American Journal of Engineering and Applied Sciences

Review

Healthcare Driven by Big Data Analytics

[1]Cheryl Ann Alexander and [2]Lidong Wang

[1]Technology and Healthcare Solutions, Inc., Vicksburg, MS, USA
[2]Institute for Systems Engineering Research, Mississippi State University, Vicksburg, MS, USA

Article history
Received: 08-07-2018
Revised: 11-09-2018
Accepted: 27-09-2018

Corresponding Author:
Cheryl Ann Alexander
Technology and Healthcare Solutions, Inc., Vicksburg, MS, USA
Email: tech.health.sol@gmail.com

Abstract: Text messages and social network posts are often included in big data and are frequently invaluable sources of health data. When machine learning or data mining is used, it is important to perform an automatic process for integrating all available and related health data. Artificial Neural Network (ANN) is one of the machine learning methods flexible in using algorithms to detect complicated nonlinear relationships within huge datasets. Pharmacokinetics uses genetics to individualize drug therapy. Because genetic and pharmacological analysis often require large scale computation, Big Data analytics has shown potential in this area. Personal data from various sensors can often be used for making health and treatment recommendations, taking appropriate action for a patient's lifestyle choices and early diagnosis vital to advancing quality of care. Big data techniques such as hadoop and spark have been used in various areas of healthcare and big data analytics has been employed in great datasets to expose hidden patterns or correlations for effective decision-making. Challenges of big data in healthcare, concerned with gathering material from multifaceted heterogonous patient sources still exist, although hadoop has been employed as the processing unit for heterogeneous data gathered from various body sensors. The most critical necessity in a healthcare big data system is the data security; however, users are continuously applying big data–driven strategies to solve the various problems and challenges.

Keywords: Big Data Analytics, Healthcare, Artificial Intelligence, Personalized Medicine, Precision Medicine

Introduction

The first national survey on the real-life clinical setting of stroke care is the J-ASPECT (Nationwide survey of Acute Stroke care capacity for Proper designation of Comprehensive stroke center) study, conducted in Japan, which used data collected from the diagnosis-procedure combination-based payment system, which established a substantial relationship between the Comprehensive Stroke Care (CSC) capacity and the volume of stroke interventions in the hospital; The study also confirmed that CSC capabilities correlated to reducing inpatient mortality rates. The use of big data in stroke care has garnered substantial notice as a vital resource for generating new evidence-based research. Several researchers attempted to uncover new verification and awareness from a real neurosurgical clinic, focusing on stroke care using Big Data, therefore developing more efficient methods for bridging the evidence-based practice gap in acute stroke knowledge (Nishimura *et al.*, 2016).

Health big data has been classified as "encompassing high volume, high diversity biological, clinical, environmental and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points" (Knoppers and Thorogood, 2017). A primary goal for healthcare organizations has recently been utilizing big data in merging -omic data from patients with behavioral, environmental, social and clinical data to increase efficiency and quality of care (Karimi *et al.*, 2016). A variety of sources can be utilized to collect large-scale data. For Example, some types of healthcare data that is available for Big Data Analytics (BDA) consists of hospital data, clinical data, genomic data, streamed data, global health survey data, clinical reference data and health publications, WHO (World Health Organization) repositories, web and social networking data (Adil *et al.*, 2015).

Big Data applications promises to consider multiple levels of health information which rests in the prospect of merging and assimilating de-identified health data to take into consideration the secondary data usage. Therefore, included is the usage and re-usage of diverse supplies of health information for such purposes as not only personal clinical treatment of exclusive patients by caregivers, or when researchers directly investigate precise biomedical research hypotheses. Some contemporary applications include: Conducting audits and benchmarking studies, financial and service planning; how researchers facilitate the recruitment of subjects to randomized controlled trials; epidemiological and pharmacovigilance analyses; and finally reinforcing the cohort of new biomedical research outcomes (Cano *et al.*, 2017).

Medical big data depends upon three types of electronic data streams: Medical encounter data, participatory syndromic data and non-health digital data. Medical encounter data include electronic records from healthcare facilities (e.g., imaging, EMRs, etc.), hospital discharge records, medical insurance claims, death certificates, etc. Participatory syndromic data are crowd-sourced data include data where volunteers self-report diverse symptoms. On the other hand, non-health digital data do not depend upon a patient or medical encounter, but are a consequence of mobile phone usage, social media networking and Internet search engines (Bansal *et al.*, 2016). Healthcare is undergoing significant changes. Critical propel agents of swift and key transformation include high-resolution imaging and whole-genome sequencing technologies (Auffray *et al.*, 2016). Digital data sources such as call data records from cellphones, medical claims data and geographically tagged tweets have appeared in infectious disease epidemiology as new sources of data to supplement outdated infectious disease surveillance. Digital data streams applicable to public health may continue to act as proxies for some time yet. In regions with meager coverage from traditional public health surveillance and conventional and digital sources of spatial big data, harnessing disease data from digital sources may permit scientists to conduct epidemiological analyses at finer spatial scales and may allow a combination to account for the bias and gaps in each (Lee *et al.*, 2016).

