Big Data Application in Biomedical Research and Health Care: A Literature Review

Jake Luo, Min Wu, Deepika Gopukumar and Yiqing Zhao

College of Health Science, Department of Health Informatics and Administration, Center for Biomedical Data and Language Processing, University of Wisconsin–Milwaukee, Milwaukee, WI, USA.

ABSTRACT: Big data technologies are increasingly used for biomedical and health-care informatics research. Large amounts of biological and clinical data have been generated and collected at an unprecedented speed and scale. For example, the new generation of sequencing technologies enables the processing of billions of DNA sequence data per day, and the application of electronic health records (EHRs) is documenting large amounts of patient data. The cost of acquiring and analyzing biomedical data is expected to decrease dramatically with the help of technology upgrades, such as the emergence of new sequencing machines, the development of novel hardware and software for parallel computing, and the extensive expansion of EHRs. Big data applications present new opportunities to discover new knowledge and create novel methods to improve the quality of health care. The application of big data in health care is a fast-growing field, with many new discoveries and methodologies published in the last five years. In this paper, we review and discuss big data application in four major biomedical subdisciplines: (1) bioinformatics, (2) clinical informatics, (3) imaging informatics, and (4) public health informatics. Specifically, in *bioinformatics*, high-throughput experiments facilitate the research of new genome-wide association studies of diseases, and with *clinical informatics*, the clinical field benefits from the vast amount of collected patient data for making intelligent decisions. *Imaging informatics* is now more rapidly integrated with cloud platforms to share medical image data and workflows, and *public health informatics* leverages big data techniques for predicting and monitoring infectious disease outbreaks, such as Ebola. In this paper, we review the recent progress and breakthroughs of big data applications in these health-care domains and summarize the challenges, gaps, and opportunities to improve and advance big data applications in health care.

KEYWORDS: big data, literature review, health care, data-driven application

CITATION: Luo et al. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights* 2016:8 1–10 doi: 10.4137/BII.S31559.

TYPE: Review

RECEIVED: August 27, 2015. **RESUBMITTED:** December 06, 2015. **ACCEPTED FOR PUBLICATION:** December 06, 2015.

ACADEMIC EDITOR: John P. Pestian, Editor in Chief

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1448 words, excluding any confidential comments to the academic editor.

FUNDING: This research was supported by GE Healthcare Catalyst Grant and UWM Foundation at the University of Wisconsin-Milwaukee. The authors confirm that the funders had no influence over the content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: jakeluo@uwm.edu

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE). Provenance: the authors were invited to submit this paper.

Published by Libertas Academica. Learn more about this journal.

Background: What is Big Data?

In the biomedical informatics domain, big data is a new paradigm and an ecosystem that transforms case-based studies to large-scale, data-driven research. It is widely accepted that the characteristics of big data are defined by three major features, commonly known as the 3Vs: volume, variety, and velocity.

First and most significantly, the *volume of data* is growing exponentially in the biomedical informatics fields.[1–7] For example, the ProteomicsDB[8] covers 92% (18,097 of 19,629) of known human genes that are annotated in the Swiss-Prot database. ProteomicsDB has a data volume of 5.17 TB. In the clinical realm, the promotion of the HITECH Act[9] has nearly tripled the adoption rate of electronic health records (EHRs) in hospitals to 44% from 2009 to 2012. Data from millions of patients have already been collected and stored in an electronic format, and these accumulated data could potentially enhance health-care services and increase research opportunities.[10,11] In addition, medical imaging (eg, MRI, CT scans) produces vast amounts of data with even more complex features and broader dimensions. One such example is the Visible Human Project, which has archived 39 GB of female datasets.[12] These and other datasets will provide future opportunities for large aggregate collection and analysis.

The second feature of big data is the *variety of data types and structures*. The ecosystem of biomedical big data comprises many different levels of data sources to create a rich array of data for researchers. For example, sequencing technologies produce "omics" data systematically at almost all levels of cellular components, from genomics, proteomics, and metabolomics to protein interaction and phenomics.[13] Much of the data that are unstructured[14] (eg, notes from EHRs,[15,16] clinical trial results,[17,18] medical images,[19] and medical sensors) provide many opportunities and a unique challenge to formulate new investigations.

The third characteristic of big data, *velocity*, refers to producing and processing data. The new generation of sequencing technologies enables the production of billions of DNA sequence data each day at a relatively low cost. Because faster speeds are required for gene sequencing,[1,20] big data technologies will be tailored to match the speed of producing data, as is required to process them. Similarly, in the public health field, big data technologies will provide biomedical researchers with time-saving tools for discovering new patterns among population groups using social media data.[21,22]

Big Data Technologies

Biomedical scientists are facing new challenges of storing, managing, and analyzing massive amounts of datasets.[23] The characteristics of big data require powerful and novel technologies to extract useful information and enable more broad-based health-care solutions. In most of the cases reported, we found multiple technologies that were used together, such as artificial intelligence (AI), along with Hadoop®,[24] and data mining tools.

Parallel computing is one of the fundamental infrastructures for managing big data tasks. It is capable of executing algorithm tasks simultaneously on a cluster of machines or supercomputers. In recent years, novel parallel computing models, such as MapReduce[25] by Google, have been proposed for a new big data infrastructure. More recently, an open-source MapReduce package called Hadoop[24] was released by Apache for distributed data management. The Hadoop Distributed File System (HDFS) supports concurrent data access to clustered machines. Hadoop-based services can also be viewed as cloud-computing platforms, which allow for centralized data storage as well as remote access across the Internet.

As such, *cloud computing* is a novel model for sharing configurable computational resources over the network[26] and can serve as an infrastructure, platform, and/or software for providing an integrated solution. Furthermore, cloud computing can improve system speed, agility, and flexibility because it reduces the need to maintain hardware or software capacities and requires fewer resources for system maintenance, such as installation, configuration, and testing. Many new big data applications are based on cloud technologies.

Research Methods

We searched four bibliographic databases to find related research articles: (1) PubMed, (2) ScienceDirect, (3) Springer, and (4) Scopus. In searching these databases, we used the main keywords "big data," "health care," and "biomedical." Then, we selected papers based on the following inclusion criteria:

1. The paper was written in English and published within the past five years (2000–2015).
2. The paper discussed the design and use of a big data application in the biomedical and health-care domains.
3. The paper reported a new pipeline or method for processing big data and discussed the performance of the method.
4. The paper evaluated the performance of new or existing big data applications.

The following exclusion criteria were used to filter out irrelevant papers:

1. The paper did not discuss any specific big data applications (eg, general comments about big data).
2. The paper was a tutorial or a course material.
3. The paper was not in the four focus areas: bioinformatics, clinical informatics, public health informatics, and imaging informatics.

