Journal of Health Informatics and Statistics

50주년 특별호
Review Article

J Health Info Stat 2022;47(Suppl 2):S31-S39 • https://doi.org/10.21032/jhis.2022.47.S2.S31

pISSN 2465-8014 | eISSN 2465-8022

# 보건의료연구를 위한 건강보험 자료의 효과적 활용방법

**박일수**

동의대학교 의료경영학과 교수

# How to Use Health Insurance Data Effectively for Healthcare Research

**Ilsu Park**

*Professor, Department of Healthcare Management, Dong-eui University, Busan, Korea*

**Objectives:** The purpose of this study is to present an understanding of the structure and characteristics of health insurance data, which should be basically known in order to use this data more effectively to derive more accurate research results across health and medical care, and an effective use method based on it. **Methods:** The National Health Insurance Sharing Service and Healthcare Big data open system was reviewed to suggest the structure and characteristics of health insurance data and an efficient use method, and Google Scholar and the Korean Studies Information Service System were used to review the overall application method according to the characteristics of healthcare big data and related literature. **Results:** The types and characteristics of each research data for public use provided by National Health Insurance Service (NHIS), Health Insurance Review & Assessment Service (HIRA), and the differences between the data of the two institutions were compared and presented. In addition, for the efficient use of health insurance data in healthcare research, understanding the structure and characteristics of health insurance data based on the health insurance system, clear operational definition of analysis variables, adjust bias between comparison groups in case-control study, and the interpretation of analysis results suggest that both statistical and clinical significance should be considered. **Conclusions:** It is very important to understand the structure and characteristics of health insurance data based health insurance system. And if we consider in more depth ways to efficiently utilize the health insurance data presented in this study, it will be of great help to future researchers and to improve the quality of the research.

**Key words:** Health insurance data, Health insurance system, NHIS, HIRA, Health care research

## 서 론

보건의료 빅데이터에 대한 정의, 종류 그리고 자료원은 관련 전문가 및 연구자들에 의해 다양한 관점에서 제시되고 있다. 이들의 정의 등을 종합해보면, 보건의료와 관련된 영역에서 발생하는 다양한 형태의 자료이며, 그 종류도 각 자료를 지칭하는 표현 등에 있어 일부 차이는 존재하지만 대부분 유사한 것으로 귀결되며, 그 중 빠지지 않는 것이 건강보험 자료이다.

우리나라 건강보험 자료는 건강보험제도권 내에서 「국민건강보험법」 제14조 및 제63조에 따라 국민건강보험공단 및 건강보험심사평가원이 그들의 고유 업무 수행을 위해 수집되는 자료로서, 넓은 범위에서는 우리나라 전체 국민의 소득 및 재산 정보, 요양기관에서의 진료 이용 정보, 건강검진 정보 등을 포함하고 있다.

많은 연구자들이 보건의료와 관련된 연구를 계획하고 설계할 때 가장 우선적으로 고려하는 사항이 연구를 통해 분석할 수 있는 자료의 존재 파악 및 습득 여부이며[1], 또한 활용하는 자료를 통해 도출된 결과에 대한 신뢰성과 일반화에 대한 가능성일 것이다. 이와 같은 관점에서 우리나라의 건강보험 자료가 가지는 가치에 대해 많은 연구자들

**Corresponding author:** Ilsu park
176 Eomgwang-ro, Busanjin-gu, Busan 47340, Korea
Tel: +82-51-890-4215, E-mail: ispark@deu.ac.kr

**Received:** July 25, 2022 **Accepted:** August 16, 2022 **Published:** August 31, 2022

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

How to cite this article:
Park I. How to use health insurance data effectively for healthcare research. J Health Info Stat 2022;47(Suppl 2):S31-S39. Doi: https://doi.org/10.21032/jhis.2022.47.S2.S31

It is identical to the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0) which permit sunrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2022 Journal of Health Informatics and Statistics

이 인정하는 이유는 건강보험 자료가 보건의료의 전문성 및 보건의료 제도권 내에서 발생하는 정확도가 매우 높은 자료이기 때문일 것이다. 또한, 전 세계에서 거의 유일하게 모든 국민의 자료가 특정 일부 기관에서 모두 관리하고 있으므로, 자료활용에 대한 제도적 근거만 마련된다면 자료의 습득, 활용 그리고 타 자료와의 연계가 매우 용이하다는 점일 것이다.

2020년 8월부터 개정·시행된「개인정보 보호법」,「정보통신망 이용촉진 및 정보보호 등에 관한 법률」,「신용정보의 이용 및 보호에 관한 법률」과 같은 데이터 3법으로 인해 자료활용 및 습득의 용이성이 과거에 비해 더욱 완화되었다. 특히, 데이터 3법은 과학적 연구, 공익적 목적 그리고 통계적 활용 등에 있어서는 개인 단위 정보의 활용을 적극적으로 수용한다는 의미를 지니고 있어, 과거 고유의 업무를 수행하는 공공기관 및 일부 연구자들에게만 제한적이었던 것이었던 건강보험 자료가 현재는 이를 활용하고자 하는 모든 사람들이 제도적 근거하에서 활용이 가능해졌다[2].

최근 위와 같은 상황변화로, 많은 연구자들이 건강보험 자료를 활용하여 연구를 계획하거나 진행하고 있다. 국민건강보험공단이 국민건강정보 데이터베이스(DB, 2021년 기준 테이블 584개, 150 Gbyte) [3]에 기반한 건강보험자료공유서비스(http://nhiss.nhis.or.kr)를 통해 지원한 연구는 2014년 이후 2020년 6월까지 총 3,553건이며[4], 데이터 3법 통과 이후 급격하게 증가하여 2021년 상반기만 총 779건이 지원되었다[5]. 또한, 건강보험심사평가원은 보건의료빅데이터개방시스템(http://opendata.hira.or.kr)을 통해 2015년부터 2021년까지 연구과제 지원 등에 총 1,136건이 제공되었으며, 이 또한 2020년 이전 제공건수가 매년 108-170건이었으나, 2021년 202건이 제공되어 2020년 대비 20% 이상 증가하였다[6]. 이와 같은 추세는 앞으로 더욱 가속화될 것으로 판단된다.

