보건의료 분야의 빅데이터 활용 동향

이지혜, 제미경, 조명지, 손현석
서울대학교

요 약

정보통신기술의 발전과 생산되는 데이터의 양적 증가에 따라 빅데이터에 대한 관심이 증대되고 있다. 빅데이터(Big Data)란 기존 데이터베이스의 데이터 저장 · 관리 · 분석 능력을 초과하는 다양한 형식을 가진 대량의 데이터를 의미한다. 여러 분야에서 빅데이터가 생성, 분석, 활용되고 있는데, 특히 보건의료 및 바이오 분야에서의 빅데이터 분석은 사회경제적으로 큰 영향력을 발휘할 수 있기 때문에 크게 주목 받고 있다. 본 연구에서는 보건의료 분야에서 생성되는 데이터의 특징과 빅데이터 분석 프로세스에 대해서 조사하였고, 국내 · 외 빅데이터 정책 및 활용사례를 분석하였다. 그리고 현재의 빅데이터 활용 장벽을 극복할 수 있는 몇 가지 전략을 제시하였다. 대량의 데이터에서 유용한 정보를 생성해내는 빅데이터 분석 기술은 보건의료 및 바이오 분야에서 국가경쟁력을 향상시키는 중요한 기반이 될 것이다.

Ⅰ. 서 론

우리나라는 최근 정보통신기술의 발전과 개인 · 정부 · 기업에 의해서 생산되는 데이터의 양이 증가되면서 빅데이터에 대한 관심이 증대되었다. 빅데이터(Big Data)란 기존에 없던 새로운 개념이 아니라 다양한 형식의 데이터가 축적되어 관계형 데이터베이스(Relational Database Management System, RDBMS)와 같은 기존의 데이터베이스로는 데이터의 저장 · 관리 · 분석의 허용 범위를 초과하는 대량의 데이터를 의미한다[1,2]. 데이터는 형식에 따라 정형, 반정형, 비정형으로 나뉠 수 있는데, 정형 데이터는 테이블의 고정된 필드에 체계적으로 저장될 수 있는 데이터를 의미하고, 반정형 데이터는 고정된 필드에 저장되진 않지만 스키마를 가지며 XML, HTML 텍스트 등이 포함된다[3]. 그리고 비정형 데이터는 고정된 필드 구조로 저장할 수 없는 텍스트, 이미지, 동영상, 음성 등의 데이터를 의미하고, 빅데이터에서 차지하는 비중이 크다[3]. 일반적으로 빅데이터는 3V인 큰 용량(Volume), 데이터 형식의 다양성(Variety), 빠른 속도(Velocity)를 주요 특성으로 가지고, 더 나아가 대량의 데이터에서 의미있는 정보를 분석하여 가치(Value)를 창출하는 특성을 가지고 있다[4][5]. 데이터의 양과 형식 측면 외에도 방대한 데이터를 분석하여 가치있는 정보를 추출하고 예측에 활용되는 기술도 또한 빅데이터 개념에 포함되어 사용되고 있다[6].

스마트 기기의 보급과 소셜 네트워크 서비스 사용의 증가로 인해 사용자의 위치 및 개인 정보들이 매일 웹 상에 저장되어 쌓이고 있으며, 이것이 빅데이터 생성의 주요 원인으로 알려지고 있다. 이를 통해 최근에는 스마트 폰 같은 개개인이 들고 다닐 수 있는 모바일 기기를 이용해서 개인 생체 데이터를 수집하고, 데이터 마이닝 기법을 적용하여 사용자에게 피드백을 제공해주는 u-Health 기술이 개발되고 있다[7]. 빅데이터 시대를 맞이하여 미국과 같은 선진국을 중심으로 공공데이터의 공개와 바이오 성과물 공유가 활발히 이루어지고 있으며, 바이오 시장경쟁력 향상을 위해서 국가적으로 빅데이터 기술 개발 및 활용을 촉진하기 위한 정책들을 추진하고 있다. 한편 컴퓨터 · 정보통신 · 통계학 등 다양한 분야에서 빅데이터 분석 프로세스에 적용하는 기술들이 개발되어 왔다. 대표적으로 방대한 비정형 데이터를 저장 · 분석 · 처리하기 위해서 하둡(Hadoop)과 같은 분산병렬처리 플랫폼과 비관계형 데이터베이스인 NoSQL이 개발되어 정형 데이터를 처리하는 데에 적합한 기존의 관계형 데이터베이스를 대체해왔다. 또한 대량의 데이터를 분석해서 특정 패턴이나 의미있는 정보를 추출하고, 예후나 영향을 예측할 수 있는 데이터 마이닝(Data Mining) 또는 기계 학습(Machine Learning)과 같은 통계적 기법들이 빅데이터 분야에서 널리 사용되고 있다.

빅데이터는 여러 분야에서 생성 및 활용되고 있는데, 특히 주목받고 있는 분야가 보건의료 분야이다. 인구 고령화에 따른 만성병 및 퇴행성 질환의 증가로 인해 보건의료 분야에서는 빅데이터를 의료비 절감, 전염병 예방, 의료 서비스의 질 향상에 활용하고자 다양한 연구들이 시도되고 있으며, 효율적인 진단 및

처치 방법의 탐색, 예후 예측 등에 효과적인 대안 제시방법이 되고 있다. 미국 의료분야에서 빅데이터를 활용 시 연 3,000억 달러의 경제 비용 절감 효과가 있을 것이라 예측되었다[2]. 의료기관에서는 의료기록 전자화로 인해 축적되고 있는 방대한 데이터를 효율적으로 저장 · 분석하고, 바이오 및 제약 업체에서는 실험 수행 전에 기존에 축적된 분자 정보를 기반으로 다양한 바이오 마커와 기계 학습 알고리즘을 이용해서 신약 및 치료 · 진단기술 개발에 빅데이터 기술을 도입하고 있다.

본 연구에서는 보건의료 분야에서 생성되는 데이터의 특징과 빅데이터 분석 프로세스에 대해서 알아보았고, 국내 · 외 빅데이터 정책 및 활용사례를 조사하였다. 먼저 국내는 각 부처별 관리 현황을 중심으로 살펴본 후, 미국, 유럽, 일본 등 국외 빅데이터 활용 현황을 살펴보았다. 특히 보건의료 및 바이오 분야에서 생성되는 환자의 진료기록, 분자생물학 데이터 등을 이용한 빅데이터 관련 연구들을 살펴보았다. 대량의 데이터에서 유용한 정보를 생성해내는 이러한 빅데이터 분석 기술은 보건의료 및 바이오 분야에서 국가경쟁력을 향상시키는 중요한 기반이 될 것이다.

Ⅱ. 본 론

2.1 보건의료 데이터의 특징 및 빅데이터 분석 프로세스

보건의료 분야에서는 생성되는 데이터는 대부분 개인정보를 포함하고 있어서 기밀유지가 필수적으로 요구된다. 특히 개인 진료 관련 데이터는 무분별한 공개 시 프라이버시 침해나 범죄에 악용될 수 있고, 사회경제적으로 개인에게 미치는 영향이 크기 때문에 보건의료 빅데이터 활용의 장벽으로 작용하고 있다[8]. 현재 규모가 큰 의료기관은 의료기록 전자화가 많이 진행되었으며, X-ray · CT 사진, 의료영상, 검체 시료 등 많은 비정형 데이터들을 소유하고 있다. 또한 정부 및 의료기관 별 진료 및 연구데이터가 분산되어 있어서 공유가 어렵고, 방대한 양의 데이터를 저장 · 관리하기 위해서 빅데이터 처리 기술의 도입을 필수적으로 요구하게 되었다.

빅데이터 분석 프로세스는 데이터의 수집, 저장, 처리, 분석, 표현의 5단계로 구성된다[9]. 첫 번째, 빅데이터 수집 단계에서는 데이터를 수동으로 입력하거나, 인터넷 링크를 통해 방문한 웹 페이지를 수집하는 크롤링(Crawling) 검색 엔진 로봇을 이용하거나, 소스 데이터에서 필요한 데이터만을 추출하여 변화시켜 저장 또는 전송하는 ETL(Extraction, Transformation, Loading)을 통해 수집할 수 있다[9,10]. 데이터로는 데이터베이스에 저장되어 있는 정형 데이터나 인터넷에서 비정형 외부 데이터들이 수집된다. 데이터 표현 방식으로는 대표적으로 JavaScript의 문법을 사용하여 기존의 XML과 비슷하지만 텍스트 기반의 웹 데이터를 정형화해서 표현하는 방식인 JSON(JavaScript Object Notation)과 JSON의 이진표현 방식인 BSON(Binary javaScript Object Notation)이 있다[9]. 수집 기법으로는 로그 데이터를 수집 및 분석하는 Chukwa와 Scribe가 있고, SQOOP(SQl-to-hadOOP)은 JDBC(Java DataBase Connectivity)를 이용하여 RDBMS(MySQL, 오라클 등)와 NoSQL(하둡 분산 파일 시스템인 Hive, HBase 등)간의 데이터 연동에 사용된다[9].

