Chapter 1

An Introduction to Healthcare Data Analytics

Chandan K. Reddy

Department of Computer Science
Wayne State University
Detroit, MI
reddy@cs.wayne.edu

Charu C. Aggarwal

IBM T. J. Watson Research Center
Yorktown Heights, NY
charu@us.ibm.com

1.1 Introduction

While the healthcare costs have been constantly rising, the quality of care provided to the patients in the United States have not seen considerable improvements. Recently, several researchers have conducted studies which showed that by incorporating the current healthcare technologies, they are able to reduce mortality rates, healthcare costs, and medical complications at various hospitals. In 2009, the US government enacted the Health Information Technology for Economic and Clinical Health Act (HITECH) that includes an incentive program (around $27 billion) for the adoption and meaningful use of Electronic Health Records (EHRs).

The recent advances in information technology have led to an increasing ease in the ability to collect various forms of healthcare data. In this digital world, data becomes an integral part of healthcare. A recent report on Big Data suggests that the overall potential of healthcare data will be around $300 billion [12]. Due to the rapid advancements in the data sensing and acquisition technologies, hospitals and healthcare institutions have started collecting vast amounts of healthcare data about their patients. Effectively understanding and building knowledge from healthcare data requires developing advanced analytical techniques that can effectively transform data into meaningful and actionable information. General computing technologies have started revolutionizing the manner in which medical care is available to the patients. Data analytics, in particular, forms a critical component of these computing technologies. The analytical solutions when applied to healthcare data have an immense potential to transform healthcare delivery from being reactive to more proactive. The impact of analytics in the healthcare domain is only going to grow more in the next several years. Typically, analyzing health data will allow us to understand the patterns that are hidden in the data. Also, it will help the clinicians to build an individualized patient profile and can accurately compute the likelihood of an individual patient to suffer from a medical complication in the near future.

Healthcare data is particularly rich and it is derived from a wide variety of sources such as sensors, images, text in the form of biomedical literature/clinical notes, and traditional electronic records. This heterogeneity in the data collection and representation process leads to numerous challenges in both the processing and analysis of the underlying data. There is a wide diversity in the techniques that are required to analyze these different forms of data. In addition, the heterogeneity of the data naturally creates various data integration and data analysis challenges. In many cases, insights can be obtained from diverse data types, which are otherwise not possible from a single source of the data. It is only recently that the vast potential of such integrated data analysis methods is being realized.

From a researcher and practitioner perspective, a major challenge in healthcare is its interdisciplinary nature. The field of healthcare has often seen advances coming from diverse disciplines such as databases, data mining, information retrieval, medical researchers, and healthcare practitioners. While this interdisciplinary nature adds to the richness of the field, it also adds to the challenges in making significant advances. Computer scientists are usually not trained in domain-specific medical concepts, whereas medical practitioners and researchers also have limited exposure to the mathematical and statistical background required in the data analytics area. This has added to the difficulty in creating a coherent body of work in this field even though it is evident that much of the available data can benefit from such advanced analysis techniques. The result of such a diversity has often led to independent lines of work from completely different perspectives. Researchers in the field of data analytics are particularly susceptible to becoming isolated from real domain-specific problems, and may often propose problem formulations with excellent technique but with no practical use. This book is an attempt to bring together these diverse communities by carefully and comprehensively discussing the most relevant contributions from each domain. It is only by bringing together these diverse communities that the vast potential of data analysis methods can be harnessed.