그러나, 이러한 추세 속에는 건강보험 자료의 정확한 활용에 대한 우려도 함께 제기되고 있다. 국민건강보험공단 및 건강보험심사평가원이 제공한 건강보험 자료를 활용한 연구들의 방법론을 검토한 Lim et al. [7]의 연구에 따르면, 2014-2020년까지 총 678편의 연구들을 검토한 결과 대부분이 단조로운 분석설계 및 분석수행에 그치고 있음을 지적하였다. 그리고 다양한 편의의 문제를 해결하기 위해 건강보험 자료를 이용하는 연구자들에 대해서는 분석방법론에 대한 심도있는 검토 그리고 자료제공 기관 등에 대해서는 자료이용에 있어 분석방법, 점검도구 개발의 필요성을 제안하였다.

전술한 것처럼, 건강보험 자료의 가치는 많은 연구에서 제시되었듯 별다른 이견이 없을 것으로 판단된다. 또한 정부의 데이터 개방정책 등으로 인해 건강보험 자료를 활용한 연구는 향후 기하급수적으로 증가될 것으로 예상된다. 그럼에도 불구하고 이를 활용하고자 하는 연구자들은 건강보험 자료의 크기에만 매료되어 그 자료가 가지는 속성 등을 충분히 고려한 연구방법론의 적용은 이루어지지 않고 있다.

이에 본 연구에서는 건강보험 자료를 보다 효과적으로 활용하여 보건의료에 전반에 걸쳐 좀 더 정확한 연구결과를 도출하기 위해 국민건강보험공단 및 건강보험심사평가원 등 각 자료보유기관의 웹사이트, 구글 스칼라 그리고 한국학술정보서비스시스템 등에서 수집된 관련 자료를 기반으로 건강보험 자료를 활용한 연구수행 시, 기본적으로 알아야 할 건강보험 자료의 구조와 특성에 대한 이해 및 효과적 활용방법을 제시하고자 한다.

## 국민건강보험 자료 특성 및 이해

건강보험 자료를 이해하기 위해 가장 기본적으로 알아야 할 것은 우리나라 건강보험 제도이다. 우리나라의 건강보험 제도는 2000년 국가에서 관장하는 단일 보험자 조직의 운영 관리체계로 통합되었으며,「국민건강보험법」제42조 따라 일부 의료기관을 제외한 대부분의 의료기관은 요양기관으로 지정되는 건강보험 당연지정제가 적용됨으로써, 의료 공급자는 건강보험에 대한 진료나 보험적용이 의무화되어있고, 진료비 산정을 위한 진료비지불제도는 주로 행위별 수가제로 운영되고 있다. 이는 2000년 이전 다수의 보험자가 가지고 있는 건강보험 자료들이 2000년에 하나의 보험자 즉, 현재의 국민건강보험공단으로 통합됨을 의미한다. 또한, 건강보험 당연지정제에 따라 우리나라 대부분의 의료기관은 건강보험이 적용되는 진료를 행하고, 이로 인해 우리나라 전체 국민의 건강보험이 적용되지 않는 비급여 및 타 보험적용건 등을 제외한 모든 의료이용 내역이 국민건강보험공단으로 집결됨을 의미한다. 또한, 건강보험료 산정을 위해 국민들의 소득, 재산, 직장가입 여부, 거주정보 등도 모두 국민건강보험공단이 보유하고 있다. 이러한 우리나라의 건강보험제도의 특성인 전국민 건강보험 의무가입대상, 단일 보험자, 당연지정제, 행위별 수가제 등에 대해서는 건강보험 제도로서의 효율성 측면이 취약할 수 있다는 부분에서는 학자들 간 이견이 존재하지만, 자료 측면에서 있어서는 더할 나위 없는 장점으로 작용하고 있다[8].

국민건강보험공단은 2020년 현재, 건강보험, 장기요양 보험, 사회보험료 징수 업무 등을 수행하며 36개 공공기관과의 자료 연계를 통한 자격 및 보험료 징수 자료, 환자 의료이용 정보 등 진료내역, 건강행태·실측정보 등으로 구성된 건강검진 정보 등 약 3조 9,000억 건의 우리나라 전체 국민과 관련된 대용량 정보를 보유하고 있다. 건강보험심사평가원도 건강보험료 징수 자료, 건강검진 정보 등을 제외한 전체 국민의 진료내역 등을 보유하고 있다[4].

## 건강보험 자료의 종류와 특징

전술한 바와 같이 우리나라 건강보험제도에 기반하여 구축된 건강보험 자료는 정부의 데이터 개방정책에 따라 국민건강보험공단 및 건강보험심사평가원에서 각각 해당 기관 고유의 역할에 따라 수집된 자료를 제공하고 있으며, 2020년 8월 데이터 3법이 시행됨에 따라 건강보험 자료와 타 기관과의 자료연계도 가능해졌다.

국민건강보험공단 및 건강보험심사평가원이 제공하는 연구용 건강보험 자료의 종류, 특징 그리고 이용방법 등을 살펴보면 다음과 같다.

### 국민건강보험공단이 제공하는 연구용 건강보험 자료의 종류와 특징

국민건강보험공단은 국민건강보험을 운영하는 과정에서 축적된 건강보험 및 노인장기요양보험 자료 등과 같은 1조 3,000억 건 이상의 자료 중 일부를 「개인정보보호법」 제2조 1호의 2에 따라 가명처리한 별도의 분석용 자료인 국민건강정보 DB로 구축하여 연구목적으로 제공하고 있다[9,10]. 국민건강정보 DB는 자격 및 건강보험료 테이블, 진료명세서 일반내역 테이블, 진료명세서 상세내역 테이블, 수진자 상병내역 테이블, 처방전 교부상세 테이블, 요양기관 테이블, 건강검진 테이블 그리고 노인장기요양보험 관련 테이블 등으로 구성되어있다[8,11,12].