두 번째, 빅데이터의 저장 단계에서는 다양한 형식을 가진 대량의 데이터를 효율적으로 저장하기 위해서 분산 파일 시스템, NoSQL(Not-only SQL) 데이터베이스, 병렬 DBMS 등이 사용되고 있다[9]. 분산 파일 시스템은 컴퓨터 네트워크로 공유된 여러 컴퓨터에 접근할 수 있는 파일 시스템으로 하둡 분산 파일 시스템인 HDFS(Hadoop Distributed File System)가 있다[11]. NoSQL은 SQL를 사용하지 않는 DBMS로 비관계형 데이터베이스를 말하고, 웹 환경의 다양한 정보를 저장하기 위해서 데이터 저장 공간의 수평적 확장이 가능하며, 데이터 모델에 따라 〈key, value〉 저장 구조, 문서(Document) 저장 구조, 열(Column) 저장 구조로 나뉜다[9]. 〈key, value〉 저장 구조를 가진 데이터베이스로는 아마존에서 개발한 DynamoDB(http://aws.amazon.com/ko/dynamodb/)가 있다. DynamoDB의 속성은 〈key, value〉 쌍으로 구성되고, key는 문자열만 가질 수 있고, value는 문자열, 숫자 등 다양한 값을 가질 수 있다. 여러 개의 속성이 모여서 항목(record)을 구성하고, 여러 개의 항목으로 테이블(table)이 구성된다. 문서 저장 구조를 가지는 데이터베이스의 예로, MongoDB(http://www.mongodb.org/)는 문서를 단위로 가지고, 각 문서들은 컬렉션에 모이며, 각 컬렉션은 데이터베이스에서 관리한다. MongoDB는 대량의 데이터를 샤딩(sharding)을 통해서 여러 개의 MongoDB에 데이터를 나누어 저장하여 분산 확장이 가능하다. 열 저장 구조를 가지는 NoSQL에는 구글의 빅테이블을 기반으로 개발된 오픈 소스 비관계형 데이터베이스인 HBase가 있다(http://hbase.apache.org/). HBase는 하나의 마스터와 여러 개의 리전 서버로 구성된다. HBase에서 테이블은 행(row), 컬럼(column), 타임스탬프(timestamp)로 구성되고, 로우키, 컬럼패밀리, 타임스탬프로 구성된 행들의 집합을 리전이라고 한다. 리전들의 집합이 테이블이며, 로우키와 컬럼으로 정렬이 되고, 타임스탬프는 작은 값일수록 최신 업데이트 값을 의미한다. HBase는 HDFS 상의 데이터 위에서 동작하고, ZooKeeper가 전반적으로 클러스터의

노드들을 관리한다. 병렬 DBMS는 다수의 프로세서를 사용해서 여러 디스크에서 동시에 질의를 수행하는 데이터베이스 시스템을 말한다[11].

세 번째, 빅데이터의 처리 단계에서 사용되는 기법으로는 비정형 데이터 분석을 위해서 구글에서 개발된 대량의 데이터를 여러 컴퓨터에 나눠서 저장하고, 나눠져서 처리한 결과들을 통합하여 제공하는 분산처리 구조인 맵리듀스(MapReduce)가 있다[10,11]. 맵리듀스는 기본적으로 맵 함수(Mapper)와 리듀스 함수(Reducer)의 입력과 출력 값으로 〈key, value〉 쌍이 사용된다. 맵리듀스 메커니즘을 사용해서 야후에서 오픈 소스 대용량 데이터 처리 분석을 지원하는 프레임워크인 하둡을 개발하였고, 최근 정부 및 연구기관에서 널리 사용되고 있다. 하둡은 맵리듀스, HDFS, HBase로 구성되어 있다[12]. 하둡은 네임 노드와 데이터 노드들로 구성된 클러스터 서버에서 운용되는데, HDFS가 대용량 파일을 데이터 노드의 여러 블록으로 나눠서 분산 저장 및 처리하고, 네임 노드에는 데이터 노드의 블록의 메타 정보를 관리하는 모니터링 시스템이 있다.

네 번째, 빅데이터의 분석 단계에서는 방대한 데이터에서 데이터 간의 내재된 관계를 탐색하여 의미있는 정보를 발견하는 데이터 마이닝과 기계 학습 기법과 같은 통계학 기법, Python과 같은 프로그래밍 언어, 데이터 마이닝을 구현할 수 있는 통계 패키지 R 등이 사용된다. 데이터 마이닝에는 클래스가 알려진 훈련 데이터 셋을 학습시켜 새로운 데이터의 클래스를 예측하는 분류(classification)와 클래스 정보를 모르고 유사성을 기반으로 데이터를 그룹 짓는 방법인 군집화(clustering) 기법이 있다. 분류에는 KNN(K-nearest neighbor), 판별분석이 있으며, 군집화 기법에는 각 데이터들 간의 거리를 계산해서 분할 합병해가는 계층적 군집화(Hierarchical Clustering), 유클리디안 거리를 이용해서 반복과정을 통해 k개의 분할 영역을 결정하는 K-means, 데이터의 변동을 잘 설명하는 주성분으로 차원을 축소시키는 주성분 분석(Principal Component Analysis)이 있다. 기계 학습 방법으로는 결정 트리(Decision Tree), 신경망(Neural Network), 은닉 마코프 모델(Hidden Markov Model), SVM(Support Vector Machine) 등이 있으며, 학습과 에러 수정과정을 반복적으로 수행해서 최적의 모델을 탐색할 수 있다. 또한 시계열 분석이나 변수들 간의 영향이나 관계를 분석하는 회귀분석(Regression Analysis)이 분석 단계에서 사용될 수 있다. 대표적인 분석 프로그램으로는 Boilerpipe, WEKA, Mahout이 있다[9]. Biolerpipe는 HTML 웹 페이지에서 불필요한 부분을 제거하고 의미있는 텍스트만을 파싱하는 Java 기반의 프레임워크이고, WEKA는 Java 기반의 오픈 소스 기계 학습 분석 프로그램으로 다양한 알고리즘을 간단한 인터페이스로 제공한다. Mahout은 하둡과 연동해서 기계 학습 알고리즘을 실행할 수 있는 오픈 소스 프레임워크이다.

다섯 째, 빅데이터의 표현 단계에서는 데이터 분석 결과를 이해하기 쉽도록 효과적으로 전달하기 위해 결과 데이터를 표나 그래프로 시각화하는 기술이 포함된다. 대표적인 프로그램으로 Fusion Tables, Tag Cloud, Tableau가 있다. 구글에서 개발한 Fusion Tables은 테이블에 저장된 데이터를 구글 맵과 연동해서 효과적으로 시각화할 수 있으며(tables.googlelabs.com/), Tag Cloud는 수집된 태그의 중요도, 빈도, 인기 등에 따라 글자의 색상 크기를 달리하여 웹 페이지나 이미지로 제공해준다[13]. Tableau은 피벗 테이블이나 크로스 테이블을 작성해서 시각화하는 데에 유용하다(http://www.tableausoftware.com/). 또한 네트워크를 생성해서 다른 노드에 얼마나 강한 영향을 주는지를 나타내는 중심성, 네트워크 응집도 등의 특성을 기반으로 네트워크에서 의미있는 정보를 해석할 수 있다[14].

2.2 국내 빅데이터 정책과 각 부처별 빅데이터 관리 현황

국내 빅데이터 정책은 공공데이터의 개방측면에서 정부 1.0을 시작으로 현재의 정부 3.0에 이르렀으며, 그 어느 때보다 활발히 추진되고 있다. 정부 1.0이 정부 중심의 일방향 서비스를 제공하였다면, 정부 3.0에서는 국민 개개인을 중심으로 능동적인 공개 · 참여 · 개방 · 공유 · 소통 · 협력을 통한 양방향 · 맞춤형 서비스를 제공함으로써 데이터 분석을 통한 융합지식을 창출하고자 하였다[15]. 이처럼 공공정보를 적극 개방 · 공유하고자 하는 정부 3.0은 정부와 국민간의 소통과 협력을 확대하고, 국민 개개인의 행복에 초점을 두며, 부처간 칸막이를 뛰어넘는 통합형 정부운영과 민간의 능동적 참여를 지향하고 있다. 이는 기존의 방식으로 풀기 어려운 복잡한 사회문제의 대두, 지식정보사회로의 전환에 따른 정부와 국민간의 관계 변화, 그리고 지식과 기술의 융 · 복합 혁명이 새로운 기회요인으로 등장함과 같은 배경에서 기인한다. 이를 위해 비공개 정보를 최소화하고 모든 정보는 공개함을 원칙으로 하며, 공개 문서는 생산 즉시 정보공개시스템에 이관되어 원문까지 공개될 수 있도록 하는 원문정보공개시스템을 구축하였다[16]. 정부는 공공 데이터 개방 정책을 통해 2015년까지 복지 사각지대 해소와 데이터 기반 정책 결정 제도화와 같은 성과를 가시화하고, 2017년까지 확산 및 정착시키며, 2018년 이후에는 이를 내재화시키는 것을 비전으로 하고 있다. 이러한 목표를 달성하기 위해서는 정부 기관들 간의 상호협력이 중요하며, 여기서 주요 기관들의 역할은 다음과 같다. 정부 3.0 추진위원회에서는 정부 3.0의 방향 및 전략 설정, 과제 우선순위를 결정하고, 행정자치부에서는 과제별 작

업그룹 지원과 자치단체 및 지방 공공기관 확산, 국무조정실에서는 부처간 갈등과제 조정 및 미이행 기관 독려, 기획재정부에서는 예산 지원, 미래창조과학부에서는 창조경제 및 국가정보화 연계 지원, 법제처에서는 과제 추진을 위한 법령 개선 지원, 문화체육관광부에서는 여론조사 지원과 선도그룹 활동 및 과제 성과 홍보, 전문기술연구단에서는 빅데이터 · 클라우드 · 보안 등 전문 인력 지원을 뒷받침 한다[17]. 또한 제도적 지원을 위해 2014년 11월, 공공기관의 정보공개에 관한 법률이 개정되어 시행되고 있으며[18], 현재 정부는 중앙 부처 및 지자체의 공공정보를 일원화된 정보 제공 창구인 공공데이터 포털(https://www.data.go. kr)을 통해 11,816개의 파일데이터와 1,618개의 오픈 API, 251개의 시각화 데이터의 형식으로 제공하고 있다. 이 중 보건의료 분야의 데이터는 파일데이터 형식으로 860건, 오픈 API 형식으로 83건, 시각화된 데이터로 23건이 해당되며, 그림 1에서 자세히 살펴볼 수 있다.

