



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

# **A Study on the Improvement for big data utilization**

**- Focused on health insurance claims data-**

Boyoung Jung

**The Graduate School**

**Yonsei University**

**Department of Public Health**

# **A Study on the Improvement for big data utilization**

**- Focused on health insurance claims data-**

A Doctoral Dissertation

Submitted to the Department of Public Health

and the Graduate School of Yonsei University

in partial fulfillment of the

requirements for the degree of

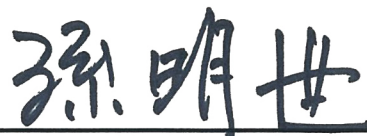
Doctor of Philosophy

Boyoung Jung

June 2018

This certifies that the dissertation of Boyoung Jung is approved.

  
Thesis Supervisor: Soyoon Kim

  
Myongsei Sohn: Thesis Committee Member #1

  
Hyeon Chang Kim: Thesis Committee Member #2

  
Tae Hyun Kim: Thesis Committee Member #3

  
Sukil Kim: Thesis Committee Member #4

Department of Public Health

The Graduate School

Yonsei University

June 2018

## Acknowledgements

The completion of this doctoral dissertation was made possible by the support of the people around me. I would like to express my deepest gratitude to all of them.

Foremost, I would like to express my deepest and sincerest gratitude to my supervisor Prof. So Yoon Kim, for her invaluable guidance, and for the constant encouragement he provided throughout my doctoral course. She has supervised me patiently and always help gave me the confidence to overcome every difficulty.

I would also like to thank members of my thesis committee: Prof. Myongsei Sohn, his valuable advice and thoughtful consideration helped me greatly improve this thesis. He enlightened me to understand health insurance claims data in various aspects as director of the former Health Insurance Review & Assessment Service; Prof. Hyeon Chang Kim, his valuable advice and thoughtful consideration helped me greatly improve this thesis; Prof. Tae Hyun Kim for his valuable guidance with cheerful passion with different point of view; Prof. Sukil Kim for giving a hint to find a right solution to the problem whenever I met barrier in writing my dissertation. Again, it was my honor to be guided by all committee members for their encouragement and insightful comments.

I wish to express gratitude to all my peers in Yonsei University, Health Insurance Review & Assessment Service and Research Institute of Korean Medicine Policy. I thank so much Kyoung Hee Cho for giving her opinions on scientific advisors. She spent their

valuable time to answer my inquiry with her kindness. I would like to thank my peers, Jin-Hyeong Kim, Jaeyeun Kim, Nayoung Park, Da Won Jung, Dong Hyun Lee and Kyungyun Sunu for being supportive during this challenging time. I was able learn knowledge on health policy during my training time from them. I would also like to thank Juchul Kim and Jang Kyung Park who always cheered me whenever I wanted to give up.

I would like to extend my sincerest appreciation to my husband, Han Wool Ko, great help. I also express my thanks to my children, A-ra and Ha-rang. They have always been patient, and supported me. I would also like to express my very profound gratitude my parents, Sun Hae Jung and Hyun Jae Shin, for their support. In particular, I thank so much my parents-in-law, Hee Jun Go and Wha Young Back, for hardships take care of my baby every weekend. Without your dedication and love, I would not have existed who had been struggling to concentrate on my thesis. I also thank my grandmother-in-law and uncle who prayed for her consistently. I thank so much my older sister Dong Sun Jung, twin sister Ji Young Jung and sister-in-law Eun Bit Ko sympathized and comforted me spiritually throughout my life.

Finally, I would also like to express my very profound gratitude to my God for supporting me spiritually throughout my life. I will continue to keep on trusting Lord.

Boyoung Jung

June 2018

## TABLE OF CONTENTS

<b>Abstract</b>	vi
<b>Chapter 1. Introduction</b>	1
1.1 Background	1
1.2 Objectives	4
1.3 Methods	5
1.3.1 Systematic Review (SR)	7
1.3.2 Comparative Study	10
1.3.3 Delphi Survey	11
<b>Chapter 2. Korean Health Insurance Claims Data</b>	14
2.1 General features of Korean Health Insurance Claims Data	14
2.2 Trend of research using Korean Health Insurance Claims Data	24
2.3 Advantages and Limitations of Health Insurance Claims Data	29

### **Chapter 3. Comparison of Big Data Utilization in Foreign Counties-- 35**

3.1 U.S -----	35
3.2 U.K -----	38
3.3 Australia -----	40
3.4 Taiwan -----	42
3.5 Korean -----	44
3.6 Summary -----	46

### **Chapter 4. Analysis in use of Health Insurance Claims Data ----- 48**

4.1 Use-cases of health insurance claims data -----	48
4.2 Issue analysis in utilizing health insurance claims data -----	52
4.3 Analysis results in utilizing health insurance claims data-----	54
4.4 Policy priorities to solve the obstacles -----	55

### **Chapter 5. Strategies in use of Health Insurance Claims Big Data -- 57**

5.1 Establishment of National Big Data governance and strategy -----	57
5.2 Legal and Institutional revision for Big Data openness -----	60
5.3 Activation of healthcare research for Big Data linkage -----	65
5.4 Summary strategies for use of health insurance claims data-----	69



<b>Chapter 6. Discussion and Conclusion</b>	<b>70</b>
6.1 Discussion	70
6.2 Limitation	74
6.3 Conclusion	76

References

Appendix

Korean Abstract

## LIST of TABLES

Table 1. Characteristics of Delphi survey participants -----	12
Table 2. Query content in Delphi survey -----	13
Table 3. Data collection and open status at NHIS and HIRA-----	17
Table 4. Characteristics of health insurance claim data released by NHIS and HIRA -----	23
Table 5. Top 30 journals published study using Health Insurance claims data -----	26
Table 6. Basic study characteristics in the SR -----	28
Table 7. Overview of health related big data in foreign countries -----	47
Table 8. Use-cases of Health Insurance claims data -----	48
Table 9. The use cases of Health Insurance claims data in the public health field --	51
Table 10. Analysis in use of health insurance claims data through PESTLE -----	52
Table 11. Obstacles to utilizing health insurance claims big data -----	53
Table 12. Policy priorities for utilizing health insurance claims big data -----	56
Table 13. Status and contents of legal system related to personal information (collection, use and sensitive information processing)-----	61
Table 14. Exceptions to the prohibition of sensitive information collection -----	63

## LIST OF FIGURES

Figure 1. Methodologic flow of the research -----	6
Figure 2. Literature selection processes-----	9
Figure 3. National Health Insurance System in Korea -----	15
Figure 4. The number of studies using Health Insurance claims Big Data by year (2007-2017)-----	25
Figure 5. Analysis results and strategy in Delphi survey -----	54
Figure 6. Analysis results and Strategy in comparative analysis -----	68
Figure 7. Summary strategies for use of health insurance claims data -----	69

## **ABSTRACT**

### **A Study on the Improvement for big data utilization**

#### **- Focused on health insurance claims data-**

The rapid development of medical technology and information and communication technology (ICT) in Korea has led to the accumulation of vast amounts of information related to healthcare. The National Health Insurance Service (NHIS) and Health Insurance Review & Assessment Service (HIRA) in Korea collect and store health insurance claims data. Despite the excitement and recent interest in healthcare big data, few empirical studies have been conducted to evaluate the potential value of health insurance claims data.

The following three methods were used to suggest strategies for optimal utilization of Korean health insurance claims data. First, Systematic Review was conducted of published studies related to Korean Health Insurance Claims Data. The PubMed and Cochrane database searches from 2007 to 2017. A total of 478 studies were included in the study after applying duplication and elimination criteria to the initial 3,951 search results.

Second, comparative analysis was conducted to draw implications for using Korean Health Insurance Claims Big Data across countries (US, UK, Australia

and Taiwan). Cross-country comparisons were performed based on horizontal as well as vertical comparison perspectives. Data analysis consisted of the constant comparison method.

Third, a Delphi survey was conducted to 42 healthcare professionals working in National Health Insurance Service (NHIS), Health Insurance Review & Assessment Service (HIRA), and relevant agencies. The questionnaire content was intended to identify the obstacles to and policy priorities for the safe use of Health Insurance claims data. This study questionnaire was approved by the IRB Institutional Review Board at Yonsei University (IRB: 2-1040939-AB-N-01-2014-228).

The results of the three methods of this study are as follows.

First, 478 studies were selected as a result of systematic review. There were 55 studies (11.5%) between 2007 and 2012, and a total of 423 (88.5%) were found over the past five years (2013–2017). The HIRA database was used a little more often than NHIS database (HIRA: 51.9%, NHIS: 47.5%). The most frequent research type was health service utilization (41.4%), and 29 (6.9%) out of 478 cases were connected with external data. These data include the information from the cause of death data (12, 41.4 %), clinical data (9, 31.0%), cancer data (7, 24.1%), cost data (6, 20.7%), Surveillance data (2, 6.9%), other data (3, 10.3%).

Second, this study shows the implications for policies in Korea through comparison of the big data utilization in the major countries. The experience of developed countries suggests important issues to be reflected in the formulation of strategies for national utilization of healthcare data; there is a national strategy and health and data governance was being built, it focuses on utilization of public interest objectives such as improvement of public health and medical quality, there is a balance between strengthening and balancing privacy and data security.

Third, 13 policies that indicate four obstacles were included through the Delphi survey. Participants responded by rating the four obstacles in this order: legal immaturity for data use, lack of consensus on providing information, technical constraints on information sharing, and lack of government support. Policy priorities include policy for the “patient’s consent to data use,” a policy for legal revision for Health Insurance Big Data utilization, an institutional improvement policy for Health Insurance Big Data utilization, an institutional consent policy for data provision, technical privacy policies such as anonymization for data sharing, and a national governance establishment policy for Health Insurance Big Data utilization.

Finally, three strategies have been proposed for each issue derived from the three methodologies. First, it is necessary to establish “National Big Data

Governance” for the successful utilization of health related big data. Second, it is necessary to develop legal institutional guidelines in the framework of the separate big data law (differentiation of personal information consent, development of legal and institutional guidelines). The method of consent should be improved to resolve the dilemma whereby utilizing and protecting personal information. Third, it is a strategy to revitalize healthcare research for big data linkage (development of personal information protection technology for data linkage, utilization of user - centered health insurance claim data).

Although Korea is aware of the global trends of big data, negative opinions are still common about the view that the use of personal information is inevitable to improve the quality of life through public well-being and public health promotion. Clear legislative and institutional grounds for the use of Health Insurance Big Data are needed and government support for the proposed policy recommendations should be established.

---

Keywords: Health Insurance Claims Data, Health Insurance Review & Assessment Service (HIRA), National Health Insurance Service (NHIS), Systemic Review (SR), Delphi Survey

## **Chapter 1. Introduction**

### **1.1 Background**

The paradigm of medicine is changing. As the science and technology develop, medicine is shifted from the intuition of the medical staff to the experience medicine. With the development of information and communication technology, the amount of information has increased exponentially, and as the collection of data is no longer restricted, the age of big data has arrived. Big Data is a key factor in determining future competitiveness of the nation because it provides a basis for predicting the future through analysis of massive data.

Major developed countries include the development of “National disease registries” and the development of interoperable “Health Information Systems” as international assessment components of value-based health care <sup>1</sup>. On such a foundation, more meaningful value can be created when big data is connected at the individual level. It has been argued that large data will enable efficiency and accountability in health care.

The first challenge shows potential ways to approach this problem by constructing a data set according to the “bigness” of various dimensions published in JAMA<sup>2</sup>: identify potential healthcare information sources; and determine the



value of connecting them. When data are combined with the right approach using sophisticated information technology, greater value can be gained.

According to a previous study, some researchers have contributed to further research by establishing a list of secondary sources that can be used for research related to health care in Korea<sup>3</sup>. The National Health Insurance Service (NHIS) and Health Insurance Review & Assessment Service (HIRA) have collected big data for national health care and health insurance claims. All of these data have been built in real time in a database format through an information and communication technology (ICT) infrastructure that has been built steadily in the process of informatization that has been proceeding rapidly since the 1990s<sup>4</sup>.

Information such as birth, death, address, work, disability, and income necessary for health insurance operations can be linked to health insurance data based on the resident registration number through the administrative network in Korea. However, the data of each institution are very limited in use and are provided to the public in the form of secondary data for research purposes.

The health insurance claims big data in NHIS and HIRA is only intended for limited in-house use and generally are not linked to other institutions<sup>5</sup>. Therefore, there are limitations on the use of the data as research material or to inform policy<sup>6,7</sup>. To transform the potential of data into actual value, the data need to open and

link precisely utilized for research.

As the elderly and chronic disease populations increase, the paradigm of the health care system is gradually shifting from the treatment of past acute disease to the prevention of disease<sup>8</sup>. There is a tendency to actively use big data to predict the outbreak of disease and to make extensive use of medical information for personalized treatment.

According to McKinsey's report on the Big Data Revolution in the healthcare field, Big Data provides a paradigm shift in providing evidence-based health care and healthcare<sup>9</sup>. The core of this revolution is the data source and the link between data sources. In Korea, it is very important to provide a basis for conducting research with big data linking claims data for evidence-based health care.

The precondition for precise diagnosis is to identify the individual characteristics of patients who are missed in evidence-based medicine and to provide optimized treatment for individual patients. Through the activation of health insurance claims data, customized treatment for the people can reduce unnecessary medical expenses and contribute to the improvement of public health care.

## 1.2 Objectives

Existing studies have segmented the new technologies and business models that can be applied to health care. Healthcare data are characterized by difficulty in voluntary change because of the nature of public services.

Setting a national strategy for innovative technologies and data only within healthcare can make a big mistake. Therefore, the strategy of activating big data within the health care system should be well coordinated. In Korea, importance of healthcare industry and data has been emphasized, but there is still insufficient analysis.

In order to lead innovation in successful health care, specific strategies for cost reduction and medical quality improvement are requested at the government level using big data <sup>10</sup>. By activating the use of health insurance claims data, accurate diagnoses and personalized treatment can be made available to the public, and unnecessary medical expenses can be reduced, contributing to the improvement of public health.

The purpose of this study is to analyze the characteristics and value of health insurance claims data and to suggest strategies for maximizing the use of Korean health insurance claims data.

### 1.3 Methods

The following methods were used to suggest strategies for utilization of Korean health insurance data as follows:

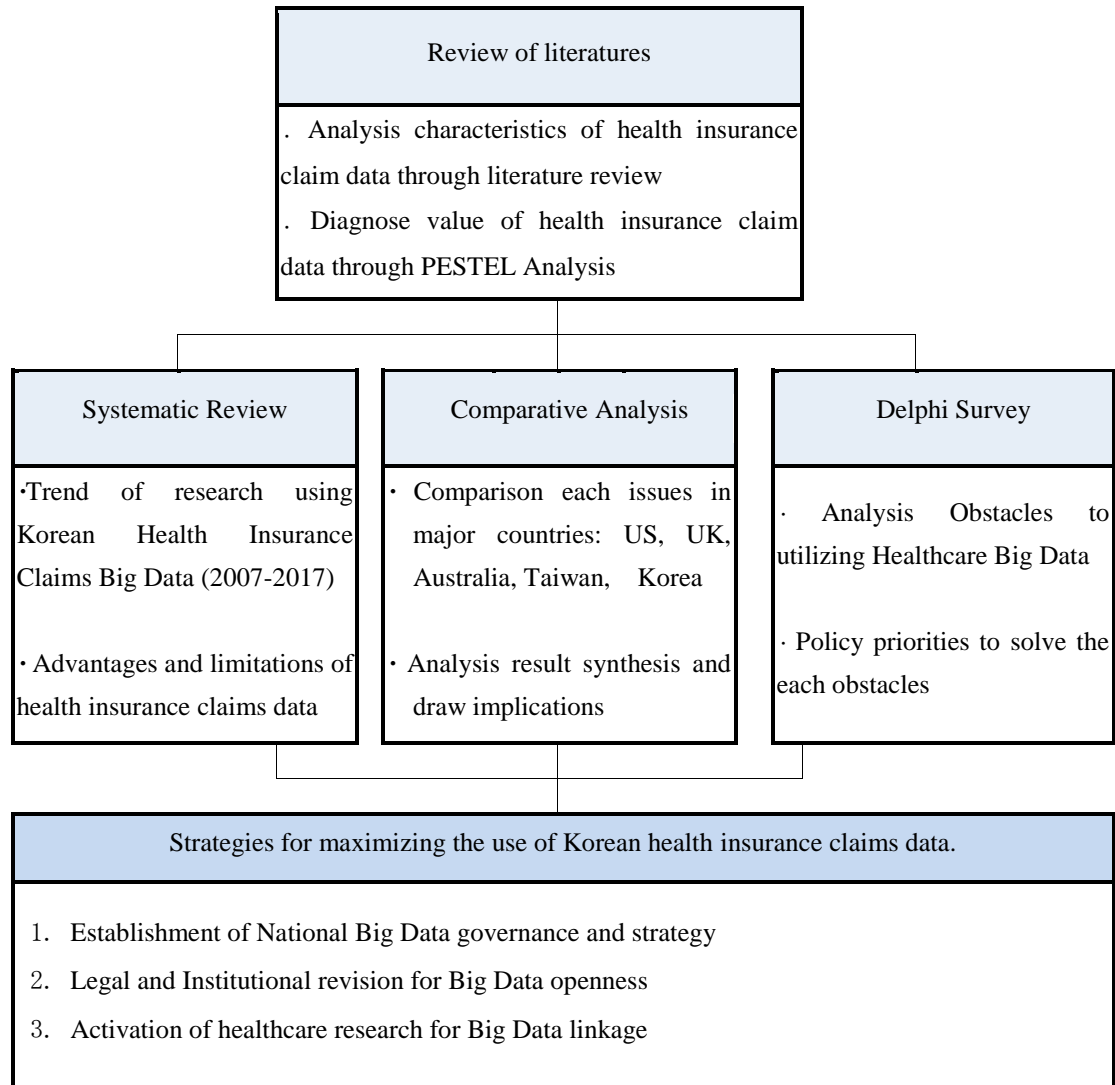
First, the formation of health insurance claim data and its characteristics were analyzed through literature review and a PESTEL analysis.

Second, a Systematic Review was conducted for the published studies related to Korean health insurance claims data over the last 10 years (2007-2017) to review the characteristics and publication trends.

Third, comparative analysis was conducted to draw implications for using Korean health insurance claims data across countries (US, UK, Australia and Taiwan). Cross-country comparisons were performed based on horizontal as well as vertical comparison perspectives. Data analysis consisted of the constant comparison method <sup>11</sup>.

Fourth, Delphi survey <sup>12</sup> was used to diagnose use-cases of Korean health insurance claims data and suggest priorities. Questioner was sent to 42 healthcare professionals working in HIRA, NHIS and relevant agencies.

Finally, three strategies have been proposed to maximize the utilization of health insurance claim data in Korea by combining the method.



**Figure 1. Methodologic flow of the research**

### **1.3.1 Systematic Review (SR)**

#### **Search Strategies**

The systematic review was conducted of published studies related to Korean health insurance claims data on public health over the last 10 years<sup>13-15</sup>. A search within the timeframe of 2007 to 2017 was considered likely to be representative of the period because HIRA and NHIS were only accessible to those performing research on a government-commissioned basis prior to 2009, and the data base has been accessible to all researchers since 2007.

