Anesth Pain Med 2017; 12: 9-14
https://doi.org/10.17085/apm.2017.12.1.9

■종 설■
pISSN 1975-5171 • eISSN 2383-7977

의료 빅데이터를 활용한 임상 연구

고려대학교 의과대학 구로병원 산부인과학교실

조금준

Clinical research using medical big data

Geum Joon Cho

Department of Obstetrics and Gynecology, Korea University Guro Hospital, Korea University College of Medicine, Seoul, Korea

In Korea, various medical big data are available for use in clinical research. More data are expected to be released and used with the increasing social interest in big data. To initiate research using medical big data, it is important to understand the characteristics of data that are suitable for down-stream research. In this review article, we suggest possible research based on published research studies. **(Anesth Pain Med 2017; 12: 9-14)**

Key Words: Big data, Research, Medical.

서 론

사회 전반에 걸쳐 빅데이터에 대한 관심이 고조되고 있는 상황에서 의료분야, 특히 임상 연구에서 빅데이터를 활용하기 위한 다양한 노력이 있다[1]. 특히 정부가 빅데이터 활용을 위한 다양한 정책을 제시하면서 더욱 좋아지고 있다. 이에 실제 임상 연구에 활용할 수 있는 의료 빅데이터를 소개하고자 한다.

Received: December 23, 2016.
Revised: December 29, 2016
Accepted: December 30, 2016.

Corresponding author: Geum Joon Cho, M.D., Ph.D., Department of Obstetrics and Gynecology, Korea University Guro Hospital, Korea University College of Medicine, 148, Gurodong-ro, Guro-gu, Seoul 08308, Korea. Tel: 82-2-2626-3145, Fax: 82-2-838-1560, E-mail: md_cho@hanmail.net

의료 빅데이터의 종류

건강보험심사평가원 데이터(심평원 데이터)

공공기관 최초로 대용량 보건의료 빅데이터 개방시스템을 구축, 운영하고 있으며, 확보된 다양한 의료 빅데이터를 연구자에게 제공하고 있다(Appendix 1).

이용 절차는 산업체 및 연구자가 원하는 맞춤형 원시데이터 분석이 가능하도록 방문 또는 원격접속 지원을 하고 있다(Fig. 1). 또한 심평원 포털(http://opendata.hira.or.kr)에 접속하여 다양한 정보와 지원을 받을 수 있다.

국민건강보험공단 데이터(공단 데이터)

공단에서는 고유 업무를 통해 다양한 빅데이터를 보유하고 있다(Table 1). 최근에는 확보된 빅데이터를 바탕으로 표본코호트 데이터베이스를 구축하여 지원하고 있으며, 이를 통해 다양한 임상 연구가 진행되고 있다.

이용절차는 국민건강보험 포털(http://www.nhis.or.kr/), 표본코호트DB (http://nhiss.nhis.or.kr/)에 접속하여 자료를 신청하거나 이와 관련된 상담을 받을 수 있다.

이외 사용 가능한 보건의료 빅데이터 자료(Appendix 2)를 활용한다면 보다 많은 연구가 가능할 것이다.

의료 빅데이터를 활용한 임상 연구

의료 빅데이터를 활용한 연구를 시작하기 위해서는 먼저 데이터의 특성을 파악하고 이를 바탕으로 가능한 연구를 이해하는 것이 무엇보다 중요하다. 이에 의료 빅데이터를 활용하여 발표된 연구들을 바탕으로 실제 가능한 연구 주제를 소개하고자 한다.