현재 44개의 중앙행정기관과 17개의 지방자치단체가 정부 3.0의 공공정보와 데이터 개방에 참여하고 있으며, 몇몇 주요 기관의 데이터 관리 현황과 전략을 살펴보면 다음과 같다(표 1). 미래창조과학부에서는 과학기술 통계, 연구장비정보, 방송산업 실태정보 등 전체 128개 공공데이터를 개방 중에 있으며, 향후 기술산업정보, 전국발명품정보 등으로 확대하여 개방할 계획이다. 정보공개는 문서, 녹음테이프, 녹화테이프, 영화 필름, 마이크로 필름, 사진, 컴퓨터 처리정보로 구분되며 대상에 따라 공개방법은 열람 또는 사본의 교부 등으로 이루어진다. 행정자치부에서는 주민등록 인구통계 정보, 지방공기업 경영공시정보, 안전시설물 정보(치안 및 비상대피시설) 등 전체 133개 공공데이터를 개방 중에 있으며, 향후 지방행정평가 DB, 국가기록원의 영구기록물 DB를 제공할 예정이다.

그 밖에 보건의료 분야와 관련된 부처인 보건복지부에서는 어린이집 정보, 통계포털 DB, 보건의료분야 R&D 사업 현황정보, 보건기관정보 DB 등 전체 136개 공공데이터를, 그리고 식품의약품안전처에서는 의약품 DB, 의료기기 품목허가정보, 식 · 의약 통계 정보 등 전체 75개 공공데이터를 개방 중에 있으며, 향후 보건서비스현황 정보, 재활원 진료정보, 식품원재료 정보, 약학정보, 희귀의약품 정보 등으로 개방 확대할 계획이다. 특히, 보건복지부에서 제공하는 국가건강 정보포털, 응급의료정보 제공, 노후준비지표와 같은 다양한 콘텐츠로 구성되어 있는 모바일앱은 국민 누구나 보다 쉽게 양질의 건강 정보를 다룰 수 있도록 한다.

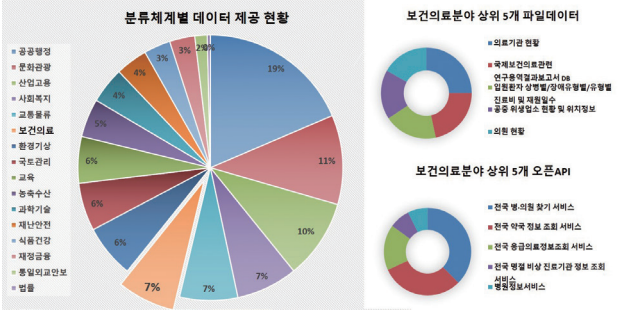

그림 1. 분류체계 별 공공 데이터 현황과 보건의료분야 상위 5개 데이터

2.3 국외 각 나라별 빅데이터 정책들

미국은 국가가 당면한 문제들에 대한 해결방안을 마련하기 위해 빅데이터의 활용을 통한 공공개혁 추진정책을 시행하고 있다.

2012년 3월, 미국 정부는 '빅데이터 R&D 이니셔티브(Big Data R&D Initiative)'를 발표하여 정부기관이 공공 정보를 개방하고 기존의 빅데이터를 활용하여 투명하고 효율적이며 혁신적인 정보 및 서비스를 제공하겠다는 계획을 발표하였다[19]. 빅데이터 연구개발 이니셔티브는 빅데이터 핵심기술 확보 및 첨단화, 사회 각 영역에의 활용, 전문인력의 양성의 3가지 측면을 중점적으로 추진함으로써 전략적 빅데이터 정책을 수행하고 있다[19]. 미국의 주요부처인 국방부, 국방부 산하의 방위고등연구계획국, 에너지부, 미국항공우주국, 미국해양대기관리처, 지질조사원, 국립보건원, 국립과학재단 등에서는 각 부처별로 빅데이터 관련 정책 및 프로젝트를 추진하고 있는데[20], 특히 국립보건원의 경우 게놈 프로젝트의 결과로 파생된 방대한 생물학적 데이터의 공개를 통해 신경과학, 생리학, 화학, 분자세포생물학, 임상학, 보건의료학 등의 관련 데이터 수집을 통한 연구 및 핵심기술 개발을 추진하고 있다[21]. 미국 연방정부는

표 1. 정부 3.0에 참여하는 주요 중앙행정기관의 빅데이터 관리 현황

기관명	현재 개방중인 정보	향후 개방 확대 정보	데이터 수
미래창조과학부	과학기술 통계, 연구장비정보, 방송산업 실태정보 등	기술산업 정보, 전국 발명품 정보 등	파일데이터 (116) 오픈API(12)
행정자치부	주민등록 인구통계 정보, 지방공기업 경영공시정보, 안전시설물 정보 등	지방행정평가 DB, 국가기록원의 영구기록물 DB 등	파일데이터 (108) 오픈API(25)
보건복지부	어린이집 정보, 통계포털 DB, 보건의료분야 R&D사업 현황정보, 보건기관정보 DB 등	보건서비스현황 정보, 재활원 진료정보 등	파일데이터 (135) 오픈API(1)
식품의약안전처	의약품 DB, 의료기기 품목허가정보, 식 · 의약 통계 정보 등	식품원재료 정보, 약학정보, 희귀의약품 정보 등	파일데이터 (43) 오픈API(32)

각 부처와 기관 및 지역정부의 공공정보를 제공하여 국가 공공정보의 접근성과 활용도를 촉진하기 위하여 다양한 공공데이터를 제공하는 포털(data.gov)을 구축하였다[22]. 미국은 이와 같은 정부 주도 하의 빅데이터 기반 공공 정책의 추진뿐만 아니라, 민간부분에 있어서도 전 세계적으로 빅데이터 시장을 주도하고 있는데, 대표적으로 Google, Facebook 등의 글로벌 인터넷기업들은 빅데이터에 기반한 마케팅전략을 활용하고 있으며 주요 빅데이터 장비 및 소프트웨어의 개발도 미국의 주도 하에 이루어지고 있다[20]. 미국의 경우 이처럼 정부와 민간부분 모두에 있어 정보 활용범위가 점차 확대됨에 따라 정보의 활용을 극대화하여 다양한 분야에서의 빅데이터의 활용이 점차 확대될 것으로 생각된다.

유럽의 경우 미국과는 달리 민간부분에 있어서는 금융 영역에 제한적으로 빅데이터의 활용이 이루어지고 있는 수준이다. 그러나 공공정보에 대해서는 개인의 접근권과 사용권을 허가하는 입장을 취하여 공공부문 데이터의 개방과 공유에 대해서는 적극적인 정책을 진행하고 있는 실정이다[20]. 2003년에 '공공 분야 정보의 재사용에 관한 지침'을 제정하고 이를 2011월 12월에 개정함으로써 공공기관의 데이터 개방 및 활용을 확장하였으며[20], '데이터 개방전략(Open Data Strategy)'을 채택하여 유럽 데이터 단일 포털(open-data.europa.eu)을 개설하였다. 이러한 유럽의 공공데이터 개방정책은 신규 사업의 창출 및 정부 행정의 투명성 및 효율성 재고를 통해 큰 사회적 · 경제적 효과를 가져올 것으로 전망되고 있다. 영국 정부는 사회와 경제 성장을 위한 새로운 원료로써의 빅데이터를 활용하기 위해 공공정보의 공유 및 활용을 위한 데이터 공유 중심의 정책을 추진하는 등 EU국가 중에서 가장 선도적인 공공데이터 정책을 펼치고 있으며, 데이터 접근성과 신뢰성을 확보하고 이를 적극적으로 활용하여 투명하고 신뢰성 있는 사회를 구현하는 것을 빅데이터 전략 및 추진 방향으로 삼고 있다[19][22]. 이에 따라 교통부, 국세청, 문화매체 체육부, 기업혁신 기술부, 노동 연금부, 보건부, 외무부 등 주요 부처별로 수집된 데이터를 토대로 빅데이터를 분류하는 작업을 추진하여 정보를 공개하는 오픈 데이터베이스(data.gov.uk)가 구축되었다. 각 주요 부처 별로 관련 데이터들을 공개하고 있으며 특히 보건부에서는 의사업무 관련 일반데이터, 암치료 데이터, 국민보건 서비스 등으로 데이터를 구분하여 각 항목별로 세부 데이터를 제공하고 있어 보건의료 분야에의 활용이 확대되고 있다.