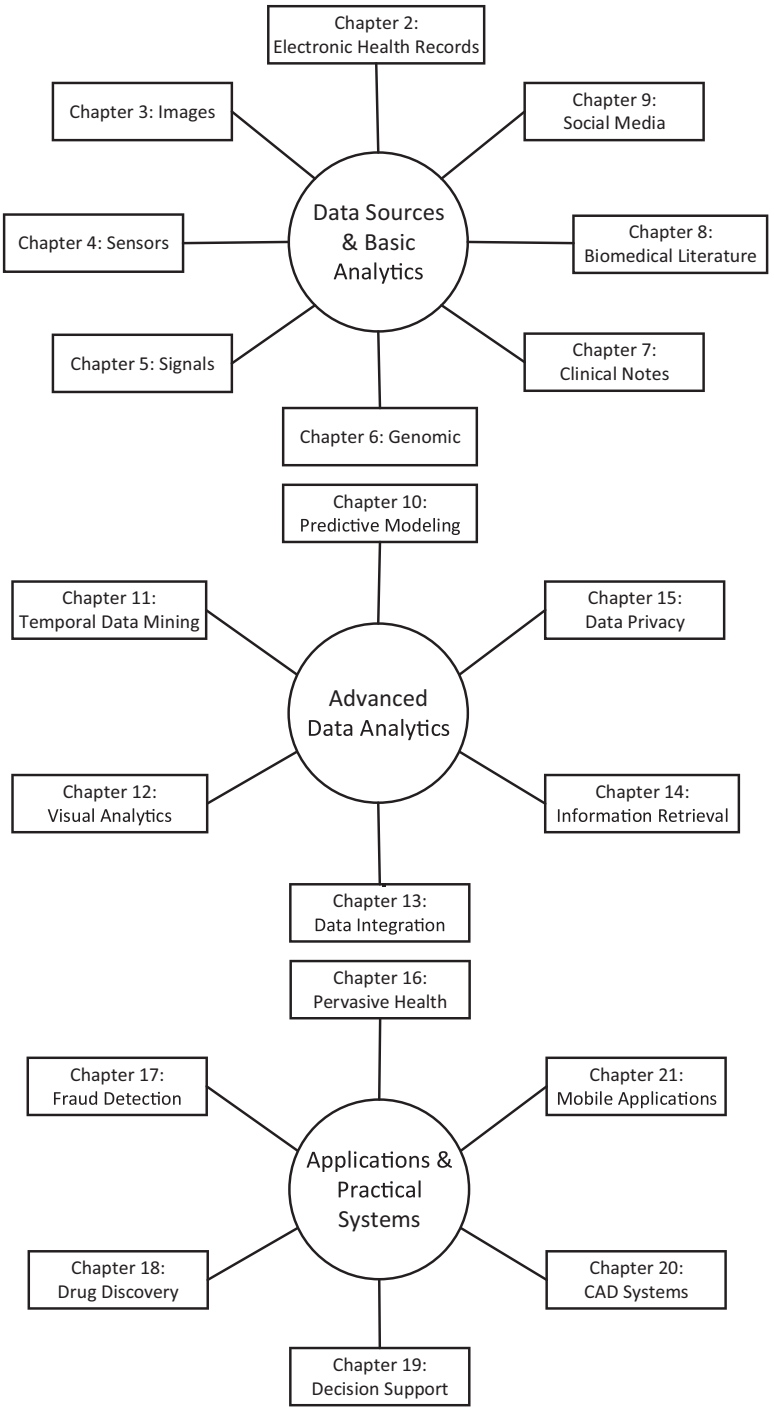

FIGURE 1.1: The overall organization of the book's contents.

Another major challenge that exists in the healthcare domain is the "data privacy gap" between medical researchers and computer scientists. Healthcare data is obviously very sensitive because it can reveal compromising information about individuals. Several laws in various countries, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, explicitly forbid the release of medical information about individuals for any purpose, unless safeguards are used to preserve privacy. Medical researchers have natural access to healthcare data because their research is often paired with an actual medical practice. Furthermore, various mechanisms exist in the medical domain to conduct research studies with voluntary participants. Such data collection is almost always paired with anonymity and confidentiality agreements.

On the other hand, acquiring data is not quite as simple for computer scientists without a proper collaboration with a medical practitioner. Even then, there are barriers in the acquisition of data. Clearly, many of these challenges can be avoided if accepted protocols, privacy technologies, and safeguards are in place. Therefore, this book will also address these issues. Figure 1.1 provides an overview of the organization of the book's contents. This book is organized into three parts:

1. *Healthcare Data Sources and Basic Analytics:* This part discusses the details of various healthcare data sources and the basic analytical methods that are widely used in the processing and analysis of such data. The various forms of patient data that is currently being collected in both clinical and non-clinical environments will be studied. The clinical data will have the structured electronic health records and biomedical images. Sensor data has been receiving a lot attention recently. Techniques for mining sensor data and biomedical signal analysis will be presented. Personalized medicine has gained a lot of importance due to the advancements in genomic data. Genomic data analysis involves several statistical techniques. These will also be elaborated. Patients' in-hospital clinical data will also include a lot of unstructured data in the form of clinical notes. In addition, the domain knowledge that can be extracted by mining the biomedical literature, will also be discussed. The fundamental data mining, machine learning, information retrieval, and natural language processing techniques for processing these data types will be extensively discussed. Finally, behavioral data captured through social media will also be discussed.

2. *Advanced Data Analytics for Healthcare:* This part deals with the advanced analytical methods focused on healthcare. This includes the clinical prediction models, temporal data mining methods, and visual analytics. Integrating heterogeneous data such as clinical and genomic data is essential for improving the predictive power of the data that will also be discussed. Information retrieval techniques that can enhance the quality of biomedical search will be presented. Data privacy is an extremely important concern in healthcare. Privacy-preserving data publishing techniques will therefore be presented.

3. *Applications and Practical Systems for Healthcare:* This part focuses on the practical applications of data analytics and the systems developed using data analytics for healthcare and clinical practice. Examples include applications of data analytics to pervasive healthcare, fraud detection, and drug discovery. In terms of the practical systems, we will discuss the details about the clinical decision support systems, computer assisted medical imaging systems, and mobile imaging systems.

These different aspects of healthcare are related to one another. Therefore, the chapters in each of the aforementioned topics are interconnected. Where necessary, pointers are provided across different chapters, depending on the underlying relevance. This chapter is organized as follows. Section 1.2 discusses the main data sources that are commonly used and the basic techniques for processing them. Section 1.3 discusses advanced techniques in the field of healthcare data analytics. Section 1.4 discusses a number of applications of healthcare analysis techniques. An overview of resources in the field of healthcare data analytics is presented in Section 1.5. Section 1.6 presents the conclusions.

1.2 Healthcare Data Sources and Basic Analytics

In this section, the various data sources and their impact on analytical algorithms will be discussed. The heterogeneity of the sources for medical data mining is rather broad, and this creates the need for a wide variety of techniques drawn from different domains of data analytics.

1.2.1 Electronic Health Records

Electronic health records (EHRs) contain a digitized version of a patient's medical history. It encompasses a full range of data relevant to a patient's care such as demographics, problems, medications, physician's observations, vital signs, medical history, laboratory data, radiology reports, progress notes, and billing data. Many EHRs go beyond a patient's medical or treatment history and may contain additional broader perspectives of a patient's care. An important property of EHRs is that they provide an effective and efficient way for healthcare providers and organizations to share with one another. In this context, EHRs are inherently designed to be in real time and they can instantly be accessed and edited by authorized users. This can be very useful in practical settings. For example, a hospital or specialist may wish to access the medical records of the primary provider. An electronic health record streamlines the workflow by allowing direct access to the updated records in real time [30]. It can generate a complete record of a patient's clinical encounter, and support other care-related activities such as evidence-based decision support, quality management, and outcomes reporting. The storage and retrieval of health-related data is more efficient using EHRs. It helps to improve quality and convenience of patient care, increase patient participation in the healthcare process, improve accuracy of diagnoses and health outcomes, and improve care coordination [29]. Various components of EHRs along with the advantages, barriers, and challenges of using EHRs are discussed in Chapter 2.

