pISSN: 2093-940X, eISSN: 2233-4718
Journal of Rheumatic Diseases Vol. 28, No. 3, July, 2021
https://doi.org/10.4078/jrd.2021.28.3.119

Review Article

Clinical Study Using Healthcare Claims Database

Jin-Su Park, M.D., Chan Hee Lee, M.D., Ph.D.
Division of Rheumatology, Department of Internal Medicine, National Health Insurance Service Ilsan Hospital, Goyang, Korea

The healthcare claims database is a database created using claims data accumulated while operating the government's health insurance system. The National Health Insurance Service (NHIS) provides benefits for health promotion, prevention, diagnosis, and disease and injury treatment, as well as for rehabilitation, birth, and death. Ninety-seven percent of the total population is enrolled in the NHIS; individuals pay a monthly insurance contribution to the system, and the NHIS pays a portion of the cost of reimbursement items to the medical institution when the subscriber receives medical services. In this process, the NHIS and Health Insurance Review Agency (HIRA) decide on payment, and claims data are documented items that medical institutions claim to these government agencies. The NHIS and HIRA have established a database to support policy and academic research, and they provide this database to researchers. Health claims data are representative of the nation, reflecting the actual medical environment. They also shorten the time and cost required for research and have several advantages as research data. However, studies should be conducted with an understanding of the limitations of claims data, a sufficient understanding of the characteristics of the Korean insurance system, and criteria for providing reimbursed services. Moreover, validating the healthcare claims database will facilitate more useful and reliable research. **(J Rheum Dis 2021;28:119-125)**

Key Words. National health insurance, Rheumatology, Republic of Korea

INTRODUCTION

Large population-based studies provide us with various information such as prevalence, incidence, disease risk factors, treatments used in clinical practice, and prognosis. As the number of participants increases, the selection and participation bias decreases; therefore, large-scale studies are preferred by researchers. Consequently, studies in which it is relatively easy to establish a cohort, such as several specific occupational groups or regions, have been actively conducted [1]. However, it is difficult to represent the entire population when specific occupations or regions are targeted. Population-based data have the advantage of reducing selection bias to a greater extent compared to data that are solely large-scale. However, such population-based data are difficult to obtain. In the 2000s, interest in big data had increased rapidly, and several advantages of using existing data for other research purposes became noticeable. Concurrently, interest in research using the healthcare claims database has increased.

The healthcare claims database comprises secondary data based on claims data accumulated while operating the government's health insurance system. Countries that have built such a database include South Korea, Japan, Taiwan, and Scandinavian countries (e.g., Sweden and Denmark) [2]. Among them, Taiwan established the National Health Insurance in 1995, and the National Health Research Institute began to build the National Health Insurance Research Database in 1997. Many papers have been published since the late 2000s [3]. In Korea, as in Taiwan, research based on databases provided by the National Health Insurance Service (NHIS) and Health Insurance Review Agency (HIRA) is rapidly increasing. Consequently, we reviewed the characteristics of the claims database, current status of research into rheumatology, and precautions taken during research.

Received：May 31, 2021, **Revised**：June 21, 2021, **Accepted**：June 21, 2021

Corresponding to：Chan Hee Lee http://orcid.org/0000-0001-5942-9029
Division of Rheumatology, Department of Internal Medicine, National Health Insurance Service Ilsan Hospital, 100 Ilsan-ro, Ilsandong-gu, Goyang 10444, Korea. E-mail：chanhee@nhimc.or.kr

MAIN SUBJECTS

Characteristics and structure of Korea's health insurance system

The health insurance system is a system in which citizens typically pay insurance premiums to the NHIS, a single insurer, and the NHIS manages and operates the insurance money and provides insurance benefits when necessary. The purpose is to prevent excessive burden on households owing to high medical expenses caused by an illness or injury. The NHIS provides benefits for health promotion, prevention, diagnosis, and disease and injury treatment, as well as for rehabilitation, birth, and death. For example, it provides free health checkups every two years to all citizens over the age of 40 years for health promotion and prevention, as well as guarantees the cost of diagnosis and treatment. However, medical services that do not meet the aforementioned purpose are not covered by insurance. Oral nutritional supplements for prophylactic purposes or surgery, procedures, and some new treatments for cosmetic purposes with no associated diseases are usually included in these non-reimbursement items.