자격 및 건강보험료 테이블에는 재산 및 소득 수준에 기반한 건강보험료 정보 외에도 거주지역, 성별, 연령, 장애정보 등 인구사회학적 정보를 포함하고 있다. 진료이용 내역을 담고 있는 테이블은 진료명세서 일반내역 테이블(T20), 진료명세서 상세내역 테이블(T30), 수진자 상병내역 테이블(T40), 처방전 교부상세 테이블(T60)로서 각각의 테이블에 포함된 공통키를 통해 서로 일 대 다(1:N)의 관계로 연결된다. 즉, 요양기관의 진료비 청구건 단위로 구성된 진료명세서 일반내역 테이블은 공통키, 요양개시일자, 요양기관, 표시과목, 주진단 및 부진단, 수술여부, 입내원일수, 심사결정요양급여비용총액 등의 정보로 구성되어있고, 이것과 일 대 다(1:N)의 관계로 연결된 진료명세서 상세내역 테이블은 청구건을 구성하는 각각의 진료행위(진찰, 처방, 검사 등)별 단가, 일일투여량/실시횟수, 총투여일수/실시횟수, 금액 등 진료비 청구건의 상세내역으로 구성된다. 또한, 진료명세서 일반내역 테이블과 일 대 다(1:N)의 관계로 연결된 수진자 상병내역 테이블은 주진단 및 부진단을 비롯하여 진료명세서 일반내역에 담고 있지 못한 추가 진단명이 모두 포함되어있다. 처방전 교부 상세내역 테이블은 원외처방내역으로 약국 의료이용에 관한 정보로 구성되어있다.[1] 건강검진 테이블에는 건강검진 등을 통한 신체계측 정보, 검진결과 그리고 과거병력, 가족력, 음주, 흡연 등과 같은 생활습관 등의 문진내용으로 구성된다. 요양기관 테이블에는 요양기관에 대한 기본적인 정보가 있는 자료로 요양기관별 의사, 간호사 등 인력 및 병실 보유 정보를 포함하고 있다. 노인장기요양보험 정보는 2008년 7월부터 시행된 노인장기요양보험제도 운영을 위해 수집된 자료로서 노인장기요양보험 대상자들에 대한 일상생활수행능력 점수, 인지환산점수, 행동환산점수, 간호환산점수, 재활환산점수, 장기요양 등급판정인정 등급 등으로 구성된 노인장기요양보험 기본테이블, 재가급여, 시설급여 등 각각의 노인장기요양급여종류에 따른 청구내역 등으로 구성된 노인장기요양 청구내역 테이블, 장기요양 인정조사표에 따른 측정점수로 구성된 장기요양 인정조사사항 테이블 그리고 노인장기요양서비스를 제공에 대한 급여를 청구한 장기요양시설정보 테이블 등으로 구성되어있다(Table 1).

**Table 1.** Composition and main contents of national health information DB

| Composition | Main contents |
|---|---|
| Qualification | - Personal ID, sex, age, location, type of subscription, social economic variable of the subject such as income rank, disability, death(date of death), and etc. |
| Treatment | |
| Statement (T20) | - Common key, personal ID, start date of medical care, medical institution, medical subject code, diagnosis, perform surgery, length of stay, medical expense, and etc.<br>※ Common key to link with T30, T40, T60 |
| Details of treatment (T30) | - Common key, personal ID, start date of medical care, medication, dosage and frequency, cost of medication, medical expense code(precedure inclued), and etc.<br>※ Common key to link with T20, T40, T60 |
| Type of disease (T40) | - Common key, personal ID, start date of medical care, medical subject code, principle diagnosis, additional diagnosis<br>※ Common key to link with T20, T30, T60 |
| Details of prescription (T60) | - Common key, personal ID, start date of medical care, medication, dosage, day of administration, and etc.<br>※ Common key to link with T20, T30, T40 |
| Medical institution | - Type of medical institution, number of physicians, number of nurses, number of beds, and etc. |
| Health screening | - Personal ID, start date of medical care, height, weight, BMI, BP level, FBS level, family history, past history, smoking status, drinking status, and etc. |
| Long term care | - Personal ID, start date of long term care, checklist of long term care, long term care score/grade, long term care providing Institution information, Days of long term care, Total cost of long term care benefits, and etc. |

국민건강보험공단의 전체 운영계 시스템(operating system) 그리고 데이터웨어하우스 시스템(data warehouse system, DW)에서 연구목적

1) 처방전 교부 상세내역 테이블은 요양기관(약국 제외)에서 원외처방전을 발행한 내역으로, 실제 약국에서 처방전에 따른 의료이용 내역은 진료명세서 일반내역 테이블(T20), 진료명세서 상세내역 테이블(T30)에 존재함.

**Table 2.** The National Health Insurance Service sample research DB type and characteristics

| Database | Subject | Population & Sample | Duration | Contents |
|---|---|---|---|---|
| NHIS-NSC | Sample Cohort DB 2.2 | - Population: Qualified individuals as of 2006<br>- Sample: 1,000,000 (2% of population)<br>- Method: Stratified sampling (sex, age, location, insurance premium, type of subscription) | 2002-2019 | - Qualification data<br>- Medical treatment data<br>- Health screening data<br>- Clinic data<br>- Long Term Care data |
| NHIS-Senior | Senior Cohort DB 2.0 | - Population: Qualified individuals as of 2008 in the age of 60-80<br>- Sample: 511,953 (8% of population)<br>- Method: Simple random sampling | 2002-2019 | - Qualification data<br>※ Excluding cause of death<br>- Medical treatment data<br>- Health screening data<br>- Clinic data<br>- Long Term Care data<br>※ 2008-2019 |
| NHIS-HealS | Health Screening Cohort DB 2.1 | - Population: Qualified individuals as of 2002 in the age of 40-79 in 2002-2003 who received general Health screening<br>- Sample: 514,866 (10% of population)<br>- Method: Simple random sampling | 2002-2019 | - Qualification data<br>- Medical treatment data<br>- Health screening data<br>- Clinic data |

NHIS, National Health Insurance Service; NSC, National Sample Cohort; HealS, Health Screening Cohort.

으로 재구성된 국민건강정보 DB는 연구자들의 요청에 따라 크게 2가지 형태로 제공된다.

첫 번째는 표본연구 DB로서, 주제별로 대표성 있는 연구대상자 선정을 위해 특정 기준의 대상자를 층화표본추출 또는 단순무작위추출 등의 방법으로 표본추출하여 주제별로 규격화된 자료이다. 현재 3가지 주제의 코호트 형태(직장여성코호트 DB 및 영유아검진코호트 DB는 자료 업데이트 불가능 등의 이유로 각각 2022년 6월 및 7월에 제공 중단)로 구축되어 제공되고 있으며, 주요 특징을 정리하면 Table 2와 같다[11,13-15].