Two searches were performed. In the first search, the first author (JL) and the second author (MW) of the present study began the search process based on the main keywords. All potentially related papers were collected by reviewing the title and abstract. This initial search resulted in 755 papers from 2000 to 2015. In the second search, the second author (MW) and the third author (DG) screened the papers based on the abovementioned inclusion and exclusion criteria and subsequently selected 94 candidate papers. Finally, each author of the present study evaluated the final selection by reading the content of the papers, and consensus was reached to review 68 papers for this study.

Big Data Applications

Bioinformatics applications. Bioinformatics research analyzes biological system variations at the molecular level. With current trends in personalized medicine, there is an increasing need to produce, store, and analyze these massive datasets in a manageable time frame. Next-generation sequencing technology enables genomic data acquisition in a short period of time.[27,28] The role of big data techniques in bioinformatics applications is to provide data repositories, computing infrastructure, and efficient data manipulation tools for investigators to gather and analyze biological information. Taylor discusses that Hadoop and MapReduce are currently used extensively within the biomedical field.[29]

This section classifies big data technologies/tools into four categories: (1) data storage and retrieval, (2) error identification, (3) data analysis, and (4) platform integration deployment. These categories are correlated and may overlap; for instance, most data input applications may support simple data analysis, or vice versa. However, our classification in the present study is based only on the main functions of each technology.

Data storage and retrieval. Nowadays, a sequencing machine can produce millions of short DNA sequencing data during one run. The sequencing data need to be mapped to specific reference genomes in order to be used for additional analysis, such as genotype and expression variation analysis. *CloudBurst*[30] is a parallel computing model that facilitates the genome mapping process. CloudBurst parallelizes the short-read mapping process to improve the scalability of reading large sequencing data. The CloudBurst model was evaluated using a 25-core cluster, and the results indicate that the speed to process seven million short-reads was almost 24 times faster than a single-core machine. The CloudBurst team have developed new tools based on CloudBurst to support biomedical research, such as Contrail[31] for assembling large genomes and Crossbow[32] for identifying single nucleotide polymorphisms (SNPs) from sequencing data.

DistMap[33] is a toolkit for distributed short-read mapping on a Hadoop cluster. DistMap aims to increase the support of different types of mappers to cover a wider range of sequencing applications. The nine supported mapper types include BWA, Bowtie, Bowtie2, GSNAP, SOAP, STAR, Bismark, BSMAP, and TopHat. A mapping workflow is integrated into DistMap, which can be operated with simple commands. For example, an evaluation test was done using a 13-node cluster, making it an effective application for mapping short-read data. The BWA mapper can perform 500 million read pairs (247 GB) in about six hours using DistMap, which is 13 times faster than a single-node mapper.

SeqWare[34] is a query engine built on the Apache HBase[35] database to help bioinformatics researchers access large-scale whole-genome datasets. The SeqWare team created an interactive interface to integrate genome browsers and tools. In a prototyping analysis, the U87MG and 1102GBM tumor databases were loaded, and the team used this engine to compare the Berkeley DB and HBase back end for loading and exporting variant data capabilities. The results show that the Berkeley DB solution is faster when reading 6M variants, while the HBase solution is faster when reading more than 6M variants.

The *Read Annotation Pipeline*®[36] by the DNA Data Bank of Japan (DDBJ) is a cloud-based pipeline for high-throughput analysis of next-generation sequencing data. DDBJ initiated this cloud-computing system to support sequencing analysis. It offers a user-friendly interface to process sequencing datasets, which supports two levels of analysis: (1) the basic-level tools accept FASTQ format data and preprocess them to trim low-quality bases and (2) during the second analysis, the data are mapped to genome references or assembled on supercomputers. This pipeline uses the Galaxy interface for advanced analysis, such as SNP detection, RNA-sequencing (RNA-seq) analysis, and ChIP-seq analysis. In a benchmark testing, DDBJ finished mapping 34.7 million sequencing reads to a 383-MB reference genome in 6.5 hours.

Hydra[37] is a scalable proteomic search engine that uses the Hadoop-distributed computing framework. Hydra is a software package for processing large peptide and spectra databases, implementing a distributed computing environment that supports the scalable searching of massive amounts of spectrometry data. The proteomic search in Hydra is divided into two steps: (1) generating a peptide database and (2) scoring the spectra and retrieving the data. The system is capable of performing 27 billion peptide scorings in about 40 minutes on a 43-node Hadoop cluster.

Error identification. A number of tools have been developed to identify errors in sequencing data; *SAMQA*[38] identifies such errors and ensures that large-scale genomic data meet the minimum quality standards. Originally built for the National Institutes of Health Cancer Genome Atlas to automatically identify and report errors, SAMQA includes a set of technical tests to find data abnormalities (eg, sequence alignment/map [SAM] format error, invalid CIGAR value) that contain empty reads. For biological tests, researchers can set a threshold to filter reads that could be erroneous (empty reads) and report them to experts for manual evaluation. A comparison of Hadoop, which was tested on a cluster, with SAMQA, which was tested on a single-core server, shows that the Hadoop cluster processed a 23-GB sample nearly 80 times faster (18.25 hours).

ART[39] provides simulation data for sequencing analysis for three major sequencing platforms: 454 Sequencing™, Illumina, and SOLiD. ART has built-in profiles of read error and read length and can identify three types of sequencing errors: base substitutions, insertions, and deletions.

CloudRS[40] is an error-correction algorithm of high-throughput sequencing data based on a parallel, scalable framework. This method is developed based on the RS algorithm.[41] The CloudRS team evaluated the system on six different datasets using the GAGE benchmarks,[42] and the results show that CloudRS has a higher precision rate compared with the Reptile[43] method.

Data analysis. In addition to the described frameworks and toolkits for sequencing data analysis, the *Genome Analysis Toolkit* (GATK)[20,44] is a MapReduce-based programing framework designed to support large-scale DNA sequence analysis. GATK supports many data formats, including SAM files, binary alignment/map (BAM), HapMap, and dbSNP. With GATK, "traversal" modules prepare and read sequencing data into the system and thus provide associated references to the data, such as ordering data by loci. The "walker" module consumes the data and provides analytics outcomes. GATK has been used in the Cancer Genome Atlas and 1000 Genomes Projects.