Most current equipment is linked to evidence-based guidelines and recommendations based on the results of large Randomized Clinical Trials (RCTs) and meta-analyses containing information applicable to the average patient. However, physicians most often do not care for average patients; therefore, generalizable knowledge useful in standardizing medical practice is generally not the most fitting for individual patient care. The difficulty is to advance more 'personalized guidelines' which reflect the patient heterogeneity and supports physicians in personalizing assessment, treatment and evaluation plans-in other words, personalized medicine for clinical decision-making (Sacristán and Dilla, 2015).

A key element of precision medicine is sharing data with the patient. The patient is both the source of granular information derived from multiple sources as well as the recipient of more health information than has traditionally been shared with patients (Rothstein, 2017). Key areas for concern can be identified as: (1) ownership, (2) informed consent, (3) privacy (including anonymization and data protection), (4) epistemology and objectivity and (5) Big Data Divides formed between people who have or lack necessary resources to analyze increasingly large datasets (Mittelstadt and Floridi, 2016). The underlying assumption in the healthcare field is that data from several bases with their diversity are likely to generate information and knowledge. A major fear linked to the introduction of Big Data is the idea of cause and effect that needs to be examined particularly where people are interested in health. Forensics may be one discipline where the role of Big Data in the development of a social medicine approach may be particularly relevant and very hopeful; data is more complete, more accurate and realistic (Dimeglio *et al.*, 2016).

The analysis of massive datasets generated by instruments contributes to difficult computational challenges. For example, it is projected that variant calling from Next-Generation Sequencing (NGS) data alone will require around two trillion CPU hours by 2025; therefore, optimized algorithms, scalable toward large input datasets, are necessary for commonly used tasks in NGS data processing. Big data algorithms are often adapted for areas such as text and data mining, as are highly scalable algorithmic techniques (Schmidt and Hildebrandt, 2017). Literature research was conducted on the databases IEEE Xplore and Scopus for this paper. Keywords combinations were used for searching papers which were published between January 2015 and May 2018. The combinations include big data and medical, big data and healthcare and big data and healthcare. Duplicated papers were removed that were found from the two databases and 316 papers were selected for literature review.

Materials and Methods

Artificial Intelligence (AI) impacts medical practice through the application of natural language processing to "read" the expanding scientific literature and collate multiple Electronic Medical Records (EMR). Machines (computers) learning directly from medical data could deter clinical errors related to human cognitive biases and positively influence patient care. However, providers are essential to cognitive medical practice because AI is neither astute nor intuitive. Artificial neural networks have started to emerge as a leading AI method due to recent advances in algorithms, computational power and access to large datasets although AI incorporates a wide range of symbolic and statistical approaches to reasoning and learning. Correcting algorithm mistakes (training) adds to AI predictive model confidence as machines learn to detect patterns by processing big data through layered mathematical models (algorithms) (Miller and Brown, 2017).

There are three types of cognitive analysis which can be divided into physiological-signals-based cognitive analysis, video-based analysis and text-based analysis. Wearable devices gather a diverse range of information on the human emotion condition (including location information, physical information, social networking information, etc.); human big data analysis collects human emotions, next a distinct set of feedbacks are utilized to impact human emotions. Machine learning approaches can be utilized to find the most optimal combination of emotional states in a variety of environments based on various types of emotional data; however, the accuracy of cognitive analysis can be improved using the information fusion theory, probability theory, decision trees, fuzzy theory, genetic algorithms and semantic web technology (Ma *et al.*, 2017). Three main functionalities of data interpretation include: (1) General clinical reporting of summaries, e.g., historical reporting, statistical analyses and time series comparisons; (2) real-time reporting such as alerts and proactive notifications, real-time data navigation, operational Key Performance Indicators (KPIs) which could be made available or sent to interested users in the form of real-time dashboards; and (3) data visualization which is critical feature of Big Data analytics for extrapolating the meaning and visualizing the information (Wang *et al.*, 2018).

A convenient and effective mathematical tool for electing the most informative feature subset from original features is Rough Set Theory (RST) which can be utilized in processing information and knowledge with uncertainty, imprecision and vagueness. Principal Component Analysis (PCA) is one of feature extraction approaches which keep a minimal number of features while maintaining high efficiency and good accuracy when representing original data. Attribute or variable reduction through feature selection is a central application in RST. However, because there are often many redundant or non-redundant, relevant or irrelevant features in practical problems, it is not easy to extract features due to high interaction and interdependency (Ding *et al.*, 2018).

There are two paradigm divisions in big data computing: batch-oriented computing and real-time oriented computing (or stream computing). Batch computing collects, stores and processes data in batches to obtain anticipated results and is efficient in handling high volume data: Apache Hadoop is one example of batch-oriented computing, but the output can be significantly delayed depending upon the volume of data being processed. On the other hand, stream computing requires constant input and data outcomes and accentuates the velocity of data. The stream computing paradigm has been used in many applications such as Yahoo S4, Tweeter Storm and Microsoft Time Stream (Ta *et al.*, 2016). Some researchers categorize big data computing as batch processing, stream processing and hybrid processing. Table 1 (Wang *et al.*, 2017) lists three generations of technologies classified by distinct processing paradigms.