1.2.2 Biomedical Image Analysis

Medical imaging plays an important role in modern-day healthcare due to its immense capability in providing high-quality images of anatomical structures in human beings. Effectively analyzing such images can be useful for clinicians and medical researchers since it can aid disease monitoring, treatment planning, and prognosis [31]. The most popular imaging modalities used to acquire a biomedical image are magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and ultrasound (U/S). Being able to look inside of the body without hurting the patient and being able to view the human organs has tremendous implications on human health. Such capabilities allow the physicians to better understand the cause of an illness or other adverse conditions without cutting open the patient.

However, merely viewing such organs with the help of images is just the first step of the process. The final goal of biomedical image analysis is to be able to generate quantitative information and make inferences from the images that can provide far more insights into a medical condition. Such analysis has major societal significance since it is the key to understanding biological systems and solving health problems. However, it includes many challenges since the images are varied, complex, and can contain irregular shapes with noisy values. A number of general categories of research problems that arise in analyzing images are object detection, image segmentation, image registration, and feature extraction. All these challenges when resolved will enable the generation of meaningful analytic measurements that can serve as inputs to other areas of healthcare data analytics. Chapter 3 discusses a broad overview of the main medical imaging modalities along with a wide range of image analysis approaches.

1.2.3 Sensor Data Analysis

Sensor data [2] is ubiquitous in the medical domain both for real time and for retrospective analysis. Several forms of medical data collection instruments such as electrocardiogram (ECG), and electroencaphalogram (EEG) are essentially sensors that collect signals from various parts of the human body [32]. These collected data instruments are sometimes used for retrospective analysis, but more often for real-time analysis. Perhaps, the most important use-case of real-time analysis is in the context of intensive care units (ICUs) and real-time remote monitoring of patients with specific medical conditions. In all these cases, the volume of the data to the processed can be rather large. For example, in an ICU, it is not uncommon for the sensor to receive input from hundreds of data sources, and alarms need to be triggered in real time. Such applications necessitate the use of big-data frameworks and specialized hardware platforms. In remote-monitoring applications, both the real-time events and a long-term analysis of various trends and treatment alternatives is of great interest.

While rapid growth in sensor data offers significant promise to impact healthcare, it also introduces a data overload challenge. Hence, it becomes extremely important to develop novel data analytical tools that can process such large volumes of collected data into meaningful and interpretable knowledge. Such analytical methods will not only allow for better observing patients' physiological signals and help provide situational awareness to the bedside, but also provide better insights into the inefficiencies in the healthcare system that may be the root cause of surging costs. The research challenges associated with the mining of sensor data in healthcare settings and the sensor mining applications and systems in both clinical and non-clinical settings is discussed in Chapter 4.

1.2.4 Biomedical Signal Analysis

Biomedical Signal Analysis consists of measuring signals from biological sources, the origin of which lies in various physiological processes. Examples of such signals include the electroneurogram (ENG), electromyogram (EMG), electrocardiogram (ECG), electroencephalogram (EEG), electrogastrogram (EGG), phonocardiogram (PCG), and so on. The analysis of these signals is vital in diagnosing the pathological conditions and in deciding an appropriate care pathway. The measurement of physiological signals gives some form of quantitative or relative assessment of the state of the human body. These signals are acquired from various kinds of sensors and transducers either invasively or non-invasively.

These signals can be either discrete or continuous depending on the kind of care or severity of a particular pathological condition. The processing and interpretation of physiological signals is challenging due to the low signal-to-noise ratio (SNR) and the interdependency of the physiological systems. The signal data obtained from the corresponding medical instruments can be copiously noisy, and may sometimes require a significant amount of preprocessing. Several signal processing algorithms have been developed that have significantly enhanced the understanding of the physiological processes. A wide variety of methods are used for filtering, noise removal, and compact methods [36]. More sophisticated analysis methods including dimensionality reduction techniques such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and wavelet transformation have also been widely investigated in the literature. A broader overview of many of these techniques may also be found in [1, 2]. Time-series analysis methods are discussed in [37, 40]. Chapter 5 presents an overview of various signal processing techniques used for processing biomedical signals.

1.2.5 Genomic Data Analysis

A significant number of diseases are genetic in nature, but the nature of the causality between the genetic markers and the diseases has not been fully established. For example, diabetes is well

known to be a genetic disease; however, the full set of genetic markers that make an individual prone to diabetes are unknown. In some other cases, such as the blindness caused by Stargardt disease, the relevant genes are known but all the possible mutations have not been exhaustively isolated. Clearly, a broader understanding of the relationships between various genetic markers, mutations, and disease conditions has significant potential in assisting the development of various gene therapies to cure these conditions. One will be mostly interested in understanding what kind of health-related questions can be addressed through in-silico analysis of the genomic data through typical data-driven studies. Moreover, translating genetic discoveries into personalized medicine practice is a highly non-trivial task with a lot of unresolved challenges. For example, the genomic landscapes in complex diseases such as cancers are overwhelmingly complicated, revealing a high order of heterogeneity among different individuals. Solving these issues will be fitting a major piece of the puzzle and it will bring the concept of personalized medicine much more closer to reality.

Recent advancements made in the biotechnologies have led to the rapid generation of large volumes of biological and medical information and advanced genomic research. This has also led to unprecedented opportunities and hopes for genome scale study of challenging problems in life science. For example, advances in genomic technology made it possible to study the complete genomic landscape of healthy individuals for complex diseases [16]. Many of these research directions have already shown promising results in terms of generating new insights into the biology of human disease and to predict the personalized response of the individual to a particular treatment. Also, genetic data are often modeled either as sequences or as networks. Therefore, the work in this field requires a good understanding of sequence and network mining techniques. Various data analytics-based solutions are being developed for tackling key research problems in medicine such as identification of disease biomarkers and therapeutic targets and prediction of clinical outcome. More details about the fundamental computational algorithms and bioinformatics tools for genomic data analysis along with genomic data resources are discussed in Chapter 6.