일본의 경우 미국이나 유럽에 비하여 민간부분에서는 전반적으로 빅데이터의 활용도가 낮은 편이며, 주로 사회현안을 해결하고 국가 경쟁력을 재고하기 위한 정부의 정책에 빅데이터의 활용도가 높아지고 있다[19]. 일본 정부는 2000년대에 이르러 시작된 국가 경제침체 및 성장저하와 2011년 동일본 대지진 등 국가 위기 상황의 극복을 위한 종합적 진흥정책으로 'Active Japan ICT'계획을 발표 및 추진하였다. 이 계획에는 총 5개 추진전략이 포함되는데 그 중 'Active Data'부문이 빅데이터를 활용하여 사회 · 경제 성장 목표를 달성하고자 하는 계획에 해당하며 빅데이터를 수집 및 분석하여 국가 재난을 비롯한 국가가 당면한 다방면의 문제 해결에 이용하고자 한 것이다[20]. 이와 같은 'Active Data'정책은 궁극적으로는 공공 분야뿐 아니라 민간 분야까지 포함하여 빅데이터의 활용을 활성화하기 위한 것으로 데이터 개방과 활용을 위한 환경마련, 데이터과학자 육성을 비롯한 7대 추진과제를 목표로 하여 전략을 추진함으로써 강력한 빅데이터 정책 수립에의 박차를 가하고 있다[20][21]. 〈표 2〉는 국외 주요국의 빅데이터 정책의 추진전략 및 정책 시스템을 요약한 것이다.

표 2. 각 주요국의 빅데이터 정책 추진전략 및 시스템

구분	미국	영국	일본
정책 및 전략	Big data R&D Initiative	Open data strategy	Active data
의사결정기관	과학기술 정책실	내각 사무처	총무성
전담기구	국가과학기술위원회	기업혁신기술부	정보통신심의회
공공데이터포털	data.gov	data.gov.uk	openlabs.go.jp
민간부문	전 세계적으로 빅데이터 시장 주도, 대표적인 글로벌 업체인 Google, Facebook 해당	금융시장영역에 제한적 활용	일부 활용사례 있으나 전반적으로 활용도 낮음
개인정보보호 관련법	개인정보보호와 상업적 이용간 균형을 지향	개인정보보호 기준을 엄격하게 준수	개인정보보호 기준을 엄격하게 준수

2.4 전자의무기록과 빅데이터 기술의 접목

전자의무기록(Electronic Medical Record)이란 의료기관에서 발생하는 모든 형태의 의료정보를 기존의 문서화 방식에 정보통신 기술을 접목하여 전산화한 의료정보시스템을 말한다[23,24]. 국민건강보험공단에서는 2011년 구축된 국민건강정보 DB를 기반으로 표본 코호트 DB인 NHID-cohort를 구축하여 유료로 제공하는데, 이는 전자의무기록을 관리하는 DB의 개인정보 보호 문제와 연구자들의 접근성 문제를 보완하여 학술용으로 제공하고자 구축되었다. 즉, 원자료에 준하는 결과물 산출을 바탕으로 국민 건강 향상 및 복지의 선진화를 위한 학술연구 지원을 목적으로 한다. 구축된 표본 코호트 DB는 2002년에서 2010년

까지의 약 100만명을 대상으로 한 환자 중심 코호트 자료로 구성되어 있으며, 진료 DB, 건강검진 DB, 요양기관 DB와 연동하여 다양한 주제의 연구에 활용할 수 있도록 데이터를 제공한다[25]. 연세대학교 의료원과 KT 합작사에서 제공하는 후(H∞H) 헬스 케어 시스템도 전자의무기록 시스템을 활용한 것으로 정보통신 기술의 발전에 힘입어 전자진료기록부, 의료영상 전송 등의 대용량 보건의료 데이터를 효율적으로 활용할 수 있도록 하는 진료 시스템이다. 이는 병원에서 이루어지는 진료, 수술을 포함한 의료행위 및 진료예약, 수납, 처방, 약제 관리 등에 대한 정보를 언제 어디서나 이용할 수 있도록 접근성을 높여 보건의료 서비스 질 향상에 이바지하고자 하는 것이다[26]. 국립암센터에서도 2010년 전자의무기록 기반의 통합의료정보시스템을 구축하여 고품질의 의료서비스와 효율적 업무를 도모하고자 하였으며[27], 2011년 9월부터 본격적으로 시행된 개인정보보호법에 의해 의료기관의 개인정보보호에 대한 중요성이 높아짐에 따라 DB 암호화 기술을 강화시켜 내부자료 유출방지와 보안관제 체계 구축 등을 위한 단계별 사업을 추진한 바 있다[28].

국외의 경우, 미국의 건강보험회사 웰포인트는 의료진의 진단과 환자의 치료에 슈퍼컴퓨팅 기술을 이용하는 시스템을 도입하여 진단 및 치료율을 향상시킴으로써 효율적인 환자치료 방식을 제시하고 있다[29]. IBM의 왓슨솔루션을 도입함으로써 보유하고 있는 건강보험 자료들과 환자에 대한 정보를 통합 및 분석하여 치료법을 검색하고 다른 의사들이 환자의 진단과 치료에 사용할 수 있는 어플리케이션의 제공을 추진하고 있다. 이를 위해 환자의 증상과 상담 및 진료 결과, 진단 결과 등 전자의무기록 정보를 활용하고, 임상실험 결과와 다양한 치료사례 등과 같은 데이터 또한 활용하여 IBM의 서버를 통해 정확한 진단 및 치료를 위한 가이드라인의 제시를 목표로 하고 있다[29][30]. 이를 통하여 진료 및 치료를 위한 불필요한 시간과 노력을 줄이고, 환자와 건강보험회사의 경제적 낭비를 줄이는 효과를 기대하고 있으며 환자 중심의 의료시장으로의 도약을 위한 발판으로서의 역할이 기대되고 있다. 미국의 eMERGE Network ACC(Administrative Coordinating Center)는 유전과학 분야에서의 진보적 발견을 위하여 전자의무기록 시스템에 발 맞춘 DNA 저장소의 활용을 탐색하기 위한 5대 기관들의 컨소시엄으로서 국립인간유전체연구센터(National Human Genome Research Institute, NHGRI)의 지원 하에 다양한 빅데이터 기반 연구를 수행하고 있다. 인구집단의 다양성과 GWAS 데이터들의 통합을 통한 네트워크의 확장을 목표로 이를 의학치료에 활용하기 위한 데이터의 통합과 EMR 데이터 처리를 위한 새로운 알고리즘 배치, 대용량의 GWAS 데이터의 메타분석 등을 수행하고 있다. 이와 같은 선진 의학 정보들과 유전체 과학, 공동회담 등을 결합시킴으로써 eMERGE 프로젝트는 데이터 기반의 접근방식의 발전을 통해 유전체 정보를 일상적인 건강관리 체계에 통합시키는 첫 번째 단계를 제시하여 주고 있다[31]. 캐나다의 온타리오 공과대학병원에서는 인큐베이터 내의 미숙아로부터 얻은 다양한 데이터를 분석하여 병원균의 감염을 예측하고 감염징후를 조기 발견하여 미숙아 및 의사 전달이 어려운 환자들을 위한 진단 및 치료 시스템을 구축하였다. 인큐베이터 내의 미숙아를 모니터링 함으로써 이로부터 생성되는 생리학적 데이터들을 분석하며 혈압, 체온, 심장박동 및 호흡을 통해 얻은 데이터들을 바탕으로 감염의 징후를 판단하여 신속하게 감염여부를 판단하고 조기치료를 가능하도록 하였다. 기업이 지원하는 기술과 소프트웨어를 이용하여 병원으로부터 얻은 실제 데이터를 대학기관에서 분석하고 이를 다시 병원의 환자 진단과 치료에 이용하는 다중 협력 시스템을 구축함으로써 효율적인 '스마트 헬스케어'시스템을 추진하고 있으며 이는 미숙아의 병원감염 방지부터 다른 질병에의 적용에까지 점차 확대될 예정이다[32]. 'MRSA and C diff'는 영국에서 개발된 유료 모바일 애플리케이션으로서 기본적으로 세균 감염자 및 병원에 대한 정보를 제공한다. 명칭에서 알 수 있듯이, 슈퍼박테리아균이라고도 불리는 메티실린 내성 황색포도알균(methicillin-resistant staphylococcus aureus, MRSA) 감염과 클로스트리듐 디피실리균(clostridium difficile, C.Diff)에 의한 감염에 관한 정보를 제공한다. 이 모바일 앱에서는 모든 영국 NHS(National Health Service) 병원의 슈퍼박테리아 감염자 수에 대한 최신의 공식적인 수치데이터를 그래픽으로 제공하며 이 수치를 감염자 수에 따라 랭킹화하여 목록으로 제시하여 줌으로써, 세균 감염에 대한 신속한 정보의 획득 및 이용이 가능하도록 하며 감염이 활발한 지역이 어디인지를 빠르게 파악할 수 있도록 한다[33].

