Healthcare and medical Big Data analytics

Blagoj Ristevski and Snezana Savoska
Faculty of Information and Communication Technologies - Bitola, University "St. Kliment Ohridski" - Bitola, Republic of Macedonia

Abstract

In the era of big data, a huge volume of heterogeneous healthcare and medical data are generated daily. These heterogeneous data, that are stored in diverse data formats, have to be integrated and stored in a standard way and format to perform suitable efficient and effective data analysis and visualization. These data, which are generated from different sources such as mobile devices, sensors, lab tests, clinical notes, social media, demographics data, diverse omics data, etc., can be structured, semistructured, or unstructured. These varieties of data structures require these big data to be stored not only in the standard relational databases but also in NoSQL databases. To provide effective data analysis, suitable classification and standardization of big data in medicine and healthcare are necessary, as well as excellent design and implementation of healthcare information systems. Regarding the security and privacy of the patient's data, we suggest employing suitable data governance policies. Additionally, we suggest choosing of proper software development frameworks, tools, databases, in-database analytics, stream computing and data mining algorithms (supervised, unsupervised and semisupervised) to reveal valuable knowledge and insights from these healthcare and medical big data. Ultimately we propose the development of not only patient-oriented but also decision- and population-centric healthcare information systems.

Keywords: Big Data; medical and healthcare Big Data; Big Data Analytics; databases; healthcare information systems

4.1 Introduction

Nowadays, using numerous diverse digital devices that generate a massive volume of heterogeneous structured, semistructured, and unstructured data results with the explosive growth of the

Applications of Big Data in Healthcare. DOI: https://doi.org/10.1016/B978-0-12-820203-6.00005-9

different types of a large amount of data, which enable the extraction of new information and inherent insights contained in data. Such an enormous variety of data collected from various and heterogeneous sources in healthcare and medicine can make valuable comprehension for patients, clinicians, hospitals, pharmacy, insurance companies, and other involved parties.

Additionally, when data from social media, banks and credit cards, census records and diverse types of data with varying quality as the Internet of things (IoT) data for measuring vital signs are attached to the healthcare and medical data, a holistic view of a patient with environmental factors, which might influence patients' health, can be obtained. These linked data can be very sensitive and with different quality attributes. But it is very useful to discover the connections from electronic health records (EHR) and coding systems to establish common criteria which will be beneficial for big data analysis and visualization for different stakeholders.

As healthcare data sources, healthcare institutions are critical data producers that demand a huge architecture for storing a wide variety of data connected to the patients, healthcare institutions, government and municipality activities associated with healthcare as well as many activities of a wide range of healthcare stakeholders. A large volume of healthcare data is generated in hospitals during clinical treatments, labs and administrative procedures. There are many healthcare big data sources that have different attributes that have to be taken into account.

This enormous volume of data collected from this sector has to be analyzed to obtain specific knowledge for all stakeholders in healthcare. Because of a large volume, veracity, variety, value, variability and velocity, healthcare data have big data properties. Usually, these data are stored as patients' EHR, medical data records, coded with known medical and pharmaceutical coding systems such as ICD10 and SNOMED.

Regarding healthcare and medical big data, the first characteristic of these data is complexity. It comes with a wide range of activities connected with patients, physicians, hospitals and clinicians, healthcare providers, healthcare insurance companies, medical instruments and medical terminology, national regulations, pharmaceutical companies, healthcare research groups, healthcare IoT appliances, and the WHO needs and directions. These data are also connected with living conditions as environmental data, transportation and communications, social media and advertising data.

All these mentioned complex data sometimes have to be associated with suitable specific conditions in the context of healthcare big data. Healthcare big data can be produced also by many types of data sources: social media and markets, scientific instruments, mobile devices and services, technological and network services, hospital medical devices, EHR, physicians' notes, medical, and pharmaceutical research.

Certain ethical healthcare obligations demand a high quality of services for patients. Medical data storage and created documents aim to support quality management in healthcare institutions. So, historical disease data for each patient (medical audit), healthcare quality system monitoring, specific clinical insights and epidemiological data should also be provided. This could be very useful to gain suitable knowledge that can be used for further training and education of healthcare professionals as well as to assess medical students. These data have to be stored in high capacity distributed repositories and used from the other stakeholders for various purposes.

Thereafter, suitable data mining techniques are applying to the medical and healthcare data sets. The specific purpose of healthcare data mining procedures should be to support anonymized patient data exchanges among healthcare staff and institutions, to support external demands like law, reimbursement procedures and documentation for planning and control of healthcare services. Additionally, applied data mining techniques should support scientific research enabling patients' analysis as well as statistical data analysis, clinical delve into the data, epidemiological data analysis, information about critical insights with using of appropriate case studies.

Many efforts are made to lead healthcare data to a unique system that codes important medical and healthcare data in general [1]. Because the highlight of the medical data is the medical and health care of the patients, these data are typical clinical data containing disease history, symptoms, clinical notes, diagnoses, therapies, and predictions or prognosis of the patient health conditions. Data also have to be connected with nursing, labs, medical knowledge, epidemiological information, and other relevant healthcare information. Clinical data management systems have to use technical language for classifying healthcare data and to use nomenclatures, making a taxonomy of medical and healthcare big data. On the other hand, the evolving healthcare standards focus on healthcare data interchange possibility to have basic communication among different healthcare information systems and their components.

Design and implementation of healthcare information systems based on suitable big data should provide to the patients more efficient and cheaper healthcare services, an enhanced knowledge-based basis for decision making intended for the managers in healthcare institutions and insurance companies and benefits for the involved stakeholders.

Additionally, dealing with the security and privacy of the patient's data, which play a central role in the healthcare information systems, must be assured.

The analysis of medical and healthcare big data gathered following the described coding systems provides a tailored analysis of specific groups of medical data. Nowadays such an analysis is more than needed to detect the best manners of treatment and potential treatment anomalies, as well as the influence of different factors to each patient.

The rest of this chapter is organized as follows. Section 4.2 describes the properties of big data with a focus on their usage in medicine and healthcare and the various data sources as a base for big data analysis. The next section depicts all stages of big data analytics, from their creation to visualization, as well as commonly used data mining algorithms. The coding systems and taxonomy of healthcare and medical data are shown in the subsequent section. Section 4.5 focuses on medical and healthcare data interchange standards. In the subsequent section, as a methodology for the development of healthcare information systems, we describe frameworks for patient-oriented information systems for data analysis based on medical and healthcare data with a highlight on their necessary components and functions. In Section 4.7, we describe the concerns about data privacy, security, and governance in medicine and healthcare and give some directions towards handling these issues. Concluding remarks and directions for future work, such as choosing of proper software development frameworks, tools, databases, in-database analytics, stream computing and data mining algorithms as well as directions towards a development patient-, decision- and population-centric healthcare information systems, are given in the last section.

4.2 Medical and healthcare Big Data

Using numerous diverse digital devices that generate a large volume of heterogeneous data results with an explosion and significant growth of the voluminous complex data [2]. These data enable the extraction of new inherent insights in many disciplines. The aim of this enormous variety of data collected from various and heterogeneous sources is to make valuable comprehensions for patients, clinicians, hospitals, pharmacy, so the medical and healthcare data analysis become an extremely challenging task [3]. The healthcare analytics tools have to integrate these data generated from numerous and heterogeneous sources, providing valuable information and a base for healthcare researchers to improve current healthcare software solutions.

