우리나라 보건복지 빅데이터 동향 및 활용 방안[1)]

송태민
한국보건사회연구원 연구위원
tmsong@kihasa.re.kr

Ⅰ. 서론

최근 스마트폰, 스마트TV, RFID, 센서 등의 급속한 보급과 모바일 인터넷과 소셜미디어의 확산으로 데이터량이 기하급수적으로 증가하고 데이터의 생산, 유통, 소비 체계에 큰 변화를 주면서 데이터가 경제적 자산이 될 수 있는 빅데이터 시대를 맞이하게 되었다. 특히, 정보통신기술(Information Communication Technology: ICT)이 다른 산업들과 융복합되면서 방대한 양의 데이터들이 생산되고 있는 가운데 사회변화에 따른 삶의 질에 대한 욕구 및 현안해결에 빅데이터의 활용은 매우 중요한 과제로 떠오르고 있다(이성훈·이동우, 2013). 빅데이터 생산의 주된 원인은 기존의 레거시 시스템의 성공적 구축과 함께 스마트기기의 보급으로 사용자의 위치정보와 온라인 및 모바일 사용기록과 SNS에서 사용자의 일상생활과 의견이 어딘가에 모두 저장됨에 따라 정보량이 폭증하는 것이다. 이러한 정보량은 2012년에 연간 2.7제타바이트(2조 7천억 Giga bytes)[2)]를 넘어서고, 페이스북 가입자가 8억 명을 돌파하고, 카카오톡의 하루 전송 메시지가 10억 건, 모바일 기기 1조 대 이상, M2M 센서 20억 대 이상 보급, 1분에 유튜브 동영상이 60시간 분량 이상 업로드됨에 따라서 빅데이터는 폭발적으로 증가할 것으로 예측하고 있다(함유근·채승병, 2012; 윤형중, 2012). 세계 각국의 정부와 기업들은 빅데이터가 향후 국가와 기업의 성패를 가름할 새로운 경제적 가치의 원천이 될 것으로 기대하고 있으며, McKinsey, The Economist, Gartner 등은 빅데이터를 활용한 시장변동 예측과 신사업 발굴 등 경제적 가치창출 사례 및 효과를 제시하고 있다. The Economist(2010)는 빅데이터를 제대로 활용하면 전 세계가 직면한 환경, 에너지, 식량, 의료 문제를 상당부분 해결할 것으로 전망하고 있고, Gartner(2011)는 빅데이터가 미래 경쟁력을 좌우하는 21세기의 원유이며, Mckinsey(2011)는 빅데이터의 활용에 따라 기업/공공분야의 경쟁력 확보와 생산성 향상, 사업혁신/신규사업 발굴의 차이가 생길 것이라고 보고 있다. 또한, Mckinsey(2011)는

1) 본 원고는 '송태민(2012. 11), "보건복지 빅데이터의 효율적 활용 방안", 「보건복지포럼」, 한국보건사회연구원'의 내용을 수정·보완함.
2) 2012년 쏟아질 데이터는 2.7제타바이트이고 매일 쏟아지는 데이터는 평균 7.7엑사바이트로 이를 저장하려면 1테라바이트 PC용 HDD가 하루에 750만 개가 필요.

빅데이터 활용 시 미국 의료분야에서 연 3,000억 달러, 유럽 공공분야에서 연 2,500억 달러의 경제적 효과가 있으며, 우리나라는 약 10.7조의 정부지출을 감소시킬 것으로 예측하고 있다. 일본 총무성(2012)은 빅데이터의 활용이 촉진되면 부가가치의 창출이나 사회적 비용의 절감에서 총 16조 원 이상의 경제적인 효과가 얻어질 것으로 예상하고 있다.

한편 인구 고령화와 만성질환 유병률의 증가로 의료비 문제와 의료서비스의 접근성 및 질에 관한 문제가 논의되면서 많은 국가에서 IT와 의료기술을 접목한 u-Health 도입을 추진하여 왔다. u-Health는 의료비 절감 등의 사회경제적 비용감소 효과와 공공보건 의료서비스와 예방관리 보건 등의 사회정책적 효과를 기대할 수 있는 가장 효과적인 대안으로 각광받고 있다. u-Health의 보급은 의료분야에서 많은 변화를 가져올 것으로 보고 있다. 유무선 통신기술과 센싱기술의 발전으로 u-Health 기기나 스마트TV 등을 통하여 의사의 건강상담 및 진료가 가능한 의료서비스를 이용할 수 있으며 개인의 건강정보를 기록하는 전자의무기록(EHR)을 통해 환자의 건강상태를 실시간으로 관찰할 수 있게 되었다. u-Health 서비스는 다양한 생체정보를 수집하기 위해 다양한 스마트센서들의 네트워크가 필수적이다. 스마트센서들이 수집한 환자의 의료정보나 건강정보는 다양한 형태로 분석·처리되어 개인의료정보에 저장되고 병원의 의사나 간호사 등에 전송되어 활용될 수 있다. 최근 이러한 u-health 서비스를 통해 생산되는 건강정보 관련 빅데이터의 관리와 활용에 대한 논의가 활발히 진행되고 있다. 따라서 본고에서는 빅데이터의 개념과 보건복지 분야의 빅데이터 적용사례에 대해 살펴봄으로써 보건복지 빅데이터의 효율적 관리와 활용을 위한 방안을 제시하고자 한다.

Ⅱ. 관련 연구

1. 빅데이터 개념

빅데이터(Big Data)는 Wikipedia(2013. 5. 30.)에서 '기존 데이터베이스 관리도구로 데이터를 수집·저장·관리·분석의 역량을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술'로 정의하고 있으며, 국가정보화전략위원회에서는 '대용량 데이터를 활용, 분석하여 가치 있는 정보를 추출하고, 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 정보화 기술'이라고 정의하고 있다(국가정보화전략위원회, 2011). 삼성경제연구소는 '빅데이터란 수십에서 수천 테라바이트 정도의 거대한 크기를 갖고 여러 가지 다양한 비정형 데이터를 포함하고 있으며, 생성, 유통, 소비가 몇 초에서 몇 시간 단위로 일어나 기존의 관리 및 분석 체계로는 감당할 수 없을 정도의 거대한 데이터의 집합으로 대규모 데이터와 관계된 인력, 조직, 기술 및 도구(수집, 저장, 검색, 공유, 분석, 시각화 등)까지 모두 포함하는 개념'으로 정의하고 있다(함유근·채승병, 2012). 이와 같은 정의를 살펴볼 때 빅데이터란 엄청나게 많은 데이터로 양적인

의미를 벗어나 데이터 분석과 활용을 포괄하는 개념으로 사용되고 있다. 인터넷이 일상화된 최근 10년 사이, 인류는 디지털 데이터가 폭증하는 데이터 홍수(Data Deluge)(IDC, 2011) 현상에 직면하여 2011년 전 세계 데이터에서 생성될 디지털 정보량이 1.8ZB(제타바이트)에 달하는 '제타바이트 시대[3]'로 진입[4]함에 따라 빅데이터의 용어가 등장하기 시작하였다(정지선, 2011). 빅데이터의 주요특성은 일반적으로 3V(Volume, Variety, Velocity)를 기본으로 1V(Value)나 1C(Complexity)의 특성을 추가하여 설명하고 있다.[5] 빅데이터 4가지 구성요소는 〈표 1〉과 같다. 비즈니스 분석 솔루션 기업인 SAS는 데이터의 가치(Value)에 중점을 두어 가치를 창출하기 위한 비즈니스 예측 및 최적화 주제를 선정하여 빅데이터로부터 어떤 가치 있는 정보를 얻을 것인가에 분석 관점을 가지고 있다(김근태, 2012).