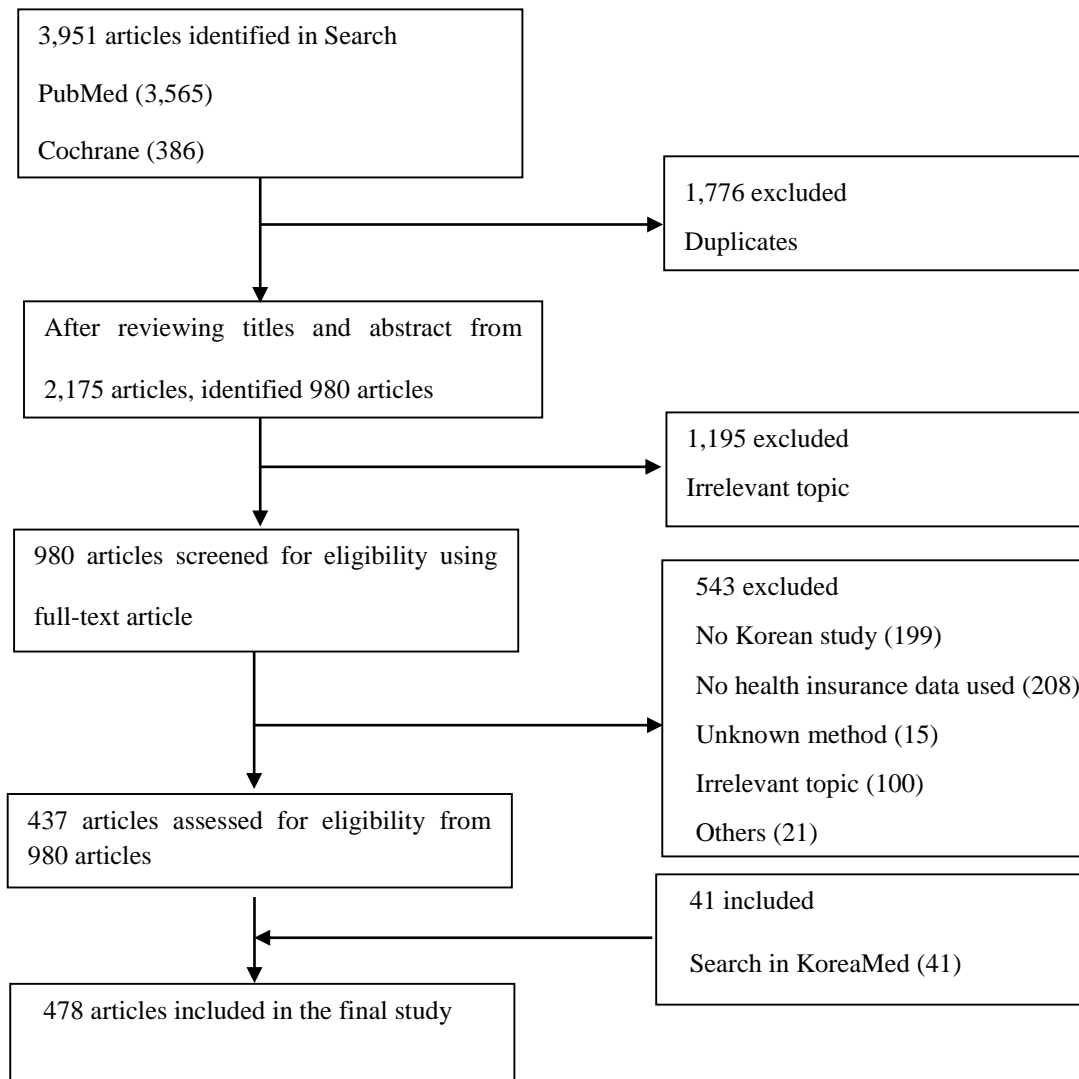
Although practice for SR recommends that at least two researchers should perform this task independently<sup>13</sup>, the same researcher took the time to select and review the literature twice in this study. The review was conducted in accordance with PRISMA guidelines<sup>16</sup>. The range of database retrieval is based on the search range suggested in the COSI (COre Standard Ideal) model adapted from the National Library of Medicine.

Only articles written in Korean or English were included. In PubMed and Cochrane, the terms 1. (“health care insurance” OR “claims data”) AND Korean, 2. secondary data AND Korean, 3. administrative data AND Korean, 4. public health insurance AND Korean, 5. “health insurance” AND Korean, and 6.

(“HIRA” OR “NHIS”) AND Korean were used within the following databases:  
Medline (PubMed), Cochrane, ProQuest Central Korea.

### **Data Extraction**

The search period began on April 19, 2018 and ended on April 28, 2018. The initial search resulted in 3,951 studies. After removing duplicates, the references, including the abstracts of all articles, were downloaded into Endnote software for further analysis. After reviewing keywords from 2,175 articles, 980 articles were identified for further analysis. Finally, 543 papers were excluded by evaluating the relevance of each research goal, and 437 papers were left for further analysis. Subsequently, 41 articles related to the topic were added to the *Korean Journal*. This resulted in a final total of 478 articles identified as relevant and included in the study.



**Figure 2. Literature selection processes**



### **1.3.2 Comparative Study**

The comparison study was used to identify the situation in industrialized countries. By observing and analyzing structures, financing and performances of each system, the fundamental factors of health care system as well as policy is derived<sup>17</sup>. The country selection criterion is the country where the big data utilization level is high. Five countries, UK, USA, Australia, Taiwan and Korea, were selected based on the review of international trends in utilization of health data and prior research.

The coding was proceeded by the constant comparison method <sup>11</sup>. The constant comparison method is a coding method for qualitative data based on grounded Theory. When a new category or topic is found in the coding process, this method returns to the beginning of the data and continuously checks whether the coding operation has been coded to a new category or topic. Categorization proceeded by the review process at least three times.

### **1.3.3 Delphi Survey**

In the final part of the study were conducted using the Delphi survey, which is a systematic, interactive method using a panel of experts who answer questions in two or more rounds <sup>12</sup>. Experts for the Delphi survey were sampled by the intentional sampling method. Intentional sampling is a representative sampling method of qualitative research, and is a suitable method for qualitative research aiming at description, interpretation, insight, and discovery of specific phenomena. In qualitative study, it has been suggested that the appropriate number of samples should be determined at a level of no more information than that collected from 30 persons.

#### **Participants' Characteristics**

Table 1 provides details about the participants in the Delphi survey: sex, age, major, current job positions. Forty-two professionals, including those invited in the First Round, were contacted directly with an invitation to take part in Second Round; of these, all 42 professionals answered the survey (100% response rate). The first-round participants answered open-ended questions, generating items for the second round. Participants from the first round (from 24 February to 14 March) were invited to take part in the second round (from 24 March to 11 April).

**Table 1. Characteristics of Delphi survey participants**

Variable		N (%)
Sex	Male	18 (42.86)
	Female	24 (57.14)
Age	20	7 (16.67)
	30	14 (33.33)
	40	11 (26.19)
	50 or more	10 (23.81)
Major	Arts and Science	8 (19.05)
	Computer Science	12 (28.57)
	Natural Science	22 (52.38)
	Music and physical education	0 (00.00)
Institution	Experts in HIRA	23 (55.00)
	Experts in NHIS	11 (26.00)
	Academic experts	3 (07.00)
	EMR company experts	5 (12.00)
Years of tenure	1–3 years	11 (26.19)
	4–7 years	13 (30.95)
	8–20 years	7 (16.67)
	≥ 21 years	11 (26.19)
Total		42 (100.00)

Abbreviation: HIRA, Health Insurance Review & Assessment Service; NHIS, National Health Insurance Service

Participants were asked to answer five open-ended questions comprising the five topics in Table 2. Responses to each question were coded and summarized using Microsoft Excel. The final list was then used to generate the items for the second round. Forty-five items were generated across the first four categories (the demographic section was not included). Table 2 summarizes the contents of the query in this study.

**Table 2. Query content in Delphi survey**

Round	Question
1st Round	The use-cases of Health Insurance Big Data
	The necessary policies for Health Insurance Big Data use-cases
	The obstacles to the utilization of Health Insurance Big Data
	The necessary policies on the utilization of Health Insurance Big Data
	Demographic information of the participant
2nd Round	Agreed or disagreed with each item
	Agreed or disagreed with each item on a 9-point scale ranging (from “strongly agree” to “strongly disagree”)

Participants rated the extent to which they agreed or disagreed with each item on a 9-point scale ranging from “strongly agree” to “strongly disagree.” Ethical approval for this study was granted by the institutional review board of the Graduate School of Public Health, Yonsei University (IRB: 2-1040939-AB-N-01-2014-228).

## **Chapter 2. Korean Health Insurance Claims Data**

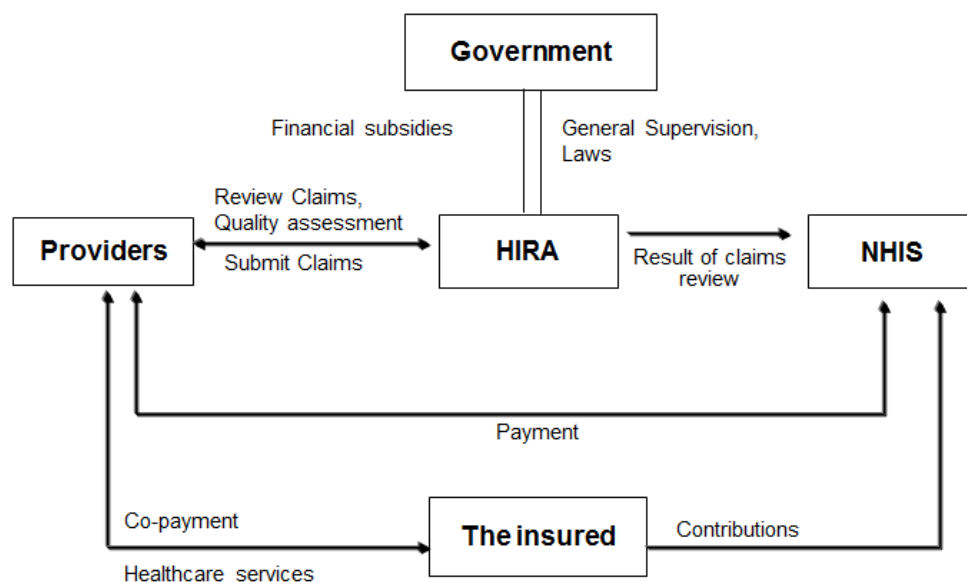
### **2.1 General features of Korean Health Insurance Claims Data**

#### **2.1.1 National Health Insurance System in Korea**

In Korea, the development of the NHI system constitutes the most important part of the change in Korea's medical system. Changes in the Korean health care system can be divided into three periods. In 1977, the Social Health Insurance corporate system (SHI) was introduced. By 1989, this system was expanded to the whole population. Finally, in 1998 the Korean government completed the first integration reform, which changed the corporate partnership to the NHI system<sup>18</sup>.

Through the enactment of this measure, the full integration of all health insurance companies, the self-employed, farmers and fishermen with the then-prevalent medical insurance service provided for government employees and private school employees, was achieved through the formation of the National Health Insurance System. The National Health Insurance Services (NHIS) and Health Insurance Review & Assessment Service (HIRA) have been established in 2000 (Figure 3).

The Korea's NHI system is compulsory for all medical providers and all citizens. The NHIS is responsible for the payment of medical expenses to providers such as medical institutions, check-up institutions, and long-term care institutions that manage the qualifications and premiums of the subscribers and provide medical care, examinations, and nursing services.



**Figure 3. National Health Insurance System in Korea**

In addition, Korea operates a health check-up service system for all citizens different from other countries. There are many countries that have a cancer screening system, but very few countries collect the blood pressure, blood sugar, and even triglyceride results for more than 10 million people every year like

Korea. As a result, the single health insurance system and health check-up service system provides an environment in which the details of the medical care utilization of 50 million people can be accumulated.

Finally, it is composed of a single insurer, the National Health Insurance Corporation, and operates a fee-for-service system as the main payment system. Compulsory enrollment, a single insurer, and a fee-for-service system can be inefficient as a health insurance system, but they are advantageous to health insurance data collection.

The NHIS and HIRA in Korea collect and store health insurance claims data. The following is a comparison of the legal evidence of data collection and the open data scope at NHIS and HIRA (Table 3).

**Table 3. Data collection and open status at NHIS and HIRA**

Process	Aspect	NHIS	HIRA
<b>Collect</b>	Legal Evidence	<ul style="list-style-type: none"> <li>• NATIONAL HEALTH INSURANCE ACT Article 14 (Services, etc.)</li> <li>• LONG-TERM CARE INSURANCE ACT Article 7 (Long-Term Care Insurance)</li> </ul>	<ul style="list-style-type: none"> <li>• NATIONAL HEALTH INSURANCE ACT Article               <ul style="list-style-type: none"> <li>- Article 42 (Health Care Institution)</li> <li>- Article 43 (Reports on Current Status of Health Care Institutions)</li> <li>- Article 47 (Claims for and Payment, etc. of Costs of Health Care Benefit)</li> <li>- Article 63 (Services, etc.)</li> </ul> </li> <li>• PHARMACEUTICAL AFFAIRS ACT               <ul style="list-style-type: none"> <li>- Article 26 (Modification and Revision of Prescriptions)</li> <li>- Article 47-2 (Submission, etc. of Expense Report on Details of Providing Economic Interests, etc.)</li> </ul> </li> <li>• MEDICAL SERVICE ACT (4)               <ul style="list-style-type: none"> <li>-Article 18 (Preparation and Issuance of Medical Prescriptions)</li> </ul> </li> </ul>
<b>Open</b>	Legal Evidence	• ACT ON PROMOTION OF THE PROVISION AND USE OF PUBLIC DATA Article 3 (Basic Principles)	• ACT ON POMOTION OF THE PROVISION AND USE OF PUBLIC DATA Article 3 (Basic Principles)
	Subject	Related Institutions and Researchers	Related Institutions and Researchers
	Range	An anonymous raw data (Qualifications and insurance premiums, medical history, medical examination, medical treatment details, long-term care)	An anonymous raw data (Medical data, drug data, medical treatment data, medical resource data, non-payment data, medical quality evaluation data)
	Methods	Online	Online
	File format	SAS data, Text data	Text data
	Personal Data Process	Provide after consulting privacy (non-discrimination) level disclosed	<ul style="list-style-type: none"> <li>• Check whether information such as private information is included in private information.</li> <li>• Judge whether technical separation and non-identification possibility is possible.</li> </ul>
	Linkage information	Information of total of 36 institutions (Ministry of Health and Welfare, Labor Welfare Corporation, Public Employee Pension Corporation, etc.)	Linkage information (examinee information, national immigration information, death information)

Abbreviation: HIRA, Health Insurance Review & Assessment Service; NHIS, National Health Insurance Service



## **2.1.2 Research Trends**

### **2.1.2.1 NHIS Sample Research DB**

NHIS provides support to research activities in various sectors of society, the economy, environment, industry, etc., as well as policy and academic research on the health sector by providing databases like the Sample Cohort DB, Customized DB, and Health Disease index through the National Health Insurance Sharing Service (NHIS) <sup>19</sup>. In addition, the NHIS will expand opportunities for personnel exchange and cooperation with researchers from a variety of sectors while adjusting to the trends of an open door and sharing of public data.

The sample research DB consists of five types of database: a sample cohort DB, medical check-up cohort DB, elderly cohort DB, working women cohort DB, and infant medical check-up cohort DB. Each cohort DB consists of the following four detailed DBs: qualification DB, treatment DB, medical check-up DB, and clinic DB.

The qualification DB refers to health insurance subscribers and Medicare recipients (excluding foreigners) and includes a total of 14 variables including gender, age, location, type of subscription, and socioeconomic variables of the subject such as income rank, disability, and death.

In the treatment DB, detailed payment data to the clinic upon the treatment of subjects that consists of 10 detailed DB, including statement (20T), details of treatment (30T), type of disease (40T), details of prescription (60T) on the data from medical institution , dental clinic, traditional Korean medicine clinic, and Pharmacy. It includes 57 variables, including a common statement, treatment, type of disease, and prescription 28 variables in 20T, 13 in 30T, 5 in 40T, and 11 in 60T.

The medical check-up DB includes major results from medical check-up and behavior and habitual data from questionnaires, including primary general medical check-up data and transition period check-up data from 2008. It comprises separate medical check-up DBs prepared for 2002–2008 and 2009–2013. Major check-up and questionnaire items have changed due to changes in the medical check-up system (2009), and include 37 variables in 2002–2008 and 41 variables in 2009–2013.

The Clinic DB includes the status, facility, equipment, and personnel data of clinics by type, establishment, and location (city and province) and comprises a total of 10 variables.

### **2.1.2.2 HIRA National Patient Sample Data**

The HIRA data are generated in the process of reimbursing the NHI provider and is specific but concerns relevant medical services, such as prescriptions and procedures, including patients' social-demographic features, surgical examinations, and treatments. The insurance claim is electronically submitted to HIRA from the medical institution, the insurance claim is examined, and a decision on redemption is made. Billing statements that have been refunded are stored in a data warehouse (DW) as records in a database consisting of multiple datasets. The datasets in DW will be the source of statistics on health care services for the development of quality indicators for each care type, as well as a source for health research<sup>20</sup>.

The HIRA provides a patient sample dataset for research purposes that has been sampled for patients based on the claims data. The sample dataset is a statistically sampled secondary dataset in which information on individuals and corporations has been removed from the raw data, and the data have been extracted from the patient unit stratification system according to sex and age (within 5-year age brackets 5 years old), including medical history and prescription details for all patients who have used medical services for one year. This dataset is available in four versions; 1) HIRA-NPS, 2) HIRA-NIS, 3) HIRA-APS, and 4) HIRA-PPS.

The sample dataset of HIRA consists of five files: 1) General Information DB (20T). 2) Medical Service DB, including hospitalization prescription (30T). 3) Diagnostic DB (40T). 4) Outpatient Prescription DB (53T). 5) Provider Information DB (YKIHO). The general information DB (20T) is a common denominator file, because it contains variables such as beneficiary socio-economic characteristics (age and gender), type of insurance (national health insurance and Medicaid), and two diagnoses (primary and secondary) to show which treatments require the most intensive resources.

Beneficiary identification (patient ID) and provider identification (hospital ID) are all stored in an encrypted format to protect personal information. It contains variables related to such events as patient-provider encounters, dates of admission and discharge, and length of stay for inpatient and cost information (patient out-of-pocket costs and payer costs).

The Medical Service DB (30T) contains detailed information on medical services provided to beneficiaries, such as prescriptions, procedures, diagnostic tests, treatments, and in-hospital prescriptions. It includes classification codes, the unit price and number of each procedure, generic name code, daily dosage information, the number of doses, and medical exception classification codes.

The Diagnostic DB (40T) contains a record of all diagnoses received by beneficiaries. This file includes diagnoses (primary and secondary) on T20 and other diagnoses coded in compliance with the Korean Classification of Disease version 6 (KCD 6), based on the International Disease Classification, Tenth Amendment (ICD-10). It is useful for identifying common morbidities and assessing general health status using numbers such as the Charlson Comorbidity Index (CCI) (8) and Elixhauser Comorbidity Index (9).

The Outpatient Prescription DB (53T) contains detailed information on each outpatient prescription and includes prescriber identification (name of drug and active ingredient), dose, quantity, supply days, and cost. If a prescription medication is the primary concern, researchers need to choose either hospitalization prescriptions, outpatient prescription tables, or both, depending on the scope of the drug being studied.

The Provider Information DB (YKIHO) contains information on health care providers such as provider identification, practice location, provider type, the number of beds and the number of providers per 50 beds. Each file can be linked via an encrypted beneficiary ID and a billing statement code assigned to an individual claim.