2.5 보건의료 및 바이오 분야에서 빅데이터 기술의 활용 사례

미국 국립 암 센터가 주도하고, 보건복지부가 국내 인체자원 종합관리 주관기관으로 지정되어 진행되고 있는 한국인체자원은행사업(http://kbn.cdc.go.kr)은 인체자원의 체계적 수집을 위한 글로벌 사업이다. 인체자원이란 공여자로부터 기증받은 DNA, 조직, 혈액, 소변 등의 인체유래물과 임상정보(진단명, 수술명, 병리조직검사결과, 혈액검사 등), 역학정보(동의서정보, 성별, 생년월일, 음주력, 흡연력 등) 및 유전정보(SNP, CNV, Exome 등)를 보건의료연구에 활용할 수 있도록 자원화 한 것을 말한다. 현재, 한국인유전체역학조사사업(2001~2012.12월) 및 국민건강영양조사사업(1998~2012.12

월)을 통하여 전국 16개 병원으로부터 36만 명의 인체자원 정보가 확보되었으며, 이를 연구자들에게 심의를 거쳐 무상으로 분양하여 병원 별로 특성화된 질환군에 대한 다양한 질병연구, 바이오 마커 개발 등의 연구를 수행할 수 있도록 하였다. 2013년 12월 통계에 따르면, 인체자원 분양현황은 DNA가 전체 85%를 차지하고, 그 다음이 혈청으로 10%, 혈장이 3%, 기타물질이 2%를 차지하는 것으로 나타났다. 이러한 한국인체자원은행사업을 통해 맞춤의료 실현 및 혁신적 신약개발과 같은 보다 발전된 보건의료서비스 제공을 기대할 수 있을 것으로 보인다[34]. 약학정보원(http://www.health.kr)에서는 의약품 정보의 표준화에 기여하기 위한 웹사이트를 제공하는데, 이 웹사이트는 의약품정보(의약품정보, 식별표시정보, 상호 작용정보, 한약정보, 의약외품정보), 성분정보(성분처방정보, KPIC 약효분류, 성분동의어 정보), 안정성정보, 낱알식별표시제도, 복약 정보(제품별 · 질환별 · 제형별 복약정보, 임신 중 약물투여, 복약지도 픽토그램), 학술정보(팜리뷰, 해외의약뉴스, 예방접종 정보, 질병 · 기타정보, 의약품 등 분류번호)와 같은 의약품 관련 정보를 광범위하게 이용할 수 있도록 구성되어 있다. 약학정보원에서는 이와 같은 의약품 전문 정보 검색 서비스를 제공함으로써 국민들의 올바른 약 사용에 이바지 하고자 하였다. 특히, 의약품 검색 시 해당 약품의 상세정보뿐만 아니라 복약정보와 동일 주성분 코드를 가지고 있는 대체 의약품의 정보를 함께 제공하여 의약품 지식이 없는 일반 국민들의 이용에 어려움이 없도록 한다. 2015년 1월 현재 등록된 의약품의 수는 44,500여개, 식별표시데이터는 16,700여개 이다. 최근에는 보건의료 빅데이터에 소셜 미디어정보(SNS)를 접목하여 웹사이트나 모바일앱 개발에 활용하기도 하는데, 대표적인 사례로 국민건강보험공단에서 제공하는 국민건강주의 알람서비스(http://forecast.nhis.or.kr)가 있다. 이 웹사이트는 국민건강정보 DB와 SNS 데이터를 연계하여 질병 연관 키워드, SNS 주요 반응에 대한 내역을 일자 별로 집계하여 확인할 수 있도록 한다. 현재는 감기(인플루엔자), 눈병, 식중독, 피부염과 같은 주요 유행성 질병에 대한 위험도를 지역과 연령에 따라 확인할 수 있도록 서비스를 제공하고 있으며, 향후 데이터 수집의 다양성과 예측의 정확도를 향상시켜 단계적으로 관리 대상 질병의 수를 확대하여 제공할 예정이다. 이러한 서비스를 제공함으로써 해당 질병이 확산되기 전에 예방 및 치료정보를 국민들이 제공받을 수 있도록 한다. 보건의료 분야에서 개발된 모바일 앱을 몇 가지 살펴보면, 보건복지부에서 제공하는 것으로는 국가건강정보포털, 금연 길라잡이, 스모크프리, 응급의료정보제공, 치매 체크, 노후준비지표, 집으로, e하늘 장사정보, 아이사랑보육포털이 있으며, 소속기관인 질병관리본부와 기타 산하기관에서 제공하는 것으로는 결핵ZERO, 예방접종도우미, 국립암센터 의약품집, 암성통증관리, 진료비 확인 등이 서비스되고 있다. 이 중 예방접종 도우미(https://nip.cdc.go.kr)에 대해서 살펴보면, 이는 웹사이트와 모바일앱으로 개발되었으며, 자녀의 예방접종 기록을 관리하고, 일정에 맞추어 알람 서비스를 받을 수 있도록 한다. 개발에는 보건복지부 산하 질병관리본부의 예방접종 위탁 의료기관의 현황 데이터가 활용되었으며, 예방접종 전 · 후 주의사항과 백신정보, 표준 예방접종 일정표와 같은 정보를 제공하여 전문적인 정보를 보다 쉽게 이용할 수 있도록 서비스한다.

전세계적으로 가장 많이 사용되는 바이오분야의 데이터베이스로는 미국의 NCBI, 유럽의 EMBL, 일본의 DDBJ가 있다. 이들 데이터베이스는 국제 뉴클레오티드 서열 데이터베이스 협력(International Nucleotide Sequence Database Collaboration)을 통해서 주기적으로 데이터가 교환되어 통합된 데이터를 제공하고 있다. NCBI 데이터베이스에는 서열, 구조, 유전체, 발현 등 세부 데이터베이스를 포함되어 있으며, 이들 세부 데이터베이스들은 데이터들이 서로 연결되어 있어서 효율적인 분석 시스템을 제공하고 있다. 그 외에 미국, 영국, 일본 정부에서 공동 운영 중인 단백질 데이터은행(Protein Databank, PDB)은 실험 분석으로 생성된 단백질 서열, 구조 및 기능 관련 데이터가 저장되어 있다. 미국은 모든 분야에서 빅데이터를 가장 많이 활용하고 있으며 보건 · 의료 분야에서도 역시 빅데이터의 활용성이 가장 높은 국가 중 하나이다. 다양한 정부부처에서 빅데이터 활용정책이 추진되고 있으며, 특히 국립보건원에서는 국립 암 연구소, 국립 심장 · 폐 · 혈액 연구소, 국립 생의학 영상 및 생체공학 연구소에서 각각 빅데이터 프로그램을 실시하고 있다. 특히 국립 암 연구소(National Cancer Institute, http://www.cancer.gov/)에서는 현재 및 과거의 미국을 비롯한 전세계적인 특정 암의 발생률, 사망률 등을 포함한 통계적 수치를 제공할 뿐만 아니라 이러한 통계자료들을 통해 생성되는 연령, 성, 인종, 지역, 시간 등의 변수들에 따른 암 관련 통계수치의 패턴을 분석한 결과들을 다양한 데이터 형태로 얻을 수 있다. 또한 암과 관련된 영상 데이터들의 공유 서비스를 제공하며, 정상세포와 전암 및 암세포들의 유전자 발현 프로필을 결정하여 궁극적으로 암환자의 진단 및 치료를 증진시키기 위한 Cancer Genome Anatomy Project(CGAP, http://cgap.nci.nih.gov)를 수행하기도 하였다[35]. 필박스(Pillbox, http://pillbox.nlm.nih.gov/pillimage/search.php)는 미국 국립보건원 산하의 국립 의학도서관(National Library of Medicine)이 운영하는 약 검색을 위한 웹사이트로써, 사용자가 요구하는 다양한 약에 대한 정보를 제공함으로써 알약의 모양, 크기, 색깔, 새겨진 글자나 숫자 등의 정보 만으로도 약에 대한

정확한 정보를 얻을 수 있도록 하고 있다. 또한 약을 개발한 제약회사가 약에 대한 정보를 직접 입력할 수 있는 시스템을 구축하고 사용자 간의 유기적 정보 공유를 가능하도록 하고 있으며, 약에 대한 정보의 수요가 많은 노년층을 위해 용이한 검색 시스템과 비디오 설명자료를 제공하여 빅데이터 기반의 의약품 사용에 대한 정보 제공의 활용성을 높이고자 하였다. 미국 질병통제예방센터(CDC, http://www.cdc.gov)는 미국의 50개 주에서 매주 보고되는 사망률 및 질병 정보 등을 제공하는 웹사이트로써 감염병과 흡연자수, 알코올 중독, 암 등과 관련된 다양한 통계자료를 비롯하여 국가적 프로그램에 대한 정보 및 관련 논문들에 대한 정보들을 포함하는 데이터베이스를 구축 및 관리하고 있다. 웹을 기반으로 하여 다양한 차트의 활용과 시각적 분석이 가능하며 사용자가 분석하고자 하는 여러 요인들을 직접 선택하여 분석할 수 있도록 서비스를 제공하고 있다. 제공되고 있는 세부적 서비스들을 살펴보면, Diseases & Conditions 항목에서는 ADHD, 암, 당뇨병, 심장질환, 인플루엔자 등 주요 질병에 대한 데이터와 통계치를 얻을 수 있으며, Healthy Living 항목에서는 청소년 건강, 식품 안전, 정상 체중 및 비만, 흡연, 백신에 대한 정보를 이용할 수 있다. 이 밖에 여행자들을 위한 건강 관련 정보와 위급상황에서의 응급처치 등에 대한 정보 또한 제공하고 있을 뿐만 아니라 최근 문제가 되고 있는 에볼라 바이러스에 대한 생물학적 임상학적 정보들과 발생현황, 예방을 위한 가이드라인, 관련논문 등 다양한 정보들을 얻을 수 있어 활용도가 높다. 미국의 Google사에서 구축한 'Google Flu Trends(http://www.google.org/flutrends/)'는 독감과 관련된 주제를 검색하는 사람의 수와 실제 독감 증상이 있는 사람의 수 간에 밀접한 관계가 있음을 이용하여 집계된 구글 검색 데이터를 기반으로 현재 전 세계의 독감 유행 수준을 예측하고자 하였다. 독감의 유행수준을 연도별로 예측하고, 전 세계 여러 국가 및 지역에서 독감의 유행 수준에 대한 예측이 가능함을 보여주며, 2004년부터 2014년까지의 연도별 독감 유행 수준에 대해 각 국가별로 실제 인플루엔자 환자 데이터와 'Google Flu Trends' 에서 예측한 수치를 비교한 그래프를 제시함으로써 이 두 수치들이 비교적 유사한 패턴을 보인다는 것을 보여주고 있다. 물론 이것은 모두 과거 수치의 예측 정확성을 보여주는 것으로 미래의 예측 정확성을 보장할 순 없으나 질병 출현의 조기감지에 이용될 수 있는 기초자료 및 시스템으로 이용될 수 있다. 또한 관련 데이터를 매일 업데이트하고 '독감이 시작할 것으로 예상되는 시기, 독감이 최고조에 이를 것으로 예상되는 시기, 이번 독감 계절 동안 예상되는 심각도' 라는 3가지의 정확도 지표를 살펴봄으로써 이 평가결과를 바탕으로 예측 모델을 생성하고 업데이트하여 그 정확성과 성과를 개선하고자 하고 있다[36].

