BRIEF REPORT
ISSN 1738-3331 eISSN 2671-826X, https://doi.org/10.7704/kjhugr.2021.0061
The Korean Journal of *Helicobacter* and Upper Gastrointestinal Research, 2022;22(1):70-74

청구 자료를 이용하는 연구의 기본 개념

박수비[1], 차재명[1,2]
강동경희대학교병원 소화기내과[1], 경희대학교 의학전문대학원 내과학교실[2]

The Basic Concept of Claim Data-based Research

Su Bee Park[1], Jae Myung Cha[1,2]
Department of Gastrointestinology, Kyung Hee University Hospital at Gangdong[1], Department of Internal Medicine, Kyung Hee University College of Medicine[2], Seoul, Korea

인류는 현재 20세기 후반에 컴퓨터와 인터넷의 등장과 함께 시작된 3차 산업혁명 시기를 지나 정보통신 기술의 융합으로 이루어지는 4차 산업혁명의 초입 단계를 직면하고 있다. 4차 산업혁명은 사람과 사물, 사물과 사물이 인터넷 통신망을 통해 연결되고 지능화되는 단계로, 핵심적인 분야로는 사물인터넷, 로봇, 인공지능, 3D 프린팅, 가상/증강현실 등이 있는데 이들을 연결하는 핵심은 데이터이다. 따라서 4차 산업혁명의 키워드는 데이터라고 할 수 있는데, 데이터 중에서도 특히 빅데이터가 핵심이다. 이러한 빅데이터들 중에서 최근 관심이 증가하고 있는 분야가 바로 의료 빅데이터이다. IBM 보고서에 의하면 세계 16,000개 병원이 환자 데이터를 수집하고 있고, 전 세계 490만 명이 원격 모니터링 디바이스를 사용하고 있다.[1] 개별 환자 모니터링 장비는 초당 1,000개의 수치를 측정하고 있고, 이는 환자 1명당 하루 86,400개의 수치가 생성되는 것을 의미한다.[1] 이처럼 의료 분야에서 발생하는 다양한 빅데이터는 연구와 진료에 이용되고 있다. 대표적인 의료 빅데이터로는 의료기관에서 진료를 할 때 생성되는 전자차트 데이터, 진료를 청구할 때 발생하는 청구 데이터, 유전자나 장내 미생물과 같은 유전체 데이터, 스마트워치와 같이 일상생활에서 환자들이 직접 생성하는 환자생성 의료 데이터들이 있다. 이들 중 최근 국내에서 활발하게 연구에 활용되고 있는 의료 빅데이터는 주로 청구 데이터이기 때문에, 이번 원고에서는 청구 데이터를 이용한 연구에 대해 주로 기술하였다.

한국과 대만의 경우 전국민 건강보험제도 의무가입을 통한 정부 주도의 단일보험 체계를 갖추고 있지만, 대만은 건강보험제도를 적용받는 요양기관으로 참여 여부를 의료인의 자유의사에 따라 결정할 수 있으며, 총액 계약제를 진료비 지불방식으로 선택하고 있다는 점이 우리나라와 차이점이다.[2,3] 우리나라는 건강보험제도를 적용받는 요양기관으로 지정을 강제하고 있으며, 행위별수가제도를 적용하고 있다는 점에서 청구 자료가 잘 구축될 수 있는 의료보험 체계를 유지하고 있다.[2] 우리나라 국민들이 의료보험료를 납부하고 의료기관에서 진료를 받으면 의료기관에서는 진료 비용에 대해 심사를 청구하고, 심사평가원에서는 심사 결과를 의료기관과 건강보험공단에 통보하게 된다(Fig. 1). 이때 발생하게 되는 심사 청구 자료를 이용하여 다양한 임상 연구를 시행할 수 있다. 청구 자료는 심사평가원(Health Insurance Review and Assessment, HIRA) 자료와 건강보험공단(National Health Insurance Sharing Service, NHISS) 자료가 있는데,[4] 두 자료는 공통으로 진료데이터베이스를 공유하고 있지만, 건강보험공단 자료에는 심사평가원 자료에 포함되어 있지 않은 수진자의 자격 정보와 건강검진 정보 등을 포함하고 있다. 청구 자료는 요양급여 비용명세서를 통해 구축되는데, 요양급여 비용명세서는 일반내역을 알 수 있는 200테이블, 상병내역을 알 수 있는 400테이블, 진료내역을 알 수 있는 300테이블 및 원외처방내역을 알 수 있는 530테이블로 구성되어 있다.[5] 200테이블의 일반내역에는 환자 기본 정보(예; 명세서 조인키, 수진자 대체키, 성별, 연령, 보험형태), 주상병 및 제1 부상병과 같은 기본 상병, 진료 정보(예; 내원 경로, 요양개시/종료일, 입원/외래, 요양기관 번호), 급여 비용(예; 청구 비용, 심사결정 비용) 등이 포함되어 있다. 400테이블에는 200테이블의 주상병 및 제1 부상병을 포함한 모든 상병 정보가 포함되어 있고, 300테이블에는 검사, 시술 및 수술, 치료 재료, 원내 조제 내역 등의 진료 내역이 포함되어 있으며, 530테이블에는 모든

