Abstract

'Big data' is massive amounts of information that can work wonders. It has become a topic of special interest for the past two decades because of a great potential that is hidden in it. Various public and private sector industries generate, store, and analyze big data with an aim to improve the services they provide. In the healthcare industry, various sources for big data include hospital records, medical records of patients, results of medical examinations, and devices that are a part of internet of things. Biomedical research also generates a significant portion of big data relevant to public healthcare. This data requires proper management and analysis in order to derive meaningful information. Otherwise, seeking solution by analyzing big data quickly becomes comparable to finding a needle in the haystack. There are various challenges associated with each step of handling big data which can only be surpassed by using high-end computing solutions for big data analysis. That is why, to provide relevant solutions for improving public health, healthcare providers are required to be fully equipped with appropriate infrastructure to systematically generate and analyze big data. An efficient management, analysis, and interpretation of big data can change the game by opening new avenues for modern healthcare. That is exactly why various industries, including the healthcare industry, are taking vigorous steps to convert this potential into better services and financial advantages. With a strong integration of biomedical and healthcare data, modern healthcare organizations can possibly revolutionize the medical therapies and personalized medicine.

Keywords: Healthcare, Biomedical research, Big data analytics, Internet of things, Personalized medicine, Quantum computing

Introduction

Information has been the key to a better organization and new developments. The more information we have, the more optimally we can organize ourselves to deliver the best outcomes. That is why data collection is an important part for every organization. We can also use this data for the prediction of current trends of certain parameters and future events. As we are becoming more and more aware of this, we have started producing and collecting more data about almost everything by introducing technological developments in this direction. Today, we are facing a situation wherein we are flooded with tons of data from every aspect of our life such as social activities, science, work, health, etc. In a way, we can compare the present situation to a data deluge. The technological advances have helped us in generating more and more data, even to a level

where it has become unmanageable with currently available technologies. This has led to the creation of the term 'big data' to describe data that is large and unmanageable. In order to meet our present and future social needs, we need to develop new strategies to organize this data and derive meaningful information. One such special social need is healthcare. Like every other industry, healthcare organizations are producing data at a tremendous rate that presents many advantages and challenges at the same time. In this review, we discuss about the basics of big data including its management, analysis and future prospects especially in healthcare sector.

The data overload

Every day, people working with various organizations around the world are generating a massive amount of data. The term "digital universe" quantitatively defines such massive amounts of data created, replicated, and consumed in a single year. International Data Corporation (IDC) estimated the approximate size of the digital universe in 2005 to be 130 exabytes (EB). The digital universe in 2017 expanded to about 16,000 EB or 16 zettabytes (ZB). IDC predicted that the digital universe would expand to 40,000 EB by the year 2020. To imagine this size, we would have to assign about 5200 gigabytes (GB) of data to all individuals. This exemplifies the phenomenal speed at which the digital universe is expanding. The internet giants, like Google and Facebook, have been collecting and storing massive amounts of data. For instance, depending on our preferences, Google may store a variety of information including user location, advertisement preferences, list of applications used, internet browsing history, contacts, bookmarks, emails, and other necessary information associated with the user. Similarly, Facebook stores and analyzes more than about 30 petabytes (PB) of user-generated data. Such large amounts of data constitute '*big data*'. Over the past decade, big data has been successfully used by the IT industry to generate critical information that can generate significant revenue.

These observations have become so conspicuous that has eventually led to the birth of a new field of science termed '*Data Science*'. Data science deals with various aspects including data management and analysis, to extract deeper insights for improving the functionality or services of a system (for example, healthcare and transport system). Additionally, with the availability of some of the most creative and meaningful ways to visualize big data post-analysis, it has become easier to understand the functioning of any complex system. As a large section of society is becoming aware of, and involved in generating big data, it has become necessary to define what big data is. Therefore, in this review, we attempt to provide details on the impact of big data in the transformation of global healthcare sector and its impact on our daily lives.

Defining big data

As the name suggests, 'big data' represents large amounts of data that is unmanageable using traditional software or internet-based platforms. It surpasses the traditionally used amount of storage, processing and analytical power. Even though a number of definitions for big data exist, the most popular and well-accepted definition was given by Douglas Laney. Laney observed that (big) data was growing in three different dimensions namely, volume, velocity and variety (known as the 3 Vs) [1]. The 'big' part of big data is indicative of its large volume. In addition to volume, the big data description also includes

velocity and variety. Velocity indicates the speed or rate of data collection and making it accessible for further analysis; while, variety remarks on the different types of organized and unorganized data that any firm or system can collect, such as transaction-level data, video, audio, text or log files. These three Vs have become the standard definition of big data. Although, other people have added several other Vs to this definition [2], the most accepted 4th V remains 'veracity'.