Hadoop, which supports distributed storage and computation of immense unstructured datasets across clusters of computers, is an Apache open source framework written in Java. Hadoop is also calculated to scale horizontally from a single server to thousands of machines, each offering local computation and storage (Adil *et al.*, 2015). The MapReduce framework is effective in enabling dynamical big datasets, however attribute reduction algorithms may lead to the high computation complexity. An attribute reduction algorithm was offered to systematically advance the attribute reduction efficacy based on the Multi-agent Consensus MapReduce (MCMAR) optimization model and co-evolutionary quantum Particle Swarm Optimization (PSO) with self-adaptive memeplexes. The results show that MCMAR has better feasibility and efficiency than previous algorithms, which can clearly enhance the superior performance of attribute reduction in big datasets (Ding *et al.*, 2018).

An acquisition service founded on Python software can devour messages from message queues. To decompose clinical data from textual to key-value format, this acquisition service runs Python parsers devoted to each medical device, which integrate data stemming from diverse medical devices in a JSON-style format. MongoDB, a document-oriented database supporting unstructured data, does not require time-consuming or expensive migrations when application requirements change (Carnevale *et al.*, 2017). The burden due to growing data volume may be frontrunner to a further utilization of big data methodology for next-generation sequencing (NGS) data analysis, but a simple application of current big data algorithms to sequencing problems will often be inadequate (Schmidt and Hildebrandt, 2017). Several big data processing technologies and tools are listed in Table 2 (Saravanakumar and Hanifa, 2017).

Table 1: Big Data technologies classified by processing paradigms

Paradigms	Technologies
Batch processing	Hadoop (HDFS, Hive, HBase), MapReduce, Spark, Sqoop, Pig, Tableau, Apache Mahout, Flume, Cascading, Dryad, Pentaho, Karmasphere, Skytree Server, Scribe, Jaspersoft BI Suite
Stream processing	S4, Strom, Flume, Splunk, Spark Streaming, Kafka, SQLstream, SAP Hana, Kestrel
Hybrid processing	SummingBird, Lambdoop

Table 2: Several big data processing technologies and platforms

Big data technologies	Description
R Studio	Open source software, powerful in data manipulations, calculations and graphs and charts generation.
Hadoop Common	Provides common utility and functions of the Java library.
Apache Hadoop	Developed for data storage, process and analysis.
Hadoop-YARN	Can separate resource management and processing components.
HDFS	Hadoop Distributed File Systems; a Hadoop storage system, fast distributing data in several nodes on a cluster.
MapReduce Arch. (Input, Splitting, Mapping, Shuffling, Reducing and Result)	Hadoop framework model for the parallel computation of massive datasets.
Spark Architecture	Supports interactive, batch and iterative streaming computation at the same run time.
Spark Environment	1. Spark MLib: ML Library includes various components and all learning algorithms. 2. Spark streaming: Real-time data processing; supporting Python Java and Scala. 3. Spark Graph X: A new API for the graph and graph parallel computation. 4. Spark SQL: Referred as Shark, a new model and technical component of Spark, including SQL and APIs.
IBM InfoSphere	Platform with an enterprise class foundation for various big data projects; IBM InfoSphere DataStage being an ETL tool and part of InfoSphere.
Hbase	Built on HDFS; able to store semi-structured and unstructured sparse data, supporting column-based DB storage (large table).
Mahout	Supporting various kinds of DM algorithms (classification, clustering and batch-based collaborative filtering).
Pig (Pig Latin)	Using textual language by Pig Latin in a large-scale analysis platform and producing a sequence of MapReduce programs on Hadoop cluster.
Storm	Providing distributed real-time computation.
Oozle	Running on Java Servlet–container-Tomcat, including workflow manager and Job coordinator.
HPCC	High performance computing cluster (HPCC) with better performance compared with Hadoop
Big top	For packaging and testing the Hadoop ecosystem.
GridGain	Alternate to MapReduce utilized in Hadoop, providing in-memory processing and therefore faster analysis.
Flume	Flexible and reliable architecture, moving large volume of log data.
Sqoop	Transferring data between RDBMS and Hadoop, individual tables/whole SQL databases being imported into HDFS.

Sensors in wearable medical devices can continuously generate immense data, often big data with a mix of structured and unstructured data; however, because big data is extremely complex, it is problematic to analyze and retrieve useful or valuable information which can assist in decision-making. To overcome this problem, a new architecture for the execution of the Internet of Things (IoT) to store and process scalable sensor big data for healthcare applications has been proposed. This novel architecture can be divided into two main sub architectures: Meta Fog-Redirection (MF-R) and Grouping and Choosing (GC) architecture. MF-R architecture employs big data technologies such as Apache Pig and Apache HB ase for the collection and storage of sensor data generated from diverse sensor devices. The anticipated GC architecture can be used to secure the integration of fog computing with cloud computing, as fog computing is used for enhancing the efficacy and decreasing the data volume which must be transferred from physical devices to the cloud (Manogaran *et al.*, 2018).

Typical applications fueled by smart clothing and big data clouds have been offered such as disease diagnosis, psychological care, emergency medical response and real-time tactile interaction, specifically electrocardiograph signals gathered by smart clothing worked for detecting emotion and monitoring mood. Most technologies can be combined in Beyond Smart Clothing on Cloud (BSC), for example, machine learning, cloud computing, big data, etc. Intelligent cloud services are provisioned for well-defined user types. BSC also functions as a public health service platform for hospitals, emergency departments and urgent care centers and other health service institutions as health big data analytics on clouds delivers intelligence for more effectual monitoring of user health and it is more sustainable (Chen *et al.*, 2016a).