두 번째는 맞춤형 연구 DB로서, 건강보험 자료 등을 정책 및 학술연구 목적으로 이용할 수 있도록 연구자의 연구목적에 따라 추출, 요약 그리고 가공하여 제공하는 맞춤형 자료 형태의 데이터이다. 이는 표본연구 DB와 달리, 연구주제별로 다양한 형태의 자료를 이용할 수 있다는 매우 큰 장점이 있으나, 제공 가능한 건강보험 자료에서의 연구대상자 기준 설정, 자료항목의 선택과 그 기준 등이 모두 해당 연구자의 몫임에 따라 건강보험 자료의 특성을 제대로 파악하지 못하고 자료를 활용할 경우 연구진행에 있어 매우 곤란한 상황에 처할 수도 있다. 맞춤형 연구 DB 또한, 표본연구 DB와 마찬가지로 건강보험 자료 범위 내에서 제공 가능하며, 개인, 법인 및 단체 등에 대한 정보식별이 불가능한 형태로 가공하여 제공함을 원칙으로 하고 있다[13].

전술한 표본연구 DB 및 맞춤형 연구 DB 이외에도 환경성 질환인 천식, 알레르기 비염, 아토피 관련 연구를 지원하기 위한 DB도 별도로 구축하여 제공하고 있으며, 2013-2017년까지 해당 질환에 대한 일자별 의료이용 데이터(외래, 입원, 응급의료에 대한 진료에피소드 형태로 재산출한 진료건수), 연도별 전국민 및 의료이용 실인원수, 주소정보 및 거주자 평균 위치 좌표데이터도 함께 제공한다[16].

표본연구 DB 및 맞춤형 연구 DB에 대한 자료이용신청은 공익적 목적의 학술연구, 정책연구, 과학적 연구를 수행하고자 하는 정부 부처, 공공기관, 공공연구기관, 민간기업, 개인 등 모두 가능하지만, 이용가능 자료범위 및 이용방법 등에 있어서는 차이점이 존재한다. 표본연구 DB는 이미 주제별로 규격화되어 구축된 자료의 범위 내에서 이용할 수 있지만, 맞춤형 연구 DB의 경우 수요자 요구에 따라 가공되어 제공되므로, 특정 기준에 따른 연구대상자의 선정 및 정보, 가장 최근까지의 진료이용 내역 등도 받을 수 있는 장점이 있다. 그러나, 표본연구 DB에는 통계청의 사망원인자료가 포함되어있으나, 맞춤형 연구 DB에는 제외되어있으므로, 필요 시 통계청의 마이크로데이터서비스(Microdata integrated Service, MDIS)를 통해 자료 연계이용 신청을 별도로 하여야 한다. 자료이용방법에 있어서는 표본연구 DB 및 맞춤형 연구 DB 모두 각각의 연구자에게 부여된 권한(가상화룸)을 통해 국민건강보험공단 시스템에 접속하여 이용하여야 한다. 다만, 표본연구 DB는 개인의 특정 공간에서 인터넷 등 정보통신망을 통해 가상 사설망(virtual private network, VPN)으로 연결된 원격연구분석시스템에 접속하여 이용하는 반면, 맞춤형 연구 DB는 국민건강보험공단이 운영하는 전국 11개의 빅데이터 분석센터의 자료분석실에 방문한 후 해당 시스템에 접속하여 이용하여야 한다[17].

## 건강보험심사평가원이 제공하는 연구용 건강보험 자료의 종류와 특징

건강보험심사평가원은 요양기관으로부터 청구받은 요양급여비용을 심사하고 요양급여 적정성 등을 평가하는 과정에서 구축된 자료인

**Table 3.** The Health Insurance Review & Assessment Service patient sample dataset type and characteristics

| Dataset | Subject | Sample | Duration | Contents |
|---|---|---|---|---|
| HIRA-NPS | Total patient | - Population: Total patient by year<br>- Sample<br>• 2009-2018: About 1,400,000/year (3% of population)<br>• 2019~: About 1,000,000/year (2% of population)<br>- Method: Stratified sampling (sex, age) | 2009-Present | - Annual medical service usage history of sample patients by subject<br>※Sample patient data sets by year cannot be linked to each other<br>• Statement (T200)<br>• Details of treatment (T300)<br>• Type of disease (T400)<br>• Details of prescription (T530) |
| HIRA-NIS | Inpatient | - Population: Inpatient by year<br>- Sample<br>• 2009-2016: About 1,000,000/year (13% of population)<br>• 2017~: About 750,000/year (10% of population)<br>- Method: Stratified sampling (sex, age) | 2009-Present | |
| HIRA-APS | Elderly patient (over age 60) | - Population: Elderly patient by year<br>- Sample<br>• 2009-2016: About 1,000,000/year (20% of population)<br>• 2017~: About 700,000/year (10% of population)<br>- Method: Stratified sampling (sex, age) | 2009-Present | |
| HIRA-PPS | Pediatrics p atient (over age 20) | - Population: Pediatrics patient by year<br>- Sample<br>• 2009-2020: About 1,000,000/year (10% of population)<br>- Method: Stratified sampling (sex, age) | 2009-2020 | |

HIRA, Health Insurance Review & Assessment Service; NPS, National Inpatient Sample; NIS, National Patient Sample; APS, Adult Patient Sample; PPS, Pediatric Patient Sample.

진료행위정보, 의약품 정보, 치료재료 정보, 의료자원 정보, 의료의 질 평가 정보, 비급여 정보 등의 일부를 학술·과학적 연구 활성화를 위한 목적 등으로 별도의 자료로서 구축하고, 이를 보건의료빅데이터개방시스템을 통해 학계, 산업계 등의 연구자들에게 제공하고 있다. 건강보험심사평가원이 제공하는 연구용 건강보험 자료의 형태는 국민건강보험공단이 제공하는 것과 유사하게 크게 2가지 형태로 나누어진다.

첫 번째, 환자표본자료는 연도별 의료이용을 한 전체 환자 단위로 층화계통 추출한 자료로서, 단년도 형태의 의료이용 내역으로 구성된 횡단면 자료이다. 현재 제공되는 환자표본자료의 형태는 주제별로 환자데이터셋(HIRA-NPS), 입원환자데이터셋(HIRA-NIS), 고령환자데이터셋(HIRA-APS), 소아청소년환자데이터셋(HIRA-PPS)으로 구분되는데, 각각의 자료에는 대상자별로 요양급여비용 청구명세서상에 기록된 명세서 일반내역(T200), 진료내역(T300), 상병내역(T400), 원외처방내역(T530) 정보로 구성되어 있다[18] (Table 3).