〈표 1〉 빅데이터의 4가지 구성요소

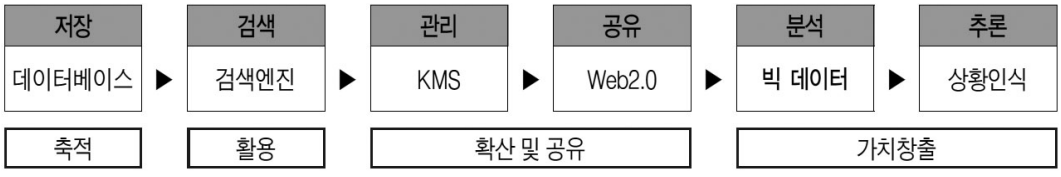
구분	주요내용
규모(Volume)의 증가	기술적인 발전과 IT의 일상화가 진행되면서 해마다 디지털 정보량이 기하급수적으로 폭증하여 제타바이트(ZB)시대로 진입
다양성(Variety)의 증가	로그기록, 소셜, 위치, 소비, 현실데이터 등 데이터의 종류의 증가와 멀티미디어 등 비정형화된 데이터 유형의 다양화
복잡성(Complexity)의 증가	구조화되지 않은 데이터, 저장방식의 차이, 중복성 문제, 데이터의 종류 확대, 데이터 관리 및 처리의 복잡성이 심화
속도(Velocity)의 증가	사물정보(센서, 모니터링), 스트리밍 정보 등 실시간 정보의 증가로 데이터의 생성과 이동(유통) 속도가 증가, 대규모 데이터 처리와 정보의 활용을 위한 데이터 처리 및 분석 속도가 중요

자료: 정지선(2011), 「新가치창출 엔진, 빅데이터의 새로운 가능성과 대응 전략」.

2. 빅데이터 기술

정보통신기술 주도권이 인프라, 기술, SW 등에서 데이터로 이동됨에 따라 빅데이터의 역할은 분석과 추론(전망)의 방향으로 진화되어 가치창출의 원천요소로 작용하고 있다([그림 1]).

[그림 1] 데이터의 진화단계

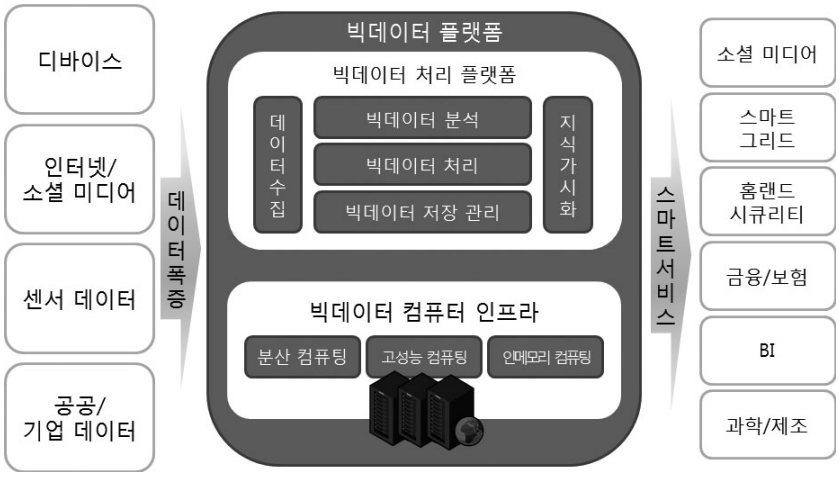


자료: 정지선(2011), 「新가치창출 엔진, 빅데이터의 새로운 가능성과 대응 전략」.

3) 1ZB(제타바이트)는 1조 GB(기가바이트)에 해당하는 양으로 미의회도서관 저장정보(235 테라바이트; '11. 4월 현재)의 4백만 배에 해당.
4) IDC(2011), *Digital Universe study.*
5) 정지선(2012), "성공적인 빅데이터 활용을 위한 3대요소: 자원, 기술, 인력", 「IT & Future Strategy」, 제3호.

빅데이터 기술은 '생성→수집→저장→분석→표현'의 처리 전 과정을 거치면서 요구되는 개념으로 분석기술과 인프라는 〈표 2〉와 같다. 빅데이터 분석기술은 통계, 데이터마이닝, 기계학습, 자연어 처리, 패턴인식, 소셜네트워크 분석, 비디오·오디오·이미지 프로세싱 등이 해당된다. 빅데이터의 활용, 분석, 처리 등을 포함하는 인프라에는 BI(Business Intelligence), DW(Data Warehouse), 클라우드 컴퓨팅, 분산데이터베이스(NoSQL), 분산 병렬처리, 하둡(Hadoop)[6] 분산파일시스템(HDFS), MapReduce 등이 해당된다.[7][8] 그리고 다양한 데이터 소스에서 수집된 빅데이터를 처리·분석하여 지식을 추출하고 이를 기반으로 지능화된 서비스를 제공하기 위해서는 빅데이터 플랫폼이 필요하다.

[그림 2] 빅데이터 플랫폼



자료: 황승구 외(2013), 「빅데이터 플랫폼 전략」, 전자신문사, p. 81.

데이터 수집에는 ETL(Extraction Transformation Loading)과 크롤링엔진(Crawling Engine)을 사용한다. ETL은 다양한 소스시스템으로부터 필요한 데이터를 추출하여 변환작업을 거쳐 저장하거나 분석을 담당하는 시스템으로 전송 및 적재하는 모든 과정을 포함한다. 크롤링엔진은 로봇이 거미줄처럼 얽혀있는 인터넷 링크를 따라다니며 방문한 사이트의 모든 페이지의 복사본을 생성함으로 문서를 수집한다. 하둡 분산파일시스템을 이용하여 수천 대의 노드들을 연결하여 수 페타바이트급의 저장용량을 제공하고, 다이나모(Dynamo) 분산데이터 관리시스템은 데이터를 분할하여 노드들을 배치함으로써 대용량의 데이터를 관리할 수 있다. 인메모리컴퓨팅은 데이터베이스 자체를 메모리에 올려서 입출력을 빠르게 하여 데이터의 분석과 저장, 제공을 빠르게 지원한다.

6) 하둡은 대용량 데이터 처리 분석을 위한 대규모 분산 컴퓨팅 지원 프레임워크로 하둡 분산파일시스템(HDFS)과 분산처리를 위한 맵리듀스가 핵심요소이며 그 외 분산DB인 Hbase, 검색엔진(Nutch), 쿼리 언어(Pig) 등을 포함한다.
7) Peter Warden(2011), *Big Data Glossary*, O'Reilly Media.
8) 장상현(2012), "빅데이터와 스마트교육", 「한국정보과학회지」, 제30권 제6호(통권 제277호), pp. 59~64.

〈표 2〉 빅데이터 처리 프로세스별 기술 영역

흐름	영역	개요
소스	내부데이터	Database, File Management System
	외부데이터	File, Multimedia Streaming
수집	크롤링(crawling)	검색엔진의 로봇을 이용한 데이터 수집
	ETL(Extraction, Transformation, Loading)	소스데이터의 추출, 전송, 변환, 적재
저장	NoSQL Databases	비정형 데이터 관리
	Storage	빅데이터 관리
	Servers	초경량 서버
처리	Map & Reduce	데이터의 추출
	Processing	다중업무처리
분석	NLP(Neuro Linguistic Programming)	자연어처리
	Machine Learning	기계 학습을 통해 데이터의 패턴 발견
	Serialization	데이터 건의 순서화
표현	Visualization	데이터를 도표나 그래픽적으로 표현
	Acquisition	데이터의 획득 및 재해석

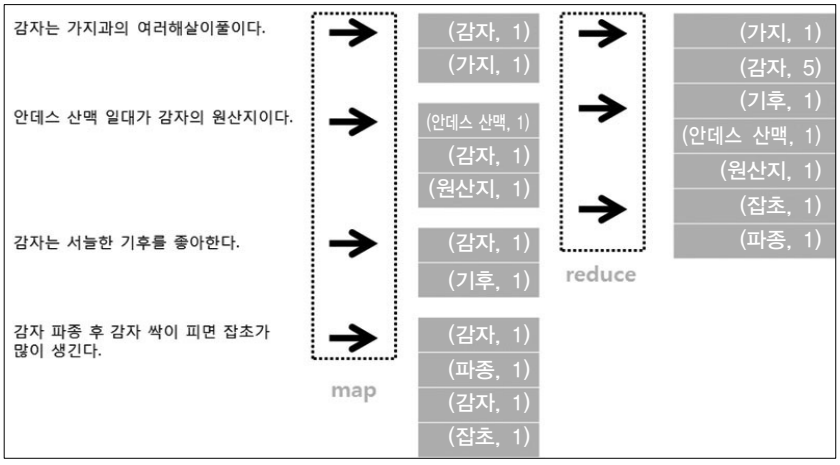
자료: Pete Warden(2011), *Big Data Glossary*, O'Reilly Media.

