beneficiary has received. This file is useful in identifying co-morbidities or assessing the general health status via such figures as the Charlson Comorbidity Index (CCI) (8) and Elixhauser Comorbidity Index (9). Diagnoses are coded in compliance with Korean Standard Classification of Diseases Version 6 (KCD6) which is based on International Classification of Diseases 10th Revision (ICD-10).

The outpatient prescription file contains detailed information on each outpatient prescription, including the following: date of drug prescribed; prescriber ID; name of drug and active ingredients; dose; quantity; number of days of supply; and cost. When prescription drugs are of main interest, researchers need to select either the healthcare service table for inpatient prescriptions, outpatient prescription table, or both tables depending on the scope of drugs studied.

The drug master file is a relational database that provides information such as product code; start and expiry date for coverage; quantity and unit of measure; manufacturer; routes of administration; coverage status indicator (e.g., covered/non-covered, deleted); indicator of eligibility for generic substitution; and therapeutic drug classification codes delineated by the Ministry of Health and Welfare of Korea. The drug master file is regularly updated by accumulating new or altered information. Thus researchers need to use an instance of the database concurrent with the period of the analysis. The drug master file is linked with the outpatient prescription file and healthcare service file by using product codes, which are unique identifiers, as foreign keys.

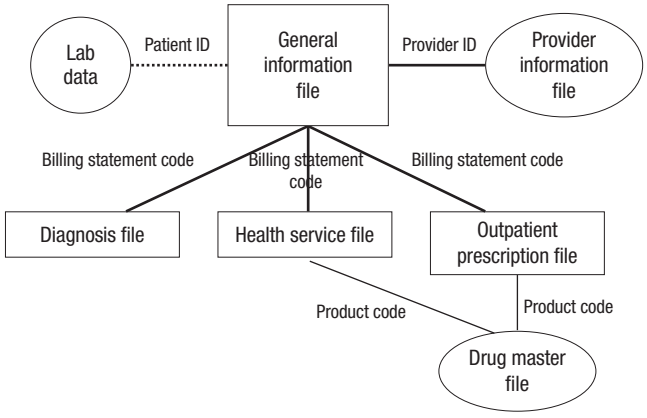

Fig. 2. Linkage of files.
ID = identification.

The provider information file carries information about healthcare providers such as the provider ID; practice location(s); provider type (i.e., primary, secondary, or tertiary); the number of beds; and inception date of institution. This file is useful in exploring patterns or behavior on the part of physicians in prescribing drugs, performing procedures, and administering treatment. Each file is linkable via encrypted beneficiary ID and billing statement codes assigned to an individual claim, as shown in Fig. 2.

ADVANTAGES AND LIMITATIONS OF HIRA DATA

Advantages

Research using HIRA data have more reliable and rigorous findings than ones containing sample data, which may have limitations in representing the populations of interest. The data include information on almost 50 million patients, covering 98% of the total population through the universal coverage system, in which all citizens are covered and all healthcare service providers provide services to patients in Korea. The collection of

data is quite complete because 99% of claims are electronically submitted by providers, and the chances of missing claims are highly unlikely. The NHI system in South Korea, which is based on a fee-for-service delivery system except in a few specific cases which presented 7 conditions are reimbursed with diagnosis related group (DRG), provides highly comprehensive benefits covering conditions ranging from mildly acute conditions — such as colds and upper respiratory infections — to severe conditions such as malignancy and leukemia. To be specific, both inpatients and outpatients are included in the data, allowing follow-up of the entire healthcare service utilization of any given patient over the course of treatment. Consequently, HIRA data contain healthcare service records from the infant to the elderly across the full range of health care settings regardless of geographic location, unlike claims data from Medicare and Medicaid programs, which cover either the elderly or those in the low income bracket in the United States.

This representativeness and comprehensiveness offer research opportunities that would not be feasible using RCTs — for instance when studying rare conditions, adverse reactions to drugs, and populations at the margins of the society, such as the elderly, the disabled, and the institutionalized — as all these demographic groups are excluded or under-represented in RCTs. Due to the large scale of the data, it is possible to have a sufficiently large sample size to secure corresponding statistical power for sophisticated analyses in studies while having narrower margins of error. Researchers are also able to derive variables from this data, which contain rich and specific information on healthcare utilization, procedures, diagnoses, treatment, and payments.

Records of healthcare service utilizations and diagnoses of individual beneficiaries are continuously accumulated in the database, and this enables researchers to track the same subjects over a period of time. This longitudinal characteristic of the data may be conducive to conducting cohort studies and exploring long-term outcomes (effects) from exposures that may not be immediately observed. Furthermore, information in HIRA data is provided exclusively by healthcare providers; consequently, studies employing this data can avoid errors arising from patient self-reporting and non-response, both of which are issues in survey-based studies (10).

