Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

Healthcare Data Analytics

Edited by
Chandan K. Reddy
Charu C. Aggarwal

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Healthcare Data Analytics

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

SERIES EDITOR
Vipin Kumar
University of Minnesota
Department of Computer Science and Engineering
Minneapolis, Minnesota, U.S.A.

AIMS AND SCOPE

This series aims to capture new developments and applications in data mining and knowledge discovery, while summarizing the computational tools and techniques useful in data analysis. This series encourages the integration of mathematical, statistical, and computational methods and techniques through the publication of a broad range of textbooks, reference works, and handbooks. The inclusion of concrete examples and applications is highly encouraged. The scope of the series includes, but is not limited to, titles in the areas of data mining and knowledge discovery methods and applications, modeling, algorithms, theory and foundations, data and knowledge visualization, data mining systems and tools, and privacy and security issues.

PUBLISHED TITLES

ADVANCES IN MACHINE LEARNING AND DATA MINING FOR ASTRONOMY
Michael J. Way, Jeffrey D. Scargle, Kamal M. Ali, and Ashok N. Srivastava

BIOLOGICAL DATA MINING
Jake Y. Chen and Stefano Lonardi

COMPUTATIONAL BUSINESS ANALYTICS
Subrata Das

COMPUTATIONAL INTELLIGENT DATA ANALYSIS FOR SUSTAINABLE DEVELOPMENT
Ting Yu, Nitesh V. Chawla, and Simeon Simoff

COMPUTATIONAL METHODS OF FEATURE SELECTION
Huan Liu and Hiroshi Motoda

CONSTRAINED CLUSTERING: ADVANCES IN ALGORITHMS, THEORY, AND APPLICATIONS
Sugato Basu, Ian Davidson, and Kiri L. Wagstaff

CONTRAST DATA MINING: CONCEPTS, ALGORITHMS, AND APPLICATIONS
Guozhu Dong and James Bailey

DATA CLASSIFICATION: ALGORITHMS AND APPLICATIONS
Charu C. Aggarawal

DATA CLUSTERING: ALGORITHMS AND APPLICATIONS
Charu C. Aggarawal and Chandan K. Reddy

DATA CLUSTERING IN C++: AN OBJECT-ORIENTED APPROACH
Guojun Gan

DATA MINING FOR DESIGN AND MARKETING
Yukio Ohsawa and Katsutoshi Yada

DATA MINING WITH R: LEARNING WITH CASE STUDIES
Luís Torgo

FOUNDATIONS OF PREDICTIVE ANALYTICS
James Wu and Stephen Coggeshall

GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY, SECOND EDITION
Harvey J. Miller and Jiawei Han

HANDBOOK OF EDUCATIONAL DATA MINING
Cristóbal Romero, Sebastian Ventura, Mykola Pechenizkiy, and Ryan S.J.d. Baker

HEALTHCARE DATA ANALYTICS
Chandan K. Reddy and Charu C. Aggarwal

INFORMATION DISCOVERY ON ELECTRONIC HEALTH RECORDS
Vagelis Hristidis

INTELLIGENT TECHNOLOGIES FOR WEB APPLICATIONS
Priti Srinivas Sajja and Rajendra Akerkar

INTRODUCTION TO PRIVACY-PRESERVING DATA PUBLISHING: CONCEPTS AND TECHNIQUES
Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu

KNOWLEDGE DISCOVERY FOR COUNTERTERRORISM AND LAW ENFORCEMENT
David Skillicorn

KNOWLEDGE DISCOVERY FROM DATA STREAMS
João Gama

MACHINE LEARNING AND KNOWLEDGE DISCOVERY FOR ENGINEERING SYSTEMS HEALTH MANAGEMENT
Ashok N. Srivastava and Jiawei Han

MINING SOFTWARE SPECIFICATIONS: METHODOLOGIES AND APPLICATIONS
David Lo, Siau-Cheng Khoo, Jiawei Han, and Chao Liu

MULTIMEDIA DATA MINING: A SYSTEMATIC INTRODUCTION TO CONCEPTS AND THEORY
Zhongfei Zhang and Ruofei Zhang

MUSIC DATA MINING
Tao Li, Mitsunori Ogihara, and George Tzanetakis

NEXT GENERATION OF DATA MINING
Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar

RAPIDMINER: DATA MINING USE CASES AND BUSINESS ANALYTICS APPLICATIONS
Markus Hofmann and Ralf Klinkenberg

RELATIONAL DATA CLUSTERING: MODELS, ALGORITHMS, AND APPLICATIONS
Bo Long, Zhongfei Zhang, and Philip S. Yu

SERVICE-ORIENTED DISTRIBUTED KNOWLEDGE DISCOVERY
Domenico Talia and Paolo Trunfio

SPECTRAL FEATURE SELECTION FOR DATA MINING
Zheng Alan Zhao and Huan Liu

STATISTICAL DATA MINING USING SAS APPLICATIONS, SECOND EDITION
George Fernandez

SUPPORT VECTOR MACHINES: OPTIMIZATION BASED THEORY, ALGORITHMS, AND EXTENSIONS
Naiyang Deng, Yingjie Tian, and Chunhua Zhang