Received: November 24, 2021 **Revised:** December 12, 2021 **Accepted:** December 12, 2021

Corresponding author: Jae Myung Cha
Department of Internal Medicine, Kyung Hee University Hospital at Gangdong, Kyung Hee University College of Medicine, 892 Dongnam-ro, Gangdong-gu, Seoul 05278, Korea
Tel: +82-2-440-6113, **Fax:** +82-2-440-6295, **E-mail:** drcha@khu.ac.kr

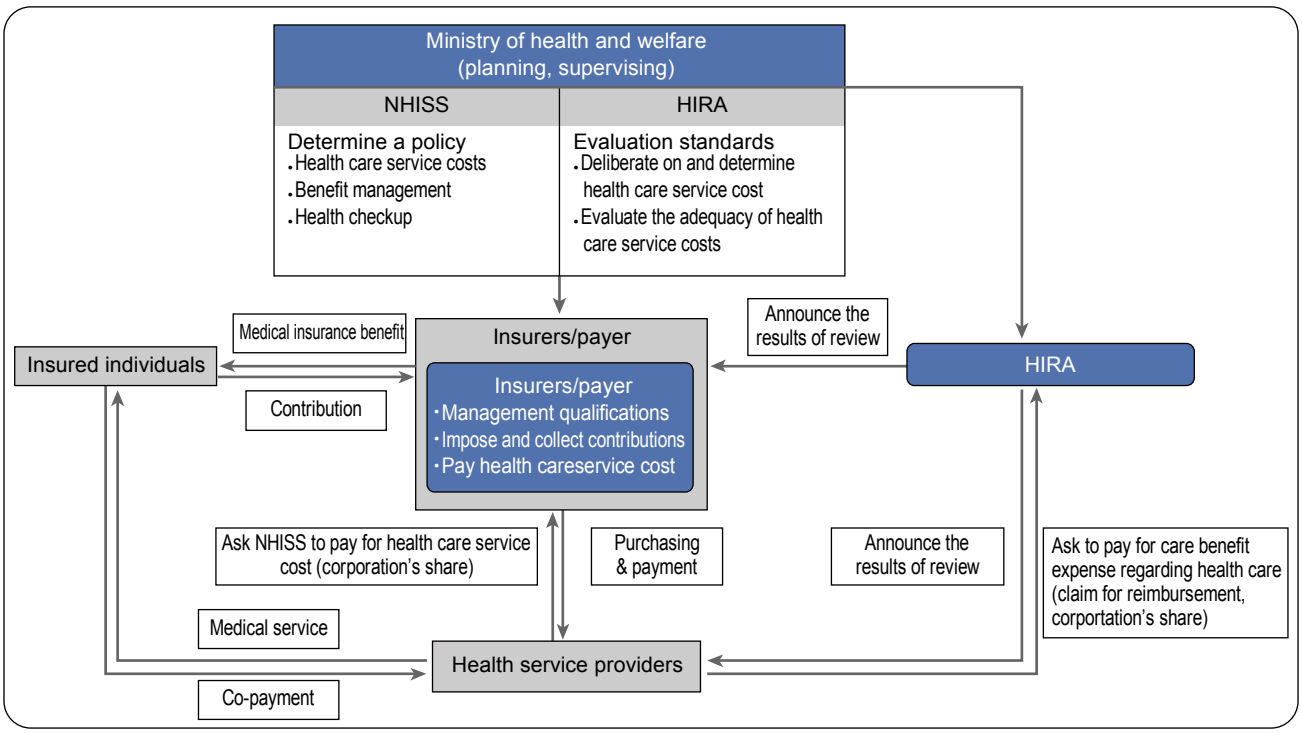

Fig. 1. The national health insurance system, including 'national health insurance review and assessment (HIRA)' and 'national health insurance sharing service (NHISS)'.[5]