Due largely to its secondary nature, HIRA data offers the key advantage that utilizing it is economical compared to primary data, because the data is already collected. Thus making it unnecessary to expend resources such as time, labor, and money for data collection. The likelihood of studies based on HIRA data of facing ethical issues is not great, unlike RCTs, where for instance the decision on whether to assign a patient to a control group vs. a treatment group is fraught with such ethical considerations whenever there are any potential risks during treatment. In contrast to those based on RCTs, studies based on HIRA data provide support for effectiveness of interventions among the general population in routine care; and HIRA-based studies are thus more useful in assessing efficacy and likely have a higher degree of external validity.

Limitations

Despite advantages, HIRA data have limitations that need to be dealt with in conducting research. Firstly, HIRA data lack some kinds of information. The data do not have records accumulated from laboratories, notably on the severity of conditions and health behavior of beneficiaries. For example, although the data may contain information on whether a diagnosis of cancer has been made, there are no indications on the severity or stage of the cancer. Information pertaining to health behavior such as smoking status, drinking, exercise, and diet is not included in the data — even while these pieces of absent information are often as important as outcomes, risk factors, or exposures.

As the claims data are generated to reimburse healthcare services eligible for coverage, services such as cosmetic surgical procedures or over-the-counter drugs that are, as matters of policy, not covered under the system are also absent in the data. Consequently, studying non-covered healthcare services is not possible when relying solely on HIRA data.

Discrepancies occur between diagnoses entered in the data and diseases that a patient has in reality; and this can be a potential source of another limitation of HIRA data. This discrepancy implies that some patients do not necessarily have the medical conditions corresponding to their diagnoses, possibly producing a sort of bias, or at best compromise in the study's validity. This problem arises from the fact that diagnostic classification is a crucial component for analyses. Such discrepancies may arise from the inherent nature of claims data, which are fashioned to obtain reimbursement, and not designed for clinical research purposes. Claims are made with the purpose of generating income for providers, and those who submit them are principally guided by this function. Such a profit-driven motivation may often result in billing practices that use the codes that would provide the highest reimbursement that can plausibly be supported by the medical records. Reimbursement policies can also be a contributing factor in these sorts of discrepancies. For instance, since the diagnosis itself is the factor that determines whether drugs will be covered, the prescribed drugs are often justified by making a diagnosis that is eligible for coverage even though a patient may actually have a condition that is excluded from coverage. Diagnoses that were given prior to ordering a diagnostic test sometimes remain in the data even if they are ruled out by a negative result. Another sort of discrepancy may be variations in diagnoses among physicians, who inevitably vary in the procedures they employ and the treatments they prescribe for any given type of medical condition.

Such discrepancies between diagnosis information and the

 https://doi.org/10.3346/jkms.2017.32.5.718

actual status of health conditions appear not only in HIRA data but also in most other claims data, though HIRA data may have a higher degree of problems largely due to the fee-for-service system and reimbursement policies. A study that examines the accordance of diagnosis in HIRA data to the actual status of health conditions by comparing medical record reports shows that, on average, 70% of diagnoses correspond to diagnoses in medical charts although the accordance rate was different depending on conditions and care setting or types providers (11). Diagnoses in HIRA data tend to be more accurate for severe conditions rather than acute and minor conditions; for inpatient settings than for outpatient settings; and in hospitals than in clinics (11). Malignant tumors and injuries had high accordance rates of 77.6% and 81.2%, respectively, while asthma and arthritis had low accordance rates of 41.2% and 59.6%, respectively (11). The accordance rate was 75.9% for inpatient settings and 55.8% for outpatient settings (11).

Another discrepancy can occur. The information about residence of beneficiaries may not be reliable because HIRA data is collected based on the location of providers. As beneficiaries are free to visit any physician without any restrictions, it is possible that the location where a beneficiary has received the healthcare service is different from the area where beneficiaries actually live.

Although HIRA data are continuously accumulated, they are available only for a 5-year period beginning from the current year: this is due to the fact that the HIRA regulates store claims for 5 years in the DW and records of claims are removed from the DW after 5 years. Research projects that require long-term follow-up exceeding 5 years are not feasible. Recently, the policy of the storage period has been changed; and accordingly, HIRA plans to expand data storage for 10 years.