빅데이터 분야에서는 '소셜 애널리틱스(Social Analytics)'가 페이스북, 트위터, SNS 등에서 수집되는 비정형 데이터를 신속하게 분석을 한다. 소셜미디어에서 정보를 뽑아내고 분석하는 방법은 크게 3가지로 나눌 수 있다.

첫째, 텍스트마이닝(Text Mining)은 인간이 언어로 쓰인 비정형 텍스트에서 자연어처리기술을 이용하여 유용한 정보를 추출하거나, 연계성을 파악, 분류 혹은 군집화, 요약 등 빅데이터의 숨겨진 의미 있는 정보를 발견하는 것이다. 둘째, 오피니언마이닝(Opinion Mining)은 소셜미디어의 텍스트 문장을 대상으로 자연어처리기술과 감성분석 기술을 적용하여 사용자의 의견을 분석하는 것으로 마케팅에서는 버즈(Buzz; 입소문)분석이라고도 한다. 셋째, 네트워크분석(Network Analytics)은 네트워크 연결구조와 연결강도를 분석하여 어떤 메시지가 어떤 경로를 통해 전파가 되는지, 누구에게 영향을 미칠 수 있는지를 파악하는 것이다.

빅데이터의 주요 기술로는 구글에서 개발한 비구조적 데이터의 획득, 조직화, 분석을 하기 위한 기술인 맵리듀스(Map & Reduce)가 있다.

[그림 3] Map & Reduce 사례



자료: 함유근·채승병(2012), 「빅데이터 경영을 바꾸다」, 삼성경제연구소, p. 81.

Ⅲ. 본론

1. 보건복지 빅데이터 활용 사례 및 수요[9)]

보건 분야의 빅데이터 국외 활용 사례로 미국 국립보건원은 다양한 질병을 연구하기 위해 유전자 데이터를 공유 분석할 수 있는 유전자 데이터 공유를 통한 질병치료체계를 마련하여 주요 관리 대상에 해당하는 질병에 대한 관리 및 예측을 실시하고 있다. 현재 1,700명의 유전자 정보를 아마존 클라우드에 저장하여 누구나 데이터를 이용 가능하게 구축하였다(www.1000genomes.org/). 미국 국립보건원 산하 국립의학도서관에서는 사용자가 요구하는 다양한 약에 대한 정보를 제공하고 제조사와 사용자 간의 쌍방향 상호작용을 통해 약의 정보를 제공하는 Pillbox 프로젝트를 통한 의료개혁을 추진하고 있다. Pillbox 서비스(pillbox.nlm.nih.gov/)로 미국 국립보건원에 접수되는 알약의 기능이나 유효기간을 문의하는 민원 수는 100만 건 이상으로 평균 한 건당 확인하는 소요 비용 50달러를 감안하면 연간 5,000만 달러의 비용 절감효과가 있는 것으로 전망하고 있다. 미국 퇴역군인국(U.S. Department of Veterans Affairs)에서는 퇴역군인의 전자의료기록 분석을 통한 맞춤형 의료서비스를 지원하는 빅데이터 분석을 위해 2년간 25개의 DW를 배치하여 2,200만 퇴역군인에게 의료서비스를 제공하고 있다. 퇴역군인 전자의무기록(EHR)을 분석하여 의사가 개별 환자를 쉽게 진료할 수 있도록 지원하고 있다. 싱가포르 PA(People's Association)는 1,800개 이상의 주민위원

9) 본 절은 일부 내용은 '한국정보화진흥원(2012), 「빅데이터로 진화하는 세상(Big Data 글로벌 선진 사례)」'에서 보건(건강한 사회)과 복지(안전한 사회)의 사례를 분석 재정리함.

회 센터(커뮤니케이션 센터)에서 진행되는 다양한 활동들을 공유하기 위해 주민위원회센터 네트워크 기반의 맞춤형 복지사회를 구현하였다. 싱가포르 PA는 빅데이터 처리를 위하여 다양한 인종, 나이, 문화, 소득, 연령에 따른 주민의 데이터를 수집·분석하여 개인별 맞춤형 서비스를 제공하고 있다.

캐나다 온타리오 공과대병원은 인큐베이터 내 미숙아에 대한 다양한 데이터를 분석하여 병원균 감염을 예측할 수 있는 시스템을 개발하여 미숙아 모니터링을 통한 감염예방 및 예측, 감염징후 등을 조기에 발견하고 다른 미숙아 등에 대한 감염을 예방하며 퇴원 후 무선센서를 이용하여 병원 밖에서도 환자들을 실시간으로 체크를 할 수 있는 시스템을 구축하였다.

IBM과 미국 건강보험회사인 웰포인트(Wellpoint)는 의사와 다른 의료진들이 진단과 환자치료에 이용할 수 있는 애플리케이션(왓슨)을 개발하여 제공하고 있다. 왓슨은 임상실험 및 우수 치료사례 등 과거 데이터를 분석하여 환자에게 가장 적절한 치료방법을 제공하고 최신 정보를 과학적인 방법으로 제시하고 있다. 구글(Google)은 감기와 관련된 검색어 분석을 통하여 독감예보 서비스 제공하고 있다. 구글 독감예보 서비스(구글 플루 트렌드; www.google.org/flutrends/)는 다양한 사용자의 검색어 분석을 통하여 사용자에게 다시 유의미한 데이터로 가공하여 정확한 정보를 실시간으로 제공하고 있다. 보건 분야 국내 활용 사례로 질병관리본부에서 운영하는 한국인체자원은행네트워크(kbn.cdc.go.kr/)는 16개 병원을 통해 36만 명의 인체자원 확보하여 질병지표 발굴 및 질병조기 진단을 위해 활용하고 있다. 한국인체자원은행네트워크는 생명연구자원의 체계적 수집과 정보 표준화, 정보공유를 통하여 질병의 예방과 진단, 맞춤치료, 신약-신기술을 위한 미래 바이오산업의 신성장동력으로서 기반을 마련하고 있다. 분당서울대병원은 빅데이터 도입을 통해 업무효율성 및 생산성향상을 위한 임상의사결정지원시스템을 개발하였다. 임상의사결정지원시스템은 환자 개인의 특이사항을 입력하여 임상적 의사결정을 지원하기 위한 서비스로 시스템이 도입된 후, 부적절한 용량의 신독성 약물 처방률이 30.6%로 감소하는 효과를 가져왔다. 임상의사결정지원시스템은 빅데이터를 분석하여 자연어 검색을 지원하고 의약품의 처방과 조제 시 의약품 안정성과 관련된 정보를 실시간으로 제공하여 부적절한 약물사용을 사전에 검사할 수 있도록 확대하고 있다. DNA Link(dnalink.com/)에서는 질병관리 분석과 개인의 유전체 염기서열 분석으로 맞춤형 건강진단 서비스를 제공하는 유전자 분석시스템을 제공하고 있다. 연세대학교의료원에서는 u-Health를 이용하여 언제 어디서나 질병예방, 진단, 치료가 가능한 후(H∞H) 헬스케어 시스템을 제공하고 있다.

