Computational Health Informatics in the Big Data Age: A Survey

RUOGU FANG*, SAMIRA POUYANFAR*, YIMIN YANG, SHU-CHING CHEN,
and S. S. IYENGAR, School of Computing & Info. Sciences, Florida International University, Miami

The explosive growth and widespread accessibility of digital health data have led to a surge of research activity in the healthcare and data sciences fields. The conventional approaches for health data management have achieved limited success as they are incapable of handling the huge amount of complex data with high volume, high velocity, and high variety. This article presents a comprehensive overview of the existing challenges, techniques, and future directions for computational health informatics in the big data age, with a structured analysis of the historical and state-of-the-art methods. We have summarized the challenges into four Vs (i.e., volume, velocity, variety, and veracity) and proposed a systematic data-processing pipeline for generic big data in health informatics, covering data capturing, storing, sharing, analyzing, searching, and decision support. Specifically, numerous techniques and algorithms in machine learning are categorized and compared. On the basis of this material, we identify and discuss the essential prospects lying ahead for computational health informatics in this big data age.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*; J.3 [**Computer Application**]: Life and Medical Sciences—*Health and medical information systems*

General Terms: Algorithm, Design, Performance

Additional Key Words and Phrases: Big data analytics, computational health informatics, 4V challenges, machine learning, data mining, clinical decision support, survey

ACM Reference Format:
Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and S. S. Iyengar. 2016. Computational health informatics in the big data age: A survey. ACM Comput. Surv. 49, 1, Article 12 (June 2016), 36 pages.
DOI: http://dx.doi.org/10.1145/2932707

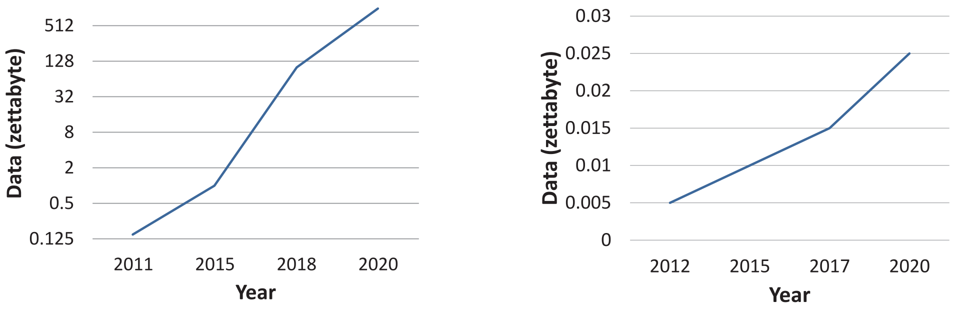
1. INTRODUCTION

Computational health informatics is an emerging research topic within and beyond the medical industry. It is a multidisciplinary field involving various sciences such as biomedical, medical, nursing, information technology, computer science, and statistics. Using Information and Communication Technologies (ICTs), health informatics collects and analyzes the information from all healthcare domains to predict patients' health status. The major goal of health informatics research is to improve Health Care Output (HCO) or patients' quality of care [Herland et al. 2014]. The healthcare industry has experienced rapid growth of medical and healthcare data in recent years. Figure 1 depicts the growth of both healthcare data and digital healthcare data. As shown in Figure 1(a), the U.S. healthcare data alone reached 150 exabytes (10^{18}) in 2011 and it will exceed the zettabyte (10^{21}) and the yottabyte (10^{24}) in the near future [Raghupathi

*Equal contribution.

Authors' addresses: R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. S. Iyengar, School of Computing and Information Sciences, Florida International University, Miami, FL 33199; emails: {rfang, spouy001, yyang010, chens, iyengar}@cs.fiu.edu.

DOI: http://dx.doi.org/10.1145/2932707



(a) U.S. healthcare data growth

(b) Digital healthcare data growth

Fig. 1. Healthcare data growth.

and Raghupathi 2014]. It is projected that the healthcare data analytics market will increase and grow 8 to 10 times as fast as the overall economy until 2017 [Perez 2013].

The rapid growth of novel technologies has led to a significant increase of digital health data in recent years. More medical discoveries and new technologies such as mobile apps, capturing devices, novel sensors, and wearable technology have contributed to additional data sources. Therefore, the healthcare industry produces a huge amount of digital data by utilizing information from all sources of healthcare data such as Electronic Health Records (EHRs, including electronic medical records) and Personal Health Records (PHRs, one subset of EHRs including medical history, laboratory results, and medications). Based on reports, the estimation of digital healthcare data from all over the world was almost 500 petabytes (10^{15}) in 2012, and it is expected to increase and reach 25 exabytes in 2020, as shown in Figure 1(b) [Sun and Reddy 2013].

The digital health data is not only enormous in amount, but also complex in its structure for traditional software and hardware. Some of the contributing factors to the failure of traditional systems in handling these datasets include:

—The vast variety of structured and unstructured data such as handwritten doctor notes, medical records, medical diagnostic images (magnetic resonance imaging (MRI), computed tomography (CT)), and radiographic films [Feldman et al. 2012]
—Existence of noisy, heterogeneous, complex, longitudinal, diverse, and large datasets in healthcare informatics [Sun and Reddy 2013]
—Difficulties to capture, store, analyze, and visualize such large and complex datasets
—Necessity of increasing the storage capacity, computation power, and processing power [Roy 2015]
—Necessity of improving medical issues such as quality of care, sharing, security of patients' data, and the reduction of the healthcare cost, which are not sufficiently addressed in traditional systems [Cottle et al. 2013]

Hence, solutions are needed in order to manage and analyze such complex, diverse, and huge datasets in a reasonable time complexity and storage capacity. Big data analytics, a popular term given to datasets that are large and complex, play a vital role in managing the huge healthcare data and improving the quality of healthcare offered to patients. In addition, it promises a bright prospect for decreasing the cost of care, improving treatments, reaching more personalized medicine, and helping doctors and physicians to make personalized decisions. Based on the definition by Gartner [2014], big data is "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making." Healthcare data definitely falls into the scope of big data. Based on the estimation of a survey by the McKinsey Global Institute, the value created by healthcare analytics could be more than $300 billion every year [Manyika et al. 2011].

Big data in health informatics is a collection of technologies, tools, and techniques that manage, manipulate, and organize large, diverse, and complex datasets in order to ameliorate the quality of patients' status. Finally, the major benefits of big data in health informatics are as follows: First, it makes use of the huge volume of data and provides timely and effective treatment to patients. Second, it provides personalized care to patients. Third, it will utilize all the medical system components including patient, payer, provider, and management [Sun and Reddy 2013].

1.1. Three Scenarios of Big Data in Health Informatics

Today, the healthcare industry is turning to big data technology to improve and manage medial systems. For this purpose, healthcare companies and organizations are leveraging big data in health informatics. What follows is a description of a selection of big data scenarios demonstrating its application and importance in the healthcare informatics and in the treatment of current medical issues.

(1) High-risk and high-cost patient management

According to the National Health Expenditure Projections 2013, the rate of health spending is anticipated to grow at 5.7%, which is faster than the expected annual growth. The cost of healthcare, or rather, health spending, in the United States is much higher ($2.5 trillion in 2009) than other developed countries [Bates et al. 2014]. Approximately half of all this spending is related to 5% of patients. Therefore, high-risk and high-cost patients need to be identified and managed carefully and effectively. Bates et al. [2014] address the issue of high-cost patients using big data solutions. Several factors have been considered to reduce cost and improve the prediction of high-risk patients. They also suggest that analytic systems using large data from high-risk patients are needed in order to develop predictive models.

(2) Risk-of-readmission prediction

One of the extremely challenging subjects in healthcare informatics is developing solutions for risk prediction. Predicting risk of readmission or rehospitalization has attracted attention as a solution to reduce costs and make quality of care better. However, predicting readmissions is complex because it involves a variety of factors such as health conditions, disease parameters, and hospital care quality parameters [Zolfaghar et al. 2013]. Recently, IBM big data analytics have been used in the University of North Carolina (UNC) healthcare system to help reduce costly and preventable readmissions [IBM 2015a]. For this purpose, Natural Language Processing (NLP) is utilized to extract key elements from unstructured data, such as notes and reports. In addition, a number of studies have been done using big data solutions for predicting risk of readmission. For instance, Zolfaghar et al. [2013] developed big data-driven solutions for both information extraction and predictive modeling to predict 30-day risk of readmission for congestive heart failure incidents.

(3) Mobile health for lifestyle recommendation

Today, ubiquitous and personalized mobile technology offers a great platform for achieving faster, cheaper, and more accessible healthcare [Cotman et al. 2007]. For example, it frequently provides a patient's information, behaviors, and health status. Moreover, doctors can monitor a patient's status including heart rate, blood pressure level, and sleep patterns using new mobile technology [Byrnes 2014]. However, effective management of this huge inflow of mobile-generated data calls for the implementation of a big data solution. Therefore, healthcare organizations leverage big data solutions to manage all of the health information, improve care, and increase access to healthcare.

1.2. Related Work

To date, various research studies have been done using big data tools and approaches in health informatics focusing on biomedical [Raghupathi and Raghupathi 2014; Feldman et al. 2012; Costa 2014; White 2014] and computational aspects [Herland et al. 2014; Merelli et al. 2014; Yoo et al. 2012; Herland et al. 2013], respectively.

In Raghupathi and Raghupathi [2014], an overview of big data in health informatics is provided. The paper emphasizes the various characteristics and advantages of big data in the following areas: clinical operations, public health, genomic analytics, remote monitoring, and so on. In addition, some tools and platforms for big data are listed. Finally, it concludes that big data in health informatics improves outcomes while reducing costs. Costa [2014] discusses the major improvements achieved in combining omics (which refers to different fields of study in biology such as genomics, proteomics, and metabolomics) and clinical health data for personalized medicine applications. In that article, the challenges associated with using big data in biomedicine and translational science are reviewed. Feldman et al. [2012] introduce readers to big data and its necessity in the healthcare landscape. By interviewing a number of companies in the emerging healthcare big data ecosystem, they researched the role of big data in health informatics. The main effort of this research is providing a general overview of big data in health informatics and reflecting the ways big data could help healthcare by grouping different companies and organizations using big data analytics.

Most popular surveys of big data in health informatics have concentrated on biomedical aspects of big data, while a smaller percentage of papers focus on the computational perspective. Herland et al. [2014] discuss various studies being done in Health Informatics branches using big data based on subfields such as bioinformatics, public health informatics, and neuroinformatics. The main focus of this survey is on recent research using big data methods to gather health informatics data at a few levels such as the molecular, tissue, and population levels. Various data mining methods are discussed in all the aforementioned levels. However, the survey is not categorized based on computational and big data approaches and only emphasizes a few healthcare examples. Merelli et al. [2014] discuss some technological aspects related to big data in health informatics including architectural solutions for big data, parallel platforms for big data, approaches for data annotation, data access measures, and security for biomedical data. Many technological issues still remain, such as data capturing, feature analyzing, and machine-learning algorithms, to name a few.

Several machine-learning and data mining algorithms have been used in health informatics. A brief overview of different data mining methods applied for association, classification, and clustering is presented in Yoo et al. [2012]. However, this article does not deal with big data aspects of data mining approaches in health informatics.

1.3. Contributions and Organization of This Article

This survey provides a structured and extensive overview of computational methods of big data in medical and health informatics. Most existing surveys on big data in health informatics only focus on high-level biomedical policies without comprehensive explanation of processing steps. In contrast, this survey offers a focused overview on the computational approaches for big data by expanding the study in several directions of health informatics processing. It explains the entire process from scratch and presents a comprehensive pipeline that discusses every processing step from capturing raw data to clinical diagnosis.

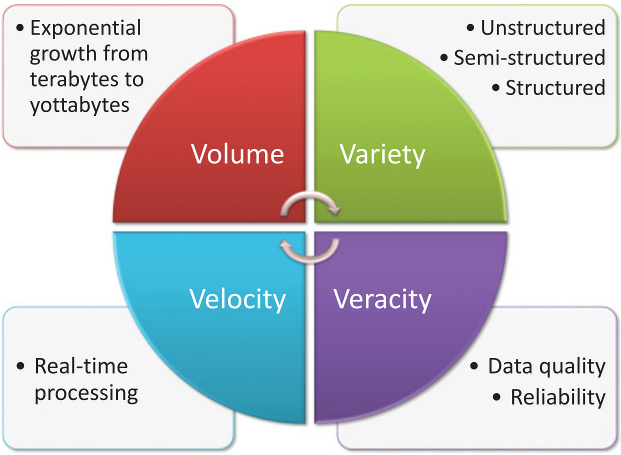


Fig. 2. Four Vs of big data.

The ultimate goal of this survey is to connect big data and health informatics communities. The primary emphasis is on computational aspects of big data in health informatics, which includes challenges, current big data mining techniques, strengths and limitations of current works, and an outline of directions for future work.

This survey is organized into three parts: challenges, processing pipeline of healthcare informatics, and future directions. In Section 2, different examples of challenges and opportunities related to big data in health informatics are provided. In addition, the four dimensions (Vs) of big data including volume, velocity, variety, and veracity are presented. Section 3 discusses the pipeline of health informatics computational processing. In Section 3.1, several capturing methods for gathering healthcare data are described. Ways of storing data and sharing it with other systems in the whole world are presented in Sections 3.2 and 3.3, respectively. Section 3.4 covers analyzing data, which includes preprocessing, feature extraction/selection, and machine-learning approaches. The remainder of Section 3, 3.5 and 3.6, present searching approaches and decision support methods, respectively. Finally, Section 4 explores the potential directions for future works based on big data's challenges.

2. CHALLENGES AND OPPORTUNITIES

The challenges of big data in health informatics could be generally described as the "4Vs": volume, variety, velocity, and veracity. It is overwhelming and challenging to deal with large volumes of healthcare data because of not only the diverse data sources but also the speed for generating and processing those data, not to mention the examination and verification of them, such as the quality and legitimacy [Feldman et al. 2012]. The four Vs characterize the essential properties of big data in health informatics (as shown in Figure 2).

2.1. Volume

With the fast development of information technology and medical devices, the world has been witnessing an exponential growth of healthcare data, whose quantity can no longer be measured in terabytes but ought to be described in zettabytes or even yottabytes. According to Health Catalyst [Crapo 2014], healthcare firms with over 1,000 employees store over 400 terabytes of data per firm (reported in the year of 2009), which qualifies healthcare as a high-data volume industry, despite the real-time streams of web and social media data.