질병의 발생 및 발생률 추이 분석

먼저 가장 쉽게 적용할 수 있는 연구 분야는 다양한 질병의 발생과 이를 바탕으로 질병의 발생 추이를 확인하는

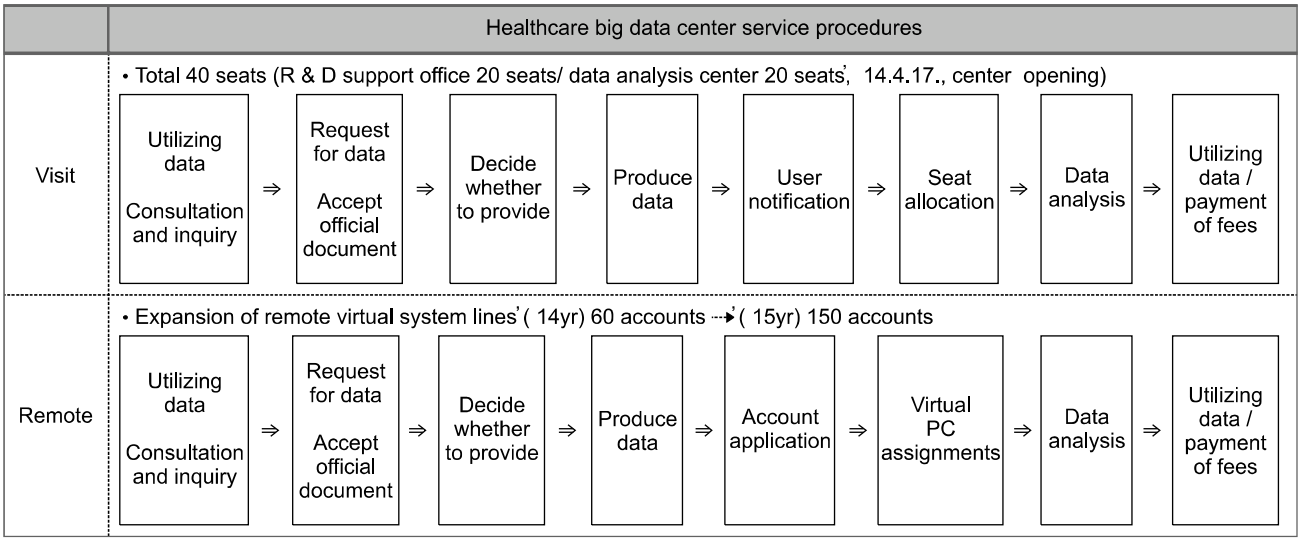

Fig. 1. Application procedure of the Health Insurance Review and Assessment service data[1].

Table 1. Composition of the NHIS Data

Medical utilization and evaluation resource		Resources for collecting insurance premiums		Resources for qualifications management	
Information	Agency	Information	Agency	Information	Agency
Billing Information	Medical Institutions	Personal income	NTS	Birth, death	HIRA
Claim review	HIRA	Property tax information	MOLIT	Business registration	NTS
Industrial disaster	KCOMWEL	Pension income	Pension institution	MB recipient registration	MW
Medical Institution	Medical Institutions	Business registration	NTS	Family relations	HIRA
Medical service evaluation	HIRA			National meritorious registration	MPVA
Examination	Medical Institutions			Immigration	MMA
Long-term care services	Long-term care services			Residence	HIRA
				Disability registration	MW

NTS: National Tax Service, HIRA: Health Insurance Review and Assessment service, MOLIT: Ministry Of Land, Infrastructure and Transport, KCOMWEL: Korea Workers' Compensation and Welfare Service, MB: Medical Benefits, MW: Ministry of Health and Welfare, MPVA: Ministry of Patriots and Veterans Affairs, MMA: Military Manpower Administration[2].

것이다. 한 연구에서는 임신성 당뇨병(gestational diabetes mellitus)이 2007년도에서 2010년까지 4배 가까이 증가함을 보고하였다[2]. 특히 의료 빅데이터를 통해서 단일 기관의 자료로 분석이 어려운 다양한 희귀 질병의 발생 및 발생률 변화를 확인할 수 있다. 한 연구에서는 중증근무력증(myasthenia gravis)의 발생을 확인하였으며, 2010년 10.42/100,000명에서 2014년 12.99/100,000명으로 증가하는 것을 보고하였다[3]. 또한 청구된 시점을 기준으로 질병 발생의 일별, 월별, 계절별 변화 분석도 가능하다. 예를 들면 한 연구에서는 가와사키병(Kawasaki disease)의 발생이 2007년 168.3/100,000명에서 2014년 217.2/100,000명으로 증가하였으며, 특히 초여름과 겨울에 그 발생이 증가함을 보고하였다[4]. 이외에도 심혈관 질환, 종양, 골다공증 및 골절, 화상 발생 등 다양한 연구가 가능하다[5-7].

하지만 이러한 의료 빅데이터는 실제 진료에 대한 청구 명세서를 바탕으로 하고 있으므로 청구 자료의 정확도에 영향을 받게 된다. 따라서 연구 시 가장 많은 노력을 기울여야 하는 부분은 진단의 정확도(validation)일 것이다. 대부분의 연구는 ICD-10 code를 통해 상병명을 확인하기 때문에

1) Health Insurance Review and Assessment service: 건강보험심사평가원.
2) NTS: 국세청, HIRA: 건강보험심사평가원, MOLIT: 지방정부/국토교통부, KCOMWEL: 근로복지공단, MB: 의료급여, MW: 보건복지부, MPVA: 국가보훈처, MMA: 병무청.