복지 분야 국외 사례로는 주로 안전과 관련한 빅데이터 활용이 주를 이루고 있다. 싱가포르에서는 국가위험관리시스템(Risk Assessment Horizon Scanning)을 구축하여 질병, 금융위기 등 모든 국가적 위험을 수집 및 분석을 하고 있다. RAHS(hsc.gov.sg)는 2004년부터 빅데이터를 기반으로 한 위험관리 계획을 추진하여 수집된 정보는 시뮬레이션, 시나리오 기법을 통해 분석하여 사전 위험 예측 및 대응 방안을 모색하고 있다. FBI는 유전자 색인 시스템 활용하여 단시간 범인을 검거하는 체계를 구축하고 있다. FBI의 유전자 정보은행 CODIS(Combined, DNA Index System)은 미제사건 용의자 및 실종자에 대한 DNA 정보 1만 3,000건을 포함하여 12만 명의 범죄자 DNA 정보가

저장되고 매년 2,200만 명의 DNA 샘플을 추가하여 범죄 수사에 활용하며 약 350만 개의 DNA 분석표가 내장되어 있다. 샌프란시스코 경찰청은 범죄발생지역 및 시각을 예측하여 범죄를 미연에 방지하기 위한 범죄예방시스템을 구축하였다(www.crimemapping.com). 범죄예방시스템은 과거 범죄를 분석하여 효율적으로 경찰을 배치하고 과거 범죄자 및 범죄유형을 SNS를 통해 지속적으로 관찰함으로써 그와 관련된 조직 및 범죄에 대한 예방을 하고 있다.

복지 분야 국내 활용현황으로는 보건복지부가 사회복지통합관리망(행복e음)을 개발하여 수요자 중심의 복지서비스 구현하였다. 사회복지통합관리망은 지자체 공무원들의 복지행정 처리를 지원하는 정보시스템으로 지자체에서 집행하는 120여 가지의 복지급여 및 서비스 이력 데이터를 이용하여 복지대상자 선정과 맞춤형 서비스를 제공하고 있다. 근로복지공단은 공공부문 고객관계관리(CRM)를 구축하여 '찾아가는 서비스'를 통한 맞춤형 서비스를 제공하고 있다. 한국정보화진흥원 빅데이터 국가전략포럼 분석팀에서는 2012년 1월부터 10월 18일까지 자살로 언급된 빅데이터 자료를 뉴스(온라인에서 게재되는 214개 웹사이트), 블로그(네이트, 네이버, 이글루스, 다음, 티스토리, 야후), 카페(네이버, 다음, 뽐뿌, 카드고릴라, SLR 클럽), SNS(트위터, 미투데이), 게시판(네이버 지식인, 네이트 지식, 다음 신지식 등) 등에서 수집하여 청소년이 작성했다고 추정되는 6만 9,886건을 분석하였다. 이를 통해 청소년들은 자살과 관련하여 온라인상에 많은 Buzz를 생성하고 있으며 Buzz의 발생패턴에 따라 보다 체계적으로 대응할 수 있는 자살예방체계를 설계할 수 있다는 가능성을 보였다.[10)]

한편, 2013년 현재, 보건의료 분야 빅데이터 시범사업으로 진행 중인 과제는 다음과 같다. 첫째, 국민건강 주의예보 시범서비스 구축사업은 국민건강공단의 건강보험 DB와 SNS 정보를 융합하여 홍역, 조류독감, SAS 등 감염별 발생 예측 모델링을 개발하고, 이를 상시 모니터링하여 위험징후 시 주의예보 서비스를 제공한다. 둘째, 빅데이터 기반 의약품 안전성 조기경보 서비스는 유해사례 DB, 치료기록, SNS 등을 연계 분석하여, 유의의약품을 추출하고, 이들의 위험도를 예측하여 병의원, 제약회사 및 유관기관 등과 정보공유를 하는 사업이다. 셋째, 보건의료 빅데이터 활용 시범사업은 포털과 질병관리본부 등과 협의된 데이터 외 병원자체 데이터를 활용하여 독감유행 예측, 심실부정맥 예측, 입원 병상배정 최적화, 신종 마약류 인지 및 감시 서비스를 제공한다.

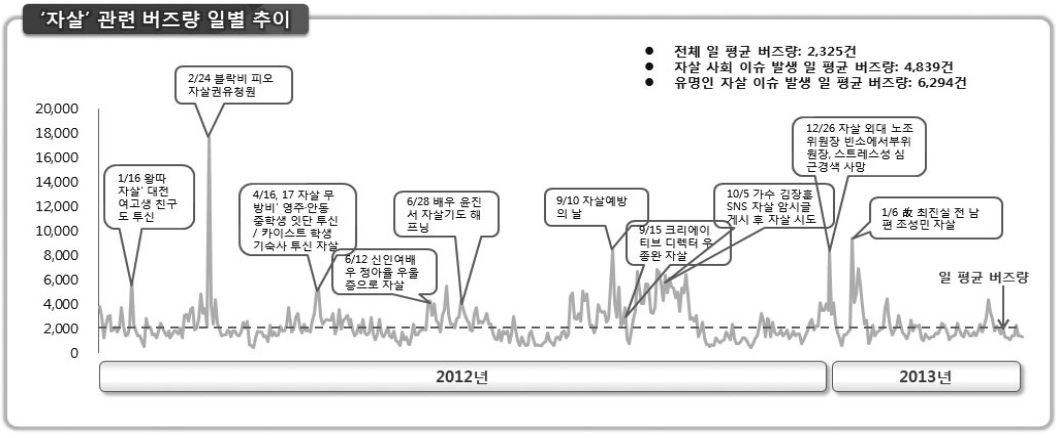
2. 빅데이터 분석 사례[11)]

우리나라는 급격한 사회·경제적 변화 속에 자살률이 2004년부터 OECD 국가 중 최고의 수준이며, 특히 청소년계층의 자살문제가 사회적 이슈로 대두되면서 정부 차원의 대책이 시급한 실정이다. 그동안 자살의 연구는 국가 간 자살률 비교나 패널 데이터의 분석을 통한 자살 원인에 대한 연구가 진행되어 왔으나 데이터 수집의 제한으로 인하여 개인과 집단의 다양한 자살 원인에 대한 분석은

10) 한국정보화진흥원·빅데이터국가전략포럼(2012. 10. 29.), 「대한민국 사회현안과 빅데이터 전략」, 제3차 빅데이터 국가전략 포럼.
11) 본 분석사례는 한국보건사회연구원과 (주)SK telecom이 공동으로 연구한 내용임을 밝힘.