TEMPORAL DATA MINING
Theophano Mitsa

TEXT MINING: CLASSIFICATION, CLUSTERING, AND APPLICATIONS
Ashok N. Srivastava and Mehran Sahami

THE TOP TEN ALGORITHMS IN DATA MINING
Xindong Wu and Vipin Kumar

UNDERSTANDING COMPLEX DATASETS: DATA MINING WITH MATRIX DECOMPOSITIONS
David Skillicorn

Healthcare Data Analytics

Edited by

Chandan K. Reddy

Wayne State University

Detroit, Michigan, USA

Charu C. Aggarwal

IBM T. J. Watson Research Center

Yorktown Heights, New York, USA

CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the

Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20150202

International Standard Book Number-13: 978-1-4822-3212-7 (eBook - PDF)

Contents

Editor Biographies

Chandan K. Reddy is an Associate Professor in the Department of Computer Science at Wayne State University. He received his PhD from Cornell University and MS from Michigan State University. His primary research interests are in the areas of data mining and machine learning with applications to healthcare, bioinformatics, and social network analysis. His research is funded by the National Science Foundation, the National Institutes of Health, Department of Transportation, and the Susan G. Komen for the Cure Foundation. He has published over 50 peer-reviewed articles in leading conferences and journals. He received the Best Application Paper Award at the ACM SIGKDD conference in 2010 and was a finalist of the INFORMS Franz Edelman Award Competition in 2011. He is a senior member of IEEE and a life member of the ACM.

Charu C. Aggarwal is a Distinguished Research Staff Member (DRSM) at the IBM T. J. Watson Research Center in Yorktown Heights, New York. He completed his BS from IIT Kanpur in 1993 and his PhD from the Massachusetts Institute of Technology in 1996. He has published more than 250 papers in refereed conferences and journals, and has applied for or been granted more than 80 patents. He is an author or editor of 13 books, including the first comprehensive book on outlier analysis. Because of the commercial value of his patents, he has thrice been designated a Master Inventor at IBM. He is a recipient of an IBM Corporate Award (2003) for his work on bioterrorist threat detection in data streams, a recipient of the IBM Outstanding Innovation Award (2008) for his scientific contributions to privacy technology, a recipient of the IBM Outstanding Technical Achievement Award (2009) for his work on data streams, and a recipient of an IBM Research Division Award (2008) for his contributions to System S. He also received the EDBT 2014 Test of Time Award for his work on condensation-based privacy-preserving data mining. He has served as conference chair and associate editor at many reputed conferences and journals in data mining, general co-chair of the IEEE Big Data Conference (2014), and is editor-in-chief of the ACM SIGKDD Explorations. He is a fellow of the ACM, SIAM and the IEEE, for "contributions to knowledge discovery and data mining algorithms."

