Studies in Big Data 66

Anand J. Kulkarni · Patrick Siarry ·
Pramod Kumar Singh · Ajith Abraham ·
Mengjie Zhang · Albert Zomaya ·
Fazle Baki *Editors*

Big Data Analytics in Healthcare

Springer

Studies in Big Data

Volume 66

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series “Studies in Big Data” (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and other. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence including neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and Operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

** Indexing: The books of this series are submitted to ISI Web of Science, DBLP, Ulrichs, MathSciNet, Current Mathematical Publications, Mathematical Reviews, Zentralblatt Math: MetaPress and Springerlink.

More information about this series at http://www.springer.com/series/11970

Anand J. Kulkarni · Patrick Siarry ·
Pramod Kumar Singh · Ajith Abraham ·
Mengjie Zhang · Albert Zomaya ·
Fazle Baki
Editors

Big Data Analytics in Healthcare

Springer

Editors
Anand J. Kulkarni
Symbiosis Institute of Technology
Symbiosis International
(Deemed University)
Pune, Maharashtra, India

Pramod Kumar Singh
ABV-Indian Institute of Information
Technology and Management Gwalior
Gwalior, Madhya Pradesh, India

Mengjie Zhang
School of Engineering and Computer
Science
Victoria University of Wellington
Kelburn, New Zealand

Fazle Baki
Odette School of Business
University of Windsor
Windsor, ON, Canada

Patrick Siarry
Université Paris-Est Créteil Val de Marne
Créteil, France

Ajith Abraham
Scientific Network for Innovation
and Research Excellence
Machine Intelligence Research Labs
(MIR Labs)
Auburn, WA, USA

Albert Zomaya
School of Computer Science
University of Sydney
Sydney, Australia

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-3-030-31671-6 ISBN 978-3-030-31672-3 (eBook)
https://doi.org/10.1007/978-3-030-31672-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The term big data can be described as the structured and unstructured data being generated from a variety of sources in huge volume and in unprecedented real-time speed. Such data becomes important when the associated analysis leads to better decision making, strategic and policy moves of the organization. The storage, processing, and analysis become critical when dealing with huge variegated data involving numerical and text documents, video, audio, pictures, etc., being generated from the sources of different modalities. The complexity of the source and associated generated data further poses challenges to correlate the relationships, generate patterns, establishing reliability, etc. As the focus of healthcare industry has now shifted from clinical-centric to patient-centric model, this necessitated efficient storage and analysis of the existing medical records and the records being generated in the numerical forms, prescriptions, graphs, images, videos, interviews, etc. There are several technical, computational, organizational, and ethical challenges being faced by the healthcare industry as well as the governments. The big data analytics in healthcare is becoming a revolution in technical as well as societal well-being view point. This edited volume intends to provide a platform to the state-of-the-art discussion on various issues and aspects of the implementation, associated testing, validation, and application of big data related to healthcare domain. The volume also aims to discuss multifaceted and state-of-the-art literature survey of healthcare data, their modalities, complexities, and methodologies along with a complete mathematical formulation.

Every chapter submitted to the volume has been critically evaluated by at least two expert reviewers. The critical suggestions by the reviewers helped and influenced the authors of the individual chapter to enrich the quality in terms of experimentation, performance evaluation, representation, etc. The volume may serve as a complete reference for big data in healthcare.

The volume is divided into two parts. The challenges, opportunities associated with the big data implementation, along with the big data platforms and tools in healthcare domain are discussed in first part. The mathematical modeling of the

healthcare problems, their solutions, existing and futuristic big data applications and platforms have been discussed in Part II of the volume. The contributions of every chapter are discussed below in detail.

First Part: Challenges, Opportunities, Platforms, and Tools of Big Data in Healthcare

Chapter "Big Data Analytics and Its Benefits in Healthcare" by Kumar et al. highlighted the limitations of traditional database management system when dealing with the unstructured data generated in real time. A critical review of the life cycle of the big data and the issues associated with its implementation such as security, dynamic classification, storage, modeling, and modalities have been discussed in detail. It underscores the need of an effective data storing and analysis system which can handle structured as well as unstructured data associated with the healthcare domain. The discussion is extended to an overview of prominent characteristics and components of fault-tolerant Hadoop, its module such as YARN, Hadoop Distributed File System, Hadoop-MapReduce for parallel processing of large datasets, etc. In addition, a critical analysis of possible applications of the big data in healthcare areas is discussed. The major examples are relevant to electronic health record keeping, real-time warning for clinical decision support system, predictive analysis, practicing telemedicine, etc.

In Chapter "Elements of Healthcare Big Data Analytics," Mehta et al. discussed the shift of the healthcare industry from clinical-centric to patient-centric model which led to the need of associated services at affordable price. The major contribution of the chapter is to highlight the systemic challenges that the healthcare organization faces to embrace the big data techniques. The challenges are classified into four levels. The data- and process-related challenges are associated with the processing of the structured data such as patient demographic details and unstructured data such as clinical notes, diagnostic images, MRI scans, and videos. It is also associated with the compatibility of the large volume of data generating from the devices and sensors of different modalities. In addition, it is also associated with data integration, storage, extraction of useful information, redundancy, and security. The manpower-related challenges are referred to the talent deficit, retention, and competition. The domain-related challenges referred to the development of efficient algorithms, adoption of novel technologies, and interpretability of results and associated human intervention and associated decisions. The authors also highlighted the managerial challenges, such as overcoming technology gap, identification of right tools, training, organizational resistance, and accepting the transparency of the data. The key elements of effective integration of big data analytics into healthcare along with foundational steps for beginning a big data analytics program within an organization are also suggested by the authors. The development and application of the preprocessing techniques of the heterogeneous

data, extraction algorithms and analytical techniques to mine the value from the data, and seamless leveraging of the enriched data across the organization are of foremost importance. The use of specially developed tools for network security and health-related data protection, along with the vulnerability management, validation of corrective actions and associated policies, are among the key necessities. The policies refer to the strategic initiatives, guidelines, iterative adoption and collaborative roadmap, planning and availability of the human resource and associated roles and responsibilities.

In Chapter "Big Data in Supply Chain Management and Medicinal Domain," Nargundkar and Kulkarni covered the significance and potential of big data techniques in medicinal industry and associated supply chain activities. The big data platforms used in supply chain associated with medicinal domain along with the prominent tool of NoSQL for processing real-time and interactive data are described in very detail. The overall process of big data analytics from data generation to visualization is exemplified with reference to the medicinal domain. Importantly, an upcoming trend of big data analytics with wearable or implanted sensors is explicated. This has reference to an architecture implementing Internet of things (IoT) to store and process huge amount of wearable sensor data being generated in real time. It provides a concise review of data collection and storage, computing and classification as well.

In Chapter "A Review of Big Data and Its Applications in Healthcare and Public Sector," Shastri and Deshpande discussed the applications of big data technologies in the fields of healthcare and public sector with focus on preventive healthcare planning and predictive analytics. The benefits for the healthcare domain discussed are enhancement in the capability of taking informed decisions based on the analysis of the historical medical data, reduction in the healthcare budget expenditure, etc. The chapter also discusses the opportunities and benefits of adopting the big data technologies in public sector, such as fraud detection, preventive healthcare and prevention of epidemics, education, boosting transparency, urban management, sentiment analysis for prediction of response to government policies, and crime prediction based on historical and real-time data. In addition, the chapter provides a rich reference to the Hadoop architecture components, viz. scalable Hadoop Distributed File System (HDFS) for distributed data storage, MapReduce for processing. The prominent and essential characteristics of these components along with the working framework have been discussed. In addition, the components such as Hive, Pig, Sqoop, Mahout, Hbase, Oozie, Zookeeper, and Cassandra have also been discussed in brief. The Apache Spark which is comparatively more efficient in iterative machine learning and interactive querying jobs is analyzed using prominent examples. Its framework along with its components and comparison with MapReduce is also discussed in detail.

Healthcare management around the world concentrates on patient-centered model rather than disease-centered; it also has approach of value-based healthcare delivery model instead volume-based. The big data processes and analysis can fill

the gap between healthcare costs and value-based outcome which is the focus of Chapter "Big Data in Healthcare: Technical Challenges and Opportunities," by Kakandikar and Nandedkar. It insisted on the necessity of big data techniques to be deployed in dealing with the overwhelming unstructured medical data as several countries have resorted to the digitization of the records. The author has highlighted four major aspects of value of the data, viz. living, care, provider, and innovation. Furthermore, the big data analysis approaches such as prescriptive analysis, diagnostic analysis for revealing hidden patterns, probable root causes, and descriptive analysis for fragmentation of the data have also been discussed. Apart from the general challenges the critical technical challenges, such as data transformation, complex event processing, multiple character complexity, semantic and contextual data handling, data replication, migration, loss and redundancy are also highlighted. This discussion is further extended to the big data applications in healthcare domains such as providing personal healthcare, fraud detection and prevention, pattern and trend analysis and associated prediction of epidemics, tailored diagnosis, and treatment decision support. Several software platforms for processing of the big data are also briefed in the chapter.

In Chapter "Innovative mHealth Solution for Reliable Patient Data Empowering Rural Healthcare in Developing Countries," Rajasekera et al. reviewed the general problems associated with collection of health data from rural areas where large percentage of population of developing countries is concentrated. The review highlighted that the application areas of mobile health (mHealth) may depend on the local characteristics and preferences of a particular country. In addition, authors insisted upon availability of frontline manpower resource, timely, credible and consistent patient data availability as the pivotal and necessary factors in successful applications of big data in mHealth in the rural areas. Authors presented several limitations and associated challenges being faced by the frontline manpower resource on mHealth platform. The chapter describes associated solution in the form of a case study on N+Care mobile application which can handle a variety of unstructured data such as photographs, prescriptions, and test details. The second case study from India referred to as Anywhere Anytime Access (A3) remote monitoring technology provides valuable insights into remote patient data monitoring system. The importance of such technology is underscored in relevance to the validation of the credibility as well as making the data available in timely manner.

Second Part: Mathematical Modeling and Solutions, Big Data Applications, and Platforms

The contribution of Chapter "Hospital Surgery Scheduling Under Uncertainty Using Multiobjective Evolutionary Algorithms" by Ripon and Nyman is motivated from narrowing down the gap between existing evolutionary approaches to machine

scheduling problems and their practical applications to real-world hospital surgery scheduling problems. Importantly, a novel variation of the surgery admission planning problem is formulated along with the development of evolutionary mechanisms to solve it with contemporary multiobjective evolutionary algorithms. The algorithms chosen are Strength Pareto Evolutionary Algorithm 2 (SPEA2) and the Non-domination Sorting Genetic Algorithm II (NSGA II). The chapter theoretically and mathematically details a complete scheduling process using Master Surgery Schedule (MSS) addressing two sources of uncertainty, viz. patient arrival uncertainty and activity duration uncertainty. The solution approaches are validated on a variety of huge test data characterized by number of rooms, days, and number of patients. The chapter provides a measure of uncertainty along with the degree of conflicts between the objectives, i.e., choice between scheduling two surgeons to work overtime in the same operating room and reserving overtime capacity for a single surgeon in two operating rooms.

The necessity of processing huge neuronal behavior data available over different human communities is discussed in Chapter "Big Data in Electroencephalography Analysis" by Yedurkar and Metkar. The work is intended to analyze different scales and dynamics of neurons which are partially responsible for logical reasoning capabilities and inclinations of the individuals. Authors have highlighted that the huge data generated from electroencephalogram (EEG) is patient specific and diversified as well as in the form of non-stationary signal, epileptic and non-epileptic patterns. The traditional data handling approaches have several limitations handling the variegated signal volume generated in real time apart from the issue of data storage for further processing. A mathematical model of exponentially big volume of data being streamed by the EEG is also exemplified along with the need of further utilization of such continuous data. Besides the big data approach to healthcare problem, the chapter also briefly covers importance of the EEG as a critical tool in neuroscience.

In Chapter "Big Data Analytics in Healthcare Using Spreadsheets," Iyengar et al. discussed big data analytics, its need and methods with special reference to the healthcare industry which may help practitioners and policy-makers to develop strategies for healthcare systems betterment. The chapter in detail discusses the analysis of big data and its subcomponent, viz. structured data type such as simple numeric data, semi-structured, and unstructured data types such as text and images. The chapter critically reviews the tools for big data analytics such as Hadoop and its constituents, spreadsheets, and add-Ins. The rationale of using spreadsheet in the

current market scenario along with its components such as Vlookup and Hlookup, pivot table, ANOVA, and Fourier analysis has been discussed along with several real-world examples.

Pune, India — Anand J. Kulkarni
Paris, France — Patrick Siarry
Gwalior, India — Pramod Kumar Singh
Auburn, USA — Ajith Abraham
Wellington, New Zealand — Mengjie Zhang
Sydney, Australia — Albert Zomaya
Windsor, Canada — Fazle Baki

Acknowledgements We are grateful to the reviewers of the volume for their valuable time and efforts in critically reviewing the chapters. Their critical and constructive reviews certainly have helped in the enrichment of every chapter. The editors would like to thank Dr. Thomas Ditzinger Springer Nature Switzerland AG, for the editorial assistance and cooperation to produce this important scientific work. We hope that the readers will find this volume useful and valuable to their research.

Contents

Challenges, Opportunities, Platforms and Tools of Big Data in Healthcare

Big Data Analytics and Its Benefits in Healthcare 3
Yogesh Kumar, Kanika Sood, Surabhi Kaul and Richa Vasuja

Elements of Healthcare Big Data Analytics 23
Nishita Mehta, Anil Pandit and Meenal Kulkarni

Big Data in Supply Chain Management and Medicinal Domain 45
Aniket Nargundkar and Anand J. Kulkarni

A Review of Big Data and Its Applications in Healthcare and Public Sector 55
Apoorva Shastri and Mihir Deshpande

Big Data in Healthcare: Technical Challenges and Opportunities 67
Ganesh M. Kakandikar and Vilas M. Nandedkar

Innovative mHealth Solution for Reliable Patient Data Empowering Rural HealthCare in Developing Countries 83
Jay Rajasekera, Aditi Vivek Mishal and Yoshie Mori

Mathematical Modeling and Solutions, Big Data Applications, and Platforms

Hospital Surgery Scheduling Under Uncertainty Using Multiobjective Evolutionary Algorithms 107
Kazi Shah Nawaz Ripon and Jacob Henrik Nyman

Big Data in Electroencephalography Analysis 143
Dhanalekshmi P. Yedurkar and Shilpa P. Metkar

Big Data Analytics in Healthcare Using Spreadsheets 155
Samaya Pillai Iyengar, Haridas Acharya and Manik Kadam

Challenges, Opportunities, Platforms and Tools of Big Data in Healthcare

Big Data Analytics and Its Benefits in Healthcare

Yogesh Kumar, Kanika Sood, Surabhi Kaul and Richa Vasuja

Abstract The main challenging task in real world is to collect huge amount of data from different sources in different format. Traditional database only helps in storing small amount of information. When the data become unstructured, it becomes difficult for the traditional database management system to extract knowledge out of it. For making an effective system, it becomes necessary to handle both structured and unstructured data. Here technology called big data solves this problem because it can extract the knowledge from structured as well as unstructured data. The purpose of big data is to collect the data that is gathered from different sources and then store this collected data in some common place. After then distributed File System is must for distributed storage and fault tolerance. Here Apache Hadoop is commonly being used these days. Another concept called Map reduce is a programming model that is most widely used in Hadoop for processing large amount of data quickly. In this paper big data are introduced in detail. Hadoop is used to process data in big data. There are many parts of Hadoop such as Hadoop common: these are the libraries of java and other modules which are included in Hadoop. Hadoop YARN which is used for cluster resource management and for job scheduling. Hadoop Distributed File System HDFS that help in providing greater amounts of access to application information and Hadoop MapReduc which is YARN based system which helps in processing parallel large data sets. The main purpose of the chapter is to use the function of big data in the fields of healthcare. Various examples as well as applications related to healthcare are discussed in this chapter. Various challenges related to big data analytics are discussed in this chapter.

Y. Kumar (✉)
CSE, Chandigarh Engineering College, Landran, Mohali, India
e-mail: Yogesh.arora10744@gmail.com

K. Sood · S. Kaul
CSE, IET Bhaddal, District Ropar, India
e-mail: kanika04sood@gmail.com

S. Kaul
e-mail: Surabhi93kaul@gmail.com

R. Vasuja
CSE, Chandigarh University Gharuan, Mohali, India
e-mail: richavasuja@gmail.com

A. J. Kulkarni et al. (eds.), *Big Data Analytics in Healthcare*,
Studies in Big Data 66, https://doi.org/10.1007/978-3-030-31672-3_1

Keywords HDFS · Big data · Hadoop · Healthcare · Map reduce

1 Introduction

Large volume of data has been generated from various sources like record keeping, patient related data in healthcare industry. Each data should be digitized in today's digital world. For getting best in new challenges the data should be analysed effectively with minimum cost. From government sector also large volume of data is generated every day. So a technology is needed that will take care of this large data set on real time. So it will help citizens for getting better results. Big data helps in providing valuable decisions by data patterns and relationship among different data set with the help of various machine learning algorithms. Likewise oil is very essential, data is also considered as much important. But unprocessed data cannot be used and is not useful. With the help of various analytical methods important information can be mined from the data. According to Hermon [1], big data can bring a lot new revolution in industry of healthcare. The large volume of data stored can help in providing better results in healthcare. Big data analytics help in processing large amount of data parallel and also help in providing solution to various hidden problems. Minimized cost can be achieved will using big data analytics for processing large volume of data. Any disease which may be occurred and cured in any part of the world, prediction for that disease can be done capably. In big data exploration, diverse statistical approaches, data mining and machine learning approaches can be implemented. Healthcare area has a lot of opportunities for providing well cure for diseases using various analytical outfits [2].

2 Introduction to Big Data

Big data may have following assets like high velocity, high volume and high variety which involve processing that help in decision making and process optimization. Big data is an expression that means large volume of data that can be either structured or unstructured and whose processing is very difficult using traditional database [3]. Its data set can be categorized with five definitions i.e. variety, velocity, volume, value and veracity as shown in Fig. 1.

Healthcare data which may include EMR reports, medical images etc. are broadly divided into structured and unstructured data. The large volume of data helps in adding the value and improving the quality of healthcare by inventive analysis as well as refining patient care. The huge amount of data related to healthcare can be computed through distributed processing with the help of cloud centers and big data [1].

Fig. 1 Big data 5 V's [2]

According to researchers more 5 important characteristics have been started which total form 10V's of big data. The five additional characteristics are Variability, Validity, Vulnerability, Volatility and Visualization. Healthcare data which may include EMR reports, medical images etc. are broadly divided into structured and unstructured data [4]. The large volume of data helps in adding the value and improving the quality of healthcare by inventive analysis as well as refining patient care. The huge amount of data related to healthcare can be computed through distributed processing with the help of cloud centers and big data.

According to researchers more 5 important characteristics have been started which total form 10V's of big data. The five additional characteristics are Variability, Validity, Vulnerability, Volatility and Visualization.

2.1 Processing of Big Data

Big data's processing can be done in four layers as shown in Fig. 2. The main challenging task is to collect huge amount of data from different sources in different format. As the data is unstructured, it becomes difficult for traditional database management system to extract knowledge out of it but big data solve this problem because it may help in extracting the knowledge from structured, semi-structured and unstructured data.

First step is to collect the data that is gathered from different sources and then store this collected data in some common place.

To provide distributed File System (HDFS) for distributed storage and fault tolerance Apache Hadoop is commonly being used these days. Map reduce is a programming model used in Hadoop for processing large amount of data quickly. In map reduce datasets are divided into two subsets which are testing and training. Machine

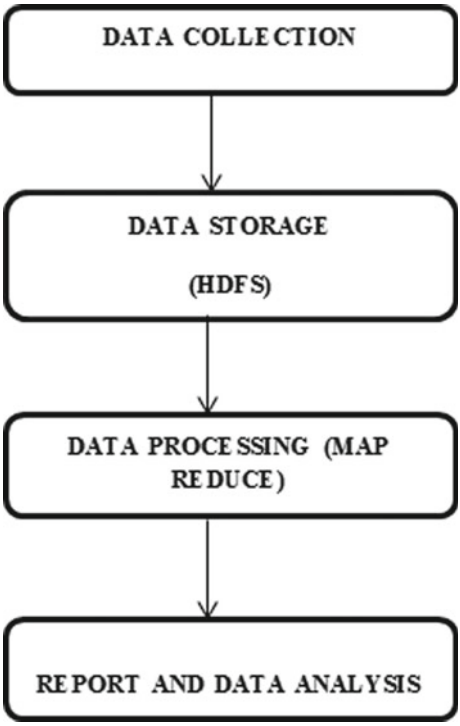


Fig. 2 Big data processing [5]

learning algorithm can be applied for achieving quick investigation on input data and create the info which can be used for producing information in processing layer.

2.2 Introduction to Hadoop

Hadoop is an Apache open source frame that is written in java language. Because as it is written in java this feature allows various distributed processing of data set across various computers using simple programming models. A Hadoop framework works in a way where it helps in providing distributed storage and computation of various computers. Hadoop structure is designed to scale up from single machine to millions of machines, each offering its local storage and computation [6].

2.3 Hadoop Architecture

Four modules are included in Hadoop framework which is discussed below.

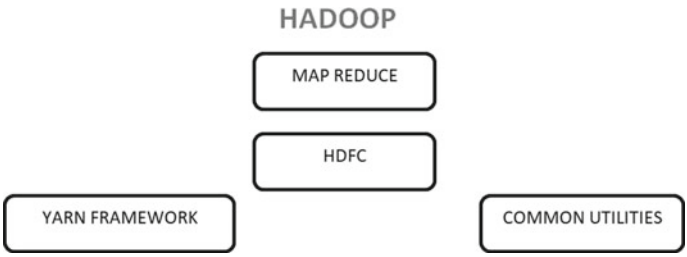


Fig. 3 Hadoop architecture [4]

Hadoop Common: these are the libraries of java and other modules which are included in Hadoop. These libraries may contain various file system and operating system abstractions. Also various java scripts and java files which are mandatory for starting the Hadoop are included in this library.

Hadoop YARN: used for cluster resource management and for job scheduling.

Hadoop Distributed File System HDFS: help in providing greater amount of access to application information.

Hadoop Map Reduce: it is YARN based system which helps in processing parallel large data sets.

Figure 3 depicts four components that are presented in Hadoop framework.

Map Reduce: For processing large amount of data in parallel cluster framework, hadoop Map Reduce is being used for easily writing such applications which are reliable and much fault-tolerant. Map reduce may consists of two different tasks that Hadoop programs perform.

The Map Task: the very first task in hadoop program which take the input data and convert that data into set of data, in which single element are fragmented into tuples keys or pairs.

The Reduce Task: it takes input as the output of map task and combines the data into smaller set. The reduce task is always performed after the completion of map task. The input and output are to be stored in file system. The hadoop framework also takes cares of scheduled tasks, monitors them and executes them again if failure occurs [4].

Map reduce framework consists of single master known as Job tracker and one slave known as Task tracker. Various responsibilities of master are resource management, tracking availability of resource, scheduling jobs to slaves and re executing failed task. The responsibilities of slave are to obey the directions and commands ordered by master and provide report to master. The single point of failure in map reduce is job tracker. It means if job tracker fails, all other running jobs that time will halt.

Hadoop Distributed File System: Hadoop is very economical; it can work with any distributed file system like HFTP, FS etc. but the common file system with which hadoop works is Hadoop Distributed File System HDFS.

The HDFS is established from Google File System which helps in providing a distributed file system which is aimed for running on thousands of computers. Master/slave architecture is being used by HDFS. In which master is name node that manages file system and slave consists of data node that help in storing of data. Various operations like read and write operations, creation of block, deletion of block all are governed by data nodes.

2.4 How Does Hadoop Works?

Step 1: For required process users submit their jobs as "job client" which involves following items:

1. The input and output files are located in a distributed file system environment.
2. The java classes used in hadoop make use of map and reduce functions.
3. The job client contains different parameters for the specific job.

Step 2: The job client then submits it jobs and provides alignment to the job tracker which has the responsibility of taking care of different slaves and monitoring their work.
Step 3: As per Map reduce implementation, task tracker executes each task and output is to be stored in output files of file system.

2.5 Advantage of Hadoop

Hadoop does not depend on its hardware to provide availability and fault tolerance infect it has its own libraries that can detect and handle the failure of its own.

Hadoop works with the interruption of adding and removing of clusters.

Hadoop is compatible with any platform.

Quick testing and working in distributed system can be done using hadoop.

2.6 HDFS in Healthcare

The large data sets can be handled very effectively by hadoop. Figure 4 shows the working between client and name node for data processing [7]. Firstly name node connects with job tracker and assigns them with jobs given by client. After that Map Reduce will analysis the data related to query asked by client and returns the results to job tracker. Map reduce also return the result in blocks (where data is stored) to the client.

Some guiding force used HDFS are:

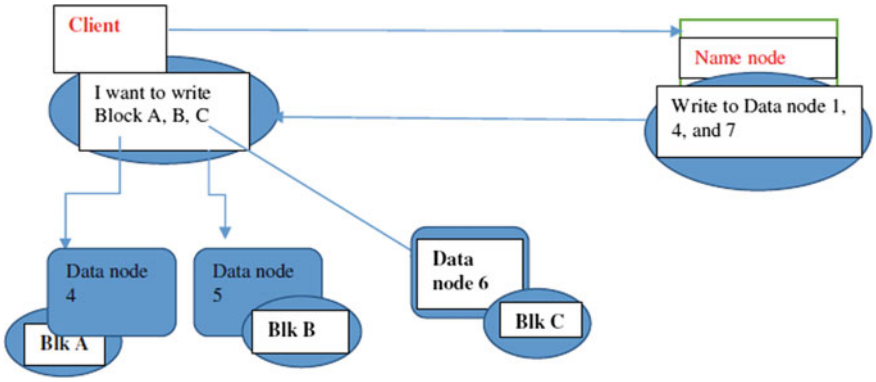


Fig. 4 HDFS file system architecture [7]

Name node: The entire request from client is received by master node. It helps in finding the suitable metadata that is appropriate for storing that data node which is related to client. The selection of a data node depends on the availability of that node whether it's free or not.
Secondary Name node: for creating the back up of name node, secondary name node is used. Those files which have detailed about that particular data node are stored in it. If name node fails, then data can be recovered from secondary name node.
Job Tracker: Maps reduce assigns jobs to data node and task tracker. Actual data is to be stored in data node and it sends heartbeat to name node about stored data.

3 Big Data in Healthcare

Big data in the field of healthcare points to the patients lab reports, X-ray reports, case history, list of doctors and nurses, list of medicine with their expiry date etc. heath care department are using help of big data technology for gaining this type of information about patients and provide with better results.

3.1 Mobile Big Data

Large data that are gathered by mobile technology may be defined as mobile big data. The data can be either offline or online or can be structured or unstructured or semi structured. Such type of data was not possible by traditional database management to manage on. Mobile big data gain its importance in day todays life for solving such problem of traditional database management system as mobile technology is widely accepted in present time. Some characteristics of mobile big data are as follows [3].

Mobile big data is huge in size. On the daily basis gigabytes and terabytes of storage is being required.

Mobile big data is rigorous. As mobile devices are portable the data must be available every time. Hence mobile data analytics should be implemented recurrently with the collected data samples at the higher speed.

Mobile big data is heterogeneous i.e. any form of data can be stored in it.

3.2 Big Data Analytics

It consists of a set of activities that are discussed below:

Data Collection: this activity may require the collection of data. In various diseases like diabetes etc. the initial sign are heart rate, blood pressure which is measured by ECG, EEG. There are many providers in market for providing body sensors. Health signals are constantly netted from on-the-body or in-the-body sensors and thus learnt by the mobile device [1].
Data Extraction: data which is gathered from Data Acquisition can either be structured or unstructured. The collected data may be preprocessed for achieving relevant information out of it through feature extraction.
Feature Selection: the data that we receive after information extraction is selected by choosing subsets of relevant features with respect to healthcare.
Predictive Modeling: data mining tools are used for predictive modeling which help in prediction of trends and patterns. In this type of modeling various predictors are used for predicting various collections of data.
Data Visualization: The consequence cultured from predictive big data analytics grows its importance from visualizations, such as time series charts, that are treasured for decision making by providers of healthcare [1].

3.3 Big Data Ecosystem for Healthcare [8]

It is basically a vast technology that includes mechanisms and tools to manage huge facts and information on it. The main purpose is to carry information from different areas, keep them in hadoop distributed file system, manage and handle this fact using modules of hadoop like PIG, Map-Reduce, SQOOP, HIVE, FLUME OZZIE etc. Various components of hadoop are mentioned in Figs. 5, 6 and 7.

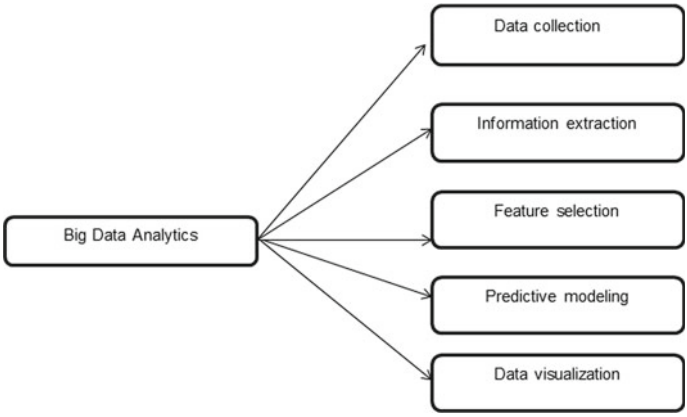


Fig. 5 Big data analytics [1]

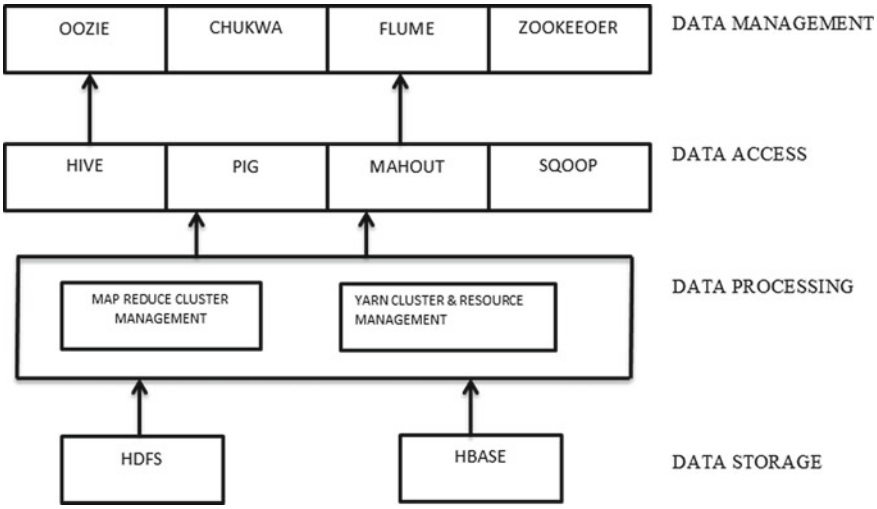


Fig. 6 Components of Hadoop [8]

3.4 Big Data Life Cycle

Data Collection: As the name suggests that data need to be collected from various places and is kept in HDFS. Here if one talks about data, it can be of any form such as medical images, social logs, sensors etc. [5].
Data Cleaning: Again name tells about this process where waste data like junk need to be deleted or washed away or it checks whether there is any requirement to delete any trash or not.

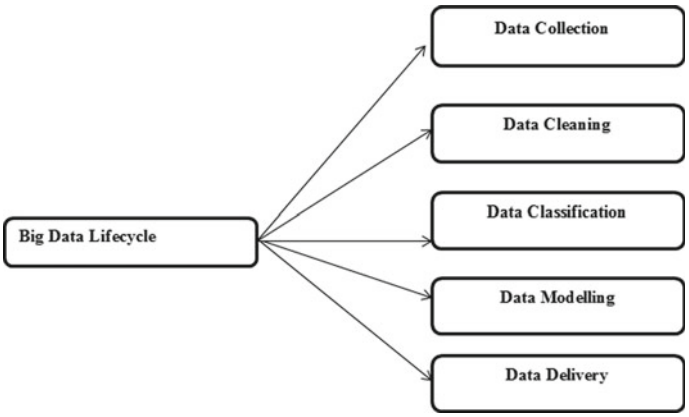


Fig. 7 Big data lifecycle [5]

Data Classification: It consists of different classification of data and their filtration. For instance big data that is used in medical side includes most of the data which is in the form of unstructured data like manually made notes. To carry out actual and effective analysis semi-structured, unstructured and structured information should be properly classified.
Data Modeling: Data modeling simply means analysis to be carried out on selected confidential data. For instance if list of underweight children is needed from any specific area then for this case their health report is needed and there is a requirement of checking information related to families that come under poverty. Thus according to this data should be processed.
Data Delivery: As name suggests data delivery means there will be a specific report related to previous instance. After analyzing the data report is made. In short in all the stages of BDLC data storage, integrity and access control is needed. Thus all the big data analysis has their own importance in order to maintain and process data should be used to present the results of investigations and large sets of figures clearly [5].

3.5 Need for Big Data Analytics in Healthcare

Big data analytics are very beneficial for healthcare and there are many factors that are responsible for improving the quality of healthcare. These are discussed below.

Provision of centric services to patients: To deliver quicker aid to the patients by giving indication related medicine distinguishing symptoms and viruses at the prior phases that depends on the medical information obtainable, reducing painkiller dosages to reduce side effect and giving effective medication created on heritable

cosmetics. These benefits in decreasing readmission degrees thus decreasing rate for the patients.

Detecting spreading diseases earlier: Calculating the viral illnesses prior formerly scattering created on the live investigation. This can be recognized by evaluating the community logs of the patients distressing from illness in a specific place. This aids the healthcare specialists to direct the sufferers by having essential defensive procedures.

Observing the hospital's quality: To check whether the clinics are arranged as per standards given by Indian medicinal assembly. It benefits administration in checking essential actions in contradiction of banning clinics.

Modifying the treatment techniques: The check-up of modified victim tells the consequences of medicines constantly and by these analysis quantities of medicines can be altered for quick results. By checking patient's energetic signs to offer active precaution to patients, creating an investigation on the documents produced by the patients who previously suffered from the similar signs, aids specialist to deliver actual tablets to other victims [13, 14].

4 Advantages of Big Data [7]

- Big data in health informatics can be offered to predict consequence of illnesses and epidemics, increase action and quality of life, and protect from untimely demises and illness growth.
- The main feature of big data is to offer data related to illness and it also provides cautioning hints for the treatment that needed to be controlled.
- By using big data not only humanity is protected but also the cost of illness treatment is reduced to very large extinct.
- As the main function of the big data is to offer large quantity of information so it is becoming very beneficial for both clinical medication and epidemiological research. It has allegations on healthcare on patients, workers, scholars, healthiness specialists. This data has been used by various organizations, companies to make formulate strategies, schemes, interposition or medicinal treatment such as medications growth. Thus everywhere big data is having very good benefits especially in health department.
- In today's time patients giving data related to healthcare decisions are highly demandable and need contribution in their fitness assessment production.
- The main function of the big data is to keep patients always updated so that they can always make best choice in healthcare. Also it also helps them to fulfill with the health related treatment.
- The Big data has capability to decrease the recency bias or recency effect bias. Recency bias means when the current measures are consider more deeply than previous measures in order to recover the condition, but It might result in improper choices.

- The real time info can also be combined into this technique called big data. It has various benefits such as some mistakes or issues in an association could be recognized directly. Also the operative problematic issues can be overwhelmed. It will definitely result in saving time, price and enhance the output. The facilities also can be additionally upgraded because the up-to-date data on particular subject substance is offered. For example, it would be easier to provide the whole data on the patients and also it will be possible to administer medicinal involvement devoid of any suspension.
- It is also castoff in prognostic investigation which is to recognize and discourse the medicinal matters before it becoming an uncontrollable issue. Healthcare specialists are capable to decrease the danger and overwhelmed the problem with the material imitative from the big data [1].
- Big data is too capable to aid recognize deceptions in healthcare particularly on indemnification rights. Fake, discrepancy and deceitful entitlements can be highlighted. This will ease indemnification corporations to avoid damages.
- It can also help healthcare via data organization, electrical medicinal histories and documents inquiry. It will benefit to discover and recognize the correct populace or objective cluster. It contains varied collection of populace and definite collection can be recognized for risk valuation and broadcasts.
- Its presence would also permit growth a s well as alteration of any software package or involvement to aim the fitness issue. It would also let the medical prosecutions to be originated instantaneously. Big data would offer a stronger image on the kind of populace and medicinal issue. The design of the dispersal or illness info would be able to offer rapid growth of interposition software and also directing the affected collection immediately.
- Information developments of pharmacological productions were evaluated from patients, caregivers, stores and Research and development. It could ease the pharmacological corporations to recognize fresh prospective and nominal medicines and bring it to the customers as soon as possible [9].

4.1 Issues in Big Data

Big data is very useful and popular technique that is having information in various forms. There are so many advantages of this technique in the fields of healthcare. Not only in medical field there are so many areas where this technique is super important. But beside with this, various issues of big data also exist. There are enormous challenges in case of information security, gathering and distribution of fitness information and information practice. Big data analytics via the usage of refined skills has the ability to convert the information storehouses and create cognizant conclusions. Problems like confidentiality, safety, values and supremacy are required to be addressed. Data like Nano particulate treatment on cancer therapy can also be combined in big data to deliver the summary and fines therapy for cancer particularly

when nanotechnology is essential in medicine distribution in cancer handling process. Separately from that contrary results of medicines usage can also be specified.

There are some major issues of big data that are really needed to mention. These problems are discussed in details and are given (Tables 1 and 2).

To avoid all such issues, there are three main components that have major significance in big data:

Data validation: This is a very important component in which it has been ensured that data is not corrupted and it is totally accurate. To check such data, validation is done. All data is validated via HDFS to check whether it is correct or not.

Table 1 Issues of big data [7, 8, 10]

Security [7]	Meanwhile the big data involved subject's individual data and their fitness all previous records, there is a requirement to prevent the databank from hacking, cyber robbery and phishing. Databank contains all the information related to the healthcare field. Intruder use this information to sell for gigantic amount. This one of the major problem that has been arrived from previous time Not only information related to medical field can be stolen but all the fields where big data has been used can be hacked for instance marketable groups like broadcastings corporations, especially banks or business organization are also in danger without the awareness of the customers. The use of big data is beneficial only when there is a proper security, safety and protection of stored information is done. The availability of the healthcare documents require to be reliably studied and checked
Data classification [7]	Big data is a huge, fewer organized and varied. There is a requirement to recognize and categorize the information so that it can be used efficiently. Though, it is arduous to explore for a particular documents in the big data. It also essential to be contextualized or joint together so that it will become much appropriate to particular person or members in a group
Cloud storage [7]	To transfer information or taking the entire scheme planned in the cloud system, cloud storage is always needed. Therefore for this purpose there should be always enough memory in the cloud and at the same time high speed is needed to transfer the information For storing graphic category for example X-ray, CT, MRI, words documents should be available in storage area. It will be only useful for the clinicians if there is always graphic presentations from the given information in order to watch and understand easily that will result in decision
Data Modeling [9]	Though big data is brilliant for demonstrating and imitation, there is also a requirement to recognize construction and pool the correct applicable documents so that it could be castoff to design the difficulties, which well along can be used for involvement. Deprived of the appropriate organized information, it is exciting to examine and envision the productivity and to extract particular data
Miscommunications gap [8]	Another crisis of big data is called the miscommunications gap which normally occurs among the customers and information experts. It is very important that every customer should have proper knowledge and understanding of information generated by the data experts. From the survey it has been observed that there is lack of communication between experts and consumers that may results in the use of big data. Data should be organized in such a way that when there is any requirement of fetching any information, it should be done quickly whether it is of medical field or another area where big data is used. But this maintenance is not possible currently. Therefore, this is a wastage time since the specialist will require data from the start, to fetch the patient's history. Meanwhile big data has the capability to foresee future medicinal disputes which is a progressive thing

(continued)

Table 1 (continued)

Data Accommodation [10]	Single basic big data scheme is needed to organize information. Thus it should be compatible and basic. The main purpose is that all customers are always getting their data easily and there is no confusion and difficulties in this process. It is a complex job to receive every applicable structure to connect with everyone. There is a belief of dissension inside every group, wherever certain revelries may handle the information for their own requirements than for the association as an entire
Technology Incorporation [10]	One of the major problems is the absence of data to support the decision making, strategy formation or rule in big data. The method of redefine and in embracing of technique is not fast and this can affect the healthcare, care distribution and investigation study. With the absence of the technology, big data is unable to generate and disseminate information [9]
Data Nature [10]	The combination of information would not only include files inside the healthcare organization but moreover outward information is also involved. Though it provides possible profits, but it is also challenging when there come confidentiality, safety and authorized troubles. The healthcare information typically contains patients who are looking for cure in the hospices or private clinic but nobody on fit persons. Through the presence of vigorous persons in the databank, it would be easy to deliver improved appreciative on the nature of the infection and involvement. Since the information is now extra present, it is essential that the facts are approved to the consumers directly for medical choice and to improve the fitness consequences

Process validation: It basically involves Map reduce that checks the correct data and sources. Process validation verifies the business logic, node by node. It checks whether the key-value pair is created accurately or not.
Output validation: In this component data is processed in the repository and the main task is to ensure that data is not distorted. This is done by comparing the data from HDFS files.

4.2 Examples of Big Data Analytics for Healthcare

Important examples of big data in healthcare are discussed in the table. There are several initiatives utilizing the potential of Big Data in healthcare. Some of the examples are listed below:

4.3 Applications of Big Data in Healthcare

Each Electronic Health Records (EHRs): It's the greatest extensive software program of big data in medication area. All patients that are taking treatments from that place have their data saved in this application which comprises demographics, medicinal records, antipathies, checkup results etc. All history is displayed through protected information schemes and is accessible for suppliers from public and private segment.

Table 2 Example of big data analytics [10–12]

Example	Description
Asthma polis [10]	For treatment of asthma company have created a tracker called global positioning system (GPS) that monitors usage of inhaler by patients. Small cap like device is to be placed at the top of the inhaler that at as a sensor and help in providing useful information. The patient that is using such type of inhaler when any time suffers from asthma attack and uses hid inhaler that time that device will records the time and place and convey the information to web site. This data is then made available to Center for Disease Control [10]. The CDC'S take the survey for why and from which allergic source the attack of asthma was caused to patient. Thus all the relevant data about the attack is gathered through the help of device. The benefit for this device to user is that he can generate the report of his attack and will be aware from that what the source he is facing attack of asthma. And now patient will be aware to face them and will be ready with all the precautions of asthma. Thus doctors can be well aware with the real time reports of their patient and provide them with better diagnoses
Battling the flu [10]	Center for Disease Control have become strong pillar in big data for influenza. Over a week 6, 80,000 flu reports are received by CDC. All these reports which are gathered by CDC include the reason for sickness of the patient, what treatments are given to them and whether that treatment is effective or not. The CDC helps general public to make this information available to them. Doctors also get the benefits of this by getting the clearer picture of how and why the disease is spreading across the world. It helps the care takers to get the information about vaccines and other antiviral medicines that can be given to patients for their faster recovery. This application of big data is not only restricted to doctors use only but the patient can himself assists for better recovery. FluNearYou an application made by the Skoll Global Threats Fund and the American Public Health Association, motivates user to input their symptoms before they fell sick completely thus proper diagnose is given at much earlier stage. The only disadvantage of this application is if some users are putting some false input or incorrect data then wrong diagnose will cause negative effect to other users also. Another application in big data analytics is "Help", I Have the Flu, planned by the pharmaceutical company Help Remedies. Application takes the benefits from social media and help in getting quick recovery from the disease

(continued)

Table 2 (continued)

Example	Description
Diabetes and big data: [11]	In big data revolution diabetes patients have also got up with lot of benefits. Common Sensing company have given GoCap application, that not only help in recording the daily dosage of insulin but also at what time dosage is given to the patient is recorded. This information is then feed over mobile devices where patient and other members can get this information. Thus data become easier for other healthcare professionals to access and allows them for identifying where the problem is and what can be the proper diagnosis for it. Another technology that has emerged with the combination of diabetes and Big Data is served by Allazo Health. Predictive analytics is being used for improving medication program in this system
GNS Healthcare and Aetna [12]	Big data analytics companies have helped lot in GNS healthcare; it has brought up health insurance company Aetna to help the people that are having metabolic syndromes. A technology called as Reverse Engineering and Forward Simulation has been developed by GNS. This technology help in providing data related to Aetna who have subscribed for insurance. Firstly the application will check for 5 sign in the patients which are: high blood pressure, low High density Lipoprotein, large waist size, high blood sugar and high triglycerides. Any patient who has the combination of these three signs lead to the result that patient is suffering from Aetna. A different combination of these signs leads to different conclusions. Like if any patient whose report include low High density Lipoprotein and high triglycerides may suffer from high-risk hypertension within the next few months

Real-time warning: It provides a vital feature which is called as real-time alerting. In hospices, an application known as Clinical Decision Support examines medicinal information at a time, offering fitness doctors with guidance since they are able to create rigid conclusions.

Help in keeping opioid abuse in the US: Another useful application is undertaking a grave issue in the US. The problem includes the tragic demise of thousands of people due to overdoes of opioids in the U.S. Another problem was highway accidents, that are earlier the greatest mutual reason of unintentional demise. A program of big data may be the response everybody is waiting for. The Researcher worked with analytics specialists at Fuzzy Logic to handle the issue. With years of assurance and apothecary documents, Fuzzy Logic analysts are now capable to classify 742 hazard issues that guess with a high degree of correctness whether somebody is in danger for mistreating opioids [13].

Improve patient appointment in their own health: Numerous of customers have an attention in high technology strategies that measure and keep the track of each phase they imitate. Phases like heart rates, napping ways are recorded in regular

basis. All this vigorous data may be joined with added recordable documents that can check possible fitness dangers in future. Customers are regularly monitored for their health, and enticements from fitness insurances can motivate them to have a fit and healthy life.

Use health data for a better-informed strategic planning: The usage of big data permits for premeditated preparation appreciations to improved visions into public's enthusiasms. Care executives can examine check-up consequences amongst persons in various demographic collections and classify what issues dishearten persons from taking over cure.

Research more widely to cure cancer: Cancer moonshot program is one of the biggest achievements of big data analytics. This application was mainly created for the treatment of the people suffering from cancer. By using this cancer can be treated in half of its actual time. The Large and effective amount of information has been used by the doctors in order to get 100% results. Medicinal investigators can follow data in big quantities on treatment tactics and improvement level of cancer people to find trends and cure that can give guarantee to give success [14].

Predictive analytics: Big data analytics are now very popular and helpful for every field. Predictive analytics are main commercial brainpower in the present time. Predictive analytics will be able to improve patient's care and treatment. Optum Labs, US research co-operatives have composed electronic health records of more than 40 billion people who are suffering from any kind of disease or illness. It has been done to make a databank for predictive analytics tools that would recover the provision of treatment. The aim of healthcare commercial intelligence is to direct surgeons make data-driven choices in milliseconds and recover people's cure.

Decrease fraud and improve information security: Fraud and damage are very big issue in every field. In order to protect societies from these malfunction, numerous have idea to make use of analytics to aid stop safety dangers by classifying variations in network circulation, or another activity that leads a cyber-attack.

Practice telemedicine: Telemedicine is available on the market-place from 40 years but it was not so popular and its benefits were not highlighted. Now technologies such as film seminars, phones, wireless policies, and wearable etc. have arrived and they have made telemedicine more popular. All the facilities of clinics are now offered by these techniques. It is castoff for main discussions and early analysis, remote patient nursing, and medicinal tutoring for health specialists.

Stop needless ER appointments: Protecting period, money and vitality by big data analytics in the field healthcare is essential. Lady in 3 years went to the ER more than 900 periods. It is the condition in California, where a lady who was having mental disease and material misuse went to a diversity of native hospices on everyday basis [9].

5 Conclusions and Future Scope

Huge capacity of information is needed in different fields. For instance record keeping, patient related data in the healthcare industry etc. The Main focus is that each data should be digitized in today's digital world. Also among these challenges the data should be analyzed effectively with minimum cost. It has been observed that government sectors are also generated large volume of data every day. So robust technology is needed that will take care of this large data set in real time to help citizens for getting better results. Big data helps in providing valuable decisions by data patterns and relationship among different data set with the help of various machine learning algorithms. Big data is nothing but the collection of information in the form of structured and unstructured. The main purpose is to carry information from different areas, keep them in the hadoop distributed file system, manage and handle this fact using modules of hadoop. In this chapter module like PIG, MapReduce, SQOOP, HIVE, and FLUME OZZIE are discussed in detail. Various issues such as Miscommunications Gap, Security, Data Classification, Cloud Storage etc. are also mentioned in the chapter. This chapter is also covering some important examples and applications of big data because big data is having so many applications in this real world. It has been mentioned in the paper that how big data is helpful and important in the field of healthcare. Big data includes some important features like map reduce, HDFS. These features have their own importance and they are also discussed in this chapter in detail.

Researchers are using machine learning concepts and tools with the big data to get the best results. Various researches are undergoing in the field of machine learning. An efficient tool can be developed to over some issues of the big data. This tool will have provision to manage noisy and imbalanced files. It will also manage uncertainty and inconsistency that will resolve the issues of big data.