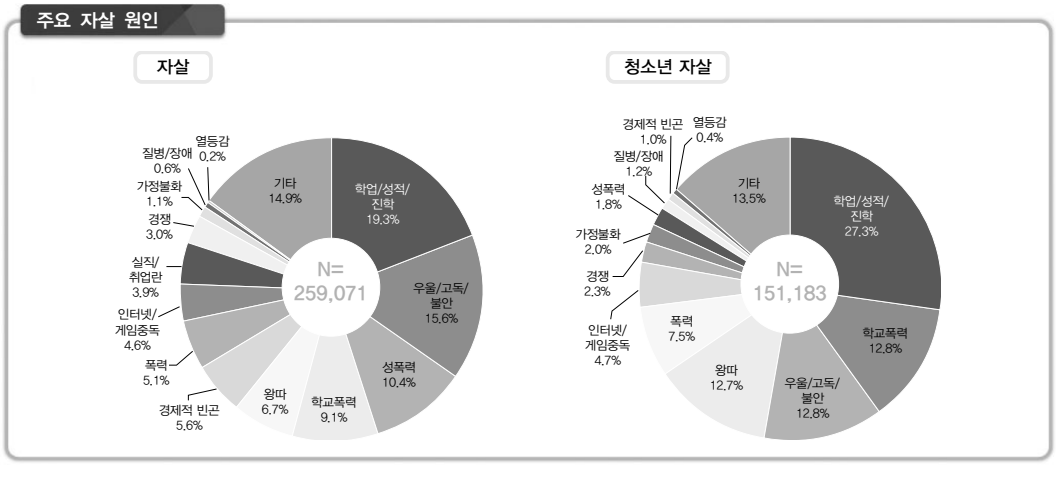
미흡한 실정이다. 따라서 빅데이터 분석을 통하여 다양한 자살의 원인과 자살에 대한 위험징후를 예측할 수 있을 것으로 본다. 본 연구는 2011년 1월 1일~2013년 3월 31일까지 인터넷 뉴스, 블로그, 카페, 게시판, SNS(트위터, 미투데이) 등의 온라인 채널에서 발생한 '청소년 자살' 관련 온라인 Buzz(본문, 댓글 포함)를 분석대상으로 하였다. 청소년 자살, 유명인 자살 등 자살과 관련된 사회적 이슈 발생 시에 자살과 관련한 커뮤니케이션이 급증하는 양상을 보이고 있으며 특히 연예인 자살 이슈 발생 시 버즈량이 급증하는 것으로 나타났다.

[그림 4] 자살 관련 버즈량 일별 추이



청소년 자살 원인의 1위는 '학업/성적/진학'으로 관련된 스트레스로 인한 청소년 자살이 우리 사회의 가장 큰 이슈가 되는 것으로 나타났다.

[그림 5] 주요 자살 원인



청소년 자살 검색의 예측모형은 〈표 3〉과 같다. [모형 1]은 2011년 일별 청소년(19세 이하) 자살률이 청소년 자살 검색에 미치는 영향을 예측하는 것으로 청소년 자살자 수가 많을수록(자살률이 높을수록) 청소년의 자살 검색은 증가하는 것으로 나타났다. [모형 2]는 일별 청소년 자살률과 스트레스 검색이 자살 검색에 미치는 영향을 예측하는 것으로 자살자 수가 많을수록, 또 스트레스 검색이 많을수록 청소년의 자살 검색은 증가하는 것으로 나타났다. [모형 3]은 일별 청소년 자살률, 스트레스 검색, 음주 검색이 청소년 자살 검색에 미치는 영향을 예측하는 것으로 자살자 수, 스트레스 검색, 음주 검색이 많을수록 청소년의 자살 검색은 증가하는 것으로 나타났다. [모형 4]는 일별 청소년 자살률, 스트레스 검색, 음주 검색, 미세먼지량이 청소년 자살 검색에 미치는 영향을 예측하는 것으로 자살자 수가 많을수록, 스트레스 검색이 많을수록, 음주 검색이 많을수록, 미세먼지량이 적을수록 청소년의 자살 검색은 증가하는 것으로 나타났다. 청소년의 자살 검색을 예측하는 모형의 그래프는 [그림 6]과 같다.

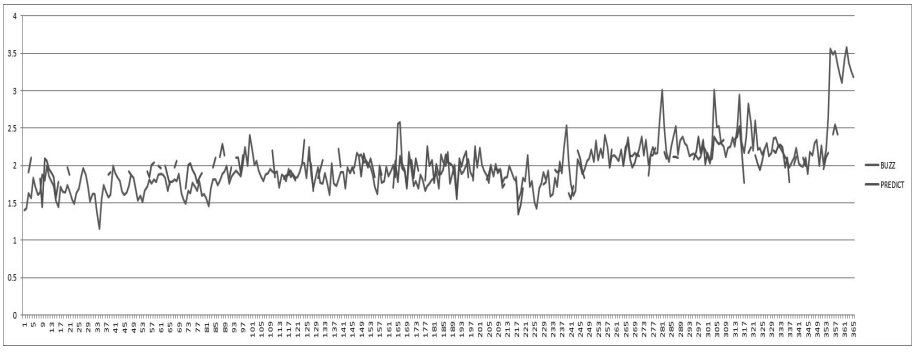
〈표 3〉 청소년의 자살 검색 예측모형

독립변수	Model 1		Model 2		Model 3		Model 4	
	β	t	β	t	β	t	β	t
(상수)	1.957	68.86**	.648	6.48**	.624	6.22**	.939	5.75**
자살률	.290	2.53*	.187	2.17*	.173	2.01*	.192	2.24*
스트레스검색량			1.003	13.40**	.913	10.94**	.935	11.25**
음주검색량					.178	2.34*	.176	2.34*
미세먼지량							−.214	−2.43*
수정된 R^2	.023		.448		.456		.467	
F	6.421**		95.430**		65.464**		51.629**	

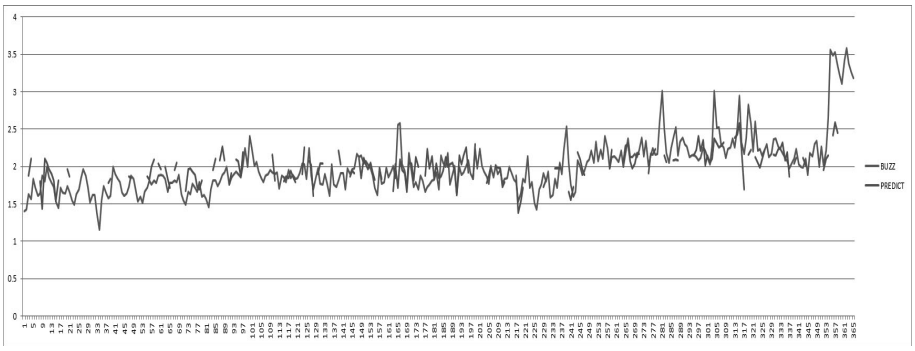
** p<0.01, * p<0.05

[그림 6] 청소년의 자살 검색 예측모형

Model 3

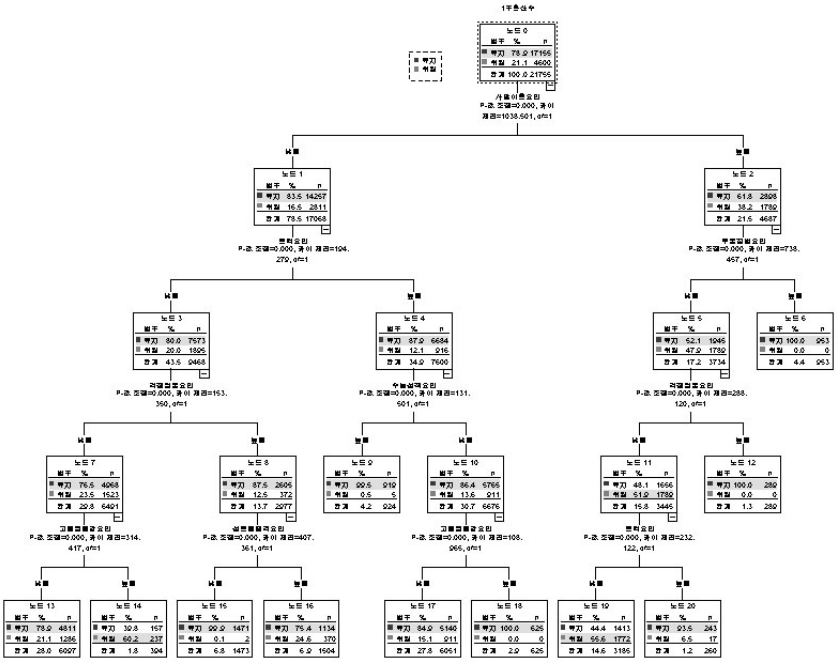


Model 4



청소년 버즈 확산의 위험도를 측정하기 위한 데이터마이닝의 의사결정나무 분석결과 사망이혼요인이 가장 큰 확산 위험요인으로 나타났으며, 사망이혼요인의 검색이 높은 집단은 우울질병요인, 걱정얼굴요인, 폭력요인 순으로 위험의 영향이 높은 것으로 나타났다. 그리고 사망이혼요인이 낮은 집단은 폭력요인, (걱정얼굴요인, 수능성적요인), (고통열등감요인, 성폭행충격요인) 순으로 위험의 영향이 높은 것으로 나타났다.