The above described limitations are not only found in HIRA data but also in other claims data because they are collected for administrative purposes, not for research. Nonetheless, benefits from utilizing HIRA data outweigh those for not utilizing them when limitations are dealt with by employing strategies that will be discussed in the next section.

APPLICATION OF HIRA DATA

Research areas and strategies for the optimal use of HIRA data

In order to find studies with HIRA data, we queried literature using Medline and Cochrane using various combinations of terms such as claims data, administrative data, insurance, health insurance, and HIRA while specifying the search term "Korea" along with the boolean operator "AND." We chose to have the search results limited to the period from 2009 to 2015. The reason is that HIRA was only accessible by those doing research on a government-commissioned basis prior to 2009,and the database has been accessible to all researchers since 2009. We excluded literature using the data obtained from the NIHS of Korea. As shown in Table 3, HIRA data were used in various research areas ranging from medication adherence, prescribing patterns, adverse events, cost-effectiveness, burden of disease, healthcare service utilization, disease incidence and prevalence, outcomes, policy evaluation, and others such as severity and health informatics.

These studies provide exemplary strategies on how to overcome limitations of HIRA data. For information absent in the HIRA database such as general health status, severity of condition, and cause of death, studies link HIRA data with other sources of governmental-owned data and lab data provided by hospitals. A study obtained causes of death by linking investigators' notes from the National Police Agency to identify suicides, which were not accounted for in the HIRA database as identified suicides (12). Studies retrieved and analyzed lifestyle factors such as smoking status and lab data such as spirometry test data by linking to the Korean National Health and Nutritional Survey (KNHANES) of Korea Center for Disease Control and Prevention (CDC) (13,14).

Some studies derive (create) variables by using the existing information in the data. Cases of derivation (or creation) of values for variables from the existing information include those in which CCI was derived using diagnosis information for comorbidities (15-18), medication possession ratio (MPR) and cumulative medication adherence (CMA) for medication adherences. These values were derived by extracting the number of days of supply and dates of dispensation from the data (19-24).

To obtain study populations based on conditions more accurately, studies developed algorithms rather than simply using diagnostic information, which is one way to address the possible discrepancies between diagnosis in HIRA records and the disease that a patient actually has. Algorithms can be made by supplementing diagnoses with procedures, treatments, or prescriptions for treatment. Studying chronic obstructive pulmonary disease (COPD) is a paradigmatic example of this approach. In studies on COPDs with the use of HIRA data, COPD patients are defined as those who are at or above 40 years old with diagnosis codes of ICD-10 in (J42.x–J44, except J430) and have at least 2 claims for COPD prescriptions per year in long-acting muscarinic antagonist (LAMA), long-acting beta-2 agonist (LABA), inhaled corticosteroids (ICS), ICS plus LABA (ICS+LABA), short-acting muscarinic antagonist (SAMA), short-acting beta-2 agonist (SABA), or theophylline (17,25-28).

Studies using claims data are mostly observational studies, which are susceptible to bias resulting from factors such as selection bias and confounding. Therefore, a key challenge in conducting observational studies with HIRA data is to address sources of these biases. In doing this, statistical methods can be adopted to minimize biases (a complete elimination of biases may