Although the term big data was introduced in the 1990s, the serious impact of database development has been achieved by extracting 3V's (volume, velocity, variety) characteristics in 2001, by Meta Group [4], defining big data as "Data-intensive technologies for data collection, data storage, data analysis, reasoning with data and discovering new data" [5]. In the next years, big data concept accepted all emerging technologies that influence human life and needs, and enhance the definition taking into account firstly 5V's and thereafter 6V's. Big data was especially useful for science and technology, business, healthcare and education. When the IoT has appeared, many sensors were connected for enhancing human life, producing a huge amount of structured and unstructured data. Big data then was just a part of the wider concept of data science that included many other methods and techniques for big data analysis and visualization usually included in the wider business intelligence and analytics (BIA) concept to create a big impact from big data [6].

Nowadays, big data usually refers to the following properties denoted as 6 "V's": volume, velocity, variety, value, variability, and veracity of the generated data that is delicate to analyze by using of standard data processing methods and platforms [7,8]. Volume characterizes the huge amount of created data, while velocity denotes data in motion as well as the frequency and speed of creation, processing and analysis of data. Complexity and heterogeneity of multiple data sets refer to the variety, while value refers to the data coherent analysis, which should be beneficial to the clients, costumers, managers, organizations, corporations and other stakeholders. Variability regards to the data consistency, while veracity refers to the data quality, relevance, uncertainty, reliability and predictive value.

Particularly, in medicine and healthcare, to obtain more medical-related data and to improve disease diagnostics, many wearable sensors, remote monitors, handheld, wearable, smart, and capturing devices are used, as well as data generated by using many novel omics technologies.

Current advances in patients' EHR, their fusion with social, behavioral, and diverse biological data have led to novel healthcare models that support personalized and precision medicine [3]. Personalized healthcare services provide individuals-tailored diagnoses, drugs, and treatments based on the psycho-physiological and spatial-temporal circumstances.

The main challenge in using healthcare big data effectively is to identify the potential sources of healthcare information and to highlight the value of linking these data together [9]. Healthcare data collection is the main data stream in healthcare big data analytics. Data sources can be curative, preventive as well as other types of healthcare and medical sources.

Curative data are medical records, lab tests, referral, and prescription data. Preventive data can be taken from growth cards, maternal and child healthcare (MCH) cards, school healthcare cards as well as family registration cards. Other data sources can deliver comprehensive contents, record filling (patient-retained), layout (self-explanatory), production forms and various environmental data. Demographic surveillance should also be taken into account in healthcare data collection because of the need of information such as causing mortality data, sex, age, age-specific fertility, perceived data for mortality and disability, expenditure for household healthcare, practices, service quality and costs covered with healthcare insurance [10].

System medicine, which combines clinical decision support systems (DSS) and EHR systems, aims toward individualized disease prognosis and treatment of the patients. These prognoses and treatments have to be based on various large amounts of data including phenotype data, omics data, and individual preferences of the patients [11]. Data used and mutually combined in system medicine can be categorized in three main groups: *personal data* (behavioral data, demographic data), *clinical data* (examination data, laboratory data, imaging data), and *omics* data (e.g., genomics, proteomics, metabolomics, transcriptomics, lipidomics, epigenomics, microbiomics, immunomics, and exposomics data) [11].

Data used in the healthcare systems and applications are categorized as unstructured, semistructured, and structured data.

Structured data have defined data type, structure, and format. Such data in healthcare systems are laboratory results, hierarchical terminologies of different diseases, their symptoms, information about diagnosis and prognosis, patients' data, drug and billing information [3].

Semistructured data, which are usually generated from various sensor devices for monitoring of patients' conditions and behavior, are organized with minimal structure properties.

Besides these two categories of data, *unstructured data* have no inherent structure. These data usually contain doctor's prescriptions, written in natural human languages, biomedical literature, clinical notes and images.

Many researchers pay attention to the symbiosis of data types collected from healthcare services and the structure of these data. Weber et al. had created the tapestry of all healthcare data sources whenever they are collected or stored, regardless of the used coding and classification system, as systemized in Table 4.1 [9].

EHR data can be categorized into two main categories: *electronic medical records (EMR)* and *sensor data*. EMR data usually consist of patients' medical history, medical features (e.g., diagnoses, lab tests, medication, procedures, unstructured text data, image data) and socio-demographic information about patients [12]. Sensor data are collected from various sensors and they originate from a huge number of users and devices, hence generating enormously large amounts of real-time data streams. The main characteristics of EMR data are high-dimensionality, missing values, sparsity, irregularity, intrinsic noise and bias [12]. Problems with missing data values and data sparsity are usually solved by using removal or imputation methods.

Medical imaging is a vigorous source of phenotypic data appropriate for further data analysis, personalized medicine, predictive analytics and artificial intelligence [13]. Medical imaging data are generated by imaging techniques such as X-ray, mammography, computer tomography (CT), ultrasonography, fluoroscopy, photoacoustic imaging and magnetic resonance imaging (MRI), histology, positron emission tomography (PET), radiography, nuclear medicine, tactile imaging, echocardiography, angiography, and elastography [7,14].

Prescription data contained in the patient EHR generated by physicians, clinical notes, medical research reports are examples of data that contain natural language terms, mathematical symbols and graphs [15]. These incompatible data structures and formats along inconsistent data semantics combined with huge data volumes make healthcare big data analytics a challenging and demanding task.

Large scale omics data can ensure clarifying of the molecular base of particular diseases and disease risks. To model and analyze complex interactions that occur between entities in biology,

Table 4.1 Simplified information sources related to individual healthcare and data classification.

Data types		Electronic pill dispenser	Prescribed medication	Medication instruction	Medication taken
Medication	OTC medications	Medication filled	Dose/Route NDC RxNorm codes	Allergies out-of-pocket expense	Diaries herbal remedies alternative therapy
Demographics			HL7		
Encounters		Employ seek days	Visit type and time		
Diagnosis		Death record	SNOMED ICD10	Differential diagnostics	
Procedures			CPT/ICD10		
Diagnostics	PHR	Home treatment, tests, monitoring	LOINC pathology histology radiology lab vital signs	Reports, imaging, digital clinical notes, physical examinations	
Genetics		SNPs, arrays			
Social history		Police and other records			Blogs, Facebook posting, tweets
Family history				Digital clinical notes	
Symptoms		OTC purchases (indirect)			
Lifestyle	Fitness clubs membership		Credit cards		
Socioeconomics			Census records		
Social network					
Environment	Climate, weather,	Public health database	GIS, EPA, health map	News feeds	
		Structured data		Unstructured data	

medicine and neuroscience, networks are fundamental and very suitable tools. The complex network nodes represent dynamic entities such as genes, microRNAs, proteins, metabolites, whereas the edges represent the links and interactions among nodes [16].