Efforts to establish a health insurance system first began in 1963, and in 1977, an occupational medical insurance system for industrial workers in workplaces with 500 or more workers was created [4]. Gradually, this system was expanded to public officials, faculty, and medical insurance in rural areas, and universal coverage was achieved in 1989. In 2000, the organization was established as a single insurer by combining various local and social insurance. While it was originally called the National Health Insurance Corporation, its name was changed to the NHIS in 2013 [5].

Of the country's total population, 97% belong to the NHIS, and about 3% are registered in the Medical Aid Program (MAP), a program for low-income families. When a person from an MAP uses medical services, little or no medical expenses are incurred, except non-reimbursement items and a certain amount of copayment. As of April 2020, 1.5 million people qualified for medical benefits based on a population of 52.38 million, accounting for 2.8% of the total population. Of the 51.34 million people enrolled in health insurance, 37.24 million (70.5%) are employees, and 14.1 million (26.7%) are self-employed local subscribers [6].

The NHIS has a unique system referred to as the Individual Copayment Beneficiaries Program (ICBP). Since the early 2000s, ICBP has reduced the burden of medical expenses for patients with cancer or rare and intractable diseases. The attending physician checks whether the patient meets the diagnostic criteria for diseases in-

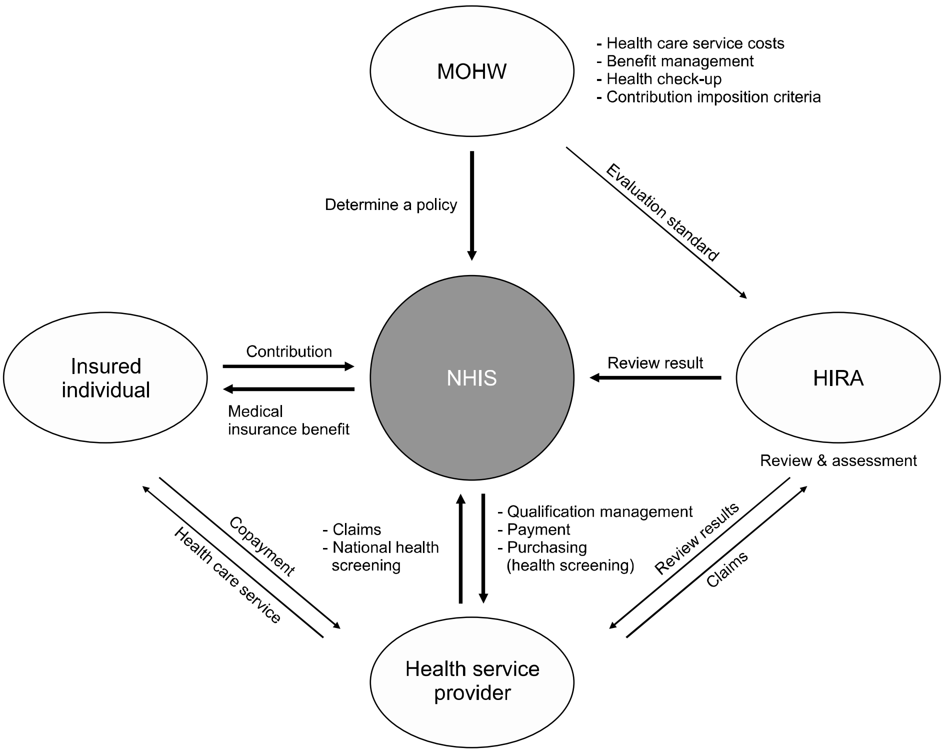

Figure 1. The operational structure of the National Health Insurance system. MOHW: Ministry of Health and Welfare, NHIS: National Health Insurance Service, HIRA: Health Insurance Review and Assessment.

cluded in the ICBP through physical, laboratory, imaging, or pathological examinations. If applicable, registration is made through the NHIS, and patients registered in this system pay 5%~10% of reimbursement items as copayment to medical institutions, while other expenses are paid for by the NHIS.