Table 3. Research areas using HIRA (or NHI) data

Class	Reference	Linkage to other data	Derived variables	Working definitions	Methods
Adherence and persistence, prescribing pattern	(21)	No	Persistency	Yes	Kaplan-Meier survival analysis
	(22)	No	Persistency	Yes	Cox proportional hazard
	(20)	No	MPR	Yes	Multivariate logistic regression analysis
	(24)	No	CMA	Yes	Multiple logistic regression analysis
	(23)	No	CMA	Yes	Multiple logistic regression analysis
	(19)	No	MPR	Yes	Cox proportional hazard
Healthcare utilization	(40)	NHIC — Medicaid-aid case management center	No	No	Logistic regression analysis (multivariate)
	(41)	No	No	No	Multivariate logistic regression analysis
	(12)	Suicides were identified by the investigator's note from National Police Agency	No	No	Multiple logistic regression analysis, repeated-measure data analysis (proc mixed procedure)
	(42)	No	Crude surgery rate	Yes	Poisson regression model
	(13)	No	No	No	Multivariate logistic regression analysis
	(43)	No	No	No	Multilevel analysis (linear mixed models with random intercept)
	(44)	No	No	No	Two-level random effect logistic regression model
	(45)	No	No	No	Multiple regression analysis
	(46)	No	No	No	Multiple logistic regression analysis
	(14)	KNHANES II	No	No	Multivariate logistic regression analysis
	(18)	No	CCI	No	Multiple logistic regression model
	(47)	No	No	No	Descriptive summary of statistics
	(48)	No	No	No	Descriptive summary of statistics
	(49)	No	No	No	Multiple regression analysis
	(16)	No	CCI	No	Descriptive summary of statistics
	(50)	No	No	No	Descriptive summary of statistics
	(51)	No	No	No	Multiple regression analysis
	(25)	No	No	No	Multiple regression analysis
Burden of disease	(52)	No	No	No	Descriptive summary of statistics
Incidence or prevalence	(30)	No	No	No	Descriptive summary of statistics
	(53)	No	No	No	Descriptive summary of statistics
	(31)	No	No	No	Descriptive summary of statistics
	(32)	No	No	No	Descriptive summary of statistics
	(33)	No	No	No	Descriptive summary of statistics
	(34)	No	No	No	Life table method
	(35)	No	No	No	Descriptive summary of statistics
	(36)	No	No	No	Descriptive summary of statistics
	(37)	No	No	No	Descriptive summary of statistics
Outcomes	(54)	No	No	Yes	PSM/Cox proportional hazard
	(17)	No	Charlson score	Yes	Descriptive summary of statistics/multivariate regression
Adverse event	(15)	No	Charlson score	No	PSM
Policy evaluation	(55)	No	No	No	Adjusted rate
	(56)	No	No	No	-
	(57)	No	No	No	Descriptive summary
Health informatics and others	(58)	Population data from Statistics Korea	Lengthiness index, costliness index, etc	No	Clustering, decision tree, stratification
	(59)	National Cancer Center	Charlson score	No	Multivariate analysis
	(60)	Statistics Korea	No	No	Multivariate logistic analyses

HIRA data is also called NHI data; The list is selective and not mutually exclusive because some fall into multiple categories.
HIRA = Health Insurance and Review Assessment, NHI = National Health Insurance, NHIC = National Health Information Center, MPR = medication possession ratio, CMA = cumulative medication adherence, KNHANES = Korean National Health and Nutritional Survey, CCI = Charlson Comorbidity Index, PSM = propensity score matching.

be ideal but often not possible). Multiple regressions, instrumental variables (IVs), and propensity score matching (PSM) are statistical methods to control biases. Each method has its own advantages and disadvantages. PSM is easy to implement but is unable to control biases from unobserved factors whereas it is hard to find suitable IVs, which are strongly correlated with exposure but not with the outcome (29). Studies adopted various statistical approaches depending on study designs, outcomes of interest and topics. Some studies presented only a descriptive summary of statistics (30-37) while other studies utilized more advanced statistical approaches to control confounders such as multiple (or logistic) regression (19,20,22-24), Kaplan-

https://doi.org/10.3346/jkms.2017.32.5.718

Meiser survival analysis and Cox proportional hazard regression (19,21,22).

Accessing HIRA data

There are 3 ways to access HIRA data. First, researchers can access the raw data, often called big data, either by visiting seven research centers located nationwide or by remote access. Raw data are available for researchers in academics and government agencies and for those in the private sector such as pharmaceutical companies and medical device companies. The scope of the raw data provided is more limited to those in the private sector. Researchers who are interested in using the raw data need to go through the following procedures. First they need to send an email requesting a consultation for the data use; and such a consultation is made in person at the HIRA. After the consultation, the researcher submits the application including a study proposal for the use of data. HIRA holds committees to review the application for approval of the use of data. Once the approval is given, HIRA extracts the data from the DW system tailored for the proposed study. ID information in the data are encrypted to protect private information. The encrypted data are uploaded in a system where analytical tools such as SAS, SPSS, R, and STATA are installed. The system is accessible only by the individual researcher for the study.

Secondly, the HIRA offers four types of patient samples: 13% of the national inpatient sample (HIRA-NIS); 3% of the national patient sample (HIRA-NPS); 20% of the aged population sample (HIRA-APS); and 10% of the pediatric patient sample (HIRA-PPS). Patient samples are available from 2009 to 2015. However, it is not possible to follow an individual over a duration of time, as each sample is cross sectional and none are linkable by individual. No review process is required to acquire samples such as raw data.

Thirdly, summary statistics related to healthcare services such as healthcare expenditures and utilization — including prescriptions, medical conditions, and healthcare providers — are publicly available through the healthcare big data open system (http://opendata.hira.or.kr).