[그림 7] 청소년 자살 검색 확산 예측모형



본 연구를 바탕으로 우리나라의 자살예방과 관련한 정책적 함의는 다음과 같다.[12)]

첫째, 본 연구의 결과는 성인과 청소년 모두 다양한 원인들에 의해 스트레스를 경험하면서 자살

을 검색하게 된다는 것이다. 따라서 청소년의 스트레스를 해소할 수 있는 학교 차원의 다양한 프로그램의 마련과 함께 성인의 경제활동으로 인한 스트레스를 해소시킬 수 있는 직장 차원의 프로그램이 개발되어야 할 것이다. 둘째, 성인과 청소년은 온라인상에서 자살과 관련한 담론을 주고받고 있으며 이러한 언급이 실제적인 자살과 관련된 심리적·행동적 특성으로 노출될 수 있기 때문에 자살예측모형에 따른 위험징후가 예측되면 실시간으로 개입할 수 있는 애플리케이션(가칭: 생명존중 온라인 게이트키퍼(Gate Keeper))이 개발되어야 할 것이다. '생명존중 온라인 게이트키퍼'는 자살에 대한 위험징후가 예측되면 소셜 담론의 분석에서 추측된 위험요인을 줄일 수 있는 맞춤형 프로그램을 실시간으로 제공할 수 있도록 개발되어야 할 것이다. 셋째, 지역별 소셜 빅데이터와 지역요인의 연계를 통한 지역별 자살 예측모형을 개발하여 실시간으로 자살을 사전에 방지할 수 있는 시스템(가칭: 지역 생명존중 예보시스템)을 구축하여야 할 것이다. '지역 생명존중 예보시스템'은 지역별 생명존중 관련 기관에 지역별 자살예보를 1주 또는 월 단위로 제공하여 지역별 자살행동의 원인에 대해 적극적 대응을 위한 대국민 홍보활동을 지속적으로 실시함으로써 지역 주민의 생명존중 인식을 강화할 수 있을 것이다. 넷째, SNS에서 주고받는 자살 관련 소셜 담론은 개인이 일상생활에서 갖는 우울한 감정이나 고민이 기록되는 '온라인 심리적 부검보고서'라 할 수 있다. 핀란드가 '오프라인 심리적 부검보고서'를 바탕으로 국가 차원의 자살예방대책을 마련하여 자살률을 줄였다면, 우리나라는 세계 최고수준의 IT 기술과 소셜 빅데이터의 활용과 분석으로 국가 차원의 자살예방 대책의 마련이 가능할 것으로 본다.

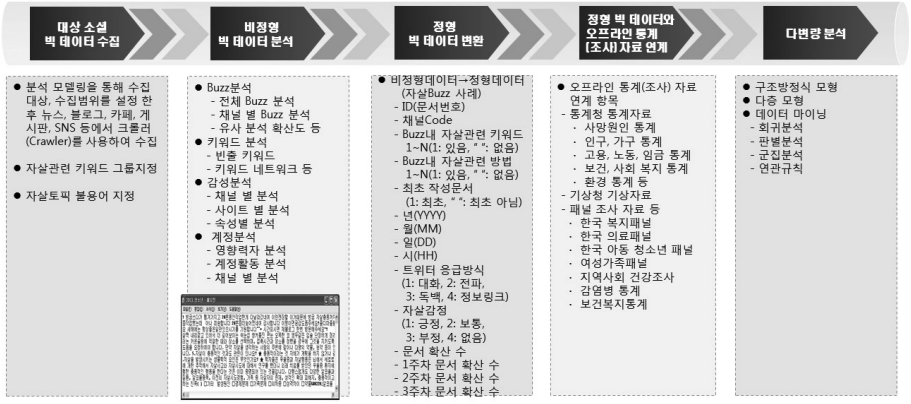
3. 빅데이터 분석 방법

소셜 빅데이터 분석 절차 및 방법은 [그림 8]과 같다. 첫째, 해당 Buzz(자살)분석 모델링을 통해 수집대상과 수집범위를 설정한 후, 대상 채널(뉴스, 블로그, 카페, 게시판, SNS 등)에서 크롤러(Crawler) 등 수집엔진(로봇)을 이용하여 수집한다. 이때 자살 관련 키워드 그룹(원인, 유형, 대상, 성별, 장소, 지역, 방법 등)과 자살 토픽에 대한 불용어 등을 지정하여 수집한다. 둘째, 수집된 비정형 데이터를 분석한다. 비정형 데이터는 Buzz 분석, 키워드 분석, 감성분석, 계정분석 등으로 진행된다. [그림 8]의 청소년 자살Buzz 수집 사례와 같이 비정형 데이터를 연구자가 분석하기는 수집된 상태로는 어렵다. 따라서 수집된 비정형 데이터는 Text Mining, Opinion Mining, Network Analysis를 통하여 비정형 데이터를 분류하는 절차가 필요하다. 셋째, 비정형 빅데이터를 정형 빅데이터로 변환해야 한다. 자살Buzz 사례를 살펴보면, 자살Buzz 각각의 문서는 ID로 Code화 되어야 하고, Buzz 내 키워드나 방법 등도 모두 Code화 되어야 한다. 넷째, 사회현상과 연계하여 분석하기 위하여 정형화된 빅데이터를 오프라인 통계(조사) 자료와 연계해야 한다. 오프라인 통계(조사) 자료는 대부분 정부나 공공기관에서 유료 또는 무료로 제공하고 있기 때문에 연계대상 자료와 함께 연

12) 본 연구의 결과와 정책적 함의는 '송태민(2013. 8), "소셜 빅데이터 분석을 통한 자살 검색 예측모형 개발", 「보건복지포럼」, 통권 제202호'의 내용 중에서 발췌한 것임을 밝힌다.

계 가능한 ID(일별, 월별, 연별, 지역별)를 확인한 후, 오프라인 자료를 수집하여 연계(Merge)할 수 있다. 다섯째, 오프라인 통계(조사) 자료와 연계된 정형화된 빅데이터의 분석은 요인간의 인과관계나 시간별 변화궤적을 분석할 수 있는 구조방정식모형이나 일별(월별, 연별), 지역별 사회현상과 관련된 요인과의 관계를 분석할 수 있는 다층분석, 그리고 수집된 키워드의 분류과정을 통해 새로운 현상을 발견할 수 있는 데이터마이닝 분석을 실시할 수 있다.