References

1. Fatt: The usefulness and challenges of big data in healthcare abstract the usefulness and challenges of big data in healthcare data modeling, mobile big data analytics in healthcare. iMedPub J. (2018)
2. Russom: Big data analytics. TDWI best practices report, fourth quarter **19**(4), 1–34 (2011)
3. Nambiar: A Look at Challenges and Opportunities of Big Data Analytics in Healthcare, pp. 17–22 (2013)
4. Shvachko, et al.: The hadoop distributed file system. In: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10 (2010)
5. Zikopoulos, et al.: Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media (2011)
6. Summary, et al.: Big Data is the Future of Healthcare, pp. 1–7 (2012)
7. Raghupathi, et al.: Big data analytics in healthcare: promise and potential. Health Inform. Sci. Syst. **2**(1), 1–3 (2014)
8. Bates, et al.: Downloaded from http://www.content.healthaffairs.org by Health Affairs on 15 Sept 2014

9. Dates, et al.: Call for Book Chapters Big Data Analytics in Healthcare Springer Series : Studies in Big Data (Link) Purpose and Scope : Editors of Book, pp. 3–4 (2019)
10. Belle, et al.: Big Data Analytics in Healthcare, pp. 12–17 (2015)
11. Archena, et al.: A survey of big data analytics in healthcare and government. Procedia Comput. Sci. **50**, 408–413 (2015)
12. Belle, et al.: Big data analytics in healthcare. BioMed Res. Int. (2015)
13. Sun, et al.: Big data analytics for healthcare. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1525–1525. ACM (2013)
14. Sarwar, et al.: A survey of big data analytics in healthcare. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **8**(6), 355–359 (2017)

Elements of Healthcare Big Data Analytics

Nishita Mehta, Anil Pandit and Meenal Kulkarni

Abstract As the focus of healthcare industry shifts towards patient-centric model, healthcare is increasingly becoming data-driven in nature. Alongside this, the newer developments in technology are opening ways for harnessing healthcare big data for improved services. The application of analytics over big data in healthcare offers a wide range of possibilities for the delivery of high quality patient care at affordable price. Although a lot has been discussed about the promise of big data analytics in healthcare, there is still a lack of its usability in the real world scenario. The healthcare organizations are starting to embrace this technology, yet many of them are still far from achieving its benefits to the full potential. The challenges that these organizations face are complex. This chapter begins with discussing about the key issues and challenges that often afflicts the utilization of big data analytics in healthcare organizations. It then highlights the essential components for effective integration of big data analytics into healthcare. It also explores important foundational steps for beginning a big data analytics program within an organization. The objective is to provide the guiding principles for successful implementation of big data technology.

1 Introduction

In the recent years, healthcare has seen a transition in its landscape from "clinical-centric" care model to more consumer-driven "patient-centric" care model. There has been a movement towards service line approach with the focus shifting from provider-centric experience to patient-centric experience. In the traditional model with hospital and healthcare providers at the center of the system, a large amount of data was available in the form of medical files and records. But, owing to the storage of data in silos along the continuum of care, there had been a limited access to this information. As we move towards more patient-centered approach, the patient

N. Mehta (✉)
Symbiosis International (Deemed University), Pune, India
e-mail: nishitamehta@sihspune.org

A. Pandit · M. Kulkarni
Symbiosis Institute of Health Sciences, Pune, India

A. J. Kulkarni et al. (eds.), *Big Data Analytics in Healthcare*,
Studies in Big Data 66, https://doi.org/10.1007/978-3-030-31672-3_2

needs are put first and patients are bestowed with greater responsibility for their own health. The key component of patient-centric healthcare—shared decision-making—has led to an increased demand for transparency in the system and hence the growth in storage of healthcare data in digital format [1, 2]. The increased pace of generation of data has brought about an explosion of digital healthcare data. The magnitude of this data, its velocity and heterogeneity contributes to such data being termed as *big data.*

Digitization of healthcare data has also opened the way for its utilization in order to improve the quality of care. The healthcare data obtained from diverse sources, can be applied across different areas for better performance management [3]. But raw data stands ineffective for decision-making and requires its conversion into meaningful information. With the advancement in technology and development of efficient analytical tools, the healthcare big data can be anatomized so as to obtain important insights.

As big data analytics is becoming the revolution in information technology, it has an immense potential to transform healthcare. The increased ability to analyze convoluted datasets and the results thus obtained, not only promises the optimization of processes, but also provides various measures for improving the quality of patient care and hence patient satisfaction. Where every second counts in a life or death situation, the use of big data analytics in healthcare also facilitates timely decision-making and hence plays a crucial part in saving lives.

Regardless of the prospects offered by application of big data technology in healthcare, there is a lack of its adoption owing to a number of challenges faced by organizations. Some of the systemic challenges need to be overcome beforehand in order to realize the full potential of big data analytics. This chapter explores the challenges in implementation and effective utilization of big data analytics. Once the challenges are identified, the vital elements for applying big data analytics into healthcare are presented. Later, the chapter also discusses essential strategic measures for successful implementation of this technology so as to accomplish its complete power.

2 Challenges in Application of Big Data Analytics for Healthcare

While the advancement in analytics offers various measures to improve healthcare, several key challenges at different levels hinder big data use in healthcare. Starting from the entry of data into digital platform till the management of results obtained through big data analytics, the organizations might face various obstacles. This section identifies those challenges and brings forth the broad areas where healthcare will confront such issues (Fig. 1).

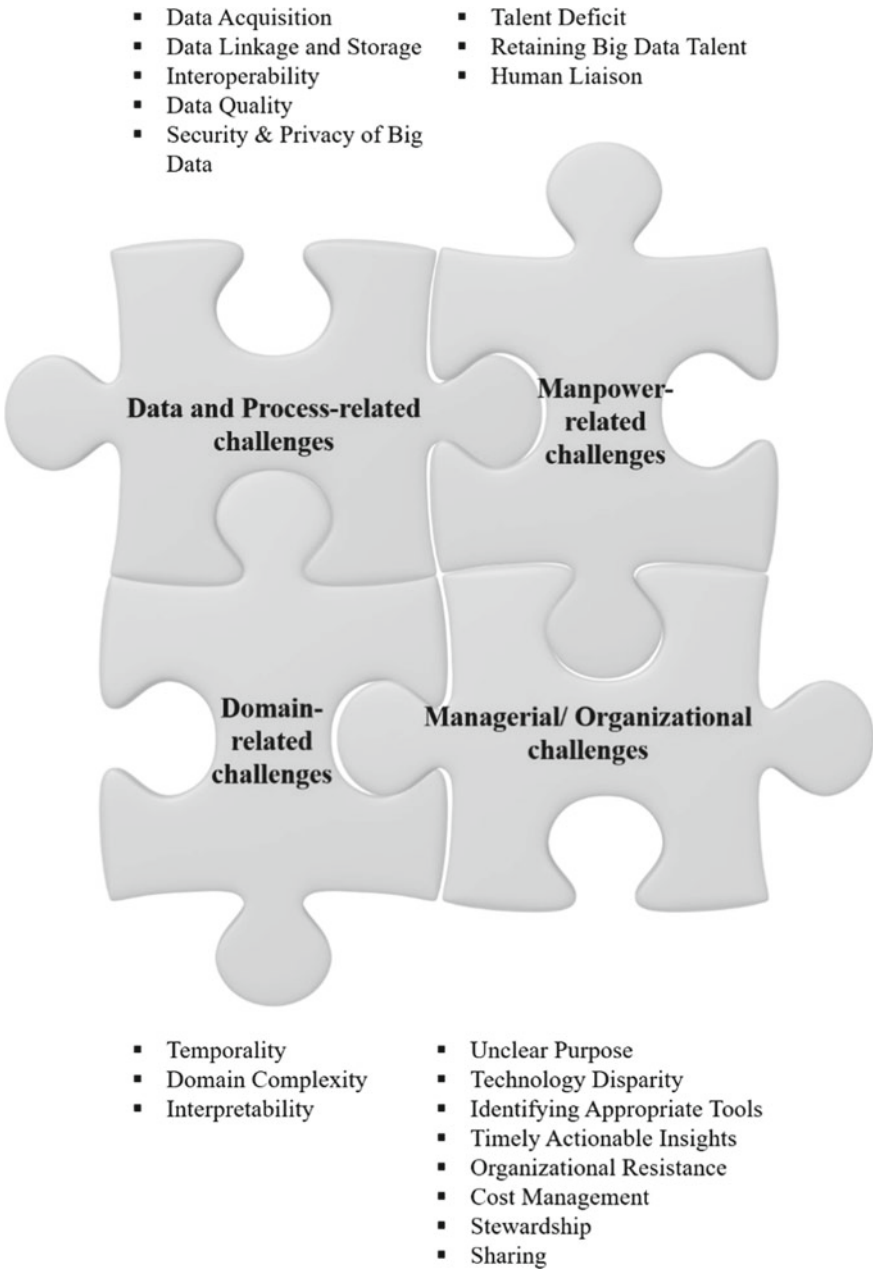


Fig. 1 Challenges for big data analytics in healthcare

2.1 Data and Process-Related Challenges

- **Data Acquisition**: Coming in from disparate sources, healthcare data is multidimensional in nature and is highly segmented [4]. Lack of synchronization amongst these data sources can create gaps and provide misleading information. Also, the data available from these sources vary widely in structure. On the one hand, where healthcare data includes patient demographic details which are structured in nature; clinician notes, diagnostic images like CT and MRI scan and videos are unstructured in nature. Unification of such data silos from different sources and its conversion into the common compatible format for storage in the system is thus a challenge [5]. Besides, the generation of data in real-time or near real-time makes continuous data acquisition difficult [6].
- **Data Linkage and Storage**: As the volume of healthcare data grows exponentially, legacy IT systems in organizations are unable to handle such large quantities [6]. Added to that, storage of data across different departments within the organization leads to the issues of data redundancy [7]. It becomes difficult to analyze such fragmented and incomplete data. In fact, even the best of the algorithms don't work on disintegrated sources. The large volume and variety of data from different sources, poses a challenge to integrate these sources and aggregate it into repositories [8]. Further, separating useful information from the voluminous raw data and data reduction would demand the use of various filters. It is an additional concern to ensure such filters do not discard the important information [9].
- **Interoperability**: Where data is produced from various healthcare equipments and devices, there arises the issue of interoperability [4, 10, 11]. Owing to the difference in platforms and software these devices work on, the data generated by them are in different formats. To be able to use this data effectively, the devices should be able to communicate and exchange data in a format which is standard and compatible with other devices.
- **Data Quality**: In bringing together different formats of healthcare data from different sources, the accuracy and integrity of data is a concern [4, 6, 12]. Patient data might come from physiological bedside monitors, pathology reports, diagnostic images like X-rays and videos from different examinations. Prior to the analysis of this diverse data, there is a need for information extraction process which pulls out the essential information and presents it in an appropriate form for analysis. Warranting the preciseness and thoroughness of this process is a technical challenge. Besides, technical faults in sensors and human errors may account for incomplete and unreliable data which might have negative impact in terms of major health risks for patients and adverse events [9]. Moreover, the heightened costs for healthcare organizations can be attributed to errors due to poor data quality. Hence, cleaning and normalization of data by removal of noise or irrelevant data, forms another issue in meaningful use of data.
- **Security and Privacy of Big Data**: From medical records to pathology reports to insurance claim details, patient information resides at multiple locations and is available from a number of endpoints. Myriads of applications are used to monitor

patient health and enable exchange of patient data among different parties, each with varied levels of security. Some dated applications can pose a threat to the security of health data, making it vulnerable to cyber-attacks. Access to personal information including name, employment and income information and its misuse can cost exorbitant amounts to healthcare providers [13]. Moreover, exposing the private health data evoke concerns for patient confidentiality [6, 14]. Due to the unrestrained acts of data breaches, even with the technical measures and safeguards in place, managing big data security and privacy is still a technical challenge.

2.2 Manpower-Related Challenges

- **Talent Deficit**: Once the organizations decide to implement big data technology, there arises the demand for qualified data experts with the skills to analyze healthcare data [15]. There is a huge demand for data scientists and analysts with healthcare domain knowledge, having the ability to apply right set of tools to the data, obtain results and interpret it in order to provide meaningful insights. But there is a scarcity of people having the skillset and expertise for applying analytics to healthcare [16].
- **Retaining Big Data Talent**: Although the organizations may find the best suitable big data talent with knowledge of healthcare, it is still difficult and expensive to hire them. Also, given the increasingly fierce competition, the retention of highly talented data scientists and analysts is another challenge [17, 18].
- **Human Liaison**: Despite of the advances in technology, there are still certain areas where human would yield faster results than what the machine does. Thus, it requires humans to be involved at all the stages of big data analytics. Experts from different areas (people who know the problem and people who know where the data resides) need to collaborate along with the big data tools to provide significant answers. The dearth of adept and proficient people again becomes an obstacle for application of big data analytics [19].

2.3 Domain-Related Challenges

- **Temporality**: Temporal information about the progression of disease over a period or advancement over the course of a hospital, is critical for healthcare. Time is a fundamental entity for crucial healthcare decision-making [19, 20]. But designing the algorithms and tools which can work on temporal health data is a difficult and cumbersome task [21].
- **Domain Complexity**: The heterogeneity of diseases, co-occurring medical conditions, difficulty in accurate diagnosis and interaction between patients and

providers, all add to the complexity of healthcare [21]. Apart from the aforementioned factors, the complexity is also affected by presence of task-related factors, team, environmental and organizational factors [22]. In fact, the newer developments with regards to certain diseases and their progression add other complications in designing big data tools.

- **Interpretability**: The decisions in healthcare are crucial since the lives are at stake. So as to ensure that the decisions can be relied on, not just the results obtained after big data analysis are important, but also the logic behind such results plays a significant part in convincing the healthcare professionals about the recommended actions [21].

2.4 Managerial/Organizational Challenges

- **Unclear Purpose**: The first obstacle in exploiting the potential of big data analytics is unclear purpose. Organizations adopt big data technology as a source of competitive advantage without having the business objectives defined explicitly. Such transformation by force fitting the new technology will lack direction [23]. Elucidating the business objectives for applying big data analytics is again a challenge.
- **Technology Disparity**: Although there is a growing interest in digitization of healthcare, most of the organizations still rely on the conventional technology. Replacing the legacy system of storing and managing data with the newest of technology is a problem. Since healthcare has a huge amount of data stored in the paper-based records, digitization would require a lot of efforts, time and human resource. Overcoming this technology gap is another challenging task [24].
- **Identifying Relevant Data and Appropriate Tools**: Once the organizations are aware of the business use cases for applying big data technology, the next hurdle is to identify the relevant data and store it. Consequently, the right tools which can be applied over such data needs to be discovered. Discerning the appropriate tools and solutions available is also a daunting task [25].
- **Timely Actionable Insights**: Processing and mining of healthcare big data using pertinent tools and analytics solutions, might yield more data or information rather than insights. Actionable insights are more valuable owing to the actions it proposes instead of providing simple answers to the questions posed [26]. Hence, the concern is to obtain insights that drive action in a timely manner. Considering the significance of timeliness of healthcare decisions, the generation of actionable insights in real-time or near real-time is another issue to contemplate [27].
- **Organizational Resistance**: Organizational impediments and the lack of adoption of technology by middle management is other stumbling block [28]. With the inadequate knowledge about advantages of using big data analytics for business as well as human resource, there is meagre motivation for implementation of big data technology. Especially in healthcare domain, where the doctors rely on their

experience and instincts for decision-making, convincing them of the benefits of big data technology is an arduous exercise.

- **Cost Management**: As the healthcare organizations adopt data-driven culture, the most important initial issue is to manage the cost of data warehouses and infrastructure needed to store huge volume of data [29]. In fact, the vast computing sources needed for analysis of big healthcare data add to higher initial investment needs. This renders it unaffordable for small and medium organizations and also the large organizations are hesitant for such an investment without much clue about the returns [30].
- **Stewardship**: Patient and health related data is highly rich in information. Such data after de-identification might be used by organizations for research purpose. With the clinical data being used for purposes other than patient treatment, there arises a concern for stewardship and curation of data [29].
- **Sharing**: Another hurdle in the way of successful implementation of big data analytics is the need for data-sharing culture [29]. To harness the benefits of big data technology for community, there is necessity for organizations to share this data with other healthcare organizations. Despite the fact that this would enable more transparency and availability of data for quicker decisions, owing to the competitiveness organizations are not keen towards data sharing. This impedes the efficacious employment of big data analytics for healthcare.

3 Key Elements of Big Data Analytics in Healthcare

With the intent of overcoming the aforementioned challenges, healthcare organizations need to begin with determining the principal elements for application of big data analytics. In fact, obtaining meaningful insights through the use of big data analytics requires the interaction between various elements. This section discusses the vital elements in healthcare big data analytics (Fig. 2).

3.1 Data Inputs

It is the data which encourages the use of analytics for obtaining actionable insights. That being so, healthcare data which is immensely rich in information can be utilized for not only enhancing the quality of care but also for improving the operational efficiency. In order to utilize such data to its potential value, all the data required for solving particular problems needs to be sourced. Identifying the sources of relevant data and collecting it, might present additional infrastructural requirements [31, 32]. Designing of such information infrastructure for collection, storage and scaling of data [33] would depend on the type of data and its structure.

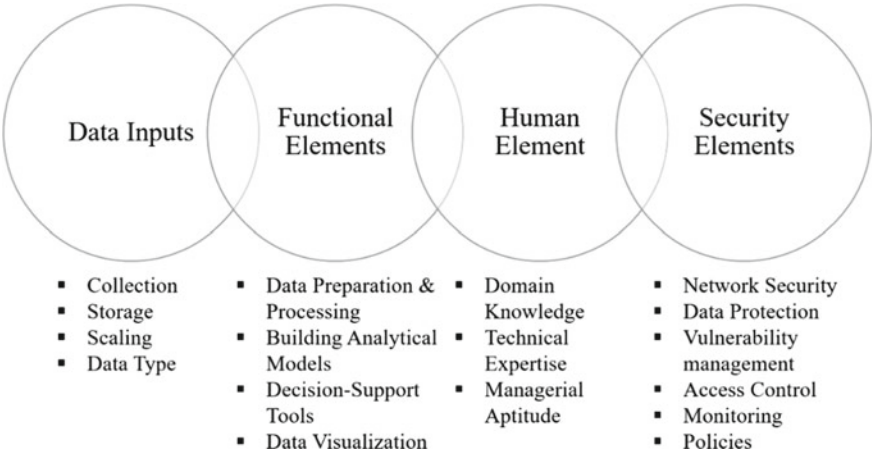


Fig. 2 Key elements of big data analytics in healthcare

Considering the heterogeneity of diseases and treatment pathways, healthcare data is highly diverse in nature. Along with the variety of health-related data, the difference in its structure makes it complex to manage. Data varies from being highly structured (demographic data), semi-structured or weakly structured (diagnostic data) to being unstructured (doctor notes) in nature. Some of the data might also be available in legacy systems [34]. Such an intricate character of healthcare data requires restructuring of data capabilities. This would also demand for more sophisticated systems with higher capacities for data storage.

3.2 Functional Elements

In order that the health-related data can produce meaningful and actionable insights, there is a need for seamless leveraging of this data across the organization. The management and analysis of information across the ecosystem involves following elements (Fig. 3).

- **Data Preparation and Processing**: This stage takes into account the cleaning and formatting of data [31] to ensure high quality results. Raw clinical data, whether structured or semi-structured, needs to be converted into flattened table format for analysis. The processing of huge amount of healthcare data would necessitate the use of distributed processing capability [35]. While the data is stored in Relational database or Hadoop Distributed File System (HDFS) or HBase, different processing frameworks can be chosen depending on the requirement. For instance, MapReduce and Spark can be used for batch processing; Storm and Spark Streaming for real-time processing; Mahout and MLib for machine learning; and

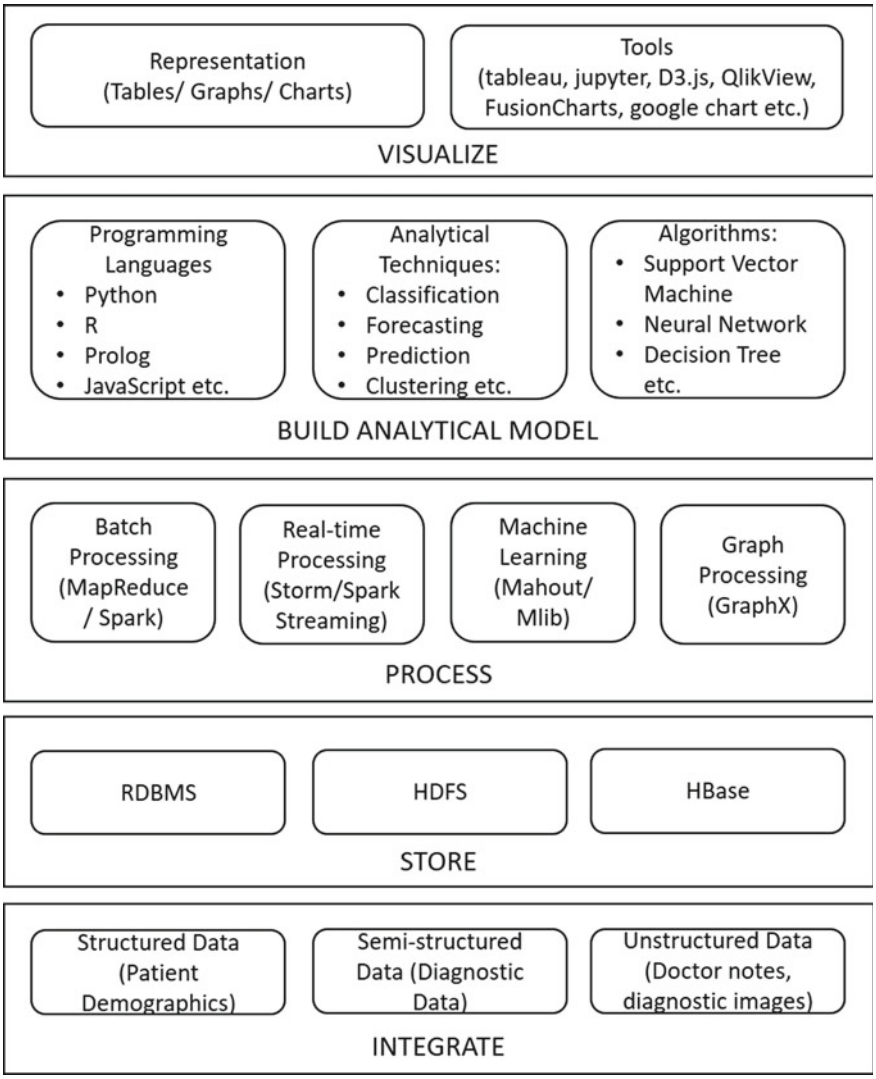


Fig. 3 Big data ecosystem

GraphX for graph processing capabilities [36]. Data processing layer, thus forms the interface between data storage and analytics.

- **Building Analytical Models**: Analytics is of paramount significance for deriving value from health-related data. To facilitate data-driven optimization and for real-time decision-making in emergency situations, advanced analytical models are essential [34]. These models should explicitly identify the areas where additional value can be created by their use; the users who will make use of them; and the measures to evade inconsistencies in scaling the model. Programming

languages such as Python, R, NoSQL, Prolog, JavaScript etc. can be used for the development of analytical models. Developing the analytical model also requires selection of right algorithms and variables to mine the data for achieving desired objectives [37]. Once the objective is clear; descriptive, predictive or prescriptive analytics can be used to design or recode models or algorithms. Different statistical techniques can be applied and machine learning algorithms can be built for pattern recognition, classification, forecasting, prediction and clustering. As an example, convolutional neural network and genetic algorithm can find application in classification of brain tumor MRI scans [38]; and random tree and support vector machine can be employed by health insurance providers for automating preauthorization/claim approval process [39].

- **Decision-Support Tools**: Before the results of analytical model can be made use of, they need to be correlated and used in synchronicity with existing organizational data in order to produce their worth [33]. Intuitive tools are essential for integrating the outputs of analytical model and operational data and converting them into tangible insights to aid decision-making [34]. In fact, the sophisticated algorithms and tools are also vital for real-time insights in case of medical emergencies.
- **Data Visualization**: Whilst the results of decision-support tools are available, they are of little value unless communicated to the decision-makers and integrated into operational processes [31]. Imparting such results to the executives require clear and concise communication in some form of visualization or reporting to streamline the access to information [35]. Results can be represented in graphical form, charts, figures or key findings; over dashboards for ease of access and interpretability. Tools such as tableau, jupyter, D3.js, QlikView, FusionCharts, google chart can be utilized for visualization.

3.3 Human Element

Though the advancement in computing technologies and analytical tools has enabled better analysis of data and prediction of events, it is the human element which is at the core. Beginning with the definition of objectives of big data analytics to identification of relevant data and interpretation of results, there is a need of people at every stage. According to EY and Forbes Insights Survey [40], data analytics demands the right set of people who can apprehend the results of analytical models and establish which information is important. It is the human intuition and skills that enable extraction of important inferences from the processed information and reason behind that information. In healthcare, where the services involve human lives, need for clinical data experts who can elucidate the output of analytical tools is of supreme importance for treating critically-ill patients and saving lives.

3.4 Security Elements

While security is one of the leading concerns in healthcare, in order to tackle this issue there is a necessity to concentrate on the elements of big data security [41]:

- **Network Security**: One of the topmost priorities of healthcare providers is to protect the confidential patient information. Compromised security can impede the day-to-day processes and can thus have life-threatening effects on patients [42]. Healthcare organizations should pay utmost attention for securing the network and systems, by creating strong network firewalls and highly secure wireless networks and VPNs. As a matter of fact, segmenting the network can also prevent the entire network from being affected even when a part of it is.
- **Data Protection**: Attempts to safeguard the healthcare data should emphasize on all the aspects of data protection. The healthcare provider [43]:
  - should have a clear and well-defined purpose for collecting data.
  - should collect the relevant and necessary data.
  - should ensure the accuracy of data.
- **Vulnerability Management**: Healthcare organizations should deploy good vulnerability management tools which have the ability to identify and monitor vulnerabilities in their assets. This system must be competent enough to scan and establish the presence of vulnerabilities, present reports about the seriousness of issues, remediate by taking corrective actions and validate them [44].
- **Access Control**: The complexity of healthcare makes it difficult to implement regular access control mechanism. For the health-related data to be readily available in emergency situations, the default access control rule of "denying access when in doubt" is replaced in healthcare by "allowing access when in doubt". Although this might pose a threat to patient confidentiality, measures for retrospective auditing must be adopted in order to strike a balance between confidentiality and availability [45].
- **Monitoring**: Even when the organizations have preceding elements in place, the continual monitoring and speedy response would ensure higher security of the healthcare system. Being aware of the points of data flow and repositories of data storage, monitoring of these strategic nodes at a convenient level and reporting the risks when observed is the suitable approach for securing the big health data [41].
- **Policies**: Ultimately having a distinct policy which describes the best practices [41] for big data security in healthcare would guide the employees to better apply the aforementioned elements.

4 Foundational Steps of Big Data Analytics Program

Having known about the primary elements for big data analytics, healthcare organizations can plan out the application of this technology. Prior to the implementation of big data analytics, there is need for designing a scheme detailing about the strategic initiatives an organization should take. This section presents a framework to aid the healthcare providers for adoption of analytics and hence provides a strong foundation for strategic viewpoint of big data (Fig. 4).

4.1 Strategic Roadmap

Healthcare organizations should design a big data roadmap which can direct the progress in this area. This roadmap should define the goals and strategy to apply analytics in order to support clinical care delivery and operational efficiency [46]. It should comprise of framing the guidelines on adoption of big data analytics, and understanding how the adoption should iterate and what steps need to be taken in order to overcome the hurdles, if any. The comprehensive roadmap should concentrate on collaborative decision-making that would yield a detailed strategic plan.

4.2 Lucid Business Case

Before its adoption, the organizations should identify the business challenge that can be addressed by making use of big data analytics. Initiatives for development of business case need to follow the recognition of business problems. The establishment of business case begins with definition of the objectives [47] for the adoption of technology, its implementation and operation [48]. This step also involves [49]:

- **Collection of Relevant Data**: As the data is generated at different points in healthcare delivery, distinguishing the right data, gathering and integrating it from varied sources is highly significant. Where the data is available in semi-structured or unstructured format, it should be converted into the usable format.
- **Determination of Suitable Analytical Models**: So as to enable predictions or insights from data, there is a need to identify the right analytical model. This should uncover the areas where the model would yield productive results and designate the people who would use it.
- **Integration of Big Data Analytics into Work Processes**: For the output of analytical model to be of value for organizations, there is need for tools that can integrate data into the operational processes and provide actionable outputs.

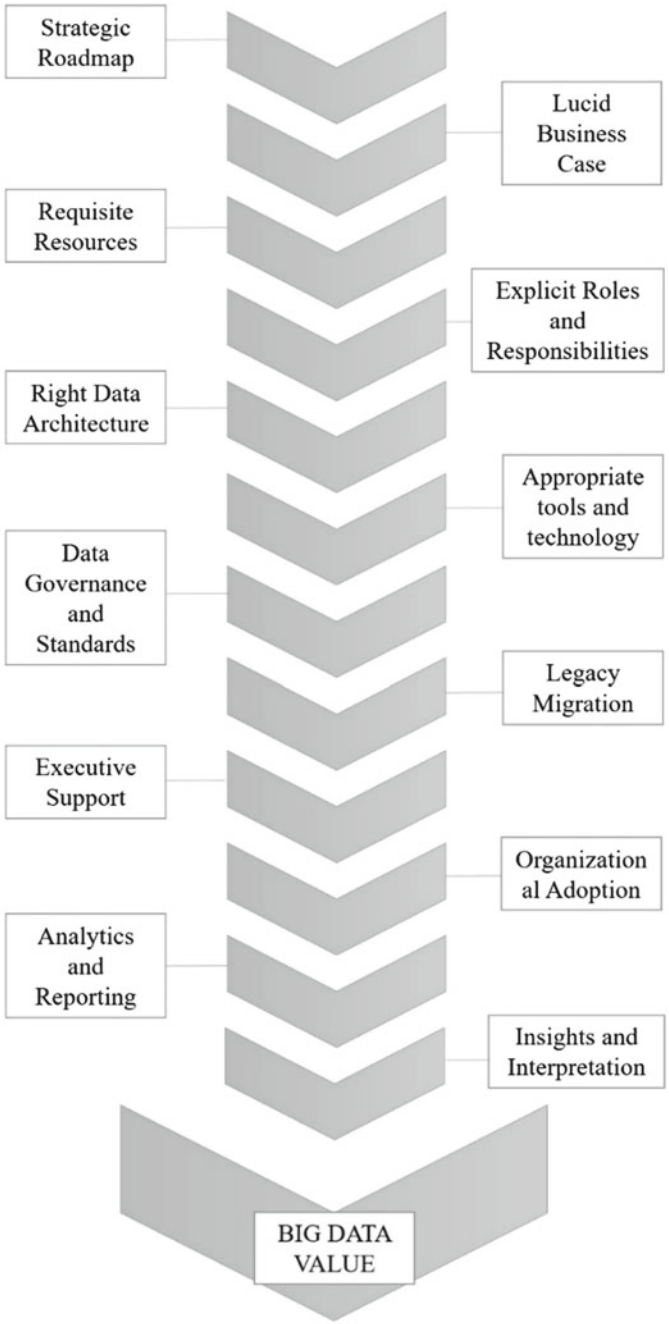


Fig. 4 Foundational steps for big data analytics program

4.3 Requisite Resources

Adequate resources necessary for implementation of big data technology should also be available with the organization. This not only includes monetary resources for procurement of big data algorithms and analytical tools, but also human resource which forms the crucial part of successful adoption of technology [47]. The absence of right people at the right place with the right organizational culture, renders all the data and analytical tools useless [49]. Hiring the people with appropriate skills and expertise, and training them for the specified role is another step towards building the right big data team [50]. Moreover, while thinking of the tangible resources, organizations must consider using third-party resources or outsourcing. Sharing of resources might benefit them in terms of time and cost-effectiveness [47, 51].

4.4 Explicit Roles and Responsibilities

Once the organization has pooled in people with domain-knowledge and analytical expertise, the subsequent act of clearly defining their roles and responsibilities [47] is paramount. There is also a need for clear authority and leadership for such big data initiative, one who is responsible for all big data activities and is accountable for its performance [52]. Besides the human resource, the key resources also need to be allocated to the specific processes so as to expedite the implementation of big data technology.

4.5 Right Data Architecture

Designing the healthcare big data architecture is a cumbersome task involving the selection of vendor, framing the strategy for implementation, planning of the infrastructure, and backup and disaster recovery plan [53]. It also takes into consideration the establishment of right source of data and formulating the clear data governance guidelines [54]. Once the right data architecture and platform are available, organizations can store, process and analyze data at scale [49].

4.6 Appropriate Tools and Technologies

It is essential for organizations to identify the appropriate tools and technologies for their business case [47]. From the acquisition of data to its processing and analysis, tools are extremely important part of generating insights from data [49]. Also, the technology solutions which are in alignment with the defined objectives should be

utilized for better outcomes. These tools and technologies vary from data repositories or cloud storage solutions and from data discovery tools to monitoring tools.

4.7 Data Governance and Standards

To leverage the data for increased productivity, organizations should form synergy amongst its people, processes and technology. This would enable improved understanding between these elements and would have a direct impact on data quality [55]. Data quality and its reliability, is critical for healthcare-decision making. Hence, appropriate controls should be established for maintaining the accuracy and consistency of healthcare data. Having the right standards and data governance plan in place would enable adequate level of integrity for patient safety and better patient experiences [56].

4.8 Legacy Migration

Migration of legacy system into big data ecosystem is needed in order to exploit the potential of big data technology. This stage involves comprehending the design of legacy systems and understanding how that can be translated into newer technology [57]. In case the healthcare data is stored in paper-based records, the migration of data would demand for additional labor to digitize this data.

4.9 Executive Support

For successful implementation of big data analytics, another important aspect is securing management support [51]. With the executives buying-in the idea of utilization of big data technology, they can convince the clinicians for advantages of its use, help build the essential infrastructure and ensure the reliability of data that would be used for clinical-decision support [58].

4.10 Organizational Adoption

Cultural adoption of big data analytics must be fully appreciated so as to accelerate the strategic journey [59]. It necessitates the organizations to build a shared vision with the physicians and healthcare professionals to bring about a transformational change in their operations [60]. The organizational alignment of big data projects; willingness of people to adopt and implement the technology; development

of information sharing culture; efficient communication among different areas; and establishment of data-driven culture [61] forms the key to success.

4.11 Analytics and Reporting

Healthcare organizations need prompt, accurate and reliable analytics and reporting solutions which can assist the clinicians in timely decision-making [62]. These solutions would allow extraction of information from the data which is analyzed and reporting of this information. Detailed or summarized information is presented in the tabular or graphical form for better visualization. This provides the decision-makers with the ability to segment and mine the information to uncover meaningful information [63].

4.12 Insights and Interpretation

Associating the healthcare data with existing organizational data and finding correlations between different data elements is crucial for clinical decision-making [33]. Drawing conclusions from the analyzed data leads to insights. For better data-driven success, extraction of actionable insights is necessary. Deriving actionable insights from amongst the insights obtained after analysis, their interpretation and application in the right context by clinical analytics experts is the ultimate step towards achieving the complete benefits of big data analytics [26].

5 Future Directions

Big data analytical tools and techniques have tremendous potential in assisting healthcare professionals and improving the care delivery process. However, the success with big data analytics for healthcare relies largely upon the access to right data which is clean, complete and accurate. In order to ensure the superior quality of data, there is a need for healthcare industry to adopt practices for better data storage, management and information exchange. Regulations for standardization of electronically captured healthcare data and interoperability of healthcare devices will also enhance handling of complex healthcare data.

Once the relevant data is available to run analytics over, big data analytics will find application in clinical care as well as operational management. Use of artificial intelligence and extraction of free-text data from clinicians' notes, will provide structured data as an input to descriptive analytical tools which can be utilized for activities related to population health management. For example, identifying disease-prone areas for better resource allocation.

Healthcare industry is also taking interest in emerging technologies including artificial intelligence and machine learning. With the potential for using descriptive data to predict outcomes, these technologies are expected to drive evidence-based practice and improve clinical decision support. Use of predictive analytics models will assist in accurate diagnoses and treatment. Moreover, it will facilitate early identification of at-risk patients for chronic diseases or for readmission. Making use of operational analytical algorithms will thereby help reduce the cost of care.

Advanced data analytics tools will not just assist in predicting outcomes, but will move a step beyond to prescribe best possible solution for the problem. Prescriptive analytics will provide healthcare professionals with evaluation of all the possible actions and suggest most appropriate course of action. It will assist healthcare practitioners to prescribe a particular treatment procedure from amongst the probable treatment procedures indicated by analytical model. Also, prescriptive analytics will promote quicker decision-making in the areas related to resource allocation, inventory management, manpower requirement etc.

6 Conclusion

As the healthcare providers proceed towards adoption of big data analytics for improving patient care and enhancing operational efficiency, above mentioned steps will form the premise for successful strategic initiatives. They will not only provide the right direction for integration of big data analytics into organizational operations and processes, but will also enable better gains from technological advancement. Although the initial measures towards adoption of big data analytics would require tremendous efforts on the part of healthcare providers, the end result will be transformational for enhancing healthcare delivery.

7 Summary Points

- In the recent past, healthcare has seen a growing interest in adoption of big data technologies.
- Healthcare organizations have not been able to exploit the full potential of big data analytics because of various challenges.
- Concerns related to health data collection and storage, its quality, security and confidentiality have formed the initial stumbling blocks.
- Dearth of people with desired skillset and expertise; complexity of healthcare domain; lack of knowledge of the benefits of technology; and resistance to adoption are some of the other challenges that impede the implementation of big data analytics.
- While the challenges persist, being aware of the key components of healthcare big data analytics marks the beginning of effective adoption of technology.

- The elemental components required for successful healthcare big data analytics includes: (i) Data inputs; (ii) Functional elements (data preparation, data processing, analytical model, visualization); (iii) Human element; and (iv) Security elements (network security, data protection, access control, policy formulation).
- As the healthcare organizations become familiar with basic elements, they can strategize the adoption of big data analytics.
- Fundamental steps of big data analytics program for healthcare organizations comprises of: (i) creating a strategic roadmap; (ii) clear definition of business use case; (iii) determining and procuring necessary resources; (iv) explicit definition of roles and responsibility of big data team; (v) designing appropriate big data architecture; (vi) acquiring suitable tools and technology; (vii) framing data governance plan; (viii) migration from traditional technology to big data analytics; (ix) backing by management; (x) culture change; (xi) analysis of data and visualization; and (xii) generation of insights and interpretation for intelligent application.

References

1. Barry, M.J., Edgman-levitan, S., Billingham, V.: Shared decision making—The Pinnacle of patient-centered care 780–781 (2016)
2. Thygeson, M., Frosch, D.L., Carman, K.L., Pande, N.: Patient + family engagement in a changing health care landscape. Gordon Betty Moore Found. Conv. Patient Fam. Engagem. (2014)
3. Cognos. Performance Management in Healthcare, Management 1–15 (2008). http://www-07.ibm.com/solutions/au/healthcare/pdf/IBM_Cognos_white_paper_performance_management_in_healthcare.pdf
4. Belle, A., Thiagarajan, R., Soroushmehr, S.M.R., Navidi, F., Beard, D.A., Najarian, K.: Big data analytics in healthcare, Hindawi Publ. Corp. 1–16 (2015). https://doi.org/10.1155/2015/370194
5. Jee, K., Kim, G.H.: Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc. Inform. Res. **19**, 79–85 (2013). https://doi.org/10.4258/hir.2013.19.2.79
6. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**, 3 (2014). https://doi.org/10.1186/2047-2501-2-3
7. Kraft, M.R., Desouza, K.C., Androwich, I.: Data mining in healthcare information systems: case study of a veterans' administration spinal cord injury population, IEEE (2002)
8. Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V.: Critical analysis of big data challenges and analytical methods. J. Bus. Res. **70**, 263–286 (2017). https://doi.org/10.1016/j.jbusres.2016.08.001
9. Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., Shahabi, C.: Big data and its technical challenges. Commun. ACM **57**(7), 86–94 (2014)
10. Keen, J., Calinescu, R., Paige, R., Rooksby, J.: Big data + politics = open data: the case of health care data in England. Policy Internet **5**, 228–243 (2013). https://doi.org/10.1002/1944-2866.POI330
11. Salas-Vega, S., Haimann, A., Mossialos, E.: Big data and health care: challenges and opportunities for coordinated policy development in the EU. Health Syst. Reform. **1**, 285–300 (2015). https://doi.org/10.1080/23288604.2015.1091538

12. Ernst & Young: Big data: changing the way businesses compete and operate (2014). http://www.ey.com/Publication/vwLUAssets/EY_-_Big_data:_changing_the_way_businesses_operate/$FILE/EY-Insights-on-GRC-Big-data.pdf
13. University of Illinois: Why data security is the biggest concern of health care (2019). https://healthinformatics.uic.edu/blog/why-data-security-is-the-biggest-concern-of-health-care/. Accessed 12 Mar 2019
14. Meingast, M., Roosta, T., Sastry, S.: Security and privacy issues with health care information technology. In: Conference on Proceedings of IEEE Engineering in Medicine and Biology Society, vol. 1, pp. 5453–5458 (2006). https://doi.org/10.1109/iembs.2006.260060
15. Ingrammicroadvisor: The top six challenges of healthcare data management (2017). http://www.ingrammicroadvisor.com/data-center/the-top-six-challenges-of-healthcare-data-management. Accessed 30 May 2018
16. Suyati: 8 challenges of big data in healthcare (2015). https://suyati.com/blog/8-challenges-of-bigdata-in-healthcare/. Accessed 3 June 2018
17. UNC Kenan-Flagler: 4 steps to bridge the big data talent gap (2016). http://execdev.kenan-flagler.unc.edu/blog/4-steps-to-bridge-the-big-data-talent-gap. Accessed 4 June 2018
18. Big data talent shortage a potential challenge at fortune 1000 organizations (2012). http://newvantage.com/wp-content/uploads/2012/11/NVP-Press-Release-Big-Data-Talent-Survey-111312.pdf
19. Agrawal, D., Bernstein, P., Bertino, E.: Challenges and opportunities with big data 2011-1. Proc. VLDB Endow. 1–16 (2011). http://dl.acm.org/citation.cfm?id=2367572%5Cnhttp://docs.lib.purdue.edu/cctech/1/
20. Zhou, L., Hripcsak, G.: Temporal reasoning with medical data—a review with emphasis on medical natural language processing. J. Biomed. Inform. **40**, 183–202 (2007). https://doi.org/10.1016/j.jbi.2006.12.009
21. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinform. 1–11 (2017). https://doi.org/10.1093/bib/bbx044
22. Kuipers, P., Kendall, E., Ehrlich, C.: Complexity in healthcare: Implications for clinical education. Focus Health (2013). https://search.informit.com.au/documentSummary;dn=155911887123225;res=IELHEA
23. TEK System: Executive summary big data: the next frontier (2013)
24. Sharma, A.: Organizations continue to face challenges with big data: lets deep dive. Anal. India Mag. (2018). https://analyticsindiamag.com/organizations-continue-to-face-challenges-with-big-data-lets-deep-dive/. Accessed 5 June 2018
25. Marx, V.: Biology: the big challenges of big data. Nature **498**, 255–260 (2013). https://doi.org/10.1038/498255a
26. Dykes, B.: Actionable insights: the missing link between data and business value. Forbes (2016). https://www.forbes.com/sites/brentdykes/2016/04/26/actionable-insights-the-missing-link-between-data-and-business-value/#4e89b7cd51e5
27. Harvey, C.: Big data challenges. Datamation (2017). https://www.datamation.com/big-data/big-data-challenges.html. Accessed 7 June 2018
28. Bean, R.: Executives report measurable results from big data, but challenges remain. Forbes (2017). https://www.forbes.com/sites/ciocentral/2017/01/10/executives-report-measurable-results-from-big-data-but-challenges-remain/#536770a92287
29. Bresnick, J.: Top 10 challenges of big data analytics in healthcare. Health IT Anal. (2017). https://healthitanalytics.com/news/top-10-challenges-of-big-data-analytics-in-healthcare (accessed June 6, 2018)
30. Balachandran, B.M., Prasad, S.: Challenges and benefits of deploying big data analytics in the cloud for business intelligence. Proc. Comput. Sci. **112**, 1112–1122 (2017). https://doi.org/10.1016/j.procs.2017.08.138
31. Marr, B.: Building your big data infrastructure. Forbes (2016). https://www.forbes.com/sites/bernardmarr/2016/06/15/building-your-big-data-infrastructure-4-key-components-every-business-needs-to-consider/#39f6d7e7577e

32. Moore, C.: Big data and analytics—five foundational elements. SiriusDecisions. (2014). https://www.siriusdecisions.com/blog/big-data-and-analytics-five-foundational-elements. Accessed 8 June 2018
33. Singh, P., Mathur, R., Mondal, A., Bhattacharya, S.: The big 'V' of big data. Analytics (2014). http://analytics-magazine.org/the-big-v-of-big-data/
34. Biesdorf, S., Court, D., Willmott, P.: Big data: what's your plan? McKinsey Co. (2013). https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-whats-your-plan. Accessed 7 June 2018
35. Güemes, C.: Data analytics as a service: unleashing the power of cloud and big data 18 (2013)
36. Grover, M.: Processing frameworks for Hadoop. O'Reilly Media (2015). https://www.oreilly.com/ideas/processing-frameworks-for-hadoop. Accessed 14 Mar 2019
37. Eckerson, W.: Analytical modeling is both science and art. TechTarget (2013). https://searchbusinessanalytics.techtarget.com/opinion/Analytical-modeling-is-both-science-and-art. Accessed 4 June 2018
38. Anaraki, A.K., Ayati, M., Kazemi, F.: Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms. Biocybern. Biomed. Eng. **39**(1), 63–74 (2019)
39. Araújo, F.H., Santana, A.M., Neto, P.D.A.S.: Using machine learning to support healthcare professionals in making preauthorisation decisions. Int. J. Med. Inform. **94**, 1–7 (2016)
40. Ernst & Young Analytics: Don't forget the human element. Forbes Insights 1–42 (2015). http://www.forbes.com/forbesinsights/ey_data_analytics_2015/
41. Ottenheimer, D.: The six elements of securing big data. MapR (2016). https://mapr.com/ebooks/securing-big-data-six-elements/chapter-6-six-elements-of-big-data-security.html
42. Cisco Healthcare Security Perspectives: Protect your patients, your practice, yourself technical implementation guide cisco healthcare security perspectives (2007). https://www.cisco.com/c/dam/global/en_ca/solutions/strategy/healthcare/assets/docs/health_security_impgd.pdf
43. European Patients Forum: The new EU regulation on the protection of personal data: what does it mean for patients? A guide for patients and patients' organisations 2 The new EU Regulation on the protection of personal data: what does it mean for patients? 23 (2017). http://www.eu-patient.eu/globalassets/policy/data-protection/data-protection-guide-for-patients-organisations.pdf
44. Eagle Consulting Partners: Health IT security risks and vulnerability management (2016). https://eagleconsultingpartners.com/newsletter/do-you-have-a-healthcare-it-vulnerability/. Accessed 8 June 2018
45. Røstad, L.: Access control in healthcare information systems (2009). http://www.idi.ntnu.no/research/doctor_theses/lilliaro.pdf
46. IBM, Health Analytics Roadmap (n.d.). https://www-03.ibm.com/industries/ca/en/healthcare/files/health_analytics_roadmap.pdf
47. Allen, B., Coates, S., Karp, B., Lerchner, H.-P., Lourens, J., Penrose, T., Rai, S.: Understanding and auditing big data 22–23 (2017). https://www.iia.nl/SiteFiles/Publicaties/GTAG-Understanding-and-Auditing-Big-Data.pdf
48. Think Big: Big data strategy definition 1–2 (n.d.). https://www.thinkbiganalytics.com/wp-content/uploads/2017/02/Think-Big-Big-Data-Strategy-Definition-Data-Sheet.pdf
49. Hodgson-Abbott, G.: 12 key components of your data and analytics capability. Capgemini (2017). https://www.capgemini.com/gb-en/2017/01/12-key-components-of-your-data-and-analytics-capability/. Accessed 11 June 2018
50. Eccella: Big data is nothing without the right people and culture (2016). http://blog.eccellaconsulting.com/human-element-big-data. Accessed 10 June 2018
51. Smith, S.: 5 elements of a successful data analytics program. Audimation Serv. Inc. (2017). www.audimation.com/Resources/Articles/5-elements-of-a-successful-data-analytics-program-3073. Accessed 12 June 2018
52. Eckerson, W.: Analytics program requires adaptability, collaboration and results. TechTarget (2012). https://searchbusinessanalytics.techtarget.com/feature/Analytics-program-requires-adaptability-collaboration-and-results. Accessed 12 June 2018

53. Pal, S.: How to design a big data architecture in 6 easy steps. Saama (2016). https://www.saama.com/blog/design-big-data-architecture-6-easy-steps/. Accessed 9 June 2018
54. Kalvakuntla, R.: How a modern data architecture will revolutionize the healthcare industry. https://cdn2.hubspot.net/hubfs/2106877/WhitePapers/MDA_revolutionizing_HC_Industry_WP_pdf?t=1528979762553&utm_campaign=ModernDataWarehouse2017&utm_source=hs_automation&utm_medium=email&utm_content=57227465&_hsenc=p2ANqtz-97yeTB_LwuzoJz6NBzmARnUIKKXmE5ADS48QDItkzwRb9ygE44b3QfCBxAAG7As_JgmOcyQi-4U2GFXTyloT932t54XQ7mL929rWBPsYuJbknNM2s&_hsmi=57227465
55. Johnson, J.: The intersection of data analytics and data governance. Infogix (2018). https://www.infogix.com/blog/the-intersection-of-data-analytics-and-data-governance/. Accessed 12 June 2018
56. Bresnick, J.: 56% of hospitals lack big data governance. Analytics plans. Health IT Anal. (2017). https://healthitanalytics.com/news/56-of-hospitals-lack-big-data-governance-analytics-plans. Accessed 11 June 2018
57. Stoltzfus, J.: What are some of the biggest challenges with legacy migration? Techopedia (2016). https://www.techopedia.com/7/32194/technology-trends/big-data/what-are-some-of-the-biggest-challenges-with-legacy-migration. Accessed 10 June 10 2018
58. Bresnick, J.: Healthcare big data analytics confuses half of providers. Health IT Anal. (2015). https://healthitanalytics.com/news/healthcare-big-data-analytics-confuses-half-of-providers. Accessed 11 June 2018
59. Infomgmtexec: Transformational leadership for big data & analytics success—part 3: organizational design & cultural adoption (2014). https://infomgmtexec.me/2014/07/20/transformational-leadership-for-big-data-analytics-success-part-3-organizational-design-cultural-adoption/. Accessed 11 June 2018
60. Bean, R.: Why cultural change is necessary for big data adoption. Forbes (2016). https://www.forbes.com/sites/ciocentral/2016/11/08/another-side-of-big-data-big-data-for-social-good-2/#1035e34a6628
61. Félix, B.M., Tavares, E., Cavalcante, N.W.F.: Critical success factors for big data adoption in the virtual retail: magazine Luiza case study. Rev. Bus. Manag. **20**, 112–126 (2018). https://doi.org/10.7819/rbgn.v20i1.3627
62. Optum Clinical Solutions: Optum perioperative analytics and reporting solutions. https://www.optum.com/content/dam/optum/resources/brochures/Analytics_Reporting_Solutions_Brochure.pdf
63. Health E Connections: Analytics & reporting (2015). https://www.healtheconnections.org/what-we-do/analytics-and-reporting/. Accessed 13 June 2018

Big Data in Supply Chain Management and Medicinal Domain

Aniket Nargundkar and Anand J. Kulkarni

Abstract This chapter presents the fundamental and conceptual overview of big data describing its characteristics. The chapter covers two domains viz. Supply Chain (SC) and Medicinal (Healthcare) industry. Under SC domain, data generation process is explained. The difference between big data and traditional analytics is clarified. Landscape of SC is described with specific case studies in central areas of application. The typical big data platforms used in supply chain are elaborated with comparison. Prominent platform NoSQL is described comprehensively. Contemporary methodologies of big data analytics in supply chain are illustrated. Second part of chapter deals with healthcare domain. Importance of big data in medicinal domain is highlighted. The overall process of big data analytics from data generation till data results visualization is exemplified. Upcoming trends of big data analytics with wearable or implanted sensors is explicated. At the end, overall big data advantages and limitations are discussed along with future direction.