Contributing to the huge volume of healthcare data are various sources of data, from traditional personal medical records and clinical trial data to new types of data such as various sensor readings and 3D imaging [Feldman et al. 2012] (Section 3.1 provides a summarization of the data sources). Recently the proliferation of wearable medical devices has significantly added fuel to the healthcare data. Those devices are able to continuously monitor a series of physiological information, such as biopotential, heart rate, blood pressure, and so forth [Hung et al. 2004].

The high volume of healthcare data creates a big challenge, the desire for scalable storage and support for distributed queries across multiple data sources. Specifically, the challenge is being able to locate and mine specific pieces of data in an enormous, partially structured dataset. Many advanced data management techniques such as virtualization and cloud computing have been widely studied and experimented in industrial companies. Those proposed platforms are capable of manipulating large volumes of data virtually or geographically distributed on multiple physical machines, enabling the universal sharing of information.

2.2. Variety

Healthcare data could be characterized by the variety of sources and the complexity of different forms of data. Generally, healthcare data could be classified into unstructured, structured, and semistructured.

Historically, most unstructured data usually come from office medical records, handwritten notes, paper prescriptions, MRI, CT, and so on. The structured and semistructured data refers to electronic accounting and billings, actuarial data, laboratory instrument readings, and EMR data converted from paper records [Feldman et al. 2012].

Nowadays, more and more data streams add variety to healthcare information, both structured and unstructured, including intelligent wearable devices, fitness devices, social media, and so on. The challenge lies in the seamless combination of old-fashioned and new forms of data, as well as the automatic transformation between the structured and unstructured data, which relies on effective distributed processing platforms and advanced data mining and machine-learning techniques. Leveraging heterogeneous datasets and securely linking them have the potential to improve healthcare by identifying the right treatment for the right individual.

2.3. Velocity

The increase in the volume and variety of healthcare data is highly related to the velocity at which it is produced and the speed needed to retrieve and analyze the data for timely decision making.

Compared with relatively static data such as paper files, x-ray films, and scripts, it is gradually becoming more important and challenging to process a real-time stream, such as various monitoring data, accurately and in a timely manner, in order to provide the right treatment to the right patient at the right time [Feldman et al. 2012]. A concrete example can be found in the prevalence of wearable monitoring devices, which provide continuous and ever-accumulated physiological data. Being able to perform real-time analytics on continuous monitoring data could help predict life-threatening pathological changes and offer appropriate treatment as early as possible.

The high velocity of healthcare data poses another big challenge for big data analytics. Although the traditional Database Management Systems (DBMS) is reported to perform well on large-scale data analysis for specific tasks [Pavlo et al. 2009], it simply cannot catch up with the pace of high-velocity data, not to mention the limitation of flexibility when facing multilevel nesting and hierarchies in data structures with high volatility, which is a common property for healthcare data. This situation creates an opportunity for introducing high-velocity data processing tools. An initial attempt includes the utilization of the Hadoop platform for running analytics across massive volumes of data using a batch process. More recently, industrial practices have approved the convergence of traditional relational databases and key-value stores. Spanner [Corbett et al. 2013], Google's globally distributed database, is an example that provides high consistency and availability.

2.4. Veracity

Coming from a variety of sources, the large volume of healthcare data varies in its quality and complexity. It is not uncommon that the healthcare data contains biases, noise, and abnormalities, which poses a potential threat to proper decision-making processes and treatments to patients.

High-quality data can not only ensure the correctness of information but also reduce the cost of data processing. It is highly desirable to clean data in advance of analyzing it and using it to make life-or-death decisions. However, the variety and velocity of healthcare data raise difficulties in generating trusted information. There are mainly two types of data quality problems depending on the causes. The first type is primarily due to IT issues (e.g., data management, audit, error reporting, and compliance). The second type reveals the underlying veracity of the data (i.e., the truthfulness, relevance, and predictive value) [Feldman et al. 2012]. It is the second type that is of greater importance and is a bigger challenge, which potentially could be handled by using big data analytic tools.

The biggest challenge is determining the proper balance between protecting the patient's information and maintaining the integrity and usability of the data. Robust information and data governance programs will address a number of these challenges. As Techcrunch [2014] points out, "while today we rely on the well-trained eye of the general practitioner and the steady hand of the surgeon, tomorrow's lifesavers will be the number-crunching data scientists, individuals with only a passing understanding of first aid."

2.5. Opportunities

Except for the "4Vs" that are most familiar to the readers, there are other emerging issues to be considered, such as the validity and the volatility of big data in health informatics. While validity is concerned with the correctness and accuracy of data, volatility refers to how long the data would be valid and should be kept for. All of these big data characteristics (including all Vs) offer great challenges to big data analytics in healthcare. At the same time, these very same challenges bring unprecedented opportunities for introducing cutting-edge technologies to make sense out of a large volume of data and provide new insights for improving the decision-making process in near real time. The enormous scale and diversity of temporal-spatial healthcare data have created unprecedented opportunities for data assimilation, correlation, and statistical analysis [Reed and Dongarra 2015].

By utilizing parallel computing platforms, various models and visualization techniques could be applied to accommodate the characteristics of big data analytics and take the most advantage of it. The following are some concrete examples of opportunities that are to be explored [Manyika et al. 2011; IBM 2012; Raghupathi and Raghupathi 2014; Priyanka and Kulennavar 2014; Issa et al. 2014; White 2014; Reed and Dongarra 2015]:

- Personalized care: Create predictive models to leverage personalized care (e.g., genomic DNA sequence for cancer care) in real time to highlight best-practice treatments for patients [Issa et al. 2014]. These solutions may offer early detection and diagnosis before a patient develops disease symptoms.

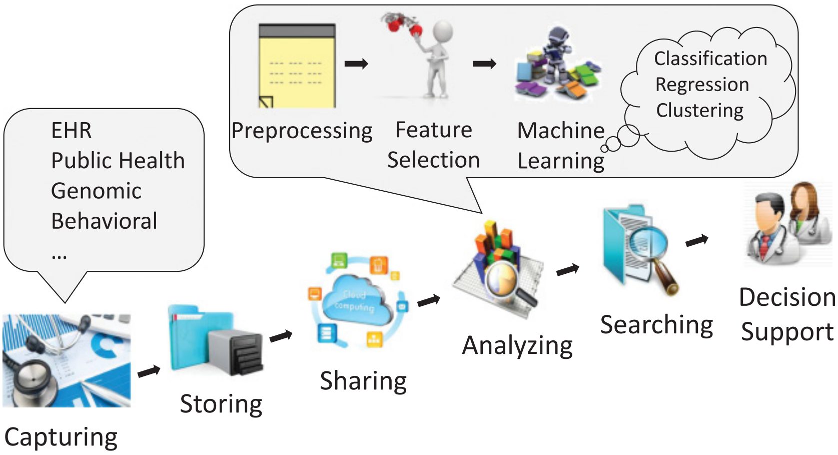


Fig. 3. Health informatics processing pipeline.

—Clinical operations: Conduct a comparative study to develop better ways for diagnosing and treating patients, such as mining large amounts of historical and unstructured data, looking for patterns, and model various scenarios to predict events before they actually happen [Issa et al. 2014; White 2014].

—Public health: Turn big healthcare data from a nationwide patient and treatment database into actionable information for timely detection and prevention of infectious diseases and outbreaks, thus benefiting the whole population [Raghupathi and Raghupathi 2014].

—Genomic analytics: Add genomic analysis to the traditional healthcare decision-making process by developing efficient and effective gene sequencing technologies. Utilize high-throughput genetic sequencers to capture organism DNA sequences and perform genome-wide association studies (GWASs) for human disease and human microbiome investigations [Reed and Dongarra 2015].

—Fraud detection: Analyze a large amount of claim requests rapidly by using a distributed processing platform (e.g., MapReduce for Hadoop) to reduce fraud, waste, and abuse, such as a hospital's overutilization of services, or identical prescriptions for the same patient filled in multiple locations [Raghupathi and Raghupathi 2014; White 2014].

—Device/remote monitoring: Capture and analyze continuous healthcare data in huge amounts from wearable medical devices both in the hospital and at home, for monitoring of safety and prediction of adverse events [Raghupathi and Raghupathi 2014].

3. HEALTH INFORMATICS PROCESSING PIPELINE

A mature computing framework of big data in health informatics involves a sequence of steps that constitute a comprehensive health informatics processing pipeline. Each step in the pipeline plays a critical role in rendering qualified and valuable outcomes of big data analytics. Specifically, the capturing, storing, and sharing of big data prepare appropriate input for the subsequent analyzing procedure, where various analytical approaches are applied to explore meaningful patterns from healthcare big data for making timely and effective decisions. This section will discuss the pipeline (as shown in Figure 3) in detail.

Since this is a survey article aiming at introducing the state of the art of health informatics in big data to the broad audience, the details of each part can be found in the references.

Table I. Healthcare Data Types and Sources

Data Type	Examples	Structured/ Unstructured	Data Format
Human-generated data	Physicians' notes, email, and paper documents	Structured & unstructured	ASCII/text
Machine-generated data	Readings from various monitoring devices	Structured & unstructured	Relational tables
Transaction data	Billing records and healthcare claims	Semistructured & structured	Relational tables
Biometric data	Genomics, genetics, heart rate, blood pressure, x-ray, fingerprints, etc.	Structured & unstructured	ASCII/text, images
Social media data	Interaction data from social websites	Unstructured	Text, images, videos
Publications	Clinical research and medical reference material	Unstructured	Text

3.1. Capturing

The discovery and utilization of meaningful patterns in data have helped to drive the success of many industries. However, the usage of analytics to improve outcomes in the healthcare industry has not gained as much traction due to the difficulty of collecting large and complex datasets, which is the very first step of big data analytics. Without the ability to capture big data, healthcare providers are unable to utilize analytics to improve outcomes.

Big data in health informatics are characterized by a variety of sources (from both internal and external points of view), diverse formats (such as flat files and database records), and different locations (either physically distributed machines or multiple sites) [Feldman et al. 2012]. The various sources and data types are listed in Table I [Cottle et al. 2013; Priyanka and Kulennavar 2014].

Recently, the introduction of EHRs to U.S. hospitals led the healthcare industry into a new, high-tech age, with a high potential for the use of analytics to improve outcomes. The EHR is well known for its benefits of delivering standardized medical records and improving patient care with reduced errors and costs. With the EHR, healthcare providers are able to access information about individual patients or populations, including demographics, medical history, laboratory test results, and personal statistics [Tippet 2014].

Taking advantage of big data means quickly capturing high volumes of data generated in many different formats with dynamic structures. However, there are issues like latency and scalability. Low latency is a highly desired property for stream processing as a big data technology, while scaling data integration is critical for adapting to the high-volume and high-velocity nature of big data. Apache Hadoop coupled with existing integration software and the Hadoop Map/Reduce framework could provide the computing environment for parallel processing. More recently, Hadoop Spark [Zaharia et al. 2010], a successor system that is more powerful and flexible than Hadoop MapReduce, is getting more and more attention due to its lower-latency queries, iterative computation, and real-time processing. Storm[1] is another scalable and fast distributed framework with a special focus on stream processing.

3.2. Storing

New technologies and the move to EHR are creating massive amounts of healthcare data with increased complexity, from various types of diagnostic images to physicians'

[1] https://github.com/nathanmarz/storm.

notes, laboratory results, and much more. It is essential to provide efficient solutions for cost-effective storage and management. More specifically, there are several critical properties for a desired healthcare data storage and management system [NetApp 2011b]:

—High availability: It is desirable for healthcare staff to access records quickly, securely, and reliably anywhere, anytime. Reliable and rapid retrieval of patients' information saves valuable time for physicians and enables them to make responses and deliver immediate care, which can mean the difference between life and death.
—High scalability: Healthcare data storage requirements can easily reach tens or hundreds of terabytes, or even petabytes, considering annual enterprise storage needs. To accommodate this explosive growth of data, the storage platforms should be incrementally scalable.
—Cost-effective: Storing and managing a high volume of healthcare data could be redundant and complex. An efficient storage infrastructure should reduce the cost and complexity and provide protection to the data without compromising performance.

Databases are the basic components of a storage infrastructure. In general, there are a number of storage options for analyzing big data. Compared with the traditional Relational DataBase Management Systems (RDBMS), the analytical RDBMSs can scale to handle big data applications. However, they are more appropriate for structured data. Hadoop technology expands the opportunity to work with a broader range of content with the Hadoop Distributed File System (HDFS) and the Map/Reduce programming framework [He et al. 2013]. Depending on what you are analyzing, the options vary. For example, Hadoop storage options are often associated with polystructured data such as text, while NoSQL databases are more appropriate for data with a multitude of representations [Madaan et al. 2013].

As mentioned earlier, electronic images and reports are among the major sources of healthcare big data. To accommodate the special needs of storing and managing these types of data, the technology of the Picture Archiving and Communication System (PACS) is developed, which has the advantages of economical storage and convenient access [Cooke Jr et al. 2003]. With the emergence and advancement of big data analytics in healthcare, many industrial companies have provided enterprise-level solutions for storing and management. For example, NetApp [2011a] provides EHR and PACS solutions for reducing storage management and complexity, transforming clinical workflows, and lowering the Total Cost of Ownership (TCO) of the data storage environment. Moreover, Intel Distribution for Apache Hadoop software [Intel 2011] provides distributed computation frameworks to store, preprocess, format, and normalize patients' free-text clinical notes, which are generally difficult and time-consuming to process and analyze. It achieves scalability by processing each patient's information separately and in parallel.

3.3. Sharing

After capturing patient and clinical data, the problem becomes how to securely exchange healthcare information with scientists and clinicians across a given healthcare organization over institutional, provincial, or even national jurisdictional boundaries [Crowne 2014]. There are a number of challenges associated with the sharing of big data in health informatics, for example:

—The ad hoc use of a variety of data formats and technologies in different systems and platforms
—The assurance of the controlled sharing of data by using secure systems

Cloud computing is one of the main reasons big data has been so ubiquitous in recent years. By ensuring storage capacity, server management, network power, and bandwidth utilities, cloud computing will help synchronize data storage with devices so that all of the information being generated automatically streams into internal systems, enabling the sharing of information. It means that security and management will be centralized and made more effective. To summarize, cloud deployments share common characteristics as follows [EMC 2011]:

—They involve an optimized or virtualized infrastructure, leverage the Internet for shared access, and charge for use based on actual consumption.
—The hardware is distributed and fault tolerant, satisfying privacy and data security requirements.
—They can enhance collaboration to mine patient and claims data.
—They allow a shared pool of computing and storage resources on a pay-as-you-go basis.