특정 질병을 가진 대상이 예상보다 많거나 또는 적게 산출되는 경우가 많다. 따라서 진단의 정확도를 높이기 위해서는 이를 뒷받침할 수 있는 추가적인 노력, 예를 들면 약물 사용, 수술 여부 등을 함께 확인함으로써 진단의 정확도를 높일 수 있을 것이다. 따라서 이러한 한계점을 연구의 시작 단계에서 고려하는 것이 가장 중요할 것이다.

질병의 위험 인자 분석

다음으로 가능한 연구 분야는 질병과 관련된 위험 인자를 분석하는 것이다. 특히 대규모의 데이터를 기반으로 하고 있으므로 기존 연구에서는 확인할 수 없었던 다양한 위험 인자의 분석이 가능하다. 위험 인자의 분석에는 몇 가지 접근 방법이 가능하다. 첫째는 질환을 진단 받은 시점에서 환자가 가지고 있는 위험 인자를 분석하는 것이다. 예를 들면 한 연구에서는 변이형 협심증(variant angina)로 인해 재입원 시 환자가 가지고 있는 위험 인자를 확인하였으며 [6], 다른 연구에서는 산후 출혈로 인한 자궁적출술(peripartum hysterectomy) 또는 자궁 동맥 색전술(uterine arterial embolization)시 환자가 가지고 있는 위험 인자를 분석하였다[8]. 다음 접근 방법은 관심 질병의 발생 시와 발생 이전의 의료 데이터를 연계함으로써 실제 다양한 질병 과거력 또는 약물 사용력 등이 관심 질병의 위험 인자로 작용하는지 분석하는 것이다. 예를 들면 한 연구에서는 초산부에 대해 다음 임신 자료와 연계를 한 후 두 번째 분만 시 전자간증(preeclampsia) 발생의 위험 인자로 첫 분만 시의 다양한 특성을 위험 인자로 분석하였다[9]. 또한, 임신 전 시행한 국가건강검진 결과와 임신 결과 데이터를 연계함으로써 임신 전 특성이 임신 합병증 발생의 위험 인자로 작용함을 보고하였다[10]. 마지막 접근 방법은 관심 질병의 발생 이후 데이터와 연계함으로써 해당 질병이 향후 다른 질병의 위험 인자로 작용하는지 분석하는 것이다. 예를 들면, 한 연구에서는 대퇴부 골절이 향후 혈전증(thromboemboizm)의 위험 인자로 작용함을 보고하였다[11]. 의료 빅데이터는 질병 발생 전후 데이터를 연계함으로써 일종의 질병 코호트를 만들 수 있다. 무엇보다도 전 국민을 대상으로 하는 우리나라의 의료 체계 특성상 중도 탈락하는 환자가 발생하지 않는다는 것이 이러한 의료 빅데이터가 갖는 중요한 장점이라 하겠다.

치료 및 처치에 대한 특성 분석

의료 데이터는 진료 및 처치에 대한 청구 내역을 바탕으로 형성된 데이터이다. 따라서 처치와 관련된 연구를 진행하기에 적합하다. 예를 들면 한 연구에서는 2006년부터 2010년까지 산후 출혈 시 자궁적출술의 시행 건수는 감소한 반면 자궁을 보전할 수 있는 자궁 동맥 색전술의 시행 건수는 상대적으로 증가하는 것을 보고하였다[8]. 이처럼 다양한 질환에 대한 치료, 특히 종양 및 희귀질환 등의 치료 패턴을 분석하는 것이 가능하다[12-14]. 또한 약물 사용에 대한 특성 분석이 가능하다. 한 연구에서는 임신성 당뇨병의 발생이 증가하고 있으며, 이와 함께 인슐린 사용 여부를 분석함으로써 인슐린이 필요한 A2 임신성 당뇨병의 경우 발생의 변화가 없는 것을 확인, 임신성 당뇨병의 증가는 A1 임신성 당뇨병의 증가에 의한 것임을 보고하였다[2]. 또 다른 연구에서는 임신 중 사용 금기 약물의 패턴, 항정신성 약물의 사용, 골다공증 치료제의 사용 패턴을 보고하였다[15-17]. 향후 마취 약제의 사용과 관련된 다양한 연구가 가능할 것으로 생각된다.