The *ArrayExpress Archive of Functional Genomics* data repository[45,46] is an international collaboration for integrating high-throughput genomics data. The repository contains 30,000 experiments and more than one million assays. About 80% of the data were extracted from the GEO data repository, and the rest 20% were directly submitted to ArrayExpress by its users. Each day, the platform is visited by more than 1,000 different users, and more than 50 GB of data are downloaded. The platform also connects with R and GenomeSpace to support data transition and analysis.

BlueSNP[47] is an R package for genome-wide association studies (GWAS) analysis, focusing on statistical tests (eg, *P*-value) to find intensive associations between large genotype–phenotype datasets. BlueSNP operates on the Hadoop platform, which reduces barriers and improves the efficiency of running GWAS analyses on clustered machines. On a 40-node cluster, BlueSNP analyzed 1,000 phenotypes on 10^6 SNPs in 10^4 individuals within 34 minutes.

Myrna[48] is a cloud-based computing pipeline that calculates the differences of gene expression in large RNA-seq datasets. RNA-seq data are m-sequencing reads derived from mRNA molecules. Myrna supports several functions for RNA-seq analysis, including reads alignment, normalization, and statistical modeling in an integrated pipeline. Myrna returns differential expression of genes into the form of *P*-value and *q*-value. This system was tested on the Amazon Elastic Compute Cloud (Amazon EC2) using 1.1 billion RNA-seq reads, and the results show that Myrna can process data in less than two hours; the cost of the test task was around $66.

The *Eoulsan* package[49] implanted a pipeline for analyzing the differential transcript expressions, including data imports from sequencer reads, data mapping to reference genomes, alignment filters, transcription expression calculations, expression normalizations using edgeR, and detection of differential expressions. Eoulsan can be run under three modes: standalone, local cluster, and cloud on Amazon Elastic MapReduce. Eoulsan was tested on Amazon EC2 using eight mouse samples of 188 million reads. The cost for processing the data was $18–$66, and the total time ranged from 109 to 822 minutes.

SparkSeq[50] is a fast, scalable, cloud-ready software package for interactive genomic data analysis with nucleotide precision. SparkSeq provides interactive queries for RNA/DNA studies, and the project is implemented on Apache Spark using the Hadoop-BAM library for processing bioinformatics files.

Platform integration deployment. The use of big data platforms usually requires a strong grasp of distributed computing and networking knowledge. To help biomedical researchers embrace big data technology, novel methods are needed to integrate existing big data technologies with user-friendly operations. The following systems have been developed to help achieve this goal.

SeqPig[51] reduces the need for bioinformaticians to obtain the technological skills needed to use MapReduce. The SeqPig project extends the Apache Pig scripts to provide feature-rich sequence processing functions. With the help of Hadoop-BAM,[52] SeqPig solves the problem of reading large BAM files to feed analysis applications. SeqPig supports commonly used sequencing formats, such as FASTQ, SAM, BAM, and QSeq. It further sustains commonly used processing tools, such as Pileup, base frequency count, read coverage, and distribution.

Current bioinformatics platform also incorporates a virtual machine. *CloVR*[53] is a sequencing analysis package that is distributed through a virtual machine. By reducing the technical barriers for analyzing large sequencing datasets, CloVR supports both local desktop and cloud systems to enable high-throughput data processing. Several automated bioinformatics workflows/pipelines are integrated into the virtual machine, including whole-genome, metagenome, and 16S rRNA-sequencing analysis. The CloVR team tested the portability of the system on a local machine (4 CPU, 8 GB RAM) and on the Amazon EC2 cloud platform (80 CPU), and the results show that CloVR is portable on both platforms, while the EC2 instance runs about five times faster. Similarly, *CloudBioLinux*[54] is a virtual machine solution that provides more than 135 bioinformatics packages for sequencing analysis, including preconfigured tools (eg, GATK, Bowtie, Velvet, FASTX) and programing libraries (eg, BioJava, R, Bioconductor).

Deploying the Hadoop cloud platform can be a big challenge for researchers who do not have a computer science background. *CloudDOE* is a software package that provides a simple interface for deploying the Hadoop cloud because the Hadoop platform is often too complex for scientists without computer science expertise and/or similar technical skills. CloudDOE[55] is a user-friendly tool for analyzing high-throughput sequencing data with MapReduce, encapsulating the complicated procedures for configuring the Hadoop cloud for bioinformatics researchers. Several packages are integrated with the CloudDOE package (CloudBurst, CloudBrush, and CloudRS), and its operation is further simplified by wizards and graphic user interfaces.

Clinical informatics applications. Clinical informatics focuses on the application of information technology in the health-care domain. It includes activity-based research, analysis of relationship between patient main diagnosis (MD) and underlying cause of death (UCD), and storage of data from EHRs and other sources (eg, electrophysiological [such as EEG] data). In this section, we classified big data technologies/tools into four categories: (1) data storage and retrieval, (2) interactive data retrieval for data sharing, (3) data security, and (4) data analysis. Compared with bioinformatics, clinical informatics does not offer many tools for error identification but pays more attention to data-sharing and data security issues. Its data analysis method is very different from bioinformatics, as clinical informatics works with both structured and unstructured data, develops specific ontologies, and uses natural language processing extensively.

Data storage and retrieval. It is critical to discuss the ways in which big data techniques (eg, Hadoop, NoSQL database) are used for storing EHRs. The efficient storage of data is especially important when working with clinical real-time stream data.[56] Dutta et al evaluated the potential of using Hadoop and HBase[35] as data warehouses for storing EEG data and discussed their high-performance characteristics. Jin et al.[57] analyzed the potential of using Hadoop HDFS and HBase for distributed EHRs.

Furthermore, Sahoo et al.[58] and Jayapandian et al.[59] proposed a distributed framework for storing and querying large amounts of EEG data. Their system, *Cloudwave*, uses Hadoop-based data processing modules to store clinical data, and by leveraging the processing power of Hadoop, they developed a web-based interface for real-time data visualization and retrieval. The Cloudwave team evaluated a dataset of 77-GB EEG signal data and compared Cloudwave with a stand-alone system; the results show that Cloudwave processed five EEG studies in 1 minute, while the stand-alone system took more than 20 minutes.

Compared with a traditional relational database that handles structured data well, the novel *NoSQL* is a great prospect for storing unstructured data. Mazurek[60] proposed a system that combines both relational and multidimensional technologies with NoSQL repositories to enable data mining techniques and provide flexibility and speed in data processing. Nguyen et al.[61] presented a prototype system for storing clinical signal data, where the time series data of clinical sensors are stored within HBase in a way that the row key serves as the time stamp of a single value, and the column stores patient physiological values that correspond with the row key time stamp. To improve the accessibility and readability of the HBase data schema, the metadata are stored in MongoDB,[62] which is a document-based NoSQL database. Google Web Toolkit is incorporated into the system to visualize the clinical signal data.