A successive usage of the cloud platform in healthcare is the “PACS-on-demand” systems. By storing and sharing medical images with cloud infrastructure, it reduces the need to invest in IT capacity as well as allowing efficient and secure collaboration with radiology specialists and affiliated practices [EMC 2011]. By using the cloud platform, it enables data sharing between healthcare entities (such as providers and payers), which often have disparate data systems that are unable to bring together different types of data to make healthcare decisions. The research society also discusses secure sensitive data sharing on big data platforms, which will help enterprises reduce the cost of providing users with personalized services and provide value-added data services [Dong et al. 2015].

3.4. Analyzing

Data mining approaches have been widely used for analyzing healthcare data. The major steps of the analyzing procedure, including preprocessing, feature extraction/selection, and machine learning, are addressed in the following sections.

3.4.1. Preprocessing. Real-world healthcare data is noisy, skewed, and heterogeneous in nature. It is impractical to directly apply analytical algorithms to the raw healthcare data due to its variety and complexity. Therefore, in order to improve data quality and prepare them for further analysis, dealing with noise and missing values in large-scale healthcare datasets through preprocessing becomes a necessity.

A typical healthcare data preprocessing procedure usually includes the following steps depending on the source and format of the data [Banaee et al. 2013; Sow et al. 2013]:

—Data cleaning: This step involves the removal of noise in healthcare data, which includes artifacts [Singh et al. 2011; Mao et al. 2012] and frequency noise in clinical data [Sow et al. 2013]. For example, thresholding methods are used to remove incompliant measurements [Fauci et al. 2008; Apiletti et al. 2009] and low-pass/high-pass filtering tools are usually applied to remove frequency noise in sensor signals [Hu et al. 2008; Frantzidis et al. 2010; Meng et al. 2013]. In Leema and Hemalatha [2011], an improved version of the SURF technique is proposed to cleanse healthcare data generated by Radio-Frequency Identification (RFID) readers.
—Missing value interpolation: Missing values may be caused by unintentional reasons (e.g., sensor failures) or intentional reasons (e.g., transportation of patients) [Fialho et al. 2012]. Usually, single missing values caused by sensor failures are interpolated by its previous and following measurements [Apiletti et al. 2009]. Data missing

for an intentional reason or because of irrelevancy to a current clinical problem is considered nonrecoverable and thus deleted [Fialho et al. 2012].

—Data synchronization: Sensor data is reported at different rates with timestamps based on their internal clocks [Sow et al. 2013]. However, the clocks across sensors are often not synchronized. It is necessary to make reasonable assumptions and derive alignment strategies to synchronize sensor data [Jané et al. 1991; Martinez Orellana et al. 2013].

—Data normalization: This step is often required to cope with differences in the data recording process [Sow et al. 2013]. For example, a daily heart rate may represent a daily average heart rate or a measurement during a specific time range. Furthermore, a normalization step is usually performed to transform the original feature set into a comparable format by adopting and mapping standardized terminologies and code sets.

—Data formatting: Each analytical approach (such as data mining algorithms) requires data to be submitted in a specified format [Eapen 2004]. Therefore, there is a necessity to transform the original healthcare data into machine-understandable format. For example, the data should be stored in the Attribute-Relation File Format (.ARFF format) and the data type of the attributes must be declared in a recognizable manner in the WEKA tool [Hall et al. 2009]. Sometimes, a discretization process should be carried out to convert the original numerical values to nominal ones for a specific algorithm. It is worth mentioning that the discretization process may cause information loss, and thus impact data quality. This is another challenge for big healthcare data analysis.

Performing preprocessing for healthcare data is a challenging task, which requires an efficient and effective big data framework. One of the solutions, the Hadoop platform, is designed to store and process extremely large datasets [Zolfaghar et al. 2013]. To simulate a scalable data warehouse and make the data extraction process more effective, big data tools such as Hive [Thusoo et al. 2009] and Cassandra are coupled with the MapReduce framework, built on top of Hadoop, to provide a powerful distributed data management system with flexibility. Specifically, first raw healthcare data will be stored as flat files on various nodes in Hadoop, which will later be loaded into a Hadoop File System (HDFS). Later, Hive commands will be invoked to create appropriate tables and develop schema to structurize the data to be queried.

3.4.2. Feature Extraction/Selection. The process of extracting and selecting a subset of important and relevant features from a large set of measured data is called feature selection, also known as attribute selection, or variable selection. In general, input data may include redundant features, which can be transformed into a small set of relevant features using dimensional reduction algorithms. By applying feature selection algorithms, complex and large datasets that are computationally expensive and need large amounts of memory are transformed to a small, relevant feature set with sufficient accuracy. Medical big and high-dimensional data may cause inefficiency and low accuracy. To overcome this issue, many researchers utilize feature extraction algorithms in healthcare informatics [Soliman et al. 2015].

Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA) are the most commonly used algorithms for dimension reduction and feature selection in the healthcare domain [Banaee et al. 2013]. In PCA, d-dimensional data is transformed into a lower-dimensional space to reduce complexities. Pechenizkiy et al. [2004] evaluate several feature transformation strategies based on PCA in healthcare diagnostics. Five medical datasets are used for evaluating PCA feature transformation: diabetes, liver disorders, heart disease, thyroid gland, and breast cancer. Experimental results show that PCA is appropriate for

Table II. Summary of Feature Extraction/Selection Algorithms in the Healthcare Informatics

Feature Extraction/ Selection Algorithm	Healthcare Examples	Pros	Cons
PCA [Pechenizkiy et al. 2004; Subasi and Gursoy 2010]	diabetes, heart disease, breast cancer	simple, nonparametric, spread-out data in the new basis, useful in unsupervised learning, most used algorithm in healthcare informatics	nonstatistical method, difficult to interpret, linear combinations of all input variables
LDA [Subasi and Gursoy 2010]	hepatitis diagnosis, coronary artery disease	multiple dependent variables, reduced error rates, easier interpretation	parametric method (assumes unimodal Gaussian likelihood), mean is the discriminating factor not variance, extremely sensitive to outliers, produces limited feature projections
ICA [Subasi and Gursoy 2010]	heart disease, genetic disease	finds uncorrelated and independent components, generalization	inappropriate for small training data, permutation and stability ambiguity
Tree based [Fialho et al. 2012; Sun et al. 2014]	ICU readmission, EHR and MRI datasets	graphical representation, simple	traps to local optimum, redundant features might be selected
CFS [Hall 1999]	breast, colon, and prostate cancer	fast, reduced error rate, better results on small datasets	ignores interaction with classifier, less scalable

certain problems related to medical diagnostics. In Subasi and Gursoy [2010], the effect of various statistical feature selection approaches such as PCA, LDA, and ICA on the medical domain are discussed.

Fialho et al. [2012] apply a tree-based feature selection algorithm combined with a fuzzy modeling approach to a large ICU database to predict ICU readmissions. The tree search feature selection algorithm builds a tree to order all the possible feature combinations. This approach appeals to clinicians due to its graphical representation and simplicity. In this article, Sequential Forward Selection (SFS) and Sequential Backward Elimination (SBE) [Mao 2004] are applied to the ICU datasets as a tree search technique. Sun et al. [2014] also apply a greedy forward selection approach to the EHR dataset. The proposed method includes a two-level feature reduction, which selects predictive features based on Information Gain (IG). In every step, the best-performing features from one concept combine with the features selected from the next proper concept. The combination process will be continued until prediction performance fails to improve. Although the greedy tree search approach is simple and easy to interpret, it may get trapped at a local optimum [Fialho et al. 2012].

Correlation Feature Selection (CFS) is a measure that evaluates the merit of feature subsets based on the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other" [Hall 1999]. To predict 5-year life expectancy in older adult,s in Mathias et al. [2013], CFS combined with greedy stepwise search is used as a feature selection strategy to find the feature subset with the best average competency. Table II summarizes some feature extraction/selection algorithms in healthcare informatics, as well as their pros and cons.

Several feature selection algorithms dealing with big data have been recently proposed. Unlike conventional feature selection algorithms, online learning, specifically online feature selection, is suitable for large-scale real-world applications in such a way that each feature is processed upon its arrival and each time the best feature set is maintained from all seen features [Li et al. 2013]. Yu et al. [2014] propose a Scalable and Accurate Online Approach (SAOLA) to select important features from large datasets with high dimensionality. Using a sequential scan, the SAOLA overcomes critical challenges in online feature selection, including computational cost of online processing specifically for large datasets, which keeps growing. Tan et al. [2014] also present a new feature selection method for extremely high-dimensional datasets on big data. In this article, the problem is transformed to a convex Semi-Infinite Programming (SIP) issue, which is solved by a new Feature Generating Machine (FGM). The FGM repeatedly extracts the most relevant features using a reduced and primal form of Multiple Kernel Learning (MKL) subproblem. The FGM is appropriate for feature selection on big data due to its subproblem optimization, which involves a small subset of features with reduced memory overhead.

3.4.3. Machine Learning. Machine learning is the study of computer science that explores the creation of algorithms that can automatically learn from data and improve through experience. Nowadays, machine-learning techniques have been applied in a variety of applications including audio processing, autonomous vehicle, and detection of fraudulent credit card activity, to name a few [Mitchell 1997]. Computerization in healthcare is growing day in and day out, which leads to complex and large medical databases. Machine-learning algorithms are able to automatically manage such large databases. A general overview of machine-learning techniques in healthcare informatics is presented in the following.

Generally, machine-learning algorithms are subdivided into two categories: supervised learning (or predictive learning) and unsupervised learning (or descriptive learning) [Yoo et al. 2012]. In supervised learning, both input and their desired outputs are presented to learn a general prediction rule that maps inputs to outputs. In other words, prediction rules are acquired from training data to predict unseen data labels. Classification and regression are the two major categories of supervised learning studied in this survey. The classification algorithms surveyed in this article are as follows: First, the decision tree is introduced. It is a simple and easy-to-implement classifier, which is useful for physicians who want to easily understand and interpret the classification results. However, it may not be applicable for very large datasets with high-dimensional features due to its space limitation and overfitting problem. Another classifier presented is Support Vector Machine (SVM), which is widely applied on image-based classification and large medical datasets, in spite of its slow training and expensive computational time. The Neural Network (NN) classifier is also presented, which has the same weaknesses as SVM, as well as its black-box nature and difficulty in interpretation. Although the NN is broadly used for medical applications, its modified version, called deep learning, has better capabilities in dealing with big data issues such as volume and velocity. Deep learning algorithms are usually used as classifiers in image-based medical research such as neuroimaging applications. In addition, different types of sparse classifiers and ensemble algorithms are introduced in the following sections, which are widely used in the big data medical datasets to overcome imbalanced data and the overfitting problem, respectively. However, most of these algorithms are computationally expensive and are not easy to interpret. The regression algorithms, another type of supervised learning, are also widely used in healthcare applications such as brain image analysis, CT scans of different organs, and battery health monitoring.

In unsupervised learning, there are no labels, and the goal is to find the structure of unknown input data by searching the similarity between records. Clustering is one type of unsupervised learning approach. Three different types of clustering algorithms are covered in this survey as follows: The partitioning algorithm is presented as it is simple, fast, and capable of handling large datasets; however, it is not recommended for noisy datasets including lots of outliers. The hierarchical algorithm is another clustering method discussed in this survey due to its visualization capability, which is requested by many physicians, though it is not sometimes appropriate for big data due to space and time limitations. The last algorithm presented in this section is density-based clustering, which handles nonstatic and complex datasets and detects outliers and arbitrary shapes specifically in biomedical images. However, it is slow for large datasets, like hierarchical clustering.

Until now, various research studies have been conducted on machine learning and data mining that outline advantages and disadvantages of multiple machine-learning algorithms in healthcare informatics [Austin et al. 2013; Dittman et al. 2013; Banaee et al. 2013]. It is worth mentioning that there is a significant overlap among machine learning and data mining as both are used in data analysis research and include supervised and unsupervised algorithms. Machine learning is used to extract models and patterns in data. However, data mining, a combination of machine learning and statistics, mainly deals with existing large datasets and analyzes massive, complicated, and structured/unstructured data. Data mining is a more general concept, which has gradually merged with and been used in database management systems. Therefore, data mining algorithms should be more scalable to discover rich knowledge from large datasets.

3.4.3.1. Classification. Classification is the problem of identifying the category of new observation records based on the training data whose categories are known. Automatic classification can be used in diagnosis systems to help clinicians with disease detection. In the following, several widely used classification algorithms and their applications in healthcare and medical informatics are discussed:

(1) Decision tree and rule-based algorithms: Decision tree is a simple and widely used classifier. It classifies instances by sorting them in a tree, which can be re-represented as sets of if-then rules [Mitchell 1997]. This learning method has been successfully used in a wide variety of medical diagnostic systems [Dittman et al. 2013; Banaee et al. 2013].

Zhang et al. [2012] introduce a real-time prediction and diagnosis algorithm. This prediction algorithm is mainly based on Very Fast Decision Tree (VFDT) [Domingos and Hulten 2000]. VFDT is a decision tree algorithm form on the Hoeffding tree that can control large and continuous data streams by remembering the mapping connection among leaf nodes and the history entries. Zhang et al. discuss that VFDT outperforms traditional decision tree algorithms, although it is not able to predict the patient illness using only the current situation on its own. Therefore, they modified VFDT as follows: Several pointers of leaf node are added in the training phase of learning. Then, a mapping table is designed to store medical data, the address, and its pointer. Therefore, whenever VFDT sends a stream to a leaf node, the corresponding pointer is used to search the table and return similar-history medical records. These similarities can be used for medical prediction and help physicians to better treat their patients.

Fuzzy Decision Tree (FDT) is compared with three other classifiers in Estella et al. [2012], which demonstrates the performance superiority of FDT in brain MRI classification problems. Unlike traditional decision trees, FDT has the ability to handle fuzzy data. In this study, an automatic classification algorithm is proposed, which includes the preprocessing step (wavelet decomposition of the image) and feature selection algorithm in order to extract a few morphological features, as well as FDT classifier. The authors conclude that their method can efficiently classify patients into the three levels of Alzheimer's (close to 90% efficiency, which is better than human experts) using a few morphological features.

Other utilizations of decision trees in medical and healthcare informatics are presented in Banaee et al. [2013]. Although decision tree techniques are simple and easy to implement, they have space limitations. Furthermore, if the dataset contains many features, it may be inefficient to create a tree. To overcome this issue, as well as dataset overfitting [Hawkins 2004], pruning algorithms are used in decision trees.