4.2.1 Exposome data

Creating an individual model can be very important for human beings. These analyses also demand a huge amount of healthcare big data sets and complex data mining tools with diverse focuses of analysis to make available important insights based on EHR, known coding system and ontologies of exposome [17]. They called expotype as a particular set of exposome features of an individual gathered throughout a certain time and/or space [17]. In this manner, the authors in Ref. [17] had stated that development of a template-driven model to identifying exposome concepts from the Unified Medical Language System (UMLS) and create expotype is important. They also defined exposotype terms as the metabolomic profile of an individual that considers an occurrence of exposure.

When wider integration of healthcare and biomedical data with environmental data is required, the term exposome is introduced as a novel conception that tends to delineate biotechnical approaches and to systematically quantify a massive subset of environmental exposures of an individual for whole lifespan [18]. Some data in this concept have genetic or clinical backgrounds. Some of the data are also associated with the integration of genotype-phenotype data, environmental risk factors at the individual level [17]. This concept can be essential for understanding the biological basis of diseases, taking into consideration the influence of ecosystems to each person. Authors in Ref. [17] stated that most diseases outcome from the multiplex interchange between genetic and environmental factors.

Some reasons of emerging of this coined word are the partition of the landscape of disciplines, which are interested in exposome characterization from different points of view such as environment, health, exposure, toxicology and health services. There are many interdisciplinary subbranches of exposome, which lead to novel coined words and terms, such as urban exposome, occupational exposome, epidemiology as public health exposome, socioexposome, nanoexposome, infectoexposome, drugexposome, psychoexposome, etc. [17].

Besides omics data, exposure data in the wide sense has to gain certain nongenetic data for the patients as data for patient behavior and habits, as social determinants of health and physicochemical exposures. They can be taken from various sources, as biomonitoring data, exposure to particular environmental agents like smoke, geographic information systems (GIS), environmental sensors, etc. Also, EHR data, digital health sensors data, mobile applications and consumer behavior data can be considered as exposome data. They can be used to create new dimensions and multi-omics data models [17].

Many databases are created for this purpose such as the US National Health and Nutrition Environmental Survey (NHANES) that treats the exposome theory and this had achieved successful results. Since 2013, the Institute of Medicine (IOM) reported that comprehending social and behavioral domains and data in EHR is important because of increasing clinical awareness of the patient's state, broadly considered, and to link clinical, public health and general public resources [19].

4.3 Big Data Analytics

Big data concept requires real-time data processing and development of real-time predictive models. Such rapidly growing amounts of data are faced with inefficient storage, preprocessing, processing and analysis by traditional relational databases [2]. To perform efficient multisite and multivariable searching and querying, high indexing and efficient data lookup of the big data sets are required. Another stage is preprocessing of the raw data that might be inconsistent, inaccurate, erroneous and/or incomplete. To make big data analysis more reliable, integration of data from heterogeneous sources must be performed to have a conventional structured form. To improve the quality of collected data from various sensors/devices and to obtain more reliable analytics results, identifying and removing incomplete, inaccurate and irrelevant data should be made. These unreliable data are usually replaced with interpolated values.

For these reasons, when we are talking about healthcare big data analysis and visualization, as most desirable techniques that create rapid information and deep insights, we have to analyze who will use results of data analyses and visualization and for which purposes. After defining the users (stakeholders), we have to analyze all necessary data sources and data formats and manner how to extract the information and knowledge from these healthcare and medical data, which demands to scrutinize design of the healthcare information system, data preparation for further data analysis and visualization and choosing suitable analytics and visualization tools and platforms.

Big data analytics acquits an enormous variety of data from former and current customers to obtain valuable knowledge to improve decision-making, to predict customer habits and behavior and to obtain real-time customer-tailored offers. Big data analytics in healthcare and medicine aims to bridge the gap between costs and outcomes in healthcare, which is a result of poor management of insights from research and poor resources management [20]. These analytics goals are achieved by *prediction, modeling, and inference.*

Moreover, data created by omics technologies can significantly improve the prediction of diabetes, heart diseases, cancer, and other diseases [21]. The main barriers in computer science related to the integration of omics data into clinic systems are the development of a model of cellular processes that cover noncorresponding omics data types, the limitation for data storage and organization of heterogeneous data sets, generated from diverse high-throughput omics technologies, and the lack of suitable multidisciplinary data scientists with wider knowledge in computer science, biology, medicine, bioinformatics and data mining [21]. To increase computing features and scalability of different omics data, 3-D memory and scalable methodologies are developed [7].

Data acquisition is followed by the *data cleansing* step that detects and removes the before-mentioned anomalies of data. In the subsequent stage, raw data must be *transformed* by data normalization and aggregation. During data transformation scaling, cleaning, splitting, translating, merging, sorting, indexing, and validating of data are performed as substeps of this stage to make data consistent and easily accessible for further data analysis stages [22]. Data transformation is important to obtain valuable data for supporting evidence-based decision making in healthcare organizations [23]. After data transformation, *their integration* in a unified standard form is required. Sensor data are generated by diverse medical devices at a different sampling rate, that makes data integration an important step [12]. *Data transmission* is a technique that transfers integrated data into data storage centers, systems or distributed cloud storage for further analysis [2].

Data reduction is a method that makes size reduction of the large data sets so that suitable data mining techniques and application could be applied. Moreover, some data mining techniques require *discretization* of the continuous attribute intervals, which means discretization techniques should be applied to the data, before further suitable analysis, interpretation and visualization.

Since physicians use voluminous clinical notes and unstructured text data, to perform an optimized full-text exploration of medical data, searching and indexing tools are needed. These tools are employed for worthwhile distributed text data management and indexing of huge amounts of data. The healthcare industry employs machine learning techniques to convert plentiful medical data into applicable knowledge by performing predictive and prescriptive analytics. Recent advance in healthcare sensor devices elicits the data processing from diverse data sources to be achieved in real-time.

To discover correlations, patterns and previously unknown knowledge in the large datasets and databases, appropriate *data mining techniques* should be applied. Data mining algorithms in medicine and healthcare enable analysis, detection and prediction of specific diseases and hence aid practitioners to make decisions about early disease detection [24]. Such early disease prediction helps to decide and use the most suitable treatment considering the symptoms, patient EHR and treatment history.

Data mining algorithms can be *supervised, semisupervised, and unsupervised learning algorithms.* The unsupervised algorithms aim to find patterns or groups (clusters) of entries within unlabeled data. Supervised learning, which uses training sets of classified data, aims to predict a known result of target and to classify or infer testing data sets. Semisupervised learning algorithms use small sets of annotated data and larger unlabeled data sets.

4.3.1 Unsupervised learning

Clustering is an unsupervised learning algorithm that is used to find groups of data entries (clusters) by using distance metrics. The clustering aims to minimize intra-cluster distances of the data entries belonging to the same clusters and, at the same time, to maximize inter-clustering distances among data entries that belong to different clusters [25]. Besides commonly used distance-based clustering algorithms, density-based clustering aims to find areas (clusters) with higher dense entries compared to the remainder of the data set. Clustering algorithms include hierarchical clustering, k-means clustering, fuzzy c-means clustering, self-organizing maps, principle-based clustering, as well as biclustering, where rows and columns of the data set matrices are clustered simultaneously, and triclustering, which uses tri-dimensional data analysis to discover coherent three-dimensional subspaces (triclusters).