처치 또는 약제 사용에 관한 연구를 계획할 시 현재 의료 데이터는 급여항목에 관해서만 확인이 가능하다는 것을 고려해야 한다. 즉 비급여 치료, 처치 항목 또는 약물에 대해서는 자료를 확인할 수 없다. 따라서 연구 계획 단계에서 관심 있는 치료 및 처치가 비급여로 이루어지지 않는지 또는 비급여 항목이 포함되어 있어 정확한 분석에 영향을 주지는 않는지 세심히 살펴야 할 것이다.

질병과 관련된 비용 분석

의료 빅데이터는 본래 진료와 처치 비용을 청구하기 위해 생성되었기 때문에 이를 통해 다양한 비용 분석이 가능하다. 우선 질병에 관한 비용 분석이 가능하다. 한 연구에서는 5세아에서 뇌성마비의 유병률은 2.6/1,000명이며, 뇌성마비 경우 정상아에 비해 생애 의료 비용이 1.8배 증가하는 것을 보고하였다[18]. 다른 연구에서도 골다공증, HIV 감염에 의한 의료적, 사회적 비용을 분석, 보고하였다[19,20]. 또한 의료 데이터를 통해 시술과 관련된 비용 분석이 가능하다. 한 연구에서는 뇌동맥류(intracranial aneurysm)에 대한 치료를 비교하였는데, 시술 및 입원 기간 등을 고려하였을 때 neurosurgical coiling의 비용이 endovascular coiling 보다 비용이 적게 소요되는 것을 보고하였다[21]. 하지만 데이터의 특성상 급여 항목에 대한 비용 분석이 가능하다는 것을 기억해야 할 것이다.

의료 빅데이터의 한계점

의료 빅데이터는 아직 여러 한계를 가지고 있다. 따라서 이러한 한계점을 이해하는 것은 올바른 연구 계획을 수립하고 결과를 해석하며 나아가 의료 빅데이터의 발전 방향을 논의하는 데 필요할 것이다.

의료 빅데이터의 이해 어려움

우선 의료 빅데이터는 진료 시 발생하는 비용에 대한 청구 데이터이다. 즉, 임상 연구를 위해 개발된 데이터가 아니다. 따라서 처음 접하는 연구자는 그 데이터의 분류 및

특성을 이해하기에는 한계가 있으며, 이로 인해 막상 연구를 시작하였을 때 어려움에 봉착하게 된다. 특히 청구 코드의 특성을 이해하고 있어야 하며 현재까지는 주로 SAS 통계 프로그램을 통해서만 분석할 수 있기 때문에 데이터를 이해하고 분석하기 위해서는 이를 분석해줄 수 있는 공동연구자의 도움이 절실하다. 반면 최근 심평원과 공단에서는 임상 연구를 위한 데이터 분석 매뉴얼과 분석 지원 프로그램을 개발 중에 있다. 또한 이미 발표된 다양한 연구의 방법론을 검토하는 것이 연구의 계획 단계에서 도움이 될 것이다.

지속적인 자료의 정도 관리 필요

다시 한번 강조하지만 의료 빅데이터는 임상 연구를 위해 수집된 데이터가 아니다. 따라서 단순히 ICD-10 코드의 상병명을 이용한 분석은 예상과 다른 결과가 도출될 수 있다. 경우에 따라 상병코드를 갖는 환자가 너무 많이 검색되거나 또는 너무 적게 검색될 수 있다. 이는 실제 진료에서 부여되는 진단명과 청구되는 진단명 사이에서 오는 다양한 문제, 또는 의료 체계의 특수성에 기인할 것이다. 이런 이유로 질병의 중등도가 높을수록, 예를 들면 단순 감기에 대한 연구보다는 암에 관한 연구를 시행하였을 때 데이터의 신뢰도가 높아지게 되어 연구에 적합할 것이다. 즉, 의료 빅데이터를 통한 임상 연구를 진행할 때에는 도출되는 결과에 대해 항상 의문점을 갖고 접근해야 할 것이다. 특히 실제 처치, 예를 들면 특정 암의 발생을 본다면 단순히 ICD-10 코드로 검색한 자료만을 사용하기보다는 암과 관련된 수술 또는 항암제 등의 처치 사용 여부를 확인하여 진단을 국한하는 것이 더욱 질병의 신뢰도를 높이는 방법이 될 것이다.