The term "*big data*" has become extremely popular across the globe in recent years. Almost every sector of research, whether it relates to industry or academics, is generating and analyzing *big data* for various purposes. The most challenging task regarding this huge heap of data that can be organized and unorganized, is its management. Given the fact that big data is unmanageable using the traditional software, we need technically advanced applications and software that can utilize fast and cost-efficient high-end computational power for such tasks. Implementation of artificial intelligence (AI) algorithms and novel fusion algorithms would be necessary to make sense from this large amount of data. Indeed, it would be a great feat to achieve automated decision-making by the implementation of machine learning (ML) methods like neural networks and other AI techniques. However, in absence of appropriate software and hardware support, big data can be quite hazy. We need to develop better techniques to handle this 'endless sea' of data and smart web applications for efficient analysis to gain workable insights. With proper storage and analytical tools in hand, the information and insights derived from big data can make the critical social infrastructure components and services (like healthcare, safety or transportation) more aware, interactive and efficient [3]. In addition, visualization of big data in a user-friendly manner will be a critical factor for societal development.

Healthcare as a big-data repository

Healthcare is a multi-dimensional system established with the sole aim for the prevention, diagnosis, and treatment of health-related issues or impairments in human beings. The major components of a healthcare system are the health professionals (physicians or nurses), health facilities (clinics, hospitals for delivering medicines and other diagnosis or treatment technologies), and a financing institution supporting the former two. The health professionals belong to various health sectors like dentistry, medicine, midwifery, nursing, psychology, physiotherapy, and many others. Healthcare is required at several levels depending on the urgency of situation. Professionals serve it as the first point of consultation (for primary care), acute care requiring skilled professionals (secondary care), advanced medical investigation and treatment (tertiary care) and highly uncommon diagnostic or surgical procedures (quaternary care). At all these levels, the health professionals are responsible for different kinds of information such as patient's medical history (diagnosis and prescriptions related data), medical and clinical data (like data from imaging and laboratory examinations), and other private or personal medical data. Previously, the common practice to store such medical records for a patient was in the form of either handwritten notes or typed reports [4]. Even the results from a medical examination were stored in a paper file system. In fact, this practice is really old, with the oldest case reports existing on a papyrus text from Egypt that dates back to 1600 BC [5].

In Stanley Reiser's words, the clinical case records freeze the episode of illness as a story in which patient, family and the doctor are a part of the plot" [6].

With the advent of computer systems and its potential, the digitization of all clinical exams and medical records in the healthcare systems has become a standard and widely adopted practice nowadays. In 2003, a division of the National Academies of Sciences, Engineering, and Medicine known as Institute of Medicine chose the term "*electronic health records*" to represent records maintained for improving the health care sector towards the benefit of patients and clinicians. Electronic health records (EHR) as defined by Murphy, Hanken and Waters are computerized medical records for patients any information relating to the past, present or future physical/mental health or condition of an individual which resides in electronic system(s) used to capture, transmit, receive, store, retrieve, link and manipulate multimedia data for the primary purpose of providing healthcare and health-related services" [7].

Electronic health records

It is important to note that the National Institutes of Health (NIH) recently announced the "All of Us" initiative (https://allofus.nih.gov/) that aims to collect one million or more patients' data such as EHR, including medical imaging, socio-behavioral, and environmental data over the next few years. EHRs have introduced many advantages for handling modern healthcare related data. Below, we describe some of the characteristic advantages of using EHRs. The first advantage of EHRs is that healthcare professionals have an improved access to the entire medical history of a patient. The information includes medical diagnoses, prescriptions, data related to known allergies, demographics, clinical narratives, and the results obtained from various laboratory tests. The recognition and treatment of medical conditions thus is time efficient due to a reduction in the lag time of previous test results. With time we have observed a significant decrease in the redundant and additional examinations, lost orders and ambiguities caused by illegible handwriting, and an improved care coordination between multiple healthcare providers. Overcoming such logistical errors has led to reduction in the number of drug allergies by reducing errors in medication dose and frequency. Healthcare professionals have also found access over web based and electronic platforms to improve their medical practices significantly using automatic reminders and prompts regarding vaccinations, abnormal laboratory results, cancer screening, and other periodic checkups. There would be a greater continuity of care and timely interventions by facilitating communication among multiple healthcare providers and patients. They can be associated to electronic authorization and immediate insurance approvals due to less paperwork. EHRs enable faster data retrieval and facilitate reporting of key healthcare quality indicators to the organizations, and also improve public health surveillance by immediate reporting of disease outbreaks. EHRs also provide relevant data regarding the quality of care for the beneficiaries of employee health insurance programs and can help control the increasing costs of health insurance benefits. Finally, EHRs can reduce or absolutely eliminate delays and confusion in the billing and claims management area. The EHRs and internet together help provide access to millions of health-related medical information critical for patient life.