4.3.2 Supervised learning

Classification is a commonly exploited method for supervised learning, that is, predictive modeling where output vector values are categorical [20]. The aim of the classification is to create rules to assign and organize data entries to those of the preidentified set of classes so the benefit of data is most efficient and effective. Commonly used classification algorithms are decision trees, Bayesian networks, neural networks, support vector machines, boosting, logistic regression and naïve Bayesian classifiers. Classification of medical and healthcare data sets is used to develop DSS for diagnosis as well as for predicting models for prognosis based on the big data analysis. Linear regression is a statistical analysis technique to represent trends in the data, quantifying the relationship between dependent variables and the independent data variables.

4.3.3 Semisupervised learning

Semisupervised learning algorithms are between unsupervised and supervised algorithms. This learning belongs to the machine learning algorithms that use a large amount of unlabeled data and a small amount of labeled data.

Data visualization aims to make a graphical representation that enables users to interact with data to extract useful information and knowledge [26]. Data visualization tools in healthcare aids to detect patterns, tendencies, outliers, clusters, to analyse time-series data and to improve clinical healthcare services and public health policy [3]. Visualization generates outputs like numerous visualization reports (e.g., charts, interactive dashboards), real-time reporting information (e.g., diverse notifications, alerts and key performance indicators) and clinical summaries (e.g., historical reports, statistical analyses, time-series comparisons) [22].

For healthcare and medical big data analysis, interoperability specification is very important, as well as the used coding systems [27], especially when healthcare clinical data, omics data and sensor data are applied [28]. They also have to be related to the patient and prescription data. Usually, hospitals and clinics have integrated information systems that produce a huge volume data suitable for data analysis, but with restricted possibilities, generating data for decision support for their stakeholders and defining interactions and workflow systems to provide healthcare quality systems, patient data privacy as well as optimization of healthcare costs. They also have an enormous quantity of reports with statistical data analysis and periodical reports. But, when they have to be analyzed at a higher level, the system has to be previously designed for storing data in desired formats for further data analysis [27].

The motivation of healthcare stakeholders such as university hospital centres is very high because they have to deal with a large amount of data. For example, one hospital has to deal with several million documents with a huge amount of different data such as clinical notes and lab results. They had to create a lot of papers, which are difficult to analyze and hence numerous data are produced. These data have to be connected with the patients' living conditions [29]. In Ref. [30], authors have proposed the concept of a four-level healthcare system that has to be followed and in which patients, care teams, organizations and environment play key roles.

Suitable system engineering tools should be employed to handle healthcare issues. First, systems design tools for implementation of parallel engineering, quality control functions deployment, human factors tools, failure analysis tools [30]. Subsequently, particular systems analysis tools are required for the following activities: modeling and simulation tools (to provide queuing methods), discrete event simulation, supply chain management, game theory and contracts, system dynamics modeling, productivity measuring and monitoring tools. For financial engineering and risk analysis, tools are required for stochastic analysis, value at risk, optimization of individual decision making and distributed decision making market models. Systems control tools are used for statistical process control and scheduling. These analysis tools have to act and to be kept together to achieve a synergy in achieving the best analysis results in healthcare and medicine.

The most important case for data analysis is bringing relevant information for decision making associated with the disease diagnoses, treatments and interventions. Also, there are large amounts of data related to the need of communication help from medical practitioners from different healthcare institutions, to the EHR between two conditions of diagnoses. Additionally, plans for treatments and storing of data about patient lab tests, plan for creating appropriate healthcare services and medical documents required for the national healthcare bodies. This is important when they have to perform statistical analysis and to create clinical reports.

The second important reason and motivation for storing huge data sets are administrative healthcare data that supports reimbursement procedures, considering the data collected from various healthcare services, interventions and treatments as well as the expenses for these services. The healthcare institutions' managers have to plan and control the working processes and these data should aid them to increase the transparency and enhance the decision making processes to manage the available resources.

Many implications arise from the need for evidence-based medicine and patient care when healthcare legislation and law are highlighted. There are many law procedures for insufficient document quantities that can have negative implications for healthcare institutions.

4.4 Healthcare and medical data coding and taxonomy

Many efforts are made to lead healthcare data to a unique system that codes important medical and healthcare data in general [1]. The need for medical and healthcare information arises over time. The healthcare stakeholders know that the healthcare documentation has to provide information, which has to be complete, without noise, on time, without missing values and outliers, in the format that has to be presented to the healthcare authorities. Indeed, the information has to be comprehensible and in form of logics for the desirable knowledge in medicine and healthcare services [1]. Because the highlight of the medical data is the medical and healthcare of the patients, these data are typical clinical data containing disease history, symptoms, clinical notes, diagnoses, therapies and predictions or prognosis of the patient health conditions. The data also have to be connected with nursing, medical knowledge, epidemiological information and other relevant information.

Clinical data management systems have to use technical language for classifying healthcare data and to use nomenclatures. The data can be classified according to data description and data mining demands. According to the subjects, classification can be done by following classification criteria [15]:

- clinical information;
- medical knowledge that abstract individual patient's insights for diseases and;
- attributes of healthcare systems as data for institutions, accidents, etc.

For instance, CDMS1 is a classification that considers the meaning of data in 5 classes: (1) contains primarily clinical facts; (2) contains primarily medical knowledge; (3) contains healthcare attributes; (4) contains a balanced mix of many types of information; (5) does not belong to any of previously mentioned classes. CDMS2 classification can have following classes (1) data-oriented to the patient or patients' groups;

(2) class according to the level of standardization (standardized or unstandardized data). Other classification criteria take into account the horizontal or vertical medical documentation (CDMS3), patients, diseases and interventions (CDMS4) and use IT tools or conventional documentation (CDMS5) [27].

Medical coding systems are used to provide fewer problems when data analysis has to be performed, shorter and accurate data entries, less memory space, possibility to grouping data according to codes and groups. The coding language is based on the medical conceptual coding system and using a thesaurus (lexicon).

The current healthcare coding systems are accepted and announced by WHO. *ICD10* (International Classification of Diseases) is the most significant coding system that contains data for death statistics, healthcare quality control, and international register of causes of death. The list of causes of death was made since 1893 by Bertillon [27] and later in 1964 ISI (International Statistical Institute) produced the document International Classification of Diseases and Causes of Death [27]. Nowadays, the current version is the tenth revision of ICD10 with a digital code length of 4−5 alphanumeric letters/digits. The first code character is a letter, while the rest 2−4 are digits. The fourth code character is divided by a decimal point.

The ICD10 contains 21 chapters for diseases [27,31]. For instance, chapter 4 classifies endocrine, nutritional and metabolic diseases into 261 groups of diseases (e.g., E10 to E14 are diabetes mellitus), 2000 classes with 3 numbers (e.g., E10 is a class of insulin-dependent diabetes mellitus), 12000 numbers with 4 digits as classes of diseases (e.g., E10.1 as insulin-dependent diabetes mellitus with ketoacidosis). The special codes that begin with U50 to U99 are reserved for research purposes. The classes are created in general according to statistical criteria (e.g., diabetes prevalence). There are no semantic features in the ICD10 classification.

Some extensions of classification are done to achieve the higher groups' granularity with ICD10-CM (clinical modification) to overcome specific organizational and terminological healthcare system demands.