(2) Support Vector Machine: SVM is another supervised learning method used for both classification and regression [Cristianini and Shawe-Taylor 2000]. As a classifier, SVM constructs hyperplanes in a multidimensional space to classify samples with different labels. In the SVM models, several kernel functions can be used including polynomial, sigmoid, Gaussian, and Radial Basis Function (RBF).

In Zhou et al. [2010], children's health is studied, and specifically, socioeconomic status on educational attainment is addressed. In order to reduce dimensionality and identify the important features, an SVM-based classifier is optimized by the particular swarms optimization (PSO) to update the hyperparameters in an automatic manner and simultaneously identify the important features via entropy regularization. To evaluate the proposed method, 21 features are extracted from 3,792 data samples, which are divided into training data (for SVM model construction), validation data (for model selection by PSO), and testing data. The method is compared with LDA and Multilayer Perceptron (MLP) [Pal and Mitra 1992], which are linear and nonlinear classification algorithms, respectively.

(3) Artificial Neural Networks: NN is a family of statistical learning and artificial intelligent approaches widely used for classification. This algorithm is inspired by biological neural networks of animals and humans, particularly the brain and central nervous system. NN includes neurons that can be trained to compute values from inputs and predict corresponding classes of test data. Until now, there have been a wide variety of decision-making processes and predictions using NN in the healthcare domain.

Vu et al. [2010] present an online three-layer NN to detect Heart Rate Variability (HRV) patterns. Specifically, HRV related to coronary heart disease (CHD) risk is recognized using electrocardiogram (ECG) sensors. When a sample enters, the hidden layer nodes find the excessive similarity between nodes and the input in a competitive manner. The proposed method outperforms other NN algorithms such as MLP, Growing Neural Gas (GNG) [Fritzke 1995], and Self-Organizing Map (SOM) [Kohonen 1998].

Yoo et al. [2012] discuss several disadvantages of NN in the healthcare domain as follows: first, NN requires numerous parameters, which is very critical for the classification result. Second, the training phase of NN is computationally expensive and time consuming. Furthermore, it lacks model transparency due to the black-box nature of the NN system and thus it is difficult for medical experts to understand its structure in order to gain knowledge from it. Finally, the accuracy is usually lower than other algorithms such as random forest and SVM, to name a few.

(4) Deep Learning: With the tremendous growth of data, deep learning is playing an important role in big data analysis. One notable success of deep learning for big data is the use of a large number of hidden neurons and parameters, involving both large models and large-scale data [Chen and Lin 2014], to model high-level abstraction in data [Nielsen 2014]. To date, various deep learning architectures including deep belief networks [Hinton 2009], deep neural networks [Larochelle et al. 2009], deep Boltzmann machine [Salakhutdinov and Hinton 2009], and deep convolution

neural network [Krizhevsky et al. 2012] have been applied in areas such as speech recognition, audio processing, computer vision, and healthcare informatics.

Liang et al. [2014] present a healthcare decision-making system using multiple-layer neural network deep learning to overcome the weaknesses of conventional rule-based models. The deep model combines features and learning in a unified model to simulate the complex producer of the human brain and thinking. To achieve this, a modified version of the deep belief network is applied on two large-scale healthcare datasets.

Deep learning models have shown success in neuroimaging research. Plis et al. [2014] use deep belief networks and restricted Boltzmann machine for functional and structural MRI data. The results are validated by examining whether deep learning models are effective compared with representative models of its class, examining the depth parameter in the deep learning analysis for this specific medical data, and determining if the proposed methods can discover the unclear structure of large datasets.

Li et al. [2014] also leverage the deep-learning-based framework to estimate the incomplete imaging data from the multimodality database. They take the form of convolution neural networks, where the input and output are two volumetric modalities. They evaluate this deep-learning-based method on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, where the input and output are MRI and PET images, respectively.

(5) Sparse Representation: Sparse Representation Classification (SRC) of signals has been attracting a great deal of attention in recent years. SRC is the problem of finding the most compact signal representation using atoms' linear combination in a given overcomplete dictionary [Huang and Aviyente 2006].

In the healthcare informatics area, several research studies have been done using SRC to improve the classification results. Marble [Ho et al. 2014] is a sparse nonnegative tensor factorization method for count data. It is used to fit EHR count data that are not always correctly mapped to phenotypes. In this model, the sparsity constraints are imposed by decreasing the unlikely mode elements. To solve the optimization problem, a periodic minimization approach (cycling through each mode as fixing others) is used.

Sparse representation has also been an effective approach to learn characteristic patterns from the medical data for image restoration, denoising, and superresolution. Sparse representation has been effective in medical image denoising and fusion using group-wise sparsity [Li et al. 2012] and image reconstruction [Xu et al. 2012; Gholipour et al. 2010]. Recently, coupled with dictionary learning, Fang et al. [2013] restored the hemodynamic maps in the low-dose computed tomography perfusion by learning a compact dictionary from the high-dose data, with improved accuracy and clinical value using tissue-specific dictionaries [Fang et al. 2015] and applying to various types of medical images [Fang et al. 2014]. The sparsity property in the transformed domain has also been important in restoring the medical information by combining with the physiological models [Fang et al. 2015].

(6) Ensemble: Ensemble is a supervised learning algorithm that combines different classification algorithms to increase the performance of single classifiers. In other words, instead of using an individual classifier, ensemble learning can be used to aggregate the prediction results of several classifiers. Ensemble learning improves the generalization and predictive performance.

Random forest [Breiman 2001] is one type of ensemble learning algorithm that constructs multiple trees at training time. This algorithm overlaps the overfitting problem of decision trees by averaging multiple deep decision trees. Recently, various research works have been done applying the random forest algorithm to the bioinformatics domain. Díaz-Uriarte and De Andres [2006] apply random forest for classification of microarray data. They also apply random forest for gene selection tasks. In this research, 10 DNA microarray datasets focusing on several parts of the body are used.

Rotation Forest Ensemble (RFE) [Rodriguez et al. 2006] with Alternating Decision Tree (ADT) [Freund and Mason 1999] is a modified version of the decision tree technique that is used as a classifier in the paper by Mathias et al. [2013]. For this purpose, 980 features are extracted from EHR data. Afterward, the greedy stepwise algorithm is used for feature selection. Finally, RFE with ADT is applied to predict the 5-year mortality rate.

Liu et al. [2012] propose ensemble learning using a sparse representation algorithm to classify Alzheimer's disease from medical images such as MRI. For this purpose, a random patch-based subspace ensemble classification is proposed. This technique builds several individual classifiers using various subsets of local patches and finally combines weak classifiers to improve performance results. Experimental results show the high performance of the proposed method on Alzheimer's MRI data.

(7) Other Classification Algorithms: In this section, other classification algorithms used in healthcare informatics are introduced.

The **Hidden Markov Model (HMM)** is a statistical model representing probability distribution over the sequences of observations [Ghahramani 2001]. This model use Markov chain to model signals in order to calculate the occurrence probability of states. Cooper and Lipsitch [2004] apply structured and unstructured HMM to analyze hospital infection data. Structured HMM is more parsimonious and can estimate important epidemiological parameters from time series data. Another work on HMM addresses the detection of anomalies in measured blood glucose levels [Zhu 2011]. Based on the experimental results, the HMM technique is accurate and robust in the presence of moderate changes.

The **Gaussian Mixture Model (GMM)** is another statistical model widely used as a classifier in pattern recognition tasks. It consists of a number of Gaussian distributions in the linear way [Wang et al. 2011]. Giri et al. [2013] present a method that uses GMM for the automatic detection of normal and coronary artery disease conditions with ECG signals. In this research, various feature selection and classification algorithms are applied to heart rate signals. The GMM classifier combined with the ICA reduction algorithm results in the highest accuracy compared to other techniques such as LDA and SVM.

Bayesian networks, also known as belief networks, are another probabilistic model corresponding to a graphical model called Directed Acyclic Graph (DAG) [Ben-Gal 2007]. Each node of the graph represents a random variable and the edge between the nodes indicates probabilistic dependencies among the related random variables. Bayesian network, for instance, could be applied to measure the probabilities of the existence of various diseases and find the relationship between diseases and symptoms.

Furthermore, some researchers have developed more complex classification algorithms in the healthcare domain, including Collateral Representative Subspace Projection Modeling (C-RSPM) [Meng et al. 2010] and the multilayer classification framework, for discovering the temporal information of biological images [Meng and Shyu 2013].

3.4.3.2. Regression. Regression is a supervised learning algorithm used to model relationships between objects and targets. The difference between regression and classification is that in regression, the target is continuous, while the latter is discrete. In other words, regression is the problem of approximating a real-valued target function [Mitchell 1997].

Yoshida et al. [2013] propose a Radial Basis Function-sparse Partial Least Squares (RBF-sPLS) regression and apply it to high-dimensional data including MRI brain images. The sPLS regression [Lê Cao et al. 2008] reduces dimension and selects features simultaneously using a sparse and linear combination of the explanatory variables. The proposed method, a combination of sPLS with basis expansion, is applicable to real data including MRI brain images with large-scale characteristics from chronic kidney disease patients. The authors evaluate the performance of RBF-sPLS by comparing it with the method without basis expansion.

Saha et al. [2007] introduce a Bayesian regression estimation algorithm implemented as a Relevance Vector Machine (RVM) via Particle Filters (PFs). This algorithm is used to integrate monitoring, diagnosis, and prediction of battery health. RVM is a Bayesian form obtaining solutions for regression and probabilistic classification [Tipping 2001], which represents a generalized linear form of the SVM. To estimate state dynamically, a general framework is provided using PF. The results show the advantage of proposed methods over conventional methods of battery health monitoring.

Regression forest is another supervised algorithm that is used for anatomy localization and detection in Criminisi et al. [2013]. The main goal of this algorithm is to train a nonlinear mapping from a complicated input to continuous parameters (from voxels to the location and size of the organ), and its difference with other forest classifiers is that it is utilized to predict multivariate and continuous outputs using a tree-based regression method. To evaluate the proposed regression algorithm, it is applied on a database including 400 three-dimensional CT scans with high-variety images, which shows the robustness and the accuracy of the trained model compared to the conventional methods.

3.4.3.3. Clustering. Clustering [Jain et al. 1999] is the task of categorizing a group of objects into subgroups (called clusters) in such a way that objects in the same category are more similar to each other compared with those in other categories. Clustering is a technique of unsupervised learning and statistical analysis that is applicable in many fields. The main difference between classification and clustering is that clustering does not use labels and finds natural grouping based on the structure of objects. Several clustering algorithms applied in healthcare data are discussed in the following. Further information regarding clustering algorithms for healthcare are discussed in Tomar and Agarwal [2013].

(1) Partitioning algorithms: Partitioning clustering algorithms divide a set of data objects into various partitions (clusters) in a way that each object is placed in exactly one partition [Tan et al. 2006]. K-means is a known partitioning clustering algorithm used in various areas such as computer vision, data mining, and market segmentation. It partitions objects into k clusters, computes centroids (mean points) of the clusters, and assigns every object to the cluster that has the nearest mean in an Expectation-Maximization fashion.

    Zolfaghar et al. [2013] present a big-data-driven solution for predicting Risk of Readmission (RoR) of congestive heart failure patients. To achieve this, they apply data mining models to predict risk of readmission. K-means is used to segment the National Inpatient Sample (NIS) dataset. Using K-means, the average income of each patient (used as a predictor variable for ROR) is calculated to map each record of a dataset to the closest cluster based on Euclidean distance.

    The graph-partitioning-based algorithm is another popular clustering method that is applicable for partitioning a graph G into subcomponents with specific properties. Yuan et al. [2009] apply a new skin lesion segmentation algorithm using a Narrow-Band Graph-Partitioning (NBGP) and region fusion to improve the efficiency and performance of the segmentation algorithm. The proposed method can produce skin lesion segmentation well even for weak edges and blurred images.

A modified version of K-means recently utilized is called subspace clustering. It partitions datasets along with the various subsets of dimension instead of the whole space, to overcome the challenge of curse of dimensionality in big data analysis [Kriegel et al. 2009]. Hund et al. [2015] apply a subspace clustering approach to the real-world medical data and analyze the patient data relationship and immunization treatment. Experimental results show how subspace clustering can effectively identify grouping of patients compared with a full space analysis such as hierarchical clustering.

(2) Hierarchical algorithms: Hierarchical algorithms build a hierarchy known as a dendrogram [Nithya et al. 2013]. There are two general strategies for hierarchical clustering: the Agglomerative (bottom-up) approach and the Divisive (top-down) approach. Both combination (for agglomerative) and splitting (for divisive) of clusters are determined in a greedy manner in which a distance or dissimilarity measure is needed by use of an appropriate metric such as Euclidean distance and Manhattan distance.

A hybrid hierarchical clustering algorithm is proposed by Chipman and Tibshirani [2006] to analyze microarray data. The proposed method utilizes mutual clusters and takes advantage of both bottom-up and top-down clustering. Belciug [2009] also applies a hierarchical agglomerative clustering for categorizing and clustering the patients based on the length of their stay in terms of days at the hospital.

(3) Density-based algorithms: Density-based clustering algorithms are extensively used to search for clusters of nonlinear and arbitrary shapes based on the density of connected points. This algorithm defines clusters by a radius that contains maximum objects based on a defined threshold. One of the most popular density-based clustering approaches is called Density-Based Clustering of Applications with Noise (DBSCAN), which was first introduced by Ester et al. [1996]. DBSCAN not only groups points with many close nearest neighbors and high-density areas but also detects outliers and noise points within low-density areas.

The density-based algorithm is also widely applied on healthcare and medical datasets such as biomedical images. In Celebi et al. [2005], an unsupervised region-based segmentation approach based on DBSCAN is used to detect homogeneous color regions in skin lesion images. The segmentation approach separates lesions from healthy skin and also detects color regions inside the lesions. Experimental results show that lesion borders are successfully identified in 80% of the tested biomedical images. As density-based algorithms cannot efficiently cluster high-dimensional datasets, a Hierarchical Density-based Clustering of Categorical data (HIERDENC) is proposed by Andreopoulos et al. [2007] to overcome some challenges of the conventional clustering approaches. HIERDENC detects clusters in a better runtime scalability that is more efficient for large datasets and big data applications. When new data is introduced, only the HIERDENC index is updated, so that there is no need to repeat clustering on all data. HIERDENC is applied on several large and quickly growing biomedical datasets such as finding bicliques of biological networks (protein interaction or human gene networks), clustering of biomedical images, and retrieving clusters of Force-Distance curves.