Fig. 1 Workflow of Big data Analytics. Data warehouses store massive amounts of data generated from various sources. This data is processed using analytic pipelines to obtain smarter and affordable healthcare options

Digitization of healthcare and big data

Similar to EHR, an electronic medical record (EMR) stores the standard medical and clinical data gathered from the patients. EHRs, EMRs, personal health record (PHR), medical practice management software (MPM), and many other healthcare data components collectively have the potential to improve the quality, service efficiency, and costs of healthcare along with the reduction of medical errors. The big data in healthcare includes the healthcare payer-provider data (such as EMRs, pharmacy prescription, and insurance records) along with the genomics-driven experiments (such as genotyping, gene expression data) and other data acquired from the smart web of internet of things (IoT) (Fig. 1). The adoption of EHRs was slow at the beginning of the 21st century however it has grown substantially after 2009 [7, 8]. The management and usage of such healthcare data has been increasingly dependent on information technology. The development and usage of wellness monitoring devices and related software that can generate alerts and share the health related data of a patient with the respective health care providers has gained momentum, especially in establishing a real-time biomedical and health monitoring system. These devices are generating a huge amount of data that can be analyzed to provide real-time clinical or medical care [9]. The use of big data from healthcare shows promise for improving health outcomes and controlling costs.

Big data in biomedical research

A biological system, such as a human cell, exhibits molecular and physical events of complex interplay. In order to understand interdependencies of various components and events of such a complex system, a biomedical or biological experiment usually gathers data on a smaller and/or simpler component. Consequently, it requires multiple simplified experiments to generate a wide map of a given biological phenomenon of interest. This indicates that more the data we have, the better we understand the biological processes. With this idea, modern techniques have evolved at a great pace. For instance, one can imagine the amount of data generated since the integration of efficient technologies like next-generation sequencing (NGS) and Genome wide association studies (GWAS) to decode human genetics. NGS-based data provides information at depths that were previously inaccessible and takes the experimental scenario to a completely