1 Introduction

The continuous increase in the usage of social media, Internet of Things (IoT), and multimedia, has produced a vast flow of data in either structured or unstructured format. Data creation is occurring at a record rate. Big data and its analysis are at the core of modern technology and business. These data are generated from online transactions, emails, videos, audios, images, and search over internet, social networking interactions, science data, sensors and mobile phones and their applications. Big data is referred as data sets that are huge for traditional data-processing application software. Data with many rows possesses better statistical insights, whereas data with higher no of columns may lead to a greater false discovery rate. The term big data has

A. Nargundkar (✉) · A. J. Kulkarni
Symbiosis Institute of Technology, Symbiosis International (Deemed University), Lavale, Pune 412115, India
e-mail: aniket.nargundkar@sitpune.edu.in

A. J. Kulkarni
e-mail: anand.kulkarni@sitpune.edu.in

A. J. Kulkarni et al. (eds.), *Big Data Analytics in Healthcare*,
Studies in Big Data 66, https://doi.org/10.1007/978-3-030-31672-3_3

been in use since the 1990s, credit to John Mashey for popularizing it [1, 2]. Big data includes data sets with sizes beyond the ability of commonly used software tools to capture and process data within a stipulated time [3]. Big data viewpoint incorporates unstructured, semi-structured and structured data. With the growing population and usage of modern tools, big data is present in almost every field.

McKinsey Global Institute identified big data in five themes [4]:

- Healthcare: decision support systems, individual analytics for patient profile, personalized medicine, analyze disease patterns, improve public health.
- Public sector: creating transparency by accessible related data, decision making with automated systems.
- Retail: in consumer behavior analysis, variety and price optimization, distribution and logistics optimization.
- Supply Chain Management (SCM): improved demand forecasting, supply chain planning, sales support, developed production operations, web search based applications.
- Personal location data: smart routing, geo targeted emergency response, urban planning.

1.1 Characteristics of Big Data

Big data is characterized by five "V" viz. volume, variety, value, velocity and veracity. Figure 1 describes the five V of big data.

Volume

Volume describes the quantity of generated and stored data. The size of the data determines whether it can be considered big data or not.

Variety

Fig. 1 Five V of big data

It is the type and nature of the data. Data can be in the form of text, images, audio, video or fusion data combining several data types.

Velocity

Velocity is the speed at which the data is generated and processed. Compared to small data, big data are produced more continually. Velocity is required not only for big data generation but also all processes to handle the big data.

Veracity

It is the extended definition for big data, which refers to the data quality. Veracity is the quality or trustworthiness of the data.

Value

When we talk about value, we're referring to the worth of the data being extracted.

1.2 Big Data Analytics

Big data analytics is defined as applying advanced analytic methods including data mining, statistical and predictive analytics, etc. on big datasets [5]. It is the procedures of scrutinizing and investigating vast data to conclude by discovering hidden patterns and correlations, trends. It ultimately helps to improve business aids, upturn operational efficiency, and explore market prospects.

The section two of this article presents overview of big data analytics in supply chain management; section three gives big data overview of healthcare and medicinal domain and at the end future directions are explored.

2 Big Data in Supply Chain Management

Supply chain refers to the chain of movement of products and/or information from raw stage to customers. Supply chain analysis involves three aspects viz. descriptive analytics, predictive analytics and prescriptive analytics. Descriptive analytics deals with the question of what has happened, what is happening, and why. It attempts to identify opportunities and problem using online analytical processing system and visualization tools supported by real time information and reporting technology. Predictive analytics deal with what will be happening or likely to happen, by exploring data pattern using statistics, simulation, and programming. It attempts to accurately predict the future and discover the reason. Prescriptive analytics deal with what should happen and how to influence it, by considering other decisions based on other analytics, using mathematical models, simulation etc. Prescriptive analytics are relatively complex and it is yet to be commercialized [6].

Table 1 Big data and traditional analytics

Dimensions	Big data	Traditional analytics
Type of data	Unstructured formats	Formatted in rows and columns
Volume of data	100 terabytes to petabytes	Tens of terabytes or less
Flow of data	Constant flow of data	Static pool of data
Analysis methods	Machine learning	Hypothesis based
Primary purpose	Data-driven products	Internal decision support and services

2.1 Traditional Analytics and Big Data Analytics

The conventional information management approach is to use analytics for providing better decision-making through reports and presentations. In contrast contemporary approach is to develop data driven products. Table 1 shows big data versus traditional data analytics in supply chain industry [7].

2.2 Landscapes of Supply Chain

The application areas of big data in supply chain varies from supplies to customers and from intrinsic to external stakeholders. There are nine important supply chain areas for big data application. Figure 2 shows these areas as landscapes of supply chain management.

Some of the above mentioned areas are explained below.

Strategic Sourcing: Strategic sourcing is company's strategic partnership which focuses on collaboration and supplier relationship management. The decision along with cost, quality, and delivery also incorporates some strategic dimensions and capabilities of the supplier [8]. Jin and Ji applied analytic hierarchy process (AHP) and fuzzy in choosing supply chain partner considering big data processing capacity as one of the evaluated factors. The objective is to select partner that can adapt to the future challenges from big data [9].

Supply Chain Network Design: Wang et al. developed a mixed integer nonlinear model that utilized big data in selecting the location of distribution centers using randomly generated big datasets for customer demand, warehouse operation and transportation. It is proved that big data could provide additional information therefore creating opportunities for designing complex distribution network [10]. Prasad et al. studied the application of big data analytics to consider intervention such as disaster relief, healthcare and education in specific supply chain network [11].

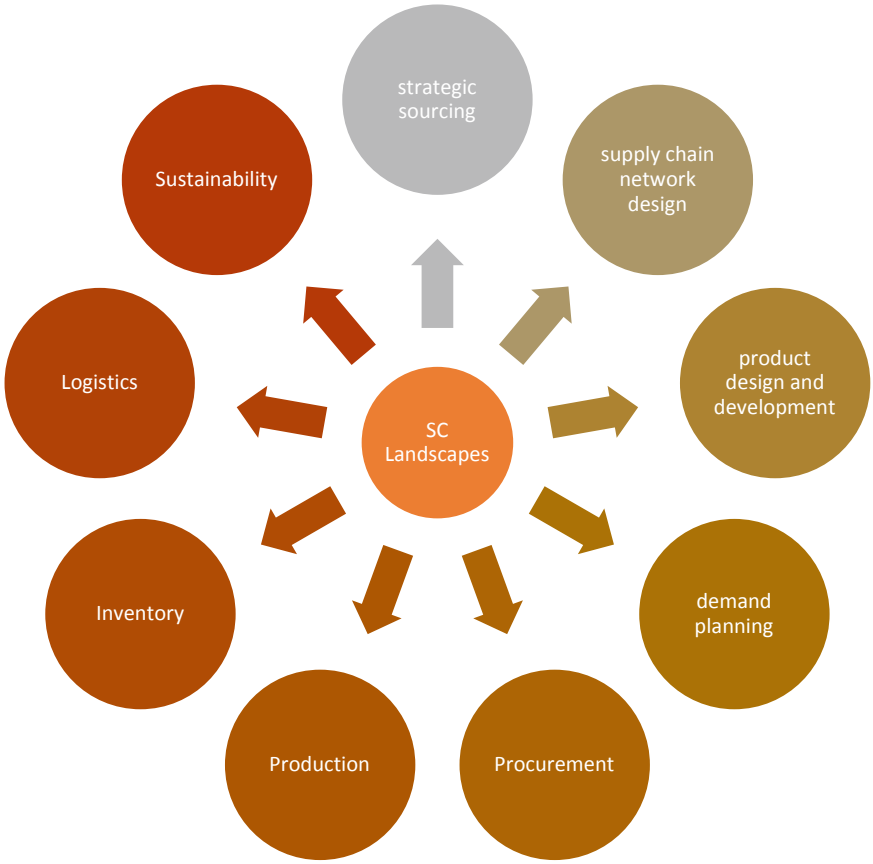


Fig. 2 Landscapes of supply chain

Product Design and Development: Big data analytics in product design and development is becoming popular. Big data analytics can improve product adaptability and give more confidence to the designer [12].
Demand Planning: Arias and Bae combined historical real-world traffic data and weather data in their forecasting model to estimate electric vehicle charging demand [13]. Kim and Shin developed a forecasting model using big data from search engine queries to estimate short-term air passenger demand. The result has helped the airport authority to set appropriate operation plans with an average forecast error of 5.3% [14].

Table 2 Types of NoSQL data stores

Type	Application	Platform
Key-value store—a simple data storage system that uses a key to access a value	Image stores Key-based file systems Object cache Systems designed to scale	Berkeley DB Memcached Redis Riak DynamoDB
Column family store—a sparse matrix system that uses a row and a column as keys	Web crawler results Big data problems that can relax consistency rules	Apache HBase Apache Cassandra Hypertable Apache Accumulo
Graph store—for relationship intensive problems	Social networks Fraud detection Relationship-heavy data	Neo4j AllegroGraph Bigdata (RDF data store) InfiniteGraph (objectivity)
Document store—storing hierarchical data structures directly in the database	High-variability data Document search Integration hubs Web content management Publishing	MongoDB (10Gen) CouchDB Couchbase MarkLogic eXist-db Berkeley DB XML

2.3 Big Data Platforms

Big data analytics of supply chain is based on two major platforms namely Hadoop and NoSQL. Though each technology is widely applied for big data, they are intended for different types of applications. NoSQL is about real-time, interactive access to data. Broadly it is about reading and writing data quickly. Hadoop, on the other hand, is large-scale processing of data. To process large volumes of data, one need to work in parallel, and typically across many servers. Hadoop manages the distribution of work across many servers in a divide-and-conquer methodology known as MapReduce. The better known NoSQL databases are MongoDB, HBase, Cassandra, Couchbase, Aerospike, DynamoDB, MarkLogic, Riak, Redis, Accumulo, and Datatomic.

Table 2 shows the different types of NoSQL data stores.

2.4 Impending Methodologies

Researchers are developing mathematical models for sustainable development of supply chain. By using contemporary metaheuristics, scientists are coming with more intelligent way of handling and analyzing data. Navin Dev et al. proposed big data architecture (BDA) for key performance indicators (KPI). The conceptual framework is proposed under RFID integrated with cloud ERP system. Predictive and prescriptive analytics is explored in the manufacturing operations. Synergies

of offline simulation, fuzzy-ANP, and TOPSIS in BDA are established. Proposed BDA attempts to handle real-time interrelated KPI problem of supply chain [15]. Rui Zhao et al. presents a multi-objective optimization model for a green supply chain management scheme that minimizes the inherent risk occurred by handling hazardous materials, associated carbon emission and economic cost. The model related parameters are capitalized on big data analysis. Three scenarios are proposed to improve green supply chain management. The first scenario divides optimization into three options: the first involves minimizing risk and then dealing with carbon emissions (and thus economic cost); the second minimizes both risk and carbon emissions first, with the ultimate goal of minimizing overall cost; and the third option attempts to minimize risk, carbon emissions, and economic cost simultaneously. Binary Dominance Matrix (BDM) is applied to assess importance of three objectives [16].

3 Big Data in Medicinal Domain

Healthcare industry is one of the world's biggest and widest industries. During, the recent years the healthcare management around the world is changing from disease-centered to a patient-centered model and also to provide more valued services to patients. To provide effective patient-centered care, it is essential to manage and analyze huge health data. The outdated data management implements are not sufficient enough to analyze big data as variety and volume of data sources have increased in the past years. The big data are used to predict the diseases before they emerge based on the medical records. Many countries' public health systems are now providing electronic patient records [17]. The practice of big data takes the prospective to encounter the upcoming market needs and trends in healthcare establishments [18]. Big data provides a great opportunity for epidemiologists, physicians, and health policy experts to make data-driven judgments that will eventually develops the patient care [19].

3.1 Process Flow

The process flow of big data analysis in healthcare in illustrated in Fig. 3.

Data Acquisition: The big data in healthcare can be in a format of the structured, semi-structured or unstructured and can be acquired from primary sources (e.g. electronic health records etc.) and secondary sources (laboratories, insurance companies, government sources, pharmacies etc.) [17].

Data Storage: The storage plays vital part in big data. As a size of data in the healthcare industry is increasing, an efficient and large storage platform is needed. Cloud is the most promising technology. Cloud computing is a powerful and promising

Fig. 3 Process of big data analytics in healthcare

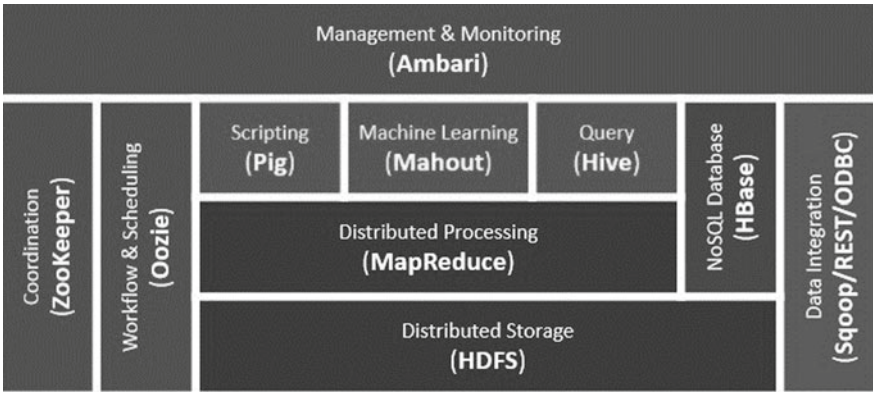


Fig. 4 Hadoop ecosystem

technology to store enormous scale of data and perform complex computing. It eradicates the need to sustain costly computing hardware, software and dedicated space. Numerous cloud storage platforms are available such as Azure S3, Smart Cloud, SQL, NoSQL, Apache etc.

Data Analytics: Data analytics is a process of transforming the raw data into information. Big data analytics in healthcare is classified into Descriptive, Diagnostic, Predictive, and Prescriptive Analytics. Typical analytics platform used for big data analytics in healthcare are Apache Pig, Apache Spark, Hadoop, Bidoop, Apache H, Cassandra etc. [20].

The basic framework and Hadoop Ecosystem is elaborated in Fig. 4.

3.2 Contemporary Methodologies/Case Studies

Researchers are developing new methodologies for big data analytics using latest mathematical models and data acquisition systems such as wearable sensors. Wearable medical devices with sensor continuously generate enormous data which is often called as big data mixed with structured and unstructured data. A case study is discussed based on Manogaran et al. [21]. They have proposed new architecture for the implementation of IoT to store and process scalable sensor data (big data) for

healthcare applications. The proposed architecture consists of two main sub architectures, namely, Meta FA og-Redirection (MF-R) and Grouping and Choosing (GC) architecture. MF-R architecture uses big data technologies such as Apache Pig and Apache H Base for collection and storage of the sensor data (big data) generated from different sensor devices. The proposed GC architecture is used for securing integration of fog computing with cloud computing. This architecture also uses key management service and data categorization function (Sensitive, Critical and Normal) for providing security services. The framework also uses MapReduce based prediction model to predict the heart diseases [21]. Second case study is based on Gui [22]. They have presented architecture for healthcare big data management and analysis. Under the prosed architecture, a prototype system is developed in which HBase and Hive perform as storage components while Kettel and Sqoop are utilized for ETL operations. Finally, Spark Streaming and Spark MLLib are integrated into the prototype system for real-time monitoring and on-line decision support for patients with elevated blood pressure [22].

4 Conclusion

The size of data at present is huge and continues to grow every day. The variety of data being generated is also expanding. The velocity of data generation and growth is increasing because of the increase of mobile devices and other device sensors connected to the internet. Thus, big data is present virtually across all domains of society. This provide opportunities that allow businesses across all industries to gain real-time business insights thru analytics. There are quite a lot benefits of big data analytics such as cost savings, real time monitoring, control and prediction, market or customer insight thru data collected from them etc. However, there exists numerous open loop holes to deal with. Research shows that several studies have addressed a number of significant problems and issues pertaining to the storage and processing of big data in clouds. The amount of data continues to increase at an exponential rate however the improvement in the processing mechanisms is relatively slow. Some of the challenges are data staging viz. dealing with unstructured nature of data, data analysis algorithms and mathematical efficient computing models and data security. These open challenges could be addressed by academia researchers and industry.

References

1. Mashey, J.R.: “Big Data and the Next Wave of Infra Stress” Slides from Invited Talk. Usenix, 25 Apr 1998
2. Lohr, S.: The origins of ‘Big Data’: an etymological detective story. The New York Times, 1 Feb 2013
3. Snijders, C., Matzat, U., Reips, U.-D.: ‘Big Data’: big gaps of knowledge in the field of Internet. Int. J. Internet Sci. **7**, 1–5 (2012)

4. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute (2011)
5. Russom, P.: Big data analytics. TDWI Best Practices Report, Fourth Quarter (2011). http://tdwi.org.
6. Tiwari, S., Wee, H.M., Daryanto, Y.: Big data analytics in supply chain management between 2010 and 2016: insights to industries. Comput. Ind. Eng. (2017). https://doi.org/10.1016/j.cie.2017.11.017
7. Mitra, A., Munir, K.: Big data application in manufacturing industry. In: Sakr, S., Zomaya, A.Y. (eds.) Encyclopedia of Big Data Technologies, pp. 1–7. Springer, UK (2019). ISBN 9783319320090. https://eprints.uwe.ac.uk/35723
8. Panchmatia, M.: Use big data to help procurement' make a real difference (2015). https://www.4cassociates.com
9. Jin, Y., Ji, S.: Partner choice of supply chain based on 3D printing and big data. Inf. Technol. J. **12**(22), 6822–6826 (2013)
10. Wang, G., Gunasekaran, A., Ngai, E.W., Papadopoulos, T.: Big data analytics in logistics and supply chain management: certain investigations for research and applications. Int. J. Prod. Econ. **176**, 98–110 (2016)
11. Prasad, S., Zakaria, R., Altay, N.: Big data in humanitarian supply chain networks: a resource dependence perspective. Ann. Oper. Res. S.I.: Big Data Analytics in Operations & Supply Chain Management (2016). https://doi.org/10.1007/s10479-016-2280-7
12. Afshari, H., Peng, Q.: Using big data to minimize uncertainty effects in adaptable product design. In: ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (2015)
13. Arias, M.B., Bae, S.: Electric vehicle charging demand forecasting model based on big data technologies. Appl. Energy **183**, 327–339 (2016)
14. Kim, S., Shin, D.H.: Forecasting short-term air passenger demand using big data from search engine queries. Autom. Constr. **70**, 98–108 (2016)
15. Dev, N.K., Shankar, R., Gupta, R., Dong, J.: Multi-criteria evaluation of real-time key performance indicators of supply chain with consideration of big data architecture. Comput. Ind. Eng. (2018)
16. Zhao, R., Liu, Y., Zhang, N., Huang, T.: An optimization model for green supply chain management by using a big data analytic approach. J. Clean. Prod. **142**, 1085–1097 (2017)
17. Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., Taha, K.: Efficient machine learning for big data: a review. Big Data Res. **2**, 87–93 (2015). https://doi.org/10.1016/j.bdr.2015.04.001
18. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**(2014), 1–10 (2014). https://doi.org/10.1186/2047-2501-2-3
19. Sessler, D.I.: Big data and its contributions to peri-operative medicine. Anaesthesia **69**, 100–105 (2014)
20. Senthilkumar, S.A., Rai, B.K., Meshram, A.A., Gunasekaran, A., Chandrakumarmangalam, S.: Big data in healthcare management: a review of literature. Am. J. Theor. Appl. Bus. **4**(2), 57–69 (2018). https://doi.org/10.11648/j.ajtab.20180402.14
21. Manogaran, G., et al.: A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system. Future Gener. Comput. Syst. (2017). https://doi.org/10.1016/j.future.2017.10.045
22. Gui, H., Zheng, R., Ma, C., Fan, H., Xu, L.: An architecture for healthcare big data management and analysis. In: Yin, X., Geller, J., Li, Y., Zhou, R., Wang, H., Zhang, Y. (eds.) Health Information Science. HIS 2016. Lecture Notes in Computer Science, vol. 10038. Springer (2016)

A Review of Big Data and Its Applications in Healthcare and Public Sector

Apoorva Shastri and Mihir Deshpande

Abstract Big Data has been a buzzword in the IT sector for a few years now. It has attracted attention from researchers, industry and academia around the world. This chapter is intended to introduce Big data and its related technologies and further trace the challenges. In this chapter, we discuss the applications of big data technologies in the fields of healthcare and public sector. Over the preceding few years, computing power has increased substantially while the storage costs have reduced significantly, leading to businesses being able to produce and store huge volumes of data. Also, increasing penetration of hand-held and internet enabled devices had led to an explosion in data generation. Social media is exemplary regarding this phenomenon. Such huge volumes of data cannot be handled using existing frameworks and requires new and innovative techniques to handle it. In this chapter, we will briefly discuss the use of big data in healthcare and its potential use cases such as preventive healthcare planning and predictive analytics. We will also discuss the potential use of big data in public sector and its applications like urban management and inclusive decision making. We will further highlight the challenges that hinder the potential use of big data technologies in these areas.

1 Introduction

Everyone is talking about Big Data these days. Leading organisations have started to recognise it as a strategic asset [1]. Let's start with a commonly agreed definition of it. Big Data is any data which is large and complex and therefore becomes difficult to process with the traditional storage and processing paradigms, loosely approximated by practitioners as data-sets around 30–50 terabytes and beyond up to petabytes [2].

A. Shastri (✉)
Lovely Professional University, Phagwara 144411, Punjab, India
e-mail: apoorva.shastri@sitpune.edu.in

Symbiosis Institute of Technology, Symbiosis International University, Pune 412115, India

M. Deshpande
School of Information Studies, Syracuse University, Syracuse, NY 13244-1190, USA
e-mail: mdeshpan@syr.edu

A. J. Kulkarni et al. (eds.), *Big Data Analytics in Healthcare*,
Studies in Big Data 66, https://doi.org/10.1007/978-3-030-31672-3_4

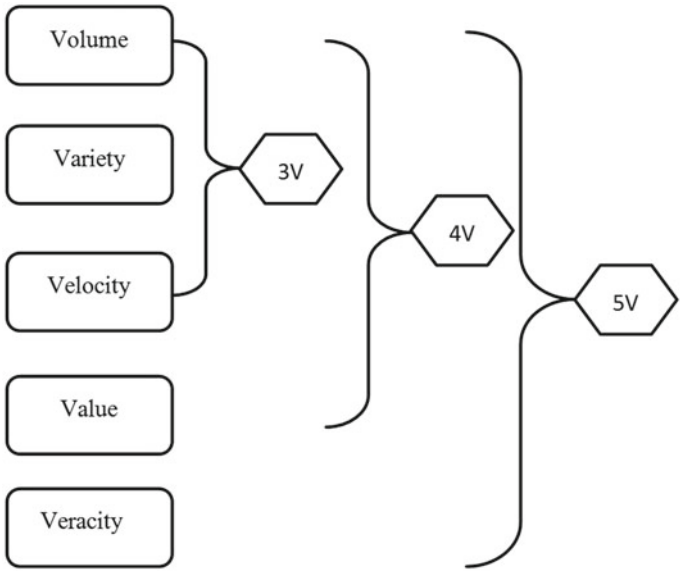


Fig. 1 The V's of big data [4]

For example, The Large Hadron Collider's 150 million sensors generate a data flow of about 15 petabytes or about 15,000,000 GB per year [3]. Therefore, traditional tools and techniques are unable to store, process and visualize it within stipulated amount of time and extract competitive insights. Big data applications can be seen everywhere from scientific community, marketing, banking, telecom to healthcare, public services and so on. It has allowed organisations to take informed decisions based on the insights derived from transactional data created at various points. Big data has been described as a set of 3V's, 4V's and even 5V's by various big data researchers as shown in Fig. 1.

1.1 5 V's of Big Data

(i) **Volume**
It refers to the humungous scale of data. The amount of data that is being generated has increased has been increasing exponentially in the past few years and is expected to continue to do so in the coming future due to reducing storage costs. By 2020, there will be around 6.1 Billion smart phones and our accumulated digital universe will be around 44 trillion gigabytes [5]. Google processes around 40,000 search queries every single second [5]. The mammoth scale of data being generated requires innovative data infrastructure, data management and processing techniques.

(ii) **Velocity**
The rate of flow of data is measured by velocity. Facebook handles around 900 million photographs every day [6]. It must absorb it, process it and later be able to retrieve it. The Data Management infrastructure that follows such high-speed data flows is a vital part of the Big Data Paradigm. Time sensitive processes like banking transactions or social media streaming data are some examples were data is generated, processed and stored in a matter of few seconds.

(iii) **Variety**
The nature of different data forms that exist. They can vary from traditional enterprise structured data, semi-structured data or unstructured data like images, text, audio and video. There are endless heterogeneous data types and sources which are to be dealt with in a Big data paradigm.

(iv) **Veracity**
It the quality associated with big data defined as data of inconsistent, incompetent, deceiving and ambiguous nature. It is also concerned with the reliability and authenticity of the data used for analyses.

(v) **Value**
It is the intrinsic value that the big data holds with respect to its size. If large volumes of imprecise data are analyzed it results in low value and if large volumes of precise data are analyzed it results in high value.

2 Big Data Technologies

"Data, I look at it as the new oil. It's going to change most industries across the board." said Intel CEO Brian Krzanich [7]. There are a plenty of tools and technology for big data processing and storage available today. Early developments like the Google File system which allowed for processing large scale distributed data-intensive applications on inexpensive commodity hardware using a fault tolerant mechanism paved way for further developments in distributed computing [8]. Later, Google developed MapReduce, a programming model based on Java language, which is useful for writing applications to process huge amounts of data, in parallel, on clusters of commodity hardware. Hadoop and Spark are the latest buzzwords in the big data universe these days. The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing in a fault-tolerant manner. The Hadoop framework is a collection of software that enables distributed processing for large sets of data across clusters of computers using simple programming models. It makes use of the Map Reduce model as one of its module. Several projects related to Hadoop that work on the top of Hadoop architecture have been developed and are available to use. Spark is yet another distributed processing framework which has its similarities with Hadoop, but is generally consider much more efficient and fast as compared to Hadoop especially when dealing with queries that are iterative in nature. Few of these technologies will further elaborated.

2.1 Hadoop

As mentioned earlier, Hadoop is a software library that enables distributed processing for large data-sets across clusters of inexpensive commodity hardware [9]. There are two main components of the Hadoop architecture- Hadoop Distributed File System (HDFS) and Hadoop MapReduce. The HDFS deals with the storage part and MapReduce deals with the processing part of the big data problem. The Hadoop ecosystem includes various projects like Pig, Hive, Hbase, Mahout etc. that run on top of the Hadoop architecture.

(i) **HDFS**

It is the distributed storage system used in the Hadoop framework. It is setup on a cluster of low-cost commodity hardware and is incrementally scalable. It is highly fault tolerant with no single point of failure. Input data is broken down into blocks and stored across the cluster. Each block has a typical size of 64 Mb and a default replication factor of 3 which means that a single data block is stored on 3 different locations on the cluster. This allows for recovery of data and continuity of work in case of failure of a node on the cluster. HDFS follows master-slave architecture. A file system cluster has a single master server called NameNode that manages the file system and regulates file access by application clients. A DataNode is run locally on all slave nodes and it manages the storage of that particular node and communicates with the NameNode.

(ii) **MapReduce**

It is the framework that handles the processing part of the Hadoop ecosystem. Owing to the large size of the input data, the computations have to be distributed across several machines in order to finish in a reasonable amount of time. The complexity of distributed processing is abstracted from the developer by the framework. The computation is split and taken to the nodes where data is residing. MapReduce is inspired the Map and Reduce functions present in many functional languages [10]. Map operation is applied on input to create a set of intermediate key/value pairs and reduce operation aggregates values with same key and consolidates the final output [10]. MapReduce programs are typically written in Java language. High level languages like Hive and Pig Latin make it easier for people without the knowledge of Java to write MapReduce applications.

(iii) **Other Components of the Hadoop Ecosystem** [9]

Hadoop Ecosystem along with its components is shown in Fig. 2. Various components of Ecosystem are as follows:

(a) **Hive**
It is a data warehousing tool for processing structured data in Hadoop. It is useful for summarization of data and querying. It has a SQL like syntax for querying.

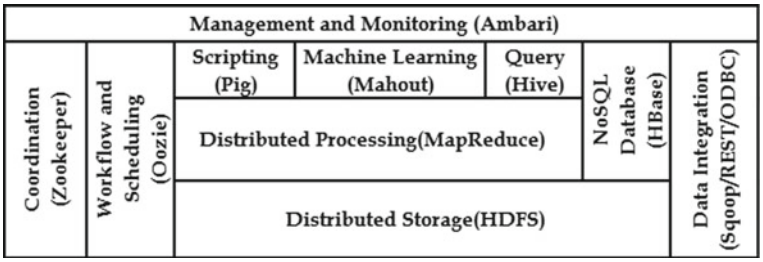


Fig. 2 Hadoop ecosystem [11]

(b) **Pig**
It is a high-level platform for analyzing big data and expressing data analysis programs. Pig's internal layer contains a compiler that produces sequences of Map Reduce programs. Pig Latin is the internal language of the Pig platform.

(c) **Sqoop**
It is a platform that allows communication between relational databases and HDFS. The name itself stands for 'SQL to Hadoop'.

(d) **Mahout**
A library containing algorithms for scalable and distributed machine learning.

(e) **Hbase**
It is a distributed and Non-relational database system for storing large tables. It provides a columnar schema and is a NoSQL database. NoSQL stands for 'Not Only SQL' which is an alternative to traditional relational databases.

(f) **Oozie**
It is a work-flow scheduler that manages Hadoop jobs.

(g) **Zookeeper**
It is a centralised service that maintains configuration information, provides group services and distributed synchronisation.

(h) **Cassandra**
It is a scalable, fault-tolerant and decentralised and a high-performance database solution. It is robust solution for load-balancing, failure detection, failure recovery; overload handling, job scheduling, system monitoring and alarming etc. It is another NoSQL database.

2.2 Apache Spark

It is a big data processing framework built with a focus on high-speed processing and advanced analytics. Spark can outperform Hadoop by 10x in iteratively machine learning and interactive querying jobs [12]. Spark was first developed in 2009 at UC Berkeley. Spark is deployed in number of ways and has local bindings for Scala, Java, Python and R programming and supports graph data processing, streaming

data and SQL [13]. It cannot be directly compared to Hadoop and competes more with MapReduce for big data processing. Spark can run on top of the HDFS [14]. Hadoop MapReduce performs fairly for ad hoc Data summarization and querying but struggles with interactive analysis and iterative jobs like Machine Learning due to its acyclic data flow model [12]. In interactive analysis, where a user would query a dataset repeatedly, each query would be treated as a MapReduce job and read data from disk each time and amount to a significant latency. Also, most machine learning algorithms apply functions repeatedly on the same dataset for optimization of its cost function. If expressed as a MapReduce job, each iteration must reload the data from disk hence incurring performance issues. Spark tries to address this by the use of Resilient Distributed Datasets (RDD) which is a read-only collection of objects partitioned across a cluster a machines [12]. The main advantage of such a mechanism is that a RDD can be explicitly cached in memory across machines and re-used in multiple iterative parallel operations [12]. RDD's are expressed as Scala objects and are fault-tolerant.

When a 39 GB Wikipedia data dump was queried interactively, the first query using spark took approximately 35 s which comparable to running a Hadoop job on MapReduce, every subsequent query took around 0.5–1 s which is significantly low [12].

When the performance of a logistic regression job was compared between Hadoop and Spark, each Hadoop iteration took approximately 127 s as each run as an independent MapReduce Job [12]. When performed in spark the first iteration took 174 s, but each subsequent iteration took only around 6 s, which is a significant improvement [12].

Components of spark architecture are the driver program, cluster manager and worker nodes as shown in Fig. 3. Driver programs acts like a central point and runs on the master node. Resource allocation is carried out by the cluster manager. The worker nodes are responsible for all the data processing tasks [15].

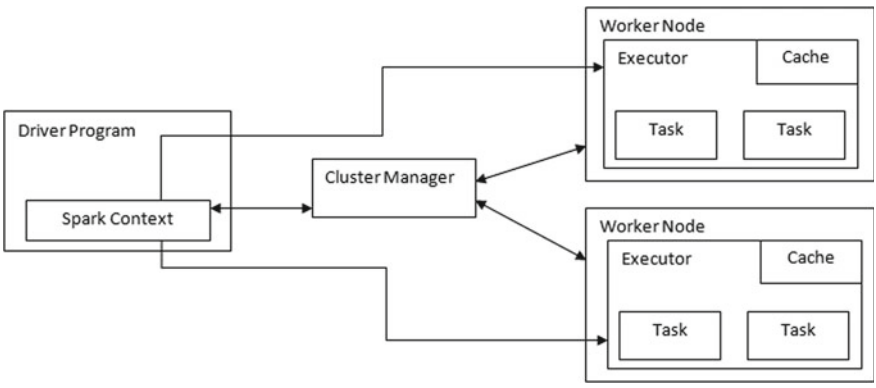


Fig. 3 Spark architecture [15]

Other Components That Run on Top of Spark are [13]

(i) **Spark MLlib**
It is a library bundle that has distributed implementations of machine learning algorithm for classification and clustering like the K-means and RandomForest.

(ii) **Spark GraphX**
It is a bundle of distributed algorithms for graph processing.

(iii) **Spark Streaming**
An approach for environment that require real-time or near real-time data processing.

3 Applications

3.1 Healthcare

In the past few years, there have been many developments in the healthcare sector. There has been a move towards digitisation of several aspects of medical data which has opened the doors for the next revolution of information management in healthcare. The data is generated under various organisations within the healthcare ecosystem like hospitals, insurance companies, governments and research institutes [16]. Mckinskey Global Institute made a suggestion that if US healthcare sector were to employ effective Big Data Solutions, it could create more than USD 300 billion in value most of which would be attributable to reduced US healthcare expenditure [16].

Traditional medical systems use various continual monitoring instruments to monitor the vitals of a patient and trigger an alarm in the event of a particular vital or set of vitals cross discrete numerical cut-off values [16]. Many such alarm mechanisms generate a considerable number of false alarms, as they depend on the isolated sources of information which usually lack context of a patient's true condition from a broader point of view. This is one of the areas where Big data can come to the rescue. Medical data can be collected through many heterogeneous sources to create a comprehensive understanding of a patient. Some of these sources are as follows [17]:

- Biometric data which includes X-rays, CT-scan, MRI, Sonography among other medical imaging techniques, fingerprints, retinal scans, various signals like Blood Pressure and ECG, DNA and genetic data etc.
- Data generated by Medical staff like Electronic Medical Records (EMR), Doctor's notes and other paper documents
- Wearable fitness devices that monitor overall fitness levels by tracking things like steps taken, sleep time, heart activity and calories spent [18]

- Data captured from various sources on the internet like social media, medical websites, blogs amongst others
- Data from Pharmacies.

They are a plenty of opportunities for Big Data technologies to be used in the healthcare space.

By analyzing historical medical patterns of patients across various diseases and various different approaches used by doctors for treatment, it is possible to take informed decisions from a broader point of view rather than an isolated one [16]. Public health-care budgets can be reduced by identifying where the resources are best needed for preventive healthcare. This can be achieved by identifying which strata of society is affected the most by certain type of diseases and develop actionable insights on where prevention and education is needed the most to produce robust populations. Remote and virtual healthcare can be promoted using Big Data hence impacting lives of millions of people without access to quality healthcare [17]. Using clustering techniques, identification and segmenting of patients with similar symptoms, indications and care patterns over longer time periods can help form patient clusters resulting in patient centric medical approaches [17].

The type of data generated in the healthcare sector mostly consists of unstructured data like medical images, graphs, doctor's notes etc. as mentioned earlier. Such variety of data types causes problem during data aggregation and pre-processing becomes challenging sometimes [17]. The existing laws and regulatory policies designed to protect and ensure patient privacy also poses a significant challenge to Big Data Analytics in healthcare. The access to such data is very limited and subject to authorization [17]. Also, Doctors' and Medical centres are unwilling to share their approaches and best practices to maintain an advantage over their competition. These are some of the challenges for Big Data in Healthcare [17].

An example use case of big data in healthcare is the collaboration between Moorfields Eye Hospital and DeepMind Health in the UK. Moorfields will be sharing 1 million eye scans with DeepMind, to be analyzed using sophisticated machine learning algorithms to predict with a stated accuracy whether a condition is worthy of immediate attention [19].

In the near future, Big Data Analytics will be widespread in the Healthcare industry and will help save millions of human lives.

3.2 Public Sector

Adopting Big Data Techniques can provide tremendous benefits to the public sector. Right from preventive healthcare, fraud detection, urban management, education to inclusive policy making and handling crime, there are plenty of opportunities for big data in public sector. Many public sector problems can be address by big data solutions such as boosting transparency and increasing productivity and efficiency. Government and its affiliate agencies are one of the biggest producers of data, a lot

of which remains unutilized or underutilized. Governments around the world are providing access to "Open Data" with a view of promoting economic growth and transparency [4]. "Open Data" can be defined as all the information relating to public sector in a free environment [4]. Services based on open data are attempting to solve real-world problem in an innovative manner. The scale and speed of creation of Open Government Data puts it within the ambit of big data.

Another interesting prospect is public decision making using big data although it has some serious limitations [20]. A lot of big data literature in public sector focuses on leveraging big data technologies to promote active citizen engagement in governance and policy making for making it an inclusive process. Sentiment Analysis using social media platforms is way to assess the response of citizen on various policy decisions. An example is the "Pulse of the nation" project carried out researchers from Harvard and Northeastern which makes use of twitter data to analyze mood for the day [20]. Social media increasingly being seen as a real-time public communication platform. Sentiment analysis using big data technologies on social media platforms like twitter can help gauge the perception of the public regarding policy and programs undertaking by the government. This can help governments prioritize effectively responding to the things that affect the citizens most.

An example of big data application in public sector is the collaboration between University of California, Los Angeles (UCLA) and the Los Angeles Police Department (LAPD) setting up the Real-Time Analysis and Critical Response Division [20]. They use historical and real-time data to "predict" probable future incidents of crime allowing the LAPD to prioritize their response. The data used is of disparate source like life video feed of traffic and inputs from local law enforcement officers. The data is high volume, variety and complexity giving it big data colours.

3.2.1 Challenges and Issues

There are many issues and challenges that hinder the use Big Data to its full potential. These challenges lead to hurdles in the implementation of Big Data technologies. We will try to discuss a few of these challenges and issues.

(1) **Security and Privacy**
It is one of the important challenges that have technical and legal ramifications. It is imperative that a massive amount of data generated nowadays does not fall at risk. Business Organisations and Social Media Companies are collecting information regarding people and creating insights by without their consent or awareness. Cambridge Analytica scandal is a prime example of how personal data on social media can be misused to profile voters to gain electoral benefits. Cambridge Analytica retained and misused private data of 50 million Facebook users [21]. There is a pressing need to address policies regarding privacy and security. Maintaining confidentiality of data and assuring that there is no unauthorized use is also a major concern and is a daunting task in the big data context. Organizations should make sure that data is always protected. They could do so

by having high grade data encryption, maintaining dedicated database servers, having multiple security levels, secure system operations, separate authentication and authorization modules and data flow control [3]. In order to realise the full potential of big data, policies relating to intellectual property, privacy, liability and security will have to be addressed [22].

(2) **Complexity**
The intrinsic complexity of Big Data makes it challenging to understand, represent and compute and perceive the problem. Complexity includes complex data types, complex data structures, computational and systemic complexity. There is a lack of proper understanding to address the problem of complexity of big data. To understand the underlying characteristics and formation of essential patterns, there is a need for a study on the topic which will further shed light on a more lucid knowledge abstraction, ease its representation and help in the further development of models and algorithms for big data [23]. The entire lifecycle of Big Data applications must be considered to address the problem of complexity.

(3) **Inter-connectedness**
Many data sources must be used in tandem to make effective use of big data. One of the challenges faced by organisations is their inability to connect multiple data points within the organisation and outside. In order to fully realise the potential Big Data, organisations need to manage data from various transaction points across its several enterprises [22]. The ability to connect unrelated datasets is of utmost importance. This also sometimes leads to privacy concerns as connected datasets often generate insights for which the individual did not consent. One particular example is the prediction of spread of H1N1 virus by Centre for Disease Control and Google by connecting search items like "flu symptoms" and "cold medicine " from Google's database to CDC's data on H1N1 virus [20].

(4) **Practical challenges**
There is a gap of Big data professionals with appropriate skill-sets require to handle big data problems that are complex in nature [3]. Most of the big data experts are usually adept at implementing big data technologies but are inept in dealing with big data management. Also, there are a plethora of tools and techniques available for analytical processing of Big Data [3]. With so many disparate options available, it becomes challenging for big data practioners to choose a consistent strategy. With humongous volumes of data available, it becomes extremely difficult to eliminate irrelevant data and ensuring that the quality of data is maintained and that too at a desired speed [3]. There is also a cost consideration while deploying a big data system. While it is much less than an enterprise data warehouse solution in terms of storage cost (a typical Hadoop cluster), it is uncertain as to what value holding such huge amount of data holds with respect to the cost of storing it [3].

4 Conclusion

In this information era, big data holds a lot of promise and opportunities while posing a set of challenges that include dealing with the scale, speed and complexity of big data along with privacy and security issues. In this chapter, we have discussed technologies like Hadoop and Spark in the big data paradigm. This study discusses some of the applications of big data technologies in Healthcare and Public Sector in brief and mentioned some of the opportunities and challenges. Although there is a long way to go for Big data technologies to be fully utilized in healthcare due to restrictions of data access owing to various regulatory policies and competitive urges, the potential impact of it is so huge, in the fields of comprehensive and preventive healthcare, that it is hard to neglect. Governments and healthcare organizations must come together and try to find the middle ground while addressing the issues plaguing the use of Big Data in healthcare. Going forward, Big Data life-cycle management skills will play a big role and will be decisive in effective implementations of Big Data solutions. Technical issues relating to scale, computational complexity, data variety and data security will also have to be addressed going forward in order to extract full potential of Big Data across domains.

References

1. Al-Sai, Z.A., Abualigah, L.M.: Big data and E-government: a review. In: 2017 8th International Conference on Information Technology (ICIT), pp. 580–587. IEEE (2017)
2. Bhosale, H.S., Gadekar, D.P.: A review paper on Big Data and Hadoop. Int. J. Sci. Res. Publ. **4**(10), 1–7 (2014)
3. Mukherjee, S., Shaw, R.: Big data-concepts, applications, challenges and future scope. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **5**(2), 66–74 (2016)
4. Fredriksson, C., Mubarak, F., Tuohimaa, M., Zhan, M.: Big data in the public sector: a systematic literature review. Scand. J. Public Adm. **21**(3), 39–62 (2017)
5. https://analyticsweek.com/content/big-data-facts/. Last accessed on 19 Jan 2019
6. https://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/. Last accessed on 19 Jan 2019
7. http://fortune.com/2018/06/07/intel-ceo-brian-krzanich-data/. Last accessed on 19 Jan 2019
8. Ghemawat, S., Gobioff, H., Leung, S.T.: The Google File System, vol. 37, no. 5, pp. 29–43. ACM (2003)
9. http://hadoop.apache.org/. Last accessed on 19 Jan 2019
10. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
11. Al-Barznji, K., Atanassov, A.: Collaborative filtering techniques for generating recommendations on big data (2017)
12. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I.: Spark: cluster computing with working sets. HotCloud **10**(10–10), 95 (2010)
13. https://www.infoworld.com/article/3236869/analytics/what-is-apache-spark-the-big-data-analytics-platform-explained.html. Last accessed on 19 Jan 2019
14. https://www.ibm.com/developerworks/library/os-spark/. Last accessed on 19 Jan 2019
15. Lakshmi, C., Nagendra Kumar, V.V.: Survey paper on big data. Int. J. Adv. Res. Comput. Sci. Softw. Eng. (2016)

16. Belle, A., Thiagarajan, R., Soroushmehr, S.M., Navidi, F., Beard, D.A., Najarian, K.: Big data analytics in healthcare. BioMed. Res. Int. (2015)
17. Patel, Sanskruti, Patel, Atul: A big data revolution in health care sector: opportunities, challenges and technological advancements. Int. J. Inf. Sci. Tech. **6**, 155–162 (2016). https://doi.org/10.5121/ijist.2016.6216
18. https://www.dummies.com/health/exercise/what-is-wearable-fitness-technology-and-how-can-it-help-you/. Last accessed on 19 Jan 2019
19. https://www.gov.uk/government/speeches/big-data-in-government-the-challenges-and-opportunities. Last accessed on 19 Jan 2019
20. Desouza, K.C., Jacob, B.: Big data in the public sector: lessons for practitioners and scholars. Adm. Soc. **49**(7), 1043–1064 (2017)
21. https://arstechnica.com/tech-policy/2018/03/facebooks-cambridge-analytica-scandal-explained/. Last accessed on 19 Jan 2019
22. Toshniwal, R., Dastidar, K.G., Nath, A.: Big data security issues and challenges. Int. J. Innov. Res. Adv. Eng. (IJIRAE) **2**(2) (2015)
23. Jin, X., Wah, B., Cheng, X., Wang, Y.: Significance and challenges of big data research. Big Data Res. **2** (2015). https://doi.org/10.1016/j.bdr.2015.01.006

Big Data in Healthcare: Technical Challenges and Opportunities

Ganesh M. Kakandikar and Vilas M. Nandedkar

Abstract Big data refers to the data generated in large volumes, with complexities, involving many parameters and their relevance. Every industry is going through transformation so as healthcare also. New diseases, new viruses are still challenges, whereas we successfully got escapes from many. The new ways of treatments based on research, new drugs, advancements in procedures, use of micro tools etc. has changed the healthcare sector to lot. Identification of diseases is more based on pathological tests than clinical practices. All these generates lot of data in healthcare sector also. So, in any sense health sector can't do away with big data. The chapter discusses of big data, scope in healthcare, challenges of implementation, characteristics of data etc. in healthcare sector.

1 Introduction

Due to research and development in diseases, pharmaceuticals, new ways of treatments, advancements in diagnostic and surgery equipment's healthcare industry has emerged as one biggest and potential industry over recent years [1, 2]. It has seen transformation from small specialized hospitals run by individual experts to multi-speciality hospitals having corporate management. Healthcare management around the world concentrates on patient-centered model rather than disease-centered, it also has approach of value-based healthcare delivery model instead volume-based. The principle behind this is to provide superior services with reduced costs [3]. The gap between healthcare costs and out-comes was analysed to be the result of poor management of insights from research, poor usage of available evidence, and poor capture of care experience, all of which led to missed opportunities, wasted resources,

G. M. Kakandikar (✉)
School of Mechanical Engineering, Dr. V.D. Karad MIT World Peace University, Pune, India
e-mail: Kakandikar@gmail.com

V. M. Nandedkar
SGGS Institute of Engineering and Technology, Nanded, India
e-mail: vilas.nandedkar@gmail.com

A. J. Kulkarni et al. (eds.), *Big Data Analytics in Healthcare*,
Studies in Big Data 66, https://doi.org/10.1007/978-3-030-31672-3_5

and potential harm to patients. This healthcare industry is composed of hospitals, druggist, pharmacist, pathologist, radiologist [4].

2 Big Data in Healthcare

When volume of data is very high, large number of transactions as well as data sources are many and complex, needing special methods and technologies in order to analyse, draw insight out of data for instance, it is popularly known as big data. In other words, big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. Traditional data warehouse solutions may fall short when dealing with big data.