Table 1. Sample Research Cohort of National Health Insurance Sharing Service (NHISS) and Health Insurance Review and Assessment (HIRA)

Type of cohort	Sample size	Duration	Computaion standard
Sample research cohort of NHISS			
Standard sample	1,000,000 (200 GB)	2002~2013 (12 years)	Qualified individuals as of 2002 (approximately a million, 2% of population)
Medical check-up	5,150,000 (100 GB)	2002~2013 (12 years)	Qualified individuals as of 2002 in the age of 40 to 79 in 2002 to 2003 who received general medical check-up (10% of population)
Elderly	5,580,000 (100 GB)	2002~2013 (12 years)	Qualified individuals as of 2002 who is over age of 60 (10% of population)
Working women	840,000	2002~2013 (12 years)	Qualified individuals as of 2002 who is age of 15 to 64 working women (5% of population)
Infant medical	1,850,000	2008~2015 (8 years)	Out of total check-up recipients who received at least one of 1st to 2nd infant medical check-up (5% of sample is extracted for each birth year of 2008 to 2015)
Sample research cohort of HIRA			
HIRA-NIS (inpatient)	1,750,000	2009~2019 (10 years)	13% of inpatient population between 2009 to 2016 10% of inpatient population since 2017
HIRA-NPS (national patient)	1,400,000	2009~2019 (10 years)	3% of national population between 2009 to 2018 2% of national population since 2019
HIRA-APS (aged patients)	1,700,000	2009~2019 (10 years)	20% of ≥65 year population between 2009 to 2016 10% of ≥65 year population since 2017
HIRA-PPS (pediatric patient)	1,000,000	2009~2019 (10 years)	10% of pediatric (<20 year) population since 2009

NIS, national inpatient sample; NPS, national patient sample; APS, adult patient sample; PPS, pediatric patient sample.

원외 처방 내역이 포함되어 있다. 한편, 각 테이블은 명세서 조인키라는 고유키로 서로 연결이 가능하고, 요양기관 현황은 요양기관 대체키로 연결이 가능하여 해당 데이터들을 통합적으로 분석할 수 있다.

청구 자료는 연구자가 원하는 자료를 원하는 형태로 제공하는 맞춤형 자료와 일정 샘플을 사전 제작하여 제공하는 표본 샘플 자료가 있다. 맞춤형 자료는 전국민 대상이며 모든 상병을 포함할 수 있기 때문에 대표성을 확인할 필요가 없다. 환자 수가 부족해 세부 분석을 못하는 경우가 없고, 사망 자료와도 연계가 가능하며, 장기추적 연구가 가능한 장점이 있다. 하지만 데이터 분량이 너무 커서 추출 후 전달받지 못할 수 있으며, 원주나 서울 센터에 직접 방문을 해서 분석해야 하는 공간적인 제약이 있다. 한편, 표본 샘플 자료는 후천성 면역결핍증이나 정신과 질환과 같이 민감상병을 제외하고 표본 인구 100만 명에 대한 자료를 제공하고 있다. 100만 명이라고 해도 희귀질환이나 표본 수가 적은 질환은 포함되는 환자 수가 적어도 연수를 수행할 수 없는 단점이 있고, 질병에 따라 실제 발생률 및 유병률의 차이가 날 수 있는 단점이 있다. 하지만 대상자 선정 후 추출이 용이하고, 맞춤형 자료에 비해 신청 후부터 자료 열람까지 대기 시간도 짧고, 클라우드 시스템으로 방문하지 않고 원격 접근할 수 있기 때문에 데이터 접근성이 용이한 장점이 있다. 무엇보다도 모든 약제에 대한 정보가 포함되어 있기 때문에 약제 연구도 가능하다. 표본 샘플 자료는 각 코호트에 따라 표본 수, 기간, 내용이 다르기 때문에 본인의 연구 목적에 맞는 샘플 자료를 잘 선택해야 한다(Table 1). 일반적으로 연구를 수행하기에는 맞춤형 자료가 좋지만 대기 시간도 오래 걸리고 직접 방문하여 분석해야 하는 시간과 공간적인 제약이 있기 때문에, 일단 표본 샘플 자료로 본인 연구를 수행할 수 없는지 검토하고 안될 때만 맞춤형 자료를 신청하는 것이 좋은 전략이다. 건강보험공단 자료와 심사평가원 자료는 별개로 다양한 표본 샘플 자료를 제공하고 있기 때문에,[6] 각 코호트 자료의 차이점을 잘 비교하여 본인 연구에 적합한 표본 샘플 자료를 선택해야 한다.