A summary of the aforementioned machine-learning algorithms with some healthcare examples, as well as their pros and cons, is shown in Table III.

3.5. Searching

Specialists and physicians use analyzed data to search for systematic patterns in patients' information, which helps them in having a more precise diagnosis and treatment.

Table III. Summary of Machine-Learning Algorithms in Healthcare Informatics

ML Category	Algorithm	Healthcare Examples	Dataset Examples	Pros	Cons
Classification	Decision Tree	brain MRI classification, medical prediction	ADNI [ADNI 2015], hemodialysis [Yeh et al. 2011]	simple, easy to implement	space limitation, overfitting
	SVM	image-based MR classification, children's health	NCCHD [NCCHD 2014]	high accuracy	slow training, computationally expensive
	Neural Network	cancer, blood glucose level prediction, Heart rate Variability recognition	Cleveland [Rani 2011], Acute Nephritis Diagnosis [Khan et al. 2013]	handle noisy data, detect nonlinear relationship	slow, computationally expensive, black-box models, low accuracy
	Sparse	EHR count data, heartbeats classification, tumor classification, gene expression	colon cancer, MIT-BIH ECG [Huang et al. 2012], DE-SynPUF [Ho et al. 2014]	efficiency, handle imbalanced data, fast, compression	computationally expensive
	Deep Learning	registration of MR brain images, healthcare decision making, Alzheimer diagnosis	ADNI, Huntington disease [PREDICT-HD 2015]	handle large dataset, deal with deep architecture, generalization, unsupervised feature learning, support multi task learning and semisupervised learning	difficult to interpret, computationally expensive
	Ensemble	microarray data classification, drug treatment response prediction, morality rate prediction, Alzheimer classification	ADNI	overcome overfitting, generalization, predictive, high performance	hard to analyze, computationally expensive
	Other Classifiers	hospital infection analysis, anomalies detection, health monitoring, drug reaction signal generation, health risk assessment	ECG data, ADRs [Bate et al. 1998]	depends on method	depends on method
Regression	-	brain imaging analysis, battery health diagnosis	CKD [Singh et al. 2012], Li-ion batteries [Saha et al. 2007]	depends on method	depends on method
Clustering	Partitioning	risk of readmission prediction, depression clustering	NIS and MHS [Zolfaghar et al. 2013], ADNI	handle large datasets, fast, simple	high sensitivity to initialization, noise and outliers
	Hierarchical	microarray data clustering, patients grouping based on length of stay in hospital	microarray datasets, HES [HSCIC 2012]	visualization capability	poor visualization for large data, slow, use huge amount of memory, low accuracy
	Density-based	biomedical image clustering, finding bicliques in a network	skin lesion images, BMC biomedical images	detect outliers and arbitrary shapes, handle nonstatic and complex data	not well for large datasets, slow, tricky parameter selection

Data mining is an analytic process that is designed to search and explore large-scale data (big data) to discover consistent and systematic patterns. One of the main challenges in big data mining in the medical domain is searching through unstructured and structured medical data to find a useful pattern from patients' information.

Information Retrieval (IR) is the process of extracting, searching, and analyzing data objects based on metadata or other content-based indexing [Salton and Harman 2003]. Data objects may be text documents, images, and audio. In this section, text and image retrieval in the medical domain are discussed.

Text mining refers to the process of extracting information and analyzing unstructured textual information [Popowich 2005]. Extracting information from textual documents is a useful information retrieval technique widely used in healthcare informatics [Holzinger et al. 2014b; Jung et al. 2014; Vijayakrishnan et al. 2014]. Using text mining, it is possible to extract information from patient records, reports, lab results, and, generally, clinical notes. The major problem is that the clinical notes are unstructured. To tackle this challenge, a wide variety of methods have been applied to analyzing these unstructured texts in the field of Natural Language Processing (NLP) [Spyns 1996].

Content-Based Image Retrieval (CBIR) reveals its crucial role in medical image analysis by providing physicians and doctors with diagnostic aid including visualizing existing and relevant cases, together with diagnosis information. Therefore, retrieving images that can be valuable for diagnosis is a strong necessity for clinical decision-support methods including evidence-based medicine or case-based reasoning. In addition, their use will allow for the exploration of structured image databases in medical education and training. Therefore, information retrieval in medical images has been widely investigated in this community.

For example, Comaniciu et al. [1999] propose a CBIR system supporting decision making in the domain of clinical pathology, in which a central module and fast color segmenter are used to extract features such as nucleus appearance (e.g., texture, shape, and area). Performance of the system is evaluated using a classification with 10-fold cross-validation and compared with that of an individual expert on a database containing 261 digitized specimens.

CBIR has been employed for histopathological image analysis. For example, Schnorrenberg et al. [2000] extend the biopsy analysis support system to consist of indexing and CBIR for retrieving biopsy slide images. A database containing 57 breast cancer cases is used for evaluation. Akakin and Gurcan [2012] propose a CBIR system using the multitiered approach to classify and retrieve microscopic images. To maintain the semantic consistency between images retrieved from the CBIR system, both “multi-image” query and “slide-level” image retrieval are enabled.

As emphasized in Zhou et al. [2008], scalability is the key factor in CBIR for medical image analysis. In fact, with the ever-increasing amount of annotated medical data, large-scale, data-driven methods provide the promise of bridging the semantic gap between images and diagnoses. However, the development of large-scale medical image analysis algorithms has lagged greatly behind the increasing quality and complexity of medical images.

Specifically, owing to the difficulties in developing scalable CBIR systems for large-scale datasets, most previous systems have been tested on a relatively small number of cases. With the goal of comparing CBIR methods on a larger scale, ImageCLEF and VISCERAL provide benchmarks for medical image retrieval tasks [Müller et al. 2005; Langs et al. 2013; Hanbury et al. 2013].

Recently, hashing methods have been intensively investigated in the machine-learning and computer vision community for large-scale image retrieval. They enable fast Approximated Nearest Neighbors (ANN) search to deal with the scalability issue. For example, Locality-Sensitive Hashing (LSH) [Andoni and Indyk 2006] uses random projections to map data to binary codes, resulting in highly compact binary codes and enabling efficient comparison within a large database using the Hamming distance. Anchor Graph Hashing (AGH) [Liu et al. 2011] has been proposed to use neighborhood graphs that reveal the underlying manifold of features, leading to a high search accuracy. Recent research has focused on data-dependent hash functions, such as the spectral graph partitioning and hashing [Weiss et al. 2009] and supervised hashing with kernels [Liu et al. 2012] incorporating the pairwise semantic similarity and dissimilarity constraints from labeled data. These hashing methods have also been employed to solve the dimensionality problem in medical image analysis. Particularly, Zhang et al. [2014, 2015b] build a scalable image-retrieval framework based on the supervised hashing technique and validate its performance on several thousand histopathological images acquired from breast microscopic tissues. It leverages a small amount of supervised information in learning to compress a high-dimensional image feature vector into only tens of binary bits with the informative signatures preserved. The supervised information is employed to bridge the semantic gap between low-level image features and high-level diagnostic information, which is critical to medical image analysis.

In addition to hashing and searching the whole image, another approach is to segment all cells from histopathological images and conduct large-scale retrieval among cell images [Zhang et al. 2015a]. This enables cell-level and fine-grained analysis, achieving high accuracy. It is also possible to fuse multiple types of features in a hashing framework to improve the accuracy of medical image retrieval. Specifically, the Composite Anchor Graph Hashing algorithm [Liu et al. 2011] has been developed for retrieving medical images [Zhang et al. 2014] (e.g., retrieving lung microscopic tissue images for the differentiation of adenocarcinoma and squamous carcinoma). Besides hashing-based methods, vocabulary tree methods have also been intensively investigated [Nister and Stewenius 2006] and employed for medical image analysis [Jiang et al. 2015].

3.6. Decision Support

Clinical Decision Support (CDS) is a process for improving the quality of healthcare. It helps physicians, doctors, and patients to make better decisions [Clayton and Hripcsak 1995]. Although CDS systems have made great contributions to improve medical care and reduce healthcare errors, they do not always improve clinical decision support systems due to some technical and nontechnical factors. Kawamoto et al. [2005] studied the literature to recognize the specific factors and features of such systems for enhancing clinical practice. Based on the results, 68% of decision support systems enhance clinical practice remarkably, and they utilize four features, automatic provision of decision support, provision of recommendations, provision of decision support during decision making and at its location, and computerized decision support, that are significantly correlated with system success. In addition, some direct experimental evidence proves the significance of three extra features, sharing decision support with patients, providing performance feedback periodically, and requesting reason documentations if system recommendations is not followed.

Healthcare systems can leverage new technologies in big data to provide better clinical decision support. Today, CDS is a hot topic and an essential system in hospitals since it improves clinical output and efficiency of healthcare. Using big data technologies, many CDSs' limitations have been broken and dynamic clinical knowledge base systems have been created to deploy more complicated models for the CDS systems. Therefore, big data makes the CDS systems more credible and effective [Xiao et al. 2014].

3.6.1. Patient Similarity. Patient similarity computation is a significant process in healthcare informatics and decision support systems, and it finds patients with similar clinical characteristics. It is very helpful for decision support applications and predicting patients' future conditions. The main goal is to find the similarity between patients by extracting distance metrics. Based on the IBM Patient Care and Insights solution [IBM 2015b], the patient similarity algorithm includes the following steps:

—Both structured (e.g., clinical factors) and unstructured data (e.g., physicians' notes) are integrated for analysis.
—Personalized healthcare delivery plans are generated based on the health history of each individual patient.
—Personalized treatment plans are created using thousands of patient characteristics examined by professionals.

—Patients with similar clinical characteristics are identified, which helps doctors to see what treatments were most effective.
—Based on patient-physician matching, each patient is paired with a doctor who is efficient for a specific condition.

Because each expert has different knowledge on patient similarities, Wang et al. [2012] propose an approach called Composite Distance Integration (Comdi) to unify the single metrics achieved for each physician into a single unified distance metric. In addition to learning a globally optimized metric, Comdi presents a technique to share knowledge of expertise without sharing private data. To achieve this, it provides the neighborhood information for each party to integrate them into a globally consistent metric. Breast cancer and Pima Indian diabetes [Lichman 2013] datasets are used to evaluate Comdi. The results show Comdi's leverage in comparison with other individual and shared methods such as PCA, LDA, and Locally Supervised Metric Learning (LSML).

Wang et al. [2014] present a Similarity Network Fusion (SNF) to integrate data samples (e.g., patients) and construct patient networks. The patient similarity network is demonstrated as a graph where nodes correspond to patients and edges correspond to the similarity weight between patients. To determine the weight of each edge, a scaled exponential similarity kernel using Euclidean distance is applied. In this article, three data types, DNA methylation, mRNA expression, and microRNA, for five cancer datasets are combined by SNF to compute and fuse patient similarity.

Information visualization of big data is a novel and important process in data mining, known as visual data mining or visual data exploration [Keim 2002]. Tsymbal et al. [2007] propose and compare three techniques for visualizing patient similarity including treemaps [Shneiderman 1992], relative neighborhood graphs [Toussaint 1980], and combined distance-heat maps [Verhaak et al. 2006], which they believe is the most promising approach in clinical workflow.

Several research works have been done recently to bring big data to personalized healthcare, specifically patient similarity. The CARE system [Chawla and Davis 2013], for instance, is developed to predict and manage patient-centered disease. CARE is a computational assistant for doctors and physicians to assess the potential disease risks of patients. This work utilizes shared experiences and similarities among a large number of patients, resulting in a personalized healthcare plan. This big data includes patient history, prognosis, treatment strategies, disease timing, disease progression, and so on.

3.6.2. Computer-Assisted Interventions. More and more computer-based tools and methodologies are developed to support medical interventions especially in the big data era. This particular field of research and practice is called Computer-Assisted Interventions (CAIs) [Johnston et al. 1994]. Examples include image processing methods, surgical process modeling and analysis, and intraoperative decision supports.

With the development of EHR and big data analytics, algorithms and approaches for healthcare big data are proposed to mine multimodal medical data consisting of imaging and textual information. To be specific, modern, scalable, and efficient algorithms are generalized to harvest, organize, and learn from large-scale healthcare datasets for automatic understanding of medical images and help in the decision-making process [Menze et al. 2014]. Schlegl et al. [2014] are able to learn from data collected across multiple hospitals with heterogeneous medical imaging equipment and propose a semisupervised approach based on convolutional neural networks to classify lung tissues with promising results. In other works [del Toro and Müller 2014], a hierarchic multiatlas-based segmentation approach is proposed and evaluated on a large-scale medical dataset for the segmentation of multiple anatomical structures in computed tomography scans.

Recently, the field of analyses and modeling of the Surgical Process (SP) has gained popularity for obtaining an explicit and formal understanding of surgery [Neumuth et al. 2009]. SP models are usually described from observer-based acquisition [Jannin and Morandi 2007] or sensor-based acquisition [Lalys et al. 2012; Nara et al. 2011; Bhatia et al. 2007]. By introducing related surgical models into a new generation of CAI systems, it improves the management of complex multimodal information, increasing the quality and efficiency of medical care.

The developments of CAI and big data technologies interact with each other and continue to assist humans in processing and acting on complex information and providing better services to patients.

4. FUTURE DIRECTIONS AND OUTLOOK

Despite many opportunities and approaches for big data analytics in healthcare presented in this work, there are many other directions to be explored, concerning various aspects of healthcare data, such as the quality, privacy, timeliness, and so forth. This section provides an outlook of big data analytics in healthcare informatics from a broader view, which covers the topics of healthcare data characteristics (e.g., high complexity, large scale, etc.), data analytics tasks (e.g., longitudinal analysis, visualization, etc.), and objectives (e.g., real-time, privacy protection, collaboration with experts, etc.).

4.1. Complexity and Noise

The multisource and multimodal nature of healthcare data results in high complexity and noise issues. In addition, there are also problems of impurity and missing values in the high-volume data. It is difficult to handle all these problems both in terms of scale and accuracy, although a number of methods have been developed to improve the accuracy and usability of data [Müller and Freytag 2003]. Since the quality of data determines the quality of information, which will eventually affect the decision-making process, it is critical to develop efficient big data cleansing approaches to improve data quality for making effective and accurate decisions [Holzinger and Simonic 2011].