DISCUSSION AND POLICY IMPLICATIONS

HIRA data have a relatively short history of use in research in Korean studies compared to those in European countries and the United States, despite offering advantages such as completeness, comprehensiveness, representativeness, and longitudinal characteristics. This is due to the fact that availability and access to HIRA data had been somewhat restricted and the knowledge that such data exists was shared among few researchers. Furthermore, even if researchers are interested in using the data, working with HIRA data is not easy. Researchers first have to understand the complex structure, characterized by a large number of variables, in order to identify information essential for research. Once they identify information needed, they should be able to extract information and make a dataset for the analysis. Researchers also need to adopt a proper statistical approach that best suits the problems related to selection bias and confounding effects that often threaten the internal validity of observational studies. Observing these caveats require a significant amount of time, knowledge, skills, and experience. Consequently, researchers neglecting or lacking any of these prerequisites may become frustrated and abandon any consideration of using HIRA data.

The use of HIRA data to support research suggests further roles for HIRA. The HIRA needs to provide data that are more research friendly and processed with immediate applicability to research. To offer a case in point, the inclusion of variables for specific conditions derived from the data can be provided with variables for conditions derived through algorithms using records of healthcare service utilization, in addition to diagnosis codes. HIRA also need to encourage studies on identifying severity distribution for conditions.

The case of the United States can be a good example in supporting the use of claims data for research. The Center for Medicare and Medicaid Services (CMS) in the United States provides systematic support for researchers who are interested in using CMS data, such as Medicare and Medicaid data, via contracts with the Research Data Assistance Center (ResDAC) (38) and the Chronic Condition Data Warehouse (CCW) (39). ResDAC provides technical support for researchers in accessing CMS data; training programs in which researchers become aware of strengths and limitations of the CMS database; and expertise on how claims-based studies can explore important health care issues (38). The CCW, another contractor for CMS data, delivers CMS data that have been processed through beneficiary matching, deduplication, and merging of the files in order to make it ready for implementation in study analyses (39).

Limitations of HIRA data suggest important policy implications. To fully realize the potential of HIRA data, linkage to external data from government sectors and private sectors should be made feasible to fill information absent in HIRA data, which in spite of its deficiencies, nevertheless constitute important components of healthcare service research. These data include causes of death recorded in the Statistics Korea Department, the cancer registry of the National Cancer Center, the Korean Genome and Epidemiology study (KoGES) of CDC, and lab data managed by hospitals. In addition, linkage with spatiotemporal data such as meteorological information supplied by the Korea Meteorological Administration and air pollution information provided by the Ministry of Environment allows studies exploring possible associations with the environment on entire populations. The linkage will significantly provide added value to HIRA data by allowing health research, which in turn can en-

hance public health.

Linkage can be possible by working on the following issues. First, government agencies need to develop the willingness to share and provide its own data for research, which would be enhanced by acknowledging the fact that data provides value that ultimately benefits public health. Institutional support for the linkage of data from different sources needs to be organized, and this can be the most challenging because it requires authorization from multiple government agencies that possess data. Institutional support can be provided via 2 possible options: establishing a new entity or designating a government agency as a service entity. The former, namely establishment of a new entity, necessitates a governmental budget, whereas the latter one would face opposition by a non-designated agency, due to the fact that transferring data to another government agency for linkage can be viewed as forfeiting or compromising an organization's ownership of the data. Last but not least, personal information protection also should be addressed; and this issue is not only laden with technical considerations but also with political considerations involving processes related to modifying regulations on matters such as the proper scope of personal information.

CONCLUSION

HIRA data, which can be described as big data containing healthcare service information for most of the Korean population, has enormous potential for research in its ability to create value in several forms such as generating effectiveness data for medicine; improving quality of care; delivering an efficient healthcare system; monitoring the safety of drugs; and evaluating health policy. As availability and access of HIRA data has grown in the recent years, HIRA data can offer a powerful tool for the assessment of healthcare services use and outcomes as long as the appropriate attention is given in regards to their potential limitations. HIRA data can enhance its value as al big data source by completing it via linkage with other data found in governmental sectors and private sectors.

DISCLOSURE

The authors have no potential conflicts of interest to disclose.

AUTHOR CONTRIBUTION

Conceptualization: Kim JA, Kim DS. Investigation: Kim JA. Writing - review & editing: Kim JA, Yoon S, Kim LA, Kim DS.