Interactive data retrieval for data sharing. Interactive medical information retrieval is expected to play an important role in sharing medical knowledge and integrating data. Many researchers have seen the need for such a role and have offered possible solutions. Deb and Srirama[63] proposed a three-tier ecosystem to improve the shortcomings of cloud-enabled social networks for eHealth Solutions. Bahga and Madisetti[64] developed a cloud-based approach for interoperable EHRs. Sharp[65] proposed an application architecture based on the cloud approach to enhance the interaction between researchers in multisite clinical trials. Chen et al.[66] discussed the present and future aspects of translational informatics based on the cloud approach. He et al.[67] provided a private cloud platform architecture for handling enormous data requests from health-care services. To handle huge amounts of online heart disease data analyses in China, Wang et al.[68] used a hybrid XML database and the Hadoop/HBase infrastructure to design the "Clinical Data Managing and Analyzing System."

Data security. Schultz[69] concluded that vast amounts of data can be collected over time and that health-care challenges could be met and addressed in response to big data opportunities. This in turn means that major data technology advancements will enable health-care practitioners to manipulate even larger amounts of data in the future. However, interactive data retrieval places greater pressure on data security. Sobhy et al.[70] proposed *MedCloud*, a model that leverages the Hadoop ecosystem to consider compliance issues with HIPAA when accessing patient data. Lin et al.[71] proposed *Home-Diagnosis*, a cloud-based framework to address challenges with privacy protection, ensure highly concurrent and scalable medical record retrieval, and conduct data analysis in a self-caring setting. To solve these major challenges, a Lucene-based distributed search cluster was used in Home-Diagnosis, while a Hadoop cluster was employed to speed up the process overall.

Data analysis. Predicting disease risk and progression over time can be very useful for clinical decision support, and building computational models for clinical prediction requires a complex pipeline. Ng et al.[72] proposed *PARAMO* as a predictive modeling platform for analyzing electronic health data. PARAMO supports the generation and reuse of a clinical data analysis pipeline for different modeling purposes. To efficiently process parallel tasks, PARAMO supports MapReduce, which analyzes data for an immense amount of medical data that can be processed in a reasonable time. Medical terminology ontologies (eg, ICD, UMLS) were integrated into the PARAMO system, and the analysis was tested on a set of EHR data from 5,000 to 300,000 patients using a Hadoop cluster; the concurrent task varies from 10 to 160. Results show that on this large dataset, 160 concurrent tasks are 72 times faster than running on 10 concurrent tasks.

In addition, Zolfaghar et al.[73] used big data techniques to study the 30-day risk of readmission for congestive heart failure patients. The patient data were extracted from the National Inpatient Dataset and the Multicare Health System. Several algorithms (eg, logistic regression, random forest) were used to build a predictive model to analyze the possibility of patient readmission. The investigators performed several tests on more than three million patient records. The results showed that the use of big data significantly increased the performance of building a predictive model: the models achieved the highest accuracy at 77% and recall at 61%.

Deligiannis et al.[74] presented a data-driven prototype using MapReduce to diagnose hypertrophic cardiomyopathy (HCM), an inherited heart disease that causes cardiac death in young athletes. Successful diagnosis of HCM is challenging due to the large number of potential variables. Deligiannis et al believed that the diagnosis rate could be improved by using a data-driven analysis. In addition to improved predictive accuracy, the experimental results showed that the overall runtime of predictive analysis decreased from nine hours to only a few minutes when accessing a dataset of 10,000 real medical records – this is a remarkable improvement over previous analyses and could lead to possible future applications for early systematic diagnoses.

Furthermore, the use of big data to analyze clinical data could have a significant impact on the medical community. A number of researchers have described future possibilities for the application of big data analytics. Ghani et al.[75] argued that the adoption of EHRs and the use of picture archiving and communication systems (PACS) have led to the capture of mass quantities of digital big data. They also inferred that urologists can use big data analytics for decision support, such as predicting whether a patient will need readmission to hospital after a cystectomy. Ghani et al anticipated that analytics of big data can also be applied to determine whether radiation therapy or prostatectomy should be used for a 75-year-old patient to avoid immediate risks from advanced prostate cancer. Wang and Krishnan[76] gave a systematic review of how big data can facilitate outcomes, such as identifying the causality of patient symptoms, predicting hazards of disease incidence or reoccurrence, and improving primary care quality.

Genta and Sonnenberg[77] provided an overview of big data in gastroenterology research, stating that the big data method is a new tool for finding significant association among large amounts of "messy" clinical data. Furthermore, the use of a large dataset will rapidly expand for gastroenterologists and advance the understanding of digestive diseases. Chawla and

Davis[78] illustrated the overall vision of the big data approach to personalized medicine and provided a patient-centered framework. Abbott[79] explained the contribution of big data to perioperative medicine. McGregor[80] contended that using big data could help predict deadly pediatric medical conditions at an early stage, leading to a breakthrough in clinical applications for neonatal intensive care units. Fahim et al.[81] proposed a system for active lifestyles and argued that a visual design engages users by enhancing their self-motivation.

Imaging informatics applications. Imaging informatics is the study of methods for generating, managing, and representing imaging information in various biomedical applications. It is concerned with how medical images are exchanged and analyzed throughout complex health-care systems. With the growing need for more personalized care, the need to incorporate imaging data into EHRs is rapidly increasing.

In this section, we classified big data technologies/tools into three categories: (1) data storage and retrieval, (2) data sharing, and (3) data analysis. Imaging informatics developed almost simultaneously with the advent of EHRs and the emergence of clinical informatics; however, it is very different from clinical informatics due to the heterogeneous data types generated from different modalities of medical images. Data security remains an important consideration in this area, but because current systems primarily rely on commercial cloud platforms and existing protocols, such as digital image communication in medicine (DICOM), there is no research focusing on improving data security in imaging informatics.

Data storage and retrieval. Imaging informatics is predominantly used for improving the efficiency of image processing workflows, such as storage, retrieval, and interoperation. PACS are popular for delivering images to local display workstations, which is accomplished primarily through DICOM protocols in radiology departments. Many web-based medical applications have been developed to access PACS, and greater use of big data technology has been improving their performance. Silva et al.[82] proposed an approach to integrate the data in PACS, given the current trend among health-care institutions to outsource the two important components of PACS (DICOM object repository and database system) to the cloud. Silva et al proposed to provide an abstract layer with a Cloud IO (input/output) stream mechanism to support more than one cloud provider despite their differences in data access standards.