1.2.6 Clinical Text Mining

Most of the information about patients is encoded in the form of *clinical notes*. These notes are typically stored in an unstructured data format and is the backbone of much of healthcare data. These contain the clinical information from the transcription of dictations, direct entry by providers, or use of speech recognition applications. These are perhaps the richest source of unexploited information. It is needless to say that the manual encoding of this free-text form on a broad range of clinical information is too costly and time consuming, though it is limited to primary and secondary diagnoses, and procedures for billing purposes. Such notes are notoriously challenging to analyze automatically due to the complexity involved in converting clinical text that is available in free-text to a structured format. It becomes hard mainly because of their unstructured nature, heterogeneity, diverse formats, and varying context across different patients and practitioners.

Natural language processing (NLP) and entity extraction play an important part in inferring useful knowledge from large volumes of clinical text to automatically encoding clinical information in a timely manner [22]. In general, data preprocessing methods are more important in these contexts as compared to the actual mining techniques. The processing of clinical text using NLP methods is more challenging when compared to the processing of other texts due to the ungrammatical nature of short and telegraphic phrases, dictations, shorthand lexicons such as abbreviations and acronyms, and often misspelled clinical terms. All these problems will have a direct impact on the various standard NLP tasks such as shallow or full parsing, sentence segmentation, text categorization, etc., thus making the clinical text processing highly challenging. A wide range of NLP methods and data mining techniques for extracting information from the clinical text are discussed in Chapter 7.

1.2.7 Mining Biomedical Literature

A significant number of applications rely on evidence from the biomedical literature. The latter is copious and has grown significantly over time. The use of text mining methods for the long-term preservation, accessibility, and usability of digitally available resources is important in biomedical applications relying on evidence from scientific literature. Text mining methods and tools offer novel ways of applying new knowledge discovery methods in the biomedical field [21][20]. Such tools offer efficient ways to search, extract, combine, analyze and summarize textual data, thus supporting researchers in knowledge discovery and generation. One of the major challenges in biomedical text mining is the multidisciplinary nature of the field. For example, biologists describe chemical compounds using brand names, while chemists often use less ambiguous IUPAC-compliant names or unambiguous descriptors such as International Chemical Identifiers. While the latter can be handled with cheminformatics tools, text mining techniques are required to extract less precisely defined entities and their relations from the literature. In this context, entity and event extraction methods play a key role in discovering useful knowledge from unstructured databases. Because the cost of curating such databases is too high, text mining methods offer new opportunities for their effective population, update, and integration. Text mining brings about other benefits to biomedical research by linking textual evidence to biomedical pathways, reducing the cost of expert knowledge validation, and generating hypotheses. The approach provides a general methodology to discover previously unknown links and enhance the way in which biomedical knowledge is organized. More details about the challenges and algorithms for biomedical text mining are discussed in Chapter 8.

1.2.8 Social Media Analysis

The rapid emergence of various social media resources such as social networking sites, blogs/microblogs, forums, question answering services, and online communities provides a wealth of information about public opinion on various aspects of healthcare. Social media data can be mined for patterns and knowledge that can be leveraged to make useful inferences about population health and public health monitoring. A significant amount of public health information can be gleaned from the inputs of various participants at social media sites. Although most individual social media posts and messages contain little informational value, aggregation of millions of such messages can generate important knowledge [4, 19]. Effectively analyzing these vast pieces of knowledge can significantly reduce the latency in collecting such complex information.

Previous research on social media analytics for healthcare has focused on capturing aggregate health trends such as outbreaks of infectious diseases, detecting reports of adverse drug interactions, and improving interventional capabilities for health-related activities. Disease outbreak detection is often strongly reflected in the content of social media and an analysis of the history of the content provides valuable insights about disease outbreaks. Topic models are frequently used for high-level analysis of such health-related content. An additional source of information in social media sites is obtained from online doctor and patient communities. Since medical conditions recur across different individuals, the online communities provide a valuable source of knowledge about various medical conditions. A major challenge in social media analysis is that the data is often unreliable, and therefore the results must be interpreted with caution. More discussion about the impact of social media analytics in improving healthcare is given in Chapter 9.

1.3 Advanced Data Analytics for Healthcare

This section will discuss a number of advanced data analytics methods for healthcare. These techniques include various data mining and machine learning models that need to be adapted to the healthcare domain.

1.3.1 Clinical Prediction Models

Clinical prediction forms a critical component of modern-day healthcare. Several prediction models have been extensively investigated and have been successfully deployed in clinical practice [26]. Such models have made a tremendous impact in terms of diagnosis and treatment of diseases. Most successful supervised learning methods that have been employed for clinical prediction tasks fall into three categories: (i) Statistical methods such as linear regression, logistic regression, and Bayesian models; (ii) Sophisticated methods in machine learning and data mining such as decision trees and artificial neural networks; and (iii) Survival models that aim to predict survival outcomes. All of these techniques focus on discovering the underlying relationship between covariate variables, which are also known as attributes and features, and a dependent outcome variable.

The choice of the model to be used for a particular healthcare problem primarily depends on the outcomes to be predicted. There are various kinds of prediction models that are proposed in the literature for handling such a diverse variety of outcomes. Some of the most common outcomes include binary and continuous forms. Other less common forms are categorical and ordinal outcomes. In addition, there are also different models proposed to handle survival outcomes where the goal is to predict the time of occurrence of a particular event of interest. These survival models are also widely studied in the context of clinical data analysis in terms of predicting the patient's survival time. There are different ways of evaluating and validating the performance of these prediction models. Different prediction models along with various kinds of evaluation mechanisms in the context of healthcare data analytics will be discussed in Chapter 10.