두 번째, 맞춤형 연구자료는 요양급여비용 청구자료를 연구목적에 맞게 가공하여 제공하는 연구자료이다. 진료개시일 기준 2007년부터 현 시점 기준 8개월 전까지 제공이 가능한 자료로서, 상병, 행위, 약품의 조건으로 추출하여 연구자에게 제공한다. 제공자료의 구성은 환자표본자료와 동일하다[19].

자료이용신청은 국민건강보험공단의 자료이용 가능자와 유사하지만, 이용방법에 있어서는 일부 차이가 존재한다. 맞춤형 연구자료가 다른 자료와 연계·결합된 경우, 개인식별 위험이 높다고 판단되는 경우 그리고 이용자의 소속이 산업체인 경우 등은 전국 11개의 건강보험심사평가원의 빅데이터센터를 방문하여 자료이용이 가능하며, 이에 해당되지 않은 맞춤형 연구자료 및 환자표본자료는 연구과제별로 할당된 가상화 ID를 통해 원격접속통계분석시스템에 접속하여 이용할 수 있다.

## 건강보험 자료의 효과적 활용방법

국민건강보험공단 자료와 건강보험심사평가원 자료는 모두 건강보험 자료에 기반한 자료이지만, 기관의 설립목적에 따라 수집되는 자료 구성 및 특성에 있어 일부 차이점이 존재한다. 국민건강보험공단은 보험자임에 따라 자격 및 보험료 정보, 현물 및 현금급여와 관련한 의료이용 정보, 건강검진 및 문진 정보, 노인장기요양 정보 등을 보유하고 있어 이를 중심으로 이용자에게 정보가 제공되지만[2], 건강보험심사평가원은 현금급여를 제외한 요양기관에서의 의료이용 내역 중심으로만 자료를 보유하고 있어 이에 대해서만 제공하고 있다. 다만, 건강보험심사평가원의 경우, 자료제공에 있어 수요자의 요구에 따라 의료영상자료, 약제 관련 부가정보, 의약품 안심서비스(Drug Utilization Review,

2) 국민건강보험공단의 건강보험자료공유서비스에서 제공하는 의료이용 기본자료는 현물급여에 대한 내역만 제공함. 이에 따라 현금급여 활용이 필요한 경우, 별도의 협의가 필요함.

DUR) 정보, 의약품 유통관리 정보, 요양급여 적정성 평가에 대한 자료 및 일부 비급여 항목에 대한 자료도 추가적으로 제공한다[20].

위와 같은 건강보험 자료를 효과적으로 이용하여 연구주제 및 목적 등에 따른 연구자가 의도하는 정확한 결과를 도출하기 위해서는 다음과 같은 부분에 대한 주의 및 검토가 필요하다.

첫째, 건강보험 자료는 건강보험제도권 내에서 행하여지는 의료이용에 대한 내용만 포함되며, 비급여는 제외된다. 건강보험 자료 중 국민건강보험공단과 건강보험심사평가원이 동일하게 보유하고 제공하는 자료는 국민들의 진료이용 내역인데, 건강보험 운영 체계에 따른 각 기관의 역할에 따라 보유하고 있으므로 자료가 가지는 의미에서 일부 차이가 존재한다. 건강보험 운영 체계에 따르면, 요양기관은 환자들에게 진료를 행한 후, 환자본인부담금을 제외한 금액을 건강보험심사평가원을 거쳐 국민건강보험공단으로 청구하게 된다. 요양급여비 청구를 요청받은 건강보험심사평가원은 해당 청구건에 대해 여러 단계의 심사를 거치게 되고 그 최종 심사결과를 국민건강보험공단으로 통보하면 국민건강보험공단은 요양기관에게 심사결정된 요양급여비를 지급하게 된다. 즉, 건강보험심사평가원은 진료이용 내역 자료에 있어 요양급여비 청구에서 심사의 단계에서 발생하는 모든 정보를 보유하고 있으나, 국민건강보험공단은 최종 심사결정된 진료이용 정보만을 보유하고 있다.

결과적으로는 두 기관 모두 요양기관에서 제공한 의료서비스에 대해 심사결정된 것을 기본적으로 연구자에게 제공한다. 다만, 최종 심사결정된 진료비는 진료명세서 일반내역 테이블에만 반영되며, 진료명세서 상세내역 테이블에 있는 각각의 진료행위에는 반영되지 않은 형태로 제공된다. 또한, 진료명세서 일반내역 테이블의 진료비는 요양기관종별 가산율이 반영된 형태이지만, 진료명세서 상세내역 테이블에는 미반영된다. 이에 따라 개념적으로는 진료명세서 일반내역 테이블의 총 진료비와 진료명세서 상세내역 테이블의 각각 진료행위별 진료비들의 합계는 서로 일치하여야 하지만 실제로는 일치하지 않는다. 이에 각 테이블 간의 연계를 통한 각 진료행위에 따른 금액이 전체 진료비에서 차지하는 비율 등과 같은 진료비 통계 산출 시에는 주의가 필요하다.

그 밖에 건강보험 의료이용 자료는 청구건 단위이므로, 요양기관에서는 특정 환자의 진료기간이 길면 동일질병에 대해 월 단위 분리청구를 할 수 있다. 그러므로 연구주제에 따라 "질병의 발생부터 종료까지를 하나의 사건으로 측정하는 단위"인 진료에피소드(episode of care)의 개념을 적용할 필요도 있다. 그러나 청구단위의 건강보험 진료이용 자료를 진료에피소드 단위의 분석자료로 재구성하는 경우, 진료 첫 시작점과 종료시점 결정에 대한 조작적 정의가 필요할 것이다. 또한, 요양기관이 급여청구에 있어 의도적 또는 비의도적 오류 등으로 환자가 여러 곳의 요양기관 진료를 받은 경우도 일부 발생할 수도 있으므로 이에 대한 면밀한 검토도 필요할 것이다.

둘째, 분석변수에 대한 명확한 조작적 정의가 필요하다. 건강보험 청구자료에 제시된 주진단명은 요양기관에서 환자진료 중 진단명이 확정되지 않은 상태에서 환자의 호소, 증세 등에 따라 기입된 일차 진단명이거나 또는 요양기관의 진료비 삭감 방지를 위한 진단명 코드 변경(업코딩)이 발생할 수 있으므로, 실제 질병과는 다를 수 있다[21].