In addition to big data technologies based on the implementation of cloud platforms with PACS, Yao et al.[83] developed a massive Hadoop-based medical image retrieval system that extracted the characteristics of medical images using a Brushlet transform and a local binary pattern algorithm. Then, the HDFS stored the image features, followed by the implementation of MapReduce. The evaluation results indicated a decreased error rate in images compared with the result without homomorphic filtering. Similarly, Jai-Andaloussi et al.[84] used the MapReduce computation model and HDFS storage model to address the challenges of content-based image retrieval systems. They performed experiments on mammography databases and obtained promising results, showing that the MapReduce technique can be effectively used for content-based medical image retrieval.

Data and workflow sharing. PACS primarily provide image data archiving and analysis workflow at single sites. Radiology groups operating under a disparate delivery model (ie, different services offered by different vendors to complete a single radiology task) face significant challenges in a data-sharing infrastructure. Benjamin et al.[85] developed *SuperPACS*, a system that enables a radiology group that serves multiple sites and has disparate PACS, RIS, reporting, and other relevant IT systems to view these sites virtually from one site and to use one virtual desktop to efficiently complete all radiology work and reporting. SuperPACS provides two approaches: (1) the federated approach, in which all patient data stay local, and (2) the consolidated approach, in which the data are stored centrally by a single agent. The agent is able to (1) provide an interface for DICOM, HL7, HTTP, and XDS standard and nonstandardized data; (2) synchronize metadata on local PACS and RIS; (3) cache images and data received from local PACS, RIS, any input tool, or another agent; (4) archive image data with compression and multitier storage, backup, disaster recovery, and image and data lifecycle management; (5) provide worklists, folders, routing logic, and mechanisms for image and nonimage data; (6) distribute image data through a web server, including compression and streaming; and (7) access local and remote data from a SuperPACS web client.

Data analysis. Seeking to overcome the challenges brought by large-scale (terabytes or petabytes) data derived from pathological images, Wang et al.[86] proposed *Hadoop-GIS*, an efficient and cost-effective parallel system. Here, GIS refers to spatially derived data management applications, which enable real-time spatial queries with the Real-time Spatial Query Engine, and integrates both MapReduce-based spatial query processing and Hive-based feature query processing. The Hadoop-GIS system also offers an easier SQL-like declarative query language that is supported by Hive. In the performance study, Wang et al used (1) a small-sized cluster for prototype tests and (2) a medium-sized cluster for scalability tests on real-world data. The results show that Hadoop-GIS increased query efficiency and decreased loading and query time.

To analyze cardiac imaging and medical data to optimize clinical diagnosis and treatment, Dilsizian and Siegel[87] proposed a framework to integrate AI, massive parallel computing, and big data mining and argued that these technologies are critical components for evidence-based personalized medicine. They also argued that big data mining techniques would be used for next-generation AI techniques in which large numbers of possible factors (eg, whether a patient had myocardial infarction) could be analyzed and a prediction could be completed in less time, thereby improving diagnosis and treatment. Using a cardiac imaging field as a focus area, Dilsizian and Siegel showed that the Formation of Optimal Cardiovascular Utilization Strategies group introduced the use of AI and big data to reduce inappropriate uses of diagnostic imaging; such cases decreased from 10% to 5% among the 55 participating sites.

In addition, Markonis et al.[88] used Hadoop to establish a cluster of computing nodes and MapReduce to speed up the process. Three cases of use were analyzed: (1) parameter optimization for lung texture classification using support vector machines (SVMs), (2) content-based medical image indexing, and (3) three-dimensional directional wavelet analysis for solid texture classification. Test results in a parallel grid search for optimal SVM parameters showed that using concurrent map tasks reduced the total runtime from 50 hours to 9 hours 15 minutes – a significant improvement in computing efficiency while maintaining a good classification performance.

Public health information. As described by Shortliffe and Cimino,[89] public health has three core functions: (1) assessment, (2) policy development, and (3) assurance. Among these, assessment is the prerequisite and fundamental function. Assessment primarily involves collecting and analyzing data to track and monitor public health status, thereby providing evidence for decision making and policy development. Assurance is used to validate whether the services offered by health institutions have achieved their initial target goals for increasing public health outcomes; as such, many large public health institutions, such as the Centers for Disease Control and Prevention and the Administration of Community Living, have collected and analyzed very large amounts of population health data.

In this section, no new approaches are introduced. Instead, we present an integrated view of big data and health from a population perspective rather than a single medical/clinical activity perspective. This section focuses on four areas: (1) infectious disease surveillance, (2) population health management, (3) mental health management, and (4) chronic disease management.

Infectious disease surveillance. Hay et al.[90] discussed the opportunities for using big data for global infectious disease surveillance. They developed a system that provides real-time risk monitoring on map, pointing out that machine learning and crowdsourcing have opened new possibilities for developing a continually updated atlas for disease monitoring. Hay et al believed that online social media combined with epidemiological information is a valuable new data source for facilitating public health surveillance. The use of social media for disease monitoring was demonstrated by Young et al.[91], in which they collected 553,186,016 tweets and extracted more than 9,800 with HIV risk-related keywords (eg, sexual behaviors and drug use) and geographic annotations. They showed that there is a significant positive correlation ($P < 0.01$) between HIV-related tweets and HIV cases based on prevalence analysis, illustrating the importance of social media (eg, Twitter, Facebook) and its potential impact on monitoring global disease occurrence.

Population health management. To study the distribution and impact of sociodemographic and medico-administrative factors, Lamarche-Vadel et al.[92] analyzed the independent association of patient MD and UCD. The MD was identified by ICD10 code, while the UCD was extracted from a death registry. If MD and UCD were different events, then those events were found to be independent. Using health insurance data, information from 421,460 deceased patients was extracted from 2008 to 2009. The results show that 8.5% of inhospital deaths and 19.5% of out-of-hospital deaths were independent events and that independent death was more common in elderly patients. The results demonstrate that large-scale data analysis can be used to effectively analyze the association of medical events.

Mental health management. Nambisan et al.[93] found that messages posted on social media could be used to screen for and potentially detect depression. Their analysis is based on previous research of the association between depressive disorders and repetitive thoughts/ruminating behavior. Big data analytics tools play an important role in their work by mining hidden behavioral and emotional patterns in messages, or "tweets," posted on Twitter. Within these tweets, we may be able to detect a disease-related emotion pattern, which is a previously hidden symptom. The authors foresee that future research could delve deeper into the conversations of the depressed users to understand more about their hidden emotions and sentiments. In addition, Dabek and Caban[94] presented a neural network model that can predict the likelihood of developing psychological conditions, such as anxiety, behavioral disorders, depression, and post-traumatic stress disorder. They also analyzed the effectiveness of their model against a dataset of 89,840 patients, and the results show that they can achieve an overall accuracy of 82.35% for all conditions.