[그림 8] 소셜 빅데이터 분석 절차 및 방법(자살Buzz 사례)



Ⅳ. 결론

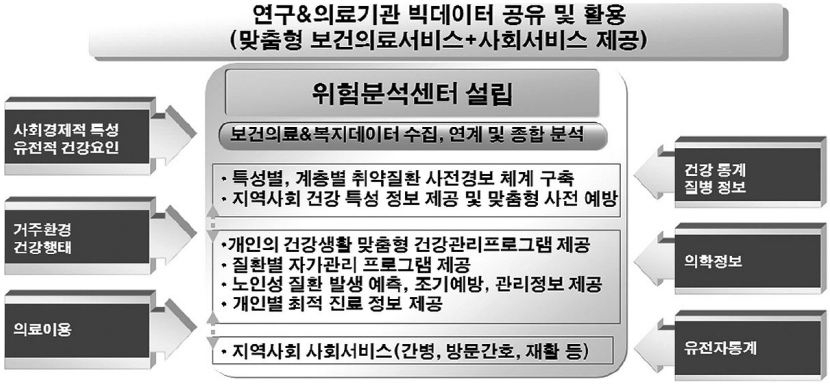
빅데이터는 신가치 창출의 엔진으로 보건복지 서비스에 새로운 패러다임을 제공할 수 있다. 국내의 보건복지 분야에서는 기존 레거시 시스템의 안정적인 구축으로 이미 수많은 빅데이터가 저장·관리되고 있다. 보건 분야에서 국민건강보험공단은 2002년부터 데이터웨어하우스를 구축하여 본부와 지역본부에서 운영 중인 급여관리시스템, 요양급여비지급시스템, 건강검진시스템, 의료보호시스템, 자격, 보험료 급여 및 사후 시스템에서 생성되는 데이터를 저장·관리하고 있다. 국민건강보험공단의 데이터웨어하우스는 보험료 시뮬레이션, 보험료 및 보험급여비 상승 추계 등의 정보를 제공하고 있다. 건강보험심사평가원에서는 2000년 의약분업 시행 이후 청구심사데이터가 비약적으로 증가하면서 2002년부터 데이터웨어하우스를 구축하여 기준정보, 요양기관정보, 지급정보에 대한 데이터를 저장·관리하고 있다. 건강보험심사평가원 데이터웨어하우스는 적시에 정보를 분석할 수 있도록 각 주제영역에 대한 통계분석, 시계열분석, 다차원분석, 추이분석 등과 같은 다양한 분석기법을 적용하고 있다. 또한 데이터의 활용목적별로 심사분석 데이터마트, 평가분석 데이터마트, 통계분석 데이터마트 등을 운영하고 있다. 국립암센터에서는 암통계(발생률, 사망률, 생존률) 산출로 암 부담 수준파악과 암 관리 정책 수립 근거를 마련하기 위한 추이 분석 등을 위하여 2002년부터 암 등록자

료의 데이터웨어하우스를 구축하여 운영하고 있다.

복지 분야에서는 사회복지통합관리망에 대부분의 복지정보가 통합·관리되고 있다. 사회복지통합관리망에는 사회복지통합관리망 채널, 희망복지, 복지행정, 복지급여통합, 새올행정, 외부영역의 6개 영역의 44개 세부업무별로 데이터가 저장·관리되고 있다. 그 외 식품의약품안전처에서는 수입식품현황이나 식품 관련 DB를 운영하고 있으며, 통계청과 국책연구기관들은 보건복지 관련 각종 통계 생산을 위한 패널 데이터를 구축하고 있다. 상기에 서술한 바와 같이 공공부분에서는 이미 수많은 정형화된 빅데이터가 저장·관리 되고 있을 뿐만 아니라 각 기관의 홈페이지나 SNS 서비스를 통해서도 많은 비정형 데이터가 관리되고 있다.

한편, 개인건강기록(Personal Health Record: PHR)이 의료서비스 소비자에게 다양한 건강정보를 제공하고 그들의 건강을 스스로 통제 관리할 수 있는 수단을 제공함에 따라 공공과 민간 차원의 PHR 구축이 지속적으로 추진되고 있다. PHR은 혈압과 같은 객관적인 자료를 수집할 수 있고 이런 자료는 측정되어 환자가 수동으로 입력하거나 u-Health 기기를 통해 직접 전송될 수도 있다. 즉, 24시간 원격 건강관리 모니터링을 위해 u-Health 기기를 통하여 전송된 PHR 정보는 건강생활습관(금연, 절주, 영양, 운동)의 확립과 자가건강관리 능력의 함양으로 국민 건강수명을 연장시키며 삶의 질을 향상시키고 만성질환에 대한 치료에서 예방 중심의 건강관리를 통한 의료비 절감의 효과를 기대할 수 있다(송태민 외, 2011). 앞에서 살펴본 바와 같이 보건복지영역과 빅데이터의 관계는 매우 밀접하다. 보건의료 분야에서는 생애주기별로 맞춤형 보건의료서비스를 제공하기 위해서는 보건의료뿐만 아니라 사회현안이나 미래중요이슈를 중심으로 빅데이터를 활용한 미래전망 및 정책의사결정모델을 도출할 필요가 있으며, 이를 위해서는 사회공동자산인 데이터의 부가가치를 높기 위한 '위험분석센터'의 설립이 필요하다. 위험분석센터에서 질병관리 및 예측, 다양한 사용자의 질병에 대한 통계데이터를 활용하여 주요 질병의 분포 및 추세를 예측함으로써 국가 차원의 조기대응이 가능할 것으로 본다(고숙자·정영호, 2012).

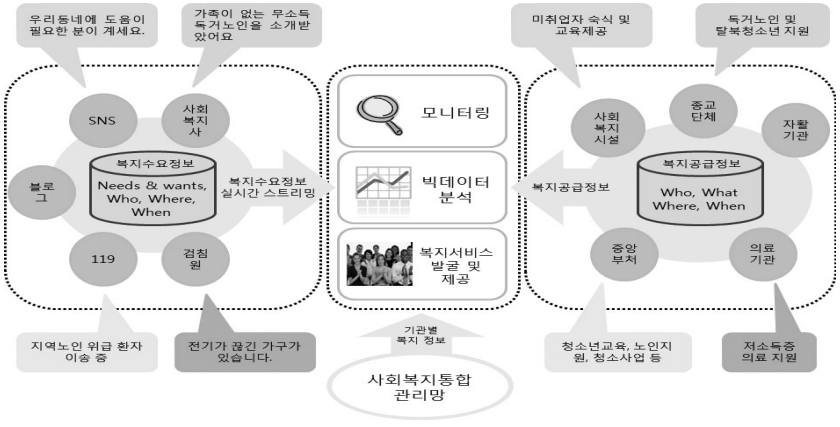
[그림 9] 위험분석센터 설립을 통한 빅데이터 활용 방안



자료: 고숙자·정영호(2012. 11), "국민건강 미래예측 시스템 구축방안", 「보건복지포럼」, 통권 제193호, 한국보건사회연구원.

복지 분야에서는 2010년 2월부터 보건복지부를 포함한 각 부처의 공공기관이 수행하는 복지사업과 수혜자정보를 통합관리하는 사회복지통합관리망(행복e음)을 가동하고 있다. 복지사각지대를 해소하고 생애주기별로 개인 맞춤형 복지서비스를 제공하려면 현재의 사회복지통합망을 전(全) 정부부처의 정보시스템과 통합 연동하는 국가 빅데이터 인프라로 확장하여야 할 것이다(황승구 외, 2013).

[그림 10] 복지 분야의 빅데이터 활용 방안



자료: 황승구 외(2013), 「빅데이터 플랫폼 전략」, 전자신문사, p. 201.

이러한 보건복지 분야 빅데이터를 효율적으로 활용하기 위해서는 다음과 같은 전략이 필요할 것이다.

첫째, 보건복지 빅데이터를 통합적으로 관리하기 위한 범부처 차원의 (가칭)보건복지 빅데이터 관리 위원회의 운영이 필요하다. 현재 보건복지 빅데이터는 보건복지부, 고용노동부, 지식경제부(현 산업통상자원부, 미래창조과학부), 식품의약품안전처, 통계청 등 많은 정부부처와 국민건강보험공단, 건강보험심사평가원, 국책연구기관 등 많은 공공기관에서 관리·운영되고 있어 각 기관에서 운영 중인 정보의 연계와 공유를 위해서는 범정부 차원의 조직이 필요할 것이다.