The massive demand of big data in healthcare organizations is resulting into high volume of data day by day [5]. In order to comply with highest quality healthcare services for patient-centered model, management and analysis of huge health data is the need of hour. Variety and volume of big data doesn't suit with old data management tools and technologies. The complexity of big data is represented in Fig. 1.

The big data can be applied for prediction of the diseases, prior their emergence referring past medical records [6]. It is imperative now in many countries to keep electronic patient's records with advanced medical imaging media. The big data has potential to cater the upcoming healthcare market as well as new trends in market. Epidemiologists, physicians, and health policy experts can have data-driven judgments in the interest of patients using big data. Most of the countries have made it mandatory to digitize the patient's records. Every new patient must be registered in electronic registration system [7]. Patient's to be issued with chip based data cards.

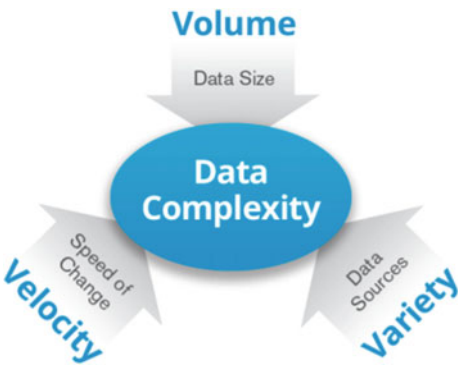


Fig. 1 Complexity of data

3 Types of Big Data

The classification of data can be done in many classes based on behaviour.

3.1 Structured Data

Relational databases with rows and columns defines structured data. The structure of these tables is defined by organizations suitable to their needs in terms of models. The model permits input, processing and output of data with certain authorities to individuals in organization. Some of them can only see the data, they don't have editing authority. Suitability of data to models is expressed in terms of characteristics of data. Storage, handling and analysis of structured data is quite easy. Structured Query Language (SQL) is applied for handling this type of data.

3.2 Unstructured Data

This type of data is completely unstructured, so cannot be stored in rows and columns. This is completely contradictory with structured one. It is growing extremely fast, so it becomes tough to manage and analyze it completely. More sophisticated tools are required for analysing this data.

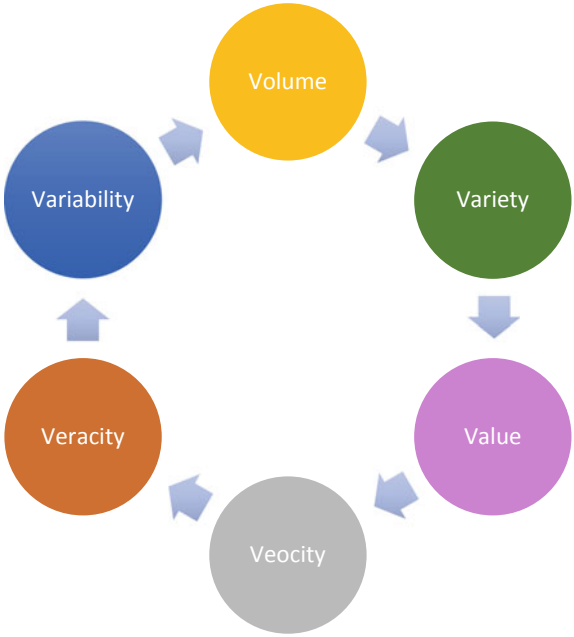
3.3 Semi-structured Data

Data which is in the form of structured data but it does not fit the data model is semi-structured data. It can be stored with some tags or markers as it is not possible to fit in data table. The data is stored in these markers with certain rankings.

4 Big Data Characteristics

Volume, Value, Velocity, Variety, Veracity and Variability are the major six characteristics of big data denoted as 6 "Vs". However, few authors and researchers also define certain additional characteristics. The importance of all six V's for big data are represented in Fig. 2.

Fig. 2 Six V's of big data



4.1 Volume

The amount of data in healthcare is going to be increased rapidly in years to come, which could be in terabytes (10^{12} bytes), petabytes (10^{15} bytes) or Exabyte's (10^{18} bytes). Basically, volume refers to the amount of data generated. Such huge amount of data creates storage as well as massive analysis issue [8].

4.2 Variety

Contrasting bases of data or different formats in which it is collected creates variety. Naturally the data in healthcare would be of various patients, departments, diseases, procedures, doctors, surgeons with inbuilt variety. Complexity and heterogeneity of multiple datasets, refer to the variety. This data could be of organized, semi organized or of totally unorganized type. Laboratory data, clinical, sensor data and data from relational data is organized one. Semi-organized data includes data that is stored in Extensible Markup Language (XML) format. Unorganized data consists of majority of data such as free text data not having any specific formats or designs e.g. manual notes. Images including from X-Ray, Radiology outputs, Electronic Medical Record, patients discharge cards, data in terms of various signals, prescriptions etc.

4.3 Velocity

Velocity refers to the speed with which data is created and supplied as well as managed [9]. Velocity thus concerns with amount of data needed to meet demand. The data can be generated in batches at certain interval on time line or it can be live data changing continuously which can be from either one source or multiple sources. In certain cases, the data generated of patients as well as demand of data for control of epidemic situations is very high.

4.4 Veracity

When data comes from multiple sources, known or unknown, how to validate the data? Veracity refers to the correctness and accuracy of information, data quality, relevance, uncertainty, reliability and predictive value. Any data cannot be 100% accurate as there would be always some unknown sources. Big data has low veracity, and it is difficult to validate. It is needed to apply certain standards to ensure the features of the data.

4.5 Variability

Data vary continuously throughout the life cycle; this change is defined as variability [10]. If data can be defined with range and variability it increases the possibility of providing more accurate data. Sometimes it could be unforeseen or hidden. Consistency of the data over time of use also matters.

4.6 Value

The inference from the data refers to the value of data. Value is irrespective of amount of data. The value of data in healthcare can be referred to five major aspects as below.

4.6.1 Right Living

Patients must be educated to take care of their own health by selecting right diet, exercise, preventive care, and other lifestyle factors.

4.6.2 Right Care

It is very right of every patient to receive appropriate, timely treatment [11].

Coordinated efforts of all health service providers with access to centralized data to provide best services to patients is needed instead of too many protocols.

4.6.3 Right Provider

All Professionals in healthcare must have strong experience to treat patients with suitable skills for getting best results.

4.6.4 Right Value

The healthcare quality provided to patients must have value.

4.6.5 Right Innovation

Innovations will lead to new therapies and approaches to health-care delivery [12].

5 Analyzing Big Data

It is necessary to analyze the data once collected to get the insight. Basically, four types analyses are applied for all types of data.

5.1 Predictive Analysis

Predictive analysis by analysing previous data patterns offers solutions for given situation. It correlates present and past data for future insights, with probabilities of outcome. This analytical method is most commonly applied in patient's relationship management data [13].

5.2 Prescriptive Analysis

Prescriptive analysis reveals necessary actions and future directions to be taken. It proposes answers to the situations in a focused way. It is one step ahead of predictive analysis by providing multiple ways to achieve likely outcomes.

5.3 Diagnostic Analysis

It is applied for analysing the past events. Diagnostic analysis provides insight to happened events as how, what and why happened. It is aimed at knowing hidden patterns of previous data to locate root causes, as well as factors affecting the root causes. This would provide opportunity to understand and correlate current data patterns with past one.

5.4 Descriptive Analysis

The simplest way to converts big data into small bytes is descriptive analysis. It is also referred as data mining. Most of the establishments use this for analysis of current data [14].

6 Challenges in Big Data Healthcare

Healthcare sector is struggling to use and apply data to fullest extent due to lot of practical challenges. They are discussed below

- The data in most of the hospitals/health providers is segmented one. To elaborate, administrative data such as claims and settlements is used by the financial and operational management teams, for business prospective. This is never applied for patients care and for treatment protocols. Clinical data such as patient history, vital signs, progress notes, and the results of diagnostic tests are stored in the Electronic Records, which is used by the physicians, nurses, and other frontline clinical staff.
- Second major challenge is to protect patient's privacy. All medical data are very sensitive and different countries consider these data as legally possessed by the patients. Sharing of data between various organizations is required to provide better healthcare. Regional health information organizations act as coordinating agency between stake holders. Patient data may be shared after de-identification, but protecting the patient from either direct or indirect identification while still maintaining the usefulness of the data is challenging. Sometimes organizations are not ready to share this data due to market competition [15].
- Patients themselves are prominent source of information these days. Collection of this data and its inclusion is very critical for further analysis. This is data collected through various monitoring systems. Real challenge is validation of this data. The risk of compromised data integrity is much higher with this patient collected data than other sources.
- Most significant challenge on operating part would be aggregation and analysis of huge amount of unstructured data. Mostly unstructured data consists of test results, scan images, progress notes or prescriptions.

- Another important challenge that must be acknowledged in data analytics is secondary use of data.
- Collection of large amount of data is challenging in itself. Obtaining high throughput is based on experimental measurements. Heterogeneity of data sources, noise in experimentations, selection of appropriate experimentation technique can affect the quality of data [16].
- The patient collected data may be high dimensional, means there could be more dimensions or features than sample size.
- The cost of establishment of Big Data Analytics architecture is very high.
- Big Data systems require data scientists with specialized experience to support design, implementation, and continued use.
- Knowledge discovery and representation is a prime issue in big data. It includes number of sub fields such as authentication, archiving, management, preservation, information retrieval, and representation [17].

7 Technical Challenges

7.1 Data Integration

In today's competitive environment, data is key asset for organizations. It is important to understand how the data can be used and applied in taking better decisions with its expression in standard formats. Organisations might also need to decide if textual data is to be handled in its native language or translated.

7.2 Complexity

Data can have multiple character sets and alphabets, which makes it complex. The transfer of external data to organization is also complex process, though it is carried out streamlined or in batches. The system should support this transfer, adequately scalable [18].

7.3 Data Transformation

Data transformation is one of the major challenge, where rules are defined for handling data from one system to another, to avoid any loss of data and making the process seamless.

7.4 Complex Event Processing

It is the process of analysing data coming from different streams to devise new inferences/events/patterns about data which can result into more complicated situations. This facilitates meaningful events and provide opportunity to respond them at the earliest. It has emerged from discrete event system simulation. Complex event processing is also known as operational excellence.

7.5 Semantic Analysis

Handling unstructured data is challenging and complicated task to make better sense out of it. Semantic analysis is the way to handle unstructured data and extract relevant and useful information. Semantic technology has always been about the meaning of data, its context, and the relationships between pieces of information.

7.6 Historical Analysis

Many times, it is required to trace past data in business processes. Like when repeating the orders already executed, we need to use same data. Historical analysis help companies to reapply the data [19].

7.7 Data Storage

Storing and handling data when in large volumes is the real challenge of today. Big data needs storage which can be easily accessed. Older technologies like magnetic disks will slowly became irrelevant in this situation.

7.8 Integrity of Data

Data Integrity refers to accurateness and completeness of data. As all the critical management and business decision are based on data, it should be updated one as far as possible. Decisions on erroneous data produces wrong results and insights. Confidence in integrity of data is very much important.

7.9 Data Lifecycle Management

For life cycle management of data, companies must know the data in detail along with its size, characteristics and purpose. There are lot of hurdles, there is lot of data and it is being continuously changed, it needs different and new approaches for data management. It is not possible always to include all the data, so we have to work with samples of data from packets. But it is to be ensured that the samples are proper representatives and provides required insight [20].

7.10 Data Replication

To avoid any loss of data due to physical damages to media data is stored at multiple locations, so as to ensure availability, known as data replication. In case of high volume of data replication is challenging task. However, Big Data technologies may take alternative approaches to handle this.

7.11 Data Migration

There is always an impact of size of data when it is to be entered to Big Data systems or when changing platforms. During this migration, data is not available for use.

7.12 Visualisation

It is very much important to present the data in visual meaningful form, without affecting the effectiveness of the system. Big Data Analytics must produce data in most appropriate way, so as to reduce any misleading. The representation should also take in into account the IT systems on which it will be displayed and their capacities.

7.13 Data Access and Security

This is major challenge to permit access to data to various levels of employees, with different authorities. This is very important with respect to data protection. It is necessary to limit the access to data as well as existence of data. Equally it is important not to lock data unnecessarily [21].

8 What Is Special About Medical Big Data?

Healthcare data is very complex due to diversity of diseases and health related ailments. The treatments also differ with respect to specialists, their knowledge and experience, so do outcomes. The intricacies of data collection, processing and interpretation also matters. The data is collected from variety of sources as administrative claims, prescriptions, treatment methodologies, surgery records, electronics depositories, medical imaging, patient reported data, as well as clinical trials [22]. Integrating all this data at central server is challenging task as it is huge and vibrant. It also exists on multiple scales, sometimes incomplete, insufficient and complex. There are several characteristics of medical big data compared to traditional clinical epidemiological data.

- Medical data is hard to access because most of the practitioners are reluctant to share due to risk of misuse or privacy issues.
- Medical data is often collected with protocols with fixed formats so as to simplify.
- Collection of medical data is costly affair because involvement of personnel, expensive instrumentation and at the discomfort of patients.
- Patient characteristics and treatment direction is core of healthcare data [23].

9 What Is Medical Big Data for?

Big data has in built importance that, it finds associations without bothering for meanings. It only indicates a signal for collected data. Big data concentrates on finding correlations and patterns rather than on casual inference amid complex data [24].

The big data has potential applications in

- Providing personal healthcare.
- Analysis of medical images for clinical decision support systems.
- Tailored diagnostic and treatment decisions.
- Study of behavioural patterns of patients.
- Big data driven population health analyses.
- Fraud detection and prevention.
- Prediction and continuous monitoring of diseases of individual health.
- Predictive modelling for risk and resource use.
- Drug and medical device safety surveillance.
- Quality of care and performance measurement.
- Population management.
- Public Health
- Research applications.

The major strength of big data is predictive analysis. It is the technology, which learns from itself i.e. data to predict the future behaviour of patients for taking better

decisions means future insights based on associations. Big data are necessary but not sufficient, and simply accumulating a large dataset is of no value if the data cannot be analysed for future insights that we can act upon [25].

10 Opportunities of Big Data in Healthcare

This section discusses the various opportunities of Big data in healthcare.

10.1 Healthcare Facilities at Low Costs

Big data can offer decrease in healthcare cost of medical treatments in many ways. The data analysis can provide data on populations which are under illness risks. This will help to take proactive steps. The data will find out need of education and prevention for mass of disease, at lower costs [26].

10.2 Promotes Research and Innovation

Data analytics will help to get insight of the current status of patients suffering from various diseases. It also reflects the status of healthcare facilities being offered. This data helps in coordinating the research activities and also promotes innovation for better healthcare practices by taking more ownership.

10.3 Personalized Treatments

It is now possible to predict individual life style diseases. Data analytics will provide more insight to symptoms and history of patient's records, which will facilitate personalize medicines and tests for necessary treatments for each patient. Early detection will lead to better care and eliminating risks for chronic diseases, as individual patients detailed data will be available [27].

10.4 Enhanced Preventive Care

As it is said Prevention is always better than cure. Availability of data will help to capture, analyze and treat patient's symptoms in earlier phases, preventing the occurrence of diseases.

10.5 Virtual Care Technologies

Technology is helping health providers to start virtual care initiatives, providing more specialized treatments and more quality services, irrespective of non-availability of local experts.

10.6 Health Trend Analysis

Data mining, text mining and many more analytical tools can be applied to health trend analysis. Big data management facilitates comprehensive patients management.

10.7 Comprehensive Management of Patients

It is easy to identify and locate patients having similar clinical indications and care patterns. Patients can be divided into groups based on the primary diagnosis. Then statistical analysis will help to divide them in further sub groups. Data will help to track the patients and determine the patterns for treatment.

10.8 Drug Efficacy Study

Data in Electronic record may also be used to study drug efficacy [28].

11 Platforms—Big Data Analytics in Healthcare

The Hadoop Distributed File System (HDFS) is one of the major platform. It enables the underlying storage for the Hadoop cluster. The data is divided into smaller parts and distributed across various servers. MapReduce is best for the distribution of sub-tasks to generate outputs. During tasks execution, MapReduce tracks the processing of each server/node [29]. Pig and PigLatin programming language can handle all types of data structured as well as unstructured. Hive is support architecture which interfaces structure query language (SQL) with the Hadoop platform. Jaql is query language designed to process large data sets. It also facilitates parallel processing. Zookeeper allows integration of various services, providing synchronization across cluster of servers. HBase is a column-oriented database management system. It uses a non-SQL approach. Cassandra, a distributed database system handles big data distributed across many utility servers. Oozie, being an open source project, streamlines

the workflow and coordination among the tasks. The Lucene project is used widely for text analytics/searches. Avro facilitates data serialization services. Versioning and version control are additional useful features [30].

12 Conclusion

It is obvious from the discussion that big data is going to play an important role in healthcare sector. There is need to organize the data in healthcare by changing the ways, the data is generated, collected, stored and used. This will facilitate ease and better services to patients at quite lower costs. This will also help for prevention of diseases than fighting with them to cure. Organised data will help to deliver customized service to patients even at greater distances through internet. The application of big data to healthcare industry will also generate lot of new jobs.

References

1. McAfee, Bryn Jolfsson, E., Davenport, T.H., Patil, D.J., Barton, D.: Big data: the management revolution. Harvard Bus. Rev. **90**(10), 60–68 (2012)
2. Lynch, C.: Big data: how do your data grow? Nature **455**(7209), 28–29 (2008)
3. Jacobs, A.: The pathologies of big data. Commun. ACM **52**(8), 36–44 (2009)
4. Manyika, J., Chui, M., Brown, B., et al.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute (2011)
5. Celi, L.A., Mark, R.G., Stone, D.J., Montgomery, R.A.: Big data in the intensive care unit: closing the data loop. Am. J. Respir. Crit. Care Med. **187**(11), 1157–1160 (2013)
6. Seibert, J.A.: Modalities and data acquisition. Practical Imaging Informatics, pp. 49–66. Springer, New York (2010)
7. Oyelade, J., Soyemi, J., Isewon, I., Obembe, O.: Bioinformatics, healthcare informatics and analytics: an imperative for improved healthcare system. Int. J. Appl. Inf. Syst. **8**(5), 1–6 (2015)
8. Ristevski, B., Chen, M.: Big data analytics in medicine and healthcare. J. Integr. Bioinform. 1–5 (2018)
9. Viceconti, M., Hunter, P., Hose, R.: Big data, big knowledge: big data for personalized healthcare. IEEE J. Biomed. Health Inform. **19**(12), 9–15 (2015)
10. Gu, D., Li, J., Li, X., Liang, C.: Visualizing the knowledge structure and evolution of big data research in healthcare informatics. Int. J. Med. Inform. **11**, 22–32 (2017)
11. Kannampallil, T.G., Franklin, A., Cohen, T., Buchman, T.G.: Sub-optimal patterns of information use: a rational analysis of information seeking behaviour in critical care. Cognitive Informatics in Health and Biomedicine, pp. 389–408. Springer, London (2014)
12. Srinivasan, U., Arunasalam, B.: Leveraging big data analytics to reduce healthcare costs. IT Prof. **15**(6), 21–28 (2013)
13. Kuo, M.H., Sahama, T., Kushniruk, A.W., Borycki, E.M., Grunwell, D.K.: Health big data analytics: current perspectives, challenges and potential solutions. Int. J. Big Data Intell. **1**(1), 114–126
14. Sandryhaila, A., Moura, J.M.F.: Big data analysis with signal processing on graphs: representation and processing of massive data sets with irregular structure. IEEE Signal Process. Mag. **31**(5), 80–90 (2014)

15. Bains, J.K.: Big data analytics in healthcare-its benefits, phases and challenges. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **6**(4), 430–434.5567 (2016)
16. Agrawal, D., et al.: Challenges and opportunities with big data. In: Big Data White Paper-Computing Research Association, February 2012. https://cra.org/ccc/docs/init/bigdatawhitepaper.pdf
17. Fatt, Q.K., Ramadas, A.: The usefulness and challenges of big data in healthcare. J. Healthc. Commun. **3**(2:21), 1–4 (2018)
18. Acharjya, D.P., Kauser, A.P.: A survey on big data analytics: challenges, open research issues and tools. Int. J. Adv. Comput. Sci. Appl. **7**(2), 511–518 (2016)
19. Kakhani, M.K., Kakhani, S., Biradar, S.R.: Research issues in big data analytics. Int. J. Appl. Innov. Eng. Manage. **2**(8), 228–232 (2015)
20. Philip, C.L., Chen, Q., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**, 314–347 (2014)
21. Kuo, M.H., Sahama, T., Kushniruk, A.W., Borycki, E.M., Grunwell, D.K.: Health big data analytics: current perspectives, challenges and potential solutions. Int. J. Big Data Intell. **1**, 114–126 (2014)
22. Balamurugan, S., Madhukanth, R., Prabhakaran, V.M., Shanker, R.G.K.: Internet of health: applying IoT and Big data to manage healthcare systems. Int. Res. J. Eng. Technol. **3**(10), 732–735 (2016)
23. Sonnati, R.: Improving healthcare using big data analytics. Int. J. Sci. Technol. Res. **6**(3), 142–146 (2017)
24. Das, N., Das, L., Rautaray, S.S., Pandey, M.: Big data analytics for medical applications. Int. J. Modern Educ. Comput. Sci. **2**, 35–42 (2018). Published Online February 2018
25. Chen, C.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inform. Sci. **275**, 314–347 (2014)
26. Althaf Rahaman, S., Sai Rajesh, K., Girija Rani, K.: Challenging tools on research issues in big data analytics. Int. J. Eng. Dev. Res. **6**(1), 637–644 (2018)
27. Lynch, C.: Big data: how do your data grow? Nature **455**, 28–29 (2008)
28. Jin, X., Wah, B.W., Cheng, X., Wang, Y.: Significance and challenges or big data research. Big Data Res. **2**(2), 59–64 (2015)
29. Chang, B.R., Lee, Y.-D., Liao, P.-H.: Development of multiple big data analytics platforms with rapid response. Sci. Program. **2017**, 1–13
30. Singh, D., Reddy, C.K.: A survey on platforms for big data analytics. J. Big Data **2**(8) (2015)

Innovative mHealth Solution for Reliable Patient Data Empowering Rural Healthcare in Developing Countries

Jay Rajasekera, Aditi Vivek Mishal and Yoshie Mori

Abstract Rural healthcare is a global issue. However, collection of health related data in a "timely and reliable" manner—as highlighted by World Health Organization in its 2018 "Monitoring Health for the Sustainable Development Goals" report remains a big challenge. This chapter reviews the general problems associated with collection of health data from rural areas where large percentages of populations of developing countries live. Two cases: one involving an innovative mHealth mobile tablet App (N+Care) designed to be used in rural areas in developing countries under a Japanese government funded research project and another a private initiative (A3: Anywhere Anytime Access) in India to provide mHealth services for providing remote medical care for remote populations. Both cases are intended to improve the credibility of data collection from rural areas in developing countries.

Keywords mHealth · Remote patient monitoring · Rural healthcare · Developing countries

1 Introduction

mHealth, an abbreviation of mobile health, is the practice of providing medical support, especially from a medical professional to a patient at a faraway location, via mobile devices and concerned multimedia technologies [1–3]. The roots of mHealth can be traced back to the age of modern communication and the invention of two-way

Jay Rajasekera and Yoshie Mori acknowledge the funding from Grant-in-Aid for Scientific Research by Japanese Government that contributed to findings in this chapter.

J. Rajasekera
Graduate School of Business and Commerce, Tokyo International University, Kawagoe, Japan
e-mail: jrr@tiu.ac.jp

A. V. Mishal (✉)
Symbiosis Institute of Operations Management, Symbiosis International University, Pune, India
e-mail: aditi.mishal@siom.in

Y. Mori
Graduate School of Health Sciences, Gunma University, Maebashi, Japan

A. J. Kulkarni et al. (eds.), *Big Data Analytics in Healthcare*,
Studies in Big Data 66, https://doi.org/10.1007/978-3-030-31672-3_6

radios [4]. The spread of telephones also helped providing medical care remotely. A term known as telemedicine had been used to refer to providing such medical help via telephones, especially to remote communities, which lacked adequate medical facilities and specialized doctors. Among developing countries, one of the early adopters of telemedicine was India, way back in 1997, when it launched "Development of Telemedicine Technology" as a national project [5].

1.1 mHealth-Revolution

With the advancement of mobile phones, especially the so-called smartphones, mHealth has taken new dimensions around the world—in developed, as well as developing countries. The modern mHealth applications rely on mobile telephone networks, WiFi networks, as well as Bluetooth that are used in short-range communications. These networks of course rely on personal computers, notebooks, mobile phones, tablets, digital cameras, sensors, etc. to connect various parties, including the health personnel and patients. Regardless, mHealth can also be viewed as an extension of existing health systems that had found applications, particularly in the developing countries as a result of limited resources to serve populations living in remote areas away from urban centers where most of the health facilities are located [6–8].

1.2 Big Data Healthcare in Developing Countries

A World Bank study by Qiang et al. [9] focusing on developing countries, identifies the operation modes of mHealth applications in three categories: non-profit, for-profit, and hybrid. The models that existed at the time of this World Bank study, year 2012, were mainly of the form of for-profit or hybrid. This study analyzed 60 mHealth applications from three developing countries, Haiti, India, and Kenya. In case of Haiti, majority are in health related surveillance; in case of India, majority are in health related treatment support, and in case of Kenya, majority were in health education and awareness. This shows that the application areas of mHealth may depend on the local characteristics of particular countries. Based on a somewhat comprehensive study by healthcare experts, Librique et al. [8], had identified 12 areas of mHealth applications mostly deployed in developing countries.

1.3 mHealth

In case of India, for example, it already had an established telemedicine systems and its expanding mobile phone market gave it an opportunity to further expand its mHealth application base. Many such mHealth applications in India enabled

consumers and health workers to receive medical advice using technology rather than having to rely on face-to-face interactions [10].

1.3.1 mHealth Applications

The mHealth application areas are in plenty, from communication methods, health support systems, baby delivery support, diagnostic system, plus more [11–16]. A more recent study by Ippoliti and L'Engle [17], noted the sexual and reproductive health (SRH) as growing area for mHealth because patients prefer not to meet health professionals face to face. Their research, while noticing the increasing interest and promise from both developed and developing countries, focused mostly on SRH mHealth applications from developing countries—countries from Africa (67%), followed by Eurasia (26%), and Latin America (13%).

A new area that comes under mHealth is the proliferation of Apps downloaded via App stores and running on smart phones; Byambasuren et al. [18], states that a 2017 study found 318,000 Apps related to health available in App stores. However, majority of such health Apps are lifestyle Apps, such as Health App found in iPhones, targeting affluent urban users. However, successful applications of mobile phones to collect data in healthcare has been reported from many developing countries as well [19–21].

1.4 Need for Reliable Patient Data

A pivotal factor in successful applications of mHealth with regards to data is their capability and contribution as an advanced tool to collect credible and consistent patient data, especially from rural areas of developing countries where large percentage of the country population lives. Latest World Health Statistics 2018 report, expounds upon critical role of credible patient data in achievement of Sustainable Development Goals (SDGs) in the area of healthcare in developing countries [22]. Director General of World Health Organization, Dr. Tedros Adhanom Ghebreyesus appropriately emphasizes the importance of this fact in his statement "*Maintaining the momentum towards the SDGs is only possible if countries have the political will and the capacity to prioritize regular, timely and reliable data collection*" in the 2018 report Monitoring Health for the SDGs, Sustainable Development Goals [23]. Such "timely and reliable" data collection from developing countries, especially geographically large ones, in Africa and Asia, is a nightmare due to multiple problems, such as lack of trained personal, lack of proper health infrastructure as well as computer or communication infrastructure, also due to sheer vastness and the sparse population spread of countries. These problems are amply highlighted by WHO [23] study report and many other studies, including Stresser et al. [24], Mills [25], Agarwal [19], Walker [26], Sukums et al. [27], Tsujimura et al. [28, 29].

1.5 Rise of mHealth

According to a 2012 study compiled by PwC, with the help of Economist Intelligence Unit, the mHealth interest in emerging countries, which included, Brazil, India, and Turkey, is considerably higher, compared to developed countries [30]. Also mentioned in the same study was the fact that more mHealth services are getting covered by health insurance schemes in emerging countries than in the developed countries, a fact that could perhaps be alluded to the government regulations necessitated due to insufficiency of hospitals and medical personnel in remote and rural areas.

1.6 Objectives of Chapter

The purpose of this article is to highlight the innovative mHealth solutions in developing countries, especially the ones which also produce reliable data that can gradually accumulate to big data in a way to empower the development of rural healthcare. As part of the study, the paper documents two cases by providing a review of current data collection methods with regards to "reliable healthcare data" and the resulting analytics for healthcare in developing countries.

2 Big Data and mHealth

The Healthcare sector, in general, is expected to be a great beneficiary of big data, where the origin goes down to collection of data into electronic records [31]. Even as far back ten years ago, studies had found that some developed countries such as UK, Netherlands etc. had more than 90% of general medical practitioners using electronic health records (EHRs) [32]. Often, electronic medical records or EMRs are synonymously used to refer to EHRs; in some cases, in a broad sense, health information systems (HISs) also covers EHRs [33–35].

However, studies show that hospitals, even among developed countries, are not in the forefront of creating EHRs and thus the usage of big data [36]; though, some countries, such as US had made great progress, especially over the last 4 years or so. The U.S. Department of Health and Human Services' health related IT information website reports data shows the adaptation among US hospitals EHR zooming from 25% of the hospitals to almost 100% during 4 years starting from 2012 to 2015 [37]. A comprehensive review of literature on large-scale implementations of health information systems, by Sligo et al. [35], found many inconsistencies in the way that EHRs are implemented and thus, the way health data is collected.

2.1 Big Data in Healthcare and mHealth Enablers

However, the recent technological advancements both in data storage and analytics have made it possible to tap into vast data sources in all kinds of areas, including healthcare. Particularly, a report in 2013, titled "The 'big data' revolution in healthcare" by consulting company Mc Kinsey [38, 39] had prompted huge interest among the IT startup communities [40]. In India, for example, the funding for healthcare startups jumped more than 60% from 2016 to 2017 [41].

2.1.1 Rising Popularity of Big Data in Healthcare and mHealth

While 20th century witnessed exponential growth in human healthcare, 21st century is shaping up to a situation where humans and machines working together for better health, with combination of human reasoning and deep learning [42]. With popularity of big data applications and particularly mHealth applications concerning rural health being on consistent rise, authors felt it was necessary and relevant to give a detailed bibliometric analysis of publications.

3 Literature Review

Current chapter provides combination of bibliometric analysis and descriptive analysis for its literature review. However, bibliometric analysis has been kept short to include the important discussions to keep chapter within readable expanse.

3.1 Bibliometric Analysis

This section provides an analytical overview from different perspectives of bibliometric analyses, which utilizes a more comprehensive hybrid approach by providing concept centric descriptive review supported by visualization empowered bibliometric analyses. Web of Science database, Scopus, and Google Scholar were used to carry out a thorough enquiry. The syntax for search purposes was *mHealth+rural healthcare*. The initial searches resulted into total of 381 articles in Scopus and 141 articles in Web of Science. The search was further narrowed for relevance with following sequential iterations.

1. Publications in English.
2. Articles from Journal, conference proceedings and book review or chapters.
3. Articles published 2010 onwards, so this criteria was untouched.
4. Articles from relevant fields were sorted like Medical, Nursing, Medicines, IT, Computational science etc.

Table 1 Bibliometric summary for mHealth and Rural healthcare searches

Documents	141
Sources (Journals, Books, etc.)	65
Keywords plus (ID)	397
Author's keywords (DE)	455
Period	2010–2018
Average citations per documents	10.7
Authors	815
Author appearances	958
Authors of single-authored documents	4
Authors of multi-authored documents	811
Single-authored documents	3
Documents per author	0.173
Authors per document	5.78
Co-authors per documents	6.79
Collaboration index	5.88

Scopus and Web of Science search results were compared for degree of relevance. Post detailed study; it was decided to use Web of Science search results for detailed bibliometric analysis. However, other databases have supported for a comprehensive descriptive analysis for literature review.

Records for 141 documents post above mentioned filtration served as feed for bibliometric analysis from sources as Journals, books, and conference proceedings totaling to 65 in numbers [43]. The effective period selected was 2010–2018. Total number of varied keywords used by authors in selected research documents being 455. Average citations per document amount to 10.7. Collaboration index as calculated by R Studio software is 5.88; with 5.78 authors on an average writing one document. More details can be referred to in following table (Table 1).

The publications, when looked at on yearly plot (Fig. 1) shows a sudden jump in 2013, which can be greatly correlated with rise in applications on big data analytics particularly in healthcare, showing a rising trend.

The authors also ran a cluster analysis to identify the country of origin of authors of the filtered out 141 papers (Fig. 2). The size of the node represents the extent of collaboration. The collaborations are strong among USA, UK, Australia, India and Bangladesh authors, as evident in the graph. China, Kenya, Canada, Uganda, Ghana, and Spain are other influential research contributor nations.

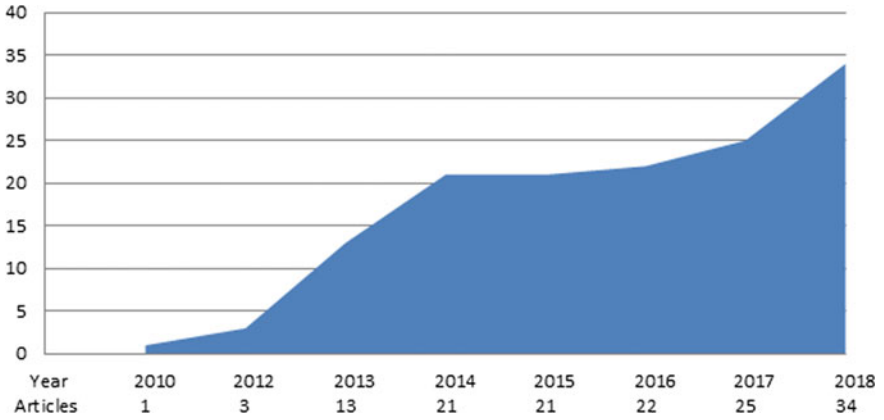


Fig. 1 Average article citations per year for mHealth in Rural healthcare

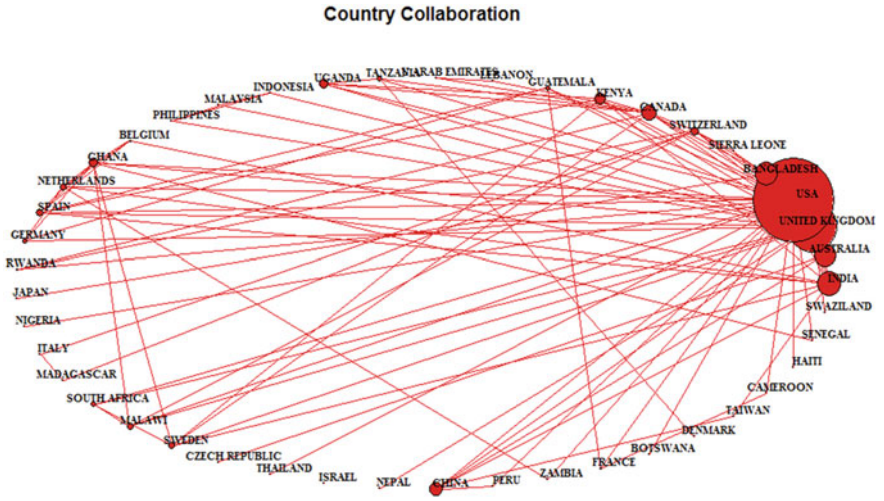


Fig. 2 Cross-Country collaboration for authors of mHealth and Rural healthcare

3.2 Descriptive Review of Big Data-Healthcare from Rural Populations in Developing Countries

Serious problem when it comes to data collection occurs in developing countries, as majority of the population still lives in rural areas. Especially in developing countries in African, Asian, and South American continents where, 70–80% of the population still resides in rural areas. It's equally true for even a big developing country such as India, where rural inhabitants account for 66 percent of the population [44], World Bank [45].

3.2.1 Challenges Concerning Big Data in Healthcare

Data Collection Challenges

When it comes to data collection, another major problem is the availability of facilities and healthcare personnel in rural areas. According to WHO statistics, Sub-Saharan countries in Africa, some countries in South Asia, and some countries in South America have severe shortages of doctors and rural medical care is in worst shape [46, 47]. In may such countries, the technology infrastructure, such as computer facilities, organized, data input methods, and communication networks needed to transmit and store data do not exist.

Some problems in data collection methods occur in developed countries as well. In a study covering data issues with government data depositories in US and Europe, [48], highlights the issues created by government regulations; they particularly point out the health related data collection problems. In another thorough study on big data in health sector, [49], identify three main areas where quality of data cannot be assured: (a) data entry errors, (b) data staging errors—errors that can happen when transferring data from one source to another; for example, from one database to another database, and (c) relevance and context, where the collected data may mean something not entirely related to the purpose expected from the data.

Challenges with Data

According to [48], 90% of world data is unstructured. This problem is prevalent in health data as well, because of various types of data collection methods and accumulation of data from various sources—where they sit in unstructured forms [50]. As mentioned a large percentage of populations in developing countries still live in rural areas. The Table 2 shows how the percentage of rural population had changed between 1960 and 2017 for some selected countries that are addressed in this paper with regards research experiences of the authors [45].

In many of these developing countries, medical care for rural populations still remains a major problem [24, 25].

Manpower Resource Challenges

In a thorough study on health workers in developing countries, Agarwal et al. [19], reported that frontline health workers play the role of midwives, nurses, pharmacists, doctors, and community health workers. Often these frontline health workers are selected from same local areas where they were born and familiar with the localities [26]. But, they usually lack proper training to do all the roles that they play; in addition, they lack proper medical equipment to collect authentic medical data.

Table 2 Percentage of rural population, 1960 versus 2017

Country	1960	2017
Afghanistan	92	72
Cambodia	90	79
China	84	42
India	82	66
Mongolia	64	26
Nicaragua	60	41
Pakistan	78	60
Sri Lanka	84	82
Uganda	96	83
Japan	37	6
USA	30	18

Source World Bank [45]

4 mHealth and Healthcare Cases

With the advancements of mobile technology, especially since the dawn of 4G mobile technologies, which allowed transferring text, images, videos, the mHealth had gained new dimensions. The most dramatic visible impact in mHealth in recent times is credited to introduction of smart phones and user-friendly health Apps which has made collection of credible patient data into computer databases easy like never before [51, 52], Andreatta, [53]. However, a vast majority of the healthcare Apps on smart phones are targeted not at developing country patients, but the affluent users in developed countries.

We propose implementation of these health Apps through front-line medical facilitators in Rural areas using smart devices such as mobile phones or tablets for collection and accumulation of data. If mHealth applications work without internet it will help collect credible patient data, and Big data analytics using this vast amount of data thus collected can truly serve as key factor to attain SDGs in Healthcare areas for developing countries [54]. Moreover, ease of interpretation empowered by AI & Visualization as depicted in the second case can empower remote patient monitoring.

4.1 Prelude-Case Studies

4.1.1 Case from Japan-Cambodia

In a study funded by Japanese government, authors conducted empirical research with extensive field work to explore propositions made. The phenomenon of front-line medical facilitators reaching out to rural communities for providing medical care was observed closely. Major field observations have revealed critical challenge

of frugally available medical facilities and those too were geographically limited to major city areas only. The front-line medical facilitators are basically nurses or mid-wives. They can only infrequently visit these rural communities and moreover; the healthcare data collection methods in these areas still remain very primitive. Unfortunately, this collected data never gets entry to any computer database. Often, they do not use any electronic device as rural areas are often out of reach from mobile networks such as 3G or 4G. Considering these field observations based on real picture in order to achieve objectives of SDG's by 2030 in healthcare area; Researchers highlight this dire need for mHealth innovations as demonstrated by our Researcher team, to become widespread applications covering large percentage of world population living in rural areas of developing countries overcoming challenge of non-reachability of mobile networks.

The traditional methods to access rural communities in developing countries for healthcare include mobile health services and telemedicine. Current chapter includes a review of such methods plus a state-of-the-art literature survey with regards to interesting modern applications of rural healthcare data collection methods. This chapter however; most critically outlines a data collection platform designed and developed by our Researcher team to work in rural areas without network coverage as successfully demonstrated through field experiments in Cambodia as part of Japanese government supported research project. Based on findings from post experiment survey of Nurses serving the rural communities in Cambodia; this platform and related App was found to be extremely useful to collect rural healthcare data where 86% of the surveyed nurses supported the fact.

4.1.2 Case from India

Authors also have an interesting perspective to offer in second case from India in rural healthcare. The second case as part of the chapter gives an overview of an innovative Remote patient monitoring system designed by a team of researchers from IIT Bombay. It is a state-of-art system that has not only had a major impact but can bring revolution in rural healthcare in India. The case talks about the company incubated in IIT Bombay which has made remarkable contribution to patient monitoring and rural healthcare with its innovative tele-diagnosis and medicine system. Our case has been crafted to first talk about historical information about the company, the need felt to improve rural healthcare which has inspired development of products and services playing significant role in improving quality of healthcare in India. The case further discusses relevance to the contextual factors in India presently in relation to nationally important agenda of 'Ayushman Bharat'. This is a world's biggest National health protection scheme launched in 2018 with an aim to provide healthcare facilities to over 100 million families covering urban or rural poor and an insurance worth Rs. 0.5 million to over 500 million citizens [55]. Case throws light upon the expanse and reach with special focus on impact A3 Remote Monitoring Technologies has created. Present case enlightens readers about the perspective of affordability especially with contextual reference to developing countries. Case also

discusses vision ahead and concrete plan to help realize this vision in congruence with national agenda of 'Ayushman Bharat'. Case also highlights 2 critical remote patient monitoring situations managed well with companies' innovative mHealth solution which demonstrate the capabilities that can be achieved in improving remote or rural healthcare.

5 N+Care App—Empowering FHWs-Case Japan-Cambodia

In a study funded by Japanese government since 2013, authors conducted empirical research with extensive field work to explore how to empower the front line health workers (FHWs) in developing countries to improve their services using mobile devices. As Agarwal et al. [19] mentions the FHWs play multiple roles as nurses, midwives, doctors, and even dispenses medicine, sometimes [56]. The observations by the authors in several developing countries, such as Mongolia, Nicaragua, Sri Lanka, and Uganda, in rural areas have confirmed the critical roles played by the FHWs.

5.1 Insights for N+Care App from Stakeholders

The discussions with doctors, nurses, and staffs at medical training institutions selected from those countries revealed interesting facts. If FHWs are to use a mobile device, such as a tablet, in order to enhance their service, including collecting patient data, consideration needs to be given to following facts:

- Tablet must be affordable
- Tablet must work in areas without Internet connections
- The App that FHWs use must have local language capabilities
- The collected records can be uploaded from mobile to a server PC at base hospital (usually located at a regional township; the FHWs travel far, sometimes 30 km or more into villages to see/monitor patients)
- The data from server PC needs to be downloaded, as needed to a tablet, prior to making next visit to see the patients.

Indeed, the current practice in many rural health facilities authors visited in those countries FHWs have no electronic means to collect any data; usually they carry some note books, but records are often lost and the data never gets to a computer.

5.2 About N+Care App Design

Taking into account the concerns stated above, the authors developed the App called N+Care (N for "Nursing"). It is designed to run on Android platform because almost all low-cost tablets run on Android operating system and they are widely used in developing countries. The database was designed to collect not only patient data, but also with option to add photos; for example, the photo of an injury. Also consideration was given to add recommended doctors based on the advice of the FHW, in case patient need to follow up with a hospital visit.

N+Care App is designed to collect the following key data:

- FHW identification
- Date and time FHW meets the patient
- Patient data (can include two photos per visit)
- Disease, Diagnostic, and description of condition
- Tests conducted and their results, medicine and prescription detail
- Doctor recommendation detail.

Part of the database design of N+Care App is shown in Fig. 3.

Some screenshots of the N+Care App that was tested in Cambodia is shown in Fig. 4.

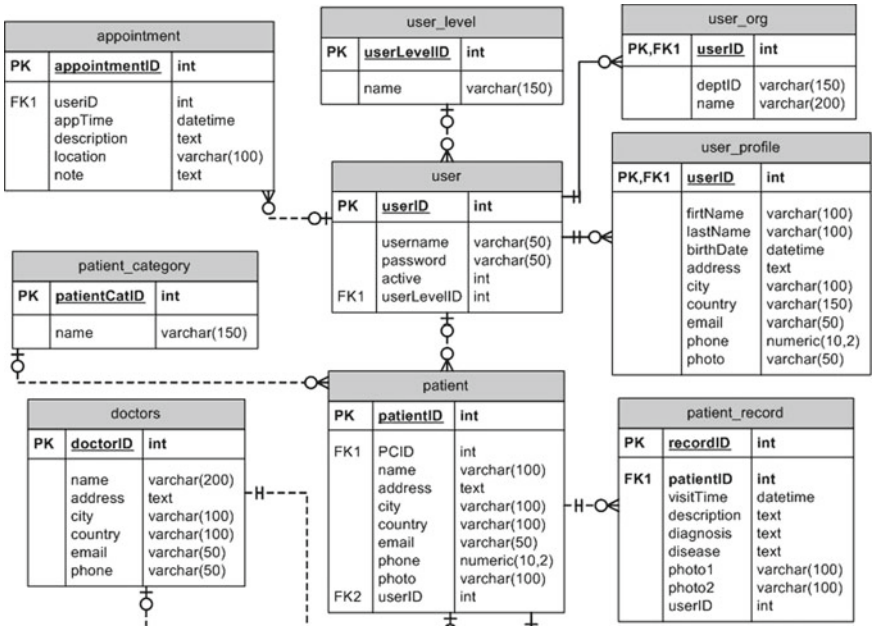


Fig. 3 Part of N+Care android app database

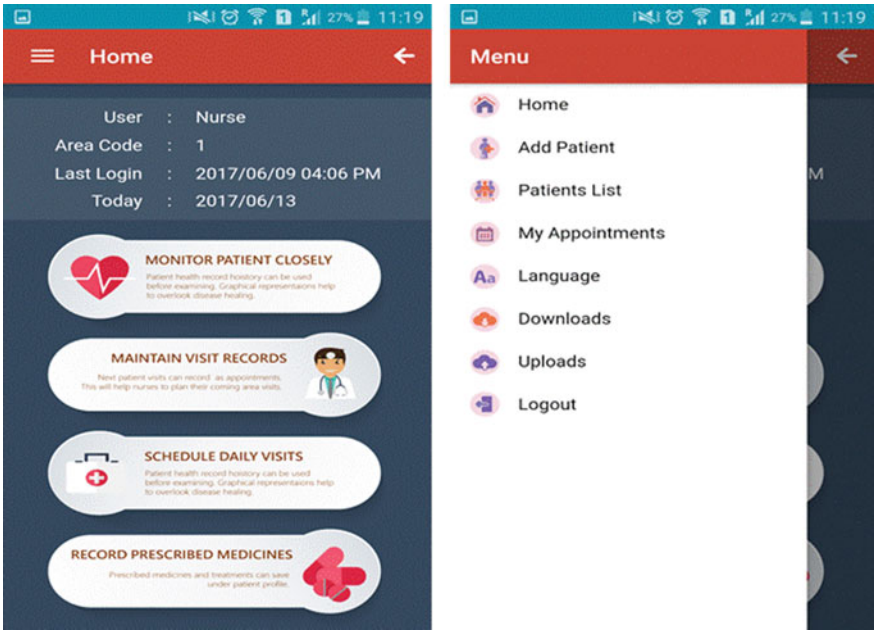


Fig. 4 Two screen-shots of N+Care Android App

5.3 Why Cambodia for N+Care App Study

The N+Care App was 1st tested in Cambodia, because this is a developing country in need of help to enhance the medical care in rural sector. Overall, the same WHO data that were used for Fig. 3, shows that in year 2011, the latest numbers the data were available, Cambodia had 8 nurses per 10,000 inhabitants [23], a considerably low number compared to world average of about 45 nurses per 10,000 inhabitants.

Cambodia health infrastructure is a typical case of disparities between urban and rural healthcare. A good overview of Cambodia health infrastructure is given in Chhea et al. [57], which cited that 54% of country's physicians are employed in the capital city of Phnom Penh, where 9.3% of the country's population lives. In rural areas 8,000 to 12,000 people are served by some entity called "health center" which provides basic treatments for common deceases and refers serious cases to "referral hospitals". Chhea et al., mentions that two-thirds of "health centers" could not provide basic services due to shortage of midwives and nurses. Though country has been making considerable progress since the end of civil war more than 20 years ago, such disparities prevail [58], as almost 80% of Cambodia's population lives in rural areas [45].

The N+Care App, with its Cambodian language turned on, was tested with the nurses from public as well as private "health centers" from rural provinces in Cambodia. Altogether 105 of the nurses were surveyed with the help of University of Health

Table 3 N+Care App, as evaluated by Cambodian nurses serving rural areas

N+Care Apps usefulness for improving nursing care service for patients		
	Frequency	Percentage
Yes	90	85.71
No	4	3.81
Not much	11	10.48
Total	105	100%

Sciences, a public university, and Norton University, a private university, both located in the capital city Phnom Penh. Among the key factors examined in the survey was N+Care Apps usefulness to improve nursing care service to the patients (Table 3).

In a related question, when asked about the actual use of N+Care App, if available on a tablet, 73 out of 105 or 69.52% of the nurses responded positively.

5.4 Results of Study in Cambodia

The results suggest that there is appeal and willingness to use N+Care App by the nurses serving in rural areas as FHWs. The App is now being considered to be distributed in Sri Lanka, in association with Health Ministry and a mobile career to be distributed to nurses serving the rural areas.

With the functionality that N+Care has to collect and store the data from rural areas would gradually build up a credible database of rural health data, which would accumulate to big data.