청구 자료를 이용한 빅데이터 연구는 관찰 연구이기 때문에 무작위 대조 연구 만큼의 근거 수준을 확보하기 어렵다. 따라서 무작위 대조 연구와 비교하여 상대적인 강점이 있는 연구 주제를 선택하여 연구를 시행하는 것이 무엇보다 중요하다. 무작위 대조 연구는 이상적인 환경에서 정해진 환자에 대해 정해진 중재만 시행할 수 있지만, 청구 자료를 이용한 연구는 다양한 임상 환경에서 다양한 환자에 대해 여러 가지 중재 결과를 취합할 수 있다(Table 2). 뿐만 아니라, 많은 환자군을 확보하는 데에도 연구 비용이 적게 들기 때문에 대장내시경 천공과 같이 대규모 모집단이 필요한 비교적 드문 합병증에 대한 연구에 보다 합당하다. 또한 무작위 대조 연구에서 윤리적인 이슈 때문에 포함하기 힘든 임산부나 고령 환자에 대한 데이터도 확보할 수 있다. 하지만 연구 목적으로 취합되지 않은 2차 자료이기 때문에 자료가 정확하지 않아 데이터의 정결 과정이 필요하며, 보험 급여로 청구되는 데이터만 취합할 수 있으며, 비뚤림과 교란 효과가 많은 데이터이기 때문에 오히려 분석이 더 어렵다는 단점이 있다. 게다가 청구 자료는 혈액 검사 결과나 시술 결과를 알 수 없다는 단점이 있다. 따라서 청구 자료를 이용하는 연구는 대규모 환자에 대한 장기 추적 연구나 최소의 비용으로 짧은 시간에 시의적절한 연구 결과를 도출해야 하는 연구에 적합하다. 한편, 청구 자료로 정의가 어려운 변수에 대한 연구, 인과관계를 규명하는 연구, 비뚤림 영향이 많은 연구, 대조군 설정이 어려운 연구 등은 연구 주제로 적합하지 않다. 최근 청구 자료를 이용하여 많이 시행되고 있는 연구 주제들은, 1) 발병률이나 유병률과 같은 역학 연구,[7] 2) 위험 인자 분석 연구,[8] 3) 약물이나 시술의 결과 분석 연구, 4) 초고령 환자나 임산부와 같이 임상 연구를 시행하기 어려운 연구, 5) 무작위 대조 연구 이전에 방향성을 가늠하기 위한 파일럿 연구, 6) 인공지능 기반 예측 모델이나 알고리즘 개발 연구 등이다.

청구 자료를 이용한 연구를 시행하기 위해서는 먼저 연구의 실현 가능성을 검토해야 하는데, 진단코드가 정확한지를 확인해야 하며, 필요하면 조작적 정의를 사용해야 한다. 예를 들어, 비교적 진단코딩이 정확한 암이나 염증성 장질환에 대한 연구

Table 2. Strengths and Weaknesses of Claim Data-based Research

Strengths	Weeknesss
Universal data	Secondary data
Real world data	Unclear accuracy and reliability of data
Low cost and time for research	Only reimbursed data
Population-based research	Limitation of space and time for data use
Fulfill unmet need of clinical trial	No laboratory data and clinical result
Enough demographic and socioeconomic data	Complex administrative process
	No accessibility for sensitive data
	Need large disk space and cost of data use