2.6 보건의료 빅데이터를 활용한 연구사례

정부의 공공정보 개방 정책과 함께 국내에서도 보건의료 분야의 빅데이터를 활용한 연구가 활발히 이루어지고 있다. 첫 번째, 한국인체자원은행의 질병기반 자원활용 성과를 그 사례로 들 수 있는데, 몇몇 연구를 살펴보면 다음과 같다. 혈청에서 약 90%에서까지 무증상 감염으로 존재하는 것으로 알려져 있는 JC(Jamestown Canyon) 바이러스는 면역억제 상황이 발생할 경우, 치명적인 탈수초질환을 일으키며, 지금까지 연구된 바로는 다양한 종류의 종양과도 연관성이 있는 것으로 알려져 있다. 조기 위암과 JC 바이러스의 T 항원과의 상관관계를 알아보기 위하여 진행된 연구에서 정상조직에서와 달리 위암조직 샘플의 70%에서 JC 바이러스 T 항원의 서열이 확인되었으므로, 이를 통해 위암 발생과정에 JC 바이러스가 작용할 것이라는 결론을 얻었다[37]. 역시, 한국인체자원은행의 질병기반 자원을 활용한 또 다른 연구에서는 초기 성인기 한국인에서의 도파민 D4 수용체의 유전적 다형성, 기질특성, 음주행동 사이의 연관성을 살펴보기 위해 대상자들의 혈액 샘플에서의 유전자형과 기질특성, 음주행동, 그리고 통계분석을 수행하였다. 연구 결과 도파민 D4 수용체의 유전형은 기질특성과 음주 정도에 영향을 미치지 않는 것으로 나타났으며, 성별에서도 유의한 차이가 나타나지 않았다. 그러나 기질특성 척도 중 하나인 새로운 것을 추구하는 경향성(NS)이 높게 나타날수록 문제적 음주군에 속할 가능성이 높음을 확인하였고, 이 척도가 음주 정도에 영향을 미치는 것으로 보인다는 결론을 얻었으며, 국외의 관련 연구에서도 유사한 결론을 얻었음을 근거로 제시하였다[38]. 두 번째, 보건의료 빅데이터인 건강보험심사평가원에서 제공하는 국가 환자 표본자료(HIRA-NPS, National Patients Sample)를 이용한 연구도 있다. 미국이나 대만에서는 이미 이러한 표본자료를 연구자들에게 제공하여 연구가 이루어지고 있음에 따라 국내에서도 관련 데이터에 대한 접근성을 높여 다양한 연구분야에 활용될 수 있도록 방안을 마련한 것으로서 당뇨병 유병률 추정 연구를 한 사례로 들 수 있다. 이 연구에서는 대표성 있는 표본자료를 이용하여 2009년 국내 당뇨병의 유병률을 추정하고 약물처방양상을 파악하고자 하였다. 표본자료와 모집단 자료에서 각각 당뇨병 유병환자 규모 및 유병률을 산출하였으며, 제2형 당뇨병 환자를 대상으로 혈당강하제 처방양상을 살펴보았다. 그 결과 표본자료를 이용한 당뇨병 유병률 추정결과가 모집단 분석결과와 일치하고, 또한 혈당강하제 각 약효군별 처방률 추정결과와 모집단 분석결과도 일치하는 것으로 확인되었다[39]. 세

번째, 건강보험공단의 빅데이터인 표본코호트 DB를 활용한 혈중바이오마커 변화에 따른 심뇌혈관질환 발생 위험도 연구가 있다. 이는 혈중 바이오마커의 변화가 심뇌혈관질환 발생 위험도에 미치는 영향을 정량적으로 평가하기 위해 이루어졌으며, 예방적 중재에 적합한 시기를 규명하는 것을 목적으로 한다. 환자군과 대조군의 혈중 바이오마커 수준의 변화와 질병 관련성을 분석하고, 혈중 바이오마커(공복혈당, 콜레스테롤, 중성지방, HDL 콜레스테롤 등)의 시점 별 변화치를 분석하였으며, 결과적으로 심혈관질환 환자군은 대조군에 비해 발병 이전 검진결과에서 혈중 바이오마커의 수준이 유의하게 높은 것으로 나타났다. 이러한 연구결과는 혈중 바이오마커의 변화 여부와 속도에 따른 심뇌혈관질환 발생의 위험도를 규명함으로써 질병의 예측력을 향상시킬 수 있을 것으로 보이며, 건강검진의 적절한 검사 주기와 항목 설정 등의 근거로 활용될 수 있을 것으로 기대할수 있다[40]. 몇 가지 연구사례로 살펴본 바와 같이 국내에서도 정보 개방과 관련 데이터베이스의 구축, 연구자들의 데이터 접근성 향상으로 보건의료 빅데이터를 활용한 연구가 지속적으로 증가하고 있음을 알 수 있다. 이러한 움직임은 기존에 제한적으로 이루어졌던 연구의 경계를 넓힘으로써, 국가적 차원에서의 보건의료 문제를 해결하는데 큰 역할을 할 것으로 생각되며, 이를 위해서는 더욱 정확하고 체계적인 데이터 생산과 지속적 관리가 뒷받침되어야할 것이다.

보건의료분야에서는 전자의료기록, 유전체 정보, 인구통계자료, 행동학적 · 사회학적 데이터, 생태학적 · 환경학적 데이터 등을 기반으로 만성 및 감염성 질환들의 인구집단 수준에서의 이해, 진단 및 치료 모니터링, 새로운 질병의 발생 및 위험요인들의 감시, 건강조사의 개선과 질병 통제 등에 활용하는 연구들이 점차 증가하고 있다. 이러한 연구 결과들을 기반으로 과학적이고 혁신적인 보건의료 정책의 수립 및 헬스케어 시스템의 구축, 질병발생 예측과 그에 대한 대비책 마련 등이 이루어지고 있다[41]. Smith K. F. et al. 의 연구에서는 인구의 전세계적 이동 증가와 국제무역의 장벽 완화 등에 의해 감염성 병원체들의 전파가 증가하고 있는 시점에서 병원체들이 가지는 특성 중 하나인 숙주범위(인간감염/다중숙주감염/인수공통감염)에 따른 지리적 분포지도를 구축함으로써 국가별, 대륙별, 지역별 감염성 질환 전파의 양상을 제시하고자 하였다. GIDEON(-Global Infectious Disease and Epidemiology Network) 데이터베이스로부터 추출한 데이터들을 이용하여 전파방식(vector-borne/nonvector-borne)이나 감염성 병원체의 분류체계(bacteria/fungi/parasites/Protistans/viruses)에 따라 사람에 특이적으로 감염성을 가지는 병원체들이 대륙적 · 국가적으로 어떠한 양상을 보이며 분포하는지를 분석하였으며, 분석 결과의 하나로써 인간특이적 감염성 병원체들과 다중숙주범위를 가지는 감염성 병원체들은 인수공통감염성 병원체들에 비해 좀 더 넓은 범위의 분포도를 가지며, 이 두 병원체의 경우 국가적 규모에서 보았을 때 가장 흔하게 존재하는 것은 바이러스인 것으로 나타났다. 이와 달리 상대적으로 낮은 분포도를 보이는 인수공통감염성 병원체의 경우 박테리아가 가장 많고, 그 다음으로 기생충, 바이러스의 순으로 흔하게 존재하는 것으로 나타났다[42]. 이와 같은 연구의 결과들은 인간에 감염되는 감염성 병원체들의 특성에 따른 세계적 유행의 양상 및 정도를 파악하고 이를 바탕으로 질병 발생의 감시 및 예측을 가능하도록 함으로써 체계적이고 과학적인 공중보건 정책과 후속연구들을 위한 기반이 될 수 있을 것이다. Kirkness C. S. et al. 의 연구에서는 Centricity Electronic Medical Records 데이터베이스를 기반으로 1995년 12월 13일부터 2007년 6월 30일까지 기록된 환자들의 데이터를 이용하여 기본 임상 수치(키, 체중, 혈압 등), 치료(처방), 진단(ICD-9 codes) 기준 등의 자료를 통해 당뇨병이나 이와 관련된 위험요인들(고혈압, 중성지방 과다, LHDL, BMI지수 등)의 여부를 밝히고, 이러한 당뇨병 유병률 및 관련 요인들과 물리치료 수요환자들과의 관련성을 비롯한 다른 변수들과의 상관관계를 밝힘으로써 이를 보건의료 계획에 활용하고자 하였다. 연구 결과, 당뇨병 발생률은 표본 인구집단의 13.2%로 나타났으며, 당뇨병 관련 위험요인들의 평가에서 고혈압의 유병률이 70.4%로 가장 높게 나타난 반면 BMI지수 초과는 39.1%로 과반수 이하로 나타났다. 표본 집단의 20%만이 모든 기본임상수치, 치료, 진단기준에 대해 정상 한계치 이내에 속하는 것으로 나타났으며 결론적으로 물리치료를 필요로 하는 사람들 중 대다수가 당뇨병을 가지고 있거나 당뇨병과 관련된 위험요인을 한 가지 이상 가지고 있다는 것을 보여주었다. 또한 당뇨병과 관련 위험요인들의 유무뿐만 아니라 인구학적 요인들에 의한 연구집단의 특성들, 지리적 위치, 건강보험상태, 물리치료를 필요로 하는 상태 등을 설명하기 위해 EMR 데이터베이스 기반의 기술역학적 연구를 수행함으로써 빅데이터를 활용한 보건의료 계획과 정책의 수립을 위한 과학적 정보를 제시할 수 있음을 보여주었다[43]. 이처럼 빅데이터의 수집 및 가공, 분석을 통한 생물학적, 의료학적 지식의 창출이 광범위하게 이루어지고 있으며, 빅데이터를 헬스케어에 활용하기 위한 분석 방법론 및 데이터마이닝 알고리즘의 개발 등 다양한 연구들이 수행되고 있다[44]. 이는 보건의료 분야에 있어 정부기관과 의료산업기관은 올바른 정책과 효율적인 헬스케어 시스템을 갖추고, 소비자에게는 'Informed Decision'을 위한 올바른 의료정보를 제공하는 효과를 가져올 것으로 기대된다.