A3 (Anywhere Anytime Access) Remote Monitoring:
Remote Monitoring Technologies is a decade old Indian Institute of Technology: Bombay (IITB) Company founded by 3 co-founders Dr. Shrikant Parikh, Dr. Sunil Nakdawala and Dr. Uday Deasi. The company traces its genesis to Indian Institute of Technology-Bombay, where it was incorporated in 2008 and incubated from 2009 to 2012 with IIT-Bombay Incubation Centre-Society for Innovation and Entrepreneurship. A3 is a pioneering State-of-Art technology company with its current operations in India and has devised a completely new approach to the wireless remote patient monitoring strategically different than its predecessors. One of the critical incident and personal experience of the founder Dr. Shrikant became a trigger for innovation of the technology and formation of the company later. Dr. Shrikant's father-in-law suffered a medical emergency situation with myocardial infarction late in the night. Dr. Shrikant struggled to reach out to his doctor at that point in time. Tis struggle left a deep impact in his mind where he realized dire need to help improve co-ordination between doctors-patients and hospitals beyond spatial constraints.

6 Case from India-A3 Remote Monitoring Technologies

The need for telemedicine in general and mHealth in particular can be attributed to global healthcare trends of global rise in population of patients with critical, chronic and emergency conditions; and towering healthcare costs-biggest challenge in inclusion as a Sustainable Development Goal (SDG). Anytime Anywhere Accessible (A3) itself very comprehensively summarizes the need where, patient information can be accessed remotely anytime and anywhere. Moreover, in developing countries as discussed in the introductory section the Doctor/Patient ratio is very low when it comes to Rural areas. Telemedicine centre did help increasing the reach but at the same time requires array of voluminous and heavy hi-tech infrastructural facilities which increases overhead costs and also makes transportation very difficult & further costly with logistics expenses. A3 Remote Monitoring Technologies provides latest technology in a compact portable model which can be accessed in Ambulance, Medical Vans, Patients House, Primary healthcare centres (PHC's). The technology is user friendly and can be operated easily with basic training by Technicians or Healthcare supporters in Anganwadis (Pre-primary Schools)/PHC's.

6.1 A3 Remote Monitoring Technology

A3 is an innovatively designed state-of-art technology with user-friendly, portable health system enabled to monitor cardiac, lung, pathology, and gynecology related parameters. The system is designed to store data on cloud servers. In case of non-availability of internet facilities; data for several patient tests can be stored in the system, which can later push this data to cloud when internet facility becomes available. A3 has data back-up and restore functions built in.

6.2 Healthcare Services Offered by A3

6.2.1 Health Data Related to Cardiology

Cardiology parameters such as ECG, BP, Heart rate, Pulse rate, Oxygen saturation, Respiration rate, etc. can be easily monitored and can be captured in screen-shots then as required are shared with concerned Doctor. The image includes Patient's name, Identity number, photo and other vital signs as depicted in Fig. 7. The powerful support of visualization in this innovative patient monitoring technology makes these images easy to be read by non-medical personnel, empowering them to decide if patients be referred for further help of medical personnel as depicted in Fig. 6. As discussed, in previous sections, patient data gathering is critical, however; authors propose that AI & Visualization empowered patient reports and data needs to empower

local guided action in remote areas leading to appropriate and timely healthcare decision (Fig. 5).

Healthcare data related to ENT-Cameras as part of A3 Remote monitoring Technologies enable monitoring of health parameters in case of eyes (Fundus), ears (Otoscope) and skin (Demoscope).

6.2.2 Health Data Related to Pregnancy

With Stethoscope, A3 system can capture Fetal heart sound as part of Doppler test to monitor heartbeats of fetus. India which has high Infant and Maternal mortality rate needs such facilities as priority. Moreover, in rural areas the rates are higher as compared to urban counterparts. Pregnant lady with BP at higher level may be at risk when traveling due to bad road conditions. Also, any technologically equivalent facility may not even be available at PHC which could be 60–70 kms farther from where pregnant lady stays.

Integrated Electronic fetal monitor is another powerful feature of A3 technologies. In the risky third trimester of pregnancy as mentioned earlier, this equipment helps monitor uterine contractions and gives with analysis the possibility of a normal delivery of baby as viable or risky.

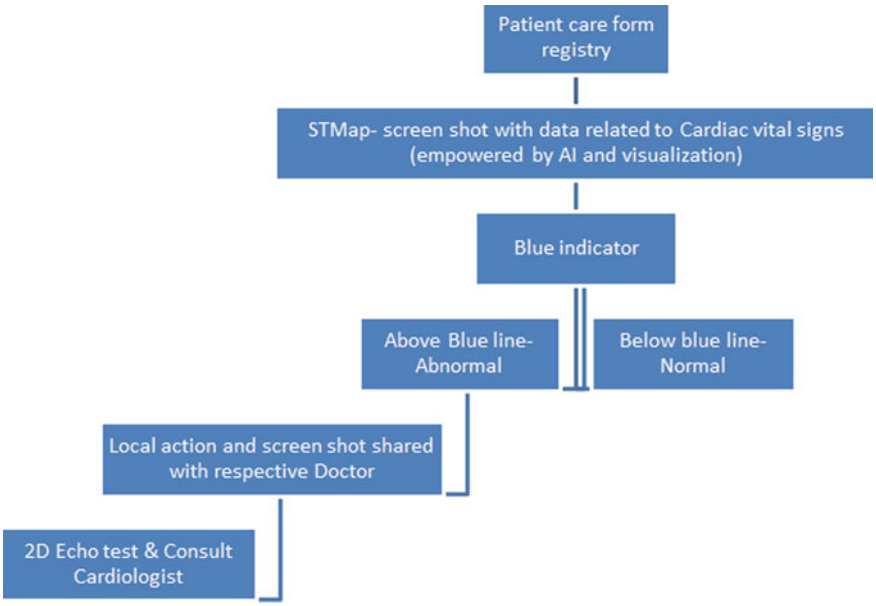


Fig. 5 Process flow map depicting ease of action in case of cardiac abnormalities observed with STMap empowered by AI & Visualization

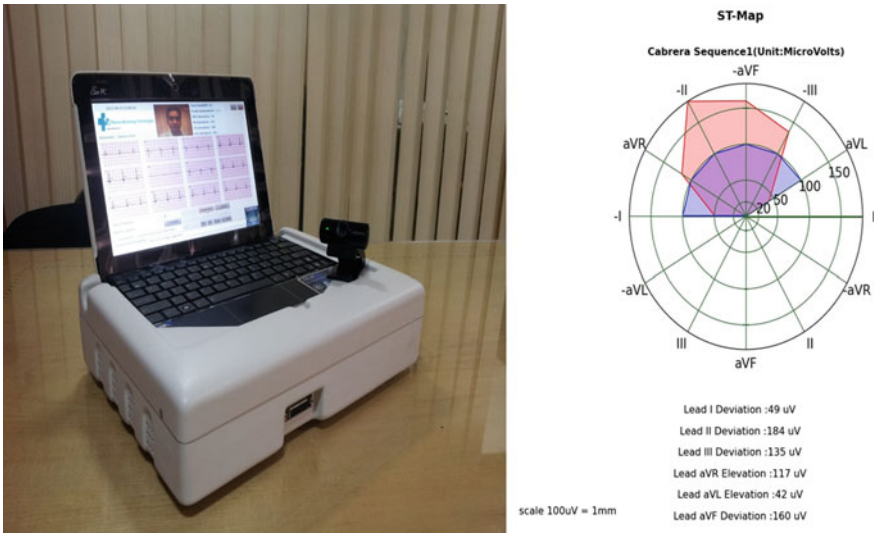


Fig. 6 Part A—A3 Remote patient monitoring system, Part B—ST-Map showing visualization empowered report for ease of interpretation beyond medical experts in Rural healthcare

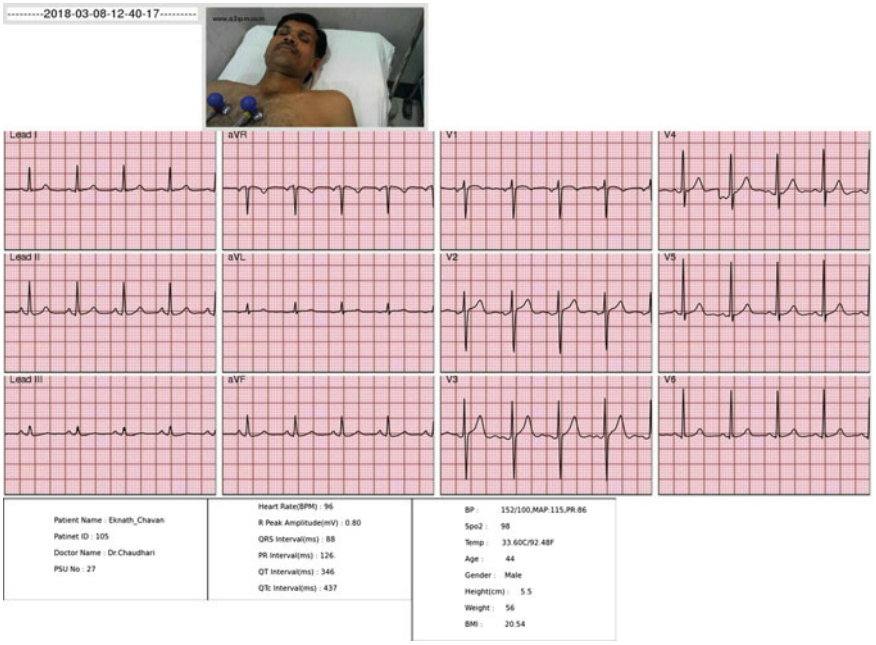


Fig. 7 A3 report with patient details in case of Cardiac care

Health data related to Pathology-A3 technology supports over 40 tests with Chemistry analyzer for blood, urine and stool samples collected. It can help monitor the Hemoglobin levels in case of again pregnant women which is another major area of concern when it comes to Maternal and infant mortality or malnutrition. The technology is fraud proof and generates report based upon samples of patients tested—disabled for any manual entry, thus, rules out possibility for fake reports and manipulation. Thus, A3 technology will also empower Government of India with true data captured real time with system being manipulation or fraud proof, empowering better Public health governance.

6.3 A3 Technology Penetration in India

A3 has been operating in over 11 states of India with 1.5 million plus population being covered and 100,000 plus patients being screened across 500 plus locations from villages, semi-urban or urban areas. A3 claims over 2000 lives saved with its efforts to bring real-time patient healthcare data to medical professionals for action-anytime and anywhere resulting into timely intervention and care. A3 technology services have been offered through various clients as governmental healthcare agencies, NGO's working towards health and education and also corporate who regularly monitor employee health as part of Occupation, Employee safety and Health regulations.

6.4 A3 Technology Affordability

A3 technology is offered via its varied clientele as mentioned in above statements depending upon their needs and modular capabilities. However, the technology is very cheap at operational level, with negligible cost of Rs. 2–3 or max. Rs. 100 including the cost of consumables used as part of the test.

6.5 A3 Technology Case-Lets Highlighting Remote Patient Monitoring Capabilities

As part of National scheme of Digital India, over 642 villages will become digital in the state of Andhra Pradesh with state government being led by Mr. Chandrababu Naidu. This experiment is being driven by Prof. Solomon from University of Berkley, who selected A3 technology for digital health. After, successful implementation at Mori, a pilot village in AP, now, A3 will be empowering sustainable smart health in over 642 villages.

In a second equally innovative endeavor, A3 technology was recommended for monitoring health for Indian army soldiers at remote place as Siachen glacier. With non-availability of medical facilities in vicinity at such places and extreme weather conditions, pose a challenge and dire need for regular remote health monitoring for Indian army personnel. A successful pilot test was carried out with foundational training to Soldiers, who are now empowered to monitor health parameters as decided by Army authority for each other. The data sharing happened with the help of Tata Net VSat in the first phase and after successful trial now will happen through dedicated Optical fiber network with Indian Army.

6.6 Vision at A3 Technology-Ayushman Bharat

A3 has vision to further penetrate all states pan India and increase its reach internationally too, especially with a focus on South-East Asia and African continent. A3 health systems are generally economically supported through a PPP model. With 2019 soon approaching, A3 has concrete plans to reach the figure of over 2.5 lakhs of patients being regularly monitored. It has shared its vision to reach a mark of saving over Million lives in near future. As part of nationally important agenda of 'Ayushman Bharat', A3 Technology aims to empower preventive healthcare which will save several invaluable human lives and will also have great economic impact by reducing health related economic burden. As per research stated in Introductory section, families not covered for health risks end up spending maximum of their savings of medical costs. A3 technology has power to change the picture of Public health governance in India. A3 technology has partnered with IBM and WATSON for AI and Machine learning tools for enabling predictive analytics based on huge amount of public health data thus available.

7 Conclusion and Future Scope

Chapter dedicated its discussion on Big data applications in healthcare with special reference to rural healthcare. Introductory sections describe the need for Big data applications in healthcare towards meeting SDG's. Further, the focus shifts to rise of mhealth applications world over, including overview of developing countries. In the next section, authors discussed a brief summary of researches published and available for overview on Big data applications in healthcare with focused on rural healthcare. The chapter highlights the current state of health information systems, EHR's and need for reliable patient data.

Next few sections document two detailed cases from Japan with its study in Cambodia and another one from India. The N+Care app case discusses the outcomes of initial study, followed by design details and results of the study. N+Care app

has its innovative features of overcoming challenges of reliable patient data collection, which authors think will be appreciated by all researchers. While, second case study from India; A3 Remote monitoring technologies provides valuable insights into remote patient data monitoring. Over 45 diagnostic services with special capabilities in Cardiac and pregnancy care make it a great learning for the audience. Case study discusses success caselets with separate discussion on affordability, penetration of technology and vision ahead.

For future research directions, authors recommend a longitudinal coverage of these cases to demonstrate their success on a larger scale and the impact it has made. Any further, collection of cases in mHealth will always be a welcome considering the developments happening in the field, especially with its applications in rural healthcare. AI in healthcare is catching up and will heavily impact developments in technology enabled healthcare and rural healthcare is no exception.

References

1. Adibi, S. (Ed.).: Mobile Health: A Technology Road Map, vol. 5. Springer (2015)
2. Istepanian, R.S., Lacal, J.C.: Emerging mobile communication technologies for health: some imperative notes on m-health. In: Proceedings of the 25th Annual International Conference of the IEEE, vol. 2, pp. 1414–1416. Engineering in Medicine and Biology Society. IEEE (2003)
3. Mechael, P.N.: The case for mHealth in developing countries. Innov. Technol. Gov. Glob. **4**(1), 103–118 (2009)
4. WHO-Unicef: World Health Organization, & Unicef: Primary Health Care: A Joint Report (1978)
5. Sood, S.P., Bhatia, J.S.: Development of telemedicine technology in India:"Sanjeevani"-An integrated telemedicine application. J. Postgrad. Med. **51**(4), 308 (2005)
6. Akter, S., Ray, P.: mHealth-an ultimate platform to serve the unserved. Yearbook of Med. Inform. **19**(01), 94–100 (2010)
7. Hampshire, K., Porter, G., Owusu, S.A., Mariwah, S., Abane, A., Robson, E., Milner, J.: Informal m-health: how are young people using mobile phones to bridge healthcare gaps in Sub-Saharan Africa? Soc. Sci. Med. **142**, 90–99 (2015)
8. Labrique, A.B., Vasudevan, L., Kochi, E., Fabricant, R., Mehl, G.: mHealth innovations as health system strengthening tools: 12 common applications and a visual framework. Global Health: Sci. Pract. **1**(2), 160–171 (2013)
9. Qiang, C.Z., Yamamichi, M., Hausman, V., Altman, D., Unit, I.S.: Mobile Applications for the Health Sector, p. 2. World Bank, Washington (2012)
10. Ganapathy, K., Ravindra, A.: mHealth: a potential tool for health care delivery in India. In: Proceedings of the Making the ehealth Connection: Global Partnerships, Global Solutions (2008)
11. Agarwal, S., Labrique, A.: Newborn health on the line: the potential mHealth applications. JAMA **312**(3), 229–230 (2014)
12. Conway, N., Campbell, I., Forbes, P., Cunningham, S., Wake, D.: mHealth applications for diabetes: user preference and implications for app development. Health Inform. J. **22**(4), 1111–1120 (2016)
13. Dunsmuir, D.T., Payne, B.A., Cloete, G., Petersen, C.L., Görges, M., Lim, J., Ansermino, J.M.: Development of mHealth applications for pre-eclampsia triage. IEEE J. Biomed. Health Inform. **18**(6), 1857–1864 (2014)

14. Huang, A., Chen, C., Bian, K., Duan, X., Chen, M., Gao, H., Xie, L.: WE-CARE: an intelligent mobile telecardiology system to enable mHealth applications. IEEE J. Biomed. Health Inform. **18**(2), 693–702 (2014)
15. Kyriacou, E.C., Pattichis, C.S., Pattichis, M.S.: An overview of recent health care support systems for eEmergency and mHealth applications. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009, EMBC. pp. 1246–1249. IEEE (2009)
16. Miller, A.S., Cafazzo, J.A., Seto, E.: A game plan: gamification design principles in mHealth applications for chronic disease management. Health Inform. J. **22**(2), 184–193 (2016)
17. Ippoliti, N.B., L'Engle, K.: Meet us on the phone: mobile phone programs for adolescent sexual and reproductive health in low-to-middle income countries. Reprod. Health **14**(1), 11 (2017)
18. Byambasuren, O., Sanders, S., Beller, E., Glasziou, P.: Prescribable mHealth apps identified from an overview of systematic reviews. npj Digital Med. **1**(1), 12 (2018)
19. Agarwal, S., Perry, H.B., Long, L.A., Labrique, A.B.: Evidence on feasibility and effective use of mHealth strategies by frontline health workers in developing countries: systematic review. Trop. Med. Int. Health **20**(8), 1003–1014 (2015)
20. Aranda-Jan, C.B., Mohutsiwa-Dibe, N., Loukanova, S.: Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa. BMC Public Health **14**(1), 188 (2014)
21. Diwan, V., Agnihotri, D., Hulth, A.: Collecting syndromic surveillance data by mobile phone in rural India: implementation and feasibility. Global Health Action **8**(1), 26608 (2015)
22. Surie, M.D.: Why India Could Make or Break the Success of SDGs. Asia Foundation (2015)
23. WHO: World Health Organization. World Health Statistics 2018: Monitoring Health for the SDGs, Sustainable Development Goals. World Health Organization, Geneva (2018). CC BY-NC-SA 3.0 IGO
24. Strasser, R., Kam, S.M., Regalado, S.M.: Rural health care access and policy in developing countries. Annu. Rev. Public Health **37**, 395–412 (2016)
25. Mills, A.: Health care systems in low-and middle-income countries. N. Engl. J. Med. **370**(6), 552–557 (2014)
26. Walker, R.: Walking beyond our borders with frontline health workers in Guatemala. Nurs. Women's Health **17**(6), 533–538 (2013)
27. Sukums, F., Mensah, N., Mpembeni, R., Kaltschmidt, J., Haefeli, W.E., Blank, A.: Health workers' knowledge of and attitudes towards computer applications in rural African health facilities. Global Health Action **7**(1), 24534 (2014)
28. Tsujimura, H., Mori, Y., Miyakoshi, S., Pathiranage, A.M.S.D., Rathnayake, U.W.S.: Distance education for supporting nurse training in developing countries—assessment of nursing techniques for postural change using skype in Sri Lanka. Kitakanto Med. J. **64**(1). 44–57 (2014a)
29. Tsujimura, H., Mori, Y., Miyakoshi, S., Pathiranage, A.M.S.D., Rathnayake, U.W.S.: Distance education for supporting nurse training in developing countries. Kitakanto Med. J. **64**(1), 57–66 (2014b)
30. Unit, E.I., Cooper, P.W.: Emerging mHealth: Paths for Growth. PwC (2014)
31. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**(1), 3 (2014)
32. Jha, A.K., Doolan, D., Grandt, D., Scott, T., Bates, D.W.: The use of health information technology in seven nations. Int. J. Med. Inform. **77**(12), 848–854 (2008)
33. Kaneko, K., Onozuka, D., Shibuta, H., Hagihara, A.: Impact of electronic medical records (EMRs) on hospital productivity in Japan. Int. J. Med. Inform. **118**, 36–43 (2018)
34. Kawaguchi, H., Koike, S., Ohe, K.: Regional differences in electronic medical record adoption in Japan: a nationwide longitudinal ecological study. Int. J. Med. Inform. **115**, 114–119 (2018)
35. Sligo, J., Gauld, R., Roberts, V., Villa, L.: A literature review for large-scale health information system project planning, implementation and evaluation. Int. J. Med. Inform. **97**, 86–97 (2017)
36. van Velthoven, M.H., Mastellos, N., Majeed, A., O'Donoghue, J., Car, J.: Feasibility of extracting data from electronic medical records for research: an international comparative study. BMC Med. Inform. Decis. Mak. **16**(1), 90 (2016)

37. HealthIT (2018). Non-federal Acute Care Hospital Electronic Health Record Adoption. US Health IT Dashboard. https://dashboard.healthit.gov/quickstats/pages/FIG-Hospital-EHR-Adoption.php. Accessed 6 Dec 2018
38. Kaka, N., Madgavkar, A., Manyika, J., Bughin, J., Parameswaran, P.: India's Technology Opportunity: Transforming Work, Empowering People. McKinsey Global Institute (2014)
39. Kayyali, B., Knott, D., Van Kuiken, S.: The Big-Data Revolution in US Health Care: Accelerating Value and Innovation, vol. 2, no. 8, pp. 1–13. Mc Kinsey & Company (2013)
40. Senthilkumar, S.A., Rai, B.K., Meshram, A.A., Gunasekaran, A., Chandrakumarmangalam, S.: Big data in healthcare management: a review of literature. Am. J. Theor. Appl. Bus. **4**(2), 57–69 (2018)
41. Velayanikal, M.: What's Putting India on the World Map for Deep Tech in Healthcare. https://www.techinasia.com/deep-tech-startups-making-inroads-into-healthcare-in-india (2018). Accessed 8 Dec 2018
42. Morgan, J.: 7 Ways Big Data Analytics Can Boost Healthcare. https://www.beckershospitalreview.com/healthcare-information-technology/7-ways-big-data-analytics-can-boost-healthcare.html. Accessed Oct 2018
43. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (2018)
44. UNO: United Nations Organization. World Urbanization Prospects: The 2018 Revision. https://esa.un.org/unpd/wup/publications/Files/WUP2018-KeyFacts.pdf (2018). Accessed 28 Aug 2018
45. World Bank: Rural Population (% of total population). https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS (2018). Accessed 28 Aug 2018
46. Nikkei: Japan plans 10 'AI hospitals' to ease doctor shortages. Nikkei Asian Rev. https://asia.nikkei.com/Politics/Japan-plans-10-AI-hospitals-to-ease-doctor-shortages (2018). Accessed 28 Aug 2018
47. WHO: World Health Organization. Global Health Observatory Data Repository. http://apps.who.int/gho/data/node.main.A1444?lang=en (2018). Accessed 12 Nov 2018
48. Kim, G.H., Trimi, S., Chung, J.H.: Big-data applications in the government sector. Commun. ACM **57**(3), 78–85 (2014)
49. Sukumar, S.R., Natarajan, R., Ferrell, R.K.: Quality of big data in health care. Int. J. Health Care Qual. Assur. **28**(6), 621–634 (2015)
50. Mathew, P. S., & Pillai, A. S. (2015, March). Big Data solutions in Healthcare: Problems and perspectives. In: 2015 International Conference in Innovations in Information, Embedded and Communication Systems (ICIIECS), (pp. 1–6)
51. Medhanyie, A.A., Moser, A., Spigt, M., Yebyo, H., Little, A., Dinant, G., Blanco, R.: Mobile health data collection at primary health care in Ethiopia: a feasible challenge. J. Clin. Epidemiol. **68**(1), 80–86 (2015)
52. Rangan, A.M., O'Connor, S., Giannelli, V., Yap, M.L., Tang, L.M., Roy, R., Allman-Farinelli, M.: Electronic dietary intake assessment (e-DIA): comparison of a mobile phone digital entry app for dietary data collection with 24-hour dietary recalls. JMIR mHealth uHealth **3**(4) (2015)
53. Andreatta, P., Debpuur, D., Danquah, A., Perosky, J.: Using cell phones to collect postpartum hemorrhage outcome data in rural Ghana. Int. J. Gynecol. Obstet. **113**(2), 148–151 (2011)
54. Murdoch, T.B., Detsky, A.S.: The inevitable application of big data to health care. Jama **309**(13), 1351–1352 (2013)
55. Economic Times (2018). Ayushman Bharat Health Insurance: Who All it Covers, How to Apply. Accessed Dec 2018 https://economictimes.indiatimes.com/articleshow/65422257.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst
56. Vélez, O., Okyere, P.B., Kanter, A.S., Bakken, S.: A usability study of a mobile health application for rural Ghanaian midwives. J. Midwifery Women's Health **59**(2), 184–191 (2014)
57. Chhea, C., Warren, N., Manderson, L.: Health worker effectiveness and retention in rural Cambodia. Rural Remote Health **10**(3), 1391 (2010)
58. Ozano, K., Simkhada, P., Thann, K., Khatri, R.: Improving local health through community health workers in Cambodia: challenges and solutions. Human Resour. Health **16**(1), 2 (2018)

Mathematical Modeling and Solutions, Big Data Applications, and Platforms

Hospital Surgery Scheduling Under Uncertainty Using Multiobjective Evolutionary Algorithms

Kazi Shah Nawaz Ripon and Jacob Henrik Nyman

Abstract Surgery scheduling is the allocating of hospital resources to surgical procedures over time. These problems are continuously encountered in hospitals delivering surgical treatment to patients, and they must be solved with great effort and care in order to utilize scarce resources, balance conflicting interests and hedge for uncertainty. Therefore, machine learning algorithms are generally not directly applicable to surgery scheduling problems. The motivation of this work is to narrow the gap between evolutionary approaches to machine scheduling problems from literature and their practical applications to real-world hospital surgery scheduling problems. We formulate a new variation of the surgery admission planning problem and develop evolutionary mechanisms to solve it with state of the art multiobjective evolutionary algorithms. Our most important contribution is a fast simulation approach for evaluating surgery schedules under surgery duration uncertainty. By using Monte-Carlo simulation to estimate the fitness of individuals during optimization, we attempt to find solutions that are robust against variations in surgery durations.

1 Introduction

Surgery scheduling problems are particular instances of large, multiobjective machine scheduling problems with uncertain processing times and multiple resource constraints. These problems are continuously encountered in hospitals delivering surgical treatment to patients. It is an operational task that requires coordination of scarce resources, thoughtful prioritization of conflicting interests and constant response to unexpected events, such as delays and cancellations [1]. Labour-intensive scheduling activities are critical in most hospitals and directly linked to the rate at which patients

K. S. Nawaz Ripon (✉) · J. Henrik Nyman
Norwegian University of Science and Technology, Trondheim, Norway
e-mail: ksripon@ieee.org

J. Henrik Nyman
e-mail: jacobhnyman@gmail.com

K. S. Nawaz Ripon
University of Oslo, Oslo, Norway

A. J. Kulkarni et al. (eds.), *Big Data Analytics in Healthcare*,
Studies in Big Data 66, https://doi.org/10.1007/978-3-030-31672-3_7

receive treatment and the quality of care provided to each patient. Optimization of the surgery scheduling process is a topic of increasing interest as it can improve schedule quality in order to utilize surgeons, nurses, equipment and operating rooms better. Thus, high-quality surgery schedules have the potential of improving hospitals core business without incurring costs associated with up-scaling of resources [2].

Hospitals maintain a list of patients in need of surgery. Each surgery must be performed by a qualified surgeon in an operating room that facilitates the required equipment. In order to simplify and control the scheduling process, it is common to adhere to a *master surgery schedule*—a periodic time table that reserves time slots in operating rooms for specific surgical subspecialties [3]. Most surgeries may be planned for well in advance to ensure smooth coordination of staff (e.g. surgeons, nurses, anesthesiologists), facilities (e.g. operating rooms, post-anaesthesia care unit, intensive care unit, hospital beds) and equipment. However, carefully designed schedules are continuously disrupted by unpredictable events such as the arrival of emergency patients and delays in the operating room [1]. It is up to the schedulers to balance the need of patients waiting for surgery against the schedule density, risk of overtime and surgery cancellations. This tedious coordination, prioritization and risk evaluation has been subject to automation in recent years [4].

Hospitals environments are dynamic, and it is governed by uncertainty, difficult priorities and careful coordination of scarce resources. Still, in most hospitals, surgery scheduling is performed manually, partly because surgery scheduling algorithms fail to capture the dynamics in real hospital scenarios accurately. Therefore, machine scheduling algorithms are generally not directly applicable to surgery scheduling problems [1, 3]. This work attempts to integrate the power of evolutionary machine scheduling algorithms with the dynamics in real-life hospitals by minimizing five objectives: (i) patient waiting time on the waiting list, (ii) surgeon overtime, (iii) surgeon idle time, (iv) operating room overtime and (v) operating room idle time.

We address how two state of the art multiobjective evolutionary algorithms (MOEAs), the strength pareto evolutionary algorithm 2 (SPEA2) [5] and the non domination sorting genetic algorithm II (NSGA-II) [6], can be applied with new evolutionary mechanisms to solve surgery scheduling problem instances that are realistic in terms of problem size, multiobjectivity and uncertainty. While multiobjectivity and uncertainty have been considered separately in many recent works, few consider them in combination. Thus, our approach instead uniquely integrates two well-acknowledged complications in an already complicated problem. For these, we also present a new variation of the existing surgery admission planning problem [3]. In the presented problem, patients from a patient waiting list are to be scheduled in operating rooms over a period of one or several days. There are multiple operating rooms available, and surgeons may switch among them in order to meet equipment requirements or tighten the overall schedule. Each solution is a detailed schedule for operating rooms and surgeons that should lead to high performance even if surgeries last longer or shorter than expected. As in general multiobjective optimization, the goal is to present a diverse set of high-quality trade-off solutions to the decision maker within a reasonable amount of computation time.

We assess the quality of the implemented MOEAs by solving instances based on the situation in a real-life Norwegian hospital. Because optimal solutions to the investigated instances are not known, the study is a comparative one, where several well-known performance measures [7, 8] are used to compare Pareto-optimality, distribution, and spread of the SPEA2 and NSGA-II. As is common in surgery scheduling literature, we model duration uncertainty by assuming lognormal distributions for surgery durations. We use a relatively simple approach for evaluating fitness under uncertainty during evolution in the MOEAs [9], namely explicit averaging using Monte-Carlo simulation. Furthermore, we investigate the benefit of incorporating uncertainty during optimization.

We model the surgery scheduling problem mathematically and modify instances from previous work on the same problem [2, 4]. This chapter is intended to be a contribution to both surgery scheduling literature and general literature on population-based methods for complex problem solving. However, this chapter is not a case study of the mentioned Norwegian hospital. The hospital serves as a context and provider of data that enables the generation of realistic problem instances.

The rest of the chapter is organized as follows. Section 2 presents background related to hospital surgery scheduling process. This Section also contains a conceptual description of our implemented metaheuristic—genetic algorithm (GA) and highlights the necessary mechanisms for applying it to scheduling problems. In Sect. 3, we review updated literature on solution methods for solving varieties of surgery scheduling problems. Section 4 details the mathematical problem formulation. Section 5 describes the implementation of the algorithms. We describe the instances and analyze the results in Sect. 6. Section 7 concludes the chapter with recommendations for future works.

2 Background

This Section gives an introduction to the general surgery scheduling process. A description of uncertainty in surgery scheduling and a conceptual description of the problem under investigation is also provided in this section.

2.1 The Surgery Scheduling Process

The surgery scheduling process is described in various ways, but may be generalized as follow [3]. It starts when the surgery of a patient is acknowledged by a specific hospital and ends when the patient leaves the hospital after completed surgery. The set of detailed activities and resources needed to complete the surgery process varies among hospitals and patients within the same hospital. Typically, once the surgery is acknowledged, the patient is assigned a surgeon and put on the patient waiting list. The schedulers look for an appropriate surgery date and inform the patient once a date is set. Next, the room and exact starting time and duration for the surgery may

be decided. Decisions regarding day, operating room and time interval for surgery may be dependent on a *master surgery schedule* (MSS). The MSS reserves time in operating rooms for specific surgical subspecialties throughout a time period (e.g. one week). The use of MSS simplifies the coordination of resources needed to perform detailed activities on the day of surgery. Daily coordination activities include personnel rostering of surgeons, anesthesiologists and nurses, facility scheduling of hospital beds, operating rooms, post-anaesthesia care unit and intensive care unit, as well as coordination of equipment such as prostheses and surgical tools. A patient may be involved with some or all of these resources before the surgery is completed.

It is essential to know if patients are (i) elective or non-elective and (ii) inpatients or outpatients in order to deal with a surgery scheduling problem. Elective patients may be planned in advance because their condition will not worsen over time if surgery is not performed. On the other hand, non-elective patients require immediate or urgent care, and postponement of surgery may critically worsen the patient's condition. The arrival of non-elective emergency or urgency cases is unpredictable and requires the swift mobilization of necessary resources. When the same set of resources are allocated to the treatment of elective and non-elective patients, this mobilization often interferes with the elective schedule. Therefore, it is not uncommon to separate treatment of elective and non-elective patients. Alternatively, schedulers can add slack in the elective patient schedule and practice a rescheduling policy for fast and efficient re-coordination of resources. The distinction between inpatients and outpatients is made between patients that must be hospitalized and patients that arrive and leave at the day of surgery, respectively. When scheduling surgery for inpatients, other resources such as hospital bed capacity and nurses must be considered in addition to the resources needed to complete the surgery. Furthermore, outpatient surgeries are usually shorter and less variable than inpatient surgeries.

2.2 Uncertainty

Two sources of uncertainty are addressed in surgery scheduling literature: arrival uncertainty and duration uncertainty. Arrival uncertainty refers to the late arrival of staff or patients as well as the random flow of non-elective patients. Duration uncertainty refers to the fact that activities performed before, during and after surgery may take a longer or shorter time than expected. Especially surgery durations are critically variable. It is possible to estimate the duration of surgeries by analyzing data, and many papers focus solely on such estimation [10–12]. As it turns out, accurate prediction is difficult, partly because the scope of surgical procedures varies based on small physical differences among patients that cannot be observed prior to surgery. The lognormal distribution is often used to model how the duration of surgical procedures varies, although other distributions are also applicable, such as the normal distribution and the gamma distribution [13]. The lognormal distribution fits data where the natural logarithm of the mean and variance constitutes a normal distribution. Formally, if the random variable, Y, follows the normal distribution with

mean λ and standard deviation, δ, then the random variable, $X = e^Y$, is lognormally distributed with the following mean, μ, and variance, σ^2 [13]:

$$\begin{aligned} E[X] &= \mu = e^{\lambda + \frac{1}{2}\delta^2}, \\ V[X] &= \sigma^2 = \mu^2(e^{\delta^2} - 1) \end{aligned} \tag{1}$$

The probability density function, providing the probability of X taking any particular value x, is usually stated as follows:

$$p(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{[ln(x)-ln(\mu)]}{2\sigma^2}} \tag{2}$$

Surgery duration uncertainty is a prominent difficulty because schedules that perform excellent, on expectation may be inadequate if surgeries are delayed or finish early. Delay early in the morning propagates throughout the day and may lead to overtime and cancelled surgeries. On the other hand, surgeries that finish early leave expensive resources idle until the next patient is ready for surgery.

2.3 *Conceptual Description of the Problem Under Investigation*

Our investigated variation of the surgery admission planning problem considers elective patients only, without any arrival uncertainty associated with non-elective patients or patient no-shows. A detailed mathematical formulation is given in Sect. 4. We do not differentiate between outpatients and inpatients explicitly, as resources outside the operating room are not part of our problem. The performance measures are the minimization of patient waiting time on the patient waiting list and maximization of the operating room and surgeon utilization. The investigated problem can be well described as a variation of the surgery admission planning problem where both operating rooms and surgeons are constraining resources. The considered source of uncertainty is surgery durations, and we concentrate within the category of heuristic combinatorial optimization. Figure 1 illustrates performance measures (except patient waiting time) and the difference between planned duration and actual duration for a daily problem. The same surgeon performs surgery C and B, and the arrow drawn between the surgeries indicates that some surgeon idle time is incurred between surgery C and B. Surgery A lasts shorter than planned and incurs some operating room idle time between surgery A and B in the operating room–1. Surgery B can start earlier than planned, and some unused operating room time is left at the end of the day. Surgery D incurs surgeon overtime and operating room overtime in the operating room–2. The figure does not show that a set of surgeries are excluded from the schedule and that there is a waiting cost associated with postponing these surgeries.

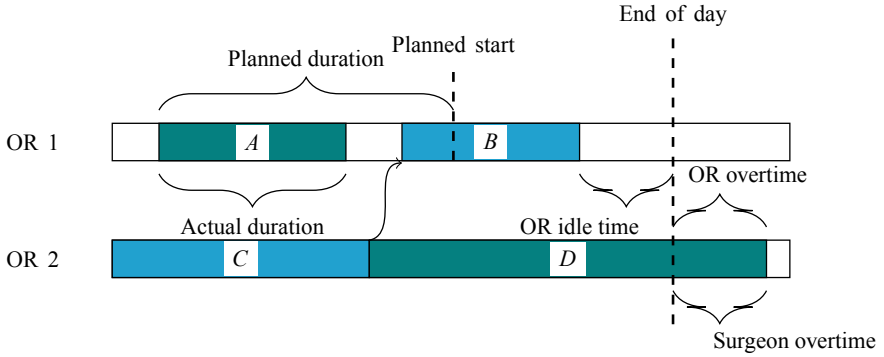


Fig. 1 The daily problem

2.4 Genetic Algorithms (GAs)

A popular metaheuristic for solving scheduling problems is the GA. It mimics mechanisms that drive evolution in nature such as mutation, selection, competition and inheritance. GA works on an evolving population of solutions (*individuals*/*chromosomes*) where several genetic/evolutionary sub-mechanisms (such as selection, crossover and mutation) are simulated over a set of generations until high-quality solutions emerge. Chromosomes possess genes that they pass on to their descendants (*offspring*). Each chromosome is characterized by a set of parameters (variables) called *genes* which carry information to form a solution. Chromosomes must be decoded from coding space (*genotype*), in which crossover, mutation (local search also) occurs, to solution space (*phenotype*) in order to perform evaluation and selection. In general, a representation may be characterized by how it enables a mapping from coding space to solution space [4, 14]. Some representations guarantee feasibility, i.e. that the solutions produced by decoding are feasible concerning constraints of the original problem. Other representations cannot guarantee feasibility, but they can guarantee the legality, i.e. that the produced solution is a complete solution to the original problem. The representation of solutions in coding space, mutation and crossover operations are problem specific. In the following, we briefly discuss how a GA can be employed in a scheduling problem.

2.4.1 GAs for Solving Scheduling Problems

In [14], nine different representations, classified as either direct or indirect, are described for the JSP. With a direct approach (e.g. operation-based, job-based, job pair relation-based, completion-based, random keys), chromosomes represent actual schedules and explicitly state the schedule design. On the other hand, with an indirect approach (e.g. preference list-based, priority rule-based, disjunctive graph-based,

machine-based), chromosomes represent a set of rules or prioritizations and some simple or complex heuristic is needed to decode the chromosome into an actual schedule. These decoders may be classified according to properties of the schedules they produce [15, 16]. A schedule may be categorized as semi-active or active [17]. In a semi-active schedule, all tasks are scheduled as early as possible in a manner that adheres to the coded sequence. In an active schedule, the predefined sequence may be altered in order to fill gaps in the schedule. Once a representation is chosen, appropriate mutation operators and crossover operators must be defined. Examples are found in [14]. The GA may then be applied to the scheduling problem by evolving a population of solutions until a stopping criterion, such as the maximum number of generations, is met.

2.4.2 Hybrid Genetic Algorithm (HGA)

When the search space is complex, the convergence of GAs may be slow, and the genetic operators are more concerned with adequate exploration than efficient exploitation of promising parts of the solution space. A popular strategy for balancing exploration and exploitation of the solution space is to hybridize GAs by performing some local search during the evolutionary process. Faster convergence to good solutions may be achieved by improving individuals in the population using a local search algorithm. In the HGA framework, the GA mechanisms distribute individuals across the solution space, while the local search procedure is used for in-depth investigation of promising solutions. For example, in [18], the advance scheduling problem is solved with a column generation based heuristic, and the allocation scheduling problem is solved with a hybrid GA. Interested readers can find more on the application of HGA on scheduling problems in [16, 19, 20].

3 Literature Review

As surgery scheduling problems are closely related to machine scheduling problems, researchers often acknowledge that realistically sized surgery scheduling problems cannot be solved exactly in a computationally tractable manner. In [21], the authors show that their daily stochastic surgery scheduling problem with multiple operating rooms is NP-hard by reducing the bin-packing problem to a variation of their problem. The advance scheduling problem and the allocation scheduling problem are both argued to be NP-hard in [22]. The combination of the two problems is even harder [3]. Therefore, the problem is typically either decomposed or solved heuristically.

The GA has attracted researchers working on scheduling both from the academia and industry due to its potential for solving real-world scheduling problems which usually involve large and unknown search space. To exemplify, the GA is the basis for solving (i) the job shop scheduling problem (JSP) in [23], (ii) the flexible job shop scheduling problem (fJSP) in [24], (iii) the flow shop scheduling problem (FSP) in

[25], (*iv*) the permutation flow shop scheduling problem (pFSP) in [26] and (*v*) the open shop scheduling problem (OSP) in [27].

Simple rules for scheduling surgeries under duration uncertainty are compared against the NSGA-II in [28]. After analyzing data to fit lognormal distributions to surgery durations, several heuristics are used to determine the sequence of surgeries at a day in an outpatient clinic. The bicriteria heuristic used for solving the surgery admission planning problem in [29] is composed of a constructive phase and an improvement phase. This approach requires very little computation, and runtime is negligible. It performs quite well, however, under the assumption of deterministic surgery durations.

In [30], two constructive heuristics for two variations of the surgical case assignment problem are combined with an improvement heuristic based on iterative neighbourhood search. A variable neighbourhood decent (VND)–based local search heuristic is developed in [3] for the surgery admission planning problem. The VND is also the basis for solving the surgery tactical planning problem [31] as well as several basic types of scheduling problems [32]. Adaptions of the tabu search, simulated annealing and multi-start method have also been implemented to solve variations of the advance surgery scheduling problem [33]. However, none of the above approaches is implemented to handle multiple objectives properly.

Several existing works consider the surgery scheduling problem in a multiobjective context. While most works use the simple weighted-sum approach [3, 21, 34, 35], several recent examples of more advanced approaches exist, such as multiobjective rule-based heuristics [28, 29] and population-based multiobjective metaheuristics (NSGA-II [28, 36], a hybrid MOEA/D [37] and a modified ant colony optimization (ACO) [38]). In [38], the load balance equilibrium objective is incorporated into an ACO implementation. Three different settings of ACO are compared for solving a daily multiobjective operating room scheduling problem. The author concludes that the hybrid Pareto-set ACO with multiple objectives is superior to the other algorithms and that it finds solutions with equal makespan as the single objective ACO focusing *only* on makespan but with better results on the other objectives. However, none of the mentioned approaches to multiobjective surgery scheduling considers uncertain surgery durations.

A majority of the reviewed papers on surgery scheduling under duration uncertainty assume that surgery durations follow the lognormal distribution. A heuristic approach to a two-stage stochastic programming model where rescheduling is done as a response to variances in surgery durations and arrival of emergency patients is developed in [39]. The heuristics are tailored to the advance scheduling problem to create feasible initial schedules, evaluate recourse strategies and improve initial schedules with a rolling horizon algorithm. For a similar problem, the sample average approximation is used to solve a two-stage stochastic program with recourse in [35]. In [40], a two-stage stochastic recourse model is solved to optimality, and it is shown that a simple heuristic where surgeries are sequenced according to increasing variance performs quite well. A two-stage, mixed-integer stochastic dynamic programming model with recourse is implemented for a detailed daily surgery scheduling in [41]. In [42], a sample average approximation approach to the advance schedul-

ing problem under duration uncertainty is compared to a deterministic model using expected values for surgery durations. Simulation results show a substantial expected cost reduction of incorporating uncertainty. An exact and a heuristic approach to a detailed daily surgery scheduling problem under duration uncertainty are implemented in [43]. The exact problem can be solved to optimality, but for the stochastic case, a decomposition approach using a constructive and an improvement heuristic is superior in terms of speed and solution quality. In [28], the authors model surgery durations with lognormal distributions and with different percentiles (i.e. different hedging levels) during optimization with NSGA-II. The experimental results suggest that using the 65th percentile during optimization yielded the best results.

4 Problem Formulation

In this Section, we mention our assumptions. We also mathematically present the multiobjective surgery admission planning model formulation.

4.1 Model Assumptions

To model the surgery admission planning problem in a comprehensible, implementable and computationally tractable manner, some assumptions must be made:

(i) the patient waiting list is static with every patient having a pre-assigned surgeon,
(ii) every patient is available at any day and time in the planning period,
(iii) all surgeons are available at any day and time in the planning period,
(iv) the considered resources are reserved for elective patients only,
(v) there is no explicit difference between inpatients and outpatients,
(vi) all patients have been categorized according to surgical subspecialties,
(vii) all surgical subspecialties are covered by the MSS,
(viii) necessary facilities and equipment are available in accordance with the MSS,
(ix) patients, equipment and staff always arrive on time,
(x) surgeons and operating rooms are the only considered resources (equipment and other staff are disregarded),
(xi) preoperative and postoperative capacity is not considered to limit the flow of patients through the surgery delivery system,
(xii) surgeons may switch between operating rooms without any turnover time,
(xiii) all scheduled surgeries must be completed without interruption,
(xiv) surgery durations follow procedure-dependent lognormal distributions and are independent of each other and all other decisions, such as scheduled sequence and allocated operating room,
(xv) there are no rescheduling activities during schedule execution.

We regard some of these assumptions as more critical than the others. For instance, rescheduling activities, such as cancellations or remobilization of staff, can take place quite frequently during a typical week of surgery. Still, we make this assumption because it significantly simplifies our model. Because no rescheduling activities are allowed, the execution of a schedule can be simulated as a single event where all durations are observed simultaneously. Surgeries are simply postponed or brought forward in time as realized surgery duration are observed. Also, by disregarding resources needed before and after surgery, solutions to our model may cause overload outside the operating room.

A variant of the classical healthcare management assignment problem is presented in [44]. The authors developed a cyclic schedule representation (the schedule is repeated every n days) with minimizing congestion in the recovery units as the primary optimization objective. This is to minimize the maximum number of patients in the recovery room so that the cost associated with related hospital resources (staff, facilities, and equipment) is minimized. However, it is broadly acknowledged that the operating room is the bottleneck in most surgery delivering system [1, 2]. Then again, the representation represented in [44] does not consider this primary objective as a priority optimization target. Though the authors reported satisfactory results in [44], it is difficult to handle the uncertainty presents in any hospital with their proposed representation. It is mainly due to the cyclic representation which repeats every n days, and all real-world hospital environments have uncertainty all around in consecutive surgery procedures. Several days seems to be a long horizon, even certainty among consecutive surgical procedures for a single day is a rare situation.

4.2 Mathematical Problem Formulation

Surgeries $i \in N$ from a patient waiting list are to be scheduled over a planning period of $d \in D$ days. Each day there are $r \in R$ available operating rooms that are reserved according to MSS. Every surgery belongs to a certain subspecialty $s \in S$ ($\sigma_{is} = 1$, if surgery i is of subspecialty s; $\sigma_{ik} = 0$, otherwise); and a qualified surgeon $k \in K$ is already assigned to each surgery ($\delta_{ik} = 1$, if surgery i is pre-assigned to surgeon k; $\delta_{ik} = 0$, otherwise). The goal is to select a set of surgeries from the patient waiting list and schedule them in rooms during the planning period so as to minimize (i) patient waiting time on the waiting list, (ii) operating room overtime, (iii) surgeon overtime, (iv) operating room idle time and (v) surgeon idle time. Formally, the decision is whether a surgery $i \in N$ should be scheduled in a room $r \in R$ on a day $d \in D$, denoted as a binary decision variable:

$$x_{idr} = \begin{cases} 1 & \text{if surgery } i \text{ is scheduled on day } d \text{ in room } r \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

To simplify some expressions, we also introduce a binary helping variable, z_{idk}, to denote whether a surgery i is performed by surgeon k on a given day, d:

$$z_{idk} = \delta_{ik} \sum_{r=1}^{R} x_{idr} = \begin{cases} 1 & \text{if surgery } i \text{ is performed by surgeon } k \text{ on day } d \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

These decisions must adhere to the MSS ($M_{drs} = 1$, if the MSS allows operating room r to be scheduled by specialty s on day d; $M_{drs} = 0$, otherwise). This is ensured by adding the following constraint:

$$x_{idr} \leq M_{drs} \tag{5}$$

At the allocation scheduling level, the decision is the sequence of surgeries on a given day (binary decision variables), and the starting time of each surgery (continuous decision variables):

$$y_{ij} = \begin{cases} 1 & \text{if surgery } i \text{ precedes surgery } j \text{ on a given day} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$t_i \geq 0 : \text{ scheduled starting time for surgery } i \tag{7}$$

The y_{ij} variable defines precedence relations among surgeries. Note that immediate precedence of two surgeries is not necessary to maintain a precedence relation between them. It is only required to ensure following constraints:

$$y_{ij} + y_{ji} \leq 1 \tag{8}$$

$$y_{ij} + y_{ji} \geq x_{idr} + x_{jdr} - 1 \tag{9}$$

$$y_{ij} + y_{ji} \geq z_{idk} + z_{jdk} - 1 \tag{10}$$

The first constraint, Eq. 8, is a basic precedence constraint stating that if surgery i precedes surgery j, then surgery j cannot precede surgery i. Equation 9 states that if two surgeries are scheduled in the same operating room on the same day, there must be a precedence relation between them. Equation 10 states that there must be a precedence relation between surgeries preformed by the same surgeon on the same day.