Contributors

Giovanni Acampora
Nottingham Trent University
Nottingham, UK

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, New York

Noha Alnazzawi
University of Manchester
Manchester, UK

Martin Alther
Wayne State University
Detroit, MI

Sophia Ananiadou
University of Manchester
Manchester, UK

Iyad Batal
General Electric Global Research
San Ramon, CA

Riza Batista-Navarro
University of Manchester
Manchester, UK

Luca Bogoni
Siemens Medical Solutions
Malvern, PA

Jesus Caban
Walter Reed National Military Medical Center
Bethesda, MD

Varun Chandola
State University of New York at Buffalo
Buffalo, NY

Annie T. Chen
University of North Carolina at Chapel Hill
Chapel Hill, NC

Diane J. Cook
Washington State University
Pullman, WA

Juan Cui
University of Nebraska-Lincoln
Lincoln, NE

Thomas M. Deserno
RWTH Aachen University
Aachen, Germany

Sanjoy Dey
University of Minnesota
Minneapolis, MN

Shobeir Fakhraei
University of Maryland
College Park, MD

Sahika Genc
GE Global Research
Niskayuna, NY

Lise Getoor
University of California
Santa Cruz, CA

Joydeep Ghosh
The University of Texas at Austin
Austin, TX

David Gotz
University of North Carolina at Chapel Hill
Chapel Hill, NC

Rohit Gupta
University of Minnesota
Minneapolis, MN

Sandeep Gupta
GE Global Research
Niskayuna, NY

Gerardo Hermosillo
Siemens Medical Solutions
Malvern, PA

William R. Hersh
Oregon Health & Science University
Portland, OR

Suresh Joel
GE Global Research
Bangalore, India

Stephan M. Jonas
RWTH Aachen University
Aachen, Germany

Siddhartha R. Jonnalagadda
Northwestern University
Chicago, IL

Georgios Kontonatsios
University of Manchester
Manchester, UK

Ioannis Korkontzelos
University of Manchester
Manchester, UK

Alexander Kotov
Wayne State University
Detroit, MI

Vipin Kumar
University of Minnesota
Minneapolis, MN

Rajesh Langoju
GE Global Research
Bangalore, India

Yan Li
Wayne State University
Detroit, MI

Shu Liao
Siemens Medical Solutions
Malvern, PA

Paulo Mendonca
GE Global Research
Niskayuna, NY

Claudiu Mihăilă
University of Manchester
Manchester, UK

Eberechukwu Onukwugha
University of Maryland
Baltimore, MD

Dirk Padfield
GE Global Research
Niskayuna, NY

Yubin Park
The University of Texas at Austin
Austin, TX

Abhijit Patil
GE Global Research
Bangalore, India

Bhushan D. Patil
GE Global Research
Bangalore, India

Zhigang Peng
Siemens Medical Solutions
Malvern, PA

Rajiur Rahman
Wayne State University
Detroit, MI

Kalpana Raja
Northwestern University
Chicago, IL

Rafal Rak
University of Manchester
Manchester, UK

Parisa Rashidi
University of Florida
Gainesville, FL

Chandan K. Reddy
Wayne State University
Detroit, MI

Marcos Salganicoff
Siemens Medical Solutions
Malvern, PA

Michael Schmidt
Columbia University Medical Center
New York, NY

Jack Schryver
Oak Ridge National Laboratory
Oakridge, TN

Xiang Sean Zhou
Siemens Medical Solutions
Malvern, PA

Yoshihisa Shinagawa
Siemens Medical Solutions
Malvern, PA

Daby Sow
IBM T. J. Watson Research Center
Yorktown Heights, NY

Michael Steinbach
University of Minnesota
Minneapolis, MN

Sreenivas Sukumar
Oak Ridge National Laboratory
Oakridge, TN

Paul Thompson
University of Manchester
Manchester, UK

Deepak S. Turaga
IBM T. J. Watson Research Center
Yorktown Heights, NY

Kiran K. Turaga
Medical College of Wisconsin
Milwaukee, WI

Athanasios V. Vasilakos
University of Western Macedonia
Kozani, Greece

Matthias Wolf
Siemens Medical Solutions
Malvern, PA

Shipeng Yu
Siemens Medical Solutions
Malvern, PA

Yiqiang Zhan
Siemens Medical Solutions
Malvern, PA

Preface

Innovations in computing technologies have revolutionized healthcare in recent years. The analytical style of reasoning has not only changed the way in which information is collected and stored but has also played an increasingly important role in the management and delivery of healthcare. In particular, data analytics has emerged as a promising tool for solving problems in various healthcare-related disciplines. This book will present a comprehensive review of data analytics in the field of healthcare. The goal is to provide a platform for interdisciplinary researchers to learn about the fundamental principles, algorithms, and applications of intelligent data acquisition, processing, and analysis of healthcare data. This book will provide readers with an understanding of the vast number of analytical techniques for healthcare problems and their relationships with one another. This understanding includes details of specific techniques and required combinations of tools to design effective ways of handling, retrieving, analyzing, and making use of healthcare data. This book will provide a unique perspective of healthcare related opportunities for developing new computing technologies.

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, image processing, 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 data analytics area. This has added to the difficulty in creating a coherent body of work in this field. The result has often been independent lines of work from completely different perspectives. This book is an attempt to bring together these diverse communities by carefully and comprehensively discussing the most relevant contributions from each domain.

The book provides a comprehensive overview of the healthcare data analytics field as it stands today, and to educate the community about future research challenges and opportunities. Even though the book is structured as an edited collection of chapters, special care was taken during the creation of the book to cover healthcare topics exhaustively by coordinating the contributions from various authors. Focus was also placed on reviews and surveys rather than individual research results in order to emphasize comprehensiveness in coverage. Each book chapter is written by prominent researchers and experts working in the healthcare domain. The chapters in the book are divided into three major categories:

- **Healthcare Data Sources and Basic Analytics:** These chapters discuss the details about the various healthcare data sources and the analytical techniques that are widely used in the processing and analysis of such data. The various forms of patient data include electronic health records, biomedical images, sensor data, biomedical signals, genomic data, clinical text, biomedical literature, and data gathered from social media.

- **Advanced Data Analytics for Healthcare:** These chapters deal with the advanced data analytical methods focused on healthcare. These include the clinical prediction models, temporal pattern mining methods, and visual analytics. In addition, other advanced methods such as data integration, information retrieval, and privacy-preserving data publishing will also be discussed.

- **Applications and Practical Systems for Healthcare:** These chapters focus on the applications of data analytics and the relevant practical systems. It will cover the applications of data analytics to pervasive healthcare, fraud detection, and drug discovery. In terms of the practical systems, it covers clinical decision support systems, computer assisted medical imaging systems, and mobile imaging systems.

It is hoped that this comprehensive book will serve as a compendium to students, researchers, and practitioners. Each chapter is structured as a "survey-style" article discussing the prominent research issues and the advances made on that research topic. Special effort was taken in ensuring that each chapter is self-contained and the background required from other chapters is minimal. Finally, we hope that the topics discussed in this book will lead to further developments in the field of healthcare data analytics that can help in improving the health and well-being of people. We believe that research in the field of healthcare data analytics will continue to grow in the years to come.

Acknowledgment: This work was supported in part by National Science Foundation grant No. 1231742.

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.