Other classifications are *ICPM* (International Classification of Procedures in Medicine) created by WHO for research purposes and *ICD-9-CM* created by the US National Center for Health Statistics (NCHS) and Health Care Financing Administration (HCFA). The code consists of 4 bits: 2 bits for a group of the procedure and 2 bits for a specific procedure and its specification. The last bit has topological meaning (e.g., 30−34 are codes of operations for the respiratory system) [31].

Next important coding system in healthcare and medicine is Systematized Nomenclature in Medicine—*SNOMED* (Systematized Nomenclature of Human and Veterinary Medicine 1975/1979/1993 from CAP), 2000 SNOMED-RT (reference technology), SNOMED-CT (clinical terms), SNOMED-RT as multidimensional nomenclature that has the two-layer alphanumeric notation: 2 bits for the base hierarchy (T-Topology, P-Procedures) and second layer bits for the concept identification [27].

The research with joint efforts resulted in gaining CAP (Clinical Audit Platform)/NSH (National Health Service), promoted in 1999 [27,32]. It integrates SNOMED-RT and NSH's version of clinical terms. This unified terminology solved many compatibility data problems and provided basic building blocks for world clinical communication.

The classification of malignant tumors (*TNM*) provides a consistent classification of anatomic spreading of malignant tumor diseases, the phases of diseases and particular common names. This oncological classification provides topological and morphological unification on the ICD10. Since 1953, Union for International Cancer Control (UICC) and International Commission for cancer statements and results presentation and cancer treatment, argued with this mainstream method of classification malignant tumours [33], which is used until now. The system takes into account the tumor spreading (T0-T4), the stadium of the nodules of metastasis (N0-N3), and the presence of metastases (M0-M1). For instance, code T2N1M0 means that the malignant tumour is in the second phase of spreading, N1 stadium of nodule metastasis and there are metastases (M0).

The next coding system in healthcare is *MeSH* (Medical Subject Heading) from the US National Library of Medicine [34]. The aim was to code the subjects of medical and healthcare literature according to the poly-hierarchical conceptual system. UMLS is another NLM (National Library of Medicine) product that brings the clinic codes and literature in metathesaurus (lexicon) to provide automatized linkage through the clinical case studies and available healthcare literature.

4.5 Medical and healthcare data interchange standards

The emerging healthcare standards focus on healthcare data interchange possibility to have basic communication between different healthcare information systems. The examples of usage of data interchange standards that affect healthcare are Health Level 7 standards (*HL7*), National Council for Prescription Drug Programs (*NCPDP*), Digital Imaging and Communications in Medicine (*DICOM*) and ANSI *X12N* standards. HL7 is developed as an HL7 messaging standard to allow interoperability among healthcare applications [35]. This standard is involved in other standards activities, but the messaging standard is denoted as HL7. Other commonly used HL7 standards are Clinical Context Management (CCM) specifications, Arden Syntax for Medical Logic Systems and Electronic Health Record functional model.

DICOM is a standard that supports communication of digital image data regardless of the device producer and it deals with picture archiving and communications systems (PACS) [35,36]. It was first published by the American College of Radiology and the National Electrical Manufacturers Association in 1985. The next NCPDP is ANSI accredited Standards Developing Organizations (SDO) who established a standard for the electronic submission of intermediary drug claims [37]. ASC X12N is the subcommittee of ASC X12 intended to deal with EDI (electronic data interchange) for the insurance sector. Its purpose is to achieve healthcare task group TG2. X12N/TG2 develops and maintains standards for healthcare EDI [38].

Health Record Content Standards are associated with creating functional standards for EHR content. These standards can be the HL7 EHR functional model or ASTM (American Society for Testing and Materials) healthcare informatics subcommittees Continuity of Care (CCR) standard. HL7 EHR functional model is adopted in 2004 as the second draft and contains three components: direct care, supportive and information infrastructure, as shown in Fig. 4.1. CCR has been developed by the ASTM Healthcare Informatics subcommittee and it tends to provide core data sets of the most relevant and timely facts about a patient's healthcare. It is completed when a patient is shifted to another healthcare provider. The third version included nine elements grouped in three main data groups, as shown in Fig. 4.2.

HIPAA (Health Insurance Portability and Accountability Act) standards also influence electronic transaction standards for healthcare and it is mainly from ASC X12N. Standard HIPAA code sets comprise: ICD-10, CDT (Code on Dental Procedures and Nomenclature), HCPCS and CPT-4 [35,39]. It also identifies deliberated standards' organizations to develop, maintain and adapt relevant EDI standards [40]. It comprises ACS X12, Dental Content committee of the ADA (American Dental Association),

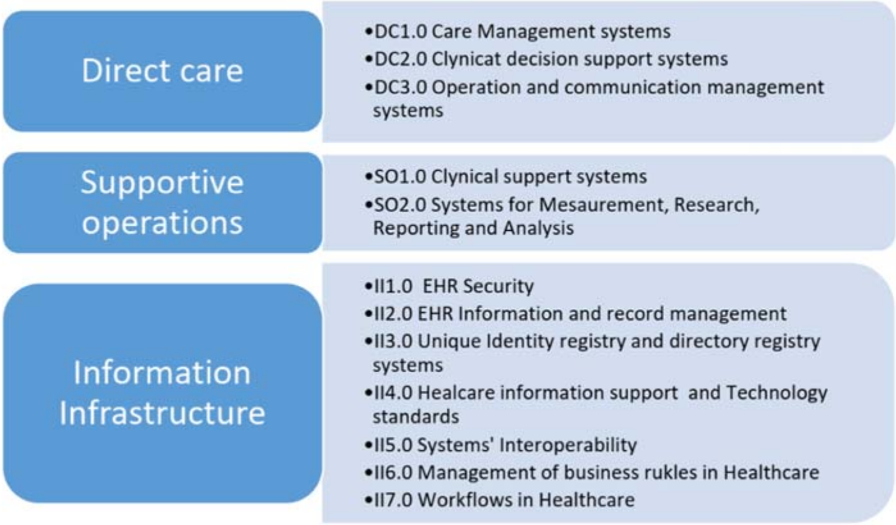
Direct care
•DC1.0 Care Management systems
•DC2.0 Clynicat decision support systems
•DC3.0 Operation and communication management systems
Supportive operations
•SO1.0 Clynical suppert systems
•SO2.0 Systems for Mesaurement, Research, Reporting and Analysis
Information Infrastructure
•II1.0 EHR Security
•II2.0 EHR Information and record management
•II3.0 Unique Identity registry and directory registry systems
•II4.0 Healcare information support and Technology standards
•II5.0 Systems' Interoperability
•II6.0 Management of business rukles in Healthcare
•II7.0 Workflows in Healthcare

Figure 4.1 Levels of healthcare information system management and base coding systems.