**Table 4. Characteristics of health insurance claim data released by NHIS and HIRA**

Class	NHIS	HIRA
Name of data	NHIS Sample Research DB	HIRA Patient Data Set
Purpose	NHIS will provide support to research activities in various sectors of society, the economy, environment, and industry, as well as policy and academic research on health	HIRA will release data requested for the assessment of medical care benefits and appropriateness assessment of medical care benefits under Article 55 of the Health Insurance Act
Provision methods	Online (An anonymous raw data)	Online (An anonymous raw data)
Years	2002–2013 (12 years)	2010–2016 (per years)
Contents	<ol style="list-style-type: none"> <li>1. Sample cohort DB</li> <li>2. Medical check-up cohort DB</li> <li>3. Elderly cohort DB</li> <li>4. Working women cohort DB</li> <li>5. Infant medical check-up cohort DB</li> </ol>	<ol style="list-style-type: none"> <li>1. HIRA-NIS</li> <li>2. HIRA-NPS</li> <li>3. HIRA-APS</li> <li>4. HIRA-PPS</li> </ol>
Advantages	<ul style="list-style-type: none"> <li>• Qualification and insurance premiums, medical history DB (medicine prescription information), health examination allowance, medical care allowance, long-term care DB for the elderly, etc.</li> <li>• It is suitable for longitudinal research by establishing the principle of representative, sustainability, inclusiveness, and completeness.</li> </ul>	<ul style="list-style-type: none"> <li>• All patients using medical services for 1 year were extracted from the patient unit stratification system based on sex and age range (5-year age brackets) including medical history and prescription details to establish the principles of inclusiveness and completeness of the entire national data</li> <li>• It is possible to check the contents of detailed medical use by establishing the billing data under fee for service (FFS), and to check rare disease and drug use information.</li> </ul>
Disadvantages	<ul style="list-style-type: none"> <li>• The size of the sample cohort DB itself is also large, so that user convenience is lowered.</li> <li>• A limited cohort of 14 years (2002–2015), only remote analysis is possible, which makes it difficult for researchers to carry out subsequent studies.</li> <li>• Since the claims data are based on the covered items, uncovered data are absent.</li> </ul>	<ul style="list-style-type: none"> <li>• There is a problem in the accuracy of the diagnosis due to differences in billing practices by the hospital or by entering an illness that does not correspond to the actual illness for the purpose of billing.</li> <li>• Cross-sectional data are not suitable for cross-sectional survey studies</li> <li>• Since the claims data are based on covered items, uncovered data are absent.</li> </ul>
Homepage	<a href="https://nhiss.nhis.or.kr/bd/ab/bdaba011eng.do">https://nhiss.nhis.or.kr/bd/ab/bdaba011eng.do</a>	<a href="http://opendata.hira.or.kr/home.do">http://opendata.hira.or.kr/home.do</a>

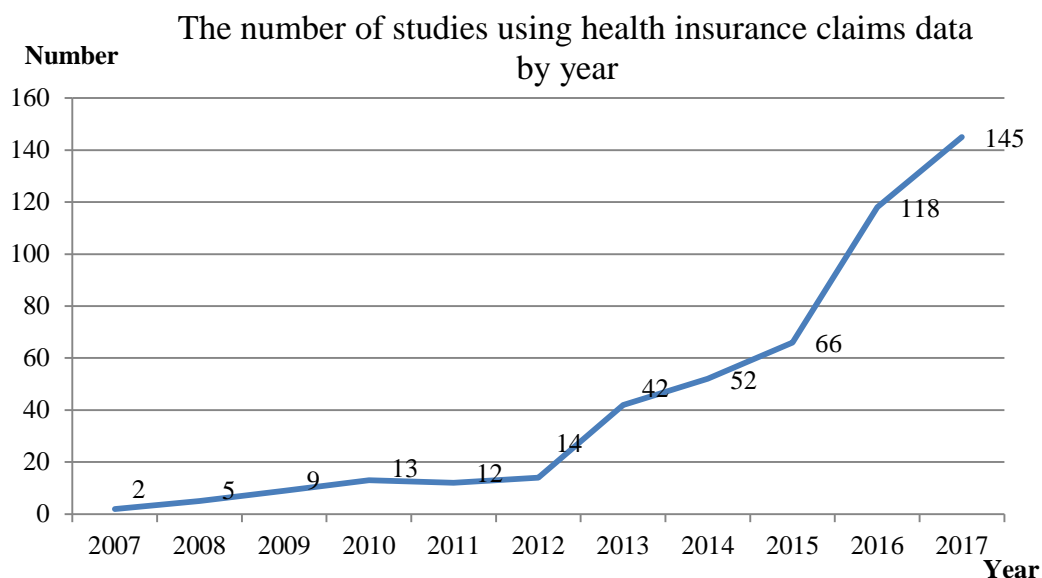
Abbreviation: HIRA, Health Insurance Review & Assessment Service; NHIS, National Health Insurance Service

## **2.2 Trend of research using Korean Health Insurance Claims Data**

### **2.2.1 Research Trends**

Health insurance claims data are a powerful resource that allows us to study routine clinical practice for a relatively large number of individuals. In particular, it has great potential to benefit investigations of the effect of changes of medical service usage patterns and fitness of doctor procedures, changes of health policies, and adverse policy effects. In this chapter, published studies using Korean health insurance claims data for the past decade (period was limited the 2007 to 2017), have been reviewed using the Systematic Review (SR). Because NHIS and HIRA could only be accessed by those doing government research before 2007 and all researchers have had access to the database since 2013.

The main purpose of the SR is to investigate research using Korean health insurance claims data in the health science field and the advantages and disadvantages of the claims data as research data. During the observation period (2007–2017), the number of publications that used Korean health insurance claims data for research purposes increased significantly (Figure 4). There were 55 studies (11.5%) between 2007 and 2012, and a total of 423 (88.5%) were found over the past five years (2013–2017)



**Figure 4. The number of studies using health insurance claims big data by year (2007–2017)**

As shown in Table 5, health insurance claims data in the studies in question were used in a variety of research fields, including health care service use, cost analysis, research on intervention and evaluation, health policy research, specific validity, and validity analysis. They were published in 245 journals, and the number of articles per journal ranged from 1 to 47. The most common journals were *J Korean Med Sci* (9.83%), *PLoS One* (7.74%), *J Prev Med Public Health* (4.81%), *Medicine* (3.76%), *Yonsei Med J* (1.67%), *J Bone Metab* (1.25%), *Asian Pac J Cancer Prev* (1.04%), *Health Policy* (1.04%), *Stroke* (1.04%), and *BMC*



*Health Serv Res* (1.04%). Interestingly, there were 166 journals that had only one of the included publications.

**Table 5. Top 30 journals that published studies using health insurance claims big data**

Rank	Journal	N	%
1	<i>J Korean Med Sci</i>	47	9.8
2	<i>PLoS One</i>	37	7.7
3	<i>J Prev Med Public Health</i>	23	4.8
4	<i>Medicine</i>	18	3.8
5	<i>Yonsei Med J</i>	8	1.7
6	<i>J Bone Metab</i>	6	1.3
7	<i>Asian Pac J Cancer Prev</i>	5	1.0
8	<i>Health Policy</i>	5	1.0
9	<i>Stroke</i>	5	1.0
10	<i>BMC Health Serv Res</i>	5	1.0
11	<i>Korean J Thorac Cardiovasc Surg</i>	5	1.0
12	<i>Endocrinol Metab</i>	4	0.8
13	<i>J Korean Acad Nurs</i>	4	0.8
14	<i>Int J Cardiol</i>	4	0.8
15	<i>Sci Rep</i>	4	0.8
16	<i>Cancer Res Treat.</i>	4	0.8
17	<i>Korean J Intern Med.</i>	4	0.8
18	<i>Korean J Fam Med</i>	4	0.8
19	<i>Rheumatol Int</i>	3	0.6
20	<i>Osong Public Health Res Perspect</i>	3	0.6
21	<i>Respir Med</i>	3	0.6
22	<i>Pharmacoepidemiol Drug Saf</i>	3	0.6
23	<i>J Clin Endocrinol Metab</i>	3	0.6
24	<i>Int J Clin Pharmacol Ther</i>	3	0.6
25	<i>Diabetes Metab J</i>	3	0.6
26	<i>Cancer Res Treat</i>	3	0.6
27	<i>Spine</i>	3	0.6
28	<i>Clin Orthop Surg</i>	3	0.6
29	<i>BMJ Open</i>	3	0.6
30	<i>Korean Circ J</i>	3	0.6

### 2.2.2 Research Features

The basic characteristics of the publications are summarized in Table 6. The HIRA was used a little more than the NHIS, but the difference was insignificant (HIRA: 51.9%, NHIS: 47.5%). Two-thirds (65.7%) of the publications studied included general populations (64.0%) and one-third (36.0 %) special populations including disease specific (e.g., patients with diabetes or hypertension) and/or age and sex-specific populations (e.g., children or senile). A high proportion of the included publications used national data 312 (65.3%), whereas 54 (11.3%) were regional studies.

The analysis period for claims data ranged from 1 to 132 months: over half of the reports studied periods of 4 years or longer. The most frequent research types were health service utilization (41.4%) and intervention and evaluation studies (31.8%). In 8.4% of the studies, claims data were used exclusively or additionally to quantify costs. Approximately half of the studies (42.5%) did not report their funding source or received no external funding. Where recorded, the major sources were public institutions or organizations (47.5%).

**Table 6. Basic study characteristics in the SR**

Characteristics	N	%
Institutional Source		
NHIS	227	47.5
HIRA	248	51.9
NHIS and HIRA	3	0.6
Population studied		
General population	306	64.0
Specific population	172	36.0
Region studied		
Regional	54	11.3
National	312	65.3
Not specified	112	23.4
Analysis period		
< 13 months (1 year)	116	24.3
13–36 months (2–3 years)	87	18.2
37–84 months (4–7 years)	141	29.5
8 years ≤	102	21.3
Not specified	32	6.7
Research Type		
Cost analysis	40	8.4
Intervention and evaluation study	152	31.8
Health policy research	37	7.7
Health service utilization	198	41.4
Specific validity and plausibility analysis	35	7.3
Others	16	3.3
Funding source		
Public	227	47.5
Industry	36	7.5
Public and industry	12	2.5
None or not specified	203	42.5
Total	478	100.0

## **2.3 Advantages and Limitations of Health Insurance Claims Data**

### **2.3.1 Advantages of Health Insurance Claims Big Data**

In this review, Korea's health insurance claims data being included in 478 studies shows that over the past decade data on public health have increasingly been used for research purposes. Many researchers have conducted epidemiological studies<sup>21-25</sup> using national health insurance claims data in order to gain generality. There is no doubt that Korea's health insurance claims data are being used for a variety of research purposes.

First, it can be useful to obtain accurate incidence and prevalence predictions which includes information on about 50 million patients covering 98% of the total population. It covers all citizens through a universal coverage system where all medical service providers provide services to patients. Since 99% of the claims are electronically submitted by the provider and the possibility of claims being lost is very low, data collection is essentially complete. As a result, Korea's health insurance application data, unlike billing data from Medicare and Medicaid programs targeting the elderly or low-income insured in the United States, contains a full range of health data regardless of geographical location, which includes medical service records from infants to the elderly throughout care facilities.

Second, since HIRA data is large-scale data, it can provide a sample size large enough to ensure corresponding statistical power. From these data, researchers can also derive variables that contain rich and specific information on medical use, procedures, diagnoses, treatments, and payments. Records of medical service usage and diagnoses of individual beneficiaries are continually accumulated in the database, allowing researchers to track the same subject over a period of time.

Third, information on HIRA data is provided only by health care providers. Due primarily to their secondary nature, the use of HIRA data is more economical than collecting primary data because the data have already been collected. Therefore, one need not spend time, effort, or resources on data collection. Fourth, in contrast to those based on RCT, studies based on claims data support the effectiveness of interventions among the general population in routine care. Therefore, HIRA-based studies are more useful in evaluating effectiveness and are more likely to have a higher degree of external effectiveness.

### 2.3.2 Limitations of Health Insurance Claims Big Data

Despite their advantages, national claims data have limitations that need to be addressed when conducting research.

First, there are several types of information missing from the claims data. There is no record in the data of the information collected from the laboratory, in particular the severity of the beneficiary's condition and health behavior. For example, even if the data contains information on whether a cancer diagnosis was made, there is no indication of cancer severity or stage<sup>21,26-29</sup>. Information on health behavior such as smoking status, drinking, exercise, and meals is not included in the data<sup>23,30-32</sup>, although the missing information is often as important as the outcome, risk factors, or exposure.

To consider severity, diagnosis files were considered in identifying co-morbidities or assessing general health status via such figures as the Charlson Comorbidity Index (CCI) and Elixhauser Comorbidity Index. Diagnoses are coded in compliance with the Korean Standard Classification of Diseases Version 6 (KCD6), which is based on International Classification of Diseases-10th Revision (ICD-10). However, there is a tendency to deliberately raise the severity of a patient's illness (up-coding) in order to prevent the payment to the medical institution from being reduced during the appraisal process. Because of this tendency, simply trying to grasp the prevalence of a specific disease only by the

health insurance injury sign code will result in an overestimation of the number of patients.

On the other hand, the region of hospital was used as a surrogate of the residences of patients. Since the claims data are collected based on the location of the provider, the information on the residence of the beneficiary may not be reliable. Beneficiaries can freely visit any doctor, so the place where beneficiaries received medical services may be different from where they actually live.

Second, a difference may arise between the diagnosis entered in the data and the disease actually suffered by the patient. Such discrepancies arise from the essential nature of claims data, which are created to obtain reimbursement and may not be designed for clinical research purposes. Claims are made for the purpose of creating provider income, and the submitter is mainly led by this function. Such profit-driven motivations can result in billing practices using codes that provide the highest reimbursement supported by medical records. The refund policy may also be a cause of such contradictions.

Another type of contradiction is the variability of diagnosis among physicians, whereby the procedures they use and the treatment regimens prescribed for any kind of medical condition are necessarily different. Differences between such diagnostic information and actual health status will appear not only in Korean Health Claims Big Data but also in most other billing data, but the discrepancies

in Korean Health Claims Big Data are largely due to the fee-for-service system and reimbursement policies. The accuracy of diagnosis in the KNHI claims data has been reported to be about 70% <sup>33</sup>.

Third, it is not covered non-reimbursable items such as traditional drugs did not generate billing data. Thus, direct medical costs based on information in the claims database were calculated. In general, there are non-medical costs such as transportation costs and lost productivity due to morbidity. Also, examination of the costs did not include items based on claims data that only contained information about medical services provided under the National Health Insurance (NHI)<sup>31,34</sup>.

Fourth, although Korean health claims data are continuously accumulated, they can only be used for 5 years from the current year. HIRA regulates shop complaints for 5 years in the DW, and the record of a complaint is deleted from the DW five years later. Recently, a policy has been announced under which HIRA plans to expand data storage for 10 years. These studies provide exemplary strategies on how to overcome limitations of health insurance claims data. For information missing from the health insurance claims data, such as general health status, severity of condition, and cause of death, studies link claims data with other sources of government-owned data and lab data provided by hospitals.

These health insurance data can be easily collaborated using the resident



registration number uniquely given to Korean individuals from birth. Information on a person's birth, death, address, workplace, disability, income, etc., necessary for the operation of health insurance can be linked with health insurance data based on the resident registration number via the administrative computer network.

In this research, 29 (6.0%) out of 478 cases were connected with external data (Appendix C). These data include the information from the cause of death recorded in the statistical department (12, 41.4 %) <sup>27,28,31,34-42</sup>, clinical data such as specific disease cohort data or hospital data (9, 31.0%) <sup>23-25,32,40,41,43-45</sup>, cancer data at the National Cancer Center (7, 24.1%) <sup>21,26-30,46</sup>, cost data at Korea Health Panel (6, 20.7%) <sup>23-25,32,40,41,43-45</sup>, KNHANES data (5, 17.2%) <sup>30</sup>, additional data such as questionnaire survey and OECD data (3, 10.3%) <sup>22,39,47</sup>, and Surveillance data from CDC Korea (2, 6.9%) <sup>42,48</sup>.

According to Oh, Yoon et al.(2011)<sup>34</sup>, the economic burden of musculoskeletal disorders is estimated using various materials such as National Health Insurance statistics, information from the Korean Health Panel (KHP), and reports on causes of death made by the National Statistical Office. Particularly when linked with other health care claims data or external sources, a wide range of questions can be answered. Appendix 3 shows an example of the studies using external data linked to the health insurance claims data.

## **Chapter 3. Comparison of Big Data Utilization in Foreign Counties**

### **3.1. U.S**

#### **3.1.1 Governance**

The Big Data market has grown rapidly under the leadership of US private companies such as Google and Facebook, and has been used in various areas such as marketing and advertising. The US government is also aiming at innovative services using information such as pursuing comparative effectiveness research in order to create a basis for health services through the National Health Institute (NHIS) and the Agency for Healthcare Research and Quality (AHRQ).

Since the United States relies on the private healthcare market, Center for Medicare and Medicaid (CMS) established the Office of the National Coordinator (ONC) to standardize and manage the information so that the information integration platform among private insurers can be active. Currently, the ONC are collaborating on the meaningful use of Electronic Health Record (EHR) <sup>49</sup>.

ONC conducts a certification program for health information technology in consultation with the head of the National Institute of Standards and Technology (NIST) <sup>50</sup>.

### 3.1.2 Privacy

The US Health Insurance Portability and Accountability Act (HIPAA) is a federal law enacted in 1996 that standardizes the electronic exchange of medical administrative and financial data <sup>51</sup>. The Protected Information includes protected health information (PHI) <sup>52</sup>. The “HIPAA Privacy Rule” defines the PHI protected by the Act, as well as the free use of medical information and the provision of partial disciplinary exemptions that allow for the provision of disciplinary exemptions <sup>51</sup>.

In the US, federal law basically does not have any regulations to protect personal information, and instead an individual approach has been adopted to protect personal information based on various individual laws according to the degree of use of personal information. Therefore, the sensitive information such as medical care can only be collected with the prior consent of the information subject.

Also, de-identified health information can be freely used and provided by anyone with full disciplinary exemption. The United States is conducting an All-of-US Research Program with the support of NIH to show that national health can be improved by providing such information (surveys, claims, electronic medical records, etc )<sup>53,54</sup>.

### 3.1.3 Research

The United States Agency for Healthcare Research and Quality (AHRQ) is a federal research institute that conducts and supports research in the field of healthcare services. One of AHRQ's research programs is National Inpatient Sample (NIS). The NIS includes all medical communities belonging to the American Hospital Association, excluding rehabilitation hospitals. NIS is based on data collected from approximately 3,900 medical institutions in 37 communities in the community. Includes total hospitalization data (about 5 million to 8 million admissions) of selected medical institutions sampled from about 20% (800-1,100 institutions) of participating medical institutions annually<sup>55,56</sup>.

US Medicare and Medicaid Service Center (CMS) systematically supports researchers interested in using CMS data such as Medicare and Medicaid data<sup>57</sup> through an agreement with the Research Data Assistance Center (ResDAC)<sup>58</sup>.

## **3.2 U.K**

### **3.2.1 Governance**

Unlike the United States, the UK operates a tax-based national health care service, which accumulates the largest amount of information on healthcare-related data. The UK has also established the NHS Digital (formerly HSCIC, Health and Social Care Information Center) for the use of healthcare at the government level, collecting and managing information and interoperating among various databases.

### **3.2.2 Privacy**

In the UK and other European countries, personal information may be used to the extent that such use complies with the OECD Principles of Privacy Protection and the EU Privacy Directive. For the protection of privacy, the principle of opt-in presupposing the prior consent of the information subject to collecting, processing, and using personal information shall be followed. However, for research related to public health or public interest, the “explicit consent” principle must be observed. (Data Protection Act, Article 30: Health, Education, and Social Work).

In the UK, the NHS Act was amended in 2006 to recognize patient and public benefits. If the patient’s consent is not available, the use of personally identifiable

information is permitted under certain principles and procedures<sup>59</sup>. In particular, the data linkage center in Scotland uses a data linkage model called “Trusted Third Party Indexing”<sup>60</sup>. The Trusted Third Party Indexing Model is of legal significance in that it aims to maximize the protection of personal information while linking data<sup>61</sup>.

### **3.2.3 Research**

Most of the information is medical records recorded by doctors at the clinic site, such as Clinical Practice Research Data (CPRD), Health Improvement Network Database, and QRESEARCH Database. Anonymized data can be easily obtained because patients already have consent for epidemiological studies and utilization of research purposes<sup>62,63</sup>. In addition, It has provided “differentiated deliberation procedures” and provision structures based on the sensitivity of personal information to prevent data leakage<sup>64</sup>.

Since research using Big Data could be done very actively, there were 749 papers published in 22 countries between 1995 and 2009<sup>62</sup>. Furthermore, starting from 2014, we launched a program called care.data that extracts information from the computer at the general clinic as well as the hospital.