2.7 보건의료 및 바이오 분야 빅데이터의 활용 전략

보건의료 및 바이오 분야에서 빅데이터를 효율적으로 활용하기 위해서는 아래와 같은 전략이 필요하다. 첫째, 보건의료 및 바이오 빅데이터 수집 및 활용을 위한 국가적인 정책 및 제도적 지원이 필요하다. 오랫동안 보건의료 및 바이오 분야의 연구개발 지원이 이루어져 왔고, 많은 논문과 데이터가 축적되면서 선진국을 중심으로 이러한 창출된 데이터의 중요성을 인지하고 활용화하기 위한 정책들이 논의되어 왔다. 바이오경제(Bioeconomy)란 바이오 연구로부터 생성된 물질 및 기술, 서비스를 통해서 창출되는 경제활동을 의미하는데, 미국은 2012년 국가 바이오경제 청사진(National Bioeconomy Blueprint)를 통해서 바이오 분야의 경제성 향상을 위한 5대 전략 목표와 실행과제를 제시하였다[45]. 그러한 전략 목표 중의 하나로 제한된 예산에서 투자의 효율성을 최대로 증대시킬 수 있는 유망기술로 생명정보학(Bioinformatics)을 제시하였다. 생명정보학은 빅데이터 기술을 포함하는 학문으로 국가적으로 유망기술에 선택적으로 집중 투자한다는 점은 국내 정책도 지향해야 할 점이다[46]. 현재 보건의료 분야 데이터는 보건복지부, 식품의약품안전처, 국민건강보험공단, 건강보험심사평가원 등 다양한 정부기관 및 의료기관에서 관리 및 운영되고 있어서 각 기관에서 보유한 데이터들의 공유와 연계를 위한 전략이 필요하다. 보건의료 데이터를 효율적으로 활용하기 위해서는 이러한 정보를 수집 · 분석할 수 있도록 공개하는 것이 중요하다. 공공데이터 포털(www.data. go.kr)에서 공공 데이터를 공개하고 있지만 아직 보건의료 분야의 공개는 미비하다. 또한 보건의료 데이터는 개인 정보를 포함하고 있어서 데이터 공개 허용 및 보안 등에 대한 제도의 정비가 시급하며, 정보 활용을 활성화하면서 개인 정보를 보호할 수 있는 적절한 정책이 마련되어야 될 것이다. 미국, 일본, 영국과 같은 나라에서는 체계적인 데이터 수집 및 저장 시스템을 구축하여 자국 내 바이오 데이터뿐만 아니라 국가 간 데이터 공유를 통해 과학기술 선도국으로서 자리매김을 하고 있다. 반면, 우리나라는 아직까지 이러한 체계적인 시스템의 부족으로 많은 연구의 성과물들이 제대로 관리되지 못하고 있는 실정이다[46]. 국내에서는 2007년부터 생물소재 및 생물정보의 한국생명공학연구원 생물자원센터, 논문 및 보고서의 한국과학기술정보연구원 등 연구성과물 전담기관을 지정하여 각 정부 및 기관, 학교에서 산출되는 수많은 연구 성과물을 수집하여 DB화하고 있으나, 데이터의 활용 및 재사용 정도가 미비한 실정이다. 정부는 한국생명공학연구원 국가생명연구자원정보센터(Korean Bioinformation Center, KOBIC)를 구축하여 국내 생물자원, 생물다양성 그리고 생명정보를 통합 관리하고, 국가과학기술정보센터(National Discovery for Science Leaders, NDSL)를 통해 학술논문, 특허, 연구보고서, 동향분석정보를 수집하는 시스템을 운영 중이다. 그러나 연구자들이 이러한 국내 시스템을 이용해서 성과물을 활용하기 위해서는 여전히 어려움이 있다. 이는 연구 결과의 낮은 활용도로 인해 중복 연구의 가능성을 높이고 국내 과학기술의 성장을 늦추는 요인으로 작용할 수 있다.

둘째, 바이오 및 보건의료 빅데이터 분석결과의 사업화 촉진이 필요하다. 이러한 사업화는 학교 · 연구소에서 창출된 빅데이터 분석 기법 및 연구 결과를 산업에 빠르게 적용하여 다양한 시장에서의 경쟁력 향상에 기여할 수 있을 것이다. 국내에서는 네이버, 다음소프트와 같은 대표적 포털과 삼성, SK, KT 등 스마트기기 관련 업체들이 빅데이터 기술을 기반으로 다양한 사업을 지원하고 있다. 그 예로, 다음소프트는 블로그, 트위터 문서를 분석하여 모니터링 정보를 제공하는 소셜 매트릭스(SOCIAL metrics) 서비스를 제공해서 사용자가 입력한 키워드에 대해서 소셜 미디어에서 노출된 빈도 및 관련 연관어 맵 등을 제공하고 있다(http://insight.some.co.kr).

셋째, 국가적 차원에서 빅데이터 시장의 우위를 선점하기 위해서는 대량의 데이터에서 의미있는 정보를 찾아낼 수 있는 빅데이터 전문 인력을 양성하는 교육 프로그램 및 기반 기술 연구에 대한 지원을 높여야 한다. 빅데이터 시대에서는 원하는 정보를 얻기 위해서 데이터를 관리하고 분석할 수 있는 전문 인력이 매우 필요하다. 발 빠른 미국에서는 2012년 '빅데이터 R&D 이니셔티브'를 통해서 국가적으로 빅데이터 활성화 전략을 추진했다. 각 부처와 기관들은 각 기관별 특성에 따른 빅데이터 인력 양성, 새로운 데이터 분석방법론 개발, 정보공유를 통한 데이터 개발, 컴퓨팅 기술 및 소프트웨어 개발, 시각화 기술 개발 등 빅데이터 기술개발을 목표로 다각적인 전략을 수행하였다. 이를 통해서 향후 미래 바이오 경제에 지속적으로 적절한 교육을 받은 빅데이터 전문인력 및 기반 기술들을 확보할 수 있도록 하였다. 국내에서는 빅데이터 분석 활성화를 위해 한국정보화진흥원에서 '빅데이터 분석활용센터(http://kbig.kr/)'를 운영하여 빅데이터 기술 인력을 양성하기 위한 교육 · 실습 인프라를 제공하고 있다. 그러나, 미국에 비해 국내 빅데이터 관련 정책 지원이 부족한 실정이고, 고가의 컴퓨팅 기자재 구입 면에서도 많은 제약이 따르고 있다.