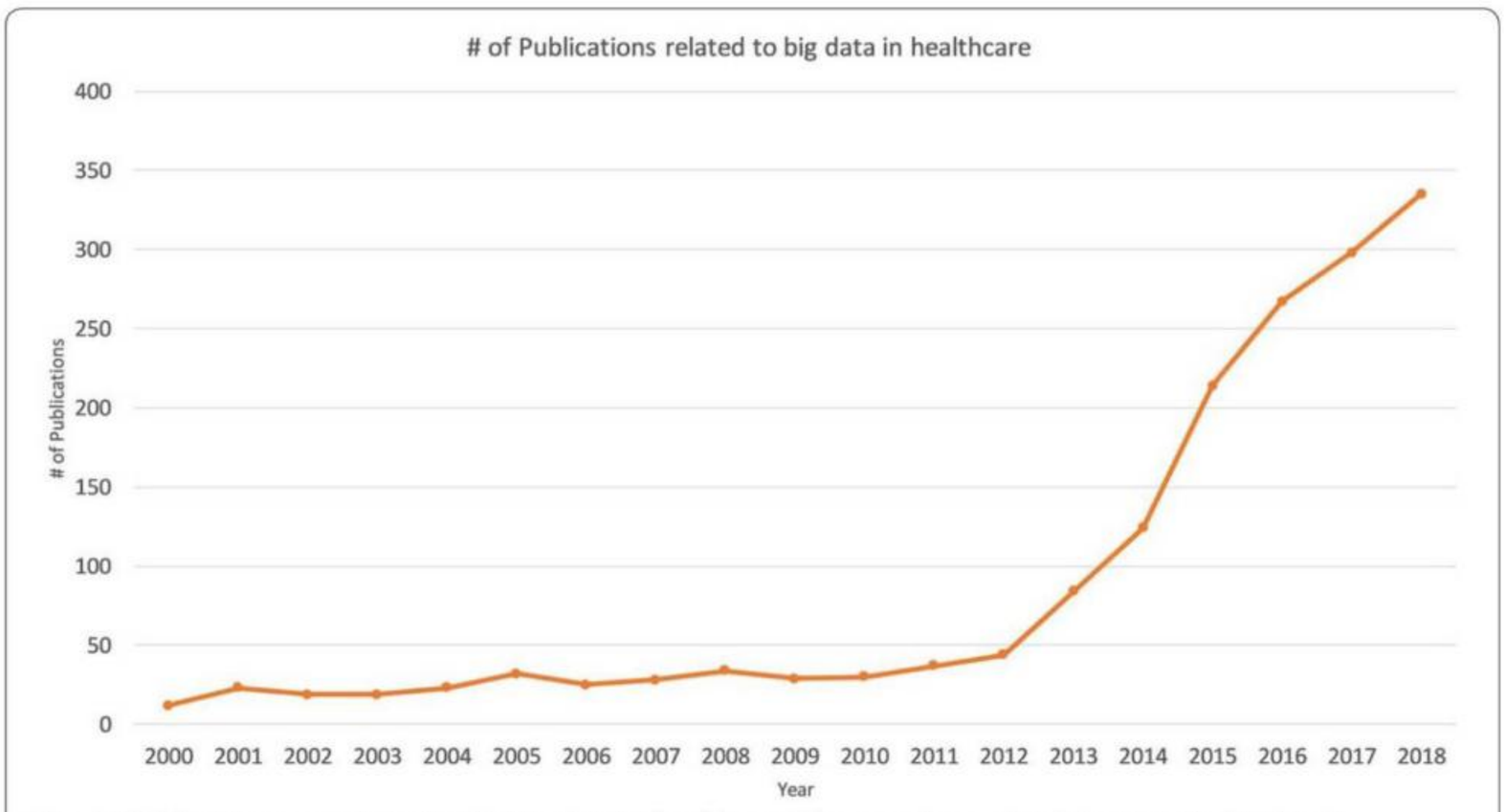

Fig. 2 Publications associated with big data in healthcare. The numbers of publications in PubMed are plotted by year

new dimension. It has increased the resolution at which we observe or record biological events associated with specific diseases in a real time manner. The idea that large amounts of data can provide us a good amount of information that often remains unidentified or hidden in smaller experimental methods has ushered-in the '*-omics*' era. The '*omics*' discipline has witnessed significant progress as instead of studying a single '*gene*' scientists can now study the whole '*genome*' of an organism in '*genomics*' studies within a given amount of time. Similarly, instead of studying the expression or '*transcription*' of single gene, we can now study the expression of all the genes or the entire '*transcriptome*' of an organism under '*transcriptomics*' studies. Each of these individual experiments generate a large amount of data with more depth of information than ever before. Yet, this depth and resolution might be insufficient to provide all the details required to explain a particular mechanism or event. Therefore, one usually finds oneself analyzing a large amount of data obtained from multiple experiments to gain novel insights. This fact is supported by a continuous rise in the number of publications regarding big data in healthcare (Fig. 2). Analysis of such big data from medical and healthcare systems can be of immense help in providing novel strategies for healthcare. The latest technological developments in data generation, collection and analysis, have raised expectations towards a revolution in the field of personalized medicine in near future.

Big data from omics studies

NGS has greatly simplified the sequencing and decreased the costs for generating whole genome sequence data. The cost of complete genome sequencing has fallen from millions to a couple of thousand dollars [10]. NGS technology has resulted in an increased volume of biomedical data that comes from genomic and transcriptomic studies. According to an estimate, the number of human genomes sequenced by 2025 could be between 100 million to 2 billion [11]. Combining the genomic and transcriptomic data with proteomic and metabolomic data can greatly enhance our knowledge about the individual profile of a patient—an approach often ascribed as "individual,