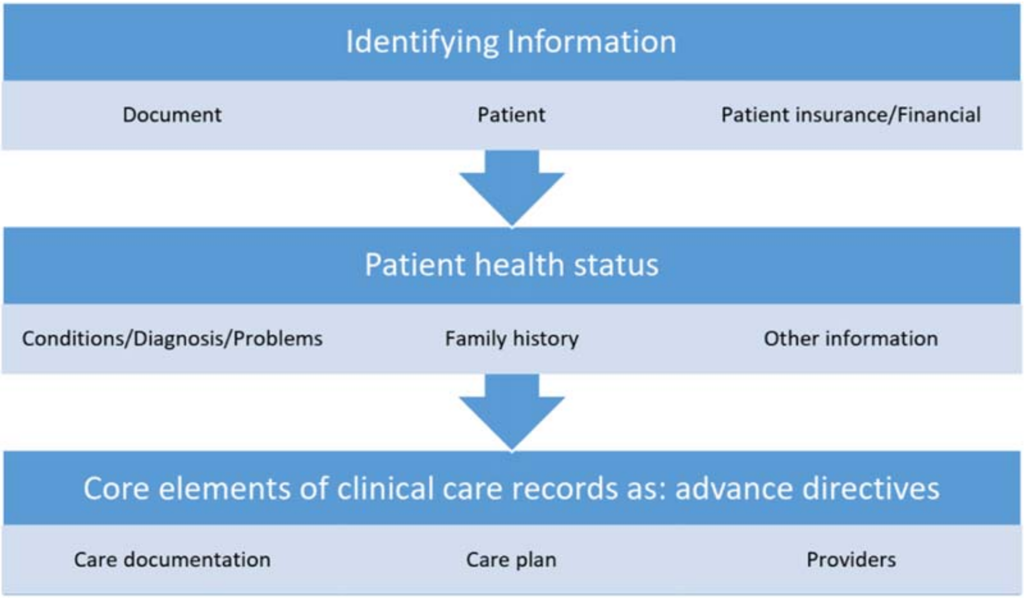

Figure 4.2 Specific clinical information care documentation and electronic health records.

HL7 and NCPDP, NUBC (National Uniform Billing Committee) and NUCC (National Uniform Claim Committee) standards.

The above-mentioned standards and coding systems function in given healthcare infrastructure as NHII (National Healthcare Information Infrastructure). NHII is neither a simple set of health information standards, nor a government agency, nor a centralized database for medical records, but it is a complete knowledge-based network of interoperable systems of clinical, public health, and personal health information [41]. This is a basis from where data are taken for further big data analytics and visualization [35]. But, these coding systems, classifications, terminologies and standards in healthcare have to support big data analysis in certain aspects. The biggest problems are related to the 6V's big data concept, which means that data might not be structured, might be in various formats, created in a huge volume and velocity, with a variety of data sets with different metadata.

4.6 Framework for healthcare information system based on Big Data

Developing healthcare information systems based on healthcare and medical big data has to take into account all stakeholders involved in healthcare and medical research.

The *patients*, who pay for health insurance and play a central role in healthcare systems, expect the healthcare institutions and hospitals to deliver a wide assortment of high-quality healthcare services at a reasonable cost. Besides physicians' diagnosis, patients can gain more medical and healthcare knowledge, such as symptoms, hospitalization, medicament information through social networks, forums, etc. [3]. Moreover, using different healthcare sensors and wearable devices provides an opportunity for telemedicine, which results in the creation of a huge amount of data.

Medical personnel, as a key stakeholder who generates various data, such as medical imaging data, CT, laboratory results and clinical notes. The medical staff makes a diagnosis based on these data and symptoms. The collected data and then integrated into the big data repository help physicians to make the right diagnosis and then to prescribe the appropriate medicaments and to observe patients' health conditions.

Hospital strategic operators should use available big data to strengthen the relationship between patient satisfaction and the offered services and to optimize using the healthcare departments and resources [3].

Pharmaceutical research based on available big omics data helps to comprehend the drugs and biological processes that lead to successful drug design. Additionally, medicaments prescriptions and recommendations for a particular disease, dosages, consumption quantity as well as sales history from a specific pharmacy should be included in big data analytics.

Clinical researchers can benefit from various clinical reports generated by big data analytics tools and data contained in the patients' EHRs.

Healthcare big data introduce opportunities to *healthcare insurance companies/organizations* to generate reports and appropriate health plans and trends for frequently occurring diseases in a particular geographic region. Such reports and plans can enable healthcare insurance funds, organizations and institutions to predict and to detect patterns of realistic claims and uncommon outliers to minimize the financial misuse costs [3]. Furthermore, analyzing the patients' behavior big data taken in real-time enables these insurers to introduce novel business models such as usage-oriented insurance, depending on the particular country's law and regulations.

Healthcare software developers play a very crucial role in the logical and physical design and development of healthcare information systems. These computer science specialists have to have a very wide interdisciplinary knowledge of computer science, data science, data mining, bioinformatics, information systems, healthcare, medicine and biomedical engineering.

The emergence of the big data and their usage in medicine and healthcare causes the development of numerous mobile healthcare services and applications that can employ and integrate data from heterogeneous sources such as biosignals (e.g., electroencephalograms, EEG; electrocardiograms, ECG), data from wearable sensor devices, laboratory data, etc. These data should be integrated along with pharmaceutical and regulatory data into models on a high level in a cloud computing environment, to address interoperability, availability and their sharing among different stakeholders such as medical physicians, patients, healthcare insurers and pharmaceutical companies [3]. Most of the proposed healthcare information system frameworks are structured from the following layers [1,3]:

- data connection layer;
- data storage layer;
- big data analytics layer; and
- presentation layer.

The role of the data connection layer is to identify, extract and integrate medical and healthcare data, while relational, nonrelational and cloud-based data are stored in the data storage layer. Big data analytics layer provides diverse analytics such as descriptive, predictive and prescriptive analytics. The presentation layer provides the developing of graphical workflows and dashboards and various kinds of data visualizations.

Besides these layers, the frameworks have to address the privacy and security issues of the system on several tiers. Sensitivity tier ensures patient information such as disease name and its status, patient mental health and biometric identifiers. The security tier authenticates patient data such as name, date of birth, doctor name, etc. To secure the privacy and security of patient data, the system should adopt a two-level security mechanism. The first security level is associated to the authorization of the user concerning retrieving patient data in the clinics by providing provisional user and patient identifiers [3]. To access the patient data at the inter- and intra-clinic levels, an OTP (one-time password) based security mechanism level should be employed.

Depending on the purpose of the analysis and data types, analysis of big data is dissociated into three parts: Hadoop MapReduce, stream computing and in-database analytics [42].

As a result of the voluminous big data and various data formats, new *NoSQL (not only SQL) database management systems* are required to integrate and retrieve data sets and to enable data transfer from standard into new operating systems. These NoSQL databases, which are used for big data storing, are classified into following 4 categories (some of them overlapping): column databases (e.g., HBase, Cassandra), document-oriented databases (e.g., MongoDB, OrientDB, Apache CouchDB, Couchbase), graph databases (e.g., Neo4j, Apache Giraph, AllegroGraph) and key-value databases (e.g., Redis, Riak, Oracle NoSQL Database, Apache Ignite).

Hadoop MapReduce is an SQL-based programming model, which can process large amounts of data sets through a Hadoop cluster by provided parallelization, distribution and scheduling services. MapReduce allows analysis of structured, semistructured and unstructured data in a massive parallel processing (MPP) environment Apache Hive is a relational model for querying, searching, analyzing huge data sets stored in Hadoop Distributed File System (HDFS). It uses HiveQL as a query language that transforms typical SQL queries into MapReduce tasks [43]. To store data in distributed and scalable databases, Apache HBase is a suitable system.