A determination of the advance scheduling decision variables (x_{idr}) and the precedence relation variables (y_{ij}) helps computing scheduled staring times (t_i) using a duration estimate ($\mathscr{D}_i$) for each surgery $i \in N$. The scheduled starting times and estimated ending times must adhere to normal opening hours of operating rooms (from starting time E^b_{dr} to ending time E^e_{dr}) and surgeons (from starting time E^b_{dk} to ending time E^e_{dk}). A certain amount of overtime, V, is allowed for surgeons and operating rooms. Expected starting times are computed with following constraints:

$$t_i \geq t_j + \mathscr{D}_j - M(1 - y_{ji}) \tag{11}$$

$$t_i \geq max \left\{ \sum_{r=1}^{R} x_{idr} E^b_{dr}, \sum_{k=1}^{K} z_{idk} E^b_{dk} \right\} \tag{12}$$

$$t_i + \mathscr{D}_i \leq max \left\{ \sum_{r=1}^{R} x_{idr} E^e_{dr}, \sum_{k=1}^{K} z_{idk} E^e_{dk} \right\} + V \tag{13}$$

Equation 11 computes the starting times. A *big M method*[1] ensures that only preceding surgeries of a surgery i take part in the computation. Equations 12 and 13 sets lower daily limits for starting times and upper daily limits for ending times, respectively. Once the starting times are computed, the starting times and ending times for surgeons may be set. Our model does not count surgeons as idle unless they are waiting in between surgeries within a daily interval of surgeries. The scheduled starting time (e^b_{dk}) and ending time (e^e_{dk}) for a surgeon coincide with the earliest and latest surgery assigned to that surgeon on a given day, respectively. These times are actually decisions in the model that may be derived directly from the timing of the surgeries:

$$e^b_{dk} = \min_{i \in N} z_{idk} t_i \tag{14}$$

$$e^e_{dk} = \max_{i \in N} z_{idk} (t_i + \mathscr{D}_i) \tag{15}$$

The waiting time objective, O_W, may be computed before schedule execution. The waiting time cost (Eq. 16) is based on the referral date (H_i) and the deadline (G_i) of surgery i in relation to the scheduled date, d, for the surgery:

$$O_W = \frac{1}{N} \sum_{i=1}^{N} \sum_{d=1}^{D} \sum_{r=1}^{R} x_{idr} \left(\frac{d - H_i}{G_i - H_i} \right)^2 \tag{16}$$

For surgeries that are not scheduled, we assume that they may be completed within the next scheduling period and assign the latest date in the next period, $d = 2|D|$, to compute their contribution to the waiting cost component.

The other objectives, (i) surgeon idle time (O^K_I), (ii) surgeon overtime (O^K_O), (iii) operating room idle time (O^R_I) and (iv) operating room overtime (O^R_O), may be computed once the realized durations, $\mathscr{D}_i(\omega)$, are observed after schedule execution. Here, ω, denotes the collective outcome of all surgery durations and allows for computation of realized starting times, $t_i(\omega)$:

$$t_i(\omega) \geq t_j(\omega) + \mathscr{D}_j(\omega) - M(1 - y_{ji}) \tag{17}$$

[1]The big M method [45] is used to keep linearity in models with binary decision variables. For large M, constraints may be turned on or off depending on the value of the decision variables.

$$t_i(\omega) \geq max\left\{\sum_{r=1}^{R} x_{idr} E^b_{dr}, \sum_{k=1}^{K} z_{idk} E^b_{dk}\right\} \tag{18}$$

Note that there is no restriction on the ending time of surgeries corresponding to Eq. 13. In other words, all surgeries must be completed, no matter how much overtime is incurred. In addition to Eqs. 17 and 18, we restrict how much earlier than planned surgeries may start. The following constraint imposes that no surgery can start T minutes earlier than scheduled:

$$t_i(\omega) \geq t_i - T \tag{19}$$

Once the realized starting times ($t_i(\omega)$) and (implicitly) completion times ($t_i(\omega) + \mathscr{D}_i(\omega)$) have been computed, the other objectives are calculated as follows.

$$O^R_O = \sum_{d=1}^{D}\sum_{r=1}^{R} \max_{i\in N}\left(max\left\{t_i(\omega)x_{idr} - E^e_{dr}, 0\right\}\right) \tag{20}$$

$$O^K_O = \sum_{d=1}^{D}\sum_{k=1}^{K} \max_{i\in N}\left(max\left\{t_i(\omega)z_{idk} - E^e_{dk}, 0\right\}\right) \tag{21}$$

$$O^R_I = \sum_{d=1}^{D}\sum_{r=1}^{R}\left[(E^e_{dr} - E^b_{dr}) - \sum_{i=1}^{N}\mathscr{D}_i(\omega)x_{idr} + \max_{i\in N}\left(max\left\{t_i(\omega)x_{idr} - E^e_{dr}, 0\right\}\right)\right] \tag{22}$$

$$O^K_I = \sum_{d=1}^{D}\sum_{k=1}^{K}\left[(e^e_{dk} - e^b_{dk}) - \sum_{i=1}^{N}\mathscr{D}_i(\omega)z_{idk} + \max_{i\in N}\left(max\left\{t_i(\omega)z_{idk} - E^e_{dr}, 0\right\}\right)\right] \tag{23}$$

The overall objective is to simultaneously minimize patient waiting time (O_W) and resource efficiency (O^R_O, O^K_O, O^R_I, O^K_I) subject to all surgery duration realizations, $\omega \in \Omega$. The objective can be formulated as follows where $\xi(\omega)$ denotes the random vector of surgery durations:

$$\begin{aligned}&\min \quad \theta(x, y, t) = O_W,\\ &E_\xi\left[O^R_O(x, y, t, \xi(\omega)), O^K_O(x, y, t, \xi(\omega)), O^R_I(x, y, t, \xi(\omega)), O^K_I(x, y, t, \xi(\omega))\right]\end{aligned} \tag{24}$$

Figure 2 illustrates some of the model notation, where colors represent two surgeons (the green surgeon performs surgeries A and D, and the blue surgeon performs surgeries C and B). Here, we have the following precedence relations: $y_{AB} = 1$, $y_{CD} = 1$ (operating room precedence) and $y_{AD} = 1$, $y_{CB} = 1$ (surgeon precedence). Figure 2 shows only one of possibly many days, $d \in D$, in the planning period. Note that operating room overtime, O^R_O, and surgeon overtime, O^K_O, often

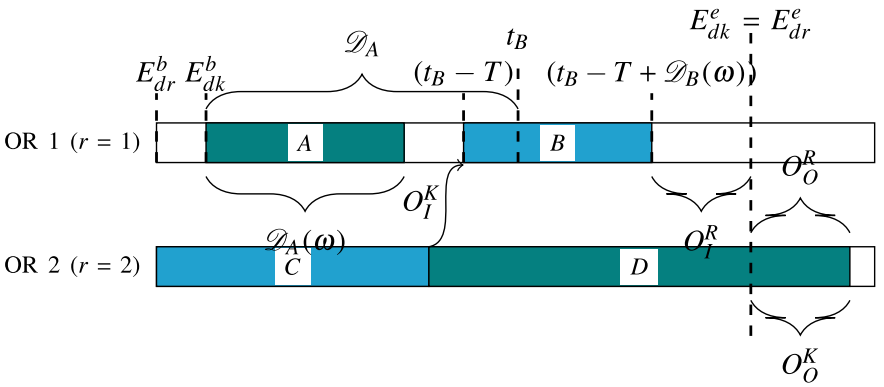


Fig. 2 Model notation

will coincide if normal surgeon ending times equal operating room ending times (as in Fig. 2; $E^e_{dk} = E^e_{dr}$).

5 Implementation

This Section presents the detailed description of our implemented algorithms: SPEA2, NSGA-II and a single objective hybrid GA. In order to efficiently solve the problem described in Sect. 4, we take inspiration from machine scheduling literature [15], observing that the operating rooms are analogous to machines, surgeons are analogous to jobs and that surgeries are analogous to operations.

5.1 Evolutionary Submechanisms

In the following, we suggest new evolutionary mechanisms for applying to MOEAs to the surgery admission planning problem under uncertainty. All the implemented algorithms are based on the similar sub-mechanisms mentioned in this sub-section.

5.1.1 Solution Representation

Using machine scheduling terminology, the surgery scheduling problem under investigation may be described as a multiobjective stochastic fJSP with non-identical machines. The selected representation must be translated into a solution of the problem defined in Sect. 4, i.e. enable a determination of the room-day allocations (x_{idr}), the precedence relations among them (y_{ij}) and the timing of surgeries (t_i). We let

Fig. 3 Two-vector chromosome representation

locus	1	2	3	4	5	6
v_1	(1,1)	(2,2)	(1,2)	(2,1)	(2,1)	(1,1)
v_2	1	4	2	3	5	6

the chromosomes dictate the combinatorial part of the problem (the binary variables: x_{idr} and y_{ij}) and use left shifting with the duration estimates ($\mathscr{D}_i$) to create active schedules and determine the continuous timing variables, t_i, during decoding.

Our selected representation is inspired by the two-vector representation used for fJSP in [46], where one vector represents machine assignments, and the other represents the operation sequence. Two critical modifications must be made. First, the machine assignment vector must consist of tuples representing combinations of rooms and days. This is also necessary to control the feasibility of chromosomes. Second, since there is no predetermined sequence among the surgeries, our scheduling problem is open, and the operation-based representation used in the sequence vector in [46] cannot consistently represent all possible sequences among operations. We, therefore, use a permutation-based representation in the sequence vector that dictates the sequence in which surgeries are fitted into the schedule during decoding. We use chromosome initialization, mutation and crossover operators that in combination with the decoding procedure guarantees feasibility. Figure 3 shows an example of a chromosome, where the room-day vector (v_1) indicates the room and day chosen for the surgery at that locus. The precedence vector (v_2) indicates the order in which the surgeries are put into the schedule.

5.1.2 Decoding

Chromosomes are translated into feasible schedules using a new variation of priority-based decoding and reordering. The decoding procedure must handle the continuous time model and ensure that operating room schedules and surgeon schedules are compatible. For every surgery, i, encountered in the chromosome from left to right in the precedence vector (v_2), the operating room (r) and day (d) for the surgery are found at the ith position in the room-day vector (v_1). The assigned surgeon (k) and an estimate for the surgery duration ($\mathscr{D}$) are found in provided surgery data. During decoding, both surgeon and operating room schedules are built from left to right while maintaining an overview of the surgeon and operating room idle time intervals. If the surgery, i, cannot be scheduled within any idle time according to its duration ($\mathscr{D}_i$), it is scheduled at the end of the day. In this way, the schedule is always active as scheduling surgery in the earliest possible idle time interval represents a global left shift. Considering the schedule under construction (Fig. 4), surgeries assigned to surgeon B are scheduled in operating room 3, 2 and 1, incurring operating room idle time at the start of the day in operating room 1 and 2. Then, a series of surgery inclusions require different kinds of modifications to the maintained set of the surgeon and operating room idle time intervals. We describe these modifications

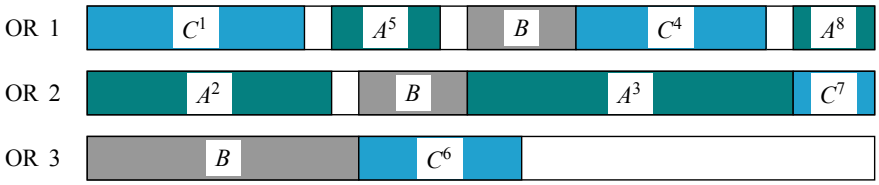


Fig. 4 Idle time intervals

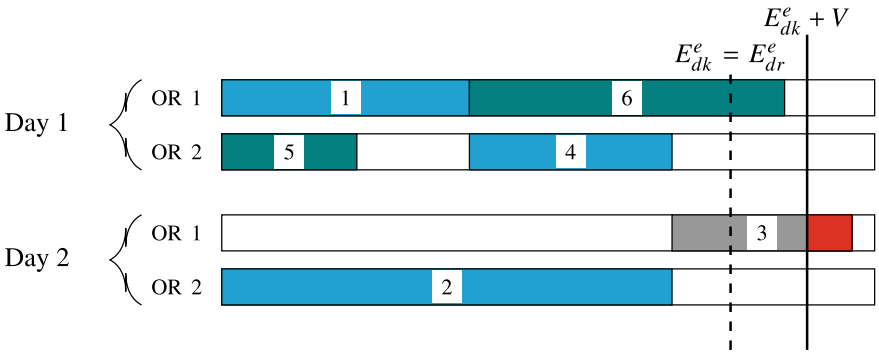


Fig. 5 The chromosome in Fig. 3 after decoding

in the following where the superscripts indicate the sequence in which surgeries are included in the schedule:

(i) including surgery C^1 and A^2 reduces the operating room idle time interval at the start of the day in operating room 1 and 2, respectively,
(ii) including A^3 and C^4 creates two surgeon idle time intervals between surgery A^2 and A^3; and C^1 and C^4, respectively,
(iii) including A^5 splits the operating room idle time interval in operating room 1 between surgery C^1 and B in two. Additionally, the surgeon idle time interval between A^2 and A^5 is shortened to be from A^5 to A^3,
(iv) including C^6 splits the surgeon idle time interval between surgery C^1 and C^4 in two,
(v) including surgery C^7 creates a new surgeon idle time interval between surgery C^4 and surgery C^7,
(vi) including surgery A^8 creates a new operating room idle time interval between surgery C^4 and A^8.

All of the exemplified scenarios must be taken care of during decoding to ensure that the idle time intervals are updated. In Fig. 5, we illustrate the decoded chromosome from Fig. 3. The surgeries are scheduled in the order specified by the precedence vector, v_2, in Fig. 3: $\{1, 4, 2, 3, 5, 6\}$:

(i) surgery 1 is scheduled in operating room 1 on day 1,
(ii) surgery 4 is scheduled in operating room 2 on day 1,

(iii) surgery 2 is scheduled in operating room 2 on day 2,
(iv) surgery 3 (performed by the same surgeon as surgery 2) is *not included* in the schedule, because it would incur overtime beyond the allowed limit of $E^e_{dk} + V$ or $E^e_{dr} + V$,
(v) surgery 5 is scheduled in the idle time interval in operating room 2 on day 1 (this is an example of a global left shift) and
(vi) surgery 6 is scheduled in operating room 1 on day 1, incurring an allowed amount of planned overtime.

5.1.3 Fitness Evaluation

After decoding into a feasible schedule, fitness evaluation is done according to

(i) Equation 16 (patient waiting time),
(ii) Equation 20 (operating room overtime),
(iii) Equation 21 (surgeon overtime),
(iv) Equation 22 (operating room idle time), and
(v) Equation 23 (surgeon idle time).

The patient waiting time objective (O_W) is independent of the outcome of surgery durations (ω), and may be deterministically computed without any further computation. Taking the example in Fig. 5 with the following referral dates–dead lines (H_i–G_i) for surgeries $\{1, 2, \ldots, 6\}$:
$\{(3, 1), (10, 1), (4, 2), (20, 3), (6, 1)\}$,
we get the following patient waiting time cost contribution from each surgery (according to the formula $\left(\frac{d-H_i}{G_i-H_i}\right)^2$):

(i) surgery 1, scheduled at day 1: $O_W \mathrel{+}= (\frac{1-(-3)}{1-(-3)})^2 = 0 + 1 = 1$,
(ii) surgery 2, scheduled at day 2: $O_W \mathrel{+}= (\frac{2-(-10)}{1-(-10)})^2 = 1 + 1.190 = 2.190$,
(iii) surgery 3, *not scheduled*[2]: $O_W \mathrel{+}= (\frac{4-(-4)}{2-(-4)})^2 = 2.190 + 1.778 = 3.968$, and so forth.

The computation of operating room overtime (O^R_O), surgeon overtime (O^K_O), operating room idle time (O^R_I) and surgeon idle time (O^K_I) requires determination of the starting time decision variables, $t_i(\omega)$. For a single collective outcome of surgery durations (ω), the computation is straightforward. Because we do not allow any rescheduling activities, the surgery sequence and allocations remain constant over all possible realizations of ω. Figure 6 shows an example of a realized schedule. The corresponding effective fitness function can thus be stated for the case where the environmental variables are uncertain:

[2] Assumes the last day in the next period, $d = 2|D| = 4$.

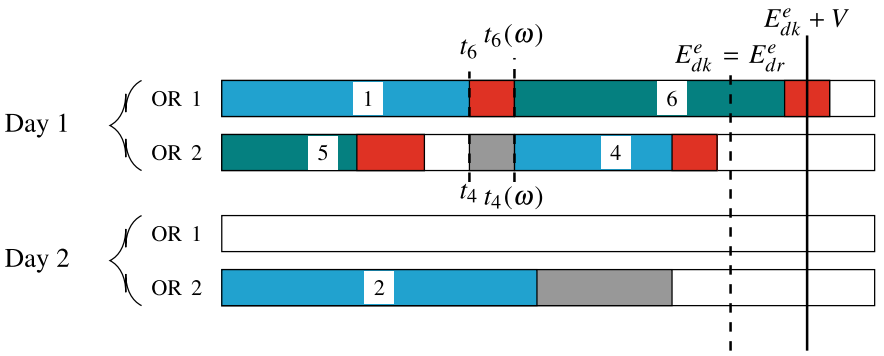


Fig. 6 Realized schedule

$$
\begin{aligned}
F(x, y, t) = \big\{ & O_W(x, y, t), \\
& \int_{-\infty}^{\infty} O_O^R(x, y, t, \omega) p(\omega) d\omega, \\
& \int_{-\infty}^{\infty} O_O^K(x, y, t, \omega) p(\omega) d\omega, \\
& \int_{-\infty}^{\infty} O_I^R(x, y, t, \omega) p(\omega) d\omega, \\
& \int_{-\infty}^{\infty} O_I^K(x, y, t, \omega) p(\omega) d\omega \big\}
\end{aligned} \tag{25}
$$

Since the effective fitness function does not have a closed form, it cannot be computed exactly. The obvious estimator for the effective fitness function is the Monte Carlo approximation:

$$
\begin{aligned}
F(x, y, t) = \big\{ & O_W(x, y, t), \\
& \frac{1}{N} \sum_{i=1}^{N} O_O^R(x, y, t, \omega_i), \\
& \frac{1}{N} \sum_{i=1}^{N} O_O^K(x, y, t, \omega_i), \\
& \frac{1}{N} \sum_{i=1}^{N} O_I^R(x, y, t, \omega_i), \\
& \frac{1}{N} \sum_{i=1}^{N} O_I^K(x, y, t, \omega_i) \big\}
\end{aligned} \tag{26}
$$

Using Monte Carlo integration to estimate the fitness of an individual in robust MOEAs is referred to as explicit averaging. The accuracy of the estimation increases

with the sample size, N. However, increasing N also increases the number of fitness evaluations. In our case, decoding is needed only once, but the simulation and calculation of O_O^R, O_O^K, O_I^R and O_O^K must be done N times.

5.1.4 Crossover

We implement a version of the enhanced order crossover operator [46] to produce a feasible child chromosome, c, from two selected parent chromosomes, p_1 and p_2. The crossover is executed in three steps. First, a subsection of the genes belonging to the first parent, p_1, is directly inherited by the child, c_1. Then, missing room-day allocations in the child, c_1, are filled with values from v_1 of the second parent (p_2). Finally, the integers missing in v_2, are filled in from the second parent (p_2) starting from the left-most locus in v_2. The child chromosome will always decode into a feasible schedule because the room-day allocations are inherited directly (the locus is unchanged) from the parents which have room-day allocations that adhere to the MSS. The crossover operation is illustrated for coding space in Fig. 7.

5.1.5 Mutation

We implement three mutation operators, and all produce feasible child: (i) room-day swap mutation, (ii) precedence swap mutation, and (iii) renew room-day mutation. The room-day swap mutation interchanges the v_1 of two randomly chosen surgeries that are categorized by the same surgical subspecialty. The precedence swap mutation interchanges two random precedence values in v_2, thus altering the sequence in surgeries to be scheduled. The renew room-day mutation assigns a new feasible room-day combination to a randomly selected surgery in v_1. A single mutation can have a significant impact that is seen only after the chromosome is decoded. Figure 8 shows the swap mutations, in general (room-day mutation is to the left and precedence mutation is to the right). In this example, it is necessary that the surgery at locus 0 and locus 3 belong to the same surgical subspecialty—they can both be scheduled in OR 1 or OR 2 on day 1.

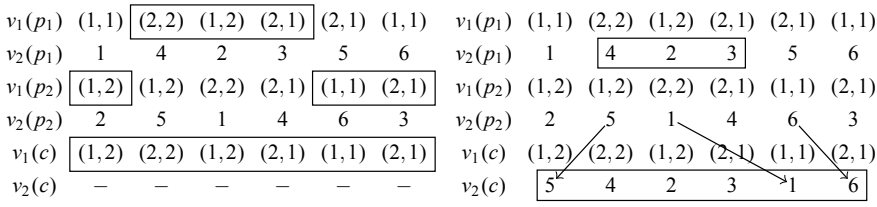


Fig. 7 Crossover of two parent chromosomes (p_1 and p_2) into a feasible child chromosome (c)

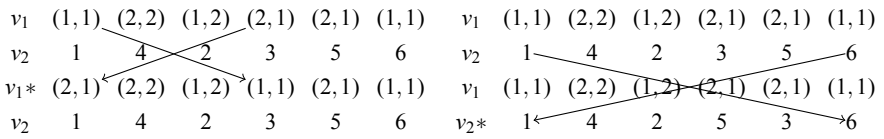


Fig. 8 Swap mutation operators with mutated vectors marked as (*)

5.2 Hybrid GA for Single Objective Optimization

The hybrid GA follows the standard procedure of simple GAs but invokes the *variable neighbourhood decent* (VND) local search procedure on every individual in a newly created generation. The motivation behind the implementation of a single objective GA is to locate the minimal possible values for each objective $m \in M$. Because these values are needed only to measure the performance of the SPEA2 and NSGA-II, computational speed is not a major concern. The VND uses three neighbourhoods based on the mutation operators presented in Sect. 5.1.5. The first neighbourhood, $\mathcal{N}_0(x)$, is defined as the set of solutions obtainable from a solution, x, by swapping the room-days of any two surgeries of the same subspecialty in the chromosome of x. The second neighbourhood, $\mathcal{N}_1(x)$, consists of all solutions like x where two integers in the precedence vector of x have been swapped. The third neighbourhood, $\mathcal{N}_2(x)$, include all solutions where a single surgery in x is assigned another feasible room-day. The procedure starts from an initial solution, x, and investigate all neighbourhoods, $\mathcal{N}_k \in \mathcal{N}$, as long as improvement is found in any of the neighbourhoods. The solution, x, is updated to the new best-found solution in a neighbourhood so that this newly found solution is the basis for generating and searching new neighbourhoods. The entire neighbourhood is searched so that the best neighbouring solution is selected as the new starting point (steepest descent). We preserve the best-known individual and copy it directly from one generation to the next (elitism).

6 Results and Discussion

This Section starts by describing the instances we experiment with and the experimental setup. Next, we perform an initial analysis to justify the use the conflicting objectives and the incorporation of uncertainty during surgery scheduling optimization. Then, we compare the two MOEAs using a selection of performance measures that do not require the Pareto-front. We turn to one of the most important contributions of this work—the effects of uncertainty during fitness evaluation of MOEAs in surgery scheduling. We then compare solution sets provided by deterministic and stochastic versions of the SPEA2 and NSGA-II.

6.1 Test Data

We create problem instances based on the surgery scheduling situation in a real-life Norwegian hospital which are modified from [2, 4]. Expected durations are assigned to surgeries according to a guideline in current use at the hospital mentioning how much time to allocate in the operating rooms for every surgical procedure. We ensure that the distribution of patients on the patient waiting list matches, to a reasonable extent, the capacity reserved in the master surgery schedule. Although the data used is not based on actual patient waiting lists and actual durations, we believe that the approximation is realistic enough for our analysis to be generalized to actual surgery scheduling settings. In summary:

Surgery scheduling instances. Our instances range in size from 2 rooms × 5 days with 34 patients on the waiting list to 8 rooms × 5 days with 255 patients on the waiting list.

Lognromal distributions. Surgery durations are modelled with procedure dependent lognormal distributions where the mean is based on estimations used by practitioners, and the coefficient of variance is set to $c_v = 0.1$.

6.1.1 Parameters

We use patient lists (N) ranging from 34 to 255 patients to be scheduled in either 2 operating rooms over 5 days, 7 operating rooms over 3 days or 8 operating rooms over 5 days. For all instances, we set the same parameters for opening hours and allowed scheduled overtime so that $E^b_{dr} = E^b_{dk} = 0$, $E^e_{dr} = E^e_{dk} = 8$ h and $V = 2$ h. The considered hospital uses a weekly MSS to schedule surgeries of 7 different subspecialties (Local, Reconstructive, Plastic, Hand, Arthroscopy, Back and Prosthesis) into 8 operating rooms. Figure 9 shows the MSS. The instances we experiment with consider subsections of the MSS (indicated by colors and brackets), with the largest instance class considering the entire schedule. The number of surgeries and surgeons is set to match the number of operating rooms and days involved in each instance.

The calibration of hyperparameters is partly based on experiences from [2, 4], partly on configurations found in literature [5, 28] and partly on preliminary experimentation with the developed algorithms. For SPEA2, population size and archive size is 500, $m = 5$ and the k-value is rounded to 32. For NSGA-II, population size is 500, tournament selection with a probability of 90% and mutation rate is 20%. The stop criterion is 250 generations for both MOEAs. We have an unconventional interpretation of the mutation rate that indicates the probability of an *individual*, and not a single gene in a chromosome, being selected for mutation. Motivated by a common interpretation of the *central limit theorem* [47], we use 30 Monte-Carlo simulation runs to estimate the fitness of an individual. We use the 65th percentile to decide scheduled starting times for surgeries during chromosome decoding. In our baseline setting, we use $T = \infty$, to infer the assumption that all resources are available, even in cases where surgeries are moved forward in time.

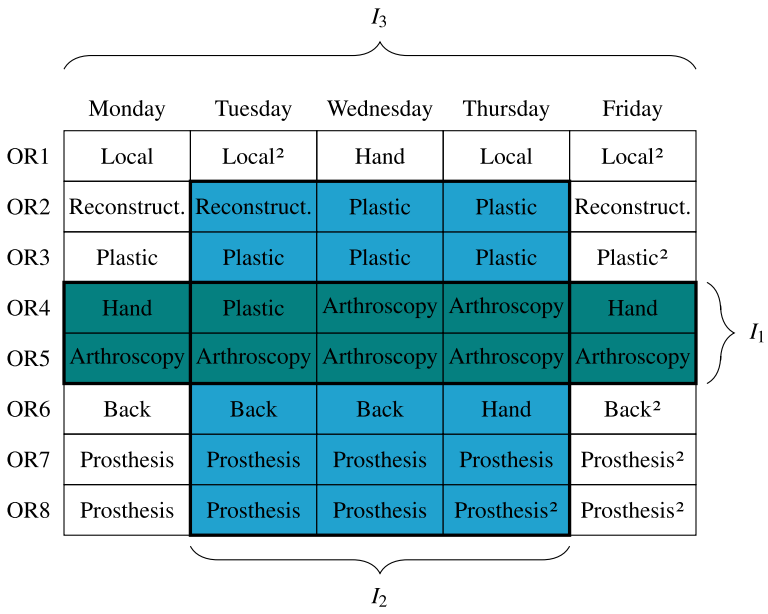


Fig. 9 Instance classes created from a real-life MSS. Brackets and colors indicate 3 different classes. The overlapping section in the middle is included in I_1 and I_2 (in the original MSS, this operating room is either closed or open for various subspecialties. The aspect of closed or open operating rooms is omitted for simplicity.)

6.1.2 Instance Generation

We categorize 9 instances into 3 different classes based on the number of room-day combinations considered (Fig. 9). *Small* instances, I_1, consider schedules with 2 operating rooms in a planning period of 5 days ($2 \times 5 = 10$ room-day combinations) and involve 3 subspecialties: Hand, Arthroscopy and Plastic. *Medium* instances, I_2, include 7 operating rooms over 3 days ($7 \times 3 = 21$ room-day combinations) and involve 5 subspecialties: Reconstructive, Plastic, Arthroscopy, Back, Hand and Prosthesis. *Large* instances, I_3, include all 8 operating rooms over a weekly period ($8 \times 5 = 40$ room-day combinations) and involve all 7 subspecialties. Table 1 summarizes this distinction within each category in 9 different instances. For each instance category (I_1, I_2, I_3), 3 different cases of patient lists are considered:

(i) the case where the load on the list matches the available resources well,
(ii) the case where the load on the list is approximately 50% larger than the available capacity, and
(iii) the case where the load on the patient waiting list matches the capacity available over two periods.

Table 1 9 different instances experimented with in this work

Instance	D	R	D × R	K	N
$I_1^{(1)}$	5	2	10	3	34
$I_1^{(2)}$	5	2	10	3	49
$I_1^{(3)}$	5	2	10	3	64
$I_2^{(1)}$	3	7	21	9	64
$I_2^{(2)}$	3	7	21	9	93
$I_2^{(3)}$	3	7	21	9	127
$I_3^{(1)}$	5	8	40	10	129
$I_3^{(2)}$	5	8	40	10	194
$I_3^{(3)}$	5	8	40	10	255

6.1.3 Surgery Properties

All surgeries are assigned a referral date (H_i) and a deadline (G_i), which are set somewhat arbitrarily between 0 and 20 days. The first day in the planning period is defined to be day 0, H_i is defined as the number of days *backward* in time, and G_i is defined as some days *forward* in time. The estimated duration of surgeries ($\mathscr{D}_i$) is based on a document schedulers in the hospital use. From [2, 4], the instances are prepared for a discrete time model and duration is given in a number of periods. We convert these durations by multiplying the number of periods with the period length. We assume that durations follow a lognormal distribution, with the estimated duration as expected value (μ) and two different cases for standard deviation:

(i) $\sigma \sim 0$, i.e. the deterministic case, and
(ii) $\sigma = c_v \mu$ where c_v is the coefficient of variance that we set to 0.1.

6.2 Degree of Conflict Between Objectives

Mutliobjective optimization makes sense when the desired properties of the sought solution are in partial conflict. We investigate how the objectives relate to each other and argue that they are indeed partially conflicting. Our analysis is based on the hybrid GA applied to the deterministic problem, as this simplifies the analysis without loss of generality. As an example, instance $I_1^{(2)}$ illustrates the schedules produced by performing variations of single-objective optimization. A solution with 0 operating room overtime is shown in Fig. 10. All surgeries end before 16:00 which is the end of regular opening hours. There is much idle time, both in operating rooms and for surgeons. The amount of idle time also leads to a high patient waiting cost, $O_W = 59.29$. However, the surgeon overtime is also 0, as might be expected when the surgeon and operating room opening hours are identical. It might seem that operating room overtime (O_O^R) and surgeon overtime (O_O^K) are non-conflicting.

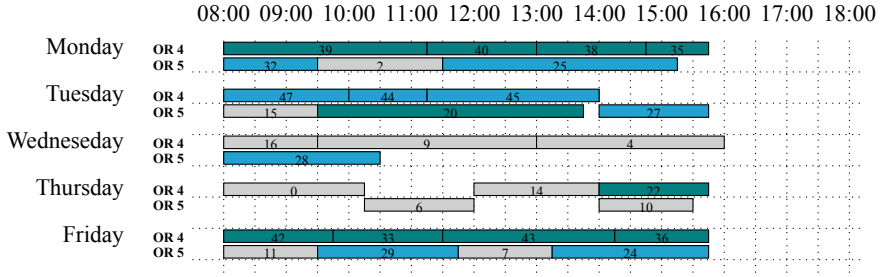


Fig. 10 Optimized operating room overtime for $I_1^{(2)}$: $\{O_O^R = 0, O_I^R = 975, O_O^K = 0, O_I^K = 345, O_W = 59.29\}$

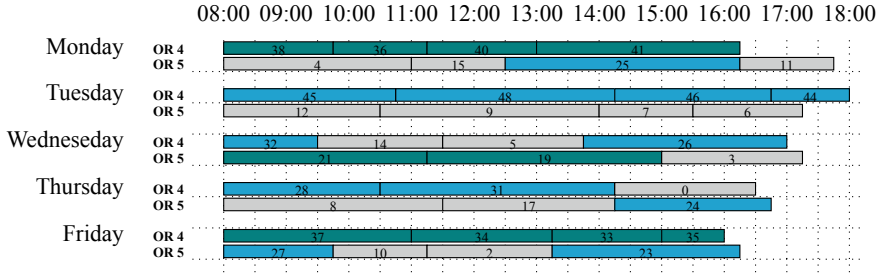


Fig. 11 Optimized operating room idle time for $I_1^{(2)}$: $\{O_O^R = 540, O_I^R = 0, O_O^K = 555, O_I^K = 765, O_W = 38.32\}$

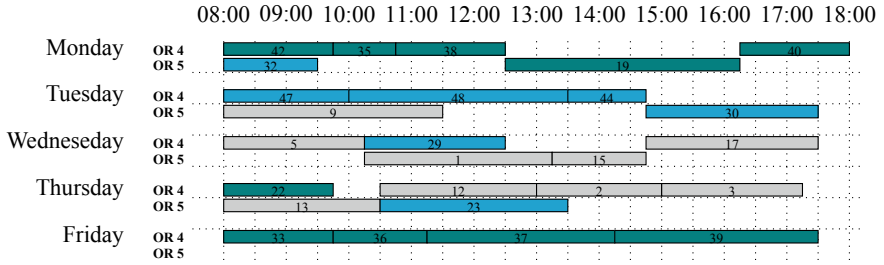


Fig. 12 Optimized surgeon idle time for $I_1^{(2)}$: $\{O_O^R = 32, O_I^R = 113, O_O^K = 31, O_I^K = 0, O_W = 70.65\}$

However, Fig. 11 and 12 present schedules with 0 operating room idle time and 0 surgeon idle time respectively, however with different relation between O_O^R and O_O^K.

In Fig. 11, a surgeon finishes a surgery in an operating room after regular opening hours and then starts a new surgery in another operating room. Overtime is then incurred in two operating rooms, but only for one surgeon. The opposite happens in Fig. 12, when a surgeon finishes a surgery in an operating room after regular opening hours before another surgeon starts a new surgery in the same operating room. Overtime is then incurred for two surgeons, but only for one operating room.

Situations can occur where a choice must be made between scheduling two surgeons to work overtime in the same operating room or reserving overtime capacity for a single surgeon in two operating rooms. Thus, the objectives are partially conflicting. However, these situations are not common in desirable schedules because it requires surgery to start after regular opening hours. Our main reason for separating the two objectives is two acknowledge that they represent two different concepts, and prepare the model and algorithms for real-life situations where the surgeon and operating room opening hours often are different. Based on the above analysis, we state that all of the five objectives are partially conflicting. It does not mean that all of them must be included in a multiobjective surgery scheduling problem. Choosing appropriate performance measures is a task for hospital managers.

6.3 The Effect of Uncertainty

Several existing works balance multiple objectives during surgery scheduling optimization by using the weighted average approach. We illustrate how a weighted average approach using expected durations leads to an efficient schedule in Fig. 13. Exposed to uncertainty, however, under the constriction that no surgery can start earlier than scheduled, the schedule quality deteriorates. The result is shown in Fig. 14. Substantial overtime is incurred, e.g. in OR 4 on Tuesday and OR 5 on Wednesday. Comparing with the planned schedule, it is evident that some of the overtime is propagated from delay earlier in the day: surgery 45 and surgery 19 contribute to most of the delay in OR 4 on Tuesday and OR 5 on Wednesday, respectively. By simulating these effects during optimization, schedules where durable surgeries are placed at the end of the day are preferred. Thus, we are convinced that proper surgery scheduling algorithms should consider duration uncertainty during the search process.

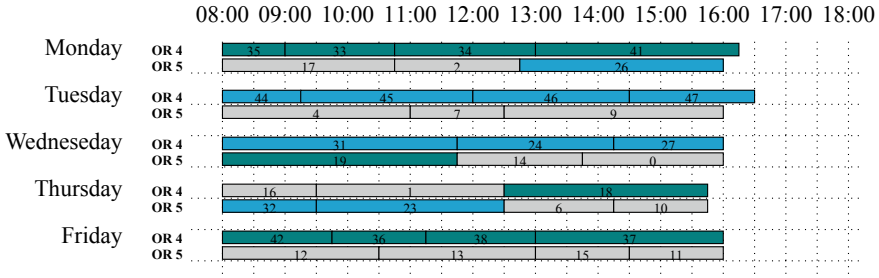


Fig. 13 Optimized weighted sum for $I_1^{(2)}$: $\{O_O^R = 45, O_I^R = 30, O_O^K = 45, O_I^K = 0, O_W = 36.76\}$

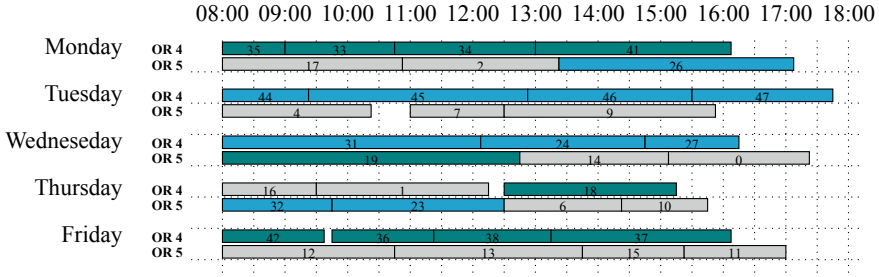


Fig. 14 Realized schedule of the schedule in Fig. 13 (coefficient of variance = 0.1)

Table 2 Critical values for Wilcoxon signed rank test for sample sizes $n = 10$ and 30

n	α			
	0.01	0.02	0.05	0.10
10	3	5	8	11
30	109	120	137	152

6.4 Comparison of SPEA2 and NSGA-II

We compare the two implemented MOEAs by a comprehensive set of performance measures:

(i) Pareto-optimality is assessed by the ratio of nondominated individuals (RNI), overall nondominated vector generation (ONVG) and the Pareto dominance indicator (NR),
(ii) Distribution of solutions is measured by uniform distribution (UD) and the number of distinct choices (NDC), and
(iii) Maximum spread is approximated using the single objective Hybrid GA.

Due to the limited time available for testing, we consider only three instances ($I_1^{(1)}$, $I_2^{(2)}$, $I_3^{(3)}$) in this analysis. We consider the hyperarea metric to be too comprehensive and out of the scope of this work.[3] Because the true Pareto-front is not known for the investigated instances, we cannot measure convergence of the obtained Pareto-front. We address the need for benchmark instances with known true Pareto-fronts as our future research goal. Results for the smallest instance ($I_1^{(1)}$) are based on 30 computation runs whereas results for the medium ($I_2^{(2)}$) and large ($I_3^{(3)}$) instances are based on 10 computation runs. We conduct the Wilcoxon signed-rank test using critical values shown in Table 2. Both 10 and 30 runs are sufficient for achieving results of statistical significance.

[3]For the considered problem, the union of five-dimensional hypervolumes must be computed in order to measure hyperarea. A fast algorithm for computing the hyperarea metric for up to 10 objectives is presented in [48].

Table 3 Comparison of SPEA2 and NSGA-II for a selection of instances

		$I_1^{(1)}$	$I_2^{(2)}$	$I_3^{(3)}$
Pareto-optimality	ONVG	SPEA2 $\sim$ NSGA-II	SPEA2 $\succ$ NSGA-II	SPEA2 $\succ$ NSGA-II
		–	$p < 0.01$	$p < 0.01$
	RNI	SPEA2 $\sim$ NSGA-II	SPEA2 $\succ$ NSGA-II	SPEA2 $\succ$ NSGA-II
		–	$p < 0.01$	$p < 0.01$
	NR	SPEA2 $\succ$ NSGA-II	SPEA2 $\prec$ NSGA-II	SPEA2 $\prec$ NSGA-II
		$p < 0.01$	$p < 0.01$	$p < 0.01$
Distribution	UD	SPEA2 $\prec$ NSGA-II	SPEA2 $\prec$ NSGA-II	SPEA2 $\prec$ NSGA-II
		$p < 0.01$	$p < 0.01$	$p < 0.01$
	$NDC_{0.2}$	SPEA2 $\prec$ NSGA-II	SPEA2 $\sim$ NSGA-II	SPEA2 $\sim$ NSGA-II
		$p < 0.01$	–	–
Maximum spread		SPEA2 $\sim$ NSGA-II	SPEA2 $\sim$ NSGA-II	SPEA2 $\succ$ NSGA-II
		–	–	$p < 0.02$
Set coverage		SPEA2 $\sim$ NSGA-II	SPEA2 $\prec$ NSGA-II	SPEA2 $\prec$ NSGA-II
		–	$p < 0.01$	$p < 0.01$

Table 3 shows the comparison results with the following notation:

(i) if H_0 cannot be rejected with significance level, $\alpha < 0.05$, we write SPEA2 $\sim$ NSGA-II indicating similar performance of two algorithms,
(ii) if H_0 can be rejected so that SPEA2 outperforms NSGA-II, we write SPEA2 $\succ$ NSGA-II,
(iii) if H_0 can be rejected so that NSGA-II outperforms SPEA2, we write SPEA2 $\prec$ NSGA-II.

We also indicate the p-value whenever H_0 is rejected. The only conclusion that is consistent for all instances is the uniform distribution measure, which indicates that the NSGA-II yields approximation fronts with less clustered solutions. From looking at actual solutions provided by the two algorithms, it seems that replicates of solutions are kept in the archive of SPEA2, which can explain why NSGA-II outperforms SPEA2 with regards to this metric. The most interesting result from Table 3, however, is the fact that NSGA-II outperforms SPEA2 on larger instances ($I_2^{(2)}$ and $I_3^{(3)}$) as measured by the set coverage metric and the NR. This is contrary to what is generally reported in literature—the SPEA2 usually outperforms the NSGA-II when more than two objectives are optimized [8]. This result indicates that this might not be the case when the fitness value of individuals is uncertain.

6.5 Dealing with Uncertainty

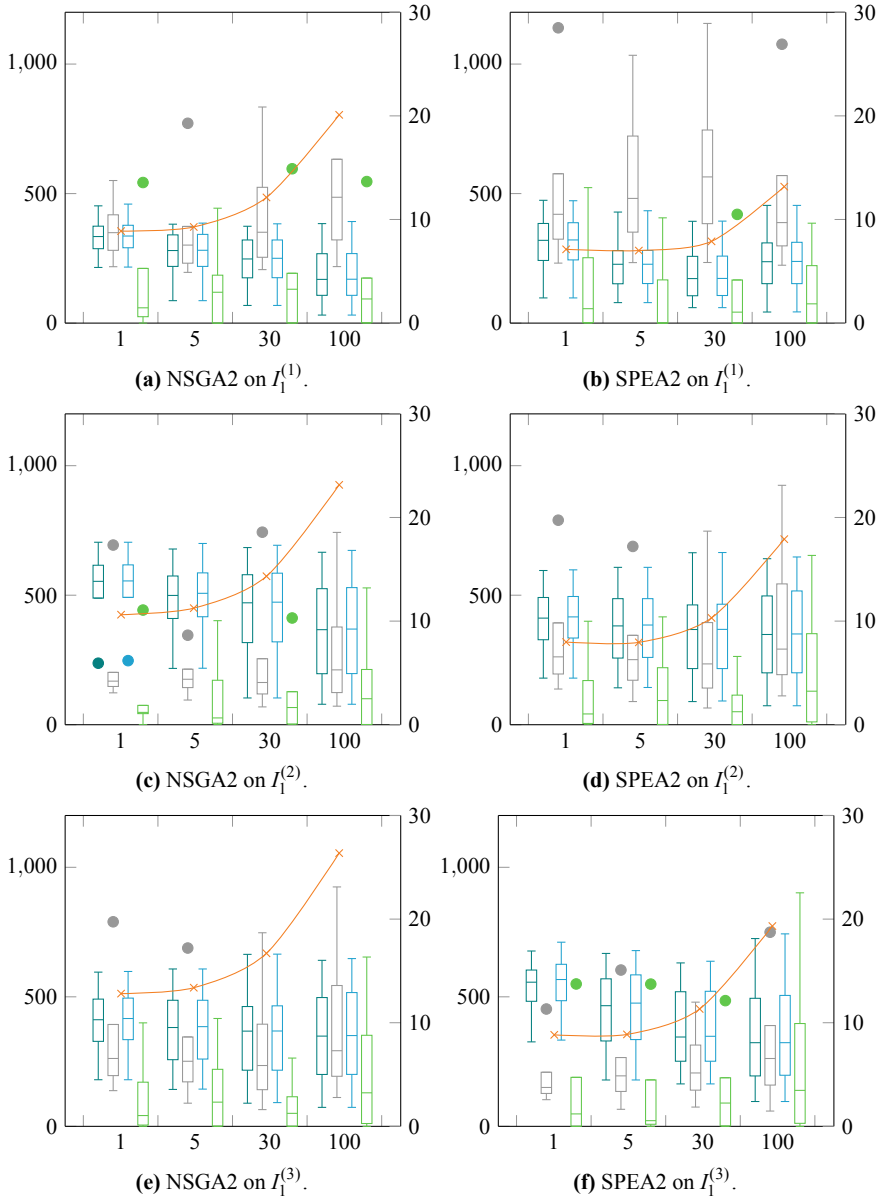
We investigate the effect of incorporating uncertainty during an evolutionary search for high-quality schedules by the following two different ways:

6.5.1 Explicit Averaging

An intuitive way of handling uncertainty during multiobjective evolutionary optimization is explicit averaging. With this approach, the fitness of an individual is estimated using Monte-Carlo simulation. An important decision when using this approach is the number of simulation runs used to estimate the fitness. There is a trade-off between the accuracy of estimation and the computation cost associated with each simulation run. Figure 15 shows the result of running the NSGA-II and SPEA2 on $I_1^{(1)}$, $I_1^{(2)}$ and $I_1^{(3)}$ with different choices for the number of simulation runs involved in each fitness evaluation: (i) a single simulation run, (ii) 5 simulation runs, (iii) 30 simulation runs, and (iii) 100 simulation runs.

Each individual in the resulting population is evaluated with 100 simulation runs, and the objectives are expressed in terms of boxplots showing the 2nd percentile, the 1st quantile, the median, the 3rd quantile and the 98th percentile of objective values present in the population. We find, analyzing with the *Shapiro–Wilk normality test*, that the objectives represented in the population or archive are generally not normally distributed. One explanation for this is frequently encountered lower bounds, e.g. 0 operating room overtime, 0 surgeon overtime or 0 surgeon idle time. We, therefore, base our percentiles and quantiles on the following nonparametric approach. All objectives, $m \in M$, are sorted in increasing order in M separate lists (A_m). Next, any percentile, ρ, of objective $m \in M$ can be estimated by taking the $||(\rho * |A_m|)||$th element from A_m, where $||x||$ denotes the nearest integer of x, and $|A_m|$ is the size of the sorted list. The computation time needed to solve each instance is indicated by the orange line referring to the values on the left y-axis in Fig. 15.

An interesting takeaway from the results in Fig. 15 is that there seems to be a linear relationship between the number of simulations and computation time. A simple linear regression analysis conducted in MS Excel gave the following R^2-values and associated gradients, Δ, in the fitted linear function for the computation time in subfigures (a)–(f): $\{R^2 = 1, \Delta = 0.1136\}$, $\{R^2 = 0.9755, \Delta = 0.0634\}$, $\{R^2 = 0.9999, \Delta = 0.126\}$, $\{R^2 = 0.9965, \Delta = 0.1027\}$, $\{R^2 = 0.9999, \Delta = 0.11369\}$, $\{R^2 = 0.997, \Delta = 0.1082\}$. Thus, a linear relationship seems likely and using the average gradient, $\overline{\Delta} = 0.1046$ may serve as an estimate for the marginal computation cost of a simulation. To exemplify, we would approximate the added computational cost of using 50 simulation runs to estimate fitness to be 50 runs $\times$ 0.1046 min/run $=$ 5 min and 14 s. The described linear relationship is not surprising, as increasing the number of simulation runs only increases the thoroughness of each fitness evaluation and not the number of fitness evaluations. The computational cost is relatively low because the chromosome under evaluation is only decoded once before the simulation



(a) NSGA2 on $I_1^{(1)}$.

(b) SPEA2 on $I_1^{(1)}$.

(c) NSGA2 on $I_1^{(2)}$.

(d) SPEA2 on $I_1^{(2)}$.

(e) NSGA2 on $I_1^{(3)}$.

(f) SPEA2 on $I_1^{(3)}$.

Fig. 15 Explicit averaging for instance $I_1^{(1)}$, $I_1^{(2)}$ and $I_1^{(3)}$. Box plots indicate, from left to right, operating room overtime, operating room idle time, surgeon overtime and surgeon idle time. The left y-axis indicates performance in minutes. The right y-axis indicates the computation time in minutes

Table 4 Estimate of computation time needed for implicit averaging

Problem	50 individuals	200 individuals	500 individuals	1000 individuals
NSGA2 on $I_1^{(1)}$ (a)	0.57	2.6	8.98	27.33
SPEA2 on $I_1^{(1)}$ (b)	0.28	1.53	7.76	28.41
NSGA2 on $I_1^{(2)}$ (c)	0.78	3.53	11.25	31.26
SPEA2 on $I_1^{(2)}$ (d)	0.33	1.83	7.9	29.28
NSGA2 on $I_1^{(3)}$ (e)	1.03	4.45	13.63	35.33
SPEA2 on $I_1^{(3)}$ (f)	0.43	2.23	8.98	31.23

is performed to determine starting times, ending times and objectives for different outcomes of surgery durations. The decoding procedure is quite comprehensive, but the simulation procedure is not.