Chronic disease management. Tu et al.[95] introduced the Cardiovascular Health in Ambulatory Care Research Team (CANHEART), a unique, population-based observational research initiative aimed at measuring and improving cardiovascular health and the quality of ambulatory cardiovascular care provided in Ontario, Canada. The research focused on identifying opportunities to improve the primary and secondary prevention of cardiovascular events in Ontario's diverse multiethnic population. The study included data from 9.8 million Ontario adults aged ≥20 years. Data were assembled by linking multiple databases, such as electronic surveys, health administration, clinical, laboratory, drug, and electronic medical record databases using encoded personal identifiers. Follow-up clinical events were collected through record linkages to comprehensive hospitalization, emergency department, and vital statistics administrative databases. The huge linked databases enable the CANHEART study cohort to serve as a powerful big data resource for scientific research aimed at improving cardiovascular health and health services delivery.

Kupersmith et al.[96] introduced the health IT infrastructure in the US Veterans Health Administration's (VHA) health information infrastructure and the factors that made it possible to achieve chronic disease management for its patients. Structured clinical data in the EHRs can be aggregated within specialized databases, while unstructured text data, such as clinician notes, can be reviewed and abstracted electronically from a central location. The rich clinical information makes it possible for professionals to extract insights; for instance, the VHA has identified a high rate of mental illness comorbidity (24.5%) among patients with diabetes. The VHA also uses EHR data to explore the influence of sex and race/ethnicity and to understand the extent to which newer psychotropic drugs contribute to poor outcomes, in the context that drugs promote weight gain and mental illness itself. The VHA also uses this information to identify and track diabetic complications, such as early chronic kidney disease without renal impairment, as indicated in the record. After identifying patients at high risk for comorbidities or amputation, the VHA distributes the information to clinicians to better coordinate patient care.

Conclusion

We are currently in the era of "big data," in which big data technology is being rapidly applied to biomedical and health-care fields. In this review, we demonstrated various examples in which big data technology has played an important role in modern-day health-care revolution, as it has completely changed people's view of health-care activity. The first three sections of this review revealed that big data applications facilitate three important clinical activities, while the last section (especially the chronic disease management section) draws an integrated picture of how separate clinical activities are completed in a pipeline to manage individual patients from multiple perspectives. We summarized recent progress in the most relevant areas in each field, including big data storage and retrieval, error identification, data security, data sharing and data analysis for electronic patient records, social media data, and integrated health databases.

Furthermore, in this review, we learned that bioinformatics is the primary field in which big data analytics are currently being applied, largely due to the massive volume and complexity of bioinformatics data. Big data application in bioinformatics is relatively mature, with sophisticated platforms and tools already in use to help analyze biological data, such as gene sequencing mapping tools. However, in other biomedical research fields, such as clinical informatics, medical imaging informatics, and public health informatics, there is enormous, untapped potential for big data applications.

This literature review also showed that: (1) integrating different sources of information enables clinicians to depict a new view of patient care processes that consider a patient's holistic health status, from genome to behavior; (2) the availability of novel mobile health technologies facilitates real-time data gathering with more accuracy; (3) the implementation of distributed platforms enables data archiving and analysis, which will further be developed for decision support; and (4) the inclusion of geographical and environmental information may further increase the ability to interpret gathered data and extract new knowledge.

While big data holds significant promise for improving health care, there are several common challenges facing all the four fields in using big data technology; the most significant problem is the integration of various databases. For example, the VHA's database, VISTA, is not a single system; it is a set of 128 interlinked systems. This becomes even more complicated when databases contain different data types (eg, integrating an imaging database or a laboratory test results database into existing systems), thereby limiting a system's ability to make queries against all databases to acquire all patient data. The lack of standardization for laboratory protocols and values also creates challenges for data integration. For example, image data can suffer from technological batch effects when they come from different laboratories under different protocols. Efforts are made to normalize data when there is a batch effect; this may be easier for image data, but it is intrinsically more difficult to normalize laboratory test data. Security and privacy concerns also remain as hurdles to big data integration and usage in all the four fields, and thus, secure platforms with better communication standards and protocols are greatly needed.

In its latest industry analysis report, McKinsey & Company predicted that big data analytics for the medical field will potentially save more than $300 billion per year in US health-care costs. Future development of big data applications in the biomedical fields holds foreseeable promise because it is dependent on the advancement of new data standards, relevant research and technology, cooperation in research institutions and companies, and strong government incentives.

Author Contributions

JL, MW conceived and designed the experiments. MW, JL jointly developed the structure and arguments for the paper. JL, MW, YZ, DG analyzed the data. JL, MW, DG, YZ wrote the first draft of the manuscript. JL, MW, YZ contributed to the writing of the manuscript. All authors reviewed and approved the final manuscript.

REFERENCES

1. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature*. 2009;458(7239):719–24.
2. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26(10):1135–45.
3. Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet*. 2010;11(1):31–46.
4. Nielsen R, Paul JS, Albrechtsen A, et al. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*. 2011;12(6):443–51.
5. Zhang J, Chiodini R, Badr A, et al. The impact of next-generation sequencing on genomics. *J Genet Genomics*. 2011;38(3):95–109.
6. Murdoch TB, Detsky AS. The inevitable application of big data to healthcare. *JAMA*. 2013;309(13):1351–2.