4.2. Heterogeneity

Traditional healthcare data usually lacks standardization, often being fragmented with multiple formats [Raghupathi and Raghupathi 2014]. Therefore, it is reasonable and critical to study and develop common data standards. However, it is a challenging task due to the complexity of generating common data standards. Not only is healthcare data diverse, but also there are various technical issues for integrating those data for special usage [Richesson and Krischer 2007]. Even with standardized data formats, the multimodal nature of data creates a challenge for effective fusion [Kambatla et al. 2014], which requires the development of advanced analytics that deal with large amounts of multimodal data. The integration and fusion of the multisource and multimodal healthcare data with increasing scale would be a great challenge [Holzinger et al. 2014a].

4.3. Longitudinal Analysis

Longitudinal data refers to the collection of repeated measurements of participant outcomes and possibly treatments or exposures [Fitzmaurice et al. 2008]; that is, “the outcome variable is repeatedly measured on the same individual on multiple occasions” [Twisk 2004]. In recent decades, longitudinal data analysis, especially the statistical analysis of longitudinal data, has attracted more and more attention. Longitudinal studies involve the characterization of normal growth and aging, and the effectiveness of the assessment of risk factors and treatments. It plays a key role in epidemiology, clinical research, and therapeutic evaluation. With big data analytic tools, it becomes promising to apply the longitudinal analysis of care across patients and diagnoses for identifying the best care approaches.

4.4. Scale

Healthcare data is rapidly growing by size and coverage [Kambatla et al. 2014]. The fact that the data volume is scaling faster than computing resources poses a major challenge in managing large amounts of data. Several fundamental shifts (from a hardware point of view) are taking place to accommodate this dramatic change [Labrinidis and Jagadish 2012]. First, over the past few years, the processor technology has gradually shifted the focus to parallel data processing within nodes and the packing of multiple sockets. Second, the move toward cloud computing enables information sharing and the aggregation of multiple workloads into large-scale clusters. Third, the transformative change of the traditional I/O subsystem from Hard Disk Drives (HDDs) to Solid-State Drives (SSDs), as well as other storage technologies, is reforming the design and operation of data processing systems.

4.5. Real Time

The velocity characteristic of big data in health informatics not only indicates the data acquisition and processing rate but also the timeliness of responses. There are many scenarios that call for a quick decision. For example, it would be extremely desirable to monitor and analyze a person's health condition to predict potential illness in real time or near real time. It would also be of great significance to raise an alarm for a potential outbreak of influenza through analyzing public health data. Although real-time analytic applications are still in their infancy in the big data era, they are the strongest trend and most promising direction in the future of health informatics [Russom 2011]. A good example is the development of Complex Event Processing (CEP) for handling streaming big data and fulfilling real-time requirements.

4.6. Privacy

The privacy of data is another big concern of future big data analytics in healthcare informatics [Weippl et al. 2006]. Although there are strict laws governing the more formalized EHR data, special attention should be paid and rules should be enforced to regularize the usage and distribution of personal and sensitive information acquired from multiple sources. "Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data" [Labrinidis and Jagadish 2012]. In addition to data privacy, there are a range of other issues, such as data protection, data security, data safety, and protection of doctors against responsibility derived from manipulated data, that require special big data analytics to handle these complex restrictions [Weippl et al. 2006; Holzinger et al. 2014a; Kieseberg et al. 2015].

4.7. Visualization

Visualization of healthcare data is critical for exploratory or discovery analytics, whose purpose is to explore and discover things that are undermined and encrypted in the data [Wong et al. 2011; Jeanquartier and Holzinger 2013]. Effective visualization tools will help clinicians and physicians to explore the data without the assistance from IT [Russom 2011]. Although visualization has been studied for several decades with relative maturity, there are still challenges and open issues to be addressed, especially for the big data analytics in healthcare data [Holzinger et al. 2014a].

4.8. Multidiscipline and Human Interaction

Big data in health informatics is predicted to be a multidisciplinary task that involves continuous efforts from multiple domain experts [Chen et al. 2014]. They include, but are not limited to, engineering scientists who provide basic big data infrastructure to collect, store, share, and manage big data; computer science data scientists who provide solutions for processing and analyzing high-volume, high-velocity healthcare data via numerous data mining and machine-learning techniques; clinicians and physicians from the medical domain who provide professional healthcare data analysis, offer personalized care, and make final decisions. Sometimes it is difficult for computer algorithms to identify patterns and analyze results; therefore, it is a desirable feature for an advanced big data analysis system to be able to support input from multiple human experts, exchange of opinions, and shared exploration of results. Furthermore, in the health domain, sometimes we do not have big data: we are confronted with a small number of datasets or rare events, where, for example, machine-learning approaches suffer from insufficient training samples. In such cases, we need more than just automatic machine learning; we need still a human in the loop. In other words, interactive Machine Learning (iML) or “human in the loop” techniques can be utilized in health informatics where automatic machine-learning approaches are not able to handle rare events alone and a human expert is needed to interact in the learning process [Holzinger 2016]. This interaction between computer algorithms and human experts can improve the learning procedure.

5. SUMMARY

This article presents a comprehensive overview of the challenges, pipeline, techniques, and future directions for computational health informatics in the big data age, by providing a structured analysis of the historical and state-of-the-art methods in over 170 papers and web articles. We have summarized the challenges of big data health informatics into four Vs: volume, variety, velocity, and veracity, and emerging challenges such as validity and volatility. A systematic data processing pipeline is provided for generic big health informatics, covering data capturing, storing, sharing, analyzing, searching, and decision support. Computational health informatics in the big data age is an emerging and highly important research field with a potentially significant impact on the conventional healthcare industry. The future of health informatics will benefit from the exponentially increasing digital health data.

REFERENCES

ADNI. 2015. Alzheimer’s disease neuroimaging initiative. http://adni.loni.usc.edu/about/. (2015). Retrieved 02-15-2015-02.

Hatice Cinar Akakin and Metin N. Gurcan. 2012. Content-based microscopic image retrieval system for multi-image queries. *IEEE Transactions on Information Technology in Biomedicine* 16, 4 (2012), 758–769.

Alexandr Andoni and Piotr Indyk. 2006. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS’06)*. 459–468.

Bill Andreopoulos, Aijun An, and Xiaogang Wang. 2007. Hierarchical density-based clustering of categorical data and a simplification. In *Advances in Knowledge Discovery and Data Mining*. Springer LNCS, 11–22.

Daniele Apiletti, Elena Baralis, Giulia Bruno, and Tania Cerquitelli. 2009. Real-time analysis of physiological data to support medical applications. *IEEE Transactions on Information Technology in Biomedicine* 13, 3 (2009), 313–321.

Peter C. Austin, Jack V. Tu, Jennifer E. Ho, Daniel Levy, and Douglas S. Lee. 2013. Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes. *Journal of Clinical Epidemiology* 66, 4 (2013), 398–407.

Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. 2013. Data mining for wearable sensors in health monitoring systems: A review of recent trends and challenges. *Sensors* 13, 12 (2013), 17472–17500.

Andrew Bate, Marie Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. Melhado De Freitas. 1998. A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology* 54, 4 (1998), 315–321.

David W. Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. 2014. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs* 33, 7 (2014), 1123–1131.

Smaranda Belciug. 2009. Patients length of stay grouping using the hierarchical clustering algorithm. *Annals of the University of Craiova-Mathematics and Computer Science Series* 36, 2 (2009), 79–84.

Irad Ben-Gal. 2007. Bayesian networks. *Encyclopedia of Statistics in Quality and Reliability.* John Wiley and Sons.

Beenish Bhatia, Tim Oates, Yan Xiao, and Peter Hu. 2007. Real-time identification of operating room state from video. In *Association for the Advancement of Artificial Intelligence (AAAI)*. Vol. 2. 1761–1766.

Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32. DOI:http://dx.doi.org/10.1023/A:1010933404324

Nanette Byrnes. 2014. MIT technology review. http://www.technologyreview.com/news/529011/can-technology-fix-medicine/. (2014). Retrieved 01-20-2015.

Mehemmed Emre Celebi, Yuksel Alp Aslandogan, and Paul R. Bergstresser. 2005. Mining biomedical images with density-based clustering. In *International Conference on Information Technology: Coding and Computing (ITCC'05)*. Vol. 1. IEEE, 163–168.

Nitesh V. Chawla and Darcy A. Davis. 2013. Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine* 28, 3 (2013), 660–665.

Min Chen, Shiwen Mao, and Yunhao Liu. 2014. Big data: A survey. *Mobile Networks and Applications* 19, 2 (2014), 171–209.

Xue-Wen Chen and Xiaotong Lin. 2014. Big data deep learning: Challenges and perspectives. *IEEE Access* 2 (2014), 514–525.

Hugh Chipman and Robert Tibshirani. 2006. Hybrid hierarchical clustering with applications to microarray data. *Biostatistics* 7, 2 (2006), 286–301.

Paul D. Clayton and George Hripcsak. 1995. Decision support in healthcare. *International Journal of Bio-Medical Computing* 39, 1 (1995), 59–66.

Dorin Comaniciu, Peter Meer, and David J. Foran. 1999. Image-guided decision support system for pathology. *Machine Vision and Applications* 11, 4 (1999), 213–224.

Robert E. Cooke Jr., Michael G. Gaeta, Dean M. Kaufman, and John G. Henrici. 2003. Picture archiving and communication system. (June 3, 2003). US Patent 6,574,629.

Ben Cooper and Marc Lipsitch. 2004. The analysis of hospital infection data using hidden Markov models. *Biostatistics* 5, 2 (2004), 223–237.

James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, J. J. Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford. 2013. Spanner: Googles globally distributed database. *ACM Transactions on Computer Systems (TOCS)* 31, 3 (2013), 8.

Fabricio F. Costa. 2014. Big data in biomedicine. *Drug Discovery Today* 19, 4 (2014), 433–440.

Carl W. Cotman, Nicole C. Berchtold, and Lori-Ann Christie. 2007. Exercise builds brain health: Key roles of growth factor cascades and inflammation. *Trends in Neurosciences* 30, 9 (2007), 464–472.

Mike Cottle, Waco Hoover, Shadaab Kanwal, Marty Kohn, Trevor Strome, and Neil W. Treister. 2013. Transforming health care through big data: Strategies for leveraging big data in the health care industry. http://c4fd63cb482ce6861463-bc6183f1c18e748a49b87a25911a0555.r93.cf2.rackcdn.com/iHT2_BigData_2013.pdf. (2013). New York: Institute for Health Technology Transformation.

Jared Crapo. 2014. Big data in healthcare: Separating the hype from the reality. https://www.healthcatalyst.com/healthcare-big-data-realities. (2014). Retrieved 01-17-2015.

Antonio Criminisi, Duncan Robertson, Ender Konukoglu, Jamie Shotton, Sayan Pathak, Steve White, and Khan Siddiqui. 2013. Regression forests for efficient anatomy detection and localization in computed tomography scans. *Medical Image Analysis* 17, 8 (2013), 1293–1303.

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press.

Andy Crowne. 2014. Preparing the healthcare industry to capture the full potential of big data. http://sparkblog.emc.com/2014/06/preparing-healthcare-industry-capture-full-potential-big-data/. (2014). Retrieved 02-25-2015.

Oscar Alfonso Jiménez del Toro and Henning Müller. 2014. Hierarchic multi–atlas based segmentation for anatomical structures: Evaluation in the VISCERAL anatomy benchmarks. In *Medical Computer Vision: Algorithms for Big Data*. Springer, 189–200.

Ramón Díaz-Uriarte and Sara Alvarez De Andres. 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7, 1 (2006), 3.

David J. Dittman, Taghi M. Khoshgoftaar, Randall Wald, and Amri Napolitano. 2013. Simplifying the utilization of machine learning techniques for bioinformatics. In *12th International Conference on Machine Learning and Applications (ICMLA'13)*. Vol. 2. IEEE, 396–403.

Pedro Domingos and Geoff Hulten. 2000. Mining high-speed data streams. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 71–80.

Xinhua Dong, Ruixuan Li, Heng He, Wanwan Zhou, Zhengyuan Xue, and Hao Wu. 2015. Secure sensitive data sharing on a big data platform. *Tsinghua Science and Technology* 20, 1 (2015), 72–80.

Arun George Eapen. 2004. *Application of Data Mining in Medical Applications*. Master's thesis. University of Waterloo, Ontario, Canada.

EMC. 2011. Managing healthcare data within the ecosystem while reducing IT costs and complexity. http://www.emc.com/collateral/emc-perspective/h8805-healthcare-costs-co mplexities-ep.pdf. (2011). Retrieved 02-25-2015.

Francisco Estella, Blanca L. Delgado-Marquez, Pablo Rojas, Olga Valenzuela, Belen San Roman, and Ignacio Rojas. 2012. Advanced system for autonomously classify brain MRI in neurodegenerative disease. In *International Conference on Multimedia Computing and Systems (ICMCS'12)*. IEEE, 250–255.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. Vol. 96. AAAI Press, 226–231.

Ruogu Fang, Tsuhan Chen, and Pina C. Sanelli. 2013. Towards robust deconvolution of low-dose perfusion CT: Sparse perfusion deconvolution using online dictionary learning. *Medical Image Analysis* 17, 4 (2013), 417–428.

Ruogu Fang, Haodi Jiang, and Junzhou Huang. 2015. Tissue-specific sparse deconvolution for brain CT perfusion. *Computerized Medical Imaging and Graphics* (2015). Available online May 21, 2015, ISSN 0895-6111, http://dx.doi.org/10.1016/j.compmedimag.2015.04.008.

Ruogu Fang, Kolbeinn Karlsson, Tsuhan Chen, and Pina C. Sanelli. 2014. Improving low-dose blood–brain barrier permeability quantification using sparse high-dose induced prior for Patlak model. *Medical Image Analysis* 18, 6 (2014), 866–880.

Ruogu Fang, Shaoting Zhang, Tsuhan Chen, and Pina Sanelli. 2015. Robust low-dose CT perfusion deconvolution via tensor total-variation regularization. *IEEE Transaction on Medical Imaging* 34, 7 (2015), 1533–1548.

Anthony S. Fauci, Dennis L. Kasper, Eugene Braunwald, Stephen L. Hauser, Dan L. Longo, J. Larry Jameson, and Joseph Loscalzo. 2008. *Harrison's Principles of Internal Medicine*. Vol. 2. New York: McGraw-Hill Medical.

Bonnie Feldman, Ellen M. Martin, and Tobi Skotnes. 2012. Big data in healthcare hype and hope. *Technical Report, Dr. Bonnie* 360 (2012).