An alternative to explicit averaging is implicit averaging that randomly perturbs the design variables of individuals in very large populations to average out peaks and ensure convergence over time. We have not implemented implicit averaging, but we can asses the computational cost of increasing the population size while using only one simulation run to estimate the fitness of an individual. The computation time results for the same instances that we investigated in Fig. 15 are shown in Table 4. Apparently, the computation time increases non-linearly with the number of individuals in the population. Indeed, the simple analysis in MS Excel indicates a better polynomial or power fit than linear fit on the computation time as a function of population size. This nonlinear relationship is in line with theory as there are sorting procedures in both the SPEA2 and NSGA2 with nonlinear complexity concerning population or archive size. Although we cannot compare explicit and implicit averaging without comparing the solution quality of each approach, we assess that explicit averaging is more comprehensible, straightforward and less computationally expensive than implicit averaging.

There are two central tendencies regarding how the population improves as the number of simulations increases. First, the operating room and surgeon overtime measures are significantly improved. By improved, we mean that the three lower box plot values (2nd percentile, the 1st quantile and the median) are lower. Occasionally, this comes at a cost, namely an increase in the surgeon and operating room idle time. However, overtime improvement seems to outweigh the idle time deterioration. Also, in most of the cases, solutions with idle times as low as the 2nd percentile in the 1-scenario population exist. Especially the surgeon idle time performance measure seems well maintained as scenarios increase. This leads us to the second significant tendency—the spread in solution space seems to increase with the number of simulations conducted during fitness evaluation.

This is because inaccurate fitness estimation leads to fluctuations in the nondomination relationships among individuals. Depending on the particular outcome of surgery durations, an individual can be nondominated in one generation and dom-

inated in the next. In such environments, solutions that are robust against duration variability will be favoured, while risky solutions will not survive occasional unfortunate variations. The underlying assumption for a larger spread with the increasing number of simulation runs is: *Increasing expected performance also increases performance variability*. This is a popular and much-applied assumption in finance where the optimal trade-off between expected return and risk of different stocks outlines the efficient Pareto-front. If only a few scenarios are used for a fitness evaluation, nondominated solutions on expectation (solutions with close to 0 overtime without too much idle time) might occasionally be dominated and left out during selection. It illustrates a major difficulty with multiobjective optimization under uncertainty—there are three notions of multiobjectivity in such problems:

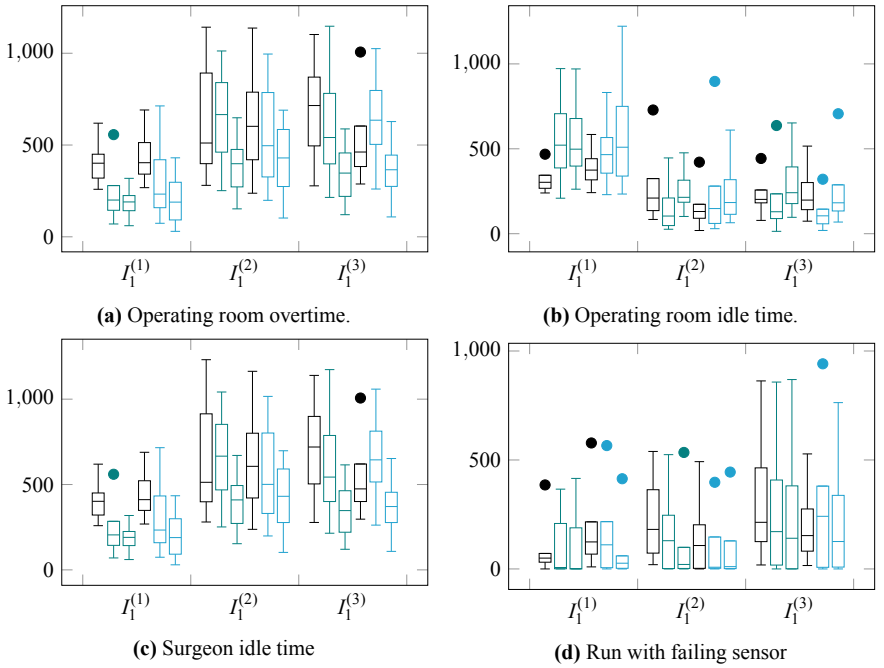
(i) the trade-off among partially conflicting performance measures,
(ii) the trade-off between closeness to, and coverage of, the Pareto-front, and
(iii) the trade-off between expected performance and performance variability.

6.5.2 Comparison with a Deterministic Approach

To analyze the benefit of incorporating uncertainty for decision makers, we compare solutions found by the developed SPEA2 and NSGA-II with their deterministic counterparts. The deterministic versions base the search process on expected values for the surgery durations. Note that this is not the same as using a single simulation run for fitness evaluation as the expected value is used both for schedule construction and fitness evaluation. The resulting solution set is then evaluated assuming uncertain surgery durations and compared against the algorithms that incorporate uncertainty during evolution. To differentiate between improvements gained by incorporating uncertainty and improvements gained from hedging, we use two different values for the scheduling confidence in the SPEA2 and NSGA 2: the expected value and the 65th percentile. Figure 16 shows results for $I_1^{(1)}$, $I_1^{(2)}$ and $I_1^{(3)}$.

In general, the overtime performance measures are substantially improved, especially when combining uncertainty incorporation and hedging. To illustrate, the mean operating room overtime is reduced from 401 to 190 min for $I_1^{(1)}$ with SPEA2; 403–189 min for $I_1^{(1)}$ with NSGA2; 510–398 min for $I_1^{(2)}$ with SPEA2; 601–429 min for $I_1^{(2)}$ with NSGA2; 715–347 min for $I_1^{(3)}$ with SPEA2 and 461–365 min for $I_1^{(3)}$. This is an average overtime decrease of 38.10%. Usually, there is an improvement without hedging as well, but this is not consistent. We expect this to be due to random performance variations of the algorithms; a reminder that this is only anecdotal evidence and that thorough statistical analysis should be conducted in the future.

The waiting cost for deterministic, uncertainty incorporating and hedging versions of the algorithms is shown in Fig. 17. It is not a surprise that uncertainty incorporation and hedging induces more waiting cost for patients, as the produced schedules should be less dense to avoid overtime. The increase in waiting cost from $I_1^{(1)}$ to $I_1^{(2)}$ to $I_1^{(3)}$ is due to the larger number of patients on the patient waiting lists associated with these instances.



(a) Operating room overtime.

(b) Operating room idle time.

(c) Surgeon idle time

(d) Run with failing sensor

Fig. 16 Deterministic versus stochastic solutions. Box plots indicate, from left to right, SPEA2 deterministic, SPEA2 without hedging, SPEA2 with hedging, NSGA-II deterministic, NSGA-II without hedging and NSGA-II with hedging. The y-axis indicates performance in minutes

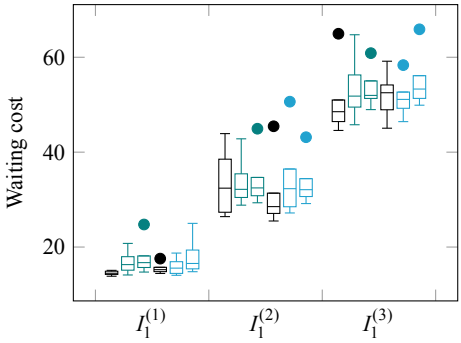


Fig. 17 Comparison of waiting time objective using deterministic and stochastic approach. Box plots indicate, from left to right, SPEA2 deterministic, SPEA2 without hedging, SPEA2 with hedging, NSGA-II deterministic, NSGA-II without hedging and NSGA-II with hedging

6.6 Discussion

The presented analysis provides some interesting insights. We see improvement over deterministic approaches, especially regarding overtime objectives. An advantage of the proposed algorithms is their flexibility. By easily changing the fitness function, selection may be based on a weighted average or a selection of performance measures.

To accommodate different levels of user risk aversion, the expected value, mean or any percentile of the simulation results may be used as fitness value during evolution. Although we acknowledge that the developed algorithms cannot be put to immediate use by any hospital, it is not hard to imagine actual use after risk adjustments and performance prioritizations have been done in tandem with hospital management at a particular department treating elective patients. In terms of optimization objectives, overtime and surgeon idle time on one hand, and operating room idle time and patient waiting time on the other pulls in different directions when it comes to exclusion and inclusion of surgeries, respectively. Thus, it is appropriate to include at least one driver of including surgeries and one driver for excluding surgeries during optimization. These performance measures must reflect the uncertainty present in realistic surgery scheduling problems.

The novel evolutionary mechanisms developed here exemplifies how MOEAs may be applied to a variation of the surgery scheduling problem. Scheduling problems are among the hardest combinatorial optimization problems, and the problem instances investigated here are both large and complex with uncertain processing times. We assess that the presented results serve as an argument for continued development and experimentation with MOEAs on surgery scheduling problems, as we find several promising solutions within 20 min of computation time for relatively large problem instances.

Our results indicate that using explicit averaging with Monte-Carlo simulation is a predictably fast way of estimating fitness during evolutionary optimization. Based on an elementary computational analysis, we do not recommend implicit averaging for this problem. However, this conclusion cannot be generalized to problems where decisions are made after uncertainty resolution. Our approach does not accurately asses the dynamics in real hospitals where decisions, such as rescheduling activities and cancellations, actually take place as uncertainty resolution is observed throughout the day. In such situations, the simulation cost is probably much higher, and implicit averaging might be a better option then explicit averaging. We have also realized, through reflection and discussion, that uncertainty incorporation comes with a new set of multiobjective considerations regarding expected value and variability of performance. It is appropriate to involve hospital management in such considerations as the task of locating the true Pareto-front for all possible trade-offs between expected performance and performance variability seems unrealistic.

7 Conclusion

Scheduling surgeries in any hospital is a dynamic process that must balance among conflicting objectives and response invariably to unexpected events. This chapter focuses on efficiently finding multiple distinguished, high-quality trade-off solutions to the surgery admission planning problem which are resistant to surgery duration variability. We apply the SPEA2 and the NSGA-II using developed evolutionary operators to the multiobjective surgery admission planning problem. In this prob-

lem, patients are selected from a patient waiting list and scheduled into operating rooms over a weekly period in a way that maximizes resource utilization and minimizes patient waiting time on the waiting list. The algorithm output is a set of the detailed and compatible surgeon and operating room schedules that are robust against surgery duration variability. Our most important contribution is uncertainty incorporation in multiobjective search for robust surgery schedules. The proposed way of estimating fitness with Monte-Carlo simulation (explicit averaging) performed after decoding shows significant improvements in the overtime performance measures without incurring too much computational cost. Our analysis exemplifies how significant overtime reduction can be achieved by incorporating uncertainty in the optimization process. The experimental results also justify that MOEAs have great potential in solving practical surgery scheduling problems. We hope that the presented work may contribute to more manageable daily operations at hospitals in general.

However, the proposed approach assumes a simplified problem structure where no decisions are made after the resolution of uncertainty. Future research is necessary to consider dynamic surgery scheduling where rescheduling activities are allowed as realized durations are observed throughout the day. Implicit averaging should be considered in future as an option if simulation becomes too time-consuming. Because simultaneous handling of multiple objectives and uncertainty quickly becomes complicated, we acknowledge that hospital managers preference regarding performance trade-offs and risk should be integrated in the future. It is particularly vital to align user preference when it comes to performance trade-offs and risk with the algorithmic search for diversity and robustness. Also, the presented hyperparameters are based on a somewhat ad-hoc trial and error procedure, and a systematic adjustment of essential parameters such as population size and mutation rate could lead to better computational performance. Furthermore, we are not able to confidently measure the convergence of our MOEAs, as all related convergence metrics require the true Pareto-front which is not known for the investigated problem instances. We, therefore, call for benchmark instance generation of practical multiobjective surgery scheduling problems that can assist accurate quality assessment of multiobjective surgery scheduling algorithms.

References

1. Cardoen, B., Demeulemeester, E., Beliën, J.: Operating room planning and scheduling: a literature review. Eur. J. Oper. Res. **201**(3), 921–932 (2010)
2. Nyman, J., Ripon, K.S.N.: Metaheuristics for the multiobjective surgery admission planning problem. In: 2018 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8. IEEE (2018)
3. Riise, A., Burke, E.K.: Local search for the surgery admission planning problem. J. Heuristics **17**(4), 389–414 (2011)
4. Nyman, J.H.: Multiobjective evolutionary surgery scheduling under uncertainty. Master's thesis, Norwegian University of Science and Technology, Norway (2018)
5. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: improving the strength pareto evolutionary algorithm. TIK-report 103 (2001)

6. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evolut. Comput. **6**(2), 182–197 (2002)
7. Ripon, K.S.N., Kwong, S., Man, K.F.: A real-coding jumping gene genetic algorithm (RJGGA) for multiobjective optimization. Inform. Sci. **177**(2), 632–654 (2007)
8. Yen, G.G., He, Z.: Performance metric ensemble for multiobjective evolutionary algorithms. IEEE Trans. Evolut. Comput. **18**(1), 131–144 (2014)
9. Jin, Y., Branke, J., et al.: Evolutionary optimization in uncertain environments-a survey. IEEE Trans. Evolut. Comput. **9**(3), 303–317 (2005)
10. Gillespie, B.M., Chaboyer, W., Fairweather, N.: Factors that influence the expected length of operation: results of a prospective study. BMJ Qual. Saf. **21**(1), 3–12 (2012)
11. Joustra, P., Meester, R., van Ophem, H.: Can statisticians beat surgeons at the planning of operations? Empir. Econ. **44**(3), 1697–1718 (2013)
12. Kayış, E., Khaniyev, T.T., Suermondt, J., Sylvester, K.: A robust estimation model for surgery durations with temporal, operational, and surgery team effects. Health Care Manag. Sci. **18**(3), 222–233 (2015)
13. Choi, S., Wilhelm, W.E.: An analysis of sequencing surgeries with durations that follow the lognormal, gamma, or normal distribution. IIE Trans. Healthc. Syst. Eng. **2**(2), 156–171 (2012)
14. Cheng, R., Gen, M., Tsujimura, Y.: A tutorial survey of job-shop scheduling problems using genetic algorithms–I. Representation. Comput. Ind. Eng. **30**(4), 983–997 (1996)
15. Ripon, K.S.N., Tsang, C.H., Kwong, S.: An evolutionary approach for solving the multi-objective job-shop scheduling problem. In: Evolutionary Scheduling, pp. 165–195. Springer, Berlin (2007)
16. Ripon, K.S.N.: Hybrid evolutionary approach for multi-objective job-shop scheduling problem. Malays. J. Comput. Sci. **20**(2), 183–198 (2007)
17. Sprecher, A., Kolisch, R., Drexl, A.: Semi-active, active, and non-delay schedules for the resource-constrained project scheduling problem. Eur. J. Oper. Res. **80**(1), 94–102 (1995)
18. Fei, H., Meskens, N., Chu, C.: A planning and scheduling problem for an operating theatre using an open scheduling strategy. Comput. Ind. Eng. **58**(2), 221–230 (2010)
19. Gonçalves, J.F., de Magalhães Mendes, J.J., Resende, M.G.: A hybrid genetic algorithm for the job shop scheduling problem. Eur. J. Oper. Res. **167**(1), 77–95 (2005)
20. Cheng, R., Gen, M., Tsujimura, Y.: A tutorial survey of job-shop scheduling problems using genetic algorithms, part II: hybrid genetic search strategies. Comput. Ind. Eng. **36**(2), 343–364 (1999)
21. Batun, S., Denton, B.T., Huschka, T.R., Schaefer, A.J.: Operating room pooling and parallel surgery processing under uncertainty. INFORMS J. Comput. **23**(2), 220–237 (2011)
22. Cardoen, B., Demeulemeester, E., Beliën, J.: Optimizing a multiple objective surgical case sequencing problem. Int. J. Prod. Econ. **119**(2), 354–366 (2009)
23. Gen, M., Tsujimura, Y., Kubota, E.: Solving job-shop scheduling problems by genetic algorithm. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, vol. 2, 1577–1582. IEEE (1994)
24. Pezzella, F., Morganti, G., Ciaschetti, G.: A genetic algorithm for the flexible job-shop scheduling problem. Comput. Oper. Res. **35**(10), 3202–3212 (2008)
25. Etiler, O., Toklu, B., Atak, M., Wilson, J.: A genetic algorithm for flow shop scheduling problems. J. Oper. Res. Soc. **55**(8), 830–835 (2004)
26. Iyer, S.K., Saxena, B.: Improved genetic algorithm for the permutation flowshop scheduling problem. Comput. Oper. Res. **31**(4), 593–606 (2004)
27. Louis, S.J., Xu, Z.: Genetic algorithms for open shop scheduling and re-scheduling. In: Proceedings of the 11th ISCA International Conference on Computers and their Applications, vol. 28, pp. 99–102 (1996)
28. Gul, S., Denton, B.T., Fowler, J.W., Huschka, T.: Bi-criteria scheduling of surgical services for an outpatient procedure center. Prod. Oper. Manag. **20**(3), 406–417 (2011)
29. Marques, I., Captivo, M.E., Pato, M.V.: A bicriteria heuristic for an elective surgery scheduling problem. Health Care Manag. Sci. **18**(3), 251–266 (2015)

30. Mateus, C., Marques, I., Captivo, M.E.: Local search heuristics for a surgical case assignment problem. Oper. Res. Health Care **17**, 71–81 (2018)
31. Dellaert, N., Jeunet, J.: A variable neighborhood search algorithm for the surgery tactical planning problem. Comput. Oper. Res. **84**, 216–225 (2017)
32. Ripon, K.S.N., Glette, K., Khan, K.N., Hovin, M., Torresen, J.: Adaptive variable neighborhood search for solving multi-objective facility layout problems with unequal area facilities. Swarm Evol. Comput. **8**, 1–12 (2013)
33. Molina-Pariente, J.M., Hans, E.W., Framinan, J.M., Gomez-Cia, T.: New heuristics for planning operating rooms. Comput. Ind. Eng. **90**, 429–443 (2015)
34. Zhou, Bh, Yin, M., Zq, Lu: An improved lagrangian relaxation heuristic for the scheduling problem of operating theatres. Comput. Ind. Eng. **101**, 490–503 (2016)
35. Jebali, A., Diabat, A.: A stochastic model for operating room planning under capacity constraints. Int. J. Prod. Res. **53**(24), 7252–7270 (2015)
36. Guido, R., Conforti, D.: A hybrid genetic approach for solving an integrated multi-objective operating room planning and scheduling problem. Comput. Oper. Res. **87**, 270–282 (2017)
37. Zhang, Z., Li, C., Wang, M., Wu, Q.: A hybrid multi-objective evolutionary algorithm for operating room assignment problem. J. Med. Imaging Health Inform. **7**(1), 47–54 (2017)
38. Xiang, W.: A multi-objective ACO for operating room scheduling optimization. Nat. Comput. **16**(4), 607–617 (2017)
39. Bruni, M., Beraldi, P., Conforti, D.: A stochastic programming approach for operating theatre scheduling under uncertainty. IMA J. Manag. Math. **26**(1), 99–119 (2015)
40. Denton, B., Viapiano, J., Vogl, A.: Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag. Sci. **10**(1), 13–24 (2007)
41. Rath, S., Rajaram, K., Mahajan, A.: Integrated anesthesiologist and room scheduling for surgeries: Methodology and application. Oper. Res. **65**(6), 1460–1478 (2017)
42. Min, D., Yih, Y.: Scheduling elective surgery under uncertainty and downstream capacity constraints. Eur. J. Oper. Res. **206**(3), 642–652 (2010)
43. Pulido, R., Aguirre, A.M., Ibáñez-Herrero, N., Ortega-Mier, M., García-Sánchez, Á., Méndez, C.A.: Optimization methods for the operating room management under uncertainty: stochastic programming vs. decomposition approach. J. Appl. Oper. Res. **201**(6), 3 (2014)
44. Kulkarni, A.J., Baki, M.F., Chaouch, B.A.: Application of the cohort-intelligence optimization method to three selected combinatorial optimization problems. Eur. J. Oper. Res. **250**(2), 427–447 (2016)
45. Winston, W.L., Venkataramanan, M., Goldberg, J.B.: Introduction to Mathematical Programming, vol. 1. Thomson/Brooks/Cole, Pacific Grove (2003)
46. Gao, J., Gen, M., Sun, L., Zhao, X.: A hybrid of genetic algorithm and bottleneck shifting for multiobjective flexible job shop scheduling problems. Comput. Ind. Eng. **53**(1), 149–162 (2007)
47. Chang, H., Huang, K., Wu, C.: Determination of sample size in using central limit theorem for weibull distribution. Int. J. Inf. Manag. Sci. **17**(3), 31 (2006)
48. While, L., Hingston, P., Barone, L., Huband, S.: A faster algorithm for calculating hypervolume. IEEE Trans. Evol. Comput. **10**(1), 29–38 (2006)

Big Data in Electroencephalography Analysis

Dhanalekshmi P. Yedurkar and Shilpa P. Metkar

Abstract Human beings have broad inclinations and logical reasoning capabilities. Some of them are developed genetically, whereas some are modelled through experience. These changes can be observed in different scales and dynamics of neurons. Therefore, it is very essential to analyse the neuronal behaviour in huge data sizes over different human communities. Big data analytics is emerging rapidly as a research area. It acts as a tool to accumulate, analyse, and manage large quantity of dissimilar, structured and unstructured data, particularly in the present medical systems. Big data concept is vastly applied in the disease detection area like epileptic seizure detection. But the adaption rate and research opportunities are delayed by some basic problems in the big data model. The main objective of this chapter is to mathematically model the data generated by an Electroencephalography (EEG) recording system. It is intended to explore the use of big data in managing a huge amount of data. Application of big data in epileptic EEG analysis is also explored.

Keywords Big data · Epilepsy · EEG · Seizure

1 Introduction

1.1 Motivation

Past techniques applied in the field of medicine were concentrated on the detection of diseases and their stages. These detections were dependant on some particular clinical procedures and methods. Hence, there is a necessity of innovative method to understand and analyse the extent of diseases [1]. However, lack of research leads to the undefined inter relationship and variations in disease structure. Due to this delay

D. P. Yedurkar · S. P. Metkar (✉)
Electronics & Telecommunications Engineering, College of Engineering Pune, Shivaji Nagar, Pune 411005, India
e-mail: metkarshilpa@gmail.com

D. P. Yedurkar
e-mail: dhanalekshmipy2013@gmail.com

A. J. Kulkarni et al. (eds.), *Big Data Analytics in Healthcare*,
Studies in Big Data 66, https://doi.org/10.1007/978-3-030-31672-3_8

in advancements, the clinical research has started to adapt recent trends in the usage of data in a discrete form. Ever since the introduction of new clinical equipment's, the information acquired from the laboratories were not used to its fullest extent [2]. Therefore, the clinical data were misapplied in many cases.

Owing to the current advancements in technology, the data produced is rich in content like images, video, audio, etc. when compared to the data used from various other sources which is limited [3]. This leads to more difficulty in generating and deriving the most valuable necessary information. Similarly, in long EEG recordings, necessary information were not extracted. Hence, the detailed examination of the EEG data produced can be uses as an improved approach to extract significant data, thereby allowing accurate disease detection/prediction.

1.2 Challenges

An EEG seizure detecting system have many difficulties. First, as EEG is a non-stationary signal, epileptic and non-epileptic patterns may change throughout the signal for different patients. Second, an EEG signal recorded with an intracranial electrodes will generate huge volume of real time data, which lead the way to the big data problem [4]. This massive amount of data produced thus requires a safe storage system which can be used for the real time data computation.

1.3 Methodology

Big data analysis is a computational technique which perform classic role in understanding the EEG patterns and relationships between the signals in huge datasets. Epilepsy detection from EEG signal is a good example to show the use of three V's of data. The three V's are velocity, Variety and Volume [5]. These V's have the ability to characterise different functionalities of the EEG data thus produced. Figure 1 shows the details of three V's in big data. In addition, truth of the data generated is also of great importance. This will help the detailed analysis, especially in clinical research. Even though the epileptic EEG signal is complex, it can be thoroughly analysed by incorporating big data aids thereby extracting worthy information.

The biological data which are complex and large in size requires in depth inspection, so that one can adapt to these data parameters and can come up with new technologically improved firm ware. By this advancements, the data can be stored and analysed by researchers successfully. All these improvisations are possible through big data analytics.

Important biological signals such as epileptogenic activities are synchronously visible due to wide range of variations of data in each EEG channel. These changes are due to inter-connections between various biological signals in the human body. These interdependency would be generated by the exchange of information between

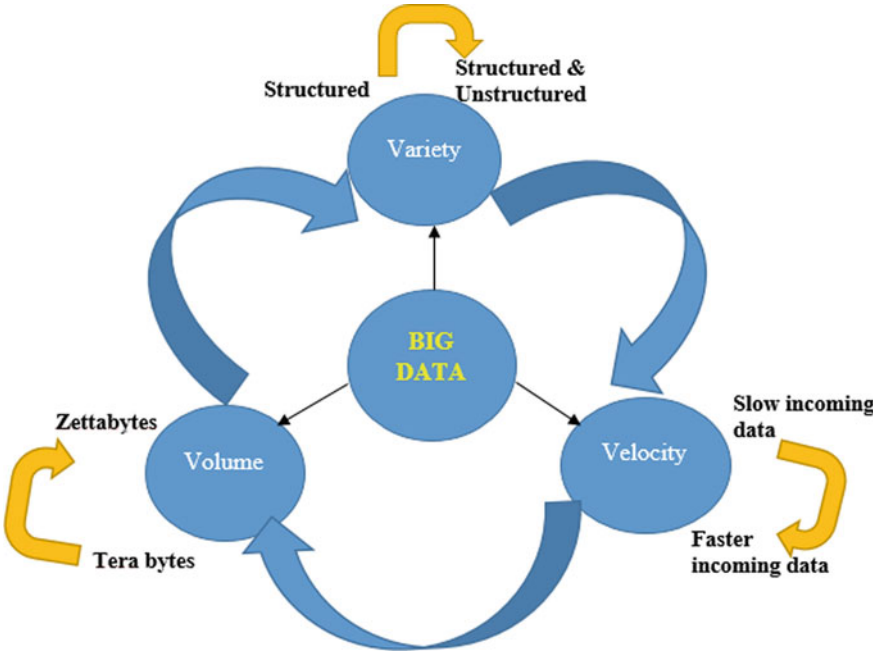


Fig. 1 Three V's of big data

blood pressure, heart rate and respiration [6]. This information can therefore act as a valuable biomarker for clinical investigation. This paves path for detailed analysis and thereby facilitating detection and prediction of epileptic EEG data through these combined biomarkers. The epileptic EEG data would be structured or unstructured one generated from a huge clinical and/or non-clinical aids. These data are further utilised to determine various stages present in an epileptic signal. In this chapter, the analysis of epileptogenic activity is carried out by utilizing the concept of big data. This understanding will provide a detailed look for possible research area in epileptic disease detection and may be prediction as well.

2 EEG as a Critical Tool in Neuroscience

EEG is an effective tool to analyse the functionality of the human brain. EEG signal provides an electrical activity of the brain on millisecond time scale and at a lower cost than other brain imaging technologies. EEG allows the recording of the human brain in natural environment other than within a laboratory set up. It facilitates the effective analysis of the emerging methods to analyse the neuronal behaviour and give deeper insights. This paves path for a wide range of opportunities by enabling

the contribution of large number of researchers and expanding the possibilities of diverse experimentation [7].

2.1 EEG Data Acquisition

It is very important to simplify the complex and interdependent behaviour of the EEG data along with the data scrutiny. Also, the secure data storage, accessing and using the data is essential. The analysis of continuous EEG requires time domain details. Static data will not give the true time information of this continuous data. Also, during the data conversion from continuous to digital the temporal information of the data has to be taken care. Even though the data storage facility is available, the EEG data acquired were rarely stored in the past since the storage and downloading is liable to proprietary software. In addition, the speed of data collection is restricted by network bandwidth, cost and scalability of the data [8].

2.2 Mathematical Model

This section provides the mathematical model to compute the amount of data generated by EEG recordings.

For 'X' hours of EEG data recording, duration of EEG recording in minutes is given by,

$$X \times 60 = M \tag{1}$$

'M' is the duration of EEG data recording in minutes.

Each 'Y' seconds of EEG data contain 's' number of samples.

Total number of samples present per second of EEG recordings are

$$N = \frac{M \times 60}{y} \times s \tag{2}$$

For each sample, if 'B' bit encoding is considered

Total number of bits B_N required to encode N number of samples is given by,

$$B_N = N \times B \text{ bits} \tag{3}$$

Thus for each patient if we consider data for one day with single channel B_N bits of data is generated. If we consider standard 10–20 electrode system for EEG recordings, total amount of EEG data generated with 'H' channel recording is given by,

Table 1 Amount of data generated by EEG recording systems

Duration	Amount of data for single channel EEG recording systems	Amount of data for 23 channels EEG recording systems
1 h	54 MB	161.8 MB
24 h	1 GB	3.8 GB
72 h	3.8 GB	11.6 GB

$$Z = B_N \times H \tag{4}$$

Table 1 shows the total data generated by EEG recording systems for various duration at a sampling frequency of 256 Hz. It can be observed that data produced will increase as the number of channels used for recording increased. With an increase in the sampling rate, larger data is produced which invariably increases the complexity of recording. Thus it is observed that, huge amount of data needs to be handled for EEG analysis.

3 EEG Signal Analytics

EEG signal monitoring devices can be found everywhere. Huge amount of data produced from these devices is not stored beyond a particular point of time. This leads to the lack of examination of the data. Recently, researchers started to utilize these datasets extensively. This paved the path for the time series based analysis and hence improvements in medical treatment and services. Streaming data analytics in medicine refers to a systematic use of continuous data and its related medical information developed through applied analytics, namely predictive analysis [8]. This analysis enables proper and accurate decision making for the treatment of the patient. For signal analytics, we need a media to capture the continuous data with large bandwidth capability and ability to manage multiple signals at various constancy. These signals from the Electronic Health Record (EHR) provides situational and contextual awareness knowledge to the analytics engine.

As the specificity of the EEG signal hugely depends on the disease type, diverse EEG processing techniques are available to get necessary features from the EEG data. These features can then be pre-trained by a machine learning algorithm which provides details of the disease. These analytics can be either diagnostic or predictive analysis. These classification details can be modelled further to provide other mechanisms such as alert systems and giving notification to clinicians. Correlating the continuous data with digital data for finding valuable patient information could be a difficult task. Also, further research for the development of new diagnosis and treatment becomes cumbersome. Several requirements have to be fulfilled for the implementation of such systems, especially bed-side application in clinical surroundings [9].

3.1 Across the Borders of Small Sample Sizes

Large scale data sets enables better functional control of intra personal diversity and application of the big data analytics which can be better used for epileptogenic activity detection using machine learning and deep learning for more accurate results throughout different applications. Effective data management is essential for a variety of data scales so as to ensure the quality parameters like data standards and data validation for a successful analysis. Big data presents an effective data management system to researchers for better managing and sharing the data efficiently without any substantial efforts in data cleaning and data arrangement. For example, in epilepsy detection, vast EEG data are available for the analysis, detection and possibly the prediction of epileptic seizures. One of the most important goals in seizure detection is to find better biomarker for the identification of disease progression. Large scale analysis of EEG enables to get insight of epileptic processes and thereby facilitating early seizure detection.

3.2 Seizure Analysis

At present, perfect seizure analysis systems is not available. Thus there is a need to improve performance of the seizure analysis systems. The main aim of an epileptic seizure detection system is to identify interictal and ictal events in the EEG. The system memory is incorporated as a circular buffer. The function of the buffer is to store the epileptic data such that the time of occurrence of the epileptic form activity can be detected and reviewed. The stored data in the computer can be accessed by a staff member in the clinics. If the epileptic activity do not occur, then the memory is overwritten by the recent data. To save an epileptic data, initiation by the expert is of importance, since a computer based detection may give inaccurate results [10]. This is because, false epileptic activity may be captured by the system and thereby analysed. For example, a circular buffer might save approximately 2–5 min of data before the save triggers and 1–3 min later. In the case of a single epileptic form potential, a second segment might be saved with the event in the centre. Also, save time of the data dependents on the Number of channels of EEG, Sampling frequency and Capacity of memory present in the computer systems.

4 EEG Data Storage and Their Management

An epileptic patient may get multiple seizures in a day. In addition to the ictal and post ictal activities, inter ictal data should also be saved with the routine EEG samples. Hence it is essential for a highly planned and managed file system [11]. The EEG file system consists of a house keeping file which will be ASCII format. This file is

the single source of record for equipment's settings and information which includes patients name, montages, and settings data of hardware (computer and amplifier), history of all the EEG files of patient specific and hospital unit number. This file type also includes time and date information regarding all the changes made.

Annotating the recorded EEG file helps in the analysis and development of the epileptic seizure detection systems. Annotations provide information about the locations and times of the seizure spikes in the EEG data. These annotations are performed by the expert clinicians or computer or sometimes together. These files come with the text descriptions which acts as an aid for further analysis and monitoring of the seizure.

Patient EEG files are also the part of the computer systems which contains the baseline EEG data including ictal, interictal information recorded and detected by computer programs or by experts. These file systems are modelled in order to provide various important functions such as: (i) Fast reviewing of routine EEG signals, (ii) To provide fast access to EEG signal from N number of channels, (iii) Supporting the development of more accurate seizure detection or prediction systems and (iv) To store both continuous and non-continuous EEG data. Some long term EEG monitoring centres uses custom montages, particularly for the intracranial patients. Over a period of time, the list of montages would have been increased in a large amount. An ideal recording system will store EEG data along with montages information. This information stored will be patient-specific. When the EEG data of a patient is moved to an archive, the information related to montages will also have moved along with the patient-specific EEG data. That is, the information related to the present patient with their montages will get displayed in the selection list of the program, whereas all the previous montages information would have been archived.

5 Digital Video EEG

Due to the advancement in the technology, VCR tape recorders were replaced by digital videos in long term video EEG recording systems. In all the digital video EEG recording systems, currently the popularly used video format standard is Moving Picture Experts Group (MPEG). It is used as the standard video format because it can accommodate larger frame sizes. The disk space is saved by using 8:1 and 30:1 compression factors. More compression factor as 100:1 may degrade the video quality. Also, when the features as well as the motion of the video increases, only less compression can be obtained. The range of compression can be changed by a factor of two or more [12]. In MPEG2, the compression range will be in the range between 3.4 and 4.5 Mbps. Other than MPEG2, other standard MPEG formats are MPEG4, MPEG7, and MPEG21.

Table 2 shows the total amount of video EEG data stored in a DVD disk [13]. For example, if 128 EEG channels are used with a sampling rate of 1000, then it would generate a data of 256 kilo bytes per second (Kbps). Therefore, one DVD of standard size 4.7 Gbps is required to store the video EEG data captured at 256 Kbps rate.

Table 2 Amount of data generated and stored in a DVD disk

Number of channels	Compression standard	Sampling rate	Kilo bytes per second	Data generated per day (GB)	Storage capacity in DVD disks (GB)
128	–	200	51.2	4. 5	0.95
128	MPEG1 with 1.5 Mbps	1000	256	22	4.7
128	MPEG2 with 4.5 Mbps	1000	818	70	15

5.1 Data Transfer and Storage

The actual size of the data generated per day will vary as the file size is dependent upon the parameters like number of samples/second per channel, storage format of the system and the rate of compression. For example, a monitoring unit which has 64 channels operating for 24 hour per day will generate 2.5 megabytes of data per minute with 200 samples per second/channel. The data produced by 16 number of channels will be calculated as: 6 pages/minute $\times$ (24 $\times$ 60) minutes/day $\times$ 365 days/year. Therefore, 32 channels would produce double the data of 16 channels and henceforth 64 channels may generate double the data of 32 channels (i.e. 12,000,000 pages of 16 channels EEG paper/year). However, it is worthy to consider the data in its digital capacity too. For example, if the EEG data recorded for seizure activity is 6–8 min and 64 channels are used, then the total amount of data occupied by the disk space would be around 4.2–12 Mbps.

To store 10 Mbps of data, it would require 5000 pages of text. One CDROM can accommodate nearly 50–60 EEG routine data. To handle the large amount of data, careful management of the data is required. On the other hand, a DVD can store nearly 4 GB of data thus reducing the burden of storage. In addition to this, digital EEG/video EEG can also cut the storage because a digital EEG recording uses 128 channels at a sampling frequency of at least 1 kHz.

As per the traditional EEG recording, data review and interpretation generally occur separately from the data acquisition process. This allows additional records to be obtained while the previous records are being interpreted by the EEG and allows a more efficient segmentation of functions.

6 Big Data in Epileptic EEG Analysis

Continuous EEG signals having high resolution has obstacles such as its quantity and rate of the dataset, and data generated from various monitors per patient. EEG signals also faces problems related to the complexity of spatio-temporal nature. Due

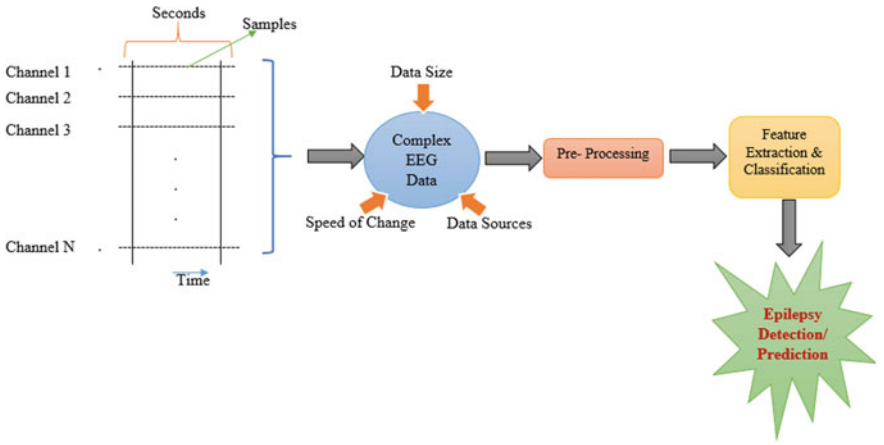


Fig. 2 Epileptic seizure detection/prediction system using big data analytics

to this, ensuring effectiveness of seizure detection/prediction systems will be difficult and hence the robustness of the continuous monitoring systems [14].

At present, there are numerous epilepsy detection systems available which utilises the change in formation from any one source of biological variation. These systems act as an alert device especially in emergency conditions. Figure 2 shows the generic epileptic seizure detection/prediction system using big data analytics. But these systems tend to be unreliable, specifically in real time epileptic monitoring systems. They can even tend to give false alarms and this may cause huge discomfort for both patients and their care takers.

These systems may not work accurately due to the lack of prior medical knowledge and analysis is done by using single biological signal's variations. There is a demand for versatile approach which can correlate these biological changes very well in the multi-channel time series EEG data [15].

Since huge amount of continuous data along with the other patient's information are generated in clinics, advanced storage systems are of great importance. With the ability to store the data, retrieving it is time consuming and expensive, big data employability facilitates rapid retrieval of huge volumes of such data and delivers with analytical modelling. Systems such as Hadoop, MapReduce and MongoDBis are commonly used in the research field [16]. MongoDB is a table based document database system. In addition to it, it is difficult, if the continuous databases are to be integrated in a single platform. MongoDB kind of database management systems can be used in these criteria and renders with good performance measure [17]. On the other hand, Apache Hadoop enables distributed data base processing in huge quantity enabling program based models. Apache Hadoop based platform is a highly scalable kind with computing modules like Spark and MapReduce. Spark platform has streaming of continuous data capability along with graphics and machine learning tools. This facility enables them to be used in the analysis of continuous epileptic

EEG signals. This in addition enables the clinical researchers to utilise data for the analysis especially in real time, effectively.

Big data based EEG processing for the development of decision support system in clinics is becoming widespread [18]. This enables an epileptologist with the knowledge of information about the patient and artifacts removed signals. These signals are delivered well in time with the improvements in patient care systems. One of the vital reasons for the development of such systems are due to the production of huge amount of EEG data in clinics per patient. Therefore, decision support systems in clinics is highly appreciated by the researcher's community.

7 Conclusion

EEG data showcases a huge variant information including data generated by movement of muscles, video information, instruments interferences, etc. the wide range of these measures, the quantity of data along with its diverse types has made the epileptic seizure detection or prediction a big data problem. Presently, improvements in data storage presents, an alternate solution while handling huge amount of data in diverse storage media. But, some storage systems are limited by the diversity of the data types. Unwanted reprocessing of the data is limited by the increased use of metadata and recording timestamps on various time, even though the past results are saved in the system level. Hence, the data model used, should be able to provide results in big data by enabling researchers to perform interpretative analysis without losing control over data. This analysis will help to acquire new insight from them. Also, while incorporating big data in the monitoring medical devices, the performance metrics has to be considered along with its time and portability. Real time epileptic detection systems can be thought in future for handling big data of EEG.

References

1. Borckardt, J.J., Nash, M.R., Murphy, M.D., Moore, M., Shaw, D., O'Neil, P.: Clinical practice as natural laboratory for psychotherapy research: a guide to case-based time-series analysis. Am. Psychol. **63**(2), 77–95 (2008)
2. Lynch, C.: Big data: how do your data grow? Nature **455**(7209), 28–29 (2008)
3. Thuraisingham, R.A., Gottwald, G.A.: On multiscale entropy analysis for physiological data. Phys. A Stat. Mech. Appl. **366**, 323–332 (2006)
4. Cuzzocrea, A., Song, I.-Y., Davis, K.C.: Analytics over large-scale multidimensional data: the big data revolution. In: Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)
5. Kaisler, S., Armour, F., Espinosa, J.A., Money, W.: Big data: issues and challenges moving forward. In: 46th Hawaii International Conference on System Sciences, pp. 995–1004 (2013)
6. Lee, J., Mark, R.G.: A hypotensive episode predictor for intensive care based on heart rate and blood pressure time series. In: Computing in Cardiology, pp. 81–84. IEEE (2010)
7. Heit, E.: Brain imaging, forward inference, and theories of reasoning. Rontiers Human Neurosci. **8**, Article 1056 (2015)

8. McCullough, J.S., Casey, M., Moscovice, I., Prasad, S.: The effect of health information technology on quality in U.S. hospitals. Health Aff. **29**(4), 647–654 (2010)
9. Chen, J., Dougherty, E., Demir, S.S., Friedman, C.P., Li, C.S., Wong, S.: Grand challenges for multimodal bio-medical systems. IEEE Circuits Syst. Mag. **5**(2), 46–52 (2005)
10. Gotman, J.: Automatic recognition of epileptic seizures in the EEG. Electroencephalogr. Clin. Neurophysiol. **54**(5), 530–540 (1982)
11. Lesser, M.P., et al.: Bleaching in coral reef anthozoans: effects of irradiance, ultraviolet radiation and temperature on the activities of protective enzymes against active oxygen. Coral Reefs **8**(4), 225–232 (1990)
12. Fogg, C., et al.: MPEG Video Compression Standard, 1st ed. Springer, US, ISBN: 978-0-412-08771-4 (1996). https://doi.org/10.1007/b115884
13. Niedermeyer, E., Lopes, F.: Electroencephalography: Basic Principles, Clinical Applications and Related Fields, 5th ed. Lippincott Williams & Wilkins Publishers (2005)
14. Drew, B.J., Harris, P., Zègre-Hemsey, J.K., et al.: Insights in to the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. PLoS ONE **9**(10), Article IDe 110274 (2014)
15. Carayon, P.: Human factors of complex sociotechnical systems. Appl. Ergon. **37**(4), 525–535 (2006)
16. Kaur, K., Rani, R.: Managing data in healthcare information systems: many models, one solution. Computer **48**(3), 52–59 (2015)
17. Yu, W.D., Kollipara, M., Penmetsa, R., Elliadka, S.: A distributed storage solution for cloud based e-healthcare information system. In: Proceedings of the IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom'13), pp. 476–480, Lisbon, Portugal (2013)
18. Belle, A., Kon, M.A., Najarian, K.: Biomedical informatics for computer-aided decision support systems: a survey. Sci. World J. **2013**, Article ID 769639 (8 pages) (2013)

Big Data Analytics in Healthcare Using Spreadsheets

Samaya Pillai Iyengar, Haridas Acharya and Manik Kadam

Abstract Advancements and sophistication in Information Technology systems have deeply impacted business operations and strategy across all domains. In a highly competitive business environment during current times, the success and growth of business have become largely dependent and highly critical with respect of business analytics and intelligence for crucial decision making. While this is true with almost all business sectors, the healthcare systems are no exception. In the information era, big data and its related cloud computing technologies form the backbone towards a data-centric ecosystem. A shift is observed from processes/methodology to data as a valuable resource, for it can neither be neglected nor disposing irresponsibly. The health ecosystem encompasses numerous stakeholders involving doctors, clinicians, patients, pharmacy on one hand while the investors and public institutions on the other at large. Systematic monitoring, processing, analysis and reporting of sensitive health data calls for a robust information technological support across its products and processes. Business analytics is an inevitable business practice in current competitive environment wherein the processes, practices, people skills and performance are key success factors for exploration and strategizing for effective and efficient operational and firm performances. This chapter discusses big data analytics, its need and methods with special reference to the healthcare industry which could help practitioners and policy makers in this domain to develop strategies for better healthcare systems and management. *Methodology* The chapter discusses the WHO health framework. It categorizes the Healthcare data. Methods of addressing Analysis of Big data techniques is done by discussing special tools and technologies like Hadoop and then Spreadsheets. The analysis of Big Data and its sub-component, viz. Structured, Semi-Structured and unstructured is discussed in detail. In Structured, analysis of simple numeric data is investigated. In unstructured, an overview of text analysis, NLP and

S. P. Iyengar (✉)
Symbiosis Institute of Telecom Management, Symbiosis International University, Pune, India
e-mail: samayapillai@gmail.com

H. Acharya · M. Kadam
Allana Institute of Management Sciences, Savitribai Phule Pune University, Pune, India
e-mail: haridas.undri@gmail.com

M. Kadam
e-mail: Msk16121612@gmail.com

A. J. Kulkarni et al. (eds.), *Big Data Analytics in Healthcare*,
Studies in Big Data 66, https://doi.org/10.1007/978-3-030-31672-3_9

one algorithms of NLP is done. In Unstructured, Image analysis, methods of pulling a sub-part of big data from cloud is discussed.

Keywords Healthcare · Big data · Cloud

1 Introduction

Big data analytics encompass complex procedures concerning scrutiny of enormous and diverse data sets also popularly identified as "big data". It involves revealing meaningful information to discover hidden patterns, unidentified correlations, market leanings, customer preferences, and other useful information such that business entities can make informed and better decisions for their operations and strategy. Integration of big data analytics tools and techniques has attained abundant prominence, acceptance, and is constantly increasing during current times to unearth valuable information from their big data repositories in favor of their businesses and holistic benefit of all stakeholders.

Big data irrespective of whether structured or unstructured, is being deeply manipulated using multiple analytical techniques for value add objective outcomes for business, such as higher operational and firm performances, cost efficiencies, and ultimately a competitive advantage in a complex and volatile economic environment. The power of big data analytics can be evidenced from a sample example such as "Hadoop" which is an open-source framework. It is generally used for massive data storage of multiple kinds, and also executing applications on several clusters of commodity hardware. It has huge processing power and the capacity to handle virtually limitless concurrent tasks or jobs. Big data promises a great future with growing data sets necessitating applications to transform into real-time and processing operations migrating to cloud systems. Also, several technologies are now available to handle storage and query/analysis of big data. In-depth mining of big data sets especially those of customer preferences, have assisted businesses managers better understand customer behaviors, and there is growing keenness to expand their historical datasets with social media data, text analysis, sensor data and browser log data which are crucial and precious material in accurately forecasting their customer choices.

Conclusively big data reposes to efficiently and effectively handle volume, velocity, variety, and veracity thereby returning high value add to corporate performance dimensions. Alternately, we can also experience that while managing and maintenance of big data analytical systems could be a costly affair with dynamic returns, it is also possible that simple processing applications such as spreadsheets or MS-Office (Excel) which is a spreadsheet software can also assist with reliably manageable datasets to provide quality and meaningful information for decision making. This chapter attempts to investigate the power of spreadsheets through the use of its functions and add-ins/plug-ins that can assist and process big data sets to provide convincing and meaningful information patterns. It is worth exploring a business domain

such as the "healthcare" through implementing the power of spreadsheet software tools/techniques to process transactional datasets and synthesizing its derived reports to assess its efficacy and acceptance as an alternate mechanism for big data analytics.

2 Integrated Health Service Delivery Systems and WHO Framework

One of the critical challenges for nations towards advancing a universal and integrated health service delivery system is the increasing need for equitable access to such quality services at reasonable/affordable costs. Populations are continuously aging and there is growing prevalence of several communication or non-communicable diseases, and hence the objective outcomes of integrated health service delivery systems need to aim at people-centric health services. In this context, the World Health Organization (WHO) in 2018 developed a regional action framework formulating a range of system and facility level action/policies for adaptation in specific contexts in order to improvise hospital management and planning strategies linking to primary healthcare.

The framework also suggests that hospitals serve the universal health policy goals, improve accountability, equity, quality, efficiency, resilience and sustainability while using regulatory, monitoring, and system design tools. Deciphering the "WHO Health Systems Framework" (refer Fig. 1), the systems blocks can be categorized into three types namely focusing upon (a) Data Source/Generation, (b) Data Analytics, and (c) Data Inference/Decision making.