자료 자체의 한계점

의료 빅데이터는 진료에 대한 청구 데이터이나 비급여 항목은 알 수 없다는 한계가 있다. 이로 인해 중요 약제 사용이나 시술에 대한 정보가 포함되어 있지 않을 수 있다. 또한, 성별, 나이 등은 확인할 수 있지만, 이외의 기본적인 정보, 예를 들면, 나이, 키, 혈압 등의 신체 정보나 음주력, 흡연력, 운동 여부 등의 사회, 경제적 정보가 포함되어 있지 않다는 것도 중요한 한계점이다. 이는 결과 분석 시 다양한 위험 인자를 바로잡지 못하게 되어 연구 결과의 정확도 및 신뢰도를 낮추는 요인으로 작용하게 된다. 따라서 이러한 데이터의 특성을 고려하여 적절한 연구 주제를 선정하고 결과를 설계하는 것이 필요하다.

의료 빅데이터의 발전 방향

향후 의료 빅데이터의 발전 방향을 이해하는 것은 더욱 나은 연구를 위해 중요할 것이다. 이에 저자가 생각하는 발전 방향에 관해 기술하고자 한다.

연구자 지원 정책 증가

가장 큰 의료 빅데이터 보유 기관인 심평원과 보험공단에서는 임상 연구자의 데이터 활용도를 높이기 위한 다양한 노력을 하고 있다. 따라서 연구자들은 이러한 정책 지원을 활용할 필요가 있다. 예를 들면 과거 직접 방문을 통해 분석 하던 것이 원격 접속을 통해 본인 연구실에서 분석할 수 있으며, 제공되는 다양한 표본 데이터를 통해 연구의 가능성을 사전에 확인해 볼 수 있다. 그리고 그러한 지원은 더욱더 확대될 것이다.

빅데이터의 연계 및 통합

향후 산재하여 있는 다양한 기관의 빅데이터는 서로 연계가 될 것이며 이를 통해 현재 빅데이터의 여러 문제점이 상당 부분 해결될 것으로 보인다. 예를 들면 국민건강영양조사 자료와 심평원 또는 공단 자료와 연계된다면 국민건강영양조사 결과를 바탕으로 향후 다양한 질병 발생 유무를 심평원 또는 공단 자료를 통해 확인할 수 있게 되어 일종의 코호트와 유사한 데이터가 만들어지게 될 것이다. 유사하게 통계청, 질병관리본부의 다양한 데이터가 연계되고 이를 통해 다양한 연구가 가능하게 될 것이다. 또한, 정부에서는 산재해 있는 의료 데이터를 한 기관에서 통합, 관리한다면 연구자들이 더욱 쉽고 효율적으로 데이터에 접근하여 연구를 진행할 수 있을 것이다.

병원 데이터와 의료 빅데이터의 연계

개별 병원의 자료와 심평원 또는 공단 자료가 연계되게 될 것이다. 이렇게 되면 병원 간의 불필요한 검사 등을 생략할 수 있으므로 의료비 절감의 효과가 있을 것이다. 또한, 실제 본인이 진료하는 관심 환자들에 대한 다양한 질병 발생에 대한 추적 관찰이 가능하게 될 것이다. 궁극적으로는 병원 간 데이터가 연계되어 국가적 차원의 의료 빅데이터가 완성될 것으로 보인다.

맺음말

우리나라에는 다양한 의료 빅데이터가 존재하며 임상 연구를 위해 활용이 가능하다. 특히 빅데이터에 대한 사회적 관심이 고조됨에 따라 더욱 많은 데이터가 공개되고 사용될 수 있을 것으로 보인다. 하지만 보다 효과적인 연구를 위해서는 여러 데이터가 연계되어야 할 것이며 이를 위해 기관 간의 효과적인 협력 및 개인정보 보호의 문제 등이 해결되어야 할 것이다. 또한, 실제 의료 데이터를 생성하고 연구를 진행하는 연구자들의 지속적인 관심이 중요하다 하

겠다.