André S. Fialho, Federico Cismondi, Susana M. Vieira, Shane R. Reti, Joao M. C. Sousa, and Stan N. Finkelstein. 2012. Data mining using clinical physiology at discharge to predict ICU readmissions. *Expert Systems with Applications* 39, 18 (2012), 13158–13165.

Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs. 2008. *Longitudinal Data Analysis*. CRC Press. Handbooks of Modern Statistical Methods. New York: Chapman and Hall.

Christos A. Frantzidis, Charalampos Bratsas, Manousos A. Klados, Evdokimos Konstantinidis, Chrysa D. Lithari, Ana B. Vivas, Christos L. Papadelis, Eleni Kaldoudi, Costas Pappas, and Panagiotis D. Bamidis. 2010. On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine* 14, 2 (2010), 309–318.

Yoav Freund and Llew Mason. 1999. The alternating decision tree learning algorithm. In *Proceedings of the 16th International Conference on Machine Learning*. Vol. 99. 124–133.

Bernd Fritzke. 1995. A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems* 7 (1995), 625–632.

Gartner. 2014. IT glossary: Big data. http://www.gartner.com/it-glossary/big-data/. (2014). Retrieved 03-06-2015.

Zoubin Ghahramani. 2001. An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence* 15, 01 (2001), 9–42.

Ali Gholipour, Judy A. Estroff, and Simon K. Warfield. 2010. Robust super-resolution volume reconstruction from slice acquisitions: Application to fetal brain MRI. *IEEE Transactions on Medical Imaging,* 29, 10 (2010), 1739–1758.

Donna Giri, U. Rajendra Acharya, Roshan Joy Martis, S. Vinitha Sree, Teik-Cheng Lim, Thajudin Ahamed, and Jasjit S. Suri. 2013. Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. *Knowledge-Based Systems* 37 (2013), 274–282.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11, 1 (2009), 10–18.

Mark A. Hall. 1999. *Correlation-Based Feature Selection for Machine Learning*. Ph.D. Dissertation. University of Waikato.

Allan Hanbury, Henning Müller, Georg Langs, and Bjoern H. Menze. 2013. Cloud-based evaluation framework for big data. In *FIA Book 2013 (Springer LNCS)*.

Douglas M. Hawkins. 2004. The problem of overfitting. *Journal of Chemical Information and Computer Sciences* 44, 1 (2004), 1–12.

Chenguang He, Xiaomao Fan, and Ye Li. 2013. Toward ubiquitous healthcare services with a novel efficient cloud platform. *IEEE Transactions on Biomedical Engineering* 60, 1 (2013), 230–234.

Matthew Herland, Taghi M. Khoshgoftaar, and Randall Wald. 2013. Survey of clinical data mining applications on big data in health informatics. In *12th International Conference on Machine Learning and Applications (ICMLA'13)*. Vol. 2. IEEE, 465–472.

Matthew Herland, Taghi M. Khoshgoftaar, and Randall Wald. 2014. A review of data mining using big data in health informatics. *Journal of Big Data, Springer* 1, 1 (2014), 2.

Geoffrey E. Hinton. 2009. Deep belief networks. *Scholarpedia* 4, 5 (2009), 5947.

Joyce C. Ho, Joydeep Ghosh, and Jimeng Sun. 2014. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 115–124.

Andreas Holzinger. 2016. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics* 3 (2016), 1–13. DOI:http://dx.doi.org/10.1007/s40708-016-0042-6

Andreas Holzinger, Matthias Dehmer, and Igor Jurisica. 2014a. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC Bioinformatics* 15, Suppl 6 (2014), I1.

Andreas Holzinger, Johannes Schantl, Miriam Schroettner, Christin Seifert, and Karin Verspoor. 2014b. Biomedical text mining: State-of-the-art, open problems and future challenges. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Springer, 271–300.

Andreas Holzinger and Klaus-Martin Simonic. 2011. Information quality in e-health. In *Proceedings of the 7th Conference of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society (USAB'11)*. Vol. 7058. Springer, Graz, Austria. DOI:http://dx.doi.org/10.1007/978-3-642-25364-5

HSCIC. 2012. Hospital episode statistics. http://www.hscic.gov.uk/hes. (2012). Retrieved 02-20-2015.

Fei Hu, Meng Jiang, Laura Celentano, and Yang Xiao. 2008. Robust medical ad hoc sensor networks (MASN) with wavelet-based ECG data mining. *Ad Hoc Networks* 6, 7 (2008), 986–1012.

Hui Fang Huang, Guang Shu Hu, and Li Zhu. 2012. Sparse representation-based heartbeat classification using independent component analysis. *Journal of Medical Systems* 36, 3 (2012), 1235–1247.

Ke Huang and Selin Aviyente. 2006. Sparse representation for signal classification. In *Advances in Neural Information Processing Systems*. 609–616.

Michael Hund, Werner Sturm, Tobias Schreck, Torsten Ullrich, Daniel Keim, Ljiljana Majnaric, and Andreas Holzinger. 2015. Analysis of patient groups and immunization results based on subspace clustering. In *Brain Informatics and Health*. Springer, 358–368. DOI:http://dx.doi.org/10.1007/978-3-319-23344-4_35

Kevin Hung, Yuan-Ting Zhang, and B. Tai. 2004. Wearable medical devices for tele-home healthcare. In *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEMBS'04).* Vol. 2. IEEE, 5384–5387.

IBM. 2012. Large gene interaction analytics at University at Buffalo. http://www-03.ibm.com/software/businesscasestudies/no/no/corp?synkey=M947744T51514R54. (2012). Retrieved 02-23-2015.

IBM. 2015a. IBM content and predictive analytics for healthcare. http://www-01.ibm.com/software/sg/industry/healthcare/pdf/setonCaseStudy.pdf. (2015). Retrieved 01-20-2015.

IBM. 2015b. IBM patient care and insights. http://www-03.ibm.com/software/products/en/IBM-care-management. (2015). Retrieved 03-05-2015.

Intel. 2011. Distributed systems for clinical data analysis. http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-hadoop-clinical-analysis-paper.pdf. (2011). Retrieved 02-24-2015.

Naiem T. Issa, Stephen W. Byers, and Sivanesan Dakshanamurthy. 2014. Big data: The next frontier for innovation in therapeutics and healthcare. *Expert Review of Clinical Pharmacology* 7, 3 (2014), 293–298.

Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys (CSUR)* 31, 3 (1999), 264–323.

Raimon Jané, Hervé Rix, Pere Caminal, and Pablo Laguna. 1991. Alignment methods for averaging of high-resolution cardiac signals: A comparative study of performance. *IEEE Transactions on Biomedical Engineering* 38, 6 (1991), 571–579.

Pierre Jannin and Xavier Morandi. 2007. Surgical models for computer-assisted neurosurgery. *Neuroimage* 37, 3 (2007), 783–791.

Fleur Jeanquartier and Andreas Holzinger. 2013. On visual analytics and evaluation in cell physiology: A case study. In *Availability, Reliability, and Security in Information Systems and HCI*. Springer, 495–502.

Menglin Jiang, Shaoting Zhang, Hongsheng Li, and Dimitris N. Metaxas. 2015. Computer-aided diagnosis of mammographic masses using scalable image retrieval. *IEEE Transactions on Biomedical Engineering* 62, 2 (2015), 783–792.

Mary E. Johnston, Karl B. Langton, R. Brian Haynes, and Alix Mathieu. 1994. Effects of computer-based clinical decision support systems on clinician performance and patient outcome: A critical appraisal of research. *Annals of Internal Medicine* 120, 2 (1994), 135–142.

Kenneth Jung, Paea LePendu, Srinivasan Iyer, Anna Bauer-Mehren, Bethany Percha, and Nigam H. Shah. 2014. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. *Journal of the American Medical Informatics Association* 22, 1 (2014), 121–131.

Karthik Kambatla, Giorgos Kollias, Vipin Kumar, and Ananth Grama. 2014. Trends in big data analytics. *Journal of Parallel and Distributed Computing* 74, 7 (2014), 2561–2573.

Kensaku Kawamoto, Caitlin A. Houlihan, E. Andrew Balas, and David F. Lobach. 2005. Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *BMJ* 330, 7494 (2005), 765.

Daniel A. Keim. 2002. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (2002), 1–8.

Irfan Y. Khan, P. H. Zope, and S. R. Suralkar. 2013. Importance of artificial neural network in medical diagnosis disease like acute nephritis disease and heart disease. *International Journal of Engineering Science and Innovative Technology (IJESIT)* 2, 2 (2013), 210–217.

Peter Kieseberg, Johannes Schantl, Peter Frühwirt, Edgar Weippl, and Andreas Holzinger. 2015. Witnesses for the doctor in the loop. In *Brain Informatics and Health*. Springer, 369–378. DOI:http://dx.doi.org/10.1007/978-3-319-23344-4_36

Teuvo Kohonen. 1998. The self-organizing map. *Neurocomputing* 21, 1 (1998), 1–6.

Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3, 1 (2009), 1–58. DOI:http://dx.doi.org/10.1145/1497577.1497578

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, 1097–1105.

Alexandros Labrinidis and H. V. Jagadish. 2012. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment* 5, 12 (2012), 2032–2033.

Florent Lalys, Laurent Riffaud, David Bouget, and Pierre Jannin. 2012. A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. *IEEE Transactions on Biomedical Engineering* 59, 4 (2012), 966–976.

Georg Langs, Allan Hanbury, Bjoern Menze, and Henning Müller. 2013. VISCERAL: Towards large data in medical imaging Challenges and directions. In *Medical Content-Based Retrieval for Clinical Decision Support*. Vol. 7723. Springer, 92–98.

Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. 2009. Exploring strategies for training deep neural networks. *Journal of Machine Learning Research* 10 (2009), 1–40.

Kim-Anh Lê Cao, Debra Rossouw, Christèle Robert-Granié, and Philippe Besse. 2008. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology* 7, 1 (2008), 1544–6115.

Anny Leema and M. Hemalatha. 2011. An effective and adaptive data cleaning technique for colossal RFID data sets in healthcare. *WSEAS Transactions on Information Science and Applications* 8, 6 (2011), 243–252.

Hai Guang Li, Xindong Wu, Zhao Li, and Wei Ding. 2013. Online group feature selection from feature streams. In *27th AAAI Conference on Artificial Intelligence*. Citeseer, 1627–1628.

Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. 2014. Deep learning based imaging data completion for improved brain disease diagnosis. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'14)*. Springer, 305–312.

Shutao Li, Haitao Yin, and Leyuan Fang. 2012. Group-sparse representation with dictionary learning for medical image denoising and fusion. *IEEE Transactions on Biomedical Engineering* 59, 12 (2012), 3450–3459.

Znaonui Liang, Gang Zhang, Jimmy Xiangji Huang, and Qmming Vivian Hu. 2014. Deep learning for healthcare decision making with EMRs. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'14)*. 556–559.

Moshe Lichman. 2013. UCI machine learning repository. (2013). http://archive.ics.uci.edu/ml. Retrieved 08-03-2015.

Manhua Liu, Daoqiang Zhang, and Dinggang Shen. 2012. Ensemble sparse classification of Alzheimer's disease. *NeuroImage* 60, 2 (2012), 1106–1116.

Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. 2012. Supervised hashing with kernels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. 2074–2081.

Wei Liu, Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. 2011. Hashing with graphs. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. 1–8.

Aastha Madaan, Wanming Chu, Yaginuma Daigo, and Subhash Bhalla. 2013. Quasi-relational query language interface for persistent standardized EHRs: Using NoSQL databases. In *Databases in Networked Information Systems*. Springer, 182–196.

James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H. Byers. 2011. Big data: The next frontier for innovation, competition, and productivity. *Technical Report, McKinsey Global Institute* (2011).

Kezhi Z. Mao. 2004. Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 34, 1 (2004), 629–634.

Yi Mao, Wenlin Chen, Yixin Chen, Chenyang Lu, Marin Kollef, and Thomas Bailey. 2012. An integrated data mining approach to real-time clinical monitoring and deterioration warning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1140–1148.

Ramon Martinez Orellana, Burak Erem, and Dana H. Brooks. 2013. Time invariant multi electrode averaging for biomedical signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'13)*. IEEE, 1242–1246.

Jason Scott Mathias, Ankit Agrawal, Joe Feinglass, Andrew J. Cooper, David William Baker, and Alok Choudhary. 2013. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. *Journal of the American Medical Informatics Association* 20, e1 (2013), e118–e124.

Tao Meng, Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen. 2010. Histology image classification using supervised classification and multimodal fusion. In *2010 IEEE International Symposium on Multimedia (ISM'10)*. 145–152.

Tao Meng and Mei-Ling Shyu. 2013. Biological image temporal stage classification via multi-layer model collaboration. In *2013 IEEE International Symposium on Multimedia (ISM'13)*. 30–37.

Tao Meng, Ahmed T. Soliman, Mei-Ling Shyu, Yimin Yang, Shu-Ching Chen, S. S. Iyengar, John S. Yordy, and Puneeth Iyengar. 2013. Wavelet analysis in current cancer genome research: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10, 6 (2013), 1442–14359.

Bjoern Menze, Georg Langs, Albert Montillo, Michael Kelm, Henning Müller, Shaoting Zhang, Weidong Tom Cai, and Dimitris Metaxas. 2014. *Medical Computer Vision: Algorithms for Big Data: International Workshop (MCV'14), held in Conjunction with MICCAI'14, Cambridge, MA, USA, September 18, 2014, Revised Selected Papers*. Vol. 8848. Springer.

Ivan Merelli, Horacio Pérez-Sánchez, Sandra Gesing, and Daniele DAgostino. 2014. Managing, analysing, and integrating big data in medical bioinformatics: Open problems and future perspectives. *BioMed Research International* 2014 (2014), 13. DOI: http://dx.doi.org/10.1155/2014/134023

Tom M. Mitchell. 1997. *Machine Learning*. Burr Ridge, IL: McGraw Hill.

Heiko Müller and Johann-Christph Freytag. 2003. *Problems, Methods, and Challenges in Comprehensive Data Cleansing*. Technical Report. Humboldt-Universitt zu Berlin. Professoren des Inst. Für Informatik.

Henning Müller, Antoine Geissbühler, and Patrick Ruch. 2005. ImageCLEF 2004: Combining image and multi-lingual search for medical image retrieval. In *Multilingual Information Access for Text, Speech and Images*. Vol. 3491. Springer, 718–727.