ORCID

Jee-Ae Kim http://orcid.org/0000-0002-3195-2552
Seokjun Yoon http://orcid.org/0000-0003-3297-0071
Log-Young Kim http://orcid.org/0000-0002-6160-8357
Dong-Sook Kim http://orcid.org/0000-0003-2372-1807

REFERENCES

1. Booth CM, Cescon DW, Wang L, Tannock IF, Krzyzanowska MK. Evolution of the randomized controlled trial in oncology over three decades. *J Clin Oncol* 2008; 26: 5458-64.
2. Dans AL, Dans LF, Guyatt GH, Richardson S. Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-Based Medicine Working Group. *JAMA* 1998; 279: 545-9.
3. Meyer RM. Generalizing the results of cancer clinical trials. *J Clin Oncol* 2010; 28: 187-9.
4. Arana A, Rivero E, Egberts TC. What do we show and who does so? An analysis of the abstracts presented at the 19th ICPE. *Pharmacoepidemiol Drug Saf* 2004; 13: S330-1.
5. Cook C, Sheets C. Clinical equipoise and personal equipoise: two necessary ingredients for reducing bias in manual therapy trials. *J Manual Manip Ther* 2011; 19: 55-7.
6. García Rodríguez LA, Pérez Gutthann S. Use of the UK General Practice Research Database for pharmacoepidemiology. *Br J Clin Pharmacol* 1998; 45: 419-25.
7. Chen YC, Wu JC, Chen TJ, Wetter T. Reduced access to database. A publicly available database accelerates academic production. *BMJ* 2011; 342: d637.
8. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 1992; 45: 613-9.
9. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care* 1998; 36: 8-27.
10. Kendler KS, Gallagher TJ, Abelson JM, Kessler RC. Lifetime prevalence, demographic risk factors, and diagnostic validity of nonaffective psychosis as assessed in a US community sample. The National Comorbidity Survey. *Arch Gen Psychiatry* 1996; 53: 1022-31.
11. Park BJ, Sung J, Park K, Seo SW, Kim SH. Studying on diagnosis accuracy for health insurance claims data in Korea. Seoul: Seoul National University, 2003.
12. Cho J, Lee WJ, Moon KT, Suh M, Sohn J, Ha KH, Kim C, Shin DC, Jung SH. Medical care utilization during 1 year prior to death in suicides motivated by physical illnesses. *J Prev Med Public Health* 2013; 46: 147-54.
13. Chung K, Kim K, Jung J, Oh K, Oh Y, Kim S, Kim J, Kim Y. Patterns and determinants of COPD-related healthcare utilization by severity of airway obstruction in Korea. *BMC Pulm Med* 2014; 14: 27.
14. Jung JY, Kang YA, Park MS, Oh YM, Park EC, Kim HR, Lee SD, Kim SK, Chang J, Kim YS. Chronic obstructive lung disease-related health care utilisation in Korean adults with obstructive lung disease. *Int J Tuberc Lung Dis* 2011; 15: 824-9.
15. Shin JY, Park MJ, Lee SH, Choi SH, Kim MH, Choi NK, Lee J, Park BJ. Risk of intracranial haemorrhage in antidepressant users with concurrent use of non-steroidal anti-inflammatory drugs: nationwide propensity score matched study. *BMJ* 2015; 351: h3517.
16. Lee JY, Jo MW, Yoo WS, Kim HJ, Eun SJ. Evidence of a broken healthcare delivery system in Korea: unnecessary hospital outpatient utilization among