Korean citizens usually pay an insurance contribution to the NHIS, equivalent to 6.86% of the monthly average wage for employees in 2021 [7]. When citizens receive medical services corresponding to reimbursement items, the NHIS will pay the total medical expenses minus the copayment to the medical institution. In addition to the NHIS, there is one more organization related to the health insurance system. The HIRA reviews medical billing and claims submitted by medical institutions to assess the adequacy of quality and quantity [8]. Based on their review and decisions, the NHIS will pay the medical institution. The HIRA and NHIS are under the Ministry of Health and Welfare (MOHW) and are influenced by the MOHW in the formulation and implementation of policies (Figure 1).

NHIS and HIRA databases

"Healthcare claims data" refer to data based on the statement of medical care benefits billed by a medical institution to receive medical expenses from the NHIS. This specification contains information on medical institutions; patients' personal information; International Classification of Diseases, 10th revision (ICD-10); medical history (tests, procedures, and surgery); prescriptions; and costs. Both the NHIS and HIRA have data based on this bill. They include not only claims data but also the results of health checkups for citizens over 40 years old and infant checkups conducted by the NHIS. To support policy and academic research, these data are protected using a personal identification code. In the case of the NHIS, a database was created in 2006 and provided to researchers in 2010. The HIRA began to provide this to researchers in 2013 [9,10].

The NHIS operates the National Health Insurance Sharing Service (NHISS) to provide support for policy and academic research using public health information. The NHISS provides a sample cohort database consisting of 2% (about 1 million people) of all citizens, including identified claims, health screening data, and mortality data [11]. Because these sample data are relatively accessible to researchers, many papers using such data have been published. In addition, there are four more cohort databases: the national health screening cohort, the senior cohort, the working women cohort, and the infant medical checkup cohort [11-13]. Currently, it is possible to conduct research targeting the whole country through

Table 1. Compositions of variables of NHIS data

Qualification	Statement (T20)	Treatment details (T30)	Disease type (T40)	Details of prescription (T60)
Sex	Start date of medical care	Start date of medical care	Start date of medical care	Start date of medical care
Age (year of birth)	Medical subject code	Medication	Medical subject code	Medication
Location	Principle diagnosis	Dosage and frequency	Principle diagnosis	Dosage
Type of subscription	1st to 4th additional diagnosis	Cost of medication	Additional diagnosis	Days of administration
Income rank	Hospitalization route	In-hospital administration of medicine	Rule out	Cost of medication
Disability	Official injury	Medical expense code (procedure included)		
Death (date of death)	Perform surgery			
	Days to visit			
	Hospitalization days			
	Prescription days			
	Treatment results			
	Medical institution			
	Medical expenses			
	Special symbol (DRG, ICBP code, etc.)			

NHIS: National Health Insurance System, DRG: diagnosis-related group, ICBP: Individual Copayment Beneficiaries Program.

a customized database, and family history research is also becoming possible through the establishment of a family tree database [14]. The HIRA provides four samples: national inpatient sample (HIRA-NIS), the national patient sample (HIRA-NPS), the aged population sample (HIRA-APS), and the pediatric patient sample (HIRA-PPS) [10].

To explain the data structure of the NHIS as an example of a customized database, it is essentially composed of qualification, statement, details of treatment, type of disease, and details of prescription. Statement, details of treatment, type of disease, and details of prescription are referred to as T20, T30, T40, and T60, respectively. The contents of each table are indicated in Table 1.