둘째, 비정형화된 보건복지 빅데이터를 관리하고 있는 민간 기관과의 협조체제가 마련되어야 할 것이다. 비정형화된 보건복지 빅데이터는 민간 기관의 검색포털이나 SNS를 통해서 생산·저장되고 있어 민간기관과의 긴밀한 협조체계(가칭: 보건복지 빅데이터 포럼)가 구축되어야 할 것이다.

셋째, 국가 차원의 오픈 API(Open Application Programming Interface)의 제공이 필요하다. 보건복지 빅데이터는 대부분 공공부문에서 독점하고 있다. 정보를 수집/분석하고 수집과 동시에 정보를 실시간으로 웹상에 공개하는 것도 중요하지만 보건복지 빅데이터를 효과적이고 효율적으로 활용하기 위해서는 정부 차원의 API 공개를 적극적으로 검토할 필요가 있다. 2013년 5월 기준으로 공유자원포털(www.data.go.kr)에서 공개되어 있는 공공정보는 1,721종으로 이중 보건의료 21종, 복지 29종에 불과하다. 국가정보화 전략위원회의에서는 폭증하는 데이터가 경제적 자산이 되는 시대

이기 때문에 정부가 능동적으로 빅데이터를 활용하고 국가지식 플랫폼을 만들기를 제안하고 있다.[13] 따라서 보건복지 빅데이터의 공개는 관련 기관과 빅데이터 전문가의 참여로 정부와 국민이 필요로 하는 정보를 분류하고 공개대상 정보는 개인정보를 철저하게 보안하여 국가지식 플랫폼에 저장할 수 있을 것이다. 넷째, 보건복지 빅데이터를 분석 처리할 수 있는 관련 기술의 개발이 필요하다. 스마트 시대에는 비관계형, 비정형 데이터의 저장과 분석, 클라우드 서비스의 확산, 시멘틱 검색 서비스, 추론에 기반한 상황인식 서비스 등의 기술이 핵심이 될 것이다. 따라서 관련 부처와 협력하여 보건복지 분야 빅데이터를 '수집→저장→분석→추론' 할 수 있는 기술개발은 물론 기술 표준화가 우선적으로 추진되어야 할 것이다. 다섯째, 구조화되지 않은 대규모 데이터 속에서 숨겨진 정보를 찾아내는 데이터 사이언티스트(Data Scientist)의 양성이 필요할 것이다. 빅데이터시대에는 데이터를 관리하고 분석할 수 있는 인력이 매우 중요하다. 이미 글로벌 IT 업체에서는 데이터 사이언티스트에 대한 인재 확보와 역량 강화에 많은 노력을 경주하고 있다.[14] 따라서 교육부와 협력하여 보건복지 분야 데이터 사이언티스트를 양성할 수 있는 전략이 마련되어야 할 것이다.

끝으로 보건복지 빅데이터의 개인정보와 기밀정보에 대한 보안정책이 마련되어야 할 것이다. 보건복지 빅데이터는 개인에 대한 거의 모든 정보가 저장되어 있지만 아직 법·제도는 미비한 상황이며 논의조차 되지 못하고 있다. 빅데이터의 활용도 중요하지만 과다한 개인정보의 유출은 프라이버시 침해는 물론 사이버 인권침해나 범죄에 악용될 수 있다. 빅데이터로부터 개인을 보호하기 위해 가장 중요한 것은 특정 개인을 식별하지 못하도록 하는 익명화와 정보접근 및 처리에 대한 통제이다. 그러나 정보접근 및 처리에 대한 통제를 강하게 하면 정보 활용이 활성화되지 않기 때문에 보건복지 빅데이터 '활용과 보호의 균형'에 대한 효과적인 정책이 마련되어야 할 것이다.

【참고문헌】

고숙자·정영호(2012. 11), "국민건강 미래예측 시스템 구축방안",「보건복지포럼」, 통권 제193호, 한국보건사회연구원.

국가정보화전략위원회(2011. 11. 7.), "지식정보 개방과 협력으로 스마트 정부 구현".

김근태(2012), "빅데이터 분석을 위한 기업의 Big Analytics 환경변화",「정보처리학회지」, 제19권 제2호, pp. 70~78.

송태민·이상영·이기호·박대순·진달래·류시원·장상현(2011),「u-Health 현황과 정책과제」, 한국보건사회연구원.

13) 국가지식플랫폼은 1,068종 공공지식정보 가운데 351종을 2013년까지 민간에 전면 공개하고 이를 통해 창출되는 경제적 부가가치는 10조 7,000억 원에 달할 것으로 예측하고 있다.

14) 이베이는 고객데이터를 분석하고 가공하는 일을 맡은 직원만 5,000명에 이르고, EMC는 경제학, 통계학, 심리학 등을 전공한 박사급 인재들인 데이터 사이언티스트로 구성된 '애널리틱스' 랩을 운영하고 있으며, 미국 IBM은 사내 200명 이상의 수학자들이 '분석학'을 집중적으로 연구하고 있음. 미국에서는 2018년까지 14만~19만 명의 전문가와 150만 명 정도의 데이터 관리자와 분석 인력이 부족할 것으로 예측하고 있음(McKinsey, 2011).

송태민(2012. 11), “보건복지 빅데이터의 효율적 활용 방안”, 「보건복지포럼」, 통권 제193호, 한국보건사회연구원.

송태민(2013. 8), “소셜 빅데이터 분석을 통한 자살 검색 예측모형 개발”, 「보건복지포럼」, 통권 제202호, 한국보건사회연구원.

윤미영·권정은(2012), 「빅데이터로 진화하는 세상-빅데이터 글로벌 선진사례-」, 한국정보화진흥원 빅데이터 전략연구센터.

윤형중(2012), 「이제는 빅데이터 시대」, e비즈북스.

이성훈·이동우(2013. 2), “빅데이터의 국내·외 활용 고찰 및 시사점”, 「디지털정책연구」, 제11권 제2호, pp. 229~233.

장상현(2012), “빅데이터와 스마트교육”, 「한국정보과학회지」, 제30권 제6호, 통권 제277호, pp. 59~64.

전자통신기술연구소(2012), “보건의료 Big Data R&D 사업 기획 연구계획서”.

정지선(2011), “新가치창출 엔진, 빅데이터의 새로운 가능성과 대응 전략”, 「IT & Future Strategy」, 제18호.

정지선(2012), “성공적인 빅데이터 활용을 위한 3대요소: 자원, 기술, 인력”, 「IT & Future Strategy」, 제3호.

한국정보화진흥원(2012. 10. 29.), “대한민국 사회현안과 빅데이터 전략”, 제3차 빅데이터 국가전략 포럼.

함유근·채승병(2012), 「빅데이터 경영을 바꾸다」, 삼성경제연구소.

황승구·최완·허성진·장명길·이미영·박종열·원희선·김달(2013), 「빅데이터 플랫폼 전략」, 전자신문사.

IDC(2011), *Digital Universe study.*

McKinsey Global Institute(2011), *Big Data: The Next Frontier for Innovation, for Innovation, Competition, and Productivity*, McKinsey Inc.

Peter Warden(2011), *Big Data Glossary,* O'Reilly Media.

総務省(2012), 「平成 24 年度版 年度版 情報通信白書」.

공유자원포털(www.data.go.kr)

구글독감예보(www.google.org/flutrends)

구글검색트렌드(www.google.org/trends/)

샌프란시스코범죄예방(www.crimemapping.com)

유전자정보제공(www.1000genomes.org/)

일본총무성(www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/)

한국인체자원은행(kbn.cdc.go.kr/)

DNA Link(dnalink.com/)

The Economist(2010)(www.economist.com/node/15557443/)
Gartner(2011)(www.gartner.com/newsroom/id/1731916/)
Pillbox service(pillbox.nlm.nih.gov/)
RAHS(hsc.gov.sg)