1.3.2 Temporal Data Mining

Healthcare data almost always contain time information and it is inconceivable to reason and mine these data without incorporating the temporal dimension. There are two major sources of temporal data generated in the healthcare domain. The first is the electronic health records (EHR) data and the second is the *sensor* data. Mining the temporal dimension of EHR data is extremely promising as it may reveal patterns that enable a more precise understanding of disease manifestation, progression and response to therapy. Some of the unique characteristics of EHR data (such as of heterogeneous, sparse, high-dimensional, irregular time intervals) makes conventional methods inadequate to handle them. Unlike EHR data, sensor data are usually represented as numeric time series that are regularly measured in time at a high frequency. Examples of these data are physiological data obtained by monitoring the patients on a regular basis and other electrical activity recordings such as electrocardiogram (ECG), electroencephalogram (EEG), etc. Sensor data for a specific subject are measured over a much shorter period of time (usually several minutes to several days) compared to the longitudinal EHR data (usually collected across the entire lifespan of the patient).

Given the different natures of EHR data and sensor data, the choice of appropriate temporal data mining methods for these types of data are often different. EHR data are usually mined using temporal pattern mining methods, which represent data instances (e.g., patients' records) as sequences of discrete events (e.g., diagnosis codes, procedures, etc.) and then try to find and enumerate statistically relevant patterns that are embedded in the data. On the other hand, sensor data are often

analyzed using signal processing and time-series analysis techniques (e.g., wavelet transform, independent component analysis, etc.) [37, 40]. Chapter 11 presents a detailed survey and summarizes the literature on temporal data mining for healthcare data.

1.3.3 Visual Analytics

The ability to analyze and identify meaningful patterns in multimodal clinical data must be addressed in order to provide a better understanding of diseases and to identify patterns that could be affecting the clinical workflow. Visual analytics provides a way to combine the strengths of human cognition with interactive interfaces and data analytics that can facilitate the exploration of complex datasets. Visual analytics is a science that involves the integration of interactive visual interfaces with analytical techniques to develop systems that facilitate reasoning over, and interpretation of, complex data [23]. Visual analytics is popular in many aspects of healthcare data analysis because of the wide variety of insights that such an analysis provides. Due to the rapid increase of health-related information, it becomes critical to build effective ways of analyzing large amounts of data by leveraging human–computer interaction and graphical interfaces. In general, providing easily understandable summaries of complex healthcare data is useful for a human in gaining novel insights.

In the evaluation of many diseases, clinicians are presented with datasets that often contain hundreds of clinical variables. The multimodal, noisy, heterogeneous, and temporal characteristics of the clinical data pose significant challenges to the users while synthesizing the information and obtaining insights from the data [24]. The amount of information being produced by healthcare organizations opens up opportunities to design new interactive interfaces to explore large-scale databases, to validate clinical data and coding techniques, and to increase transparency within different departments, hospitals, and organizations. While many of the visual methods can be directly adopted from the data mining literature [11], a number of methods, which are specific to the healthcare domain, have also been designed. A detailed discussion on the popular data visualization techniques used in clinical settings and the areas in healthcare that benefit from visual analytics are discussed in Chapter 12.

1.3.4 Clinico–Genomic Data Integration

Human diseases are inherently complex in nature and are usually governed by a complicated interplay of several diverse underlying factors, including different genomic, clinical, behavioral, and environmental factors. Clinico–pathological and genomic datasets capture the different effects of these diverse factors in a complementary manner. It is essential to build integrative models considering both genomic and clinical variables simultaneously so that they can combine the vital information that is present in both clinical and genomic data [27]. Such models can help in the design of effective diagnostics, new therapeutics, and novel drugs, which will lead us one step closer to personalized medicine [17].

This opportunity has led to an emerging area of integrative predictive models that can be built by combining clinical and genomic data, which is called clinico–genomic data integration. Clinical data refers to a broad category of a patient's pathological, behavioral, demographic, familial, environmental and medication history, while genomic data refers to a patient's genomic information including SNPs, gene expression, protein and metabolite profiles. In most of the cases, the goal of the integrative study is biomarker discovery which is to find the clinical and genomic factors related to a particular disease phenotype such as cancer vs. no cancer, tumor vs. normal tissue samples, or continuous variables such as the survival time after a particular treatment. Chapter 13 provides a comprehensive survey of different challenges with clinico–genomic data integration along with the different approaches that aim to address these challenges with an emphasis on biomarker discovery.

1.3.5 Information Retrieval

Although most work in healthcare data analytics focuses on mining and analyzing patient-related data, additional information for use in this process includes scientific data and literature. The techniques most commonly used to access this data include those from the field of information retrieval (IR). IR is the field concerned with the acquisition, organization, and searching of knowledge-based information, which is usually defined as information derived and organized from observational or experimental research [14]. The use of IR systems has become essentially ubiquitous. It is estimated that among individuals who use the Internet in the United States, over 80 percent have used it to search for personal health information and virtually all physicians use the Internet.

Information retrieval models are closely related to the problems of clinical and biomedical text mining. The basic objective of using information retrieval is to find the *content* that a user wanted based on his requirements. This typically begins with the posing of a *query* to the IR system. A *search engine* matches the query to content items through metadata. The two key components of IR are: *Indexing*, which is the process of assigning metadata to the content, and *retrieval*, which is the process of the user entering the query and retrieving relevant content. The most well-known data structure used for efficient information retrieval is the inverted index where each document is associated with an identifier. Each word then points to a list of document identifiers. This kind of representation is particularly useful for a keyword search. Furthermore, once a search has been conducted, mechanisms are required to rank the possibly large number of results, which might have been retrieved. A number of user-oriented evaluations have been performed over the years looking at users of biomedical information and measuring the search performance in clinical settings [15]. Chapter 14 discusses a number of information retrieval models for healthcare along with evaluation of such retrieval models.

