

내과에서의 빅데이터: 자료원과 도전 과제

¹경북대학교 의과대학 내과학교실, ²경북대학교병원 소화기내과

장세영^{1,2}

Big Data in Internal Medicine: Sources and Challenges

Se Young Jang^{1,2}

¹Department of Internal Medicine, School of Medicine, Kyungpook National University, Daegu;

²Division of Gastroenterology and Hepatology, Department of Internal Medicine, Kyungpook National University Hospital, Daegu, Korea

Big data has become an important research resource in internal medicine. Major data sources in Korea include national claims databases, health screening records, registry data, mortality statistics, survey data, hospital electronic health records, and emerging biobig data resources. These datasets provide new opportunities for disease surveillance, outcome prediction, and evaluation of real-world practice across a wide range of chronic diseases. Nevertheless, large-scale data do not guarantee valid evidence. Diagnostic misclassification, limited clinical details, incomplete medication information, residual confounding, and difficulties in data linkage and standardization remain major challenges. Therefore, the value of big data in internal medicine depends not only on data volume but also on data quality, appropriate linkage, validated definitions, and rigorous study design. Big data should be recognized as a complementary tool that can strengthen, but not replace, clinical reasoning and conventional clinical research. (**Korean J Med 2026;101:165-168**)

Key Words: Big data; Internal medicine; Insurance, health; Routinely collected health data; Observational study

서론

보건의료 빅데이터는 건강, 질병, 진료, 생활습관, 환경 요인 등 인간의 건강과 관련된 정보를 대규모로 수집, 저장, 분석한 자료를 의미한다. 고령화와 만성 질환의 증가, 장기 추적 관찰의 필요성, 실제 진료 현장을 반영한 근거에 대한 요구가 커지면서 빅데이터는 내과 연구의 중요한 자원으로 자리 잡고 있다. 특히 내과학은 당뇨병, 심혈관질환, 만성 콩팥병, 만성 간질환, 만성 호흡기질환, 감염질환, 암과 같이 유병률이 높고 장기 경과를 보이는 질환을 주로 다루기 때문에 넓은 모집단과

장기간 추적이 가능한 자료의 가치가 더욱 크다. 이러한 점에서 빅데이터는 단일 기관 연구나 전통적인 임상시험만으로는 충분히 확인하기 어려운 질병 부담, 진료 패턴, 예측, 의료 이용의 변화, 정책 효과를 분석하는 데 중요한 역할을 한다.

본론

국내에서 가장 널리 활용되는 보건의료 빅데이터는 국민건강보험공단과 건강보험심사평가원의 청구 자료이다[1,2]. 이들 자료는 전국민 단위의 의료 이용과 처방 정보를 포괄하고 있어

Received: 2026. 3. 30 **Revised:** 2026. 6. 29 **Accepted:** 2026. 7. 6

Correspondence to:

Se Young Jang, M.D., Ph.D.

Division of Gastroenterology and Hepatology, Department of Internal Medicine, Kyungpook National University Hospital, School of Medicine, Kyungpook National University, 130 Dongdeok-ro, Jung-gu, Daegu 41944, Korea

Tel: +82-53-200-5505, Fax: +82-53-426-2046, E-mail: vocjsy@knu.ac.kr

Copyright©2026 The Korean Association of Internal Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