Data Source/Generation

The blocks "Healthcare Financing", "Health workforce", "Medical products, technologies" and "Service Delivery" are primary activities which lead to lot of data

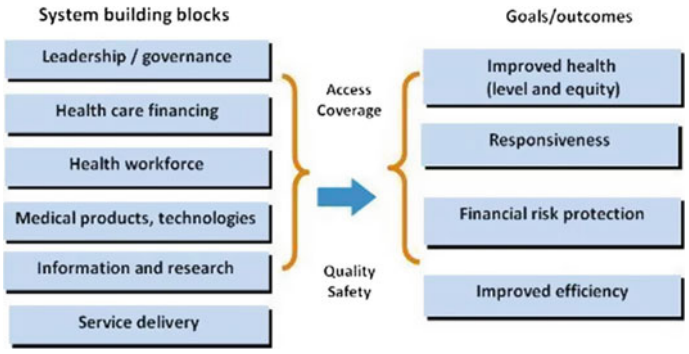


Fig. 1 The WHO health systems framework (*source* WHO web portal)

generation. Big data generated by public organizations is shared and readily available for students to learn analytics. A separate section on shared data sources is given at the end of this chapter as Appendix.

Data Analytics

Data which needs to be analyzed, such as "Healthcare financing", "Information and Research" and "Service Delivery" are considered here. "Healthcare Financing"—transactional data from in-house-hospital i.e. "Health Workforce" and "Service Delivery" needs to analyzed. "Information Research" are transactional data dealing with government policies, pharmaceutical organizations, etc. which form the datasets for analysis.

Data Inference/Decision Making

The blocks on "Leadership/governance" and "Information Research" can enhance decision making. A well-performing health workforce is the one which are quick to respond, unbiased and proficient in their functionality and can achieve better health-outcomes. This is in reference to given available resources and circumstances. A good operating health information structure or system is the one which can guarantee the creation, investigation, propagation and usage of reliable and timely information on health systems performance, health determinants, and health status.

Other Aspects of Integrating Technology with Health Service Delivery Systems

In recent times, several disruptive technologies have significantly impacted the healthcare systems, such as:

1. Internet of Things (IoT) and wearables (technology gadgets)
2. Artificial Intelligence (AI)
3. Telemedicine
4. Cloud Technology
5. Big Data

While there are several other innovative technological disruptions and challenges, we will however focus specifically on the effects of cloud technology and big data only as these are the primary focus of this chapter.

Cloud Computing

Cloud computing is defined as the applications and services that execute on a distributed network using virtualized resources and used by common Internet protocols and networking standards (Sosinsky, B. (2010). Cloud computing bible (Vol. 762). John Wiley & Sons). The concept and technology of Cloud computing has been centered on technological practices of harnessing the power of several remote servers networked and hosted over the internet rather than a personal computer or local server. These provide abundance of computer system resources (namely tools, applications, data storage, servers, databases, networking, and software) available for users to

store, manage, and process data without any active direct management by the end users. They also act as data centers for a host of users across the internet. Cloud computing relies upon high end technologies such as infrastructure as a service (IaaS), software as a service (SaaS), and platform as a service (PaaS) or even Database, Storage, Process, Information, Application, Integration, Management, Security, and Testing-as-a-service.

Big Data

Big data as the term implies are data resources with enormous sizes and types beyond compare to any traditional relational databases. Big data usually engages several techniques to systematically analyze, extract, or process data sets that are quite complex and voluminous and cannot be compared with any traditional data-processing application software. Big data can be analyze computationally in order to reveal intricate data/information patterns, trends, and associations, especially relating to human interactions and behavior. The key characteristics of big data include high volume, velocity, variety, and veracity. These in turn generate high value to the business.

3 Nature of Healthcare Data

To comprehend the nature of healthcare data it is beneficial to understand how data is generated in Healthcare: Referring to the WHO framework, the 2nd Block "Healthcare Financing", 3rd Block "Health workforce", 4th Block "Medical products, technologies" and 6th block "Service Delivery" are the points in healthcare which generate data in healthcare.

Def.: Healthcare Data:	The data in healthcare is a combination of {Structured Data, Semi structured Data, Unstructured Data}.
Def.: Structured Data:	The data that has a structure and is generated by SQL-based transaction processing applications. This is further stored in well-organized containers like Relational Database tables.
Def.: Semi Structured Data:	Data cannot be stored in Relational database tables without processing. E.g. XML, JSON files.
Def.: Un-Structured Data:	The data is unorganized. It refers to data that cannot be stored in the traditional row and column structure of relational databases tables. E.g.: emails, videos, Photographs, audio files, web pages, and social media messages.

4 Tools and Technologies Used for Big Data

We classify the technologies used for Big Data Analysis into two as follows: (a) The Hadoop Framework and its constituents (b) The Spreadsheets and Add Ins.

(a) Hadoop: Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs (Fig. 2).

Constituents of Hadoop.

4.1 Hadoop Core Components

The Hadoop system consists of 4 fundamental components.

The Hadoop Common-Apache Foundation has proposed a set of libraries and utilities. These can be utilized by other components of the HS (Hadoop system).

(1) Hadoop Distributed File System (HDFS)
(2) Replicas of HDFS component for the purpose of reliable and quick data access.
(3) Map Reduce—Framework of Apache Hadoop for Data Processing
(4) YARN-YARN is a dynamic resource utilization on Hadoop framework (Table 1).

Pig and Hive. Pig—provides a high level data flow language, Hive—is put together on top of Hadoop and provides a simple query language known as HiveQL. They are the Data Access Components.

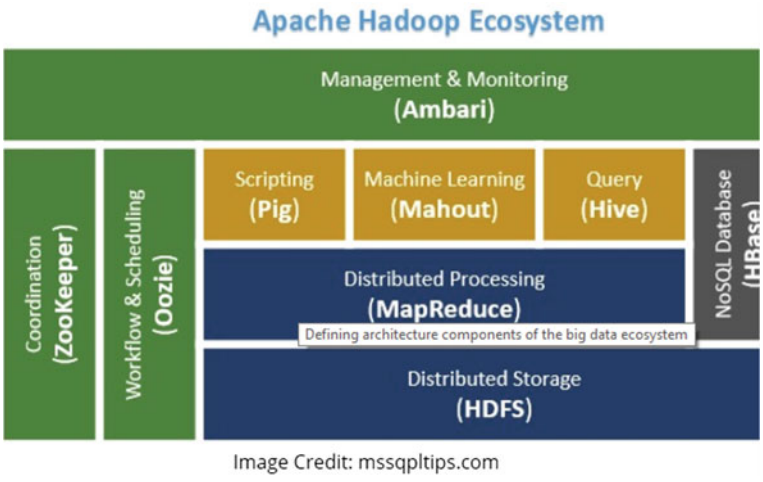


Fig. 2 Components of Hadoop

Table 1 The components of Hadoop ecosystem are

Purpose	Component name
Access	Pig, Hive, HiveQL
Integration	Sqoop and Flume
Storage	HBase
Monitoring, management and orchestration	Oozie and Zookeeper
Data storage component	*HBase*
Monitoring, management and orchestration	Oozie and Zookeeper
Web user interface	*Apache AMBARI*
Machine learning algorithms	*Apache MAHOUT*
Feeds of messages	*Apache Kafka*

Sqoop and Flume. The Sqoop element is used for importing data from external sources into HDFS, HBase or Hive. It is also be used the reverse task of exporting data from Hadoop.

Flume is utilized to collect and aggregate big quantities of data. Apache Flume is utilized for gathering data from its origin and transferring it back to the resting location (HDFS). These are considered as the Data Integration Components.

HBase—HBase is a column-oriented database that utilizes HDFS for underlying storage of data. IT helps in carrying out random reads and batch computations using MapReduce. This is Data Storage Component.
Oozie and Zookeeper—Oozie-Oozie is a workflow scheduler. Here the workflow content is represented as Directed Acyclic Graphs. The workflow in Oozie is conducted based on data and time dependencies.

Zookeeper-Zookeeper provides reliable, easy, simple, fast, and ordered operational services for a Hadoop cluster. It is accountable for distributed configuration service, synchronization service and supplementing a naming registry for distributed systems. These perform Monitoring, Management and Orchestration.

Apache AMBARI—It provides easy to use web user interface for Hadoop management.
Apache MAHOUT—It is a crucial component for machine learning. It implies several, varied machine learning algorithms.

Cloudera Inc. was founded by big data geniuses from Facebook, Google, Oracle and Yahoo in 2008. It was the first company to develop and distribute Apache Hadoop-based software. Hortonworks, founded in 2011, is a leading vendor of Hadoop. It provides analyzing, storing and managing big data. It is an open source platform based on Apache Hadoop.

In October 2018, both the companies Cloudera and Hortonworks announced that they would be entering into a merger.

4.2 Spreadsheets

A spreadsheet is also termed as "Table" or 2-dimensional structure. It's a document which is made up of rows and columns and their intersections are called cells. Spreadsheets are designed to ease the management of numbers and calculations. They calculate totals, averages, percentages, budgets, and complex financial and scientific formulas. Spreadsheet as an application has the ability to generate a million rows by 16,000 columns. This can be seen as a sufficient space to store text and data of this huge value. It can create and analyze statistical values such as the mean, standard error, percentile and rank. The spreadsheet also allows other word processor tasks like the expansion, merging of cells, rows and columns.

Few Significant Functions in Spreadsheets:
Vlookup & Hlookup

Both the above mentioned are termed as "lookup functions". It is useful for comparing values of data sets. Helps to retrieve data based on some common attribute existing in the same dataset, different dataset, physically present in one dataset, physically present in different datasets.

Pivot Table

Pivot tables helps to compare two different datasets. It helps to check if a particular value exists in the datasets. Analysis similar to Step reports or wise reports like for example:

Item-wise, Supplier-wise quantities; Zone-wise, region-person wise quantities, salesman wise etc.

Pivot is a table which allows arranging data as per requirement. It supports functions like count, average, sum etc. It performs other calculations according to the reference feature selected. It can convert a data table to inference table which is like a report. And which enables decision making.

Apart from the simple pivot tables we have the power pivot, power query, power maps. All these are assembled into one single tool by Microsoft. This is termed as "Power BI".

Statistical analysis in Excel can also be done by using the Analysis Tool Pack. This is a different and distinct Microsoft excels plugin. The Analysis Tool Pack reduces the time required to generate huge or small, complex statistical or engineering analysis. We can use the Tool pack to increase the efficiency. We need to give data and related parameters for the analysis. The tool will choose from the in-built, appropriate functions for calculating. It includes many tools, some of them are discussed here.

Anova: It stands for "analysis of variance". It provides several types of variance analysis. The factors and the samples decide which tool to use in the study.
Descriptive Statistics: This tool helps to generate various reports of univariate statistics. The reports can be charts, graphs.
Fourier Analysis: It is an analysis method used majorly in engineering. It solves problems in topics like linear systems. It also helps to analyze periodic data.

5 Preference of Doctors and Researchers

It is motivating to note that, in spite of an offshoot foray of several tools for visualization and analytics, the preference in healthcare organizations and the stakeholders still seems to be Spread Sheet Centric.

One in four healthcare organizations still practice some form of the core Excel or spreadsheet to do their major tasks like reporting on their clinical and financial analytics. This is stated as per a new survey from TCS Healthcare Technologies. This is found in spite of a health IT market crammed with corporations contributing analytical software which also provide visual analysis. Dashboards, analytics-as-a-service packages, and cloud-based databanks. It reflected that 39% of organizations still use Excel, Crystal Reports was utilized by 20% of respondents, and 17% preferred Access. Only 25% of providers are currently using predictive modeling applications, though that market is expected to grow as organizations shift their focus away from EHR implementation on to the medical and economic optimization tools.

The major stakeholders like the doctors, nurses and admin staff have always shown an inclination towards tools they are conversant with, rather than learning new tools in reference to the patient data management.

6 Rationale for Using Spreadsheets

Analytics has caught the imagination of the business world like a fire. Analytics involves skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. There are special tools, frameworks in the market designed for Analytics like Hadoop. But, for stakeholders to use these new tools and frameworks requires training, time and patience. And even after the training, the preference of the stakeholders may not lie in using these tools. In healthcare the stakeholder's preference goes for the age old tools like spreadsheet.

Since most users are conversant with Spreadsheets, training them for big data analysis using spreadsheets would be easy. Since Spreadsheets comes both in open source and with vendor based soft wares like windows, no extra efforts on installations or configurations is required. They are much more cost effective than the huge frameworks. They would serve a good purpose of Analysis, and the actual data storage can be done using cloud services.

6.1 Current Market of Spreadsheet Upgrades Reflecting Big Data Add Ins

Since mid-1980s, spreadsheets have been the backbone in the end-user data management, which included gathering, studying, and dissemination of data. This was always owing to its ease of use, fast learning curve and the skill of the users to swiftly get reports without having to depend on the IT department.

Microsoft which has the Excel 2013 with data explorer supports big data. The product is a high-end dashboard which allows users to complete mashups of structured data from databases and unstructured data. It also supports to format data in the standard spreadsheet, summarize the data in visual display like bar charts and maps. IBM's BigSheets conglomerates Hadoop with any web-based frontend. It then permits the users to execute their own mashups of structured and unstructured data; the sources could be from single or multiple sources. Datameer found analytics solution for both structured and unstructured data in 2010. It then joined efforts with 1010 Data to provide spreadsheet analytics featuring up to a trillion rows per spreadsheet via a cloud-based service. Google's BigQuery (PDF link) allows businesses run SQL queries against billions of rows of data in the cloud and assimilates them.

As per Dresner Report (2017), 50% for the first time in 2017 stated of adopting Big Data, 36% of respondents say they may use big data in the future. Just 11% of respondents have "no plans to use big data at all". A survey from TCS Healthcare Technologies states that 39% of organizations use Excel to track and report on their clinical and financial analytics.

6.2 Logical Reasoning in Implementation of Spreadsheets for Big Data Operations

As discussed earlier, Big Data is huge data with variety, velocity, veracity and value. How to analyze this enormous bulk of data without using some software tools? Tools which may be costly and mostly managed by experts and requiring special training? We need not essentially need to—Instead, you can use the concept of "data samples". Data Samples is a subset of a huge "data population."

Consider a huge dataset, such as 10 million rows or more of patient visits for more than 10 years, both (in-patient-department) ipd and (out-patient-department) opd of a hospital. 20 million daily-test reports of patients. Suppose you want to examine and analyze this data.

How do we study this? It is highly impossible to analyze this huge data all at once. It has to done using either random samples or using sampling techniques and creating samples from the population of data.

For testing these samples, you do not have to work with a high-end software tools. These samples can be investigated and tested using spreadsheets is what the authors are claiming.

With these sample-testing using spreadsheets we can not only analyze the data for basic findings like descriptive statistics or more in-depth statistical tests, but also search for patterns like relations, collections or groups, trends, differences or anything else that might be of concern to you. This can be done by posing at least three legitimate questions as per the concept:

1. How many records are required to have a sample which are required w to make accurate estimates?
2. How to extract random records from the main dataset?
3. How to check the reliability of data?

The confidence level tells us that if we extract 100 random samples of 66,327 records each from the same population, 99 samples may be assumed to reproduce the underlying characteristics of the dataset they come from. The 0.5% error level says the values we obtain should be read in the plus or minus 0.5% interval, for instance after transforming the records in contingency tables.

Findings of Spreadsheets being used as major tool in Analytics of Fig. 3.

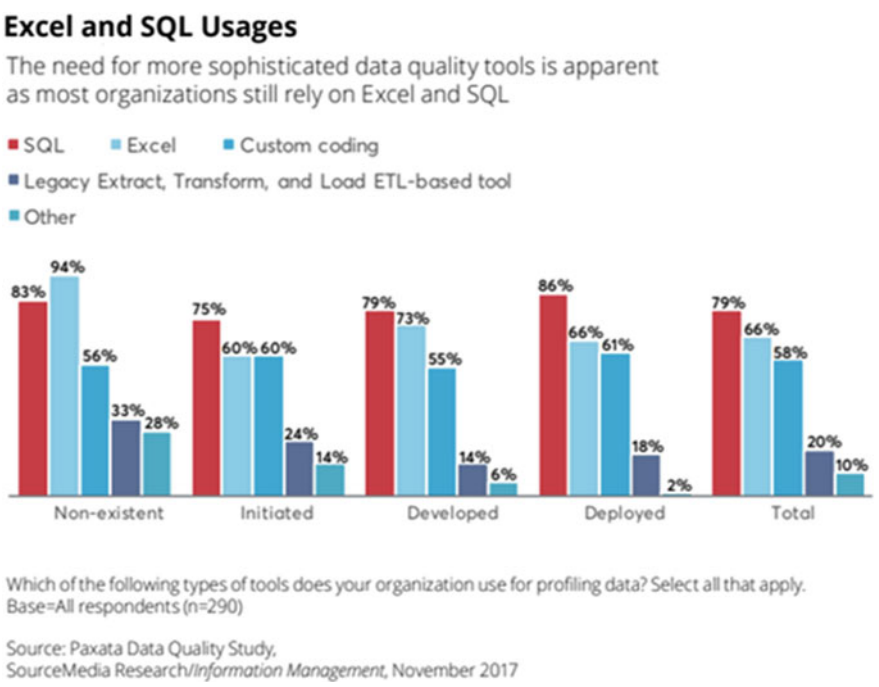


Fig. 3 Excel and SQL usages

7 Does Healthcare Data Validate Itself to Become "Big Data"?

Referring the WHO framework when we consider the Data Source/Generation: The 2nd Block "Healthcare Financing", 3rd Block "Health workforce", 4th Block "Medical products, technologies" and sixth block "Service Delivery" can be categorized under this.

The Data Source has to have the following properties of Big data viz. the 5 V's Volume—scale of data, Velocity—Streaming in data (Analysis), Variety—Different forms of data, Veracity—uncertainty of data, and Value—giving the ultimate benefit and visualization to the organization.

To be defined under the category of big data, the data must have the above discussed properties like: Volume, Velocity, Variety, Veracity and Value.

If we map, the healthcare data to the above said properties we find some gaps. The healthcare data does not reflect. (a) Velocity—Speed of data generation per unit time. Here the data of from any of the above said sources is never streaming in and continuous. The data generated is deferred (b) The attribute Volume—if we look at the data in terms of the above sources, this property holds truth in few cases. But in most cases, the volume is not huge it is manageable. (c) Veracity—Data in healthcare is not uncertain or imprecise as compared to the data in social media.

The other two Vs i.e. Variety and Value are supported by the healthcare data. So this is a big Question to be asked, whether we give the same treatment to healthcare data as we give to other big data such as, social media, weather or traffic data?

Similarly when we analyse whether the healthcare data can be classified as under the aegis of structured/semi-structured or unstructured data, we get following answers:

In the overall data, some data is structured. There is some amount of data which can be defined as unstructured like images, scans, MRI's etc. And there is few datum which can be categorized as semi-structured like Prescriptions, reports of tests, diagnosis remarks, emails etc. i.e. the textual content which cannot be easily slotted to structured data.

As per this definition, the healthcare data does validate itself to be defined as Big Data.

Hence we can deductively conclude to state that some portion of the healthcare data can be classified as Big Data. It is the Images, Scans, MRIs etc. and the Semi-structured textual Data having a flexi-structure or no structure. Hence for healthcare we should be ideally performing the Un-structured data analysis.

Unstructured data is divided into Text data analysis and Image Data analysis. The semi-structured data as a thumb rule is considered as unstructured data, since it cannot be stored in the traditional containers i.e. the Database tables.

8 Big Data Analysis

Def.: Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights (www.sas.com) (Fig. 4).

Patient records, health plans, insurance information and other types of information can be difficult to manage—but are full of key insights once analytics are applied. That's why big data analytics technology is so important to healthcare. By analyzing large amounts of information—both structured and unstructured—quickly, healthcare providers can provide lifesaving diagnoses or treatment options almost immediately.

In the Healthcare domain there is a huge amount of unstructured data in the form of Medical Records, which represent Volume and Variety. The Electronic Health Records (EHR), Hospital Information Systems—which includes the Billing, inventory, Operations aspect and Financial Data can be grouped under Structured Data Analysis.

The Medical Reports, Emails, Written Communication etc. can be grouped as Semi-Structured or Unstructured Data Analysis (Text Data Analysis).

And the various Picture Archiving including various image reports, Sound Records can be grouped as Unstructured Data (Image Data Analysis).

Fig. 4 Significance of big data analytics (www.sas.com)

8.1 Structured Data Analysis

The data that has a structure and is well organized either in the form of tables or in some other way is known as structured data.

Lookup: When you use a lookup function in any Spread Sheet, you are basically saying, "Here's a value. Go to another location and find the same value. Then show me specific information related to that value."

VLOOKUP (search_key, range, index, [is_sorted])
HLOOKUP (search_key, range, index, [is_sorted])

Pivot Table: You can use pivot tables to narrow down a large data set or see relationships between data points. Cross tables are analyst's basic requirements.
Refer for Solved Examples to Appendix.
Refer solved examples which are executed on data which is not huge in volume, but the analysis can be mirrored to the huge, voluminous data. UnStructured Data Analysis. Refer Appendix.

8.2 Unstructured Data Analysis—Text Data Analysis

Def.: Semi Structured Data:	Semi-structured data is information that doesn't reside in a relational database. With some process you can store them in relation database (it could be very hard for some kind of semi structured data). It does have some organizational properties that make it easier to analyze. E.g. XML, JSON files.
Def.: Un-Structured Data:	The data that is unstructured or unorganized. It refers to data that doesn't fit neatly into the traditional row and column structure of relational databases. E.g. of unstructured data include: emails, videos, Photographs, audio files, web pages, and social media messages. Such Semi-structured or unstructured text data reanalysis is termed as "Text data Analysis".
Text data Analysis:	Uncover insights hidden in massive volumes of textual data using NLP i.e. Natural language Processing.

What is NLP?

Apart from an approach to communication, personal development, and psychotherapy (Neuro Linguistic Programming), NLP stands for Natural Language Processing, which is defined as the application of computational techniques to the analysis and synthesis of natural language and speech.

The input/output of a NLP system can be: (a) written text (b) speech.

To process written text, we need: lexical, syntactic, semantic knowledge about the language discourse information, real world knowledge.

To process spoken language, we need everything required to process written text, plus the challenges of speech recognition and speech synthesis.

Components of NLP: There are broadly two components of Natural Language Processing: (a) Natural Language Understanding (NLU): Here representing the given input in natural language into useful depictions is carried out. This helps to analyzing all facets of the language. (b) Natural Language Generation (NLG): Here the stress is on generating "meaning" from the input. This stage consists of: (i) Text planning: here the objective to identify relevance from the content. (ii) Sentence planning: Selection of word is carried out for Sentence generation. (iii) Text Realization: Here the plan is mapped to the structure in reference to the Sentence.

The Fundamentals:

The input to natural language processing is in the form of Unicode characters (generally UTF-8). (This is converted to lexical token which is made of phrases, words and syntactic markers).

Structure extraction—This process initially finds blocks of content based on tagging. It identifies and marks sentence, phrase, and paragraph boundaries—These can serve the purpose of a boundary or block for further analysis.
Language identification—This process detects which language is used for the entire document and for each paragraph or sentence. This is any human language. Language detectors are critical to determine what algorithms and dictionaries need to be applied to the text.
Tokenization—This process divides the character streams in tokens which may or may not be used for further processing and understanding. Token are numbers, words, identifiers or punctuation (depending on the use case).
Lemmatization/Stemming—This process decreases the disparities in words to much easier and simpler forms. This process may help to escalate the coverage of NLP utilities.
Decompounding—This process is used mainly in languages like Cyrillic, Germanic, Scandinavian languages. In such languages there is high usage of compound word. These compound words are required to be broken down in smaller parts to allow for accurate usage of NLP.
Entity extraction—This process helps first to identify the entities, and then to extract these entities. Entities are for example people, places, companies, etc. This is a crucial step to simplify the process of down streaming. This can be carried out in variety of ways like:
Regex extraction—This extraction method is better for ID numbers, phone numbers.
Dictionary extraction—This extraction method uses a dictionary made up of token of sequences. This works well for the known entities, such as employee name, department name etc.
Complex pattern-based extraction—This process works well for people names, business names and context-based extraction scenarios.

Statistical extraction—This process practices the statistical analysis to perform context extraction.
Phrase extraction—This process deducts sequences of tokens (phrases) that have a strong meaning. This sequence is independent of the words when treated separately.

Applications of NLP:

- Machine Translation—This is Translation between two natural languages.
- Information Retrieval—This is Web search. It can be uni-lingual or multi-lingual.
- Query Answering/Dialogue—Natural language interface with a database system, or a dialogue system.
- Report Generation—Generation of reports for healthcare as follows Patient details, patient bill, patient room etc.
- Some Small Applications—Grammar Checking, Spell Checking, Spell Corrector, Part-of-Speech (POS) Tagging, Lexical Processing, Word-level ambiguity, Syntactic Processing i.e. Parsing, Semantic Analysis.

Algorithm used in NLP: Latent Dirichlet Allocation (LDA): LDA or latent Dirichlet allocation is a "generative probabilistic model" of a collection of composites made up of parts.

In natural language processing, latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

Original LDA is unsupervised learning algorithm, while Labeled-LDA and Multi-Grain LDA, another topic model for classification and sentiment analysis, are supervised algorithm.

Why use LDA? If you view the number of topics as number of clusters and the probabilities as the proportion of cluster membership then using LDA is a way of soft clustering your composites and parts. LDA provides a more nuanced way of recommending similar items, finding duplicates, or discovering user profiles/personas.

Tools for text Analysis: Refer Appendix.

Examples of Unstructured Analysis: Refer the Appendix.

8.3 Unstructured Data Analysis—Image Data Analysis

Def.: Un-Structured Data: The data that is unstructured or unorganized. It refers to data that doesn't fit neatly into the traditional row and column structure of relational databases. E.g. of unstructured data include: videos, Photographs, Images.

Image processing is a method to convert an image into digital form. And to perform operations on this image so as to get an enhanced image. This is done in order to get an enhanced image or to extract some useful information from it. Usually

Image Processing system includes treating images as two dimensional signals while applying already set signal processing methods to them.

Applications

Visual information is the most important type of information perceived, processed and interpreted by the human brain. Digital image processing, as a computer-based technology, carries out automatic processing, manipulation and interpretation of visual information. It plays an important role in a wide variety of disciplines of science and technology, with applications such as photography, robotics, remote sensing, medical diagnosis.

1. Computerized photography (e.g., Photoshop)
2. Medical/Biological image processing (e.g., interpretation of X-ray images, Blood/cellular microscope images)
3. Automatic character recognition (zip code, license plate recognition)
4. Finger print/face/iris recognition.

Detailed Applications in Healthcare.

1. *Biomedical Imaging techniques*—The medical domain is undergoing tremendous changes due to the technological upgrades. In diagnosis, several types of imaging tools such as Ultrasound, X-ray, computer aided tomography (CT), 2D Echo etc. are used. Few applications of biomedical imaging are as discussed:
    a. Heart disease identification—There are at least 280 different types heart related ailments found. The diagnosis of heart diseases has undergone tremendous improvement with the help of image analysis techniques which are attached to the radiographic images.
    b. Lung disease identification—X-rays play a major role in lung related ailments. The X-ray film is a special type of film where the heart, the ribs, thoracic spine, and the diaphragm are clearly demarked. These X-rays help in further diagnosis.
    c. Digital mammograms—This technique is used in identification of breast tumour. Image processing techniques such as segmentation, shape analysis, contrast enhancement, feature extraction, etc. play a major role in this.

Image Analysis Using Excel (Spreadsheets):

Spreadsheet is not a traditional application or software for image processing. But since it can handle large amounts of data, graphic capabilities and due to its wide acceptance it can be used by applying mathematics in image data processing. It supports image pre-processing, image enhancement, image classification, analysis of change over time, and image data fusion.

This requires a simple graphical user interface (GUI) application was developed in MATLAB. That application, RGBExcel or the later RGB2X, extracts RGB image data from image files of any format and file size, and exports to Excel for processing.

Deployed as standalone applications, both versions can be installed on a 64-bit windows computer and run without MATLAB. (Larbi, P. A. (2018). Advancing Microsoft Excel's potential for teaching digital image processing and analysis.)

Although Excel does not currently have its own automated means of importing image data directly from an image file, the data can be copied from other resources. However, a more efficient means of acquiring the image data from multiple files all at the same time is by using the RGB2X 1.0 software application. RGB2X offers the advantages of multiple image file handling and creation of Excel files saved in the same location and having the same base name(s) as the original image file(s) (Larbi, P., & Larbi, D. (2018). Adopting Microsoft Excel for Biomedical Signal and Image Processing. In Computer Methods and Programs in Biomedical Signal and Image Processing. IntechOpen.)

Solved Examples of Unstructured Data Analysis and Tools for Image Processing. Refer Appendix **(Page 83).**

9 XLMiner

XLMiner: **XLMiner**™ is a comprehensive data mining add-in for Excel. **XLMiner** is a **tool** which supports analysis of data in variety of ways. It has extensively covers statistical as well as machine-learning techniques for further segmentation, classification, prediction and further analysis of given data. It is easy to work in it if you've prior knowledge of MS Excel or SPSS. **Tasks XLMiner can do**: XLMiner can do lot of things which can be done in any of the following languages: R, Python or Julia. Here we need not write extensive code but use the in-built libraries and functions. It offers a great deal in machine learning and data mining tasks. XLMiner supports Excel 2007, Excel 2010 and Excel 2013 (32-bit and 64-bit). XLMiner can perform Data Exploration and Visualization, Feature Engineering, Text Mining, Time Series Analysis, Machine Learning.

Note: It is not available for free. You can download it on 15 days trial period and later purchase two year license for $2495.

Note: Use the Google Drive from your mail-id. In this drive in the Google-Sheets use the option of Add-Ins for adding XLMiner or Google Analytics completely free of cost. Here we can do all the statistics required.

Examples of Unstructured Analysis—Refer the Appendix**. Here some free and open websites are referred, wherein big data is available which is in terabytes. Hence it cannot be reproduced on spreadsheets directly. These sites allow a user to select fields from the given source. Here the user can decide which fields and what sample size and download data into spreadsheets for further analysis**.

10 Conclusions and Future Scope

Big Data is a vast concept. The chapter gives an overview of the sub-topics. It gives an insight to readers about big data in healthcare. The available tools used for analysis. The chapter focusses more on spreadsheets rather than high-end frameworks for big data analysis. In each sub-topic i.e. structured—the chapter summarizes spreadsheet functions like vlookup, hlookup, pivot. In semi-structured—the chapter summarizes text analysis discusses an algorithm of text analysis (NLP) and in unstructured—Image analysis, the chapter discusses the significance of image analysis.

The reader can further explore each of these sub-topics independently. And each of these sub-topics can be discussed with case studies having statistical inferences.

Appendix

Structured Data Analysis: Solved Examples for Spreadsheets

(1) **V-Lookup**

Marks	Grade	Class	Grade	Class
60	C	First	A	Honors
67	C	First	B	Distinction
69	C	First	C	First
62	C	First	D	Pass+
45	E	Pass	E	Pass
55	D	Pass+	F	Fail
39	F	Fail		
99	A	Honors		
55	D	Pass+	**We are given a set a marks for 50 students, for whom we have to classify their class based on their grades. To achieve this task we did a simple if-else analysis function to find the grade first. Next we applied VLOOKUP function to derive the class based on their grade and classified the students accordingly**	
62	C	First		
98	A	Honors		
86	B	Distinction		
68	C	First		

(continued)

(continued)

Marks	Grade	Class	Grade	Class
38	F	Fail		
98	A	Honors		
68	C	First		
81	B	Distinction		
64	C	First		
56	C	First		
72	C	First		
71	C	First		
72	C	First		
55	D	Pass+		
93	A	Honors		
83	B	Distinction		
93	A	Honors		
78	B	Distinction		

Formulas used:
To allocated Grades:
=IF(A2 > 90,“A”,IF(A2 > 75,“B”,IF(A2 > 55,“C”,IF(A2 > 49,“D”,IF(A2 > 40,“E”,“F”)))))
To allocate class:
=VLOOKUP(B2,$H$1:$I$7,2,0)

(2) **H-Lookup**

Student	A	B	C	D	E	F	G
Subject 1	44	35	31	42	25	39	25
Subject 2	43	45	29	44	26	26	30
Subject 3	33	29	47	32	28	30	31
Subject 4	38	43	28	31	25	31	46
Subject 5	46	26	38	43	26	33	25
Subject 2							
43	45	29	44	26	26	30	

=HLOOKUP(C2,$B$2:$I$7,3,FALSE)

(3) Discussion: The systolic blood pressure was measured for 30 people of different ages. A nonzero intercept seems appropriate here, since even a very young person can have a high blood pressure. There are 30 rows of data.

Age	Systolic blood pressure
17	114
19	124
20	116
21	120
25	125
29	130
34	110
36	136
39	144
39	120
42	124
42	128
44	160
45	138
45	135
46	142
47	220
47	145
48	130
50	142
53	158
56	154
56	150
59	140
63	144
64	162
65	162
67	170
67	158
69	175

Analysis:
Since we have been given two variables namely Age and Systolic Blood Pressure, we try to find if a relationship exists between the two. It is generally used when we

wish to predict one value based on other value. The variable we wish to predict is called the dependent variable (or sometimes, the outcome variable). The variable we are using to predict the other variable's value is called the independent variable (or sometimes, the predictor variable).

Taking, SBP as dependent variable and age as an independent variable, we carry out the regression and get the following outputs:

Variable details			
Model	Variables entered	Variables removed	Method
1	Age[b]		Enter

[a]Dependent variable: SBP
[b]All requested variables entered

Model summary[b]				
Model	R	R square	Adjusted R square	Std. error of the estimate
1	0.658[a]	0.432	0.412	17.31375

[a]Predictors: (constant), age
[b]Dependent variable: SBP

This table provides the R and R^2 values. The R value is 0.658 and it represents the simple correlation (the "**R**" Column). The value indicates a high degree of correlation. The R^2 value (the "**R Square**" column) indicates how much of the total variation in the dependent variable, SBP, can be explained by the independent variable, Age. In this case, 43.2% can be explained.

ANOVA[a]						
Model		Sum of squares	df	Mean square	F	Sig.
1	Regression	6394.023	1	6394.023	21.330	0.000[b]
	Residual	8393.444	28	299.766		
	Total	14,787.467	29			

[a]Dependent variable: SBP
[b]Predictors: (constant), age

This table indicates that the regression model predicts the dependent variable significantly well. We need to check "**Regression**" row and go to the "**Sig.**" column. The statistical significance of the regression model can be seen here. Here, $P < 0.0005$, which is less than 0.05, and indicates that, overall, the regression model statistically and significantly predicts the outcome variable (i.e., it is a good fit for the data).

The **Coefficients** table provides us with the necessary information to predict SBP from age as well as determine whether age contributes statistically significantly to the model (by looking at the "**Sig.**" column).

Coefficients[a]						
Model		Unstandardized coefficients		Standardized coefficients	t	Sig.
		B	Std. error	Beta		
1	(Constant)	98.715	10.000		9.871	0.000
	Age	0.971	0.210	0.658	4.618	0.000

Coefficients[a]			
Model		95.0% confidence interval for B	
		Lower bound	Upper bound
1	(Constant)	78.230	119.200
	Age	0.540	1.401

[a]Dependent variable: SBP

to present the regression equation as:

$$\mathbf{SBP} = \mathbf{98.715} + \mathbf{0.971}(\mathbf{Age})$$

Solved Examples for Structured Data.

(1) A study tested whether cholesterol was reduced after using a certain brand of margarine as part of a low fat, low cholesterol diet. The subjects consumed on average 2.31 g of the active ingredient, stanol easter, a day. This data set contains information on 18 people using margarine to reduce cholesterol over three time points. The data set can be used to demonstrate paired t-tests, repeated measures ANOVA and a mixed between-within ANOVA using the final variable. The dataset is also good for discussion about meaningful differences as the difference between weeks 4 and 8 is very small but significant.

ID	Before	After 4 weeks	After 8 weeks	Margarine
2	6.76	6.2	6.13	A
4	4.8	4.27	4.15	A
6	7.49	7.12	7.05	A

(continued)

(continued)

ID	Before	After 4 weeks	After 8 weeks	Margarine
8	5.05	4.63	4.67	A
10	3.91	3.7	3.66	A
13	6.17	5.56	5.51	A
14	7.67	7.11	6.96	A
15	7.34	6.84	6.82	A
17	5.13	4.52	4.45	A
1	6.42	5.83	5.75	B
3	6.56	5.83	5.71	B
5	8.43	7.71	7.67	B
7	8.05	7.25	7.1	B
9	5.77	5.31	5.33	B
11	6.77	6.15	5.96	B
12	6.44	5.59	5.64	B
16	6.85	6.4	6.29	B
18	5.73	5.13	5.17	B

Solution.
Margarine type A

ID	Before	After 4 weeks	After 8 weeks	Margarine
2	6.76	6.2	6.13	A
4	4.8	4.27	4.15	A
6	7.49	7.12	7.05	A
8	5.05	4.63	4.67	A
10	3.91	3.7	3.66	A
13	6.17	5.56	5.51	A
14	7.67	7.11	6.96	A
15	7.34	6.84	6.82	A
17	5.13	4.52	4.45	A

Paired t-test for before and after 4 weeks for A		
t-Test: paired two sample for means		
	Variable 1	*Variable 2*
Mean	6.035555556	5.55

(continued)

(continued)

Paired t-test for before and after 4 weeks for A		
t-Test: paired two sample for means		
Variance	1.858402778	1.744175
Observations	9	9
Pearson correlation	0.995718578	
Hypothesized mean difference	0	
Df	8	
t Stat	11.09802124	
P(T <= t) one-tail	1.93991E−06	
t Critical one-tail	1.859548038	
P(T <= t) two-tail	**3.87982E−06**	
t Critical two-tail	2.306004135	

Looking at the table above, we find that *P*-value is negative and thus it is less than alpha value 0.05. Which also means that we reject null hypothesis and accept alternate one. We infer that cholesterol level before and after 4 weeks in type A margarine is significantly different.

Paired sample t-test for before and after 8 weeks for A		
t-Test: paired two sample for means		
	Variable 1	*Variable 2*
Mean	6.035555556	5.488888889
Variance	1.858402778	1.708986111
Observations	9	9
Pearson correlation	0.993766013	
Hypothesized mean difference	0	
Df	8	
t Stat	10.30041841	
P(T <= t) one-tail	3.40115E−06	
t Critical one-tail	1.859548038	
P(T <= t) two-tail	**6.80229E−06**	
t Critical two-tail	2.306004135	

Looking at the table above, we find that *P*-value is negative and much lower than 0.05. Hence, we accept alternate hypothesis to infer that yes there is difference in cholesterol levels.

Margarine type B

ID	Before	After 4 weeks	After 8 weeks	Margarine
1	6.42	5.83	5.75	B
3	6.56	5.83	5.71	B
5	8.43	7.71	7.67	B
7	8.05	7.25	7.1	B
9	5.77	5.31	5.33	B
11	6.77	6.15	5.96	B
12	6.44	5.59	5.64	B
16	6.85	6.4	6.29	B
18	5.73	5.13	5.17	B

Paired t-test for before and after 4 weeks for B		
t-Test: paired two sample for means		
	Variable 1	*Variable 2*
Mean	6.78	6.133333333
Variance	0.844575	0.746
Observations	9	9
Pearson correlation	0.989579645	
Hypothesized mean difference	0	
Df	8	
t Stat	13.85714286	
P(T <= t) one-tail	3.55598E-07	
t Critical one-tail	1.859548038	
P(T <= t) two-tail	**7.11197E−07**	
t Critical two-tail	2.306004135	

Looking at the table above, we find that *P* value is very much low than 0.05 thus we accept alternate hypothesis. It also means that both values for cholesterol are not same.

Paired t-test for before and after 8 weeks for B		
t-Test: paired two sample for means		
	Variable 1	*Variable 2*
Mean	6.78	6.068888889
Variance	0.844575	0.681986111
Observations	9	9

(continued)

(continued)

Paired t-test for before and after 8 weeks for B		
t-Test: paired two sample for means		
Pearson correlation	0.98782641	
Hypothesized mean difference	0	
df	8	
t Stat	12.9444903	
P(T <= t) one-tail	6.0057E−07	
t critical one-tail	1.85954804	
P(T <= t) two-tail	**1.2011E−06**	
t critical two-tail	2.30600414	

Same test between two variables Before and after 8 weeks *P* value which is lower than 0.05 is a clear indicator of difference between cholesterol levels at both times.

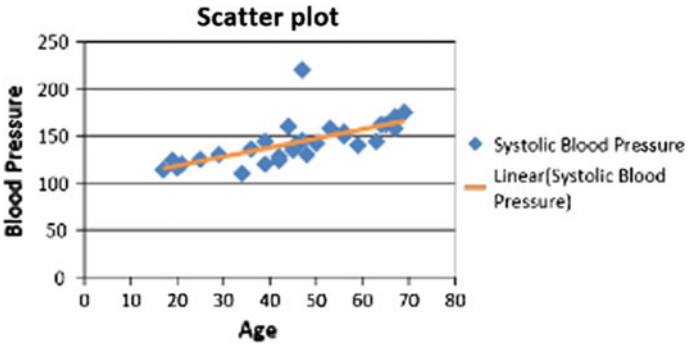
(2) **Regression Analysis: Numeric Healthcare Data Analysis**

Index	Age	Systolic blood pressure	Index	Age	Systolic blood pressure
1	39	144	16	48	130
2	47	220	17	45	135
3	45	138	18	17	114
4	47	145	19	20	116
5	65	162	20	19	124
6	46	142	21	36	136
7	67	170	22	50	142
8	42	124	23	39	120
9	67	158	24	21	120
10	56	154	25	44	160
11	64	162	26	53	158
12	56	150	27	63	144
13	59	140	28	29	130
14	34	110	29	25	125
15	42	128	30	69	175

Data Given—The systolic blood pressure was measured for 30 people of different ages. A nonzero intercept seems appropriate here, since even a very young person can have a high blood pressure. There are 30 rows of data. The data include: 3 columns and 30 rows.

Objective—To analyse the given data according to own judgement, draw insights and present the same.

Analysis Done—Plot a scatter plot which will give you distribution of blood pressure agents age, analysis of same shows there is positive relation between blood pressure and age.



Regression analysis was chosen since the data given was of age and blood pressure and the business decision to be taken could be of analysing the age/blood pressure patterns and to address age-related issues and resolve troubles, say in the case of a hospital, medical services provider, etc., through the patterns observed.

1. **Regression analysis**—A regression model of the data was formed with the below details:

Null Hypothesis: Slope = 0;

There is no significant relation between age of a person and systolic blood pressure.

Alternate Hypothesis: Slope! = 0;

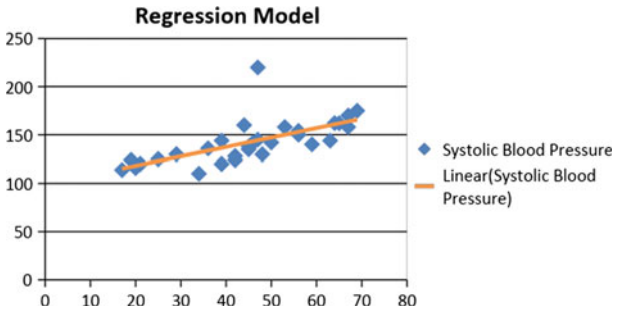
There is a significant relation between age of a person and systolic blood pressure.

P-value In the below image of the analysis, it can be seen that since the *P*-value is much less than 0.05, hence, it can be concluded that the regression model is statistically significant and alternate hypothesis holds true. Therefore, it can be said that, age of a person is directly proportional to the blood pressure.

Blood Pressure = 0.97087 * (Age) + 98.7142

If the value of c in y = mx + c/Blood Pressure = 0.9709(Age) + 98.715 was 0 instead of 98.715, then with every 1 step/year of increment in age, the Blood Pressure would be affected by factor of 0.9709.

Coefficient		Standard err.	T Stat	*P*-value	Lower 95%	Upper 95%	Power 95%	Upper 95.0%
Intercept	98.71472	10.00047	9.87101	1.28093 E−10	78.22968896	119.1997	78.22969	119.1997
X variable	0.97087	0.210216	4.618447	7.86726E–05	0.540262924	1.401478	0.540263	1.401478
Blood pressure = 0.97087 (age) + 98.71472								



Solved examples for Unstructured Data—Text Analysis

(1) Open Source Text Analysis Tools

No	Analytical tools	Description
1.	GATE	• GATE is an open source software toolkit capable of solving almost any text processing problem • It has a mature and extensive community of developers, users, educators, students and scientists • It is used by corporations, SMEs, research labs and Universities worldwide • It has a world-class team of language processing developers
2.	Rapid Miner	• Rapid Miner is a great tool for non-programmers to do data mining and text analysis • It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization
3.	Apache Mahout	• Mahout is primarily a library of machine learning algorithms that can help in clustering, classification and frequent pattern mining • It can be used in a distributed mode that helps easy integration with Hadoop. Mahout is currently being used by some of the giants in the tech industry like Adobe, AOL, Drupal and Twitter, and it has also made an impact in research and academics

(continued)

(continued)

No	Analytical tools	Description
4.	MOA	• Massive Online Analysis (MOA), as the name suggests, is primarily data stream mining software that is well suited for applications that need to handle volumes of real-time data streams at a high speed • It is a rich compilation of machine learning algorithms and has proved to be a great choice during the design of real-time applications
5.	KNIME Text Processing	• KNIME Analytics Platform is the open source software for creating data science applications and services • Intuitive, open, and continuously integrating new developments, KNIME makes understanding data and designing data science workflows and reusable components accessible to everyone

Software for Text Analysis (Vendor Based)

1.	SAS Text Miner
2.	OpenText
3.	Google Cloud Prediction API
4.	NetOwl
5.	Oracle Social Cloud

Amazon comprehend	AWS console

(2) **File Name—Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions**

Tool used	SketchEngine
Word count	11,776
Number of commas	807
Number of nouns	230
Number of adjectives	57
Number of keywords	1002

(3) **Example of LDA**:

Suppose you have the following set of sentences:

I like to eat Potato and apples.
I ate an apple and spinach smoothie for breakfast.
Puppies and tiny tots are cute.
My elder sister adopted a puppy yesterday.
Look at this cute hamster enjoying eating a piece of carrot.

What is latent Dirichlet allocation?
It's a way of automatically discovering topics that these sentences contain. For example, given these sentences and asked for 2 topics, LDA might produce something like

Sentences 1 and 2: 100% Topic A
Sentences 3 and 4: 100% Topic B
Sentence 5: 60% Topic A, 40% Topic B
Topic A: 30% Apples, 15% potato, 10% breakfast, 10% eating, … (at which point, you could interpret topic A to be about food)
Topic B: 20% Puppies, 20% Tiny Tots, 20% cute, 15% hamster, … (at which point, you could interpret topic B to be about cute animals).

Unstructured Analysis: Big Data

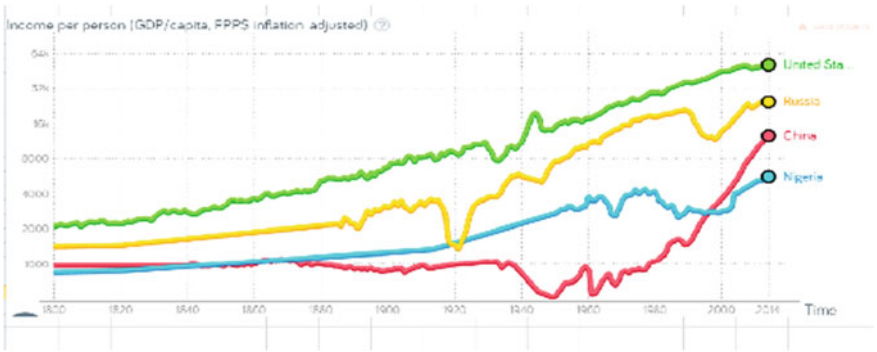
(1) Tool used is gapminder
Link: https://www.gapminder.org/data/

Gapminder: Gapminder Tools is free and comes with built-in data. You can use it online or download it. The offline version also allows you to create your own animated bubble charts, line charts and so on. Also have a look at Google's free Motion Chart Gadget and Public Data Explorer, where you can visualize your own data.

Data: CO_2 emission (metric tons per person).

Since the data is huge Snap shot of data not given here.

Row labels	Max of 2003
Qatar	60.3
United Arab Emirates	28.6
Kuwait	27.6
Trinidad and Tobago	24.9
Luxembourg	22.1
Bahrain	21.1
United States	19.6
Canada	17.5
Australia	17.1
Saudi Arabia	14.5
Oman	13.6
Finland	13.2
Brunei	13



(2) **Image Processing Tools**

Open source	URL
GemIdent	http://gemident.com/
CellCognition	http://www.openbioimage.org/wiki/cellcognition
CellProfiler	https://cellprofiler.org/
FMRIB Software Library	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki
Endrov	https://imagej.net/Endrov

(continued)

(continued)

Open source	URL
ImageJ	https://imagej.nih.gov/ij/
ITK	https://itk.org/
FreeSurfer	https://surfer.nmr.mgh.harvard.edu/
InVesalius	https://www.cti.gov.br/pt-br/invesalius
GNU Octave	https://www.gnu.org/software/octave/
ilastik	https://www.ilastik.org/
3D Slicer	https://www.slicer.org/
Tomviz	https://tomviz.org/
Open Lab	https://www.open-lab.com/

License	URL
Lead tools	https://www.leadtools.com/sdk/image-processing
Amira	https://www.fei.com/software/amira-for-life-sciences/
Analyze	https://analyzedirect.com/analyze-12-0/
Aphelion	https://www.aphelionsoftwares.com/
Bitplane	http://www.bitplane.com/
Mathematica	https://www.wolfram.com/mathematica/
MATLAB	https://www.mathworks.com/products/matlab.html
Mimics	https://www.materialise.com/en/medical/software/mimics
MountainsMap	https://www.digitalsurf.com/free-trial/
Visage SDK	http://visagetechnologies.com/products-and-services/visagesdk/