또한, 특정 질병이 분석변수라 가정한다면, 전술한 것처럼, 건강보험 청구자료는 실제 정확한 진단명이 아닐 수도 있으므로, 이를 충분히 감안하여 다양한 제공된 자료항목을 종합적으로 고려하여 조작적 정의를 내릴 필요도 있다. 당뇨병, 고혈압, 이상지질혈증의 경우, 만성질환임에 따라 건강검진에 포함되는 공복혈당, 혈압수치, 혈중 콜레스테롤을 함께 고려하거나 진료명세서 상세내역에 있는 정보(주성분 코드 등)를 이용한 투약 내역정보를 함께 고려하는 등의 노력이 필요하다.

셋째, 코호트 내 환자-대조군 연구를 위한 국민건강보험공단의 자료 활용 시 특정 요인에 노출된 집단(환자군)과 노출되지 않는 집단(대조군)의 선정에 있어 두 집단 간의 차이를 반영한 연구설계가 필요하다. 특히, 노출매칭 코호트 연구의 경우 코호트 형태로 비노출대상자가 매우 많거나 접근이 현실적으로 불가능하여 분석 불가능 상태이면 노출대상자의 인구사회학적 요인, 건강검진, 동반질환 등을 파악하여 그에 비슷한 비노출대상자와의 매칭을 통해 약 1-10배수를 선정하는 방식이다[7]. 이러한 연구방법론은 이미 후향적 코호트 연구가 가능한 형태로 제공되는 국민건강보험공단의 자료를 활용한 연구에 많이 적용되고 있는데, 코호트 내 환자-대조군 연구 또는 노출매칭코호트 연구 등에서 해당 집단 선정 시 교란편의(confounding bias), 정보편의(information bias), 출판편의(publication bias) 등이 발생할 수 있으므로 집단 간 편의에 대한 차이보정이 필요하다. 왜냐하면, 교란인자들을 통제하지 못할 시, 불완전한 실험설계를 초래하게 되어 인과론적인 추론을 편향시킬 수 있기 때문이다[22]. 그러나, Lim et al. [7]의 연구결과, 검토된 건강보험 자료를 활용한 전체 논문 중 편의 보정을 위한 연구설계가 적용되지 않은 비율이 81.4%이며, 특히 편의에 대한 보정이 매우 중요한 연구인 노출매칭코호트 연구, 코호트 내 환자-대조군 연구 그리고 환자-대조군 연구에서도 편의 보정 미적용률이 각각 31.7%, 84.1% 그리고 94.4%로 나타났다. 해당 연구에서는 향후 편의 보정에 대한 필요성과 중요성을 강조하였고, 이에 따른 방법으로 성향점수매칭법(propensity score matching, PSM) 및 랜드마크 분석방법을 제시하기도 하였다.

넷째, 건강보험 자료를 이용할 경우, 실제 통계적으로 유의하지 않음에도 불구하고, 대표본 효과(large sample size effect)로 인한 통계적으로 유의한 결과가 나올 가능성이 높으므로, 통계적 유의성뿐만 아니라

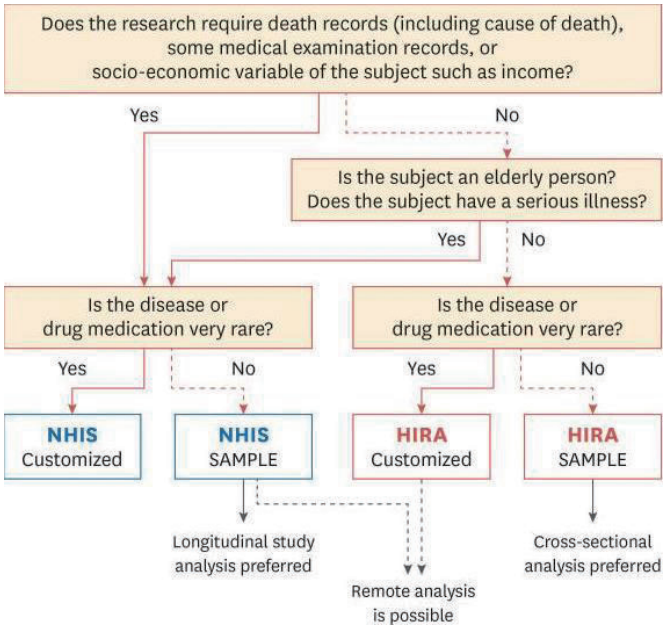


**Figure 1.** Selection of appropriate claims data according to research purpose.

임상적 유의성도 함께 고려한 해석이 필요하다. 이에 대해 Chung et al. [23]의 연구에서는 임상연구 수행 시 분석과정의 과학적 신뢰와 결과 도출 시 합당성을 판단하는 과정에 임상의료 전문가의 개입은 물론 데이터사이언티스트, 역학 및 통계전문가와의 공동연구를 수행하는 것을 제안하였다[23].

다섯째, 국민건강보험공단 및 건강보험심사평가원이 제공하는 건강보험 자료의 선택과 이용 측면에 있어서는 연구주제, 연구목적, 연구설계 등에 따라 두 기관의 자료 활용도는 달라질 수 있으며, 해당 기관에서 제공하는 표본 자료와 맞춤형 자료에 대한 선택도 달라질 수 있다. 구체적으로는 수행하고자 하는 연구가 사회경제적 특성, 건강검진 및 문진 내용, 사망원인(통계청과 자료연계 필요), 희귀질환 및 이와 관련한 세부 의료행위 그리고 연구설계가 종단연구(longitudinal study)이면, 국민건강보험공단의 자료 중 맞춤형 연구 DB가 적합할 것이다. 또한, 건강보험심사평가원이 제공하는 자료는 단년도별로 구성되어 각 연도 간 연계가 불가능하므로 횡단면 연구(cross sectional study)에 보다 적합할 것이다. 이에 대해 Kyung and Kim [24]의 연구에서는 위 내용을 중심으로 Figure 1과 같은 자료선택 체계도를 제시하기도 하였다.

그 밖에 최근 자료이용신청자 급증, 가상화 계정 수 부족 그리고 자료제공 인력의 부족 등으로 최소 3개월에서 최대 6개월 이상 대기기간이 필요함에 따라 건강보험 청구자료를 활용한 연구를 계획할 경우 해당 대기시간을 충분히 고려할 필요가 있다. 또한, 국민건강보험공단의 건강보험 표본코호트의 경우, 최근 민감상병 및 이에 따른 수가코드에 대한 비식별 조치(특수상병 2,645개 및 법정감염병 570개) [11]가 증가하여 이와 관련된 연구를 진행하고자 하는 경우 사전검토가 필요하다.