Precision medicine seeks personalized solutions tailored to each unique combination of patient features such as genetics (DNA, RNA, etc.), diet, exercise, patient history, other -omics (proteomics, metabolomics, etc), medications, etc. It is also key to establish linkages between systems and precision medicine to translate its principles into clinical practice for practitioners. A high-genomic content in Electronic Health Records (EHRs) could be very useful to uncovering existing knowledge discrepancies in diabetes data analyses (Capobianco, 2017). Some current applications of Big Data analytics in precision medicine include (Wu *et al.*, 2017):

1. EHR data pre-processing: Data embedded in the EHR is ample but not well organized; consequently, EHR data necessitates preprocessing such as imputing missing values
2. EHR data mining: Two strategies comprised of static endpoint prediction and temporal data mining are most often applied to mine actionable data from complex EHR big data
3. Systems biology modeling using –omic data: Systems biology modeling using either static network analysis or dynamic temporal analysis may be performed based on –omic features to attain perceptions about a complex molecular system
4. -Omic data pre-processing which is computationally intensive
5. General analytics for biomedical big data: Most EHR and -omic data are high-dimensional, not only demanding long computation time, but also disturbing the analysis accuracy. Feature extraction is most often completed to transform available features into a more compact set of dimension
6. Biomarker identification using –omic data: Dissimilar groups of samples are gathered for many biological conditions or various time points (e.g., pre vs. post treatment). Most –omic biomarkers are identified through discovering statistically significant differences among the groups

A method called “patient similarity analysis”, which attempts to identify patients experiencing interesting health outcomes and displaying comparable clinical characteristics, treatment pathways and risk factors is a big data methodology which can make more precise Health Economic Outcomes Research (HEOR). Conventional HEOR for medications can be ineffective when using “one-size-fits-all” solutions because it is more focused on the typical patient and often examines data in terms of the mode, median, or mean and fails to recognize the real-world heterogeneity in patient populations (Chen *et al.*, 2016b). However, integrating Big Data analytics and validating drugs in silico has the potential to advance the cost-effectiveness of the drug development pipeline, while computational prediction of drug toxicity and pharmacodynamic/pharmacokinetic properties based on the integration of numerous data varieties potentially reduces costs and helps prioritize compounds for *in vivo* and human testing (Wooden *et al.*, 2017).

Results

Cloud computing enables efficient capture, storage and manipulation of huge volumes of data. Vital sign data are high velocity because it is collected in real time, however, vital signs and equivalent correlations vary for individual patients, also known as variety causing a high degree of ambiguity in correlations within these evolving data streams, which is known as veracity. To assist providers and other caregivers to deliver accurate, real-time and data-drive decisions about patient diagnosis and treatment, there has been an advance of techniques for early discovery of knowledge using intrinsic patterns in vast quantities of vital sign data and the scalable power of cloud computing. This novel method of data management will have an incredible impact on patient care, diagnosis and treatment which will result in a reduction in patient morbidity and mortality as well as population outbreaks and in nosocomial infections within hospitals (Forkan *et al.*, 2017). Big Data offers opportunities in clinical medicine including: (1) patient involvement is increased by creating accessible and understandable data, (2) personalized medicine is translated into healthcare and (3) new knowledge is generated and disseminated (Mathias *et al.*, 2016). Table 3 (Nydegger *et al.*, 2016) shows advantages, disadvantages/drawbacks and measures (for drawbacks) of big data in EHRs.

Sequencing is a well-known high throughput technique; examination of gene chips and proteomics are vital in analyzing the cells, tissues and diseases for personalized medicine and high throughput experiments are essential to these processes. However, these genomics experiments cause challenges to the management of big data. A patient-centered data diagnosis framework, which offers a method to advise the best medication choices for any given patient based on numerous crucial features of the smaller clusters of historical data extracted from the massive amount of historical data, was proposed to offer decision-support to general practitioners. Data clustering is vital to divide the data for improved training of key machine learning algorithms (Babar *et al*., 2016). Medical genetics is primed to incorporate big data and become standard practice, otherwise known as personalized medicine through looking at key factors belonging to each individual patient for evaluation, diagnosis and treatment. Genomic medicine incorporates genomics-based diagnostics into practice. The entire human genomic and exomic sequence (i.e., whole-genome sequencing), is now available and can be used for less expensive and rapid individual analysis (Mathias *et al*., 2016). Table 4 (Wooden *et al*., 2017) shows some publicly available big data resources and their description.

Some traditional parallelized processing frameworks, for example, Hadoop MapReduce, Graphlab and Pregel, are functionally limited and structurally constrained in processing real-time stream big data because their designs are engineered to access and process the static input data. When input data is delivered in a stream flow, no built-in iterative module can be used.