1.3.6 Privacy-Preserving Data Publishing

In the healthcare domain, the definition of privacy is commonly accepted as "a person's right and desire to control the disclosure of their personal health information" [25]. Patients' health-related data is highly sensitive because of the potentially compromising information about individual participants. Various forms of data such as disease information or genomic information may be sensitive for different reasons. To enable research in the field of medicine, it is often important for medical organizations to be able to share their data with statistical experts. Sharing personal health information can bring enormous economical benefits. This naturally leads to concerns about the privacy of individuals being compromised. The data privacy problem is one of the most important challenges in the field of healthcare data analytics. Most privacy preservation methods reduce the representation accuracy of the data so that the identification of sensitive attributes of an individual is compromised. This can be achieved by either perturbing the sensitive attribute, perturbing attributes that serve as identification mechanisms, or a combination of the two. Clearly, this process required the reduction in the accuracy of data representation. Therefore, privacy preservation almost always incurs the cost of losing some data utility. Therefore, the goal of privacy preservation methods is to optimize the trade-off between utility and privacy. This ensures that the amount of utility loss at a given level of privacy is as little as possible.

The major steps in privacy-preserving data publication algorithms [5][18] are the identification of an appropriate privacy metric and level for a given access setting and data characteristics, application of one or multiple privacy-preserving algorithm(s) to achieve the desired privacy level, and postanalyzing the utility of the processed data. These three steps are repeated until the desired utility and privacy levels are jointly met. Chapter 15 focuses on applying privacy-preserving algorithms to healthcare data for secondary-use data publishing and interpretation of the usefulness and implications of the processed data.

1.4 Applications and Practical Systems for Healthcare

In the final set of chapters in this book, we will discuss the practical healthcare applications and systems that heavily utilize data analytics. These topics have evolved significantly in the past few years and are continuing to gain a lot of momentum and interest. Some of these methods, such as fraud detection, are not directly related to medical diagnosis, but are nevertheless important in this domain.

1.4.1 Data Analytics for Pervasive Health

Pervasive health refers to the process of tracking medical well-being and providing long-term medical care with the use of advanced technologies such as wearable sensors. For example, wearable monitors are often used for measuring the long-term effectiveness of various treatment mechanisms. These methods, however, face a number of challenges, such as knowledge extraction from the large volumes of data collected and real-time processing. However, recent advances in both hardware and software technologies (data analytics in particular) have made such systems a reality. These advances have made low cost intelligent health systems embedded within the home and living environments a reality [33].

A wide variety of sensor modalities can be used when developing intelligent health systems, including wearable and ambient sensors [28]. In the case of wearable sensors, sensors are attached to the body or woven into garments. For example, 3-axis accelerometers distributed over an individual's body can provide information about the orientation and movement of the corresponding body part. In addition to these advancements in sensing modalities, there has been an increasing interest in applying analytics techniques to data collected from such equipment. Several practical healthcare systems have started using analytical solutions. Some examples include cognitive health monitoring systems based on activity recognition, persuasive systems for motivating users to change their health and wellness habits, and abnormal health condition detection systems. A detailed discussion on how various analytics can be used for supporting the development of intelligent health systems along with supporting infrastructure and applications in different healthcare domains is presented in Chapter 16.

1.4.2 Healthcare Fraud Detection

Healthcare fraud has been one of the biggest problems faced by the United States and costs several billions of dollars every year. With growing healthcare costs, the threat of healthcare fraud is increasing at an alarming pace. Given the recent scrutiny of the inefficiencies in the US healthcare system, identifying fraud has been on the forefront of the efforts towards reducing the healthcare costs. One could analyze the healthcare claims data along different dimensions to identify fraud. The complexity of the healthcare domain, which includes multiple sets of participants, including healthcare providers, beneficiaries (patients), and insurance companies, makes the problem of detecting healthcare fraud equally challenging and makes it different from other domains such as credit card fraud detection and auto insurance fraud detection. In these other domains, the methods rely on constructing profiles for the users based on the historical data and they typically monitor deviations in the behavior of the user from the profile [7]. However, in healthcare fraud, such approaches are not usually applicable, because the users in the healthcare setting are the beneficiaries, who typically are not the fraud perpetrators. Hence, more sophisticated analysis is required in the healthcare sector to identify fraud.

Several solutions based on data analytics have been investigated for solving the problem of healthcare fraud. The primary advantages of data-driven fraud detection are automatic extraction

of fraud patterns and prioritization of suspicious cases [3]. Most of such analysis is performed with respect to an episode of care, which is essentially a collection of healthcare provided to a patient under the same health issue. Data-driven methods for healthcare fraud detection can be employed to answer the following questions: Is a given episode of care fraudulent or unnecessary? Is a given claim within an episode fraudulent or unnecessary? Is a provider or a network of providers fraudulent? We discuss the problem of fraud in healthcare and existing data-driven methods for fraud detection in Chapter 17.

1.4.3 Data Analytics for Pharmaceutical Discoveries

The cost of successful novel chemistry-based drug development often reaches millions of dollars, and the time to introduce the drug to market often comes close to a decade [34]. The high failure rate of drugs during this process, make the trial phases known as the "valley of death." Most new compounds fail during the FDA approval process in clinical trials or cause adverse side effects. Interdisciplinary computational approaches that combine statistics, computer science, medicine, chemoinformatics, and biology are becoming highly valuable for drug discovery and development. In the context of pharmaceutical discoveries, data analytics can potentially limit the search space and provide recommendations to the domain experts for hypothesis generation and further analysis and experiments.

Data analytics can be used in several stages of drug discovery and development to achieve different goals. In this domain, one way to categorize data analytical approaches is based on their application to pre-marketing and post-marketing stages of the drug discovery and development process. In the pre-marketing stage, data analytics focus on discovery activities such as finding signals that indicate relations between drugs and targets, drugs and drugs, genes and diseases, protein and diseases, and finding biomarkers. In the post-marketing stage an important application of data analytics is to find indications of adverse side effects for approved drugs. These methods provide a list of potential drug side effect associations that can be used for further studies. Chapter 18 provides more discussion of the applications of data analytics for pharmaceutical discoveries including drug-target interaction prediction and pharmacovigilance.