## 결론 및 제언

보건의료분야의 자료는 타 영역과 견주어볼 때 국가 단위로 생산되는 자료가 매우 많다. 통계청의 국가승인통계 현황만 보더라도 2022년 7월 현재 전체 1,291종 중 보건영역에 해당하는 통계수가 63종(4.8%)으로 타 영역(인구 0.8%, 사회일반 1.9%, 범죄·안전 2.5%, 노동 2.9%, 소득·소비·자산 0.5%, 복지 2.2%, 교육·훈련/문화·여가 3.0%, 주거/국토이용 1.9%, 경제일반·경기/기업경영 3.2%, 농림/수산 4.4%) [25]에 비해 해당 통계를 산출하기 위한 국가 단위의 자료원이 많다는 것을 간접적으로 유추할 수 있다. 또한 최근 스마트 헬스케어(smart healthcare)의 등장, 정보통신기술의 지속적인 발전과 2019년 보건복지부에서 발표한 바이오헬스 산업 혁신전략에 따른 5대 빅데이터 플랫폼 구축 사업 등과 같은 보건의료정보와 관련된 국가의 다양한 정책적 시도에 따라 보건의료분야의 자료는 향후 급속하게 증가할 것으로 예상된다.

이러한 보건의료분야의 자료를 논할 때 가장 빈번히 등장하는 것이 건강보험 자료일 것이며, 이는 많은 연구에서 보건의료 빅데이터의 대표적인 사례로 제시되고 있다.

보건의료 빅데이터라는 개념이 사전적으로 명확하게 정의되지는 않았지만, 다양한 연구자들이 제시한 정의 그리고 우리나라 「보건의료기본법」 제3조에 따른 "보건의료"는 "국민의 건강을 보호·증진하기 위하여 국가·지방자치단체·보건의료기관 또는 보건의료인 등이 행하는 모든 활동"이라는 것 그리고 빅데이터의 주요 특성이라 제시되는 크기, 속도, 다양성, 정확성, 가치 등을 종합적으로 고려해 보면, 우리나라의 보건의료자료 중 건강보험 자료가 빅데이터인 것은 분명해 보인다. 다만, 일부 연구에서는 국민건강보험공단 및 건강보험심사평가원에서 수집하는 자료에 있어 그들의 건강보험 관련 업무 자체가 비정형 데이터를 포함한 다양한 형태의 데이터를 수집할 만한 것이 아니며, 또한 현재 빅데이터 시스템에서 요구하는 신속한 정보처리를 위한 인프라가 미흡하다는 점을 고려할 때 국민건강보험공단 및 건강보험심사평가원의 자료를 빅데이터라고 부르기에는 한계가 있음을 지적하고 있다[26]. 그러나 이는 자료의 측면이라기보다는 정보시스템 측면에서 바라본 관점이 더욱 강하므로, 자료 자체만 본다면 건강보험 자료는 빅데이터임은 이견이 없을 것으로 판단된다. 또한, 일부 논란의 여지가 있을지라도 우리나라의 건강보험 자료가 우리나라 국민 전체에 해당하는 자료로서의 가치만을 고려한다면, 우리나라 보건의료분야의 소중한 자료임에는 틀림이 없다고 판단된다.

최근 데이터 3법의 통과 등을 비롯한 정부의 다양한 공공 데이터 개

방 정책에 따라 국민건강보험공단 및 건강보험심사평가원의 대규모 자료원을 활용한 연구 진행이 가능하게 되었다. 이들은 전체 국민의 의료이용을 기반으로 한 자료원이기 때문에 우리나라의 실제 보건의료 상황에 대한 실태 파악을 위해 활발히 활용되고 있다. 본 자료원들에서는 행위별 수가제하의 세부 의료이용 내역 확인이 가능하고 약물 및 처치에 대한 자세한 정보가 포함되어 있는 것도 장점이다. 일부 사용에 제한이 있지만, 명확한 연구의 목적과 적절한 행정절차를 수반한다면, 드물게 발생하는 질병이나 사건에 대한 연구 그리고 타 기관과의 자료 연계도 가능할 것이다. 그러나, 질병군별 포괄수가제 대상 질병의 경우에는 구체적 진료내역이 없는 단점이 있고[27], 급여가 인정된 의료이용만 포함되므로 비급여 항목 사용에 대한 연구는 불가능하다. 또한, 상병코드로 입력된 진단명이 부정확할 수 있으므로 질병에 대한 조작적 정의를 명확하게 하려는 노력이 선행되어야 한다.

또한, 질병 관련 연구 시 진료에피소드 관점에서의 자료 재구성, 환자-대조군 연구 시 집단 간 편의 보정을 고려한 연구설계, 통계적 유의성과 함께 임상적 유의성 등도 함께 고려한 연구결과의 해석이 필요할 것이다. 이와 관련하여, 2014-2020년간 건강보험 자료를 활용한 678편의 논문을 분석한 Lim et al. [7]의 연구에서 건강보험 청구자료를 활용하는 연구자들의 분석방법에 대한 지식 및 정보가 미비함을 지적함과 동시에 자료제공기관 등에서는 연구의 수월성과 보다 정확한 결과의 도출을 위해 분석방법 점검도구 개발이 필요함을 제시하기도 하였다.

그러나 전술된 내용보다 더욱 더 중요한 것은 연구주제에 따른 우리나라 건강보험제도에 대한 명확한 이해를 바탕으로 건강보험 자료구조와 속성을 이해하고 연구를 수행해야 한다는 것이다. 최근 매년 1,200여 건 이상의 건강보험 자료를 활용한 연구가 진행되고 있으며, 이에 따른 연구성과물이 2019년 기준 국내외 저널 등에 913편이 게재되었으며, 이 또한 최근에는 더욱 더 급증하는 추세이다[28]. 또한, 건강보험 자료에 머신러닝 등과 같은 최신 기법을 적용한 연구들도 활발히 진행되고 있어 향후 보다 의미 있는 분석결과가 도출될 것으로 기대된다. 다만, 보다 의미 있고 정확하며, 고도화된 연구결과 도출을 위한 방법론 측면에서 다양한 시도가 이루어지는 것은 매우 바람직한 현상이라 판단된다. 그러나 복잡한 기교 위주의 분석방법론에만 치중한 연구는 지양할 필요가 있다. 또한, 건강보험 자료에 대한 명확한 조작적 정의, 연구설계 및 활용된 연구방법에 대한 명확하고 자세한 기술 등이 수반된다면 연구결과물의 가독성 제고 및 연구의 질적 향상은 물론 이를 활용하고자 하는 후속 연구자들에게 많은 도움을 줄 수 있을 것으로 판단된다.