7. Lynch C. Big data: how do your data grow? *Nature*. 2008;455(7209):28–9.
8. Wilhelm M, Schlegl J, Hahne H, et al. Mass-spectrometry-based draft of the human proteome. *Nature*. 2014;509(7502):582–7.
9. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med*. 2010;363(6):501–4.
10. Botsis T, Hartvigsen G, Chen F, et al. Secondary use of EHR: data quality issues and informatics opportunities. In: AMIA Summits on Translational Science Proceedings, San Francisco, California; AMIA. 2010:1.
11. Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. *J Biomed Inform*. 2012;45(4):763–71.
12. Ackerman MJ. The Visible Human Project: a resource for education. *Acad Med*. 1999;74(6):667–70.
13. Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*. 2006;7(3):198–210.
14. Feldman R, Sanger J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press; 2007.
15. LePendu P, Iyer SV, Fairon C, et al. Annotation analysis for testing drug safety signals using unstructured clinical notes. *J Biomed Semantics*. 2012; 3(suppl 1):S5.
16. Rosenbloom ST, Denny JC, Xu H, et al. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc*. 2011;18(2):181–6.
17. Luo Z, Duffy R, Johnson S, et al. Corpus-based approach to creating a semantic lexicon for clinical research eligibility criteria from UMLS. In: AMIA Joint Summit of Translational Informatics, San Francisco, California; AMIA. 2010:26–31.
18. Weng C, Wu X, Luo Z, et al. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc*. 2011;18:i116–24.
19. Reiner BI. Medical imaging data reconciliation, part 3: reconciliation of historical and current radiology report data. *J Am Coll Radiol*. 2011;8(11):768–771.
20. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–303.
21. Chou W-YS, Hunt YM, Beckjord EB, et al. Social media use in the United States: implications for health communication. *J Med Internet Res*. 2009;11(4):e48.
22. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin Infect Dis*. 2009;49(10):1557–64.
23. Margolis R, Derr L, Dunn M, et al. The National Institutes of Health's Big Data to Knowledge (BD2 K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc*. 2014;21(6):957–8.
24. White T. *Hadoop: The Definitive Guide*. Sebastopol, CA: O'Reilly Media, Inc.; 2012.
25. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM*. 2008;51(1):107–13.
26. Armbrust M, Fox A, Griffith R, et al. A view of cloud computing. *Commun ACM*. 2010;53(4):50–8.
27. Schuster SC. Next-generation sequencing transforms today's biology. *Nature*. 2007;200(8):16–8.
28. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. *Genomics*. 2008;92(5):255–64.
29. Taylor R. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics*. 2010;11(suppl 12):S1.
30. Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*. 2009;25(11):1363–9.
31. Schatz M, Sommer D, Kelley D, et al. Contrail: assembly of large genomes using cloud computing. In: CSHL Biology of Genomes Conference, Cold Spring Harbor, New York; CSHL. 2010.
32. Gurtowski J, Schatz MC, Langmead B. Genotyping in the cloud with crossbow. *Curr Protoc Bioinformatics*. 2012;Chapter 15:Unit15.3.
33. Pandey RV, Schlötterer C. DistMap: a toolkit for distributed short read mapping on a Hadoop cluster. *PLoS One*. 2013;8(8):e72614.
34. O'Connor BD, Merriman B, Nelson SF. SeqWare query engine: storing and searching sequence data in the cloud. *BMC Bioinformatics*. 2010;11(suppl 12):S2.
35. George L. *HBase: The Definitive Guide*. Sebastopol, CA: O'Reilly Media, Inc.; 2011.
36. Nagasaki H, Mochizuki T, Kodama Y, et al. DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. *DNA Res*. 2013;20(4):383–90.
37. Lewis S, Csordas A, Killcoyne S, et al. Hydra: a scalable proteomic search engine which utilizes the Hadoop distributed computing framework. *BMC Bioinformatics*. 2012;13(1):324.
38. Robinson T, Killcoyne S, Bressler R, et al. SAMQA: error classification and validation of high-throughput sequenced read data. *BMC Genomics*. 2011;12(1):419.
39. Huang W, Li L, Myers JR, et al. ART: a next-generation sequencing read simulator. *Bioinformatics*. 2012;28(4):593–4.
40. Chen C-C, Chang Y-J, Chung W-C, et al. CloudRS: an error correction algorithm of high-throughput sequencing data based on scalable framework. In: 2013 IEEE International Conference on Big Data, Santa Clara, California; IEEE, 2013:717–22.
41. Gnerre S, MacCallum I, Przybylski D, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A*. 2011;108(4):1513–8.
42. Salzberg SL, Phillippy AM, Zimin A, et al. GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res*. 2012;22(3):557–67.
43. Yang X, Dorman KS, Aluru S. Reptile: representative tiling for short read error correction. *Bioinformatics*. 2010;26(20):2526–33.
44. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;11(1110):11.10.1-11.10.33.
45. Rustici G, Kolesnikov N, Brandizi M, et al. ArrayExpress update – trends in database growth and links to data analysis tools. *Nucleic Acids Res*. 2013;41(D1): D987–90.
46. Brazma A, Parkinson H, Sarkans U, et al. ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*. 2003;31(1):68–71.
47. Huang H, Tata S, Prill RJ. BlueSNP: R package for highly scalable genome-wide association studies using Hadoop clusters. *Bioinformatics*. 2013;29(1):135–6.
48. Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol*. 2010;11(8):R83.
49. Jourdren L, Bernard M, Dillies M-A, et al. Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. *Bioinformatics*. 2012;28(11):1542–3.
50. Wiewiórka MS, Messina A, Pacholewska A, et al. SparkSeq: fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision. *Bioinformatics*. 2014;30(18):2652–3.
51. Schumacher A, Pireddu L, Niemenmaa M, et al. SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop. *Bioinformatics*. 2014;30(1): 119–20.
52. Niemenmaa M, Kallio A, Schumacher A, et al. Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. *Bioinformatics*. 2012;28(6):876–7.
53. Angiuoli SV, Matalka M, Gussman A, et al. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics*. 2011;12(1):356.
54. Krampis K, Booth T, Chapman B, et al. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC Bioinformatics*. 2012;13(1):42.
55. Chung W-C, Chen C-C, Ho J-M, et al. CloudDOE: a user-friendly tool for deploying Hadoop clouds and analyzing high-throughput sequencing data with MapReduce. *PLoS One*. 2014;9(6):e98146.
56. Dutta H, Kamil A, Pooleery M, et al. Distributed storage of large-scale multidimensional electroencephalogram data using Hadoop and HBase. In: Fiore S, Aloisio G, eds. *Grid and Cloud Database Management*. Berlin: Springer; 2011:331–47.
57. Jin Y, Deyu T, Yi Z. A distributed storage model for EHR based on HBase. In: 2011 International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII), Shenzhen, China; IEEE. 2011:369–72.
58. Sahoo SS, Jayapandian C, Garg G, et al. Heart beats in the cloud: distributed analysis of electrophysiological 'Big Data' using cloud computing for epilepsy clinical research. *J Am Med Inform Assoc*. 2014;21(2):263–71.
59. Jayapandian CP, Chen C-H, Bozorgi A, et al. Cloudwave: distributed processing of "Big Data" from electrophysiological recordings for epilepsy clinical research using Hadoop. In: AMIA Annual Symposium Proceedings, Washington, DC; AMIA. 2013:691.
60. Mazurek M. Applying NoSQL databases for operationalizing clinical data mining models. In: Kozielski S, Mrozek D, Kasprowski P, Małysiak-Mrozek B, Kostrzewa D, eds. *Beyond Databases, Architectures, and Structures*. Berlin: Springer; 2014:527–36.
61. Nguyen AV, Wynden R, Sun Y. HBase, MapReduce, and integrated data visualization for processing clinical signal data. In: AAAI Spring Symposium: Computational Physiology, Palo Alto, California; AAAI, 2011.
62. Banker K. *MongoDB in Action*. Greenwich, CT: Manning Publications Co; 2011.
63. Deb B, Srirama SN. Social networks for eHealth solutions on cloud. *Front Genet*. 2013;4:171.
64. Bahga A, Madisetti VK. A cloud-based approach for interoperable electronic health records (EHRs). *IEEE J Biomed Health Inf*. 2013;17(5):894–906.
65. Sharp J. An application architecture to facilitate multi-site clinical trial collaboration in the cloud. In: Proceedings of the 2nd International Workshop on Software Engineering for Cloud Computing, Honolulu, Hawaii; ACM. 2011:64–8.
66. Chen J, Qian F, Yan W, et al. Translational biomedical informatics in the cloud: present and future. *Biomed Res Int*. 2013;2013:658925.
67. He C, Fan X, Li Y. Toward ubiquitous healthcare services with a novel efficient cloud platform. *IEEE Trans Biomed Eng*. 2013;60(1):230–4.
68. Wang Y, Wang L, Liu H, et al. Large-scale clinical data management and analysis system based on cloud computing. In: Li S, Jin Q, Jiang X, Park JJJH, eds. *Frontier and Future Development of Information Technology in Medicine and Education*. Berlin: Springer; 2014:1575–83.
69. Schultz T. Turning healthcare challenges into big data opportunities: a use-case review across the pharmaceutical development lifecycle. *B Am Soc Inform Inf Sci Technol*. 2013;39(5):34–40.