1.4.4 Clinical Decision Support Systems

Clinical Decision Support Systems (CDSS) are computer systems designed to assist clinicians with patient-related decision making, such as diagnosis and treatment [6]. CDSS have become a crucial component in the evaluation and improvement of patient treatment since they have shown to improve both patient outcomes and cost of care [35]. They can help in minimizing analytical errors by notifying the physician of potentially harmful drug interactions, and their diagnostic procedures have been shown to enable more accurate diagnoses. Some of the main advantages of CDSS are their ability in decision making and determining optimal treatment strategies, aiding general health policies by estimating the clinical and economic outcomes of different treatment methods and even estimating treatment outcomes under certain conditions. The main reason for the success of CDSS are their electronic nature, seemless integration with clinical workflows, providing decision support at the appropriate time/location. Two particular fields of healthcare where CDSS have been extremely influential are pharmacy and billing. CDSS can help pharmacies to look for negative drug interactions and then report them to the corresponding patient's ordering professional. In the billing departments, CDSS have been used to devise treatment plans that provide an optimal balance of patient care and financial expense [9]. A detailed survey of different aspects of CDSS along with various challenges associated with their usage in clinical practice is discussed in Chapter 19.

1.4.5 Computer-Aided Diagnosis

Computer-aided diagnosis/detection (CAD) is a procedure in radiology that supports radiologists in reading medical images [13]. CAD tools in general refer to fully automated second reader tools designed to assist the radiologist in the detection of lesions. There is a growing consensus among clinical experts that the use of CAD tools can improve the performance of the radiologist. The radiologist first performs an interpretation of the images as usual, while the CAD algorithms is running in the background or has already been precomputed. Structures identified by the CAD algorithm are then highlighted as regions of interest to the radiologist. The principal value of CAD tools is determined not by its stand-alone performance, but rather by carefully measuring the incremental value of CAD in normal clinical practice, such as the number of additional lesions detected using CAD. Secondly, CAD systems must not have a negative impact on patient management (for instance, false positives that cause the radiologist to recommend unnecessary biopsies and follow-ups).

From the data analytics perspective, new CAD algorithms aim at extracting key quantitative features, summarizing vast volumes of data, and/or enhancing the visualization of potentially malignant nodules, tumors, or lesions in medical images. The three important stages in the CAD data processing are candidate generation (identifying suspicious regions of interest), feature extraction (computing descriptive morphological or texture features), and classification (differentiating candidates that are true lesions from the rest of the candidates based on candidate feature vectors). A detailed overview of some CAD approaches to different diseases emphasizing the specific challenges in diagnosis and detection, and a series of case studies that apply advanced data analytics in medical imaging applications is presented in Chapter 20.

1.4.6 Mobile Imaging for Biomedical Applications

Mobile imaging refers to the application of portable computers such as smartphones or tablet computers to store, visualize, and process images with and without connections to servers, the Internet, or the cloud. Today, portable devices provide sufficient computational power for biomedical image processing and smart devices have been introduced in the operation theater. While many techniques for biomedical image acquisition will always require special equipment, the regular camera is one of the most widely used imaging modality in hospitals. Mobile technology and smart devices, especially smartphones, allows new ways of easier imaging at the patient's bedside and possess the possibility to be made into a diagnostic tool that can be used by medical professionals. Smartphones usually contain at least one high-resolution camera that can be used for image formation. Several challenges arise during the acquisition, visualization, analysis, and management of images in mobile environments. A more detailed discussion about mobile imaging and its challenges is given in Chapter 21.

1.5 Resources for Healthcare Data Analytics

There are several resources available in this field. We will now discuss the various books, journals, and organizations that provide further information on this exciting area of healthcare informatics. A classical book in the field of healthcare informatics is [39]. There are several other books that target a specific topic of work (in the context of healthcare) such as information retrieval [10], statistical methods [38], evaluation methods [8], and clinical decision support systems [6, 9].

There are a few popular organizations that are primarily involved with medical informatics research. They are American Medical Informatics Association (AMIA) [49], International Medical Informatics Association (IMIA) [50], and the European Federation for Medical Informatics (EFMI)

[51]. These organizations usually conduct annual conferences and meetings that are well attended by researchers working in healthcare informatics. The meetings typically discuss new technologies for capturing, processing, and analyzing medical data. It is a good meeting place for new researchers who would like to start research in this area.

The following are some of the well-reputed journals that publish top-quality research works in healthcare data analytics: *Journal of the American Medical Informatics Association* (*JAMIA*) [41], *Journal of Biomedical Informatics* (*JBI*) [42], *Journal of Medical Internet Research* [43], *IEEE Journal of Biomedical and Health Informatics* [44], *Medical Decision Making* [45], *International Journal of Medical Informatics* (*IJMI*) [46], and *Artificial Intelligence in Medicine* [47]. A more comprehensive list of journals in the field of healthcare and biomedical informatics along with details is available here [48].

Due to the privacy of the medical data that typically contains highly sensitive patient information, the research work in the healthcare data analytics has been fragmented into various places. Many researchers work with a specific hospital or a healthcare facility that are usually not willing to share their data due to obvious privacy concerns. However, there are a wide variety of public repositories available for researchers to design and apply their own models and algorithms. Due to the diversity in healthcare research, it will be a cumbersome task to compile all the healthcare repositories at a single location. Specific health data repositories dealing with a particular healthcare problem and data sources are listed in the corresponding chapters where the data is discussed. We hope that these repositories will be useful for both existing and upcoming researchers who do not have access to the health data from hospitals and healthcare facilities.