### **3.3 Australia**

#### **3.3.1 Governance**

Australia has established a national strategy (2013) based on the use of data combining and provision services (1995), and supports the improvement of R & D policies. In Australia, Big Data is used as a strategic tool to provide enhanced services at the existing level <sup>65</sup>. In Australia, data are provided to researchers by uniquely identifying and combining individual data such as medical use data, cancer registration data, demographic data, birth and death data, and parents' information. These activities are conducted through the Big Data Working Group.

#### **3.3.2 Privacy**

The Privacy Amendment Act was amended in 1990 and the Privacy Amendment (Private Sector) Act was enacted in 2000, and it now functions as a general law that regulates the public and private sectors together (Australian Privacy Act, Schedule 3). As in the United States, it is the industry's responsibility to make personal identification items self-judging. Data masking, pseudonymization, aggregation, derived data items, and banding for identification of personal information. Unsupervised personal data is no longer treated as

personal information, so no separate consent is required for its use under the Data Protection Act.

### **3.3.3 Research**

In Western Australia, data sources are provided in a timely manner through distributed management in the management institutions, and in the Ministry of Health and the research organizations under the supervision of researchers <sup>66</sup>. It is possible to maintain confidentiality by not using a personal identification code, and it is possible to carry out research for the purpose of public interest by concluding a memorandum of understanding once <sup>67</sup>. If separate approval is required, it must be granted by the Ethics Committee of the Western Australian Department of Health, and all staff and researchers must sign a confidentiality agreement and go through the formal approval process from the data manager before the data are made available to researchers<sup>6</sup>.



### **3.4 Taiwan**

#### **3.4.1 Governance**

Taiwan has a similarity with Korea in that it operates a medical dualism system in which traditional medicine and modern medicine coexist. Taiwan started single national health insurance on 1995 and collects health insurance claims big data called the National Health Insurance Research Database (NHIRD). It is used as a basis for policy decisions in the healthcare sector and covers 99% of Taiwan population. Therefore, Taiwan is the most active in the area of healthcare in Asia<sup>68,69</sup> and is an important example for us to utilize Big Data in the future.

#### **3. 4. 2 Privacy**

Theoretically, it is impossible to query the data alone to identify individuals at any level using this database. All researchers who wish to use the NHIRD and its data subsets are required to sign a written agreement declaring that they have no intention of attempting to obtain information that could potentially violate the privacy of patients or care providers. The use of NHIRD is limited to research purposes only. Applicants must follow the Personal Information Protection Act, (PDPA) integrated and renamed from Computer-Processed Personal Data Protection Law (CPPDPL) and<sup>70</sup> related regulations of National Health Insurance

Administration and NHRI (National Health Research Institutes), and an agreement must be signed by the applicant and his/her supervisor upon application submission.

### **3.4.3 Research**

The NHIRD is large population-based database which can inform us on the prevalence, incidence, natural history, treatment, correlates, and associations of disease, as well as the pattern of health care utilization. It includes 'insurer registration data' and 'health insurance claim data' for reimbursement of medical care costs. Sampling data on health claims data are being constructed in various ways, from monthly sample data extracted on the basis of claims, one-year sample data extracted on patient basis, and 5-year unit panel data. Since 2003 in Taiwan, the National Health Insurance Research Database has been providing the name to the private sector ahead of Korea. Using these data, 2,638 papers were published from 2003 to 2015 <sup>71</sup>. In 2011, the Center for Health Information Cooperation (CCHIA) was established with the aim of improving the quality of public health policy and supporting related academic research, linking data from the NHRI and other public institution data <sup>72</sup>.

### **3.5 Korean**

#### **3.5.1 Governance**

In Korea, with the rapid growth of information technology and changes in the medical market, the introduction of electronic medical records in medical institutions has spread and the basis for medical informatization has been established. The government of the Republic of Korea, after 2013, has processed various databases of the public sector under the banner of "Government 3.0" and provides it to the public.

As part of national strategy for Healthcare Information Use, “Health Insurance Big Data Connection Platform Project” was conducted to link and integrate data scattered among various institutions in 2016–2017 by Korean Ministry of Health and Welfare. However, it was difficult to carry out the project due to various problems; strict privacy regulation, poor data quality & standardization and lack of strategy and governance.

Governance previously operating in Korea includes the Public Data Strategy Committee of the Prime Minister and the Personal Information Protection Committee of the President.

### 3.5.2 Privacy

In Korea, personal information protection has been continuously strengthened through the enactment of the Personal Information Protection Act (PIPA), and the use of big data containing personal information is becoming increasingly difficult. Since the Personal Information Protection Act is a general law in the application of the law on personal information, the provisions of this Act shall take precedence except as otherwise provided in other laws.

Under current legal system, personal information can be used only for statistical analysis and academic research purposes, and before provision, information must be unidentified or deleted. However, in the case of deleting the individual identification information, it is difficult to associate the individual with specific data, and there are restrictions on combining the data between institutions.

Although Korea has excellent healthcare data, the linkage and integration system between institutions is insufficient. Since personal data such as social security numbers are required for data connection, guidelines based on the identification of personal information have been developed in Korea<sup>73-76</sup>. However, there is a limit to utilize it because it does not contain concrete contents about data linkage.

### **3.5.3 Research**

According to the Ministry of Health and Welfare of Korea, 4.6 billion will be invested in a project to build a foundation for research in connection with healthcare information held by public institutions in 2018. The project is divided into two major areas.

2.4 billion will be invested in building a network to share and analyze data between institutions. This is the basis for promoting the results of analysis of data held by medical institutions and public institutions. Second, 1.9 billion will be invested in Health Insurance Big Data linkage and utilization enhancement research to support R & D for dataset and service development using the platform.

### **3.6. Summary (US, UK, Australia, Taiwan, Korea)**

The following table summarizes comparative analysis of big data policies in overseas major countries (US, UK, Australia, Taiwan, and Korean) discussed above. As a result of examining the cases of developed countries, the important points to be reflected in the strategy formulation are as follows: public good purposes, strengthening and balancing of data security, building national big data governance, ensuring the right to informational self-determination, and improving access to patient information

**Table 7. Overview of health related big data in foreign countries**

Class	Korea	Australia	Taiwan	US	UK
Grade	☆☆☆☆	★★☆☆	★★★★☆	★★★★☆	★★★★★
Strategy	Healthcare Big data linking platform business (2018- 2020)	The Australian Public Service Big Data Strategy	DIGI+ 2025 (2017-2015)	Big Data Initiative - BD2K Big Data to Knowledge) (2012-present)	Personalized Health and Care 2020 (2013-present)
Health care Provision	NHI (National Health Insurance)	NHS (National Health Service)	NHI (National Health Insurance)	NHI (National Health Insurance)	NHS (National Health Service)
Open data	<a href="https://www.data.go.kr/">https://www.data.go.kr/</a>	<a href="https://www.datalinkage-wa.org.au/">https://www.datalinkage-wa.org.au/</a>	<a href="https://data.gov.tw/">https://data.gov.tw/</a>	<a href="https://www.data.gov/">https://www.data.gov/</a>	<a href="https://data.gov.uk/">https://data.gov.uk/</a>
Law Ground	<ul style="list-style-type: none"> <li>• Constitution</li> <li>• Personal Information Privacy Act (PIPA)-2011</li> <li>• Act on promotion of the provision and use of public data-2013</li> </ul>	<ul style="list-style-type: none"> <li>• Federal Privacy Act -1998</li> <li>• Act on Personal Information Processing Regulations -2000</li> <li>• Australian Privacy Principles (2012): 13 Principles</li> </ul>	<ul style="list-style-type: none"> <li>• Computer-Processed Personal Data Protection Act (CPDPA)-2007</li> <li>• Personal Information Protection Act (PDPA)-2012</li> <li>• The Freedom of Government Information Law-2005</li> </ul>	<ul style="list-style-type: none"> <li>• HITECH Act (Health Information Technology for Economic and Clinical Health Act)</li> <li>• HIPAA (Health Insurance Portability and Accountability Act)</li> </ul>	<ul style="list-style-type: none"> <li>• Health and Social Care Act (2012)</li> <li>• Care Act</li> <li>• Data Protection Act</li> </ul>
National Big data	<ul style="list-style-type: none"> <li>• Health Insurance Review and Assessment Service (HIRA) <a href="http://opendata.hira.or.kr/home.do">http://opendata.hira.or.kr/home.do</a></li> <li>• National Health Insurance Service (NHIS) <a href="https://nhiss.nhis.or.kr/bd/ab/bdaba01leng.do">https://nhiss.nhis.or.kr/bd/ab/bdaba01leng.do</a></li> </ul>	<ul style="list-style-type: none"> <li>• Western Australian Data Linkage System (WADLS)</li> </ul>	<ul style="list-style-type: none"> <li>• National Health Insurance Research Database (NHIRD) <a href="http://nhird.nhri.org.tw/en/Research.html">http://nhird.nhri.org.tw/en/Research.html</a></li> <li>• Cell Bank</li> <li>• Health Research Information Network (HINT)</li> <li>• Bioinformatics (Gene Bank, GDB, Swiss-prot, ExPSAY, GCG)</li> </ul>	<ul style="list-style-type: none"> <li>• CMS Chronic Conditions Data Warehouse(CCW)</li> <li>• Surveillance, Epidemiology, and End Results-Medicare data (SEER-Medicare data)</li> <li>• NIH-NHGRI eMERGE network (Electronic Medical Records and Genomics)</li> <li>• Research Data Center</li> <li>• Ginger.io <a href="https://ginger.io">https://ginger.io</a></li> <li>• Wellpoint</li> </ul>	<ul style="list-style-type: none"> <li>• Clinical Practice Research Data (CPRD) <a href="http://www.cprd.com/governance">http://www.cprd.com/governance</a></li> <li>• i-sense <a href="http://www.i-sense.org.uk">http://www.i-sense.org.uk</a></li> <li>• Connecting care</li> <li>• Summary Care Record (SCR)</li> </ul>
Governance	—	<ul style="list-style-type: none"> <li>• Information Management Agency</li> <li>• Western Australian Data linkage</li> </ul>	<ul style="list-style-type: none"> <li>• CCHIA (Collaboration Center of Health Information Application)</li> <li>• Bureau of National Health Insurance in Taiwan.</li> </ul>	<ul style="list-style-type: none"> <li>• ONC (Office of the National Coordinator)</li> <li>• Data Science Institute</li> <li>• NIH Scientific Data Committee</li> </ul>	<ul style="list-style-type: none"> <li>• NHS Digital = HSCIC (Health &amp; Social Care)</li> <li>• National Information Commission</li> <li>• National Statistical Office</li> <li>• Ministry of Health</li> </ul>

\*Source: Kang, Study on basic plan for utilization of healthcare Big Data, 2015<sup>77</sup>

Source: Yoon, Health industry trend research and issue analysis, 2016<sup>78</sup>

## Chapter 4. Analysis in use of Health Insurance Claims Data

### 4.1 Use-cases of Health Insurance Claims Data

Health related services provided by HIRA, hospital information service, non-payment medical information service, pharmacy information service, detailed information service by medical institution, medical treatment information service, has been developed through applications and websites; 'Ask your doctor for thyroid cancer (please ask)', 'Find a hospital', 'Find Pharmacy', 'Find Doctor', 'Haidak', 'Gooddak', 'Medilate'. Both institutions provided data for research, disease prediction, and personal health-related services to the public (Table 8). The use of health insurance claim big data can be largely utilized in the following three ways.

**Table 8. Use-cases of Health Insurance Claims Data**

Class	NHIS	HIRA
Use as research data	<ul style="list-style-type: none"> <li>• Opening Sample Cohort DB</li> <li>• Providing customized materials</li> <li>• Health Insurance Big Data Platform</li> </ul>	<ul style="list-style-type: none"> <li>• Big Data Utilization Research Project</li> <li>• Study of Drug Interaction Detection</li> <li>• Verification of the consistency between KNHANES and HIRA NPS</li> <li>• Virtual Big Data Analysis Platform</li> </ul>
Disease forecasting service	<ul style="list-style-type: none"> <li>• Infectious disease monitoring system using Big Data</li> </ul>	<ul style="list-style-type: none"> <li>• DUR real time information</li> </ul>
Personal health risk assessment	<ul style="list-style-type: none"> <li>• Personal Health Record (PHR) Service</li> <li>• Provide health checkups and medical use indicators</li> <li>• National Health Alarm Service</li> </ul>	<ul style="list-style-type: none"> <li>• Health Information Application (App) Ex. Haidak, Gooddak, Medilate</li> <li>• Customized Hospital Finder</li> <li>• Medical Rewards</li> </ul>
Other	<ul style="list-style-type: none"> <li>• Crime forecasting service using scientific investigation</li> </ul>	

Source: Kang, Study on basic plan for utilization of healthcare Big Data, 2015<sup>77</sup>

First, health insurance claims big data can be used for health promotion and disease prevention. Currently, the use of health claim data in the aspect of health promotion and disease prevention in Korea focuses on clinical decision support (knowledge of cause and mechanism of disease and disease monitoring) mainly through building knowledge base. In Korea, combined with mobile health care, chronic diseases and health behavior management projects are actively being carried out.

Second, health insurance claims big data can contribute to the improvement of healthcare value. It can be used to improve the system for improving service quality by linking scattered medical institutions and insurance claim data. It is possible for an individual to manage the details received from the individual medical institutions so that unnecessary medical care can be prevented and medical care can be linked.

Third, the health insurance claims data contribute to the development of the healthcare industry. In terms of industrial development, it is possible to contribute to health promotion and disease prevention by using health information that directly inputs individual's medical record in combination with mobile industry. It uses a supercomputer for various clinical guidelines and research information, various medical journals can contribute to clinical decision support and personalized treatment.



Use-case results for health insurance claims big data are presented in Table 9. After analyzing the responses in the first round, 15 items were asked for evaluation in the second round. According to the results of the first questionnaire, 8 items were answered with strong consent (95%). The proposed use cases are explained by 4Ps<sup>79-81</sup> of 'Personalized Medicine'<sup>82-84</sup>; Predictive, Preventive, Personalized, Participatory.

The participants thought that it was most necessary in the case of “Information for selecting the medical center.” This use-case reflects “preventive” and “participatory” characteristics when using Big Data in the fourth industrial revolution. “General screening instead of health screening” ranked second in the first round, and “Personal life style monitoring” and “Personal health risk assessment” ranked third. These use cases reflect the characteristics of “preventive” and “personalized” when using Big Data.

While “Emergency risk management system” and “Adverse drug reaction detection using DUR” were recommended by one expert in the first round, they ranked first and second in the second round, respectively. “Life cycle and time-dependent disease prediction services” reflects the characteristics of “Predictive” and “Preventive”<sup>85</sup>.

**Table 9. The use cases of Health Insurance Claims Data in the public health field**

ID	Item	First round		Second round	
		N (%) <sup>1</sup>	Rank <sup>3</sup>	Consensus <sup>2</sup> (%)	Rank <sup>3</sup>
<b>1</b>	<b>Information for selecting the medical center</b>	<b>9 (21.43)</b>	<b>1</b>	<b>40 (95.24)</b>	<b>4</b>
2	Evidence for health policy development	1 (2.38)	5	38 (90.48)	9
3	Unfair billing agency monitoring	1 (2.38)	5	36 (85.71)	11
4	Disease forecasting service by region	1 (2.38)	5	38 (90.48)	7
5	<b>Disease forecasting service by time (season)</b>	<b>3 (7.14)</b>	<b>4</b>	<b>40 (95.24)</b>	<b>4</b>
6	<b>Disease forecasting service by life cycle</b>	<b>4 (9.52)</b>	<b>3</b>	<b>41 (97.62)</b>	<b>2</b>
7	<b>Adverse drug reaction detection using DUR</b>	1 (2.38)	5	<b>41 (97.62)</b>	<b>2</b>
8	<b>Personal life style monitoring</b>	<b>4 (9.52)</b>	<b>3</b>	<b>40 (95.24)</b>	<b>4</b>
9	<b>Personal health risk assessment</b>	<b>4 (9.52)</b>	<b>3</b>	<b>40 (95.24)</b>	<b>4</b>
10	Information from the medical services available in one place (medical history, medications, and payments)	1 (2.38)	5	37 (88.10)	10
11	<b>General screening instead of health screening</b>	<b>6 (14.29)</b>	<b>2</b>	<b>40 (95.24)</b>	<b>4</b>
12	Crime forecasting service using scientific investigation	1 (2.38)	5	36 (85.71)	11
13	Determine body standards using biometric information	1 (2.38)	5	36 (85.71)	11
14	<b>Emergency risk management system</b>	1 (2.38)	5	<b>42 (100)</b>	<b>1</b>
15	Use as research data	2 (4.76)	4	37 (88.10)	10
16	No answer (Missing)	<b>2 (4.76)</b>			
Total		42 (100)			

<sup>1</sup>The number of participants who suggested each item as use-case

<sup>2</sup>The number of participants agreed to be important for items derived from first round.

<sup>3</sup>Rank for each item. “1” is the highest rank.

## 4.2 Issue analysis in utilizing health insurance claims data

### 4.2.1 PESTEL Analysis in utilizing health insurance claims data

Health insurance claims big data is evaluated as highly useful health information in terms of inclusiveness that reflects the patient's medical care path for all citizens<sup>86</sup>. A PESTEL analysis is recommended for identifying the internal and external factors that influence on using health insurance claims big data (Table 8). The factors are grouped into six categories: political, economic/financial, social, technical, legal, and environmental (PESTEL)<sup>87</sup>.

**Table 10. Analysis in use of health insurance claims data through PESTEL**

Factor	Condition analysis
<b>Political factor</b>	<ul style="list-style-type: none"> <li>• Absence of national strategy for healthcare use of public data</li> <li>• Lack of national level governance for healthcare data management</li> </ul>
<b>Economic factor</b>	<ul style="list-style-type: none"> <li>• Rising economic value through the use of research data</li> <li>• Creation of new jobs in the field of research data</li> </ul>
<b>Social factor</b>	<ul style="list-style-type: none"> <li>• Increased demand for personalized health and medical services</li> <li>• Social needs through health care big data-based solutions</li> </ul>
<b>Technological factor</b>	<ul style="list-style-type: none"> <li>• Establish a virtual system that allows researchers to analyze data</li> <li>• Linking skills with health insurance claims big data and external data</li> </ul>
<b>Environmental factor</b>	<ul style="list-style-type: none"> <li>• Healthcare paradigm shifts from treatment to prevention</li> <li>• Data disclosure and sharing environment according to government 3.0</li> </ul>
<b>Legal factor</b>	<ul style="list-style-type: none"> <li>• Legal immaturity related to use personal information as linkage key</li> <li>• Lack of legal guideline to resolve the conflicting use and protection of personal information</li> </ul>

#### 4.2.2 Obstacles in utilizing health insurance claims data

The participants suggested four obstacles to utilizing health insurance claims data (Table 11). In the first round, 20 (47.62%) considered “legal institutional immaturity for data use” to be an obstacle. In the second round, all participants agreed that it was a top-priority obstacle to overcome. “Technical constraints for data sharing” and “lack of consensus for data provision” were suggested to be obstacles by 10 participants (23.81%). The former obstacle received 97.62% of support with a higher priority (2.14) than the latter (2.79) in the second round. The fourth obstacle, “lack of governmental support for data utilization,” was suggested by two participants in the first round and was considered to have the lowest priority, yet received 76.19% of support.

**Table 11. Obstacles in utilizing health insurance claims big data**

Obstacles	First round	Second round	
	N <sup>1</sup> (%)	Consensus <sup>2</sup> (%)	Priority <sup>3</sup>
I. Legal institutional immaturity for data use	20 (47.62)	42 (100.00)	1.33
II. Technical constraints for data sharing	10 (23.81)	38 (90.48)	2.79
III. Lack of consensus for data provision	10 (23.81)	41 (97.62)	2.14
IV. Lack of governmental support for data utilization	2 (04.76)	32 (76.19)	2.93

<sup>1</sup>The number of participants who suggested each item as an important issue

<sup>2</sup>The number of participants agreed to be important for items derived from first round

<sup>3</sup>Average priority on a 9-point scale. “1” is the first priority.