 https://doi.org/10.3346/jkms.2017.32.5.718

patients with a single chronic disease without complications. *J Korean Med Sci* 2014; 29: 1590-6.
17. Kim J, Kim K, Kim Y, Yoo KH, Lee CK, Yoon HK, Kim YS, Park YB, Lee JH, Oh YM, et al. The association between inhaled long-acting bronchodilators and less in-hospital care in newly-diagnosed COPD patients. *Respir Med* 2014; 108: 153-61.
18. Kim SY, Park JH, Kim SG, Woo HK, Park JH, Kim Y, Park EC. Disparities in utilization of high-volume hospitals for cancer surgery: results of a Korean population-based study. *Ann Surg Oncol* 2010; 17: 2806-15.
19. Shin S, Jang S, Lee TJ, Kim H. Association between non-adherence to statin and hospitalization for cardiovascular disease and all-cause mortality in a national cohort. *Int J Clin Pharmacol Ther* 2014; 52: 948-56.
20. Han E, Suh DC, Lee SM, Jang S. The impact of medication adherence on health outcomes for chronic metabolic diseases: a retrospective cohort study. *Res Social Adm Pharm* 2014; 10: e87-98.
21. Ahn SH, Choi NK, Kim YJ, Seong JM, Shin JY, Jung SY, Park BJ. Drug persistency of cholinesterase inhibitors for patients with dementia of Alzheimer type in Korea. *Arch Pharm Res* 2015; 38: 1255-62.
22. Cho SK, Sung YK, Choi CB, Bae SC. Impact of comorbidities on TNF inhibitor persistence in rheumatoid arthritis patients: an analysis of Korean National Health Insurance claims data. *Rheumatol Int* 2012; 32: 3851-6.
23. Shin DW, Park JH, Park JH, Park EC, Kim SY, Kim SG, Choi JY. Antihypertensive medication adherence in cancer survivors and its affecting factors: results of a Korean population-based study. *Support Care Cancer* 2011; 19: 211-20.
24. Park JH, Shin Y, Lee SY, Lee SI. Antihypertensive drug medication adherence and its affecting factors in South Korea. *Int J Cardiol* 2008; 128: 392-8.
25. Rhee CK, Yoon HK, Yoo KH, Kim YS, Lee SW, Park YB, Lee JH, Kim Y, Kim K, Kim J, et al. Medical utilization and cost in patients with overlap syndrome of chronic obstructive pulmonary disease and asthma. *COPD* 2014; 11: 163-70.
26. Kim JH, Park JS, Kim KH, Jeong HC, Kim EK, Lee JH. Inhaled corticosteroid is associated with an increased risk of TB in patients with COPD. *Chest* 2013; 143: 1018-24.
27. Kim J, Rhee CK, Yoo KH, Kim YS, Lee SW, Park YB, Lee JH, Oh Y, Lee SD, Kim Y, et al. The health care burden of high grade chronic obstructive pulmonary disease in Korea: analysis of the Korean Health Insurance Review and Assessment Service data. *Int J Chron Obstruct Pulmon Dis* 2013; 8: 561-8.
28. Kim J, Lee JH, Kim Y, Kim K, Oh YM, Yoo KH, Rhee CK, Yoon HK, Kim YS, Park YB, et al. Association between chronic obstructive pulmonary disease and gastroesophageal reflux disease: a national cross-sectional cohort study. *BMC Pulm Med* 2013; 13: 51.
29. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology* 2006; 17: 260-7.
30. Kim HA, Kim S, Seo YI, Choi HJ, Seong SC, Song YW, Hunter D, Zhang Y. The epidemiology of total knee replacement in South Korea: national registry data. *Rheumatology (Oxford)* 2008; 47: 88-91.
31. Kim SA, Kilgore PE, Lee SY, Nyambat B, Ki M. Trends in pneumonia and influenza-associated hospitalizations in South Korea, 2002-2005. *J Health Popul Nutr* 2011; 29: 574-82.
32. Kwon GY, Lee H, Gwack J, Lee SW, Ki M, Youn SK. Regional distribution of hepatitis C virus infection in the Republic of Korea, 2007-2011. *Gut Liver* 2014; 8: 428-32.
33. Lee EJ, Park HM. Trends in laparoscopic surgery for hysterectomy in Korea between 2007 and 2009. *J Obstet Gynaecol Res* 2014; 40: 1695-9.
34. Lee JH, Park YS, Choi JS. The epidemiology of appendicitis and appendectomy in South Korea: national registry data. *J Epidemiol* 2010; 20: 97-105.
35. Lee WJ, Ko Y, Cha ES. Acute pesticide poisoning among children in South Korea: findings from National Health Insurance claims data, 2006-2009. *J Trop Pediatr* 2014; 60: 4-9.
36. Lee YK, Ha YC, Yoon BH, Koo KH. National trends of hip arthroscopy in Korea. *J Korean Med Sci* 2014; 29: 277-80.
37. Lee YK, Yoon BH, Nho JH, Kim KC, Ha YC, Koo KH. National trends of surgical treatment for intertrochanteric fractures in Korea. *J Korean Med Sci* 2013; 28: 1407-8.
38. Research Data Assistance Center (ResDAC) (US) [Internet]. Available at http://www.resdac.org/ [accessed on 6 November 2016].
39. Chronic Condition Data Warehouse (CCW) (US) [Internet]. Available at Available at http://www.ccwdata.org/ [accessed on 6 November 2016].
40. Ahn YH, Kim ES, Ham OK, Kim SH, Hwang SS, Chun SH, Gwon NY, Choi JY. Factors associated with the overuse or underuse of health care services among medical aid beneficiaries in Korea. *J Community Health Nurs* 2011; 28: 190-203.
41. Cho GJ, Kim LY, Hong HR, Lee CE, Hong SC, Oh MJ, Kim HJ. Trends in the rates of peripartum hysterectomy and uterine artery embolization. *PLoS One* 2013; 8: e60512.
42. Choi HG, Hah JH, Jung YH, Kim DW, Sung MW. Influences of demographic changes and medical insurance status on tonsillectomy and adenoidectomy rates in Korea. *Eur Arch Otorhinolaryngol* 2014; 271: 2293-8.
43. Chung W. Psychiatric inpatient expenditures and public health insurance programmes: analysis of a national database covering the entire South Korean population. *BMC Health Serv Res* 2010; 10: 263.
44. Chung W, Chang HS, Oh SM, Yoon CW. Factors associated with long-stay status in patients with schizophrenia: an analysis of national databases covering the entire Korean population. *Int J Soc Psychiatry* 2013; 59: 207-16.
45. Han K, Cho M, Chun K. Determinants of health care expenditures and the contribution of associated factors: 16 cities and provinces in Korea, 2003-2010. *J Prev Med Public Health* 2013; 46: 300-8.
46. Hong JS, Kang HC. Relationship between the use of new or used computed tomography scanners and image retake rates in South Korea. *Acta Radiol* 2013; 54: 428-34.
47. Kim V, Kim H, Lee K, Chang S, Hur M, Kang J, Kim S, Lee SW, Kim YE. Variation in the numbers of red blood cell units transfused at different medical institution types from 2006 to 2010 in Korea. *Ann Lab Med* 2013; 33: 331-42.
48. Koh IJ, Kim TK, Chang CB, Cho HJ, In Y. Trends in use of total knee arthroplasty in Korea from 2001 to 2010. *Clin Orthop Relat Res* 2013; 471: 1441-50.
49. Lee HS. The impact of emergency room utilization by depression patients on medical treatment expense in Korea. *Osong Public Health Res Perspect* 2013; 4: 240-5.
50. Lee K, Kim H, Heo JH, Bae HJ, Koh IS, Chang S. Application of magnetic resonance imaging and magnetic resonance angiography as diagnostic measures for the first attack of suspected cerebrovascular diseases in Korea. *Yonsei Med J* 2011; 52: 727-33.
51. Park HS, Choi BY, Kwon YD. Rapid increase in the national treatment costs for hepatitis A infections in Korea. *Tohoku J Exp Med* 2012; 226: 85-93.