는 합당한 주제이지만, 진단코딩이 정확하지 않은 기능성 소화불량이나 역류성 식도질환에 대한 연구는 합당한 주제가 아니다. 청구 자료를 이용하는 연구에 조작적 정의를 많이 사용하게 되는데, 조작적 정의는 측정 가능한 구체적 형태의 정의하고 할 수 있다. 예를 들어, 사랑을 정의할 때 다른 사람을 좋아하는 마음은 측정할 수 없기 때문에, 하루 한번 이상 포옹하는 것으로 정의하는 것이 조작적 정의의 예이다. 조작적 정의는 진단코드, 약물이나 시술 코드, 기관 코드, 치료 기간으로 정의할 수 있는데, 예를 들어, 염증성 장질환을 진단 코드인 K50(크론병) 또는 K51(궤양성 대장염)만으로 진단하는 것보다는 진단 코드뿐만 아니라 염증성 장질환의 약물 1가지 이상을 처방받고 2년 동안 1회 이상의 입원 또는 3회 이상의 외래 방문으로 정의를 하게 되면 염증성 장질환에 대한 진단 정확도가 훨씬 높아지게 된다. 그러나 너무 많은 추출 조건이 주어지게 되면 보다 많은 연구 대상자가 누락될 수 있기 때문에, 예상보다 너무 작은 수가 추출되었다면 조작적 정의의 오류를 점검해야 한다. 예를 들어, 해당 질환을 정의할 때 주진단만으로 정의하거나, 제1 부상병, 제2 부상병까지 정의하는 것에 따라 추출되는 환자 수가 달라질 수 있다. 연구 주제를 정한 후에는 맞춤형 또는 표본 샘플 자료 중 어떤 자료를 이용해야 할지 결정해야 하고, 연구를 잘 구현할 수 있는 연구 방법을 선택해야 한다. 청구 자료 연구는 청구 자료의 특성을 잘 파악해야 하는데, 청구 자료는 시간 순서를 알 수 없어서 동일 명세서 내에서의 검사, 처치, 약물 등의 시간적 순서를 파악할 수 없으며, 동일 입원도 분리하여 청구될 수 있어서 확인이 필요하다. 포괄수가제나 요양기관 정액제와 같이 의료 정책에 따라 파악하기 어려운 연구 주제는 가급적 피해야 한다. 게다가 급여 기준이나 급여 횟수의 변경에 따라 청구 자료도 달라질 수 있고, 연구 기간 동안에 청구 코드 자체가 변경되는 경우도 있기 때문에 보험 급여의 변화를 모니터해야 한다.

청구 자료를 이용하는 연구는 임상 연구 근거수준에서 레벨 4 또는 5의 관찰 연구로 대조군이 없는 단순 기술 연구와 대조군이 있는 환자-대조군 연구 및 코호트 연구로 분석할 수 있다. '단면조사 연구(cross-sectional study)'는 한 시점에서 한 집단의 질병 여부를 위험 요인 노출 여부에 대해 조사하는 방법으로 전후관계는 알 수 없지만 상관관계를 알 수 있는 기술 연구이다. 한 시점에서 조사하기 때문에 발생률은 구할 수 없지만 유병률을 구할 수 있으며, 무작위 대조 연구의 파일럿 연구에 적합한 연구 방법이다. '환자-대조군 연구(case-control study)'는 특정 질병의 유무로 환자군과 대조군을 선정하여 위험 인자에 대한 노출 여부를 조사하고, 두 군 간 노출 정도의 차이를 비교하는 연구 방법으로 오즈비를 이용하여 연관성을 기술한다. 환자군과 대조군 사이에 위험 인자 노출의 차이가 존재한다면, 그 요인이 질병 발생과 연관이 있다고 추론하게 된다. 이 방법은 질병 위험군 전체가 아니라 위험 집단을 대표하는 샘플만 연구하기 때문에 효율적이고 빠르게 결과를 도출할 수 있고, 질병과 관련된 한 개 이상의 위험 인자들을 조사할 수 있는 장점이 있다. 그러나 환자군과 대조군의 전체 모집단을 알 수 없으므로 위험도(risk)나 위험차(risk differences)를 측정할 수 없으며, 비교성 있는 대조군의 선정이 환자-대조군 연구의 성패를 결정하는 중요한 변수가 된다. 이 방법은 질병의 발생률이 낮은 질환 연구에 적합하며, 연관성, 선후관계를 알 수 있지만 인과성은 알 수 없고, 비뚤림과 교란변수가 많이 발생할 수 있다는 위험이 있다. '코호트 연구(cohort study)'는 서로 다른 노출 병력을 가진 집단에서 새로운 질병의 발생률을 비교할 수 있으며, 위험 인자를 갖고 있는 군이 그렇지 않은 군에 비해 질병의 발생률이 얼마나 더 높은지 비교위험도(relative risk)로 제시할 수 있는 연구 방법이다. 위험 인자에 노출된 경우, 노출되지 않은 경우보다 질병에 걸릴 확률이 몇 배 높다고 해석하게 된다. 코호트 연구는 질병 발생률이 높으면서 샘플이 큰 연구에 적합한데, 연관성, 선후관계, 인과성을 모두 확인할 수 있는 장점이 있다.