실제 진료 기반의 연구에 강점을 가진다. 국민건강보험공단 자료는 건강 검진 자료, 자격 자료와 연계할 수 있다는 점에서 장기 추적 연구와 예방의학적 연구에 유리하다[1]. 반면 건강보험심사평가원 자료는 심사 과정에서 생성되는 보다 상세한 청구 정보를 활용할 수 있어 의료 이용 양상과 진료 행태를 세밀하게 파악하는 데 장점이 있다[2]. 즉 국민건강보험공단 자료가 자격, 건강 검진, 사망 자료와의 연계를 통한 장기 추적과 인구집단 기반 연구에 강점이 있다면 건강보험심사평가원 자료는 청구 및 심사 과정에서 생성되는 상세한 진료 및 의료 이용 정보를 분석하는 데 강점이 있다. 여기에 국가암등록자료, 사망 자료, 국민건강영양조사 자료[3]가 더해지면 질병 발생과 사망뿐 아니라 건강 행태, 영양, 신체 계측, 일부 검사 결과까지 함께 분석할 수 있어 연구 범위가 한층 넓어진다. 최근에는 공공 자료 연계를 지원하는 플랫폼이 운영되고 있으며[4], 유전체 자료 역시 국가 단위 구축사업을 통해 축적이 진행되고 있어 향후 정밀의료 연구 기반으로 활용될 가능성이 있다[5].

이러한 자료원은 내과학 연구에 여러 장점을 제공한다. 첫째, 대규모 모집단을 대상으로 희귀 사건이나 장기 예후를 평가할 수 있다. 예를 들어 심혈관 사건, 암 발생, 간경변 합병증, 말기 콩팥병과 같이 임상적으로 중요한 결과는 단일 기관 자료만으로 충분한 통계적 검정력을 확보하기 어려운 경우가 많다. 실제로 국내 청구 및 등록 자료를 활용한 연구들은 이러한 장점을 잘 보여준다. 심혈관 분야에서는 국민건강보험공단 건강 검진 자료를 이용하여 혈압, 혈당, 콜레스테롤, 체질량지수의 방문 간 변동성이 사망, 심근경색, 뇌졸중과 관련됨을 분석한 전국민 기반 연구가 보고되었으며[6], 신장 분야에서는 국내 말기 콩팥병의 발생률, 유병률, 사망률의 장기 추세를 평가한 연구가 수행되었다[7]. 간질환 분야에서는 국민건강보험공단 자료를 이용하여 국내 간세포암종 발생률의 장기 변화와 고령층에서의 증가 양상을 분석한 연구가 보고되었고[8], 종양 분야에서는 국민건강보험 자료와 중앙암등록자료를 비교하여 청구 자료 기반 암 진단 조작적 정의의 타당성을 평가한 연구가 제시되었다[9]. 이러한 연구들은 국내 보건의료 빅데이터가 내과 주요 질환의 발생, 예후, 사망 및 장기 합병증을 평가하는 데 실제로 활용될 수 있음을 보여준다. 둘째, 실제 임상 현장을 반영한 근거(real-world evidence)를 생산할 수 있다. 임

상시험은 엄격한 선택 기준 때문에 고령 환자, 다질환 환자, 취약 계층을 충분히 반영하지 못하는 경우가 많지만 빅데이터는 이러한 환자군을 포함한 실제 의료 현장을 보다 폭넓게 보여준다. 셋째, 시간에 따른 질병 추이와 치료 변화, 정책 개입의 효과를 분석하는 데 유용하다. 건강검진 제도의 확대, 약제 급여 기준의 변화, 국가 암 검진 정책, 만성 질환 관리 사업 등은 빅데이터를 통해 평가할 수 있는 대표적인 주제이다.

그러나 빅데이터의 규모가 곧 연구의 타당성을 보장하는 것은 아니다. 가장 먼저 고려해야 할 문제는 진단 정확도이다. 많은 연구가 국제질병분류 코드에 기반하여 환자를 정의하지만 청구 자료의 진단코드는 임상적 진단 자체보다 보험 청구와 행정적 필요에 영향을 받기 때문에 실제 질환 상태와 완전히 일치하지 않을 수 있다. 같은 질환이라도 기관마다 코딩 관행이 다를 수 있고 확진 전 의심 단계의 코드가 포함되거나 반대로 실제 환자가 누락될 수도 있다. 따라서 진단코드만으로 질환을 정의할 경우 오분류 가능성을 항상 염두에 두어야 한다.