1.6 Conclusions

The field of healthcare data analytics has seen significant strides in recent years because of hardware and software technologies, which have increased the ease of the data collection process. The advancement of the field has, however, faced a number of challenges because of its interdisciplinary nature, privacy constraints in data collection and dissemination mechanisms, and the inherently unstructured nature of the data. In some cases, the data may have very high volume, which requires real-time analysis and insights. In some cases, the data may be complex, which may require specialized retrieval and analytical techniques. The advances in data collection technologies, which have enabled the field of analytics, also pose new challenges because of their efficiency in collecting large amounts of data. The techniques used in the healthcare domain are also very diverse because of the inherent variations in the underlying data type. This book provides a comprehensive overview of these different aspects of healthcare data analytics, and the various research challenges that still need to be addressed.

Bibliography

[1] Charu C. *Aggarwal. Data Streams: Models and Algorithms*. Springer. 2007.

[2] Charu C. Aggarwal. *Managing and Mining Sensor Data*. Springer. 2013.

[3] Charu C. *Aggarwal. Outlier Analysis*. Springer. 2013.

[4] Charu C. Aggarwal. *Social Network Data Analytics*. Springer, 2011.

[5] Charu C. Aggarwal and Philip S. Yu. *Privacy-Preserving Data Mining: Models and Algorithms*. Springer. 2008.

[6] Eta S Berner. *Clinical Decision Support Systems*. Springer, 2007.

[7] Richard J. Bolton, and David J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–249, 2002.

[8] Charles P. Friedman. *Evaluation Methods in Biomedical Informatics*. Springer, 2006.

[9] Robert A. Greenes. *Clinical Decision Support: The Road Ahead*. Academic Press, 2011.

[10] William Hersh. *Information Retrieval: A Health and Biomedical Perspective*. Springer, 2008.

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

[12] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute Report, May 2011.

[13] Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31:2007.

[14] W. Hersh. *Information Retrieval: A Health and Biomedical Perspective.* Springer, 2009.

[15] R. B. Haynes, K. A. McKibbon, C. J. Walker, N. Ryan, D. Fitzgerald, and M. F. Ramsden. Online access to MEDLINE in clinical settings: A study of use and usefulness. *Annals of Internal Medicine*, 112(1):78–84, 1990.

[16] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, J. Diaz, L. A., and K. W. Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, 2013.

[17] P. Edn, C. Ritz, C. Rose, M. Fern, and C. Peterson. Good old clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *European Journal of Cancer*, 40(12):1837–1841, 2004.

[18] Rashid Hussain Khokhar, Rui Chen, Benjamin C.M. Fung, and Siu Man Lui. Quantifying the costs and benefits of privacy-preserving health data publishing. *Journal of Biomedical Informatics*, 50:107–121, 2014.

[19] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Modeling spread of disease from social interactions. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*, pages 322–329, 2012.

[20] L. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: From information retrieval to biological discovery. *Nature Reviews Genetics*, 7(2):119–129, 2006.

[21] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. Cohen. Frontiers of biomedical text mining: Current progress. *Briefings in Bioinformatics*, 8(5):358–375, 2007.

[22] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*, pages 128–144, 2008.

[23] Daniel Keim et al. *Visual Analytics: Definition, Process, and Challenges*. Springer Berlin Heidelberg, 2008.

[24] K. Wongsuphasawat, J. A. Guerra Gmez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman. LifeFlow: Visualizing an overview of event sequences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1747-1756. ACM, 2011.

[25] Thomas C. Rindfieisch. Privacy, information technology, and health care. *Communications of the ACM*, 40(8):92–100, 1997.

[26] E. W. Steyerberg. *Clinical Prediction Models*. Springer, 2009.

[27] E. E. Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223, 2009.

[28] Min Chen, Sergio Gonzalez, Athanasios Vasilakos, Huasong Cao, and Victor C. Leung. Body area networks: A survey. *Mobile Networks and Applications*, 16(2):171–193, April 2011.

[29] Catherine M. DesRoches et al. Electronic health records in ambulatory carea national survey of physicians. *New England Journal of Medicine* 359(1):50–60, 2008.

[30] Richard Hillestad et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Affairs* 24(5):1103–1117, 2005.

[31] Stanley R. Sternberg, Biomedical image processing. *Computer* 16(1):22–34, 1983.

[32] G. Acampora, D. J. Cook, P. Rashidi, A. V. Vasilakos. A survey on ambient intelligence in healthcare, *Proceedings of the IEEE*, 101(12):2470–2494, Dec. 2013.

[33] U. Varshney. Pervasive healthcare and wireless health monitoring. *Mobile Networks and Applications* 12(2–3):113–127, 2007.

[34] Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* 9 (3):203–214, 2010.

[35] R. Amarasingham, L. Plantinga, M. Diener-West, D. Gaskin, and N. Powe. Clinical information technologies and inpatient outcomes: A multiple hospital study. *Archives of Internal Medicine* 169(2):108–114, 2009.

[36] Athanasios Papoulis. *Signal Analysis*. McGraw-Hill: New York, 1978.

[37] Robert H. Shumway and David S. Stoffer. *Time-Series Analysis and Its Applications: With R Examples*. Springer: New York, 2011.

[38] Robert F. Woolson and William R. Clarke. *Statistical Methods for the Analysis of Biomedical Data*, Volume 371. John Wiley & Sons, 2011.

[39] Edward H. Shortliffe and James J. Cimino. *Biomedical Informatics*. Springer, 2006.

[40] Mitsa Thephano. *Temporal Data Mining*. Chapman and Hall/CRC Press, 2010.

[41] `http://jamia.bmj.com/`

[42] `http://www.journals.elsevier.com/journal-of-biomedical-informatics/`

[43] `http://www.jmir.org/`

[44] `http://jbhi.embs.org/`

[45] http://mdm.sagepub.com/

[46] http://www.ijmijournal.com/

[47] http://www.journals.elsevier.com/artificial-intelligence-in-medicine/

[48] http://clinfowiki.org/wiki/index.php/Leading_Health_Informatics_and_Medical_Informatics_Journals

[49] http://www.amia.org/

[50] www.imia-medinfo.org/

[51] http://www.efmi.org/