52. Kim C, Yoo KH, Rhee CK, Yoon HK, Kim YS, Lee SW, Oh YM, Lee SD, Lee JH, Kim KJ, et al. Health care use and economic burden of patients with diagnosed chronic obstructive pulmonary disease in Korea. *Int J Tuberc Lung Dis* 2014; 18: 737-43.
53. Kim HA, Koh SH, Lee B, Kim IJ, Seo YI, Song YW, Hunter DJ, Zhang Y. Low rate of total hip replacement as reflected by a low prevalence of hip osteoarthritis in South Korea. *Osteoarthritis Cartilage* 2008; 16: 1572-5.
54. Kim YJ, Choi NK, Kim MS, Lee J, Chang Y, Seong JM, Jung SY, Shin JY, Park JE, Park BJ. Evaluation of low-dose aspirin for primary prevention of ischemic stroke among patients with diabetes: a retrospective cohort study. *Diabetol Metab Syndr* 2015; 7: 8.
55. Lee K, Lee S. Effects of the DRG-based prospective payment system operated by the voluntarily participating providers on the cesarean section rates in Korea. *Health Policy* 2007; 81: 300-8.
56. Bae G, Park C, Lee H, Han E, Kim DS, Jang S. Effective policy initiatives to constrain lipid-lowering drug expenditure growth in South Korea. *BMC Health Serv Res* 2014; 14: 100.
57. Heo JH, Suh DC, Kim S, Lee EK. Evaluation of the pilot program on the real-time drug utilization review system in South Korea. *Int J Med Inform* 2013; 82: 987-95.
58. Kim YJ, Oh Y, Park S, Cho S, Park H. Stratified sampling design based on data mining. *Healthc Inform Res* 2013; 19: 186-95.
59. Seo HJ, Yoon SJ, Lee SI, Lee KS, Yun YH, Kim EJ, Oh IH. A comparison of the Charlson comorbidity index derived from medical records and claims data from patients undergoing lung cancer surgery in Korea: a population-based investigation. *BMC Health Serv Res* 2010; 10: 236.
60. Park HK, Yoon SJ, Ahn HS, Ahn LS, Seo HJ, Lee SI, Lee KS. Comparison of risk-adjustment models using administrative or clinical data for outcome prediction in patients after myocardial infarction or coronary bypass surgery in Korea. *Int J Clin Pract* 2007; 61: 1086-90.

 https://doi.org/10.3346/jkms.2017.32.5.718