Publications based on research using the health claims database

The NHIS and HIRA actively provide data through the establishment of organizations. As they provide several sample datasets that are easy to access, the number of research papers using health claims data has increased quite rapidly. In PubMed, we searched for papers with Korea as the affiliation and “national health insurance” or “health insurance review” in the title or abstract (Figure 2A). We confirmed that, toward the second half of the 2010s, papers using NHIS data were frequently published, and their number of studies was increasing rapidly. When we added representative rheumatic diseases to the above results as a search word (search keywords: rheumatoid arthritis, lupus, ankylosing spondylitis, Behçet, osteoarthritis, gout, Sjögren, myositis, vasculitis, fibromyalgia, systemic sclerosis, antiphospholipid, and adult-onset Still disease), we found additional studies using NHIS data. Among studies using NHIS data, rheumatoid arthritis was the most studied topic with 54 studies. Lupus was the second most studied topic (30 studies), followed by ankylosing spondylitis (23 studies), Behçet’s disease (17 studies), and osteoarthritis (15 studies). There were fewer than 10 or no studies targeting other diseases. Among studies using HIRA data, rheumatoid arthritis was also the most studied disease (14 studies), followed by ankylosing spondylitis (10 studies). Studies of other topics were limited or had not been conducted (Figure 2B).

When reviewing the conducted studies, we confirmed that the following studies were conducted according to the disease course:

1) Risk factor and behavior

If we add several types of data (particularly national health checkup data) to the existing data, we can check the results of people's behavior/habits and laboratory results and use them for research [15]. When the temporal relationship of these results can be interpreted, the risk factors for disease occurrence can be estimated.

2) Epidemiology

The prevalence and incidence survey is considered to be close to the actual data; thus, it is the most frequently conducted study and is useful for rare disease surveys [16,17]. In addition, it is possible to investigate comorbid

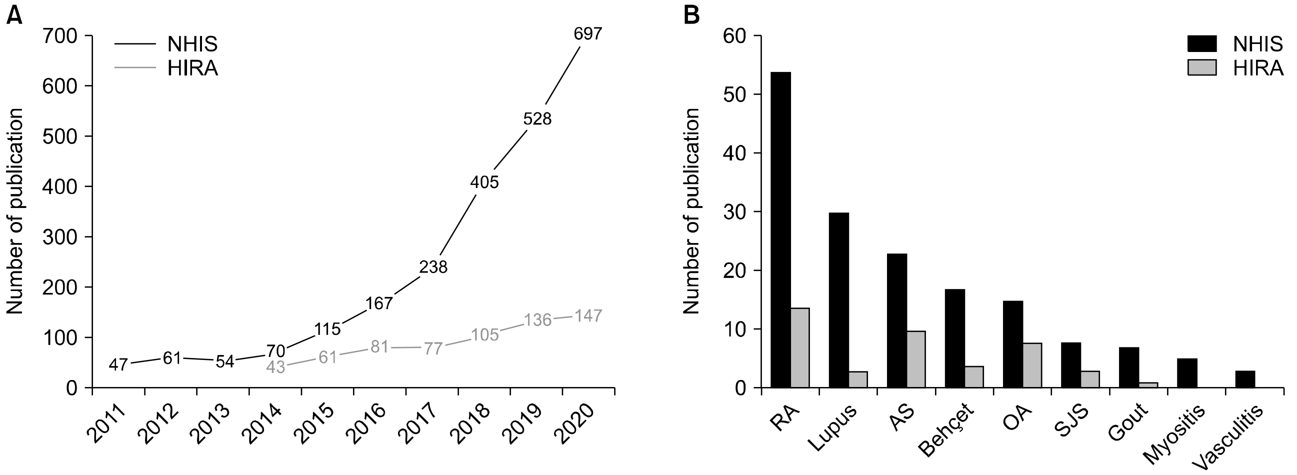

Figure 2. Publications of research using the health claims database. (A) Year-specific number of publications using the NHIS (2011～2020) and Health Insurance Review and Assessment (HIRA) databases (2014～2020). (B) Rheumatologic disease-specific number of publications using the NHIS (2011～2020) and HIRA databases (2014～2020). NHIS: National Health Insurance Service, RA: rheumatoid arthritis, AS: ankylosing spondylitis, OA: osteoarthritis, SJS: Sjögren’s syndrome.

diseases, and care should be taken when interpreting causal relationships [18].