본 연구는 우리나라 건강보험제도하에서 국민건강보험공단과 건강보험심사평가원에서 수집하여 학술 및 연구목적 등으로 제공하고 있는 건강보험 자료의 특성과 활용방법 등에 대해 보건의료연구 수행 시 고려해야 할 내용을 중심으로 제시하였다. 그러나, 전반적인 건강보험 자료를 중심으로 제시함에 따라 각각의 부문별 자료내용에 대한 보다 세부적인 활용방법의 검토에는 한계가 있다. 특히, 국민건강보험공단의 노인장기요양보험 관련 자료, 건강보험심사평가원의 의료영상자료, 약제 관련 부가정보, 의약품 안심서비스 정보, 의약품 유통관리 정보, 요양급여 적정성 평가에 대한 자료 및 일부 비급여 자료 등의 구체적인 활용방법에 대해서는 검토가 미흡함에 따라 후속 연구를 통한 검토가 필요할 것이다.

## ORCID

Ilsu Park *https://orcid.org/0000-0003-0445-8556*

## REFERENCES

1. Kim JA, Kim RY. Introduction and utilization of claim data by Health Insurance Review and Assessment Service for medical health research. OLD 2014;2(1):3-9 (Korean).
2. Lee JA, Oh JW, Moon SJ, Lim JT, Lee JS, Lee JY, et al. Assessment of needs and accessibility towards health insurance claims data. Korean J Health Policy Adm 2011;21(1):77-92 (Korean). DOI: 10.4332/KJHPA.2011.21.1.077
3. National Health Insurance. Medical device and service development seminar using public medical big data. The 37th Korea International Medical & Hospotal Equipment Show; 2022 (Korean).
4. Medifonews. NHIS DB research support increased 12 times in 6 years. Available at https://www.medifonews.com/news/article.html?no=154301 [accessed on July 22, 2022].
5. National Health Insurance Sharing Service. 2021 National health information data review committee minutes. Available at https://nhiss.nhis.or.kr/bd/ay/bdaya001iv.do [accessed on July 20, 2022].
6. Health Insurance Review and Assessment Service. HIRA big data analysis educational material–Introduce to HRA bigdata and support service. Available at https://opendata.hira.or.kr/op/opb/selectRfrmList.do?rfrmTpCd=&searchCnd=&searchWrd=&sno=0&pageIndex=2 [accessed on July 16, 2022].
7. Lim HS, OH HC, Jang JH, Yoon SL, Lee JK, Park SH, et al. Research on the development of an analysis method inspection tool to improve the quality of big data research using the national health information DB. Ilsan: NHIS Ilsan hospital research center; 2020 (Korean).

8. Park JH. How to use big data from NHIS bigdata. The 5th Clinical Research Methodology Workshop, Korean Association for the Study of the Liver; 2016 (Korean).
9. National Health Insurance Service. Operational regulations for provision of national health information data; 2020 (Korean).
10. National Health Insurance Service. National Health Insurance Sharing Service. Available at https://nhiss.nhis.or.kr/bd/ab/bdaba016lv.do [accessed on August 15, 2022].
11. National Health Insurance Service. Sample Cohort DB 2.2 user manual (Ver 1.3) (Korean).
12. Park JS, Lee C. Clinical study using healthcare claims database. J Rheum Dis 2021;28(3):119-125. DOI: 10.4078/jrd.2021.28.3.119
13. National Health Insurance Service. National Health Insurance Sharing Service. Available at https://nhiss.nhis.or.kr/bd/ab/bdaba002cv.do [accessed on August 15, 2022].
14. National Health Insurance Service. Senior Cohort DB 2.0 user manual (Ver 1.0) (Korean).
15. National Health Insurance Service. Health Screening Cohort DB 2.1 user manual (Ver 2.1) (Korean).
16. National Health Insurance Service. Environmental diseases DB user manual (Ver 1.0) (Korean).
17. National Health Insurance Service. National Health Insurance Sharing Service. Available at https://nhiss.nhis.or.kr/bd/ab/bdabd003cv.do [accessed on August 15, 2022].
18. Health Insurance Review & Assessment Service. Patient Sample Data user guide. 2022 (Korean).
19. Health Insurance Review & Assessment Service. HIRA customized research analysis user guide. 2022 (Korean).
20. Son W. Choosing the right data for the type of health and medical big data and my research. The 7th Clinical Research Methodology Workshop, Korean Association for the Study of the Liver; 2018 (Korean).
21. Park EC, Jang SI. Assessment and improvement of health insurance claim disease code and medical record consistency. Seoul: Yonsei University; 2017, p. 287 (Korean).
22. Soni PD. Selection bias in population registrybased comparative effectiveness research. Int J Radiat Oncol Biol Phys 2019;103(5):1058-1060. DOI: 10.1016/j.ijrobp.2018.12.011
23. Chung HS, Kin SY, Kim HS. Clinical research from a health insurance database: practice and perspective. Korean J Med 2019;94(6):463-470 (Korean). DOI: 10.3904/kjm.2019.94.6.463
24. Kyoung DS, Kim HS. Understanding and utilizing claim data from the Korean National Health Insurance Service and Health Insurance Review & Assessment database for research. J Lipid Atheroscler 2022;11(2):103-110. DOI: 10.12997/jla.2022.11.2.103
25. Statistics Korea. Korean Statistical Information Service. Available at http://www.narastat.kr/pms/pub/scs/css/selectConfmStatsStatusRealm.do [accessed on July 1, 2022].
26. Oh SW. Medical use of health insurance big data. Medical Policy Forum 2014;12(3):18-23 (Korean).
27. Ryu DR. Introduction to the medical research using national health insurance claims database. Ewha Med J 2017;40(2):66-70 (Korean). DOI: 10.12771/emj.2017.40.2.66
28. Kim HK, Song SO, Noh JH, Jeong IK, Lee BW. Data configuration and publication trends for the Korean National Health Insurance and Health Insurance Review & Assessment database. Diabetes Metab J 2020;44(5):671-678. DOI: 10.4093/dmj.2020.0207