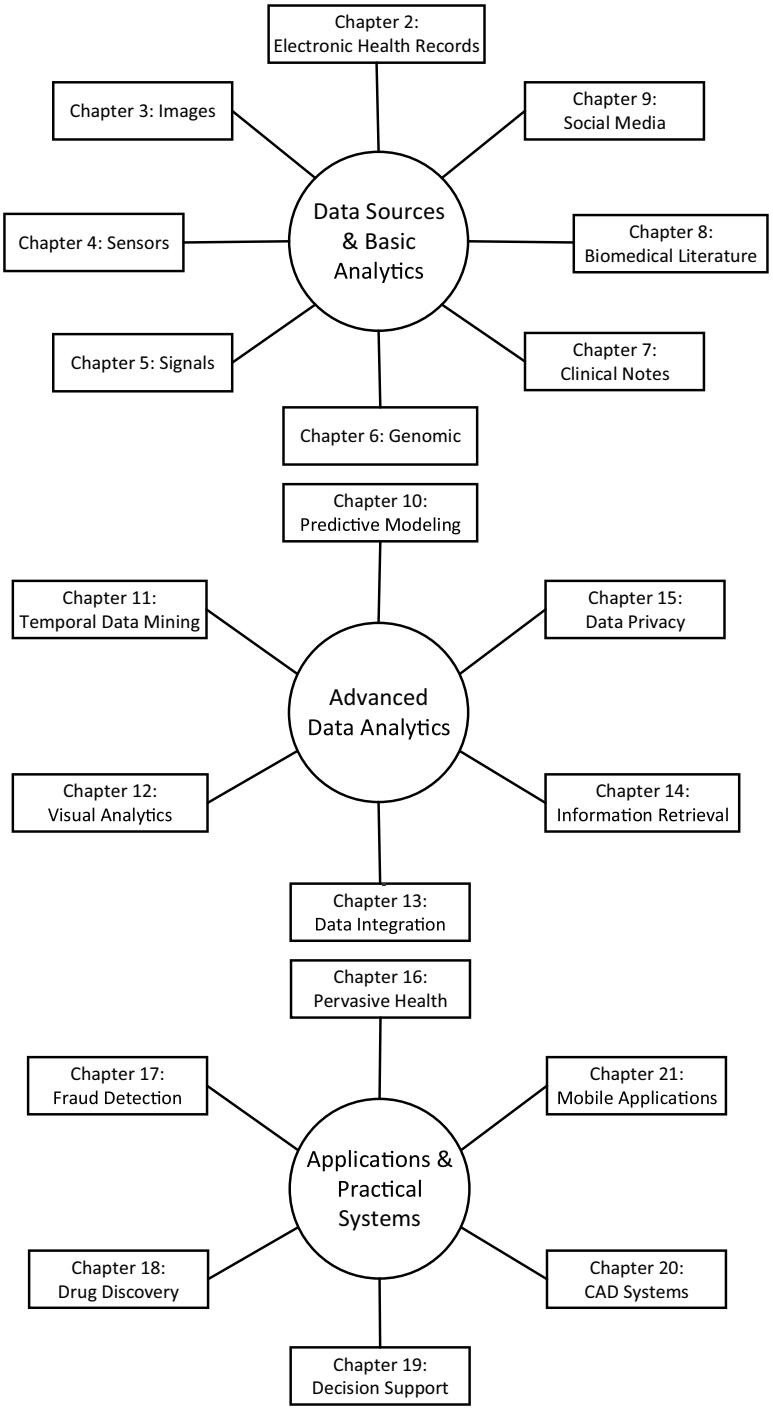

FIGURE 1.1: The overall organization of the book's contents.

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/

Part I

Healthcare Data Sources and Basic Analytics

Chapter 2

Electronic Health Records: A Survey

Rajiur Rahman

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

Chandan K. Reddy

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

2.1 Introduction

An Electronic Health Record (EHR) is a digital version of a patient's medical history. It is a longitudinal record of patient health information generated by one or several encounters in any healthcare providing setting. The term is often used interchangeably with EMR (Electronic Medical Record) and CPR (Computer-based Patient Record). 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, immunizations, laboratory data, radiology reports, personal statistics, progress notes, and billing data. The EHR system automates the data management process of complex clinical environments and has the potential to streamline the clinician's workflow. 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. An EHR system integrates data for different purposes. It enables the administrator to utilize the data for billing purposes, the physician to analyze patient diagnostics information and treatment effectiveness, the nurse to report adverse conditions, and the researcher to discover new knowledge.

EHR has several advantages over paper-based systems. Storage and retrieval of data is obviously 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. It also reduces cost by eliminating the need for paper and other storage media. It provides the opportunity for research in different disciplines. In 2011, 54% of physicians had adopted an EHR system, and about three-quarters of adopters reported that using an EHR system resulted in enhanced patient care [1].

Usually, EHR is maintained within an institution, such as a hospital, clinic, or physician's office. An institution will contain the longitudinal records of a particular patient that have been collected at their end. The institution will not contain the records of all the care provided to the patient at other venues. Information regarding the general population may be kept in a nationwide or regional health information system. Depending on the goal, service, venue, and role of the user, EHR can have different data formats, presentations, and level of detail.