두 번째 한계는 임상적 세밀성의 부족이다. 청구 자료는 의료 이용과 청구 명세를 중심으로 구성되기 때문에 영상 검사 결과, 병리 소견, 반복 측정된 검사 수치, 증상의 정도, 질병의 중증도와 같은 중요한 임상 정보가 충분히 담기지 않는 경우가 많다. 건강 검진 자료나 설문조사 자료가 이를 일부 보완할 수 있지만 측정 시점이 제한적이고 반복 추적이 충분하지 않다는 한계가 있다. 또한 비급여 검사와 비급여 약제, 실제 복약 여부, 환자의 순응도, 생활습관 변화와 같은 요소는 자료만으로 정확히 파악하기 어렵다. 내과 질환의 예후는 이러한 요소들에 크게 영향을 받기 때문에 자료에 보이지 않는 임상적 맥락을 어떻게 해석할 것인지가 중요하다.

세 번째 문제는 관찰 연구 고유의 방법론적 한계이다. 대부분의 빅데이터 연구는 후향적 관찰 연구이며 잔여 교란, 선택 편향, 역인과성의 영향을 완전히 배제하기 어렵다. 통계적으로 성향 점수 매칭이나 가중치 분석을 적용하더라도 측정되지 않은 교란변수는 여전히 남아 있을 수 있다. 내과 영역의 주요 질환은 흡연, 음주, 비만, 사회경제적 수준, 동반 질환과 복잡하게 얽혀 있기 때문에 단순한 연관성 결과를 인과관계처럼 해석하는 데에는 신중함이 필요하다. 최근에는 이러한 한계를 보완하기 위하여 표적 임상시험 모사(target trial emulation)

와 같이 관찰 자료를 보다 엄밀한 인과 추론의 틀에서 해석하려는 방법론도 제시되고 있으며[10], 실제 적용 연구의 설계 및 분석 관행을 체계적으로 검토하려는 시도와 실제 진료 근거 연구의 계획 및 보고를 표준화하려는 논의도 지속적으로 이루어지고 있다[11,12].

네 번째로는 자료 연계와 표준화의 문제를 들 수 있다. 공공 자료와 병원 자료, 등록 자료를 연계하면 연구의 깊이는 커지지만 기관마다 코드 체계와 변수 정의, 자료 구조가 달라 동일한 연구 질문이라도 결과가 달라질 수 있다. 따라서 공통 데이터 모델(common data model)과 표준 용어 체계의 확산이 중요하며 환자 정의 역시 진단코드 단독이 아니라 처방, 시술, 검사, 등록 자료를 함께 활용하는 방식으로 정교화할 필요가 있다[13]. 또한 국내 공공 청구 자료는 개인정보 보호를 위하여 가명 처리와 통제된 활용 절차 아래 운영되고 있다. 이러한 체계는 연구의 신뢰성을 높이는 데 중요하지만 한편으로는 자료 결합과 활용의 속도 및 범위를 제한하는 현실적인 요인이 되기도 한다.

결국 빅데이터 연구의 발전을 위해서는 단순히 더 많은 자료를 축적하는 것만으로는 충분하지 않다. 서로 다른 자료원을 적절히 연계하고 연구 질문에 맞는 환자 정의를 검증하며 자료의 한계를 이해한 상태에서 엄밀한 연구 설계를 적용해야 한다. 또한 통계적 유의성만이 아니라 임상적 의미를 함께 해석하려는 태도가 필요하다. 앞으로는 자료 연계를 통해 보다 풍부한 임상 정보를 확보하고, 이를 바탕으로 질병의 발생, 경과, 치료 반응을 더 정확하게 이해하려는 노력이 중요해질 것이다.