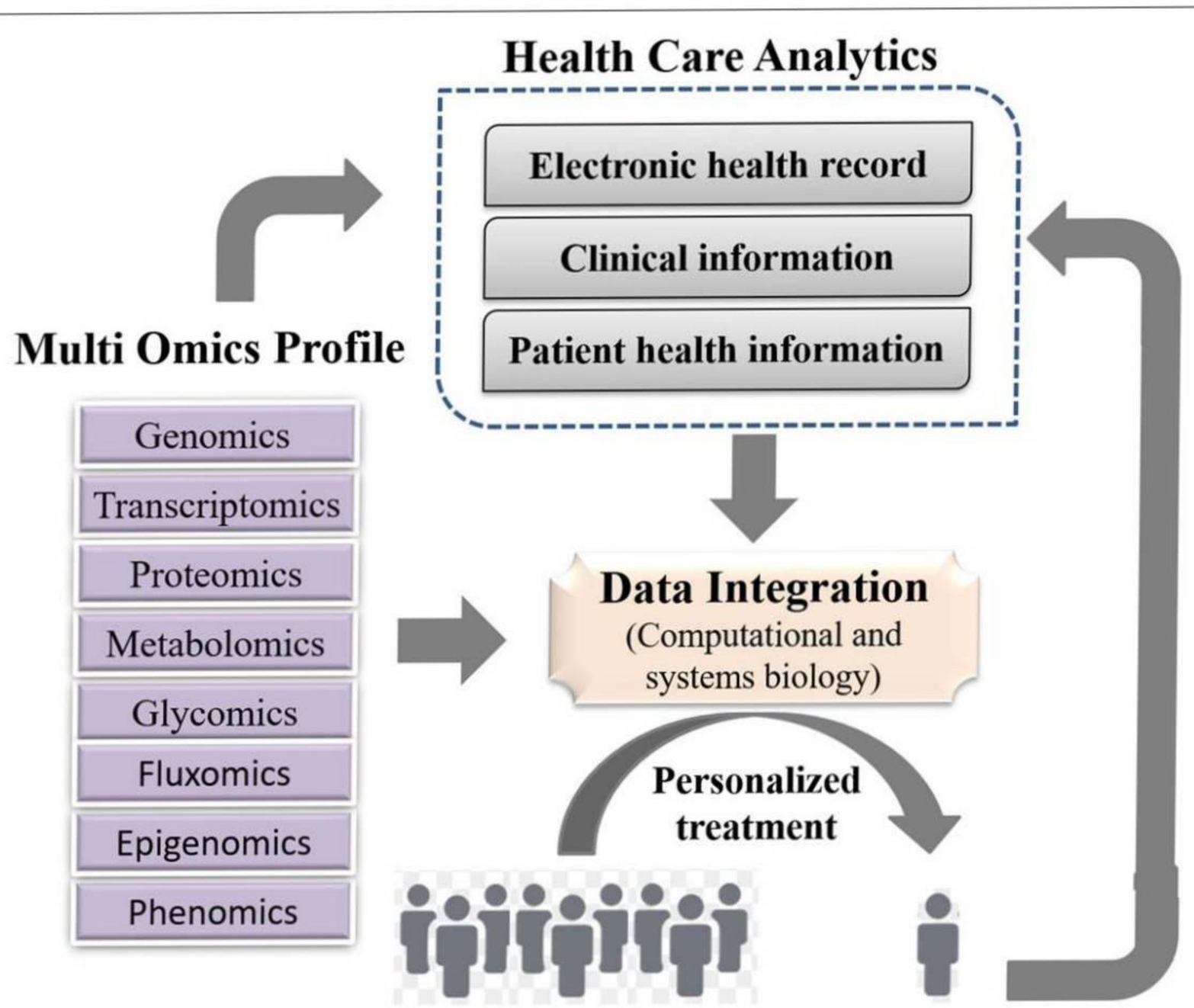

Fig. 3 A framework for integrating omics data and health care analytics to promote personalized treatment

personalized or precision health care". Systematic and integrative analysis of omics data in conjugation with healthcare analytics can help design better treatment strategies towards precision and personalized medicine (Fig. 3). The genomics-driven experiments e.g., genotyping, gene expression, and NGS-based studies are the major source of big data in biomedical healthcare along with EMRs, pharmacy prescription information, and insurance records. Healthcare requires a strong integration of such biomedical data from various sources to provide better treatments and patient care. These prospects are so exciting that even though genomic data from patients would have many variables to be accounted, yet commercial organizations are already using human genome data to help the providers in making personalized medical decisions. This might turn out to be a game-changer in future medicine and health.