The remainder of this chapter is organized as follows. Section 2.2 discusses a brief history of EHR development and Section 2.3 provides the components of EHRs. Section 2.4 presents a comprehensive review of existing coding systems in EHR. The benefits of using EHRs are explained in more detail in Section 2.5, while the barriers for the widespread adoption of EHRs are discussed in Section 2.6. Section 2.7 briefly explains some of the challenges of using EHR data. The prominent phenotyping algorithms are described in Section 2.8 and our discussion is concluded in Section 2.9.

2.2 History of EHR

The first known medical record can be traced back to the fifth century B.C. when Hippocrates prescribed two goals for medical records [2]:

- A medical record should accurately reflect the course of disease.
- A medical record should indicate the probable cause of disease.

Although these two goals are still appropriate, EHR has a lot more to offer. Modern EHR can provide additional functionalities that could not be performed using paper-based systems.

Modern-day EHR first began to appear in the 1960s. Early EHRs were developed due to physicians' concerns about the increasing complexity and size of medical data. Data retrieval was much faster using digital format. In 1967, Latter Day Saints Hospitals in Utah started using Health Evaluation through Logical Programming (HELP) software. HELP is notable for its pioneering logical decision support features. In 1969, Harvard Medical School developed its own software Computer Stored Ambulatory Record (COASTER) and Duke University began to develop The Medical Record (TMR).

In 1970, Lockheed unveiled the Technicon Medical Information Management System/ Technicon Data System (TDS). It was implemented at El Camion Hospital in California. It came with a groundbreaking Computer Provided Order Entry (CPOE) system. In 1979, Judith Faulkner, a computer programmer established Human Services Computing Inc., which developed the Chronicles data repository. The company later became Epic Systems. It was initially based on a single longitudinal patient record and designed to handle enterprise-wide data from inpatient, ambulatory, and payer environments.

In 1985, The Department of Veterans Affairs launched the automated data processing system, Decentralized Hospital Computer Program (DHCP), which includes extensive clinical and administrative capabilities within its medical facilities. It received the Smithsonian Award for best use of Information Technology in Medicine in 1995. The current variant of DHCP is VistA (Veterans Health Information Systems and Technology Architecture). By providing care to over 8 million veterans operating in 163 hospitals, 800 clinics, and 135 nursing homes, VistA manages one of the largest medical system in the United States [4]. In 1983, Epic Systems launched a patient scheduling software program called Cadence. This application helped clients to improve resource utilization and manage patient access. In 1988, Science Application International Corporation (SAIC) secured a $1.02 billion dollar contract from the U.S. Government to develop a composite healthcare system. In 1992, Epic Systems introduced the first Windows-based EHR software named EpicCare. Allscripts released the first software with an electronic prescribing solution for physicians in 1998.

From 2000 and beyond, EHR software has been increasingly trying to incorporate other functionalities to become an interactive companion for physicians and professionals. In January 2004, President George W. Bush launched an initiative for the widespread adaptation of EHRs within the next 10 years. He said in his State of the Union Address, "By computerizing health records, we can avoid dangerous medical mistakes, reduce costs, and improve care" [5]. In January 2009, in a speech at George Mason University, President Barack Obama said "[EHRs] will cut waste, eliminate red tape, and reduce the need to repeat expensive medical tests. It just won't save billions of dollars and thousands of jobs – it will save lives by reducing the deadly but preventable medical errors that pervade our health care system" [6]. The data from a National Ambulatory Medical Care Survey (NAMCS) and Physicians Workflow mail survey shows that in the year 2011, 54% of the physicians had adopted an EHR system. About three-quarters of the adopters reported that their system meets the federal "meaningful use" criteria. Almost half (47%) of the physicians said they were somewhat satisfied, and 38% reported being very satisfied with their system. About three-quarters of the adopters reported that EHR has resulted in enhanced patient care. Nearly one-half of physicians without an EHR system at the time of the survey said they had plans for purchasing one within the next year [1].

2.3 Components of EHR

The main purpose of EHR is to support clinical care and billing. This also includes other functionalities, such as improving the quality and convenience of patient care, improving the accuracy of diagnoses and health outcomes, improving care coordination and patient participation, improving cost savings, and finally, improving the general health of the population. Most modern EHR systems are designed to integrate data from different components such as administrative, nursing, pharmacy, laboratory, radiology, and physician' entries, etc. Electronic records may be generated from any department. Hospitals and clinics may have a number of different ancillary system providers; in that case, these systems are not necessarily integrated to the main EHR system. It is possible that these systems are stand-alone, and different standards of vocabularies have been used. If appropriate interfaces are provided, data from these systems can be incorporated in a consolidated fashion; otherwise a clinician has to open and log into a series of applications to get the complete patient record. The number of components present may also vary depending on the service provided. Figure 2.1 shows different components of an EHR system.

2.3.1 Administrative System Components

Administrative data such as patient registration, admission, discharge, and transfer data are key components of the EHR. It also includes name, demographics, employer history, chief compliant, patient disposition, etc., along with the patient billing information. Social history data such as marital status, home environment, daily routine, dietary patterns, sleep patterns, exercise patterns, tobacco use, alcohol use, drug use and family history data such as personal health history, hereditary diseases, father, mother and sibling(s) health status, age, and cause of death can also be a part of it. Apart from the fields like "comments" or "description," these data generally contain <name-value> pairs. This information is used to identify and assess a patient, and for all other administrative purposes. During the registration process, a patient is generally assigned a unique identification key comprising of a numeric or alphanumeric sequence. This key helps to link all the components across different platforms. For example, lab test data can create an electronic record; and another record is created from radiology results. Both records will have the same identifier key to represent a single patient. Records of a previous encounter are also pulled up using this key. It is often referred to as the medical record number or master patient index (MPI). Administrative data allows the aggregation of a person's health information for clinical analysis and research.