Ⅲ. 결 론

빅데이터는 축적된 데이터의 활용 및 의료 비용 절감 측면에

서 보건의료 분야에 새로운 패러다임을 제공할 수 있을 것이다. 국내 보건복지 분야에서는 국민건강보험공단, 건강보험심사평가원 등에서 빅데이터를 활용한 서비스를 제공하고 있으며, 미국에서는 국립보건원의 의약품 검색 서비스인 필박스와 웰포인트의 인공지능 컴퓨터 시스템 왓슨이 빅데이터 분석 능력을 활용하여 약물정보 제공, 정확한 진단과 치료 효과 향상에 기여했다. 또한 정보통신 기술의 발달로 스마트 기기를 이용한 개인 데이터를 기반으로 개인 맞춤형 서비스 제공이 가능해지고, 축적된 데이터에서 패턴을 탐색하여 '정상'과 '비정상' 클래스를 구분할 수 있는 모델을 구축하여 위험 분자나 이상 현상을 감지하는 데에 이용되고 있다. 바이오 분야에서 미국, 유럽, 일본 등 선진국들은 연구 성과물을 NCBI, EMBL, DDBJ 등 바이오 특화 데이터베이스에 체계적으로 저장하는 시스템이 구축되어 있다. 그 결과 방대한 빅데이터들이 축적되어 왔고, 후속 연구자들에게 연구기반을 제공하고 있다. 이러한 시스템은 이들 나라를 과학기술 강대국으로 만드는데에 큰 역할을 하였다. 우리나라는 보건의료 및 바이오 성과물 데이터에 대한 관리 시스템이 아직 체계적으로 구축되지 않고 있다. 특히, 보건의료 및 바이오 분야에서 축적된 빅데이터는 국가 과학기술의 경쟁력을 향상시킬수 있기 때문에 보건의료 및 바이오 데이터가 효과적으로 활용될 수 있도록 체계적으로 수집 · 공유하는 작업이 반드시 필요하다. 또한 빅데이터 수집을 위한 플랫폼 구축 및 분석 전문가를 양성할 수 있는 국가적인 인재 개발 시스템이 필요하다. 마지막으로 개인정보가 포함된 보건의료 데이터에 대한 개인정보 유출 차단 기술의 개발이 필요하며, 개인 정보를 철저히 은닉하면서 보건의료 빅데이터의 활용성을 높일 수 있는 적절한 방안이 강구되어야 할 것이다.

Acknowledgement

이 논문은 미래창조과학부의 재원으로 한국연구재단의 일반연구자지원사업(MEST2012008344)과 바이오-의료기술개발사업(No.2012M3A9D1054622)의 지원을 받아 수행된 연구의 결과물입니다.

참 고 문 헌

[1] 이성훈, 이동우. "빅데이터의 국내 · 외 활용 고찰 및 시사점," 한국디지털정책학회 디지털정책연구, 11(2), pp. 229~233. 2013.

[2] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A. H. "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, 2011.

[3] 김정숙. "빅데이터 활용과 관련기술 고찰," 한국콘텐츠학회, 10(1), pp. 34-40. 2012.

[4] Laney D. "3D Data Management: Controlling Data Volume, Velocity and Variety," Application Delivery Strategies, 949, 2001.

[5] 정지선. "성공적인 빅데이터 활용을 위한 3대요소: 자원, 기술, 인력," IT & Future Strategy, 3, 2012.

[6] 국가정보화전략위원회. "지식정보 개방과 협력으로 스마트 정부 구현," 2011.

[7] Varshney U. "Pervasive Healthcare," IEEE Communication, pp. 138-140.

[8] 송태민. "보건복지 빅데이터의 효율적 활용 방안," 한국보건사회연구원 보건복지포럼, 193, 2012.

[9] 박두순, 문양세, 박영호, 윤찬현, 정영식, 장형석. "빅데이터 컴퓨팅 기술," 한빛아카데미, 2014.

[10] Warden P. "Big Data Glossary," O'Relly Media, 2011

[11] 김정미. "빅데이터 시대의 데이터 자원 확보와 품질 관리 방안," 한국정보화진흥원, IT & Future Strategy, 5, pp. 1-21, 2012.

[12] IDC. "빅 데이터 분석: CIO를 위한 미래지향적 아키텍처 기술 그리고 로드맵," 2011

[13] Halvey M, Keane M. T. "An Assessment of Tag Presentation Techniques," poster presentation at WWW 2007. 2007

[14] Freeman L. "Centrality in Social Networks: a Sonceptual Classification," Social Neworks, 1, 1979.

[15] 국가정보화전략위원회. "빅데이터를 활용한 스마트 정부 구현안," 2011.

[16] 관계부처 합동. "정부3.0 추진 기본계획," 2013.

[17] 정부3.0 추진위원회. "정부3.0 발전계획," 2014.

[18] 국가법령정보센터(http://www.law.go.kr).

[19] 윤미영. "주요국의 빅데이터 추진전략 분석 및 시사점," 과학기술정책, 23(3), pp. 31-43, 2013

[20] 배동민, 박현수, 오기환. "빅데이터 동향 및 정책 시사점," 방송통신정책, 25(10), pp. 37-74, 2013.

[21] 한국정보화진흥원 빅데이터전략연구센터. "新가치창출을 위한 주요국의 빅데이터 추진 전략 분석," IT & Future Strategy, 11, 2012.

[22] 이정아. "스마트 정부의 공공정보 개방과 이용활성화 전략," IT&SOCIETY, 28, pp. 1-19, 2010.
[23] Han Y., Huh S. J., Ju S. G., Ahn Y. C., Lim D. H., Lee J. E., Park W. "Impact of an electronic chart on the staff workload in a radiation oncology department," Jpn J Clin Oncol., pp. 470-474, 2005.
[24] Furhang E. E., Dolan J., Sillanpaa J., Harrison L. B. "Automating the initial physics chart checking process," J Appl Clin Med Phys., pp. 129-135, 2009.
[25] 이준영, 김기환, 이지성. "국민건강정보 데이터베이스를 이용한 표본 코호트 DB 구축," pp. 1-15, 2014.
[26] 윤미영, 권정은. "빅데이터로 진화하는 세상: big data 글로벌 선진사례," 한국정보화진흥원, pp. 1-159, 2012.
[27] 국립암센터. "국립암센터, EMR기반 통합의료정보시스템 사업종료 보고회 성료," 사이버홍보센터 보도자료, 2010.
[28] 홍현철. "국립암센터, EMR시스템(전자의무기록시스템) DB 암호화 적용," 지역정보화, pp. 65-67, 2014.
[29] IBM. "Wellpoint and IBM announce agreement to put Watson to work in health care," (http://www-03.ibm.com/press/us/en/pressrelease/35402.wss), 2011.
[30] Arnaout R. "Elementary, My Dear Doctor Watson," Clinical Chemistry, 58(6), pp. 986-988, 2012.
[31] McCarty C. A., Chisholm R. L., Chute C. G., Kullo I. J., Jarvik G. P., Larson E. B. et al. "The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies," BMC Medical Genomics, 4(13), 2011.
[32] IBM. "University of Ontario Institute of Technology: Leveraging key data to provide proactive patient care," 2010.
[33] Visvanathan A., Hamilton A., Brady R. R. W. "Smartphone apps in microbiology - is better regulation required?," Clinical Microbiology and Infection, 18, E218-E220, 2012.
[34] 질병관리본부. "바이오뱅크 최신 동향과 한국인체자원은행사업," 2010.
[35] Strausberg R. L. "Cancer Genome Anatomy Project," eLS, 1999.
[36] Ginsberg J., Mohebbi M. H., Patel R. S., Brammer L., Smolinski M. S., Brilliant L. "Detecting influenza epidemics using search engine query data," Nature, 457(19), 2009.
[37] Son H. S., Jeon S. W., Jung M. K., Kim S. K., Kim J. T., Choi Y. J., Xu Z. G., Bae H. I. "Association of T-antigen of jamestown canyon virus with early gastric cancer," Korean Journal of Helicobacter and UPPER Gastrointestinal Research, pp. 27-31. (2010).
[38] Nam Y. W., Lee S. I., Shin C. J., Son J. W., Kim S. K. "The association among the genetic polymorphism of dopamine D4 receptor, temperament and alcohol drinking behavior in young Korean adults," Korean J Biol Psychiatry, pp. 1-8. (2011).
[39] 박병주. "우리나라 당뇨병 유병률 추정 및 DPP-4 억제제 사용양상평가," 건강보험심사평가원 국가환자표본자료(HIRA-NPS : National Patients Sample)를 이용한 학술심포지엄, 2011.
[40] 김현창. "혈중 바이오마커 변화에 따른 뇌혈관질환 발생 위험도," 국민건강보험공단 빅데이터 시범연구 학술 심포지엄, 2013.
[41] Pentland A., Reid T. G., Heibeck T. "Revolutionizing medicine and Public Health," Report of the Big Data and Health Working Group, 2013.
[42] Smith K. F., Sax D. F., Gaines S. D., Guernier V., Gue´gan J. F. "Globalization of Human Infectious Disease," Ecology, 88(8), pp. 1903-1910, 2007.
[43] Kirkness C. S., Marcus R. L., LaStayo P. C., Asche C. V., Fritz J. M. "Diabetes and Associated Risk Factors in Patients Referred for Physical Therapy in a National Primary Care Electronic Medical Record Database," Physical Therapy, 88(11), 2015.
[44] Raghupathi W., Raghupathi V. "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, 2(3), 2014.
[45] 한국산업기술진흥원. "미국의 바이오 산업 현황 및 정책 동향," 2013.
[46] 배세은. "빅데이터 DB구축을 통한 과학기술 정책 효율화 연구: 국가 바이오사업 분야를 중심으로," 한국과학기술기획평가원. 2014.

약 력

이 지 혜

2013년 서울대학교 협동과정 생물정보학 석사
2013년~2015년 서울대학교 협동과정 생물정보학 박사 수료
관심분야: 생명정보학

제 미 경

2008년 경북대학교 자연과학대학 지질학사
2015년 서울대학교 보건대학원 보건학 석사
관심분야: 생명정보학

조 명 지

2008년 건국대학교 동물생명과학대학 동물생명공학사
2013년~현재 서울대학교 보건대학원 보건학과
관심분야: 생명정보학

손 현 석

2003년~현재 서울대학교 보건대학원 보건학과 & 자연과학대학 협동과정 생물정보학 교수
관심분야: 생명정보학, 보건학, 역학시뮬레이션