Table 3: Big data in electronic health records (EHRs)

Advantages	Disadvantages	Measures
Patient’s own medical records	Patient-driven medical updates necessary	Explain medical terminology to patients
Patient records download by hospital W-LAN	Hacker friendliness	Limit time of accessibility
Ubiquitous access to EHR	Code readability not yet universal	National health offices regulations in progress
Real-time health profiles	Hacker friendliness	Encryption
Post-marketing surveillance of medical devices	Criteria selection	LOINC^{TM} coding
Distribution pattern of virulence of the same bacterial strain	Exchange of DNA: Most strains have overlapping genome	(1) update bioinformatic resource; (2) use hybridization of identification

Table 4: Some publicly available big data resources

Category	Name	Description
Literature mining	PolySearch 2.0	Web-based text mining tool
Machine learning	Weka	Extensive library of machine Learning algorithms
Omics data repositories	Gene Expression Omnibus (GEO)	Repository of raw and processed omics data
	ArrayExpress	Repository of raw and processed omics data
	Sequence Read Archive (SRA)	Repository of sequencing data
	The Cancer Genome Atlas	Repository of genomic, proteomic, histological and clinical data for a wide variety of cancers
Molecular pathway knowledgebase/analysis tool	Molecular Signatures Database (MSigDb)	Repository of molecular signatures from curated databases, publications and research studies
	NDEx	Biological network knowledge Base
	DAVID	Searchable/downloadable database of molecular pathway knowledge base
Functional perturbation data repository	ChemBank	Database/knowledge base of high-throug compound screens and other small molecule–related information
	Connectivity Map (CMap)	Database of drug perturbation gene expression signatures
	Library of Integrated Cellular Signatures (LINCS)	Database of functional cellular responses to genetic and pharmacological perturbations measured in multiple types of biomolecules
Cheminformatics	PubChem	Comprehensive database of structural, pharmacological and biochemical activity data
	DrugBank	Database of drug chemical, structural, pharmacological and target information
	Protein Data Bank	Repository of protein structural data
	SIDER	Database of drug adverse Effects
	admetSAR	Web tool predicting pharmacological and toxicology parameters based on chemical structures
	The Drug Gene Interaction Database (DGIdb)	Database of known drug-gene connections for selected genes

A task-level adaptive MapReduce framework for processing stream data in healthcare scientific applications, which expands the traditional Hadoop MapReduce framework and deals with the varied arrival rate of big data splits has been proposed (Zhang *et al.*, 2015). A generic architecture for big data healthcare analytics was also proposed using open sources, which includes Hadoop, NoSQL Cassandra, Kafka and Apache Storm. This combination of high throughput publishes subscribed messaging for streams, distributed real-time computing and distributed storage systems and can effectually analyze enormous rapidly incoming healthcare data (Ta *et al.*, 2016). The most commonly used programming model in Big Data analytics that provides the ability to process large volumes of data in batch form cost-effectively is MapReduce and it fosters the analysis of both unstructured and structured data in Massively Parallel Processing (MPP) environments. Users can track data in motion, respond to unexpected events as they happen and quickly decide the next-best action with a real-time analysis (Wang *et al.*, 2018).

IoT provides a foundation for smart health using heterogeneous sensors such as temperature sensors, heartbeat sensors, glucometers, etc. Sensors are used to measure vital sign data such as heartbeat, temperature, oxygen saturation, glucose, etc.; once the vital signs are gathered using the sensors, they are then transmitted via the IoT to the provider for analysis. Energy harvesting is based on the piezoelectric effect produced by human body vibrations or pressure while big data processing is performed using Hadoop and MapReduce. Because health monitoring sensors must be functional 24 hours-a-day seven days-a-week with a zero need for maintenance, the importance and necessity for energy harvesting in IoT cannot be understated. The piezoelectric devices can be attached to different human body parts identified as pressure areas which cause the piezoelectric devices to produce the piezoelectric effect, causing the generation of electric energy supplied to the wearable health monitoring sensors planted on the human body (Babar *et al.*, 2017).

A Semantic Interoperability Model for Big-data in IoT (SIMB-IoT) has been proposed with the key goal to provide semantic interoperability in big data among heterogenous IoT devices through the data model of Resource Description Framework (RDF). RDF is a semantic web framework recycled to communicate things using Triples, so it will be semantically significant. Providers prescribed selected patients' medications concurrently with the diagnosis of diseases using IoT devices, the annotated information then suggested medications matched from the pharmaceutical industry and transmitted the prescribing information on each medicine to the patient's personal IoT device on the intelligent health cloud (Ullah *et al.*, 2017).

The subsystems and operating procedure of a mobile physiological sensor system (MoPSS) platform have been developed so that the Wireless Body Sensor Network System (WBSNS) has three subsystems: Wearable physiological sensors, a smart phone and LAN. The Data Collection and Classification System (DCCS) has data collection and data classification servers. The healthcare monitor system contains two subsystems: The healthcare cloud and healthcare monitor website/system (You *et al.*, 2018).

Four new system architecture components required on top of traditional block chain were proposed and technology challenges in the block chain platform were discussed: (1) Data management component of block chain application for data integrity, big data integration and integrating disparity of medical related data; (2) a new block chain-based general distributed and parallel computing paradigm component to study and devise parallel computing methodology for big data analytics; (3) trust data sharing management component to enable a trust medical data ecosystem for collaborative research; and (4) verifiable anonymous identity management component for identity privacy for both the individual and IoT devices and secure data access to make patient-centric medicine possible (Shae and Tsai, 2017).

Discussion

An important, but often overlooked, privacy issue involves the merger of genomic data with other disparate datasets. Besides health records and biological and environmental monitoring data, many other types of information are arguably relevant to an individual's health, including the following: (1) employment records, including exposure and biological monitoring data; (2) travel information and geo-location data, also relevant to exposures; (3) health histories and vital statistics of family members; (4) educational records, including behavioral health information and student health service records; (5) social media postings, including behavioral and mental health self-reports; (6) financial information; (7) various government records; and (8) military service records, which could include health records and data on hazardous exposures (Rothstein, 2017).