### 4.3. Analysis results and strategy in Delphi survey

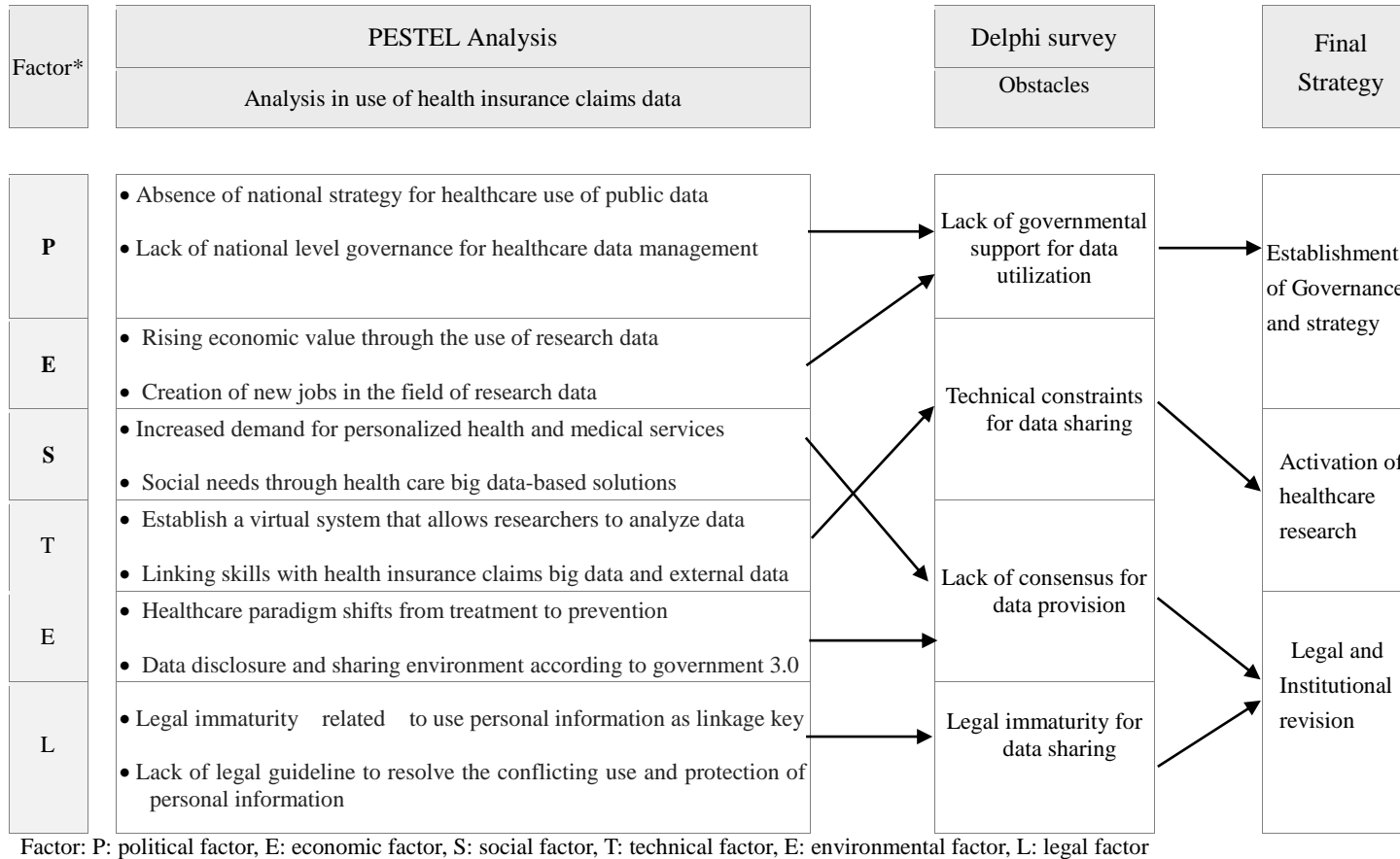


Figure 5. Analysis results and strategy in Delphi study

#### 4.4 Policy priorities to solve the obstacles

The participants proposed 13 policies to solve the four obstacles in the first round. Only one participant suggested “patients’ consent to the use of data” and “providers’ consent to the use of data” as policy on consensus for data provision in the first round. However, 42 (100.00%) and 40 participants (95.24%) agreed with these needs and ranked them as the first (1.86) and fourth (2.02) priorities in the second round, respectively.

The policies that were suggested as solutions to legal immaturity for data use, a “law revision policy” (including privacy information) and an “institutional improvement policy” (for the release of data by institutions), were ranked as the second and third priorities in the second round.

“Non-identification for information linkage” was suggested as the sole solution for technical constraints on data sharing, with the highest consensus (12 participants, 28.57%) in the first round. However, 38 participants (90.48%) ranked this policy as the fifth priority (2.14).

“Establishment of national governance for health data utilization” was suggested by five participants (11.90%). It was ranked as the sixth priority overall and had the highest priority among the policies suggested as solutions to the fourth obstacle, which concerned governmental support.

**Table 12. Policy priorities for utilizing health insurance claims big data**

Obstacle	Policy	First round	Second round	
		N <sup>1</sup> (%)	Consensus <sup>2</sup> (%)	Priority (Rank) <sup>3</sup>
I	Law revision policy	6 (14.29)	41 (97.62)	1.88 (2)
I	Institutional improvement policy	4 (9.52)	41 (97.62)	1.93 (3)
II	De-identification for data linkage	12 (28.57)	38 (90.48)	2.14 (5)
II	Expansion of infrastructure for managing data	2 (4.76)	38 (90.48)	2.55 (8)
II	Technical back-up for the data provider	1 (2.38)	39 (92.86)	2.36 (7)
III	Reward for the data provision	6 (14.29)	37 (88.10)	2.74 (10)
III	Patients' consent to data provision	1 (2.38)	42 (100.00)	1.86 (1)
III	Providers' consent to data provision	1 (2.38)	40 (95.24)	2.02 (4)
IV	Establishment of national governance for health data utilization	5 (11.90)	38 (90.48)	2.19 (6)
IV	Establishment of center to maintain data	1 (2.38)	32 (76.19)	2.98 (11)
V	Award for best practice of use case	1 (2.38)	30 (71.43)	3.05 (12)
V	Demonstrating effectiveness of Health Insurance Big Data	1 (2.38)	37 (88.10)	2.62 (9)
V	User training on the Health Insurance Big Data	1 (2.38)	40 (95.24)	2.55 (8)

<sup>1</sup>The number of participants who suggested each item as an important issue as policy solutions to solve the related obstacles in the first column

<sup>2</sup>The number of participants agreed to be important for items derived from first round

<sup>3</sup>Average priority on a 9-point scale. "1" is the first priority and "1" is the first rank.

## **Chapter 5. Strategies in use of Health Insurance Claims Data**

### **5.1 Establishment of National Big Data governance and strategy**

#### **5.1.1 Current diagnostics and problems**

Governance previously operating in Korea includes the Public Data Strategy Committee of the Prime Minister and the Personal Information Protection Committee of the President. The current state of governance related governance has been conducted in two opposing forms, supporting the nature of management reinforcement to protect patients' medical information and supporting the use of big data.

National governance has been pursuing the next plan to actively link and integrate domestic healthcare medical data and enhance its value; establish a platform to open big data held by public institutions including NHIS and HIRA, build a research platform for specialized medical centers (diabetes, dementia, etc.), active development of healthcare and health service models that can be put to practical use based on R & D research contents, establishment of information protection system such as purpose of use of information including anonymization and legal basis, expansion of R & D support based on big medical data.



### 5.1.2 Drawing implications from overseas cases

Since 2004, the United States has been promoting a federal government-based strategy, legislation, standards, data, and other healthcare data. The government's strategies include the Medical Information Technology Promotion Plan (04), the Big Data Promotion Plan ('12)<sup>88</sup>, and the Precise Health Promotion Plan ('16).

Since 2011, the HITECH Act has provided financial incentives for suppliers to 'meaningful use' of the EHR system. The EHR Incentive Program, which imposes penalties on suppliers who fail to show meaningful use since 2015, has been implemented. In the U.S, there is strategy of linking the development of precision medicine and improving quality of care. The ONC under the CMS collaborated to promote the strategy on the “EHR is a meaningful use strategy”.

Also, in the U.K there is strategy of integrating and sharing the segmented data into "patient-centered" to improve the outcomes of patient care at the NHS. “care.data strategy”. The NHS Digital (formerly HSCIC) is generated for national data integration and sharing with the aim of improving the outcomes of NHS medical care.

### 5.1.3 Development Direction and Strategy

It is necessary to establish “National Big Data Governance” for the successful utilization of health related big data. The value of health insurance claims big data increases when patient care is provided with large amounts of data that enable long-term analysis of various data, including personal level information and individual health determinants. “Big Data Governance”<sup>89,90</sup> should establish strategies for utilizing health related data and ensure the quality of the information and optimize the data results for decision-making<sup>90,91</sup>. In major countries, there is a strategy for expanding utilization of big medical data and a new governance structure to utilize national big data<sup>6,49,92</sup>.

Due to the complex and diverse environment of health care organizations, a multi-prong approach is needed regardless of the type of governance structure. Healthcare organizations prefer “a standing, hierarchical governance model” that manages information as well as data. This model consists of the following components; executive council, strategic committee, working groups.

In addition to support for regulations, trained personnel, and aggressive investment in infrastructure critical for maintaining them is also needed. These recommendations should be backed by Big Data Governance

## **5.2 Legal and Institutional revision for Big Data openness**

### **5.2.1 Current diagnostics and problems**

Health Information should be protected in terms of personal information. It is because the right of self-determination on his or her own personal information is guaranteed as a fundamental right in the constitution. Table 14 shows the status of the legal system related to personal information collection and handling of sensitive information. The efforts to resolve the dilemma whereby utilizing and protecting personal information are conflicted between each other have been made continuously.

In Korea, it allows limited use of personal information for government administrative purposes such as the Cancer Control Act, the Health and Medical Technology Promotion Act, and the Social Welfare Act. In December 2014, Big Data Privacy Guidelines were published by the Ministry of Health and Welfare<sup>93</sup>. However, the guideline is a guideline with normality that is the minimum judgment criterion rather than the compulsory one to observe. In accordance with the act on promotion of the provision and use of public data, it does not cover the general principles and management methods of data use and personal information processing in the big data era.

**Table 13. Status and contents of legal system related to personal information (Collection, use and sensitive information processing)**

Related Laws	Contents
Medical Service Act	<ul style="list-style-type: none"> <li>- (Article 22) The medical practitioner must record and sign medical treatment items and opinions such as the patient's main symptom, diagnosis and treatment contents.</li> <li>· Medical or medical institutions are allowed to collect and use patient's medical records without consent for medical purposes.</li> </ul>
National health Insurance law	<ul style="list-style-type: none"> <li>- (Article 96 (1) (2), Providing data) The Corporation and the HEDA may provide the data specified by the Presidential Decree for the performance of the health insurance business from the public institutions.</li> </ul>
Personal Information Protection Act	<ul style="list-style-type: none"> <li>- (Article 3: Protection Principle) Personal information is collected properly and legitimately only to the extent necessary for the purpose, and cannot be used for purposes other than purpose. Minimize the invasion of privacy of information subjects and possibly anonymity</li> <li>- (Article 15 (1): Collecting and Using) If the consent of the information subject is obtained, it may collect personal information for the purpose of carrying out the duties stipulated by laws and ordinances by the public agency</li> <li>- (Article 23, Article 24 (3)) When handling sensitive information and unique identification information, it can be processed only when it receives the informed consent of the information or it is permitted by the law. However, unique identification information needs to be taken to ensure safety, such as encryption.</li> </ul>
E-government Act	<ul style="list-style-type: none"> <li>- (Article 4 (2)) Personal information shall not be used against the will of the parties except as provided by laws and regulations.</li> <li>- (Article 36 (1)) Administrative agencies, etc. should be used jointly with other administrative agencies that need collected administrative information.</li> </ul>
Act on Promotion of Information and Communication Network utilization and Information Protection, etc.	<ul style="list-style-type: none"> <li>- (Article 23) When collecting personal information, only minimum personal information should be collected. Sensitive information is only allowed if it is permitted by the consent of the information entity or the law.</li> </ul>
Big Data Privacy Guideline	<ul style="list-style-type: none"> <li>- (Article 4) If personal information is included, it can be collected and used after non-identification action</li> <li>- (Article 7) Sensitive information can only be processed if it is permitted by the user's prior consent and law</li> </ul>
Medical institution privacy guidelines	<ul style="list-style-type: none"> <li>- Only the minimum amount of personal information necessary for medical purposes should be collected.</li> </ul>

Source: National Law Information Center (<http://law.go.kr/LSW/eng/engMain.do>)

### **5.2.2 Drawing implications from overseas cases**

The Privacy Act was established in 1974 by the U.S. federal government in order to protect the use of personal information without prior consent for the collection and storage of personal information. The US Health Insurance Portability and Accountability Act (HIPAA) is a federal law enacted in 1996 that standardizes the electronic exchange of medical administrative and financial data<sup>51</sup>.

The UK adopted the Information Protection Act for the first time in 1984, and adopted it again in 1998, reflecting the European Union member privacy guidelines. The UK laws related to the protection of personal information in the field of health care were established in Access to Medical Reports Act (1998) and the Access to Health Records Act (1990).

Australia is a federal state, and laws governing the protection of personal information are run in the federal and parking yards, respectively. The Australian Federal Privacy Act was enacted in 1988 and forms the basis of the Australian Privacy Act. In the collection of personal information, sensitive personal information is prohibited from collection but permits the collection of personal information if the person has his or her consent or if the information has a direct relationship with the function and role of the agency (Australian Privacy Principles-3 principle).

Exceptions to the prohibition of sensitive information collection by the Australian Privacy Act Schedule-3 are shown in the following table.

**Table 14 Exceptions to the prohibition of sensitive information collection**

No	Exceptions to the prohibition of collecting sensitive information
1	Research related to public health or public safety
2	Preparation or analysis of statistics related to public health or public safety
3	Collection of information that does not identify an individual in the case of health service management, fund appearance, or supervision, or information that cannot reasonably be identified as an individual's identity
4	Where it is impossible to obtain individual consent for the collection and the information is collected in accordance with the provisions of the law or in accordance with the rules set by the health organization or medical institution dealing with occupational confidentiality obligations applicable to the organization
5	collected in accordance with the guidelines approved by the Privacy Commissioner

Source: Australian Privacy Act

Sixth Principles In the use or disclosure of personal information, we have prohibited the use or disclosure of data for any purpose other than the original purpose of data collection. (I) the person has consented; (ii) the secondary purpose is related to the original purpose; the sensitive information is directly related to the original purpose; or (iii) it is used or disclosed if it is stipulated in Australian law. Seventh principles require that direct marketing does not use or disclose personal information. However, sensitive information may also be used and disclosed if it is agreed to use and disclose the information for marketing purposes.

### 5.2.3 Development Direction and Strategy

In accordance with the PIPA, only the consent of the information subject allows the linking and combining of information. This has hindered the availability of big data due to issues such as significant time and cost. On the other hands, foreign countries allow the use of medical information and personal information without the explicit consent of the information subject<sup>60,63,94-96</sup>.

First, the method of consent should be improved to resolve the dilemma whereby utilizing and protecting personal information. To solve dilemma, some governmental agencies or departments in South Korea should announce series of policies related to “comprehensive agreement”. Even if personal data agreed to the collection and use of data, if Big Data is used outside of the preliminary purpose, it is necessary to flexibly solve the requirements of notice and consent by obtaining a “comprehensive agreement”.

Also, consideration should be given to the opt-in / opt-out system by classifying the methods of consent according to risk. In other words, it is necessary to take measures to minimize the risk of flexible consent by introducing a “differentiated consent system” according to the level of confidentiality of personal information in the claim data. Finally, it is necessary to develop "Healthcare Big Data Usage Guidelines” in the framework of the separate big data laws.

## **5.3 Activation of healthcare research for Big Data linkage**

### **5.3.1 Current diagnostics and problems**

Berners-Lee emphasized not only the connection of existing data but also the spread of linked data<sup>97</sup>. The use of personal information is essential when linking billing data with other data. Previous studies regarding data linkage have used sensitive information as linkage variables<sup>6,98-101</sup>. Using resident registration numbers directly on data connections can be dangerous because of the high risk of re-identification<sup>102</sup>. It is urgent to provide service contents to the people with effective data linkage and diversification.

The academic databases provided by the HIRA and NHIS will become the infrastructure for future statistical system data and related institution data linkage. Currently, the two institutions provide pre-consultation at the time of application, and guidance on the purpose of analysis and application materials. The Big Data Open System operates a community of users who can exchange opinions while using data among users who are performing the same tasks<sup>103</sup>. However, it can be said that data users are limited in solving methodological difficulties or sharing information.



### 5.3.2 Drawing implications from overseas cases

In major countries, there is a mechanism to prevent re-identification. Anonymized information is combined with other information, and identification of an individual is called re-identification<sup>104</sup>. Examples of popular re-identification include re-identification of Massachusetts Governor medical records<sup>105</sup>, AOL query re-identification<sup>106</sup>, and Netflix movie review re-identification<sup>107</sup>. In order to prevent re-identification, there was a concern about the individual unique identifier such as Master Patient Identifier<sup>99,108</sup>, Master Linkage Key<sup>99</sup>, or Trusted Third Party Indexing<sup>60,109</sup>. In Europe, there is a Data Protection Officer who implements public data security policies including data privacy<sup>110,111</sup>.

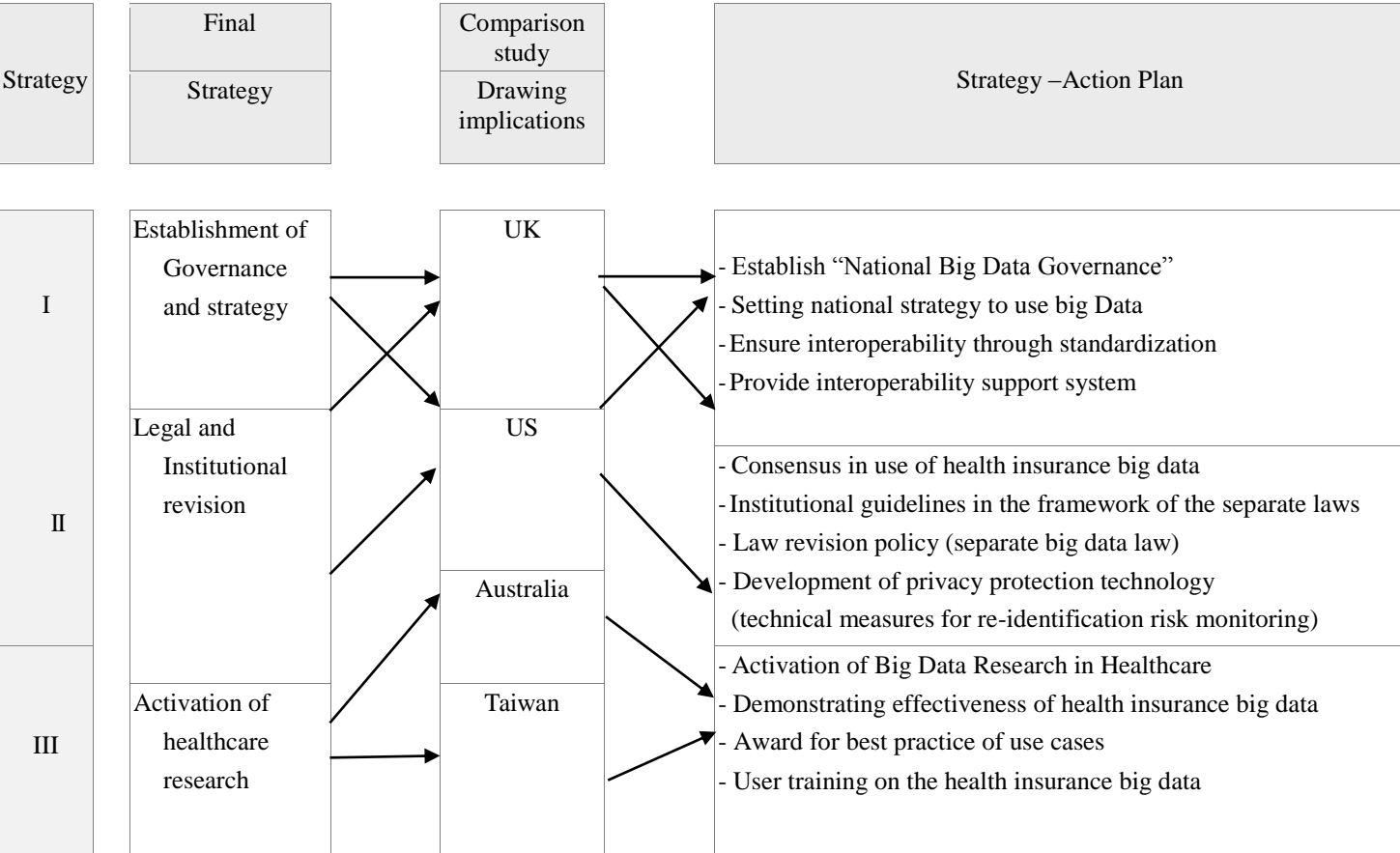
Major countries provide services for data linkage. In the case of CPRD, ISAC has methodological conditions for research projects that use data through advisory agencies<sup>112</sup>. Therefore, an advisory committee is needed to advise on the understanding of the data, the epidemiological and the statistical methodology. If the researchers make academic achievements using the requested data, a motivation strategy such as giving a certain incentive is needed to activate the use.