2.3.2 Laboratory System Components & Vital Signs

Generally, laboratory systems are stand-alone systems that are interfaced to the central EHR system. It is a structured data that can be expressed using standard terminology and stored in the form of a name-value pair. Lab data plays an extremely important part in the clinical care process, providing professionals the information needed for prevention, diagnosis, treatment, and health management. About 60% to 70% of medical decisions are based on laboratory test results [7]. Electronic lab data has several benefits including improved presentation and reduction of error due to manual data entry. A physician can easily compare the results from previous tests. If the options are provided, he can also analyze automatically whether data results fall within normal range or not.

The most common coding system used to represent the laboratory test data is Logical Observation Identifiers Names and Codes (LOINC). Many hospitals use their local dictionaries as well to encode variables. A 2009–2010 Vanderbilt University Medical Center data standardization study found that for simple concepts such as "weight" and "height," there were more than five internal representations. In different places there are different field names for the same feature and the values

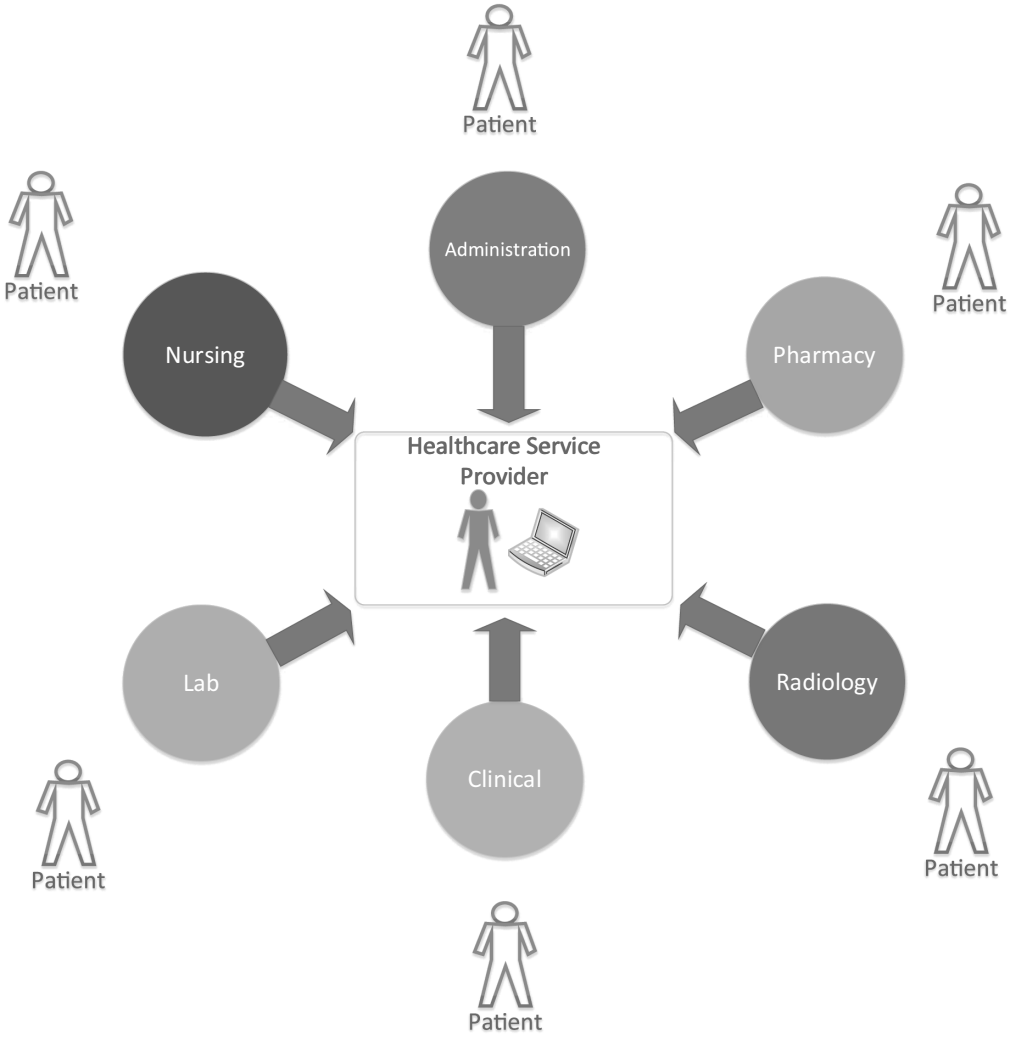

FIGURE 2.1: Various components of EHR.

are stored with different units (e.g., kilograms, grams, and pounds for weight; centimeters, meters, inches, and feet for height).