A challenge of preclinical and clinical research is to attain and acquire access to adequate informative high-quality data; however, many health data cannot be directly used for secondary purposes (Auffray *et al.*, 2016). When a patient is assessed, diagnosed, treated, etc., obtaining clinical notes and perceiving them in the right context, organizing medical imaging data, grasping data concerning biomarkers and understanding the substantial amounts genomic data available for each patient that are useful in clinical settings are challenges of big data in healthcare. Nevertheless, data from various sensors and social media sites can be accessed to provide data about the patient's behavioral, psychological and social patterns (Kalantari *et al.*, 2018). In the emergency department, patient information, triage and prioritization become a complex decision-making process during peak times or when many patients need to be assessed at once. For example, patients with chronic heart disease are harder to prioritize and establishing a triage priority is challenging because it is necessary to asses each patient with chronic heart disease according to multiple symptoms such as vital signs, oxygen saturation, activity level, edema, etc (Salman *et al.*, 2017).

High-dimensional spaces may arise from an extensive set of biomarkers, health attributes and sensor fusion, creating a bottleneck in analyzing big data which is to obtain fast inference in real-time from large and high-dimensional observations. From a software point of view, processing big data usually requires parallel programming paradigms such as Map Reduce. If challenges to analyze big data can be addressed in a coherent manner, it will create more value to both patients and healthcare systems (Perez *et al.*, 2015). Although data analytics can improve the quality of patient care and reduce healthcare costs, one significant challenge facing us is the exponential growth of data volume. The RNA Inference (RNAI) experiment, which is marking cells with variable fluorescent dyes to capture three desirable components, namely, DNA, Actin and PH3 channels is one popular type of such data that are generated. Efficient access to such data is essential for expert systems to extract useful medical information (Karimi *et al.*, 2016).

Big Healthcare Data (BHD) often has complementary dimensions: Disparate sources, complexity, large size, multiple scales, incompleteness and incongruence's. No universal protocol exists which compares, models, or benchmarks the performance of various data analytics. Big data often include heterogeneous data elements where small sub-samples might capture specific cohorts, including extreme data or outliers (Dinov, 2016). The following issues require much closer scrutiny in the immediate future (1) meaningful access rights to individual data subjects lacking necessary resources is difficult to provide (2) the analysis of aggregated datasets will generate intellectual property and ownership could be called into question; (3) the changing nature of fiduciary relationships which have become increasingly data saturated; (4) the dangers of ignoring group-level ethical harms; (5) it is necessary to distinguish between academic and commercial Big Data practices according to potential harm to data subjects; and (6) the importance of epistemology in assessing the ethics of Big Data (Mittelstadt and Floridi, 2016).

In the quest for precision medicine the recent scale of data collection and usage in clinical care and biomedical research via the use of next generation sequencing technologies is seemingly limitless within the multidisciplinary contexts of data sharing that reinforce and enable both discovery and infrastructure science. However, as genomic data enters the clinic, will health data flow to and from the medical record to the research context and back in a learning healthcare system? We have to consider not only the ethical imperative to pursue "the good" or the volume of big data, but its multivariate nature that inspires us to re-examine the "classical" socio-ethical issues surrounding the risks and benefits of data collection, access and sharing and their impact on privacy and discrimination (Knoppers and Thorogood, 2017).

Conclusion

The heterogeneity of healthcare data causes unavoidable differences in data formats, data type variability and population characteristics. Home healthcare services need the crucial support of sensors and remotes monitors as a critical approach for the growth of public health and provider services in rural and disparate populations, for those who live geographically distant from Emergency Medical Systems (EMS) and hospitals. Wearable sensors have a wide variety of applications and functions, from smart clothing to sensors in wrist watches and smart phones. These vital tools are key to advancing personalized medicine because data can be transmitted immediately from the patient to the provider for diagnosis and treatment. Currently there is much work being done to learn the value of various Electrocardiograph (ECG) tracings in exercise, heart disease and other diseases. New applications for sensor data are constantly being developed. Although security remains a highly important factor in the use of sensor and remote monitors, as well as social media, new laws and software is aimed at controlling the leaks of identifiable data which can be traced back to the individual patient. Another new technology is smart clothing, which is becoming critical to realizing sustainable health monitoring for health big data collection over an extended period. Smart clothing is convenient and can provide a range of sensor activities.

Stream computing supports high performance stream data processing in real-time or near real-time. Streaming input datasets are from different sources and have various arrival rates and unfortunately almost none of the existing data processing frameworks can handle the situations that arise from this technology. There are many challenges of Big Data concerning patient and EHR data, especially in triage and assigning appropriate priority. The key lies in the simultaneous consideration of crucial data such as vital signs, clinical features, etc. Other challenges include the veracity and variety of the structured and unstructured nature of health big data and the fact laws must be adhered to, data must be deidentified and EHR systems must be protected from both internal and external threats. Security is a significant obstacle in interoperability among heterogeneous IoT devices and must be a serious aspect for the solution of interoperability. Unstructured data processing, real-time processing of stream data, cybersecurity and privacy protection, etc. can be future studies of Big Data analytics in healthcare.

Acknowledgement

This study was supported in part by Institute for Information Technology Innovations and Smart Healthcare in Mississippi, USA. No conflict of interest to disclose.

