빅데이터 활용 임상연구 시 유의사항

김현창

연세대학교 의과대학 예방의학교실

Hyeon Chang Kim

Department of Preventive Medicine, Yonsei University College of Medicine, Seoul, Korea

빅데이터(big data)란 기존 데이터베이스 관리도구로 처리할 수 있는 역량을 넘어서는 대량의 정형 및 비정형 데이터 집합을 말하거나, 또는 이러한 데이터로부터 유용한 정보를 추출·분석· 해석하는 기술을 말한다. 2012년 세계 경제 포럼이 10대 유망 기술 중 첫 번째로 빅데이터 기술을 선정하고, 우리나라 정부도 정보통신(IT) 10대 핵심기술 가운데 하나로 빅데이터를 선정한 바 있다. 의학계에서 빅데이터에 큰 관심을 보이는 이유는 보건의료분야는 자체적으로 방대한 양의 데이터를 만들어내는 빅데이터의 원천 중에 하나이며, 불확실성(uncertainty)을 줄이기 위해 항상 더 많은 정보를 원하는 특성이 있기 때문이다. 우리나라 보건의료분야 빅데이터 소스로는 의료기관의 의무기록(health record), 영상정보, 유전정보, 역학정보, 건강검진 데이터, 건강보험 청구 및 심사 자료, 각종 질환 레지스트리, 사망원인 통계정보 등을 꼽을 수 있다. 특히 최근 건강보험공단과 건강보험심사평가원에서 보험 청구와 심사용으로 수집된 정보를 공개하면서 건강보험 빅데이터를 이용한 임상연구가 증가하고 있다. 동시에 건강보험 빅데이터의 효용과 한계에 대한 논의와 이 데이터를 적절히 분석하고 해석하고 있는지에 대한 논란도 증가하고 있다. 이에 맞추어 건강보험 빅데이터를 이용하여 임상연구를 할 때 주의사항 및 극복해야 할 문제들을 다루고자 한다.

가장 먼저 꼭 건강보험 빅데이터가 필요한 연구인지 자문해 보아야 한다. 건강보험 빅데이터의 최대 장점은 대표성이 높고 샘플사이즈가 크다는 것이다. 그러나 연구 목적으로 수집하지 않은 이차자료의 단점도 많이 있다. 샘플사이즈가 큰 것보다는 측정의 표준화와 정밀성이 더 중요한 연구라면 굳이 전국민의 건강보험자료가 필요치 않을 것이다. 다음으로 연구하려는 질환이 건강보험청구 자료 연구에 적합한 특성을 가지고 있는지 확인해 보아야 한다. 간암, 간경변처럼 발병한 환자의 대부분이 의료이용을 하는 질환은 건강보험자료로 연구하기에 적합하지만, 비알코올성지방간처럼 초기에 증상이 없어서 진단율이 낮은 질환은 질병 유병상태보다는 의료이용 행태의 영향을 많이 받기 때문에 건강보험자료로 연구하기에는 제한점이 많다. 또한, 진단율이 높은 질환이라도 객관적인 진단검사가 없고 진단이 표준화되어 있지 않아서 병원간, 의사간 진단기준에 차이가 큰 경우라면 건강보험자료 이용 연구에 적합하지 않다.

건강보험 빅데이터를 이용한 연구 중 임상의학분야에서 큰 관심을 가지고 있는 것이 특정 치료법(특히 약

물치료)의 유효성과 안전성을 비교 평가하는 것이다. 이러한 연구에서는 다양한 바이어스의 개입 가능성이 있는지를 검토하고, 이를 극복할 수 있는 방안을 찾아야 한다. 예를 들어, 최근 국내 건강보험공단의 청구자료를 분석한 연구보고서에서 대표적 심혈관질환 예방을 위해 고지혈증 치료제(스타틴계 약물)를 투여 받은 군이 그렇지 않은 군에 비하여 당뇨병 발생 위험도가 높다고 보고하여 큰 관심을 받은 적이 있다. 이런 연구에서 유의할 점이 indication bias의 가능성이다. 만약 의사들이 애초에 당뇨병 발생 위험도가 높은 사람들에게 스타틴 처방을 더 많이 하였다면 연구결과가 왜곡될 수 있다. Indication bias를 극복하려면 환자 개인의 당뇨병 위험도에 대한 정보를 파악하여 보정해주어야 하는데 건강보험 빅데이터에 그 정보가 부족하다면 바이어스를 극복하기 어렵게 된다. 또 다른 종류의 바이어스인 differential outcome ascertainment (혹은 differential outcome misclassification)의 가능성도 있다. 만약 스타틴을 복용하는 사람들이 복용하지 않는 사람들에 비하여 더 병원을 자주 방문하고 혈액검사를 자주하게 된다면 당뇨병이 진단될 확률이 높아지게 된다. 특정 치료를 받은 환자군과 비교군이 질병 발생 확인에 차이가 생겨서 연구결과가 왜곡될 가능성이 있다. 비교대상 군간에 질병 발생 확인의 방법이나 빈도에 차이가 없음을 입증하거나, 차이가 있다면 그 차이를 극복할 수 있는 대안이 필요하다.

마지막으로 데이터의 분석에 앞서 데이터베이스에 대한 이해와 기초분석에 충분한 시간을 할애할 것을 강조한다. 연구자가 프로토콜에 따라 직접 수집한 데이터도 아니고, 처음부터 연구를 목적으로 수집한 데이터도 아니므로, 데이터를 이해하는 데 많은 시간과 노력을 들여야 원하는 결과를 얻을 수 있다. 복잡한 데이터베이스의 구조도 파악하여야 하며, 데이터베이스에 포함된 개별 변수들이 어떻게 수집되어 코딩 되었는지를 파악하여야 한다. 데이터베이스에 포함된 모든 변수를 완벽히 이해하는 것은 불가능하다. 그러므로 연구에 필요한 핵심 변수들을 골라서 그 특성을 파악하여야 한다. 연구가설에 따라서 주요 원인(독립)변수와 결과(종속)변수, 그리고 혼란변수일 가능성이 있는 변수들을 나열하고, 이 핵심 변수들을 대상으로 충분한 기술통계분석과 변수간 상호 관련성을 파악이 선행되어야 한다. 기술통계분석과 변수간 관련성 분석을 수행할 때 통계수치만 계산하는 것보다는 다양한 그래프를 그려서 변수의 분포와 변수간 관련성을 시각적으로 파악하는 것이 큰 도움이 된다.