청구 자료 기반의 연구에서는 데이터를 정제하는 과정이 필수적이다. 같은 환자가 이중으로 청구되거나 심지어 입원이나 약제 에피소드도 중복 청구될 수 있어서 데이터를 정제해야 한다. 데이터를 정제하는 과정은 단순 반복 작업으로 지난한 과정인 경우가 많고 시간 소요가 많기 때문에 꼼꼼한 점검이 필요하다. 통계 분석에는 SAS나 R 통계 프로그램을 주로 사용하기 때문에 이에 대한 지식도 필요하며, 필요에 따라서는 임상 연구자와 의학통계 전문가의 협업도 고려할 수 있다. 논문 기술 과정은 일반적인 관찰연구를 작성하는 지침인 STROBE 지침과 유사한데, 청구 자료 연구는 특히 진단 알고리즘, 데이터 유형 및 연계, 환자 선택, 데이터 정리, 비뚤림 해결에 대해 자세히 기술하도록 제시하고 있는 reporting of studies conducted using observational routinely collected health data (RECORD) 지침을 준수하여 작성해야 한다.[9] 논문을 작성하는 과정에 있어서도 통계 방법론은 무작위 대조 연구보다 더 어려울 수 있기 때문에 다양한 통계 전문가의 도움을 받는 것도 좋고, 심사평가원 또는 건강보험공단에 대한 기존 문헌을 참고문헌으로 잘 제시해야 한다.

요약하면 심사평가원 또는 건강보험공단의 청구 자료를 맞춤형 자료나 표본 샘플 자료를 신청하여 청구 자료 기반의 연구를 시행할 수 있다. 하지만 이 연구는 데이터를 얻기는 쉽지만 분석하고 해석하는 것은 오히려 무작위 대조 연구보다 더 어려울 수 있다는 점을 기억해야 한다. 따라서 청구 자료의 장점을 살릴 수 있는 연구 주제를 선정하는 것이 가장 중요하며, 적합하지 않는 연구 주제임에도 유행처럼 남들을 따라하는 오류를

피해야 한다. 청구 자료의 불확실성 때문에 연구 자체를 기피하는 일부 연구자의 시각도 있지만, 전국민 유병률이나 발생률과 같이 청구 자료가 아니면 제시할 수 없는 소중한 데이터도 있기 때문에 청구 자료 기반의 연구는 지속될 것으로 생각한다.

CONFLICT OF INTEREST

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

ORCID

Su Bee Park https://orcid.org/0000-0002-4638-413X
Jae Myung Cha https://orcid.org/0000-0001-9403-230X

REFERENCES

1. Park A, Song J, Lee SB. Healthcare service analysis using big data. JKSCI 2020;25:149-156.
2. Shin YS. 30 years of Korean health insurance - its success, failure, and future directions. J Korean Med Assoc 2007;50:568-571.
3. Kim KJ, Kim KH. A study on the global budget payment system in Germany and Taiwan [Internet]. Research Institute for Healthcare Policy; 2011 [cited 2022 Feb 25]. Available from: https://rihp.re.kr/bbs/board.php?bo_table=research_report&wr_id=181&page=10.
4. 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;32:718-728.
5. Kim HK, Song SO, Noh J, Jeong IK, Lee BW. Data configuration and publication trends for the Korean national health insurance and health insurance review & assessment database. Diabetes Metab J 2020;44:671-678.
6. Seong SC, Kim YY, Park SK, et al. Cohort profile: the national health insurance service-national health screening cohort (NHIS-HEALS) in Korea. BMJ Open 2017;7:e016640.
7. Yen HH, Weng MT, Tung CC, et al. Epidemiological trend in inflammatory bowel disease in Taiwan from 2001 to 2015: a nationwide population based study. Intest Res 2019;17:54-62.
8. Kim JH, Chung HS, Kim HS, et al. Research using big data in gastroenterology-based on the outcomes from big data research group of the Korean Society of Gastroenterology. Korean J Gastroenterol 2020;75:4-10.
9. Benchimol EI, Smeeth L, Guttmann A, et al. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. PLoS Med 2015;12:e1001885.