3) Treatment

It is possible to investigate treatment modalities and treatment drugs that are being implemented [19]. In addition, treatment effects and adverse reactions can be indirectly confirmed, and the cost of treatment can be estimated [20,21].

4) Prognosis

When data for each claim are accumulated, a long-term cohort is created. Thus, it is possible to study the long-term effects, prognosis, and complications of the treatment [22]. In addition, there is code and death information concerning disability; thus, it is possible to study this as well [23].

Points to consider when writing theses using claims data

Claims data provide information on a large number of patients and incidents targeting the entire nation. An extremely large sample size has the advantage of allowing for the performance of several studies that cannot be performed in conventional small clinical studies. Conversely, because the data were not constructed for the purpose of research from the beginning, the data collected by this method contain various errors in interpretation. Limitations have been mentioned often by existing authors, but it is worth reorganizing and mentioning other restrictions.

1) Claims data

This data format of items is provided by medical institutions to receive expenses from the NHIS. Therefore, clinical information may be omitted or inaccurate if it is not important to the claim [24]. In addition, the results cannot be confirmed because the data were not tailored to the study. Physical examination results including blood pressure, laboratory tests, imaging tests, and pathology results could not be confirmed. Adding the results of the national health checkup to the existing data shows some of the results, but only the results related to cardiovascular diseases are included. Regarding the national health checkup, since 2009, the checkup items have been changed, and the display of the results of the questionnaire on lifestyle has also been changed [15].

2) Operational definition

Because the results cannot be viewed, it is difficult to apply the diagnostic/classification criteria used to select participants in general studies. Therefore, researchers must define the diagnosis, namely, the operational definition. Diagnosis can be defined through disease codes, which is the easiest method; however, these codes cannot represent 100% of patients' diseases. The claim disease code accounts for approximately 70%, which is consistent with the diagnosis of the medical record. The degree of concordance between them decreases in outpatient patients, compared to inpatients, mild diseases, compared to severe diseases, and primary care, compared to general hospitals [3,25]. In addition, the number of participants may vary depending on the scope of the disease code investigation, as more mild diseases are included in additional diagnosis rather than principal diagnosis [26-28]. As described above, the low degree of agreement can be increased by adding processes such as searching for codes of drugs specifically used for diseases and the number of visits to the outpatient department [20]. However, when the aforementioned process is added, it is important for the researcher to develop an appropriate operational definition according to the research topic, as it may be inappropriate for prevalence and incidence research because the sensitivity to disease is lowered. Therefore, it is recommended to perform a self-validation or refer to the algorithms of previously published papers [29].

3) Criteria for providing reimbursed services in the NHIS

When medical institutions claim and receive expenses from the NHIS, they will receive expenses after the HIRA reviews whether the claim is appropriate to the criteria for providing reimbursed services in the NHIS. When the HIRA screens a claim, the provision of benefits is often determined by the presence or absence of a specific disease code; thus, there may be cases in which a disease code that is not directly related to the underlying major disease is added. As the formulary approach is changed, the use of drugs may increase or decrease [30]. Moreover, medications and procedures may vary depending on the diagnosis-related group (DRG) policy and the fee for service policy [31]. Therefore, the interpretation of research conducted using claims data should be based on a long-term and detailed understanding of insurance items.

4) Access to healthcare services

If the copayment is reduced owing to enrollment in the ICBP, MAP, etc., registered persons' accessibility to medical services increases. As their use of medical services increases, their prevalence/incidence rates and drug use for diseases may differ from those of other income and disease groups [26]. Therefore, care should be taken when interpreting the results of these groups.

5) Causal relationship

The claims data contain several samples; thus, the statistical power is high. However, these data show a phenomenon, and it is difficult to determine a causal relationship [28].