70. Sobhy D, El-Sonbaty Y, Abou Elnasr M. MedCloud: healthcare cloud computing system. In: 2012 International Conference for Internet Technology and Secured Transactions, London, UK; IEEE. 2012:161–6.
71. Lin W, Dou W, Zhou Z, et al. A cloud-based framework for home-diagnosis service over big medical data. *J Syst Software*. 2015;102:192–206.
72. Ng K, Ghoting A, Steinhubl SR, et al. PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. *J Biomed Inform*. 2014;48:160–70.
73. Zolfaghar K, Meadem N, Teredesai A, et al. Big data solutions for predicting risk-of-readmission for congestive heart failure patients. In: 2013 IEEE International Conference on Big Data, Santa Clara, California; IEEE. 2013:64–71.
74. Deligiannis P, Loidl H-W, Kouidi E. Improving the diagnosis of mild hypertrophic cardiomyopathy with MapReduce. In: Proceedings of Third International Workshop on MapReduce and its Applications Date, Delft, Netherlands; ACM. 2012:41–8.
75. Ghani KR, Zheng K, Wei JT, et al. Harnessing big data for healthcare and research: are urologists ready? *Eur Urol*. 2014;66(6):975–7.
76. Wang W, Krishnan E. Big data and clinicians: a review on the state of the science. *JMIR Med Inform*. 2014;2(1):e1.
77. Genta RM, Sonnenberg A. Big data in gastroenterology research. *Nat Rev Gastroenterol Hepatol*. 2014;11(6):386–90.
78. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. *J Gen Intern Med*. 2013;28(3):660–5.
79. Abbott R. Big data and pharmacovigilance: using health information exchanges to revolutionize drug safety. *Iowa L Rev*. 2013;99:225.
80. McGregor C. Big data in neonatal intensive care. *Computer*. 2013;46(6):54–9.
81. Fahim M, Idris M, Ali R, et al. ATHENA: a personalized platform to promote an active lifestyle and wellbeing based on physical, mental and social health primitives. *Sensors*. 2014;14(5):9313–29.
82. Silva LA, Costa C, Oliveira JL. A PACS archive architecture supported on cloud services. *Int J Comput Assist Radiol Surg*. 2012;7(3):349–58.
83. Yao Q-A, Zheng H, Xu Z-Y, et al. Massive medical images retrieval system based on Hadoop. *J Multimed*. 2014;9(2):216–2.
84. Jai-Andaloussi S, Elabdouli A, Chaffai A, et al. Medical content based image retrieval by using the Hadoop framework. In: 2013 20th International Conference on Telecommunications (ICT), Casablanca, Morocco; IEEE. 2013:1–5.
85. Benjamin M, Aradi Y, Shreiber R. From shared data to sharing workflow: merging PACS and teleradiology. *Eur J Radiol*. 2010;73(1):3–9.
86. Wang F, Lee R, Liu Q, et al. Hadoop-GIS: A High Performance Query System for Analytical Medical Imaging with MapReduce: Technical Report; 2011; Atlanta: Emory University.
87. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep*. 2014;16(1):1–8.
88. Markonis D, Schaer R, Eggel I, et al. Using MapReduce for large-scale medical image analysis. In: 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB), La Jolla, California; IEEE. 2012:1.
89. Shortliffe EH, Cimino JJ. *Biomedical Informatics*. Berlin: Springer; 2014.
90. Hay SI, George DB, Moyes CL, et al. Big data opportunities for global infectious disease surveillance. *PLoS Med*. 2013;10(4):e1001413.
91. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Prev Med*. 2014;63(0):112–5.
92. Lamarche-Vadel A, Pavillon G, Aouba A, et al. Automated comparison of last hospital main diagnosis and underlying cause of death ICD10 codes, France, 2008–9. *BMC Med Inform Decis Mak*. 2014;14(1):44.
93. Nambisan P, Luo Z, Kapoor A, et al. Social media, big data, and public health informatics: ruminating behavior of depression revealed through twitter. In: 2015 48th Hawaii International Conference on IEEE System Sciences (HICSS), Honolulu, Hawaii; IEEE. 2015:2906–13.
94. Dabek F, Caban JJ. Brain informatics and health. In: Guo Y, Friston K, Aldo F, Hill S, Peng H, eds. *A Neural Network Based Model for Predicting Psychological Conditions*. Cham: Springer International Publishing; 2015:252–61.
95. Tu JV, Chu A, Donovan LR, et al. The Cardiovascular Health in Ambulatory Care Research Team (CANHEART) using big data to measure and improve cardiovascular health and healthcare services. *Circ Cardiovasc Qual Outcomes*. 2015;8(2):204–12.
96. Kupersmith J, Francis J, Kerr E, et al. Advancing evidence-based care for diabetes: lessons from the Veterans Health Administration[J]. *Health Affairs*. 2007; 26(2): w156–68.