Atsushi Nara, Kiyoshi Izumi, Hiroshi Iseki, Takashi Suzuki, Kyojiro Nambu, and Yasuo Sakurai. 2011. Surgical workflow monitoring based on trajectory data mining. In *New Frontiers in Artificial Intelligence*. Vol. 6797. Springer, 283–291.

NCCHD. 2014. Births: Data from the National Community Child Health database. http://gov.wales/statistics-and-research/births-national-community-child-health-database/?lang=en. (2014). Retrieved 02-15-2015.

NetApp. 2011a. http://www.netapp.com/us/solutions/industry/healthcare/. (2011). Retrieved 02-24-2015.

NetApp. 2011b. NetApp EHR solutions: Efficient, high-availability EHR data storage and management. http://www.netapp.com/us/system/pdf-reader.aspx?cc=us&m=ds-3222.pdf&pdfUri=tcm:10-61401. (2011). Retrieved 02-20-2015.

Thomas Neumuth, Pierre Jannin, Gero Strauss, Juergen Meixensberger, and Oliver Burgert. 2009. Validation of knowledge acquisition for surgical process models. *Journal of the American Medical Informatics Association* 16, 1 (2009), 72–80.

Michael Nielsen. 2014. Neural networks and deep learning. *Determination Press*. Vol. 1.

David Nister and Henrik Stewenius. 2006. Scalable recognition with a vocabulary tree. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2. IEEE, 2161–2168.

N. Nithya, K. Duraiswamy, and P. Gomathy. 2013. A survey on clustering techniques in medical diagnosis. *International Journal of Computer Science Trends and Technology (IJCST)* 1, 2 (2013), 17–23.

Sankar K. Pal and Sushmita Mitra. 1992. Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on Neural Networks* 3, 5 (1992), 683–697.

Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden, and Michael Stonebraker. 2009. A comparison of approaches to large-scale data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*. 165–178.

Mykola Pechenizkiy, Alexey Tsymbal, and Seppo Puuronen. 2004. PCA-based feature transformation for classification: Issues in medical diagnostics. In *Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04)*. IEEE, 535–540.

Ken Perez. 2013. MedeAnalytics. http://www.thefreelibrary.com/MedeAnalytics%27+Ken+Perez+Predicts+Rapid+Growth+of+Healthcare+Data...-a0328478771/. (2013). Retrieved 03-02-2015.

Sergey M. Plis, Devon R. Hjelm, Ruslan Salakhutdinov, Elena A. Allen, Henry J. Bockholt, Jeffrey D. Long, Hans J. Johnson, Jane S. Paulsen, Jessica A. Turner, and Vince D. Calhoun. 2014. Deep learning for neuroimaging: A validation study. *Frontiers in Neuroscience* 8 (2014).

Fred Popowich. 2005. Using text mining and natural language processing for health care claims processing. *ACM SIGKDD Explorations Newsletter* 7, 1 (2005), 59–66.

PREDICT-HD. 2015. PREDICT-HD project. https://www.predict-hd.net/. (2015). Retrieved 04-10-2015.

K. Priyanka and Nagarathna Kulennavar. 2014. A survey on big data analytics in health care. *International Journal of Computer Science and Information Technologies* 5, 4 (2014), 5865–5868.

Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems* 2, 1 (2014), 3.

K. Usha Rani. 2011. Analysis of heart diseases dataset using neural network approach. *arXiv preprint arXiv:1110.2626* (2011).

Daniel A. Reed and Jack Dongarra. 2015. Exascale computing and big data. *Communications of the ACM* 58, 7 (2015), 56–68.

Rachel L. Richesson and Jeffrey Krischer. 2007. Data standards in clinical research: Gaps, overlaps, challenges and future directions. *Journal of the American Medical Informatics Association* 14, 6 (2007), 687–696.

Juan José Rodriguez, Ludmila I. Kuncheva, and Carlos J. Alonso. 2006. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 10 (2006), 1619–1630.

Roy. 2015. Roy tutorials. http://www.roytuts.com/big-data/. (2015). Retrieved 01-20-2015.

Philip Russom. 2011. Big data analytics. *TDWI Best Practices Report, Fourth Quarter* (2011).

Bhaskar Saha, Kai Goebel, Scott Poll, and Jon Christophersen. 2007. An integrated approach to battery health monitoring using Bayesian regression and state estimation. In *2007 IEEE Autotestcon*. 646–653.

Ruslan Salakhutdinov and Geoffrey E. Hinton. 2009. Deep Boltzmann machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS'09)*. Vol. 5. 448–455.

Gerard Salton and Donna Harman. 2003. Information retrieval. In *Encyclopedia of Computer Science*. John Wiley and Sons, Chichester, UK, 858–863.

Thomas Schlegl, Joachim Ofner, and Georg Langs. 2014. Unsupervised pre-training across image domains improves lung tissue classification. In *Medical Computer Vision: Algorithms for Big Data*. Vol. 8848. Springer, 82–93.

Frank Schnorrenberg, Constantinos S. Pattichis, Christos N. Schizas, and Kyriacos Kyriacou. 2000. Content-based retrieval of breast cancer biopsy slides. *Technology and Health Care* 8, 5 (2000), 291–297.

Ben Shneiderman. 1992. Tree visualization with tree-maps: 2-D space-filling approach. *ACM Transactions on Graphics (TOG)* 11, 1 (1992), 92–99.

Rajiv Ranjan Singh, Sailesh Conjeti, and Rahul Banerjee. 2011. An approach for real-time stress-trend detection using physiological signals in wearable computing systems for automotive drivers. In *14th International IEEE Conference on Intelligent Transportation Systems (ITSC'11)*. 1477–1482.

Sanjay K. Singh, Adeel Malik, Ahmad Firoz, and Vivekanand Jha. 2012. CDKD: A clinical database of kidney diseases. *BMC Nephrology* 13, 1 (2012), 23.

Ahmed T. Soliman, Tao Meng, Shu-Ching Chen, S. S. Iyengar, Puneeth Iyengar, John Yordy, and Mei-Ling Shyu. 2015. Driver missense mutation identification using feature selection and model fusion. *Journal of Computational Biology* 22, 12 (2015), 1075–1085. DOI:http://dx.doi.org/10.1089/cmb.2015.0110

Daby Sow, Deepak S. Turaga, and Michael Schmidt. 2013. Mining of sensor data in healthcare: A survey. In *Managing and Mining Sensor Data*. Springer, 459–504.

Peter Spyns. 1996. Natural language processing in medicine: An overview. *Methods of Information in Medicine* 35, 4 (1996), 285–301.

Abdulhamit Subasi and M. Ismail Gursoy. 2010. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications* 37, 12 (2010), 8659–8666.

Jimeng Sun, Candace D. McNaughton, Ping Zhang, Adam Perer, Aris Gkoulalas-Divanis, Joshua C. Denny, Jacqueline Kirby, Thomas Lasko, Alexander Saip, and Bradley A. Malin. 2014. Predicting changes in hypertension control using electronic health records from a chronic disease management program. *Journal of the American Medical Informatics Association* 21, 2 (2014), 337–344.

Jimeng Sun and Chandan K. Reddy. 2013. Big data analytics for healthcare. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*. ACM, New York, NY, 1525–1525. DOI:http://dx.doi.org/10.1145/2487575.2506178

Mingkui Tan, Ivor W. Tsang, and Li Wang. 2014. Towards ultrahigh dimensional feature selection for big data. *Journal of Machine Learning Research* 15, 1 (2014), 1371–1429.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Vol. 1. Boston: Pearson Addison Wesley.

Techcrunch. 2014. Healthcare's Big Data Opportunity. http://techcrunch.com/2014/11/20/healthcares-big-data-opportunity/. (2014). Retrieved 02-20-2015.

Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. 2009. Hive: A warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment* 2, 2 (2009), 1626–1629.

Andy Tippet. 2014. Data capture and analytics in healthcare. http://blogs.zebra.com/data-capture-and-analytics-in-healthcare. (2014). Retrieved 02-08-2015.

Michael E. Tipping. 2001. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* 1 (2001), 211–244.

Divya Tomar and Sonali Agarwal. 2013. A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology* 5, 5 (2013), 241–266.

Godfried T. Toussaint. 1980. The relative neighbourhood graph of a finite planar set. *Pattern Recognition* 12, 4 (1980), 261–268.

Alexey Tsymbal, Martin Huber, Sonja Zillner, Tamás Hauer, and Shaohua Kevin Zhou. 2007. Visualizing patient similarity in clinical decision support. In *LWA*. Martin-Luther-University Halle-Wittenberg, 304–311.

Jos W. R. Twisk. 2004. Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. *European Journal of Epidemiology* 19, 8 (2004), 769–776.

Roel G. W. Verhaak, Mathijs A. Sanders, Maarten A. Bijl, Ruud Delwel, Sebastiaan Horsman, Michael J. Moorhouse, Peter J. van der Spek, Bob Löwenberg, and Peter J. M. Valk. 2006. HeatMapper: Powerful combined visualization of gene expression profile correlations, genotypes, phenotypes and sample characteristics. *BMC Bioinformatics* 7, 1 (2006), 337.

Rajakrishnan Vijayakrishnan, Steven R. Steinhubl, Kenney Ng, Jimeng Sun, Roy J. Byrd, Zahra Daar, Brent A. Williams, Christopher Defilippi, Shahram Ebadollahi, and Walter F. Stewart. 2014. Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *Journal of Cardiac Failure* 20, 7 (2014), 459–464.

Thi Hong Nhan Vu, Namkyu Park, Yang Koo Lee, Yongmi Lee, Jong Yun Lee, and Keun Ho Ryu. 2010. Online discovery of heart rate variability patterns in mobile healthcare services. *Journal of Systems and Software* 83, 10 (2010), 1930–1940.

Bo Wang, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* 11, 3 (2014), 333–337.

Fei Wang, Jimeng Sun, and Shahram Ebadollahi. 2012. Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 5, 1 (2012), 54–69.

Wei Wang, Honggang Wang, Michael Hempel, Dongming Peng, Hamid Sharif, and Hsiao-Hwa Chen. 2011. Secure stochastic ECG signals based on gaussian mixture model for e-healthcare systems. *IEEE Systems Journal* 5, 4 (2011), 564–573.

Edgar Weippl, Andreas Holzinger, and A. Min Tjoa. 2006. Security aspects of ubiquitous computing in health care. *e & i Elektrotechnik und Informationstechnik* 123, 4 (2006), 156–161.

Yair Weiss, Antonio Torralba, and Rob Fergus. 2009. Spectral hashing. In *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.). Curran Associates, 1753–1760. http://papers.nips.cc/paper/3383-spectral-hashing.pdf.

Susan E. White. 2014. A review of big data in health care: Challenges and opportunities. *Clinical, Cosmetic and Investigational Dentistry* 6 (2014), 45–56.

B. L. William Wong, Kai Xu, and Andreas Holzinger. 2011. Interactive visualization for information analysis in medical diagnosis. In *Information Quality in e-Health. Lecture Notes in Computer Science*. Vol. 7058. Springer, 109–120.

Leon Xiao, Judy Hanover, and Sash Mukherjee. 2014. Big data enables clinical decision support in hospital settings. http://www.idc.com/getdoc.jsp?containerId=CN245651. (2014). Retrieved 03-09-2015.

Qiong Xu, Hengyong Yu, Xuanqin Mou, Lei Zhang, Jiang Hsieh, and Ge Wang. 2012. Low-dose X-ray CT reconstruction via dictionary learning. *IEEE Transactions on Medical Imaging* 31, 9 (2012), 1682–1697.

Jinn-Yi Yeh, Tai-Hsi Wu, and Chuan-Wei Tsao. 2011. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems* 50, 2 (2011), 439–448.

Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. 2012. Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems* 36, 4 (2012), 2431–2448.

Hisako Yoshida, Atsushi Kawaguchi, and Kazuhiko Tsuruya. 2013. Radial basis function-sparse partial least squares for application to brain imaging data. *Computational and Mathematical Methods in Medicine* 2013 (2013), 7. DOI:http://dx.doi.org/10.1155/2013/591032

Kui Yu, Xindong Wu, Wei Ding, and Jian Pei. 2014. Towards scalable and accurate online feature selection for big data. In *2014 IEEE International Conference on Data Mining (ICDM'14)*. 660–669.

Xiaojing Yuan, Ning Situ, and George Zouridakis. 2009. A narrow band graph partitioning method for skin lesion segmentation. *Pattern Recognition* 42, 6 (2009), 1017–1028.

Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: Cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*. 10–11.

Xiaofan Zhang, Wei Liu, M. Dundar, S. Badve, and Shaoting Zhang. 2014. Towards large-scale histopathological image analysis: Hashing-based image retrieval. *IEEE Transactions on Medical Imaging* 34, 2 (2014), 496–506.

Xiaofan Zhang, Hai Su, Lin Yang, and Shaoting Zhang. 2015a. Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5361–5368.

Xiaofan Zhang, Fuyong Xing, Hai Su, Lin Yang, and Shaoting Zhang. 2015b. High-throughput histopathological image analysis via robust cell segmentation and hashing. *Medical Image Analysis* 26, 1 (2015), 306–315.

Xiaofan Zhang, Lin Yang, Wei Liu, Hai Su, and Shaoting Zhang. 2014. Mining histopathological images via composite hashing and online learning. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'14)*. Vol. 8674. Springer, 479–486.

Yang Zhang, Simon Fong, Jinan Fiaidhi, and Sabah Mohammed. 2012. Real-time clinical decision support system with data stream mining. *Biomedicine and Biotechnology* 2012 (2012), 8. DOI:http://dx.doi.org/doi:10.1155/2012/580186

Shang-Ming Zhou, Ronan A. Lyons, Owen Bodger, Joanne C. Demmler, and Mark D. Atkinson. 2010. SVM with entropy regularization and particle swarm optimization for identifying children's health and socioeconomic determinants of education attainments using linked datasets. In *The 2010 International Joint Conference on Neural Networks (IJCNN'10)*. IEEE, 1–8.

Xiang Sean Zhou, Sonja Zillner, Manuel Moeller, Michael Sintek, Yiqiang Zhan, Arun Krishnan, and Alok Gupta. 2008. Semantics and CBIR: A medical imaging perspective. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*. ACM, 571–580.

Ying Zhu. 2011. Automatic detection of anomalies in blood glucose using a machine learning approach. *Journal of Communications and Networks* 13, 2 (2011), 125–131.

Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin, and Brian Muckian. 2013. Big data solutions for predicting risk-of-readmission for congestive heart failure patients. In *2013 IEEE International Conference on Big Data*. 64–71.

Received May 2015; revised December 2015; accepted March 2016