Health claims data have the above limitations, but, as real-world data, they also have many advantages that other data cannot have. Health claims data are representative of the nation and are easy to generalize. Since health claims data are population-based data, selection bias can be reduced to a greater degree compared to other large-scale data. They reflect the actual healthcare environment rather than a limited experimental environment and show the current status and trends because they are long-term follow-up data. If health claims data are used in the thesis writing process, data that have already been established are used, thus reducing the time and cost required for data construction, which takes up the most time during such a process. In addition, it is easy to obtain detailed information on medication use, access to actual treatment costs, and research on rare incidents [32].

To overcome the above limitations, other data (Statistics Administration, National Institute of Environmental Sciences, Meteorological Administration, etc.) can be linked, or data validation for "operational definitions" can be added. Adding these multiple processes will lead to more accurate results. In addition, more studies should be conducted as claims data are continuously building an additional database based on the NHIS.

CONCLUSION

Health claims data in Korea contain ample medical treatment data, and various studies can be conducted. Considering data limitations and the need for further validation, these data are critical for medical research.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

AUTHOR CONTRIBUTIONS

J.S.P. was involved in the conception and design of the study. J.S.P and C.H.L. were involved in critically drafting and revising the manuscript for important intellectual content and final approval of the version to be published.

REFERENCES

1. Ng B, Chu A. Factors associated with methotrexate dosing and therapeutic decisions in veterans with rheumatoid arthritis. Clin Rheumatol 2014;33:21-30.
2. Hsing AW, Ioannidis JP. Nationwide population science: lessons from the Taiwan National Health Insurance Research Database. JAMA Intern Med 2015;175:1527-9.
3. Lin LY, Warren-Gash C, Smeeth L, Chen PC. Data resource profile: the National Health Insurance Research Database (NHIRD). Epidemiol Health 2018;40:e2018062.
4. National Health Insurance Service Korea. History of the NHIS [Internet]. Wonju: National Health Insurance Service Korea [cited 2021 May]. Available from: https://www.nhis.or.kr/english/wbheaa01300m01.do.
5. Song SO, Jung CH, Song YD, Park CY, Kwon HS, Cha BS, et al. Background and data configuration process of a nationwide population-based study using the Korean national health insurance system. Diabetes Metab J 2014;38:395-403.
6. National Health Insurance Service Korea. Population coverage [Internet]. Wonju: National Health Insurance Service Korea [cited 2021 May]. Available from: https://www.nhis.or.kr/nhis/policy/wbhada01700m01.do.
7. National Health Insurance Service Korea. Contributions [Internet]. Wonju: National Health Insurance Service Korea [cited 2021 May]. Available from: https://www.nhis.or.kr/english/wbheaa02500m01.do.
8. Seong SC, Kim YY, Khang YH, Park JH, Kang HJ, Lee H, et al. Data resource profile: the National Health Information Database of the National Health Insurance Service in South Korea. Int J Epidemiol 2017;46:799-800.
9. Lee CH, Sung NY. The prevalence and features of Korean gout patients using the National Health Insurance Corporation database. J Rheum Dis 2011;18:94-100.
10. Kim JA, Yoon S, Kim LY, Kim DS. 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. J Korean Med Sci 2017;32:718-28.
11. Lee J, Lee JS, Park SH, Shin SA, Kim K. Cohort profile: the National Health Insurance Service-National Sample Cohort (NHIS-NSC), South Korea. Int J Epidemiol 2017;46:e15.
12. Kim YI, Kim YY, Yoon JL, Won CW, Ha S, Cho KD, et al. Cohort Profile: National health insurance service-senior

(NHIS-senior) cohort in Korea. BMJ Open 2019;9:e024344.