Vital signs are the indicators of a patient's general physical condition. It includes pulse, respiratory rate, blood pressure, body temperature, body mass index (BMI), etc. A typical EHR system must provide the option to accommodate these kinds of variables.

2.3.3 Radiology System Components

In hospital radiology departments, radiology information systems (RIS) are used for managing medical imagery and associated data. RIS is the core database to store, manipulate, and distribute patient radiological data. It uses Current Procedural Terminology (CPT) or International Classification of Diseases (ICD) coding systems to identify procedures and resources. Generally, an RIS consists of patient tracking, scheduling, result reporting, and image tracking capabilities. RIS is usually used along with a picture archiving communications system (PACS), which is a medical technology for

providing economical storage and convenient access to the digital images. An RIS can generate an entire patient's imagery history and statistical reports for patients or procedures. Although many hospitals are using RIS, it may or may not be integrated with the central EHR system.

2.3.4 Pharmacy System Components

In hospitals and clinics, the pharmacy department's responsibility is to maintain the inventory, prescription management, billing, and dispensing medications. The pharmacy component in EHR will hold the complete medication history of a patient such as drug name, dosage, route, quantity, frequency, start and stop date, prescribed by, allergic reaction to medications, source of medication, etc. Pharmacists serve an important public health role by administering immunizations and must have the capabilities to document these services and share this information with other healthcare providers and public health organizations. They assure safe and effective medication and supporting patient-centered care. Pharmacies are highly automated in large hospitals. Again, it may be independent of central EHRs. The Food and Drug Administration (FDA) requires all the drugs to be registered and reported using a National Drug Code (NDC). Coding systems used are NDC, SNOMED, and RxNorm.

2.3.5 Computerized Physician Order Entry (CPOE)

Computerized Physician Order Entry (CPOE) is a very important part of EHRs. It is a system that allows a medical practitioner to enter medical orders and instructions for the treatment of a patient. For example, a doctor can electronically order services to laboratory, pharmacy, and radiology services through CPOE. Then it gets propagated over a network to the person responsible for carrying out these orders. As a digital system, CPOE has the potential to reduce medication-related errors. It is possible to add intelligent rules for checking allergies, contradictions, and other alerts. The primary advantages of CPOE are the following: overcomes the issue of illegibility, fewer errors associated with ordering drugs with similar names, more easily integrated with decision support systems, easily linked to drug-drug interaction warning, more likely to identify the prescribing physician, able to link the adverse drug event (ADE) reporting systems, able to avoid medication errors like trailing zeros, create data that is available for analysis, point out treatment and drug of choice, reduce under- and overprescribing, and finally, the prescriptions can reach the pharmacy quicker. While ordering, a professional can view the medical history, current status report from a different module, and evidence-based clinical guidelines. Thus, CPOE can help in patient-centered clinical decision support.

If used properly, CPOE decreases delay in order completion, reduces errors related to handwriting or transcriptions, allows order entry at point-of-care or off-site, provides error checking for duplicate or incorrect doses or tests, and simplifies inventory and positing of charges. Studies have shown that CPOE can contribute to shortened length of stay and reduction of cost [8]. There are some risks involved in adopting CPOE as well. It may slow down interpersonal communication in an emergency situation. If each group of professionals (e.g., physicians and nurses) works alone in their workstations, it may create ambiguity about the instructions. These factors led an increase in mortality rate by 2.8%–6.5% in the Children's Hospital of Pittsburgh's Pediatric ICU when a CPOE system was introduced [8]. Frequent alerts and warnings may also interrupt workflow. The adaptation rate of CPOE is slow. It may be partly due to physicians' doubt about the value of CPOE and clinical decision support.

2.3.6 Clinical Documentation

A clinical document contains the information related to the care and services provided to the patient. It increases the value of EHR by allowing electronic capture of clinical reports, patient assessments, and progress reports. A clinical document may include [9]

- Physician, nurse, and other clinician notes
- Relevant dates and times associated with the document
- The performers of the care described
- Flow sheets (vital signs, input and output, and problems lists)
- Perioperative notes
- Discharge summaries
- Transcription document management
- Medical records abstracts
- Advance directives or living wills
- Durable powers or attorney for healthcare decisions
- Consents (procedural)
- Medical record/chart tracking
- Release of information (including authorizations)
- Staff credentialing/staff qualification and appointments documentations
- Chart deficiency tracking
- Utilization management
- The intended recipient of the information and the time the document was written
- The sources of information contained within the document

Clinical documents are important because documentation is critical for patient care, serves as a legal document, quality reviews, and validates the patient care provided. Well-documented medical records reduce the re-work of claims processing, compliance with CMS (Centers for Medicare and Medicaid Services), Tricare and other payer's regulations and guidelines, and finally impacts coding, billing, and reimbursement. A clinical document is intended for better communication with the providers. It helps physicians to demonstrate accountability and may ensure quality care provided to the patient. A clinical document needs to be patient centered, accurate, complete, concise, and timely to serve these purposes.

The clinical document architecture (CDA) [10] is an XML-based electronic standard developed by the Health Level 7 International (HL7) to define the structure. It can be both read by human eyes and processed by automatic software.