### 5.3.3 Development Direction and Strategy

First, privacy protection technology must be developed for health insurance claims big data open and sharing. In order to link data from various sources, it is necessary to clearly define the target and scope of the non-identification of personal identification codes. Therefore, technical measures for re-identification risk monitoring in response to non-identification measures should also be undertaken.

Second, user-centric service strategy that maximizes health insurance claims big data is needed. Finally, in the early stage where big data utilization is limited institutionally, it is necessary to expand the opportunity to integrate and analyze distributed data between participating institutions by inducing the establishment of multi-purpose data networks by various research subjects. The first major data project in the United States began with the Comparative Effectiveness Research (CER) project through the NIH and the Agency for Healthcare Research & Quality (AHRQ)<sup>113</sup>.

### 5.3.4 Analysis results and Strategy in comparative analysis



\* Factor: P: political factor, E: economic factor, S: social factor, T: technical factor, E: environmental factor, L: legal factor

**Figure 6. Analysis results and Strategy in comparative analysis**

## 5.4 Summary strategies for use of health insurance claims data

Since health claim data is administrative data, it can maximize its value through linkage with other data. In order to link health insurance claims data with other data, it is necessary to legal revision strategy related to privacy and consent, institutional improvement strategy to solve technical problems, establishment of governance for data management. After this strategy is backed up, data is collected, shared, and analyzed to yield meaningful results. The strategies fall into three categories, summarized below.



Figure 7. Summary strategies for use of health insurance claims data

## **Chapter 6. Discussion & Conclusion**

### **6.1 Discussion**

The results of the three methods of this study are as follows.

First, this study clearly demonstrates the academic benefits of a publicly available, Korean health insurance claims big data. It is one of the large-scale, nationwide administrative health care databases around the world. During the observation period (2007-2017), the number of publications using Korean health claims data has risen sharply since 2013 (Figure 4). Korean health insurance claims greatly encourage scientific production in a variety of research fields; including health service utilization, cost analysis and health policy (Table 6).

As we encourage the use of public data at the national level, it is necessary to track the use of similar databases in other countries and draw implications. In previous foreign studies, there are studies that show introduction or trend of healthcare claim data in Taiwan<sup>62,71</sup> and Germany<sup>114</sup>. In Taiwan, the National Health Insurance Research Database (NHIRD) has been producing government-oriented policy to encourage national or institutional level data holders to consider re-using their administrative databases for academic purposes<sup>69</sup>.

In order to maximize the value of health insurance claims big data in Korea, it is necessary to address related problems. Professional knowledge is still crucially

required in handling database specific problems<sup>115,116</sup> and technical barriers<sup>117</sup>. It will be necessary to establish a long-term plan for consumer-oriented data disclosure, linkage and utilization.

Second, this study shows the implications for policies in Korea through comparison of the big data utilization in the major countries. Although Korea has excellent health insurance claims big data and uses resident registration numbers, there are many restrictions on utilizing data. This is due to such limitations as lack of data governance, strict personal information protection law, insufficient linkage and integration systems between institutions (Table 7).

The experience of developed countries suggests important issues to be reflected in the formulation of strategies for national utilization of healthcare data. The ONC has been created to support interoperability in US. The NHS Digital was established to promote “Personalized Health and Care 2020” in UK. The CCHIA was established with the aim of improving the quality of public health policy in Taiwan. Finally, national strategy and data governance are needed to manage and utilize health insurance claim big data in Korea.

Also, health insurance claims big data should be focused on the use of the public good, such as improving the public interest. In the US, the public system is pursuing a variety of information utilization projects (meaningful use of EHR,

medical quality evaluation including patient-centered medical linkages)<sup>49</sup>. The NHS in UK is encouraging people to participate and choose to use medical information to provide better care (sharing medical information between doctors through care.data)<sup>94,96</sup>. The CCHIA in Taiwan support related academic research, linking data from the NHRI and other public institution data<sup>72</sup>.

Finally, there is a balance between strengthening and balancing privacy and data security. In Australia, there is a mechanism to strengthen sensitive data protection and data security in the country to encourage public trust and participation<sup>6,65,99</sup>. By guiding the use of information for public interest purposes, the right to opt-out of inclusion of his information in the utilization data is recognized.

Third, this study suggests 13 policies and 4 obstacles in using health insurance claims big data through Delphi survey. Participants responded by rating the four obstacles in this order: legal immaturity for data use, lack of consensus on providing information, technical constraints on information sharing, and lack of government support (Table 11). Policy priorities are as follows; a policy for the consent policy for data provision, a policy for law revision, an institutional improvement policy, technical policies such as anonymization for data sharing, and a national governance establishment policy (Table 12).

Since the healthcare industry in Korea is expanding around the national health insurance system, a strategic approach is needed to create industrial added value based on utilization in public systems. The use cases presented in this study are explained by 4P which characterized in customized medicine (Table 10). Many scholars emphasized the value of the use of healthcare data in personalized care<sup>118-120</sup>. The policy presented in this study is a key approach for activating the public utility of health insurance claims big data<sup>6</sup>. It needs to encourage public use on the basis of the policy presented in this study.

Finally, three strategies have been proposed for each issue derived from the three methodologies. First, it is necessary to establish “National Big Data Governance” for the successful utilization of health related big data. Second, it is necessary to develop legal institutional guidelines in the framework of the separate big data law (differentiation of personal information consent, development of legal and institutional guidelines). The method of consent should be improved to resolve the dilemma whereby utilizing and protecting personal information. Third, it is a strategy to revitalize healthcare research for big data linkage (development of personal information protection technology for data linkage, utilization of user - centered health insurance claim data).



## 6. 2 Limitation

This study had three major limitations. First, SR should be reviewed by two independent researchers, but there is a problem of validity because only one researcher performed this study. To overcome these limitations, same researcher extracted the literature twice over time.

Second, the participants included in the Delphi survey may not have been evenly distributed to the HIRA and NHIS, which may have had a biased effect on the response results. However, the number of participants is 42 respondents who were more than 30 experts, at least in qualitative research methodology.

Thirds, more diverse countries were not included in the comparison analysis. Among the OECD countries, country with high health information utilization rank Iceland, Korea, and Singapore in top 3<sup>121</sup>. A review of the countries not covered in this study should be reviewed in further study.

Despite many limitations, this study has the following strength.

Previous studies were mostly epidemiological studies using their own health insurance data at the HIRA and NHIS. There are no studies to consider NHIS data, only research that introduces HIRA data which was mainly limited to introducing HIRA-NPS data<sup>20,122</sup>. There is a lack of study on in-depth review on the use of

data compared to the interest in health insurance claims data<sup>123,124</sup>. The strategies for use of health insurance claim big data useful for researchers who are interested in using the data.

From 2018, the Ministry of Health and Welfare will conduct a pilot project on healthcare data for the next two years. This “Health Insurance Big Data Pilot Project” will link the healthcare information held by public institutions such as the NHIS and the HIRA for use for research purposes. This project has carried out in order to expand customized healthcare services through the establishment and utilization of the Korean medical care infrastructure by 2020. Currently, committee for public opinion that can discuss big-data-related issues is established under the Ministry of Health and Welfare.

Timely information is essential for policymakers to reach decisions. The strategy presented in this study is a key approach for activating the public utility of health insurance claims big data. Based on the strategy presented in this study, it is necessary to developing policies and long-term strategies<sup>125</sup>. In a similar context, this study provides useful advice on the current use of health insurance claims big data.

### 6.3 Conclusion

Globally, the rising cost of health care due to aging populations is threatening the sustainability of health care systems. Healthcare systems are shifting from volume-based to value-based. In major developed countries, the national registry is being developed and the interoperable health information system is being developed to realize valuable healthcare services. It is urgent to prepare to enable the strengths of Health Insurance Claims Big Data use through the mechanisms of national governance.

Korean health insurance big data can provide a powerful tool for evaluating use and outcome of a health care service, with due caution with regard to potential value. It can enhance the value as a valuable data source by completing the data in conjunction with other data found in government and the private sector. In order to increase the utilization value of data in the future, it is necessary to establish a strategy that returns the benefit to the public and balance with the industrial development.

## References

1. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: a systematic review. *JMIR medical informatics* 2016;4.
2. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *Jama* 2014;311:2479-80.
3. Heo JH, Suh DC, Kim S, Lee E-K. Evaluation of the pilot program on the real-time drug utilization review system in South Korea. *International journal of medical informatics* 2013.
4. Adler-Milstein C. Better Measurements for Realizing the Full Potential of Health Information Technologies.
5. National\_Evidence-based\_Healthcare\_Collaborating\_Agency. A round-table conference for utilizing personal health information while protecting its privacy and integrity
6. Brook EL, Rosman DL, Holman CAJ. Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System. *Australian and New Zealand journal of public health* 2008;32:19-23.
7. Hornnes E, Jansen A, Langeland Ø. How to develop an open and flexible information infrastructure for the public sector? *Electronic Government: Springer*; 2010. p.301-14.
8. Hudson KL, Collins FS. The 21st Century Cures Act—a view from the NIH. *New England Journal of Medicine* 2017;376:111-3.
9. Groves P, Kayyali B, Knott D, Van Kuiken S. The ‘big data’revolution in healthcare. *McKinsey Quarterly* 2013;2:3.
10. Omachonu VK, Einspruch NG. Innovation in healthcare delivery systems: a conceptual framework. *The Innovation Journal: The Public Sector Innovation Journal* 2010;15:1-20.
11. Glaser BG. The constant comparative method of qualitative analysis. *Social problems* 1965;12:436-45.
12. Linstone HA, Turoff M. *The Delphi method: Techniques and applications*. 1975.
13. Walshe K, Rundall TG. Evidence-based management: from theory to practice in health care. *The Milbank Quarterly* 2001;79:429-57.
14. Klassen TP, Jadad AR, Moher D. Guides for reading and interpreting systematic reviews: I. Getting started. *Archives of pediatrics & adolescent medicine* 1998;152:700-4.

15. Lobiondo-Wood G, Haber J, Krainovich-Miller B. The research process: Integrating evidence-based practice. *Nursing research: Methods and critical appraisal for evidence-based practice*. St. Louis, MO: Elsevier 2006;29-38.
16. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine* 2009;6:e1000097.
17. Lijphart A. Comparative politics and the comparative method. *American political science review* 1971;65:682-93.
18. Lee S-Y, Kim C-W, Seo N-K, Lee SE. Analyzing the Historical Development and Transition of the Korean Health Care System. *Osong public health and research perspectives* 2017;8:247.
19. Service NHI. National Health Insurance Sharing Service. Available at <https://nhiss.nhis.or.kr/bd/ab/bdaba011eng.do>
20. Kim L, Kim J-A, Kim S. A guide for the utilization of health insurance review and assessment service national patient samples. *Epidemiology and health* 2014;36.
21. Kim SJ, Lee J, Jee SH, Nam CM, Chun K, Park IS, et al. Cardiovascular risk factors for incident hypertension in the prehypertensive population. *Epidemiology and health* 2010;32.
22. Cho H-S, Kim Y-W, Park H-W, Lee K-H, Jeong B-G, Kang Y-S, et al. The relationship between depressive symptoms among female workers and job stress and sleep quality. *Annals of occupational and environmental medicine* 2013;25:12.
23. Kim M-Y, Jee SH, Yun JE, Baek SJ, Lee D-C. Hemoglobin concentration and risk of cardiovascular disease in Korean men and women-the Korean heart study. *Journal of Korean medical science* 2013;28:1316-22.
24. Lee T, Kim J, Kim S, Kim K, Park Y, Kim Y, et al. Risk factors for asthma-related healthcare use: longitudinal analysis using the NHI claims database in a Korean asthma cohort. *PloS one* 2014;9:e112844.
25. Yi S-W, Hong J-S, Ohrr H, Yi J-J. Agent Orange exposure and disease prevalence in Korean Vietnam veterans: the Korean veterans health study. *Environmental research* 2014;133:56-65.
26. Kang JI, Sung NY, Park SJ, Lee CG, Lee BO. The epidemiology of psychiatric disorders among women with breast cancer in South Korea: analysis of national registry data. *Psycho-Oncology* 2014;23:35-9.
27. Kang S, Kim H-S, Choi E-S, Han I. Incidence and treatment pattern of extremity soft

tissue sarcoma in Korea, 2009-2011: a nationwide study based on the health insurance review and assessment service database. *Cancer research and treatment: official journal of Korean Cancer Association* 2015;47:575.

28. Lim Y, Lee J-O, Bang S-M. Incidence, survival and prevalence statistics of classical myeloproliferative neoplasm in Korea. *Journal of Korean medical science* 2016;31:1579-85.
29. Kim SH, Joung JY, Suh YS, Kim YA, Hong JH, Kuark TS, et al. Prevalence and survival prognosis of prostate cancer in patients with end-stage renal disease: a retrospective study based on the Korea national database (2003–2010). *Oncotarget* 2017;8:64250.
30. Kim J, Hahm MI, Park EC, Park JH, Park JH, Kim SE, et al. Economic burden of cancer in South Korea for the year 2005. *J Prev Med Public Health* 2009;42:190-8.
31. Ko S-K, Yoon S-J, Oh I-H, Seo H-Y, Kim E-J. The economic burden of inflammatory heart disease in Korea. *Korean circulation journal* 2011;41:712-7.
32. Ahn IM, Park D-H, Hann HJ, Kim KH, Kim HJ, Ahn HS. Incidence, prevalence, and survival of moyamoya disease in Korea: a nationwide, population-based study. *Stroke* 2014;45:1090-5.
33. Kim J. Basis of using health insurance data, strategic, and assignment from the computation of health statistic. The 4th statistically innovation forum; Facts and figures on utilization of insurance statistics and pension statistics. Seoul: National Medicine Health Insurance; 2005.
34. Oh I-H, Yoon S-J, Seo H-Y, Kim E-J, Kim YA. The economic burden of musculoskeletal disease in Korea: a cross sectional study. *BMC Musculoskeletal Disorders* 2011;12:157.
35. Park K, Lee JS, Kim Y, Kim YI, Kim J. The socioeconomic cost of injuries in South Korea. *J Prev Med Public Health* 2009;42:5-11.
36. Jung H-k, Jang B, Kim YH, Park J, Park SY, Nam M-H, et al. Health care costs of digestive diseases in Korea. *The Korean Journal of Gastroenterology* 2011;58:323-31.
37. Lee S, Chung W, Hyun K-R. Socioeconomic costs of liver disease in Korea. *The Korean journal of hepatology* 2011;17:274.
38. Oh I-H, Yoon S-J, Yoon T-Y, Choi J-M, Choe B-K, Kim E-J, et al. Health and economic burden of major cancers due to smoking in Korea. *Asian Pacific Journal of Cancer Prevention* 2012;13:1525-31.
39. Song YS, Shim SR, Jung I, Sun HY, Song SH, Kwon S-S, et al. Geographic distribution of urologists in Korea, 2007 to 2012. *Journal of Korean medical science* 2015;30:1638-45.

40. Kim HK, Lee JB, Kim SH, Jo MW, Kim EH, Hwang JY, et al. Association of prediabetes, defined by fasting glucose, HbA1c only, or combined criteria, with the risk of cardiovascular disease in Koreans. *J Diabetes* 2016;8:657-66.
41. Lee Y-R, Moon K, Kim Y, Park S-Y, Oh C-M, Lee K-S, et al. Disability-adjusted life years for communicable disease in the Korean Burden of Disease Study 2012. *Journal of Korean medical science* 2016;31:S178-S83.
42. Moon S, Han JH, Bae G-R, Cho E, Kim B. Hepatitis A in Korea from 2011 to 2013: current epidemiologic status and regional distribution. *Journal of Korean medical science* 2016;31:67-72.
43. Rhee SY, Hong SM, Chon S, Ahn KJ, Kim SH, Baik SH, et al. Hypoglycemia and medical expenses in patients with type 2 diabetes mellitus: an analysis based on the Korea National Diabetes Program Cohort. *PLoS One* 2016;11:e0148630.
44. Jung S-Y, Hwang B, Yang BR, Kim Y-J, Lee J. Risk of motor vehicle collisions associated with medical conditions and medications: rationale and study protocol. *Injury prevention* 2017;23:356-.
45. Lim JU, Kim K, Kim SH, Lee MG, Lee SY, Yoo KH, et al. Comparative study on medical utilization and costs of chronic obstructive pulmonary disease with good lung function. *International journal of chronic obstructive pulmonary disease* 2017;12:2711.
46. Cho E, Kang MH, Choi KS, Suh M, Jun JK, Park E-C. Cost-effectiveness outcomes of the national gastric cancer screening program in South Korea. *Asian Pac J Cancer Prev* 2013;14:2533-40.
47. Lee H, Lee KS, Sim SB, Jeong HS, Ahn HM, Chee HK. Trends in Percutaneous Coronary Intervention and Coronary Artery Bypass Surgery in Korea. *The Korean journal of thoracic and cardiovascular surgery* 2016;49:S60.
48. Kang H-Y, Yoo H, Park W, Go U, Jeong E, Jung K-S, et al. Tuberculosis notification completeness and timeliness in the Republic of Korea during 2012–2014. *Osong public health and research perspectives* 2016;7:320-6.
49. Henricks WH. “Meaningful use” of electronic health records and its relevance to laboratories and pathologists. *Journal of pathology informatics* 2011;2.
50. Regenscheid A, Scarfone K. Recommendations of the National Institute of Standards and Technology. *NIST special publication* 2011;800:155.
51. Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. *Annu. Rev. Med.* 2006;57:575-90.