13. Seong SC, Kim YY, Park SK, Khang YH, Kim HC, Park JH, et al. Cohort profile: the National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS) in Korea. BMJ Open 2017;7:e016640.
14. Kim YY, Hong HY, Cho KD, Park JH. Family tree database of the National Health Information Database in Korea. Epidemiol Health 2019;41:e2019040.
15. Lee CH, Sung NY, Lee J, Bae SC. Factors associated with gout in South Koreans: analysis using the National Health Insurance Corporation and the National Health Screening Exam databases. Clin Rheumatol 2013;32:829-37.
16. Kwak SG, Park SH, Kim JY. Incidence and prevalence of juvenile systemic lupus erythematosus in Korea: data from the 2017 National Health Claims Database. J Rheumatol 2021;48:258-61
17. Chung MK, Park JS, Lim H, Lee CH, Lee J. Incidence and prevalence of systemic lupus erythematosus among Korean women in childbearing years: a nationwide population-based study. Lupus 2021;30:674-9.
18. Hong M, Moon DS, Chang H, Lee SY, Cho SW, Lee KS, et al. Incidence and comorbidity of reactive attachment disorder: based on National Health Insurance claims data, 2010-2012 in Korea. Psychiatry Investig 2018;15:118-23.
19. Won S, Cho SK, Kim D, Han M, Lee J, Jang EJ, et al. Update on the prevalence and incidence of rheumatoid arthritis in Korea and an analysis of medical care and drug utilization. Rheumatol Int 2018;38:649-56.
20. Kang EH, Choi HK, Shin A, Lee YJ, Lee EB, Song YW, et al. Comparative cardiovascular risk of allopurinol versus febuxostat in patients with gout: a nation-wide cohort study. Rheumatology (Oxford) 2019;58:2122-9.
21. Nam JH, Lee C, Kim N, Park KY, Ha J, Yun J, et al. Impact of continuous care on health outcomes and cost for type 2 diabetes mellitus: analysis using National Health Insurance cohort database. Diabetes Metab J 2019;43:776-84.
22. Choi J, Kim HJ, Lee J, Cho S, Ko MJ, Lim YS. Risk of hepatocellular carcinoma in patients treated with entecavir vs tenofovir for chronic hepatitis B: a Korean nationwide cohort study. JAMA Oncol 2019;5:30-6.
23. Cho WK, Lee NY, Han K, Suh BK, Park YG. The population prevalence, associations of congenital heart defect and mortality risk for Down's syndrome in South Korea based on National Health Insurance Service (NHIS) data. Clin Epidemiol 2020;12:519-25.
24. Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. J Epidemiol Community Health 2014;68:283-7.
25. Kimm H, Yun JE, Lee SH, Jang Y, Jee SH. Validity of the diagnosis of acute myocardial infarction in Korean national medical health insurance claims data: the Korean heart study (1). Korean Circ J 2012;42:10-5.
26. Park JS, Kang M, Song JS, Lim HS, Lee CH. Trends of gout prevalence in South Korea based on medical utilization: a National Health Insurance Service Database (2002～2015). J Rheum Dis 2020;27:174-81.
27. Kim JW, Kwak SG, Lee H, Kim SK, Choe JY, Park SH. Prevalence and incidence of gout in Korea: data from the national health claims database 2007-2015. Rheumatol Int 2017;37:1499-506.
28. Kestle JR. Administrative database research. J Neurosurg 2015;122:441-2
29. Sentinel Initiative. Health outcomes of interest [Internet]. Sentinel Initiative [cited 2021 May]. Available from: https://www.sentinelinitiative.org/methods-data-tools/health-outcomes-interest.
30. Kim JW, Kwak SG, Park SH. Prescription pattern of urate-lowering therapy in Korean gout patients: data from the national health claims database. Korean J Intern Med 2018; 33:228-9.
31. Fetter RB, Freeman JL. Diagnosis related groups: product line management within hospitals. Acad Manage Rev 1986;11:41-54.
32. Seo HY, Yoon SJ, Kim EJ, Oh IH, Lee YH, Kim YA. The economic burden of rheumatic heart disease in South Korea. Rheumatol Int 2013;33:1505-10.