Stream computing supports high-performance big data processing in real-time or almost real-time. Stream computing analysis of healthcare big data can respond to unexpected events that occur, such as customer account misusing and to determine quickly the most appropriate actions. Suitable non-Hadoop processing tools for streaming data processing are Spark, Hive, Storm and GraphLab [44].

In-database analytics provides high-speed parallel processing, scalability and optimization ensuring a safe environment for sensitive data. Results of the in-database analysis are not real-time and this analysis in healthcare supports preventive healthcare practice and evident-based medicine.

4.7 Big Data security, privacy, and governance

Medical and healthcare data are generated from diverse multiple sources, which means that patients' *data security* is a big concern, as well their *privacy* due to the major risk of data leakage since their massive usage of third-party infrastructures and services. Cloud storage of these data can be vulnerable by potential malicious outsiders who can access the cloud platform and for instance, can act man-in-the-middle attacks. The primary goal is to provide confidentiality, availability and integrity of the patients' data with achieving security in healthcare systems [45,46]. Confidentiality of the data can be attained by protecting data from accidental and unauthorized users. When big data are stored in databases, encryption methods for data protection can be categorized into table encryption, disk encryption and data encryption [2].

Another major legal and ethical issues are related to the ownership of data and the developed applications that employ patients' data. Particularly, whether patients' data, which are used for development and validation of application and analytics models, can be reused, shared and/or even sold [13]. Especially, concerns are raised when the application that employs patients' data for development and validation should be sold for profit [13]. Patients, whose data are crucial for application development and validation, should be preacquainted that the data cannot be reused for other purposes and will not be misused [13]. Data recycling, data repurposing, data recontextualization, data sharing and data portability are the most commonly used forms for data reusing, as well as "the right to be forgotten" [47].

Data *recycling* covers using the same data, in the same manner, more than once. When a patient chooses another medical practitioner or health insurance company, it should not be allowed to the previous ones to use the data for that patient. When patients' data are used for a different purpose than the main one, it is categorized as data *repurposing*. While interpreting big data in a different context than in which they were primarily collected, for example, the same data physicians and health insurance companies can be interpreted differently or may have a different meaning, is classified as data *recontextualization*. Data *sharing* of medical data is sharing or disclosing in a specific context for particular purposes to other people or institutions, while data *portability* is the capability for patients to reuse their data through different devices and services. The "right to be forgotten" is the right of the data owner to invoke or block secondary using of his/her data.

It is essential for the healthcare organization to secure personal data and to address the risks and legal responsibilities associated with personal data processing, according to the valid national and international laws, policies and regulations for data privacy [48]. Storing healthcare big data in a public cloud, which is a cost-saving alternative, requires solving security risks and patient privacy control since the data access is controlled by third parties. Differently, storing data in a private cloud, which keeps the sensitive data in-house, is a more secure option, but is a more expensive choice. This means that healthcare managers have to make a trade-off between the project budget and the security and privacy of sensitive patient data.

Data governance is also a very important part of a healthcare system framework. Typically, it is composed of master data management, data life-cycle management and data privacy and security management [42]. Master data management refers to the governance, policies, standards and tools for data management, while life-cycle management manages the business data lifecycle from achieving data, warehousing data, testing and providing diverse application systems. Data security and privacy management deliver activities related to discovery, monitoring, configuration appraisal, auditing and protection of healthcare big data.

4.8 Discussion and further work

All the mentioned issues in this chapter cannot cover all the needs of healthcare and medical data for global users but intend to give a holistic view of the problems of healthcare big data analysis and visualization.

The human exposome, as big data will increase the healthcare and medical repository, and IT has to deal with the analysis of these data, we are convinced that this will be the challenge in the following years. The purpose is evident, to produce data for personal health risk assessment and for better living conditions for every human being.

Taking into consideration previous work in many countries and current healthcare information systems organization, as well as the needs of integration of cross border healthcare systems, that are considered in a couple of ongoing European Union projects [49,50], such as cloud-oriented cross border solutions, healthcare big data analysis and visualization have to be considered as a global challenge. It is necessary to create wide range taxonomy for healthcare and medical big data analytics as well as to create standards for usage of big data from different stakeholders with maximal security and privacy of patients' data as one of the priorities. Then, we have to take into account all big data methods and tools to create the most suitable tools for analysis and visualization that are tailored to the various stakeholders' needs.

As further work, due to the huge amounts of generated data, we suggest that more efforts should be made towards big data governance that manages with the rules and control over data, their integrity and standardization. More efforts have to be made to improve the quality of the patients' EHR, sensor and omics data, which is still a demanding task in big data analytics. These data have to be integrated into a unique clinical system, which would result in reducing of waste of resources and therefore to provide to the patients more efficient and cheaper healthcare services. The newest concept of Personal Health Record (PHR), where the patient is the data owner and plays a key role ins data collections, enables a new vision for personalized medicine and healthcare due to the data ownership. Combining PHR with a new concept of exposome and IoT data provides a wider horizon for using big data analytics methods for patient's healthcare risk assessment as well as predicting the risk of some diseases and proposing appropriate preventing actions. It is a challenging task for the next decade to employ a wide community of medical practitioners, computer scientists as well as other specialists.

As a suitable software framework for application development, we suggest Hadoop MapReduce that can process large amounts of data sets through a Hadoop cluster by provided parallelization, distribution and scheduling services. MapReduce allows analysis of structured, semistructured and unstructured data in the MPP environment. Because many of the generated data are streaming data, stream computing principles should be considered. Stream computing supports high-performance big data processing in real-time or almost real-time and using suitable tools such as Spark, Hive, Storm and GraphLab [44].

Because gathering of health and medical big data grows exponentially, to improve clinical decision making, development of decision-centric information systems is needed. Besides patient- and decision-centric healthcare information systems, nowadays when huge amounts of infective disease data are generating, there is an urgent need for development of population-oriented information systems. These information systems have to integrate GIS, particularly on a worldwide level to detect patterns of disease spreading, and to stop the propagation of a particular disease and hence to help public healthcare institutions to handle the novel disease.

References

[1] K. Beaver, Healthcare Information Systems, Auerbach Publications, 2002.

[2] A. Siddiqa, I.A.T. Hashem, I. Yaqoob, M. Marjani, S. Shamshirband, A. Gani, et al., A survey of big data management: taxonomy and state-of-the-art, J. Netw. Comput. Appl. 71 (2016) 151–166.

[3] V. Palanisamy, R. Thirunavukarasu, Implications of big data analytics in developing healthcare frameworks–A review, J. King Saud. Univ. Comput. Inf. Sci (2017).

[4] S. Mukherjee, Ovum decision matrix: selecting a business intelligence solution, 2014–15. Ovum, (July 2014), Product code: IT0014–002923, 2014.

[5] T.A. Keahey, Using visualization to understand big data. IBM Business Analytics Advanced Visualisation, 2013.

[6] H. Chen, R.H. Chiang, V.C. Storey, Business intelligence and analytics: From big data to big impact, MIS Q. 36 (4) (2012).

[7] G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K.M. Abbas, R. Sundarsekar, Big data knowledge system in healthcare, Internet of Things and Big Data Technologies for Next Generation Healthcare, Springer, Cham, 2017, pp. 133–157.