52. Barnes M, Kulynych J, Hermes C. HIPAA and Human Subjects Research: A Question & Answer Reference Guide: Barnett International; 2003.
53. Sankar PL, Parker LS. The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. *Genetics in Medicine* 2017;19:743.
54. Kaufman DJ, Baker R, Milner LC, Devaney S, Hudson KL. A survey of US adults' opinions about conduct of a nationwide Precision Medicine Initiative® cohort study of genes and environment. *PLoS One* 2016;11:e0160461.
55. Agency\_for\_Healthcare\_Research\_and\_Quality. Design of the nationwide inpatient sample (NIS). 2005.
56. Agency\_for\_Healthcare\_Research\_and\_Quality. Design of the nationwide inpatient sample (NIS). 2007.
57. CCW. Chronic Condition Data Warehouse. Available at <http://www.ccwdata.org/> [Accessed 6 November 2017]
58. ResDAC. Research Data Assistance Center. Available at <http://www.resdac.org/> [Accessed 6 November 2017]
59. POSNOTE HoP. Big data and public health. *Houses of Parliament POSNOTE* 2014:474.
60. Whyte A. Emerging infrastructure and services for research data management and curation in the UK and Europe. *Research data management. Facet, London* 2012:205-34.
61. Beresford AR. Privacy issues in geographic information technologies. *Frontiers of Geographic Information Technology: Springer*; 2006. p.257-77.
62. Chen Y-C, Wu J-C, Haschler I, Majeed A, Chen T-J, Wetter T. Academic impact of a public electronic health database: bibliometric analysis of studies using the general practice research database. *PloS one* 2011;6:e21404.
63. Mathur R, Grundy E, Smeeth L. Availability and use of UK based ethnicity data for health research. 2013.
64. Williams T, van Staa T, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Therapeutic advances in drug safety* 2012;3:89-99.
65. Summons P, Regan B. Social impact of big data in australian healthcare. *Les Cahiers du numérique* 2016;12:13-30.
66. Australia GoW. the WA Data Linkage System (WADLS) Available at



<http://www.datalinkage-wa.org2018>]

67. Bae J, Jee S, Nam M, Kim S, Park J, Choi H, et al. Round-table conference of data utilization for public good healthcare study. Seoul: National Evidence-based Healthcare Collaborating Agency 2011.
68. Hsing AW, Ioannidis JP. Nationwide population science: lessons from the Taiwan national health insurance research database. JAMA internal medicine 2015;175:1527-9.
69. Database TNHIR. Taiwan National Health Insurance Research Database. Available at <http://nhird.nhri.org.tw/en/Research.html> [Accessed June 17 2018]
70. Computer-Processed Personal Data Protection Act. Taiwan; 2007.
71. Chen Y-C, Yeh H-Y, Wu J-C, Haschler I, Chen T-J, Wetter T. Taiwan's National Health Insurance Research Database: administrative health care database as study object in bibliometrics. Scientometrics 2011;86:365-80.
72. Yu R-r. Data Sharing in Taiwan: Policies and Practice. IASSIST Conference; 2013.
73. Safety Motla. Guidelines for the Protection of Personal Information by Opening and Sharing Public Information by Opening and Sharing Public Information. Ministry of the Interior and Safety; 2013.
74. Agency KIT. Case Study on Identification of Personal Information for the Use of Big Data. Ministry of the Interior and Safety; 2014.
75. Safety Motla. Big Data Privacy Guideline. Ministry of the Interior and Safety; 2014.
76. Agency KIT. Guide for Self - Assessment of Personal Information Nomination Self - Assessment for Personal Information Nomenclature. Korea Information Technology Agency; 2014.
77. Kang H-J. Study on basic plan for utilization of healthcare Big Data. 2015.
78. Young YJ. Health industry trend research and issue analysis analysis. 2016.
79. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nature reviews Clinical oncology 2011;8:184.
80. Hamburg MA, Collins FS. The path to personalized medicine. New England Journal of Medicine 2010;363:301-4.
81. Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. New biotechnology 2012;29:613-24.

82. Katsios C, Roukos DH. Individual genomes and personalized medicine: life diversity and complexity. *Personalized Medicine* 2010;7:347-50.
83. Whirl-Carrillo M, McDonagh EM, Hebert J, Gong L, Sangkuhl K, Thorn C, et al. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics* 2012;92:414-7.
84. Jain K. Personalized medicine. *Current opinion in molecular therapeutics* 2002;4:548-58.
85. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: The next frontier for innovation, competition, and productivity. 2011.
86. Statistics OH. Data visualisation. Available at <http://www.oecd.org/dac/financing-sustainable-development/datavisualisations/>
87. Yüksel İ. Developing a multi-criteria decision making model for PESTEL analysis. *International Journal of Business and Management* 2012;7:52.
88. Marcinko DE, Hertico HR. *Financial Management Strategies for Hospitals and Healthcare Organizations: Tools, Techniques, Checklists and Case Studies*: CRC Press; 2013.
89. Sanders D. *7 Essential Practices for Data Governance in Healthcare*. 2013.
90. Fleissner B, Jasti K, Ales J, Thomas R. The importance of data governance in healthcare. 2014.
91. Agency NE-bHC. A foundational study for the system establishment to link data sources in healthcare. 2010. p.125.
92. Johnson JE. Big data+ big analytics= big opportunity: big data is dominating the strategy discussion for many financial executives. As these market dynamics continue to evolve, expectations will continue to shift about what should be disclosed, when and to whom. *Financial Executive* 2012;28:50-4.
93. Jung Y-Y. Legal protection and viewing of healthcare information. *Medical law* 2012;13:359-95.
94. Taskforce AD. *The UK administrative data research network: improving access for research and policy*. Economic and Social Research Council, London 2012.
95. Shadbolt N, O'Hara K, Berners-Lee T, Gibbins N, Glaser H, Hall W. Linked open government data: Lessons from data. gov. uk. *IEEE Intelligent Systems* 2012;27:16-24.
96. Biobank U. *UK Biobank ethics and governance framework*. version; 2007.

97. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. *International journal on semantic web and information systems* 2009;5:1-22.
98. Park K, Bae J, Lee H, Kim J, Jang E, Choi J, et al. A strategic study of data linkage for evidence development. Seoul: National Evidence-based Healthcare Collaborating Agency 2010.
99. Holman CDAJ, Bass JA, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Australian Health Review* 2008;32:766-77.
100. Herman A, McCarthy B, Bakewell J, Ward R, Mueller B, Maconochie N, et al. Data linkage methods used in maternally-linked birth and infant death surveillance data sets from the United States (Georgia, Missouri, Utah and Washington State), Israel, Norway, Scotland and Western Australia. *Paediatric and perinatal epidemiology* 1997;11:5-22.
101. Christen P, Churches T, Hegland M. Febrl—a parallel open source data linkage system. *Advances in Knowledge Discovery and Data Mining*; Springer; 2004. p.638-47.
102. Kim H, Kim S. Legislation direction for implementation of health information exchange in Korea. *Asia-Pacific Journal of Public Health* 2012;24:880-6.
103. HIRA. HIRA Big Data Open System [Accessed 17 June 2018]
104. Narayanan A, Shmatikov V. De-anonymizing social networks. *Security and Privacy, 2009 30th IEEE Symposium on*; IEEE; 2009. p.173-87.
105. Sweeney L. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics* 1997;25:98-110.
106. Barbaro M, Zeller T, Hansell S. A face is exposed for AOL searcher no. 4417749. *New York Times* 2006;9:8For.
107. Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. *Security and Privacy, 2008. SP 2008. IEEE Symposium on*; IEEE; 2008. p.111-25.
108. Jung Sun Park , Kwang Keun Lee , Seon Ju Ahn , Dong Kyun Park , Young Ran Jin , Han Seok Kim, et al. Establish infrastructure to foster medical-IT convergence industry. Korea Health Industry Development Agency; 2013.
109. Froomkin AM. The essential role of trusted third parties in electronic commerce. *Or. L. Rev.* 1996;75:49.
110. Fromholz JM. European Union Data Privacy Directive, *The. Berk. Tech. LJ* 2000;15:461.
111. Bignami F. Privacy and Law Enforcement in the European Union: The Data Retention

- Directive. *Chi. J. Int'l L.* 2007;8:233.
112. (NIHR) NfHR. CPRD Governance. Available at <https://www.cprd.com/governance/> [Accessed 17 June 2018]
  113. Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. *Annals of internal medicine* 2009;151:203-5.
  114. Hoffmann F. Review on use of German health insurance medication claims data for epidemiological research. *Pharmacoepidemiology and drug safety* 2009;18:349-56.
  115. Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part I. *Value in Health* 2009;12:1044-52.
  116. Harpe SE. Using secondary data sources for pharmacoepidemiology and outcomes research. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 2009;29:138-53.
  117. Diamond CC, Mostashari F, Shirky C. Collecting and sharing data for population health: a new paradigm. *Health affairs* 2009;28:454-66.
  118. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs* 2014;33:1163-70.
  119. Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. *Health affairs* 2014;33:1178-86.
  120. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. *Health affairs* 2014;33:1115-22.
  121. OECD. Recommendation of the OECD Council on Health Data Governance. OECD; 17 January 2017.
  122. Kim L, Sakong J, Kim Y, Kim S, Kim S, Tchoe B, et al. Developing the inpatient sample for the National Health Insurance claims data. *Health Policy and Management* 2013;23:152-61.
  123. Kim J, Yoon S, Kim L-Y, Kim D-S. 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. *Journal of Korean medical science* 2017;32:718-28.
  124. Kim R. Introduction of HIRA-NPS (Health Insurance Review and Assessment Institute).

HIRA Research:37.

125. Baek H, Park S-K. Sustainable development plan for Korea through expansion of green IT: policy issues for the effective utilization of big data. *Sustainability* 2015;7:1308-28.
126. Ko MJ, Jo AJ, Park CM, Kim HJ, Kim YJ, Park D-W. Level of blood pressure control and cardiovascular events: SPRINT criteria versus the 2014 hypertension recommendations. *Journal of the American College of Cardiology* 2016;67:2821-31.

## **Appendix**

Appendix A. Delphi questionnaire (1 round)

Appendix B. Delphi questionnaire (2 round)

Appendix C. Summary of SR studies

## **Appendix A. Delphi questionnaire (1st round)**

Hello!

I am a student majoring in health science at Yonsei University.

I am a studying Health Insurance Big Data in the era of the fourth industrial revolution. In particular, I will study the legal and ethical constraints on the use of public information and conduct research to establish ways of making reasonable use of Health Insurance Big Data.

This survey is based on the Delphi method and will be conducted in two rounds. I would like to ask you to reply sincerely to the questionnaire so that it will help me carry out research.

In addition, I promise that this questionnaire is only for the collection of basic data for research purposes and will not be used in any form other than this research purpose. I would like to thank you once again for your valuable time and wish you all the best for your happiness.

※ If you have any questions regarding this survey, please contact the following address.

Tel: (02) -2182-2528

E-mail: happiness630@hiramail.net

■ The following is an open questionnaire.

1. Please freely write down Health Insurance Big Data examples using public health information as shown in the following example.

[write here ]

2. What do you think is the first thing you need in order to realize the case like first question?

[write here ]

3. What do you think is the most limiting factor in utilizing public information? [Number]

- ① Lack of consent of the institutions
- ② Personal information privacy issue
- ③ Legal and ethical constraints
- ④ Lack of government support
- ⑤ Others (Please specify if you have any other)

4. What policy do you think should be the first priority to promote Health Insurance Big Data?

[ write here ]

## **Appendix B. Delphi questionnaire (2nd round)**

Hello!

I am a student majoring in public health at Yonsei University.

In order to understand issues of health care big data and how to use it reasonably, we conducted a first survey for three weeks from February 24th. This questionnaire is a second questionnaire based on an analysis of the responses of the 42 respondents in the first Delphi survey. This questionnaire is the same as the first questionnaire survey. It consists of the statistical data of 42 specialists who responded to the first questionnaire, and the answers on the first questionnaire. Even if you are busy, please join us once again, as it will be very helpful to my research. Thank you for your cooperation.

In addition, I promise that this questionnaire is only for the collection of basic data for research purposes and will not be used in any form other than this research purpose. I would like to thank you once again for your valuable time and wish you all the best for your happiness.



Based on the answers of the open questions held in the first round of the survey, the results of the responses were listed reflecting the duplicate answers. Please rank them according to their importance.

1. Please describe the big data case using public health information.

Health Insurance Big Data Use Case				Consensus			Priority	Reference
				Indicate whether you agree with an item			estimate the degree to which the item is considered important	The first answer
No	class	serial	Item	agree	disagree	reason for disagree	1<----->9 Average of priority of 9 point scale. "1" is the first priority.	
1	Provide Information	1-1	Information for selecting the medical center	1	2		1 2 3 4 5 6 7 8 9	9
		1-2	Evidence for Health Policy Development	1	2		1 2 3 4 5 6 7 8 9	3
		1-3	Unfair billing agency monitoring	1	2		1 2 3 4 5 6 7 8 9	1
2	Forecasting system	2-1	Disease forecasting service by region	1	2		1 2 3 4 5 6 7 8 9	1

		2-2	Disease forecasting service by time (season)	1	2		1	2	3	4	5	6	7	8	9	4
		2-3	Disease forecasting service by life cycle	1	2											1
		2-4	Adverse drug reaction detection using DUR	1	2		1	2	3	4	5	6	7	8	9	1
3	customized medical services	3-1	Personal life style monitoring	1	2		1	2	3	4	5	6	7	8	9	4
		3-2	Personal health risk assessment	1	2		1	2	3	4	5	6	7	8	9	4
		3-3	Information from the medical services available in one place (medical history, medications, and payments)	1	2		1	2	3	4	5	6	7	8	9	1
4	Utilized by related organizations	4-1	General screening instead of health screening	1	2		1	2	3	4	5	6	7	8	9	2
		4-2	Crime forecasting service using scientific investigation	1	2		1	2	3	4	5	6	7	8	9	1

		4-3	Determine body standards using biometric information	1	2		1	2	3	4	5	6	7	8	9	1
		4-4	Emergency risk management system	1	2		1	2	3	4	5	6	7	8	9	1
5	research	5-1	Use as research data	1	2		1	2	3	4	5	6	7	8	9	2
6	Others															Unanswered (4)

2. What do you think is the first thing you need to realize the same thing as number one?

Health Insurance Big Data Use Case				Consensus			Priority									Reference
				Indicate whether you agree with an item			estimate the degree to which the item is considered important									The first answer
No	class	serial	Item	agree	disagree	reason for disagree	1<----->9 Average of priority of 9 point scale. "1" is the first priority.									
1	Sufficient rewards and incentives	1-1	Public awareness campaign	1	2		1	2	3	4	5	6	7	8	9	5
		1-2	Financial incentives	1	2		1	2	3	4	5	6	7	8	9	7
		1-3	Suggest ways to utilize healthcare data	1	2		1	2	3	4	5	6	7	8	9	1
2	Legal basis needed	2-1	regulations on the provision of information	1	2		1	2	3	4	5	6	7	8	9	4
		2-2	Need for scope of institutional information disclosure.	1	2		1	2	3	4	5	6	7	8	9	4
		2-3	Revision laws such as the PIPA	1	2		1	2	3	4	5	6	7	8	9	6

3	Provide a separate governance to manage Health Insurance Big Data	3-1	Data collection and processing	1	2		1	2	3	4	5	6	7	8	9	2
		3-2	Standardization and quality management of collected data	1	2		1	2	3	4	5	6	7	8	9	3
		3-3	Finding Healthcare Big Data	1	2		1	2	3	4	5	6	7	8	9	3
		3-4	Establishment of medium and long-term plans	1	2		1	2	3	4	5	6	7	8	9	1
4	Other	4-1	Training of experts in the Big Data	1	2		1	2	3	4	5	6	7	8	9	1
		4-2	Health Insurance Big Data Efficiency Measurement	1	2		1	2	3	4	5	6	7	8	9	1
5	other	Other														Unanswered (3)

### 3. What do you think is the most obstacles in utilizing Health Insurance Big Data?

Obstacle		Consensus			Priority	Reference
		Indicate whether you agree with an item			estimate the degree to which the item is considered important	
No	item	Agree	disagree	reason for disagree	1<----->9 Average of priority of 9 point scale, “1” is the first priority.	The first answer
1	Legal institutional immaturity for data use	1	2		1 2 3 4 5 6 7 8 9	10
2	Technical constraints for data sharing	1	2		1 2 3 4 5 6 7 8 9	22
3	Lack of consensus for data provision	1	2		1 2 3 4 5 6 7 8 9	10
4	Lack of governmental support for data utilization	1	2		1 2 3 4 5 6 7 8 9	2
5	Other	Please feel free to write if you have any other				

**4. What policy do you think should be the first priority to promote Health Insurance Big Data?**

Priority to promote Health Insurance Big Data				Consensus			Priority	Reference
				Indicate whether you agree with an item			estimate the degree to which the item is considered important	The first answer
No	class	serial	Item	agree	disagree	reason for disagree	1<----->9 Average of priority of 9 point scale. "1" is the first priority.	
1	Legislative revision strategy	1-1	Law revision policy	1	2		1 2 3 4 5 6 7 8 9	5
		1-2	Institutional improvement policy	1	2		1 2 3 4 5 6 7 8 9	
2	Institutional Improvement strategy	2-1	De-identification of policies for data sharing	1	2		1 2 3 4 5 6 7 8 9	6
		2-2	Expansion of infrastructure for managing data	1	2		1 2 3 4 5 6 7 8 9	4
		2-3	Technical back-up for the data provider	1	2		1 2 3 4 5 6 7 8 9	

3	Consent revision strategy	3-1	Reward for the data provision	1	2		1	2	3	4	5	6	7	8	9	5
		3-2	Patients' consent for data provision	1	2		1	2	3	4	5	6	7	8	9	1
		3-3	Providers' consent for data provision	1	2		1	2	3	4	5	6	7	8	9	1
4	Governance building strategy	4-1	Establishment of national governance for public health data utilization	1	2		1	2	3	4	5	6	7	8	9	
		4-2	Establish an center to maintain data	1	2		1	2	3	4	5	6	7	8	9	
5	Other strategy	5-1	Award for best practice of use case	1	2		1	2	3	4	5	6	7	8	9	
		5-2	Demonstrating effectiveness of Health Insurance Big Data	1	2		1	2	3	4	5	6	7	8	9	
		5-3	User training on the Health Insurance Big Data	1	2		1	2	3	4	5	6	7	8	9	



■ The following is information about the respondent. Please indicate or list the applicable items.