Author's Contributions

Both the authors contributed equally to prepare, develop and complete this study and manuscript.

Ethics

This article is original. Authors declare that are not ethical issues that may arise after the publication of this manuscript.

References

Adil, A., H.A. Kar, R. Jangir and S.A. Sofi, 2015. Analysis of multi-diseases using big data for improvement in health care. Proceedings of the IEEE UP Section Conference on Electrical Computer and Electronics, Dec. 4-6, IEEE Xplore Press, Allahabad, India, pp: 1-6. DOI: 10.1109/UPCON.2015.7456696

Auffray, C., R. Balling, I. Barroso, L. Bencze and M. Benson *et al*., 2016. Making sense of big data in health research: Towards an EU action plan. Genme Med., 8: 71-71. DOI: 10.1186/s13073-016-0323-y

Babar, M., A. Rahman, F. Arifand and G. Jeon, 2017. Energy-harvesting based on internet of things and big data analytics for smart health monitoring. Sustainable Computing: Inform. Syst. DOI: 10.1016/j.suscom.2017.10.009

Babar, M.I., M. Jehanzeb, M. Ghazali, D.N. Jawawi and F. Sher *et al*., 2016. Big data survey in healthcare and a proposal for intelligent data diagnosis framework. Proceedings of the 2nd IEEE International Conference on Computer and Communications, Oct. 14-17, IEEE Xplore Press, Chengdu, China. DOI: 10.1109/CompComm.2016.7924654

Bansal, S., G. Chowell, L. Simonsen, A. Vespignani and C. Viboud, 2016. Big data for infectious disease surveillance and modeling. J. Infect. Dis., 214: S375-S379. DOI: 10.1093/infdis/jiw400

Cano, I., A. Tenyi, E. Vela, F. Miralles and J. Roca, 2017. Perspectives on big data applications of health information. Curr. Opin. Syst. Biol., 3: 36-42. DOI: 10.1016/j.coisb.2017.04.012

Capobianco, E., 2017. Systems and precision medicine approaches to diabetes heterogeneity: A big data perspective. Clin. Translat. Med., 6: 23-23. DOI: 10.1186/s40169-017-0155-4

Carnevale, L., A. Celesti, M. Fazio, P. Bramantiand and M. Villari, 2017. How to enable clinical work flows to integrate big health care data. Proceedings of the IEEE Symposium Conference on Computers and Communications, Jul. 3-6, IEEE Xplore Press, Heraklion, Greece. DOI: 10.1109/ISCC.2017.8024634

Chen, M., Y. Ma, J. Song, C.F. Lai and B. Hu, 2016a. Smart clothing: Connecting human with clouds and big data for sustainable health monitoring. Mobile Netw. Applic., 21: 825-845. DOI: 10.1007/s11036-016-0745-1

Chen, Y., G.F. Guzauskas, C. Gu, B. Wang and W.E. Furnback *et al*., 2016b. Precision health economics and outcomes research to support precision medicine: Big data meets patient heterogeneity on the road to value. J. Personalized Med., 6: 20-20. DOI: 10.3390/jpm6040020

Dimeglio, C., M.K. Irving, T. Lang and C. Delpierre, 2016. Expectations and boundaries for big data approaches in social medicine. J. Forensic Legal Med., 57: 51-54. DOI: 10.1016/j.jflm.2016.11.003

Ding, W., C.T. Lin, S. Chen, X. Zhang and B. Hu, 2018. Multiagent-consensus-MapReduce-based attribute reduction using co-evolutionary quantum PSO for big data applications. Neurocomputing, 272: 136-153. DOI: 10.1016/j.neucom.2017.06.059

Dinov, I.D., 2016. Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. GigaScience, 5: 12-12. DOI: 10.1186/s13742-016-0117-6

Forkan, A.R.M., I. Khalil and M. Atiquzzaman, 2017. ViSiBiD: A learning model for early discovery and real-time prediction of severe clinical events using vital signs as big data. Comput. Netw., 113: 244-257. DOI: 10.1016/j.comnet.2016.12.019

Kalantari, A., A. Kamsin, S. Shamshirband, A. Gani and H.A. Rokny *et al*., 2018. Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions. Neurocomputing, 276: 2-22. DOI: 10.1016/j.neucom.2017.01.126

Karimi, N., S. Samavi, S.M.R. Soroushmehr, S. Shirani and K. Najarian, 2016. Toward practical guideline for design of image compression algorithms for biomedical applications. Expert Syst. Applic., 56: 360-367. DOI: 10.1016/j.eswa.2016.02.047

Knoppers, B.M. and A.M. Thorogood, 2017. Ethics and big data in health. Curr. Opin. Syst. Biol., 4: 53-57. DOI: 10.1016/j.coisb.2017.07.001

Lee, E.C., J.M. Asher, S. Goldlust, J.D. Kraemer and A.B. Lawson *et al*., 2016. Mind the scales: Harnessing spatial big data for infectious disease surveillance and inference. J. Infect. Dis., 214: S409-S413. DOI: 10.1093/infdis/jiw344

Ma, Y., Y. Wang, J. Yang, Y. Miao and W. Li, 2017. Big health application system based on health internet of things and big data. IEEE Access, 5: 7885-7897. DOI: 10.1109/ACCESS.2016.2638449