REFERENCES

1. Jhun HJ, Ahn K, Lee SC. Estimated number of self-reported intervertebral disc disorders in Korean adults based on the data from the third Korea national health and nutrition examination survey. Anesth Pain Med 2009; 4: 208-213.
2. Cho GJ, Kim LY, Sung YN, Kim JA, Hwang SY, Hong HR, et al. Secular trends of gestational diabetes mellitus and changes in its risk factors. PLoS One 2015; 10: e0136017.
3. Lee HS, Lee HS, Shin HY, Choi YC, Kim SM. The epidemiology of myasthenia gravis in Korea. Yonsei Med J 2016; 57: 419-25.
4. Ha S, Seo GH, Kim KY, Kim DS. Epidemiologic study on kawasaki disease in Korea, 2007-2014: based on health insurance review & assessment service claims. J Korean Med Sci 2016; 31: 1445-9.
5. Lim Y, Lee JO, Bang SM. Incidence, survival and prevalence statistics of classical myeloproliferative neoplasm in Korea. J Korean Med Sci 2016; 31: 1579-85.
6. Kim HL, Lee SH, Kim J, Kim HJ, Lim WH, Seo JB, et al. Incidence and risk factors associated with hospitalization for variant angina in Korea. Medicine (Baltimore) 2016; 95: e3237.
7. Park SB, Kim J, Jeong JH, Lee JK, Chin DK, Chung CK, et al. Prevalence and incidence of osteoporosis and osteoporotic vertebral fracture in Korea: nationwide epidemiological study focusing on differences in socioeconomic status. Spine (Phila Pa 1976) 2016; 41: 328-36.
8. Cho GJ, Kim LY, Hong HR, Lee CE, Hong SC, Oh MJ, et al. Trends in the rates of peripartum hysterectomy and uterine artery embolization. PLoS One 2013; 8: e60512.
9. Cho GJ, Kim LY, Min KJ, Sung YN, Hong SC, Oh MJ, et al. Prior cesarean section is associated with increased preeclampsia risk in a subsequent pregnancy. BMC Pregnancy Childbirth 2015; 15: 24.
10. Cho GJ, Park JH, Shin SA, Oh MJ, Seo HS. Metabolic syndrome in the non-pregnant state is associated with the development of preeclampsia. Int J Cardiol 2016; 203: 982-6.
11. Lee S, Hwang JI, Kim Y, Yoon PW, Ahn J, Yoo JJ. Venous thromboembolism following hip and knee replacement arthroplasty in Korea: a nationwide study based on claims registry. J Korean Med Sci 2016; 31: 80-8.
12. Kang S, Kim HS, Choi ES, 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 Res Treat 2015; 47: 575-82.
13. Lee KW, Lee JH, Kim JW, Kim JW, Ahn S, Kim JH. Population-based outcomes research on treatment patterns and impact of chemotherapy in older patients with metastatic gastric cancer. J Cancer Res Clin Oncol 2016; 142: 687-97.
14. Kang JK, Kim MS, Jang WI, Kim HJ, Cho CK, Yoo HJ, et al. The clinical status of radiation therapy in Korea in 2009 and 2013. Cancer Res Treat 2016; 48: 892-8.
15. Song I, Choi SH, Shin JY. Trends in prescription of pregnancy-contraindicated drugs in Korea, 2007-2011. Regul Toxicol Pharmacol 2016; 75: 35-45.
16. Park SC, Lee MS, Kang SG, Lee SH. Patterns of antipsychotic prescription to patients with schizophrenia in Korea: results from the health insurance review & assessment service-national patient sample. J Korean Med Sci 2014; 29: 719-28.
17. Lee YK, Ha YC, Park C, Yoo JJ, Shin CS, Koo KH. Bisphosphonate use and increased incidence of subtrochanteric fracture in South Korea: results from the National Claim Registry. Osteoporos Int 2013; 24: 707-11.
18. Park MS, Kim SJ, Chung CY, Kwon DG, Choi IH, Lee KM. Prevalence and lifetime healthcare cost of cerebral palsy in South Korea. Health Policy 2011; 100: 234-8.
19. Choi HJ, Shin CS, Ha YC, Jang S, Jang S, Park C, et al. Burden of osteoporosis in adults in Korea: a national health insurance database study. J Bone Miner Metab 2012; 30: 54-8.
20. Yang BM, Kim DJ, Byun KS, Kim HS, Park JW, Shin S. The societal burden of HBV-related disease: South Korea. Dig Dis Sci 2010; 55: 784-93.
21. Chang HW, Shin SH, Suh SH, Kim BS, Rho MH. Cost-effectiveness analysis of endovascular coiling versus neurosurgical clipping for intracranial aneurysms in Republic of Korea. Neurointervention 2016; 11: 86-91.