1. What is your gender? [     ]

① male ② female

2. What is your age range? [     ]

① 20 units ② 30 units ③ 40 units ④ 50 units or more

3. Please enter your occupation, the name of your institution, and the number of years you have worked.

Occupation [             ] Organization Name [             ] Working years [     ]

4. What is the department of your major? [     ]

① Humanities and Social Sciences ② Engineering ③ Department of Natural Health ④ Department of Arts and Physical Education

**Thank you for answer.**

### Appendix C. Cases were connected with external data in SR studies

Research type	Study	Title	Link Source	Main Finding
Cost analysis	Kim, Hahm et al. 2009 <sup>30</sup>	Economic burden of cancer in South Korea for the year 2005	National Cancer Center  KNHANES  Statistics Korea	To estimate the cost of cancer, use the data from the following four sources  - National Cancer Center: Cancer Registry  - National Health Insurance Corporation: Cancer qualification data,  -Korean National Health Insurance Corporation : Cost data and qualification related to the cancer,  - Korea National Health and Nutrition Examination Survey (KNHANES): Clinical data  - Statistics Korea: Statistics of causes of death
Cost analysis	Park, Lee et al. 2009 <sup>35</sup>	The socioeconomic cost of injuries in South Korea	Automobile insurance  IACI  Statistics Korea	Estimate socio-economic costs by matching the patient's unique identifier  - Korean National Health Insurance Corporation : claims data  - Automobile insurance : claims data  - Industrial accident compensation insurance (IACI) : claims data  - Statistics Korea: Statistics of causes of death (2001-2003)

Cost analysis	Jung, Jang et al. 2011 <sup>36</sup>	Health care costs of digestive diseases in Korea	Statistics Korea	<ul style="list-style-type: none"> <li>- HIRA: claim data of patients with gastrointestinal diseases</li> <li>- Health Insurance Statistical Yearbook : the medical cost of gastrointestinal diseases</li> <li>- Statistics Korea: Statistics of causes of death</li> </ul>
Cost analysis	Ko, Yoon et al. 2011 <sup>31</sup>	The economic burden of inflammatory heart disease in Korea	Korea Health Panel (Cost data)  Statistics Korea	<ul style="list-style-type: none"> <li>- Korean Health Panel: The total costs of inflammatory heart diseases were estimated as the sum of direct medical care costs, direct non-medical care and indirect costs</li> <li>- NHIC claims data: a number of resources to obtain data, national health insurance statistics,</li> <li>- Korean National Statistical Office: the causes of death report.</li> </ul>
Cost analysis	Lee, Chung et al. 2011 <sup>37</sup>	Socioeconomic costs of liver disease in Korea	Korea Health Panel (Cost data)  Statistics Korea	<ul style="list-style-type: none"> <li>- NHIC claims data: Direct medical costs</li> <li>- Korea Health Panel study: Direct non-medical costs</li> <li>-Korean Statistical Information Service (KOSIS) : Indirect costs</li> <li>- Statistics Korea: annual report on the cause of death statistics</li> </ul>
Cost analysis	Oh, Yoon et al. 2011 <sup>34</sup>	The economic burden of musculoskeletal disease in Korea: a cross sectional study	Korea Health Panel (Cost data)	<ul style="list-style-type: none"> <li>To estimate the economic burden of musculoskeletal disease.</li> <li>- NHIC claims data: nationally representative of</li> </ul>

				<p>medical care costs covered by the Korean insurance</p> <p>- Korea Health Panel study: prevalence of musculoskeletal disease, proportion of costs, Direct non-medical costs</p>
Cost analysis	Cho, Kang et al. 2013 <sup>46</sup>	Cost-effectiveness outcomes of the national gastric cancer screening program in South Korea.	<p>Korea Health Panel (Cost data)</p> <p>Statistics Korea</p>	<p>- National Statistical Office: Mortality information (7-year follow-up period_</p> <p>-Korean National Health Insurance Corporation : Cost data related to the gastric cancer screening directly or indirectly were collected from the internal accounts of screening units in hospitals, published studies, and national statistics.</p>
Cost analysis	Ahn, Park et al. 2014 <sup>32</sup>	Incidence, prevalence, and survival of moyamoya disease in Korea: a nationwide, population-based study.	<p>Clinical data (Rare Intractable Disease registration program)</p>	<p>- HIRA: Data from nationwide, population-based claims database</p> <p>-Rare Intractable Disease registration program: physician-certified diagnoses based on uniform criteria for moyamoya disease from 2007 to 2011.</p>
Cost analysis	Lim, Kim et al. 2017 <sup>45</sup>	Comparative study on medical utilization and costs of chronic obstructive pulmonary disease with good lung function.	<p>KNHANES</p> <p>Clinical data (KOCOSS)</p>	<p>- Korea National Health and Nutrition Examination Survey (KNHANES): EuroQol 5-dimension questionnaire index scores of patients with COPD</p> <p>-HIRA: Data including the number of outpatient clinic visits, admission to hospitals, COPD-related medications, and medical costs</p> <p>- Korean COPD Subtype Study (KOCOSS) cohort: data of patients with COPD with FEV1 ≥60%</p>

Intervention and evaluation study	Kim, Lee et al. 2010 <sup>21</sup>	Cardiovascular risk factors for incident hypertension in the prehypertensive population	National Cancer Center	- NHIC data: Korean Cancer Prevention Study: The data from participants were examined at baseline and at follow-up health examinations in 1998, 2000, 2002, and 2004.
Intervention and evaluation study	Cho, Kim et al. 2013 <sup>22</sup>	The relationship between depressive symptoms among female workers and job stress and sleep quality.	Additional Survey data (KOSS-SF)	- National Health Insurance Service (NHIS) : worksite-based health checkup  - Korean Occupational Stress Scale-Short Form(KOSS-SF): questionnaire survey
Intervention and evaluation study	Kim, Jee et al. 2013 <sup>23</sup>	Hemoglobin concentration and risk of cardiovascular disease in Korean men and women - the Korean heart study.	Clinical data (examination centers)	-17 Korean nationwide health examination centers : Clinical data (physical examination)  - Korean National Health Insurance database : Data regarding CVD incidence
Intervention and evaluation study	Lee, Kim et al. 2014 <sup>24</sup>	Risk factors for asthma-related healthcare use: longitudinal analysis using the NHI claims database in a Korean asthma cohort.	Clinical data (Korean asthma cohort)	- Korean National Health Insurance database : asthma-related claims database  - Korean asthma cohort: 736 patients registered
Intervention and evaluation study	Yi, Hong et al. 2014 <sup>25</sup>	Agent Orange exposure and disease prevalence in Korean Vietnam veterans: the Korean veterans health study.	Clinical data (Korean Vietnam veterans)	- Korean Vietnam veterans: The Agent Orange exposure was assessed by a geographic information system-based model. A total of 111,726 were analyzed for  -Korea National Health Insurance claims data : prevalence

Intervention and evaluation study	Song, Shim et al. 2015 <sup>39</sup>	Geographic Distribution of Urologists in Korea, 2007 to 2012.	Statistics Korea	<ul style="list-style-type: none"> <li>- National Health Insurance Service : County level data</li> <li>- National Statistical Office : ecological study.</li> <li>- American Medical Association (AMA) Master file: the number of physicians</li> <li>- Population Census Division, National Statistical Office: Population data was obtained from</li> <li>- National Atmospheric Administration: local temperature</li> </ul>
Intervention and evaluation study	Ko, Jo et al. 2016 <sup>126</sup>	Level of Blood Pressure Control and Cardiovascular Events: SPRINT Criteria Versus the 2014 Hypertension Recommendations.	KNHANES	<ul style="list-style-type: none"> <li>- KNHANES (Korean National Health and Nutrition Examination Survey): 2008 -2013 patient(n = 13,346)</li> <li>- Korean National Health Insurance Service health examinee cohort: 2007 (n = 67,965)</li> </ul>
Intervention and evaluation study	Jung, Hwang et al. 2017 <sup>44</sup>	Risk of motor vehicle collisions associated with medical conditions and medications: rationale and study protocol.	Clinical data (Traffic accident data)	<p>A retrospective cohort will be constructed for individuals who died in</p> <ul style="list-style-type: none"> <li>- Korean Traffic Accident Analysis System database: MVCs between 2005 and 2014</li> <li>- Korean National Health Insurance database : diseases and medications between 2002 and 2014</li> </ul>
Health service utilization	Oh, Yoon et al. 2012 <sup>38</sup>	Health and economic burden of major cancers due to smoking in Korea	<p>Statistics Korea</p> <p>Korea Health Panel (Cost data)</p>	<ul style="list-style-type: none"> <li>- National Health Insurance Corporation,: Cancer-related direct medical cost</li> <li>- Statistics Korea: cause of death</li> <li>- Korea Health Panel: direct non-medical cost (caregivers 'cost, transportation cost etc)</li> </ul>

Health service utilization	Kang, Sung et al. 2014 <sup>26</sup>	The epidemiology of psychiatric disorders among women with breast cancer in South Korea: analysis of national registry data.	National Cancer Center	<ul style="list-style-type: none"> <li>- NHIS database: diagnosed with breast cancer</li> <li>-National Cancer Center: epidemiology of psychiatric disorders</li> </ul>
Health service utilization	Kang, Kim et al. 2015 <sup>27</sup>	Incidence and Treatment Pattern of Extremity Soft Tissue Sarcoma in Korea, 2009-2011: A Nationwide Study Based on the Health Insurance Review and Assessment Service Database.	National Cancer Center Statistics Korea	<ul style="list-style-type: none"> <li>-Korea National Cancer Incidence Data Base : the nationwide incidence and treatment patterns of extremity STS</li> <li>- Korea National Cancer Incidence (KNCI) database</li> <li>- Health Insurance Review and Assessment Service (HIRA) database.</li> </ul>
Health service utilization	Lee, Lee et al. 2016 <sup>47</sup>	Trends in Percutaneous Coronary Intervention and Coronary Artery Bypass Surgery in Korea.	Additional Survey data (OECD Health Data)	<ul style="list-style-type: none"> <li>-OECD Health Data: country-specific ratios of procedure volumes of PCI per 100,000 population in relation to CABG between 2004 and 2013</li> <li>-National Health Insurance Service (NHIS): surgery statistics, procedure volumes, the number of Korean hospitals providing medical services for PCI and for CABG, the hospital-specific procedure volume, and the regional distribution among hospitals in Korea</li> </ul>
Health service utilization	Lee, Moon et al. 2016 <sup>41</sup>	Disability-Adjusted Life Years for Communicable Disease in the Korean Burden of Disease Study 2012.	Statistics Korea Clinical data(Dismod-II program)	<ul style="list-style-type: none"> <li>- Statistic Korea : cause-of-death statistical data</li> <li>- National Health Insurance Service (NHIS): calculate the incidence rate</li> <li>- Dismod-II program: duration and age at onset of disease</li> </ul>

Health service utilization	Lim, Lee et al. 2016 <sup>28</sup>	Incidence, Survival and Prevalence Statistics of Classical Myeloproliferative Neoplasm in Korea.	National Cancer Center  Statistics Korea	-Korea National Cancer Incidence Data Base (KNCIDB): Incidence data of Myeloproliferative neoplasm (MPN)  - Statistic Korea : mortality database  - HIRA : Number of cases of each diagnosis, the prevalence of each disease, prescription data
Health service utilization	Moon, Han et al. 2016 <sup>42</sup>	Hepatitis A in Korea from 2011 to 2013: Current Epidemiologic Status and Regional Distribution.	Surveillance (Centers for Disease Control and Prevention)  Statistics Korea	-Korea Centers for Disease Control and Prevention (KCDC): National Infectious Diseases Surveillance  - HIRA : reimbursement data with hepatitis A virus  - Statistic Korea : national population data
Health service utilization	Rhee, Hong et al. 2016 <sup>43</sup>	Hypoglycemia and Medical Expenses in Patients with Type 2 Diabetes Mellitus: An Analysis Based on the Korea National Diabetes Program Cohort.	Clinical data (Korea National Diabetes Program)	- Korea National Diabetes Program (KNDP): incidence, clinical characteristics, and medical expenses of hypoglycemia  - HIRA: KNDP data were merged with claims data from the Health Insurance Review and Assessment Service (HIRA) of Korea.
Health service utilization	Kim, Joung et al. 2017 <sup>29</sup>	Prevalence and survival prognosis of prostate cancer in patients with end-stage renal disease: a retrospective study based on the Korea national database(2003-2017)	National Cancer Center	- Nationwide Korean Health Insurance System: reimbursement data with PC(prostate cancer) and end-stage renal disease (ESRD)  - Korean Central Cancer Registry data: patients with PC(prostate cancer) and end-stage renal disease (ESRD)



Specific validity and plausibility analysis	Kang, Yoo et al. 2016 <sup>48</sup>	Tuberculosis Notification Completeness and Timeliness in the Republic of Korea During 2012-2014.	Surveillance (Centers for Disease Control and Prevention)	<p>-NHIS: reimbursement data of Tuberculosis(TB) ca (2012-2014)</p> <p>- Korean National Tuberculosis Surveillance System (KNTSS): surveillance data (2011-2015) cases were matched using Resident Registration Numbers.</p>
Specific validity and plausibility analysis	Kim, Lee et al. 2016 <sup>40</sup>	Association of prediabetes, defined by fasting glucose, HbA1c only, or combined criteria, with the risk of cardiovascular disease in Koreans.	<p>Clinical data (Asan Medical center)</p> <p>Statistics Korea</p>	<p>- Health Screening &amp; Promotion Center (Asan Medical Center): general health examination</p> <p>- Nationwide Health Insurance Claims Database : Cardiovascular(CVD) events</p> <p>- Statistics Korea: death due to CVD</p>

## Korean Abstract

### 빅 데이터 이용 개선 방안연구

#### -건강보험 청구데이터를 중심으로-

**서론:** 1989년 전국민 의료보험제도와 정보 통신 기술 (ICT)로 인해 국민건강보험공단과 건강보험심사평가원에서는 건강보험 데이터가 많이 축적되었다. 그러나, 건강보험 빅데이터를 보유하고 있는 두 기관인 건강보험공단과 건강보험심사평가원은 각각 제한된 목적으로만 데이터를 사용해왔고, 다른 기관과 연결되어 사용하기에 한계가 있었다. 왜냐하면, 건강보험 청구 데이터는 국민의 개인정보와 민감 정보를 포함하고 있는 행정데이터이기 때문에 사용에 제약이 따르기 때문이다. 보건의료 빅데이터에 대한 관심에 비해 건강보험 청구 빅데이터의 활용에 대한 실증적 연구는 거의 수행되지 않았다.

**연구목적:** 본 연구의 목적은 한국의 건강보험 청구 빅데이터를 대상으로 특징과 활용 현황을 살펴보고, 활용을 극대화를 위한 전략을 모색하는 것이다.

**연구방법:** 본 연구는 크게 3가지 방법론을 통해 수행되었다.

첫째, 체계적인 문헌검토 (SR)를 이용하여 지난 10 년 동안 보건학 분야에서 한국 건강보험 청구 빅데이터를 대상으로 한 연구를 검토하였다. 2007 년부터 2017 년까지 PubMed 및 Cochrane 데이터베이스를 대상으로 검색하였다. 초기 검

색결과에서 중복 및 제거 기준을 수행 한 후, 총 478 건의 연구가 포함되었다.

둘째, 의도적 표집방법을 통해 선출된 42 명의 전문가(학계 전문가, 건강보험 기관 전문가, EMR 전문가 등)를 대상으로 델파이 설문을 수행하였다. 설문 기간은 각각 1차는 3주(2014년 2 월 24 일 ~ 3 월 14 일)동안, 2차는 참여자들은 1차 항목에 대해 3주 동안 (2014년 3 월 24 ~ 4 월 11일) 진행되었다. 본 연구 설문지는 연세대학교의 IRB 기관 검토위원회에서 승인을 받았다(IRB: 2-1040939-AB-N-01-2014-228). 전문가 설문지내용은 건강보험 빅데이터 활용사례, 활용에 있어 장애요인, 장애를 해소하기 위한 정책 우선순위다. 정책 우선순위 도출기준은 전문가 동의여부와 9점 리커트 척도를 사용하였다.

셋째, 국가별 비교 방법론을 통해 주요국(미국, 영국, 호주, 대만)의 빅데이터 활용 국가전략 및 현황을 검토하였다. 이 결과를 바탕으로 시사점을 찾아, 한국의 건강보험 빅데이터 활용을 위한 전략을 제안하였다.

**연구결과:** 본 연구의 3 가지 방법론에 대한 결과는 아래와 같다.

첫째, 체계적인 문헌검토 대상인 478 건의 연구는 2007 년과 2011 년 사이에 55 건의 연구 (11.5 %)와 지난 5 년간 (2013 ~ 2017 년) 총 423 건 (88.5 %)이었다. 주로 J Korean Med Sci (9.83 %)저널에 등재되었고, 청구자료 제출기관은 건강보험심사평가원(HIRA)이 건강보험공단(NHIS)보다 많았다 (HIRA: 51.9 %, NHIS: 47.5 %). 연구대상은 대부분 전체인구를 대상으로 하였고(65.7 %), 연구기간은 4 년 이상인 경우가 과반수를 넘었다(50.8%). 연구분류는 건강 서비스 활용이었고

(41.4 %), 외부 데이터와 연결하여 사용한 경우는 478건 중 29 건 (6.0 %)이었다. 외부 연계자료는 통계청 사망 자료(41.4 %), 임상자료 (31.0%), 암등록 자료 (24.1%), 비용자료 (20.7%), 국민건강영양조사와 같은 국가통계(17.2 %), 기타 설문조사자료 (10.3 %), 기타 감시자료 (6.9%)순이다.

둘째, 텔파이 연구 결과로 4가지 장애요소(데이터 사용을 위한 법적 미성숙, 정보제공을 위한 공감대 부족, 정보 공유를 위한 기술적 제약, 데이터 활용을 지원할 정부지원 부족 순)와 13 가지 정책(정보제공을 위한 국민의 동의를 위한 정책, 정보활용을 위한 법 개정 정책, 정보활용을 위한 제도개선 정책, 정보제공을 위한 기관 동의정책, 정보공유를 위한 익명화 등 기술정책, 정보 활용을 위한 국가거버넌스 설립 정책 순 등)이 도출되었다.

셋째, 텔파이 연구결과 도출된 이슈별로 국가 비교제도결과는 다음과 같다.

미국은 경제위기를 극복하기 위한 경기부양법(ARRA)에 따라, 경기부양정책의 일환으로 의료정보기술에 관한 법률인 HITECH Act(Health Information Technology for Economic and Clinical Health Act)를 제정, 국가적 차원의 빅데이터 활용전략인 Health Data Initiative(HID)을 시행한다. 민간 보험사들을 중심으로 다양한 정보 통합 플랫폼이 활성화되도록 간접적인 조정 역할로서 ONC(Office of the National Coordinator)기관을 만들어 EHR의 의미 있는 활용(meaningful use)을 위한 상호운용성(interoperability) 지원하고 있다.

영국의 보건부는 2013년에 ‘Personalized Health and Care 2020’ 을 발표하고, 영역별로 분리되어 있는 국민의 보건의료 데이터를 실시간으로 사용 가능하게 하여 국민 건강수준 향상시키는 계획이다. 독립 조직으로 NHS Digital(예전 HSCIC)를 설립하여 환자와 시민의 의료와 복지 정보에 대한 통제권을 강화시키고, 분산된 사회보장 데이터들을 수집, 저장, 연계, 분석을 지원한다.

호주 정보관리청에서는 빅데이터 전략을 수립하고, 공공서비스 수준향상을 위한 정책수단으로 빅데이터를 활용한다. 빅데이터를 운영/활용하는 전(全)과정에서 개인별 master linkage key를 사용하기 때문에 이용자의 프라이버시와 데이터를 보호하는 특징이 있다. 또한, 서호주주립대학, 커틴대학 및 테프론연구소가 협력하여 개인정보에 대한 비밀유지를 전제로 하는 연계 체계인 “Western Australian Data Linkage System(WADLS)”가 연구자료로 활용되고 있다.

대만은 위생복지부 산하 중앙건강보험서(NHIA), 국가위생연구원(NHRI)을 중심으로 일관된 보건의료 데이터를 관리 체계 및 관련 정책을 확립하고 수요자 중심의 데이터 제공을 위해 노력한다. 특히, 위생복지부가 건강정보협력센터(CCHIA)를 설립하여 국가위생연구원(NHRI)의 자료와 타공공기관 데이터를 연계하여 국민의 복지 증진을 위해 노력한다. 대만은 국가차원의 거버넌스(CCHIA)와 수요자 중심의 데이터 접근성 확대에 의해 아시아 정보화 성공 국가로 자리매김하고 있다.

한국은 국가 차원의 전략부재로 인해 보건의료 산업 등에 제한적으로 빅데이터를 활용하고 있다. 이는 기관별 낮은 데이터 공개 수준, 과도한 개인정보보호법,

기관간 연계 및 통합체계 미흡, 데이터 거버넌스 부재 등의 한계점에 기인한다. 결론적으로 세가지 방법론을 통해 이슈별로 크게 3가지 정책전략을 제안하겠다. 국가차원 빅데이터 거버넌스 및 전략수립, 빅데이터 공개 활성화를 위한 법적 제도적 개선(개인정보동의방법 차등화, 법적 제도적 지침개발 전략), 빅데이터 연계를 위한 보건의료 연구 활성화(데이터 연계를 위한 개인정보 보호기술의 개발, 사용자 중심의 건강보험 청구자료 이용)전략이다.

**결론:** 최근 보건복지부에 2018년부터 향후 2년 동안 보건의료 빅데이터 시범사업을 실시할 예정이라고 밝혔다. 보건의료 빅데이터 시범사업은 국민건강보험공단(이하 공단)과 간경보험심사평가원(이하 심평원) 등 공공기관이 보유한 건강보험 빅데이터를 연계해 연구 등에 활용하는 것으로 2020년까지 개인 맞춤형 정밀진단 및 치료기반을 마련하는 것이다.

한국은 정보화 선진국으로, 단일 보험체계와 전국민을 식별할 수 있는 주민등록번호를 사용하고 있어 건강보험 청구 빅데이터를 활용하기에 좋은 환경을 갖추고 있다. 건강보험 청구 빅데이터를 다른 데이터와 연계가 수월한 여건이 마련된다면, 더 큰 가치를 창출해낼 수 있다. 이를 시행할 정부의 법률 및 제도적 정책이 필요하다.

---

**핵심어:** 건강보험 청구데이터, 건강보험심사평가원, 국민건강 보험공단, 텔레이 기법, 체계적 문헌고찰(SR)