Manogaran, G., R. Varatharajan, D. Lopez, P.M. Kumar and R. Sundarasekar *et al*., 2018. A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system. Future Generat. Comput. Syst., 82: 375-387. DOI: 10.1016/j.future.2017.10.045

Mathias, B., G. Lipori, L.L. Moldawer and P.A. Efron, 2016. Integrating "big data" into surgical practice. Surgery, 159: 371-374. DOI: 10.1016/j.surg.2015.08.043

Miller, D.D. and E.W. Brown, 2017. Artificial intelligence in medical practice: The question to the answer? Am. J. Med., 131: 129-133. DOI: 10.1016/j.amjmed.2017.10.035

Mittelstadt, B.D. and L. Floridi, 2016. The ethics of big data: Current and foreseeable issues in biomedical contexts. Sci. Eng. Eth., 22: 303-341. DOI: 10.1007/s11948-015-9652-2

Nishimura, A., K. Nishimura, A. Kada and K. Iihara, 2016. Status and future perspectives of utilizing big data in neurosurgical and stroke research. Neurol. Medico-Chirurgica, 56: 655-663. DOI: 10.2176/nmc.ra.2016-0174

Nydegger, U., T. Lung, L. Risch, M. Risch and P.M. Escobar *et al*., 2016. Inflammation thread runs across medical laboratory specialities. Mediators Inflammat., 2016: 10-10. DOI: 10.1155/2016/4121837

Perez, J.A., C.C.Y. Poon, R.D. Merrifield, S.T.C. Wong and G.Z. Yang, 2015. Big data for health. IEEE J Biomed Health Infor., 19: 1193-1208. DOI: 10.1109/JBHI.2015.2450362

Rothstein, M.A., 2017. Structural challenges of precision medicine: currents in contemporary bioethics. J. Law Med. Eth., 45: 274-279. DOI: 10.1177/1073110517720655

Sacristán, J.A. and T. Dilla, 2015. No big data without small data: learning health care systems begin and end with the individual patient. J. Evaluat. Clin. Pract., 21: 1014-1017.DOI: 10.1111/jep.12350

Salman, O.H., A.A. Zaidan, B.B. Zaidan, Naserkalid and M. Hashim, 2017. Novel methodology for triage and prioritizing using big data patients with chronic heart diseases through telemedicine environmental. Int. J. Inform. Technol. Decision Mak., 16: 1211-1245. DOI: 10.1142/S0219622017500225

Saravanakumar, M.V. and S.M. Hanifa, 2017. BIGDATA: Harnessing insights to healthier analytics-a survey. Proceedings of the International Conference on Algorithms, Methodology, Models and Application Emerging Technology, Feb.16-18, IEEE Xplore Press, Chennai, India. DOI: 10.1109/ICAMMAET.2017.8186648

Schmidt, B. and A. Hildebrandt, 2017. Next-generation sequencing: big data meets high performance computing. Drug Discovery Today, 22: 712-717. DOI: 10.1016/j.drudis.2017.01.014

Shae, Z. and J.J.P. Tsai, 2017. On the design of a blockchain platform for clinical trial and precision medicine. Proceedings of the IEEE 37th International Conference on Distributed Computing Systems, Jun. 5-8, IEEE Xplore Press, Atlanta, Georgia, USA, pp: 1972-1980. DOI: 10.1109/ICDCS.2017.61

Ta, V.D., C.M. Liu and G.W. Nkabinde, 2016. Big data stream computing in healthcare real-time analytics. Proceedings of the IEEE International Conference on Cloud Computing and Big Data Analysis, Jul. 5-7, IEEE Xplore Press, Chengdu, China. DOI: 10.1109/ICCCBDA.2016.7529531

Ullah, F., M.A. Habib, M. Farhan, S. Khali and M.Y. Durrani *et al*., 2017. Semantic interoperability for big-data in heterogeneous IoT infrastructure for healthcare. Sustainable Cities Society, 34: 90-96. DOI: 10.1016/j.scs.2017.06.010

Wang, Y., L. Kung and T.A. Byrd, 2018. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technol. Forecast. Soc. Change, 126: 3-13.

Wang, Y., L. Kung, W.Y.C. Wang and C.G. Cegielski, 2017. An integrated big data analytics-enabled transformation model: Application to health care. Inform. Manag., 55: 64-79. DOI: 10.1016/j.im.2017.04.001

Wooden, B., N. Goossens, Y. Hoshida and S.L. Friedman, 2017. Using big data to discover diagnostics and therapeutics for gastrointestinal and liver diseases. Gastroenterology, 152: 53-67. DOI: 10.1053/j.gastro.2016.09.065

Wu, P.Y., C.W. Cheng, C.D. Kaddi, J. Venugopalan and R. Hoffman *et al*., 2017. Omic and electronic health record big data analytics for precision medicine. Trans. Biomed. Eng., 64: 263-273. DOI: 10.1109/TBME.2016.2573285

You, I., K.K.R. Choo and C.L. Ho, 2018. A smartphone-based wearable sensors for monitoring real-time physiological data. Comput. Electr. Eng., 65: 376-392. DOI: 10.1016/j.compeleceng.2017.06.031

Zhang, F., J. Cao, S.U. Khan, K. Li and K. Hwang, 2015. A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications. Future Generat. Comput. Syst., 43: 149-160. DOI: 10.1016/j.future.2014.06.009