ICDC WEEKLY 2018년 33주차·통권 제25호 | 전문가칼럼

감염병과 빅데이터?

이우주 부교수
인하대학교 통계학과 /
의학통계 및 역학 연구

"빅데이터"는 더 이상 통계학과 컴퓨터 공학의 전유물이 아니다.

의학 연구 분야에서도 "빅데이터"라는 용어는 자주 사용되고 있다. 한 예로 2017년 G20 정상회의 공동선언문에서 빅데이터를 활용한 감염병 확산 방지 프로젝트가 언급되었고 실제 규모가 큰 정부 사업으로 진행되고 있다는 사실은 주목할 만하다. 또한, 한 동안 정부기관과 건강보험심사평가원의 협력을 통해 빅데이터 기반 감염병 조기 파악 시스템 구축과 관련한 기사를 다양한 매체를 통해 접할 수 있었다. 한편, 선도적으로 빅데이터 활용의 기치를 올린 의학 분야의 여러 응용 사례에서 기대만큼 성과가 나오지 못하자 여러 전문가들은 빅데이터의 유용성에 대해 쓴 소리를 뱉어내고 있다. 이러한 상황에서 **감염병 탐지 문제에 초점**을 맞추어 우리가 "빅데이터"에 기대해볼 것이 남아있는지 한 번 간단히 점검해보고자 한다.

실시간 감염병 탐지와 관련지어 **빅데이터의 가능성**을 보여준 가장 인상적인 사례는 **2008년 구글의 플루 트렌드(Google Flu Trends)**였다. 그러나 2009년부터 플루 트렌드의 예측이 실제와는 너무 달라서 많은 사람들은 이를 "빅데이터의 역대급 실패" 사례로 언급해왔다. 이유는 분명하지 않으나 2015년부터 구글 플루 트렌드는 더 이상 예측치를 제공하지 않고, 관련 자료를 특정 감염병 관련 연구 기관에 제공만하는 것으로 알려져 있다. 그러나 중요한 이야기는 여기서 시작이다. 2015년 하버드 대학교 통계학과의 Samuel Kou 교수와 그의 연구그룹에서 기존 플루 트렌드 모형의 문제점-사람들의 검색 방법의 동적 특성이 무시됨, 해당 질병의 계절성이 무시됨 등-을 지적하고 이를 해결하는 모형으로 AutoRegression with GOogle search (ARGO)를 발표하였다. 이 모형의 예측 성능은 현재까지도 미국 질병관리본부의 실제 보고된 수치와 매우 잘 맞는 것으로 알려져 있다. 플루 이외에도 ARGO는 여러 나라의 **뎅기열에 대한 실시간 트렌드에 대해서도 굉장히 정확한 예측**을 하는 것으로 보고되었다. 최근에는 전자건강기록(electronic health record) 정보까지 반영하여 **실시간 감염병 탐지의 정확도가 더 개선**되었다는 결과도 저널에 보고되어있다.

위의 사례에서 이야기하는 것은 감염병 탐지와 관련하여 빅데이터의 실패 사례는 빅데이터가 유용하지 않아서 라기보다는, 빅데이터 사용방법에 문제가 있었다는 점이다. 즉, 빅데이터가 엄청난 정보의 원천이더라도 이를 활용하는 소프트웨어적인 기법이 부적절하다면 마치 우리에게 빅데이터 자체가 유용하지 않다는 거짓된 인상을 줄 수 있다는 것이다. 또한 현재 많은 전문가들에 따르면, 빅데이터를 축적하는데 기여한 사람들의 집단이 전체 집단에 대해 대표성이 없는 경우, 즉 선택 편향(selection bias)이 있는 경우, 빅데이터의 엄청난 크기에 비례하여 통계적 추론 결과가 심각하게 왜곡될 수 있다고 한다. 따라서 **실시간 감염병 탐지 문제에서 빅데이터를 활용하는 경우 얼마나 과학적으로 타당하고 논리적인 근거로 보정하여 분석**하는가가 빅데이터로부터 가치를 만들어내는 가장 중요한 부분이라고 볼 수 있다. 빅데이터는 여전히 우리의 손길을 기다리고 있으며, 아직 기대할 것이 많은 재미난 "무엇"임에는 틀림없다.