[8] B. Ristevski, M. Chen, Big data analytics in medicine and healthcare, J. Integr. Bioinforma. 15 (3) (2018).

[9] G.M. Weber, K.D. Mandl, I.S. Kohane, Finding the missing link for big biomedical data, JAMA 311 (24) (2014) 2479–2480.

[10] J.J. Baker, Activity-Based Costing and Activity-Based Management for Health Care, Jones & Bartlett Learning, 1998.

[11] M. Gietzelt, M. Löpprich, C. Karmen, M. Ganzinger, Models and data sources used in systems medicine, Methods Inf. Med. 55 (02) (2016) 107–113.

[12] C. Lee, Z. Luo, K.Y. Ngiam, M. Zhang, K. Zheng, G. Chen, et al., Big healthcare data analytics: challenges and applications, Handbook of Large-Scale Distributed Computing in Smart Healthcare, Springer, Cham, 2017, pp. 11–41.

[13] P. Balthazar, P. Harri, A. Prater, N.M. Safdar, Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics, J. Am. Coll. Radiol. 15 (3) (2018) 580–586.

[14] L. Hong, M. Luo, R. Wang, P. Lu, W. Lu, L. Lu, Big data in health care: applications and challenges, Data Inf. Manag. 2 (3) (2018) 175–197.

[15] K. Wan, V. Alagar, Characteristics and classification of big data in health care sector, in: 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 1439–1446, IEEE, 2016.

[16] B. Ristevski, S. Savoska, P. Mitrevski, Complex network analysis and big omics data, in: Web proceedings of 11th ICT Innovation Conference 2019, Ohrid, Macedonia, 2019.

[17] M. Househ, A.W. Kushniruk, E.M. Borycki, Big Data, Big Challenges: A Healthcare Perspective: Background, Issues, Solutions and Research Directions, Springer, 2019.

[18] S. Savoska, B. Ristevski, N. Blazheska-Tabakovska, I. Jolevski, Towards Integration Exposome Data and Personal Health Records in the Age of IoT, in: Web proceedings of 11th ICT Innovation Conference 2019, Ohrid, Macedonia, 2019.

[19] Institute of Medicine (US). Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records. Capturing social and behavioral domains and measures in electronic health records: phase 2. National Academies Press, 2014.

[20] C.H. Lee, H.J. Yoon, Medical big data: promise and challenges, Kidney Res. Clin. Pract. 36 (1) (2017) 3.

[21] E.S. Boja, C.R. Kinsinger, H. Rodriguez, P. Srinivas, Integration of omics sciences to advance biology and medicine, 2014.

[22] Y. Wang, N. Hajli, Exploring the path to big data analytics success in healthcare, J. Bus. Res. 70 (2017) 287–299.

[23] Y. Wang, L. Kung, W.Y.C. Wang, C.G. Cegielski, An integrated big data analytics-enabled transformation model: application to health care, Inf. Manag. 55 (1) (2018) 64–79.

[24] B.M. Bai, B.M. Nalini, J. Majumdar, Analysis and detection of diabetes using data mining techniques—a big data application in health care, Emerging Research in Computing, Information, Communication and Applications, Springer, Singapore, 2019, pp. 443–455.

[25] B. Ristevski, S. Loskovska, A survey of clustering algorithms of microarray gene expression data analysis, in: Proceedings of the 10th International Multiconference Information Society – IS 2007, 2007, pp. 52–55, Ljubljana, Slovenia.

[26] J.F. Rodrigues Jr, F.V. Paulovich, M.C. de Oliveira, O.N. de Oliveira Jr, On the convergence of nanotechnology and Big Data analysis for computer-aided diagnosis, Nanomedicine 11 (8) (2016) 959–982.

[27] J. Tan (Ed.), Healthcare Information Systems and Informatics: Research and Practices: Research and Practices, IGI Global, 2008.

[28] D. Schaefer, A. Chandramouly, B.D.D. Owner, I.B. Carmack, K. Kesavamurthy, Delivering self-service BI, data visualization, and Big Data analytics. Intel IT: Business Intelligence, 2013.

[29] Big Digital Data, Analytic Visualization and the Opportunity of Digital Intelligence, 2014, white paper, SAS Institute Inc.

[30] G. Fanjiang, J.H. Grossman, W.D. Compton, P.P. Reid (Eds.), Building a Better Delivery System: A New Engineering/Health Care Partnership, National Academies Press, 2005.

[31] https://www.icd10data.com/ICD10CM/Codes

[32] K.A. Spackman, K.E. Campbell, R.A. Côté, SNOMED RT: a reference terminology for health care, in: Proceedings of the AMIA annual fall symposium, p. 640, American Medical Informatics Association, 1997.

[33] https://www.uicc.org/resources/tnm

[34] https://www.nlm.nih.gov/mesh/meshhome.html

[35] K.A. Wager, F.W. Lee, J.P. Glaser, Health Care Information Systems: a Practical Approach for Health Care Management, John Wiley & Sons, 2017.

[36] Mustra, M., Delac, K., & Grgic, M., Overview of the DICOM standard, in: 2008 50th International Symposium ELMAR, Vol. 1, pp. 39–44, IEEE, 2008.

[37] https://ncpdp.org/Standards-Development/Standards-Information

[38] www.x12.org/x12org/subcommittees/X12N/N0200_X12N_TG2Charter.pdf

[39] http://www.hipaasurvivalguide.com/hipaa-standards.php

[40] https://www.edibasics.com/edi-resources/document-standards/hipaa/

[41] Final Report NHII – Information for Health: A Strategy for Building the National Health Information Infrastructure, 2001. https://ncvhs.hhs.gov/reports/reports

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

[43] F. Bajaber, R. Elshawi, O. Batarfi, A. Altalhi, A. Barnawi, S. Sakr, Big data 2.0 processing systems: taxonomy and open challenges, J. Grid Comput. 14 (3) (2016) 379–405.

[44] N. Mehta, A. Pandit, Concurrence of big data analytics and healthcare: a systematic review, Int. J. Med. Inform. 114 (2018) 57–65.

[45] A. Sajid, H. Abbas, Data privacy in cloud-assisted healthcare systems: state of the art and future challenges, J. Med. Syst. 40 (6) (2016) 155.

[46] N.M. Shrestha, A. Alsadoon, P.W.C. Prasad, L. Hourany, A. Elchouemi, Enhanced e-health framework for security and privacy in healthcare system, in: 2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC), pp. 75–79, IEEE, 2016.

[47] B. Custers, H. Uršič, Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection, Int. Data Priv. Law 6 (1) (2016) 4–15.

[48] K. Abouelmehdi, A. Beni-Hessane, H. Khaloufi, Big healthcare data: preserving security and privacy, J. Big Data 5 (1) (2018) 1.

[49] Z. Savoski, S. Savoska, E-health, need, reality or mith for R. of Macedonia, in: Proceedings/8 th International conference on applied internet and information technologies, Vol. 8, No. 1, pp. 56–59, 2018.

[50] S. Savoska, I. Jolevski, Architectural model of e-health system for support the integrated cross-border services, in: Proceedings of 12th Information Systems and Grid Technologies ISGT 2018, Sofia, Bulgaria, pp. 42–49, 2018.