결론

보건의료 빅데이터는 내과 연구에서 이미 중요한 자료원이 되었으며 앞으로는 실제 진료 현장의 복잡성을 반영하는 근거를 생산하는 데 더 큰 역할을 할 것이다. 그러나 빅데이터 연구의 핵심은 자료의 크기 자체가 아니라 임상적으로 의미 있는 질문을 설정하고 그 질문에 맞는 자료원과 환자 정의, 연구 설계를 선택하는 데 있다. 따라서 임상적 빅데이터 연구 결과를 해석할 때 진단코드의 정확성, 측정되지 않은 교란, 자료에 포함되지 않은 임상 정보의 한계를 함께 고려해야 한다. 동시

에 연구자로서는 청구 자료, 등록 자료, 검진 자료, 병원 자료를 적절히 연계하고 가능한 경우 조작적 정의를 검증함으로써 더 신뢰도 높은 근거를 만들어야 한다. 결국 내과 영역에서 빅데이터의 가치는 단순히 많은 환자를 분석하는 데 있는 것이 아니라 실제 환자 진료와 정책 결정에 도움이 되는 임상적으로 해석 가능한 근거를 생산하는 데 있다.

중심 단어: 빅데이터; 내과학; 건강보험; 일상적으로 수집되는 보건의료 자료; 관찰 연구

CONFLICTS OF INTEREST

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

FUNDING

None.

AUTHOR CONTRIBUTIONS

Se Young Jang conceived and designed the manuscript, reviewed the literature, drafted and critically revised the manuscript, approved the final version, and agrees to be accountable for all aspects of the work.

ACKNOWLEDGEMENTS

None.

REFERENCES

1. National Health Insurance Service. National Health Insurance Service [Internet]. Wonju (KR): National Health Insurance Service, c2018 [cited 2026 Mar 27]. Available from: <https://nhiss.nhis.or.kr>.
2. Health Insurance Review & Assessment Service. HIRA healthcare big data hub [Internet]. Wonju (KR): Health Insurance Review & Assessment Service, c2023 [cited 2026 Mar 27]. Available from: <https://opendata.hira.or.kr/home.do>.

3. Korea Disease Control and Prevention Agency. Korea National Health and Nutrition Examination Survey [Internet]. Cheongju (KR): Korea Disease Control and Prevention Agency, c2025 [cited 2026 Mar 27]. Available from: <https://www.kdca.go.kr/eng/4428/subview.do>.
4. Korea Health Information Service. Health information big data integration platform [Internet]. Sejong (KR): Korea Health Information Service, c2019 [cited 2026 Mar 27]. Available from: <https://hcdl.mohw.go.kr/>.
5. Republic of Korea National Bio Bigdata Bureau. National Integrated Bio Big Data Project [Internet]. Seoul (KR): BioBigData.Korea, c2024 [cited 2026 Mar 27]. Available from: <https://www.biobigdata.kr/>.
6. Kim MK, Han K, Park YM, et al. Associations of variability in blood pressure, glucose and cholesterol concentrations, and body mass index with mortality and cardiovascular outcomes in the general population. *Circulation* 2018;138:2627-2637.
7. Lee MJ, Ha KH, Kim DJ, Park I. Trends in the incidence, prevalence, and mortality of end-stage kidney disease in South Korea. *Diabetes Metab J* 2020;44:933-937.
8. Chon YE, Park SY, Hong HP, et al. Hepatocellular carcinoma incidence is decreasing in Korea but increasing in the very elderly. *Clin Mol Hepatol* 2023;29:120-134.
9. Yang MS, Park M, Back JH, et al. Validation of cancer diagnosis based on the National Health Insurance Service Database versus the National Cancer Registry Database in Korea. *Cancer Res Treat* 2022;54:352-361.
10. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016;183:758-764.
11. Bigirimurame T, Hiu SKW, Teare MD, Wason JMS, Bryant A, Breckons M. Current practices in studies applying the target trial emulation framework: a protocol for a systematic review. *BMJ Open* 2023;13:e070963.
12. Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ* 2021;372:m4856.
13. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015;216:574-578.