2.4 Coding Systems

Standards play an important role in enhancing the interoperability of health information systems and the purposeful use of EHR systems. Collecting and storing information following standard coding systems provide better and accurate analysis of the data, seamless exchange of information, improved workflow, and reduced ambiguity. A complete healthcare system is complex and requires various EHR products. Different vendors have implemented standards in their own way. This practice has resulted in a significant variation in the coding practices and implemented methods for which systems cannot interoperate. To create an interoperable EHR, standardization is critical in the following four major areas:

- Applications interaction with the users
- System communication with each other
- Information processing and management
- Consumer device integration with other systems and application

Interoperability between the different EHR systems is a crucial requirement in the "meaningful use of certified EHR technology" to receive incentives. That is why conforming to a standard coding system is very important. In a practical EHR, we need standards for

- Clinical vocabularies
- Healthcare message exchanges
- EHR ontologies

There are three organizations mainly responsible for developing the related standards: Health Level Seven (HL7), Comité Europeen de Normalisation-Technical Committee (CEN-TC), and the American Society of Testing and Materials (ASTM). HL7 develops healthcare-related standards that are widely used in North America. CEN-TC is a prominent standard developing organization working in 19 member states in Europe. Both HL7 and CEN-TC collaborate with ASTM. Along with the standards developed by these organizations, EHR systems must comply with the Health Insurance Portability and Accountability (HIPAA) Act [11] to conserve the security and privacy of patient information.

2.4.1 International Classification of Diseases (ICD)

ICD stands for International Classification of Diseases, which is the United Nations-sponsored World Health Organization's (WHO) official coding standard for diseases, diagnoses, health management, and clinical purposes [12]. It first appeared as the International List of Causes of Death in 1893, adopted by the International Statistical Institute. Since then it has been revised according to advancements in medical science and healthcare. Since the creation of WHO in 1948, WHO has maintained ICD. WHO published ICD-6 in 1949, and it was the first coding system in which morbidity was incorporated [13]. It also included mental disorders for the first time. The U.S. Public Health Services issued International Classification of Diseases, Adapted for Indexing of Hospitals Records and Operation Classification (ICDA) in 1959. It was revised regularly and used to classify diseases and mortality until WHO published the ninth revision of ICD.

The 1967 WHO Nomenclature Regulations specified that the member nations should use the most recent ICD version for mortality and morbidity statistics. Along with the storage and retrieval

of epidemiological and clinical information, it allows for the compilation of morbidity statistics for more than 100 WHO member nations. About 70% of the world's health expenditure in reimbursement and resource allocation is also done using ICD codes [14]. It is used to classify diseases and related problems, and provides a system of codes for a wide variety of diseases, signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease. It is the global foundation for providing common language in disease and health-related information and statistics exchange. ICD is comprehensive and organizes information into standard groups that allows for the following [15]:

- Easy storage, retrieval, and analysis of health information for evidence-based decision-making.
- Sharing and comparing health information between hospitals, regions, and countries.
- Data comparison in the same location across different time periods.

2.4.1.1 ICD-9

ICD ninth revision is the most popular coding system published by WHO in 1978. It was designed to promote comparability of classification, collection, processing, and presentation of mortality statistics. Its clinical modification, ICD-9-CM, was published by the U.S. Public Health Services in the following year to meet the statistical needs. The modified version had expanded the number of diagnostic codes and developed a procedure coding system. It has more than 13,000 codes and uses more digits representing the codes compared to ICD-9. It is the system that is used to encode all the diagnoses for healthcare services in the United States. It is maintained by the National Center for Health Statistics (NCHS) and the Center for Medicare and Medicaid Services (CMS). Both the departments are part of the federal department of Health and Human Services. The ICD-9-CM code set is organized in three volumes and consists of tabular lists and alphabetical indices.

- Volume 1: Disease and Injuries Tabular List
- Volume 2: Disease and Injuries Alphabetical Index
- Volume 3: Procedures Tabular List and Alphabetic Index

ICD-9-CM is updated every year to keep up-to-date with medical trends and diseases. NCHS has the responsibility to update Volumes 1 and 2, and CMS maintains Volume 3. Concerned parties from both the public and private sectors can propose changes to it. The major updates take effect on October 1 every year and minor updates occur on April 1. It is a statistical tool that converts the diagnoses and procedures into number codes. Its primary applications are

- Reporting and research
- Monitoring the quality of patient care
- Communication and transactions
- Reimbursement
- Administrative uses

2.4.1.2 ICD-10

The tenth version was endorsed by WHO in 1990 during the 43rd World Health Assembly. The first full version of ICD-10 was released in 1994. The first step of implementing ICD-10 was taken by NCHS awarding a contract to the Center for Health Policy Studies (CHPS) to evaluate ICD-10 for morbidity purposes within the United States. A prototype of clinically modified ICD-10 was developed after a thorough evaluation of ICD-10 by a technical advisory panel. After strong recommendations, NCHS proceeded with implementing a revised version of ICD-10-CM. During 1995–1996, further work on the enhancement of ICD-10-CM was done incorporating experiences from ICD-9-CM and through collaborating with many speciality groups like American Association of Dermatology, American Academy of Neurology, American Association of Oral and Maxillofacial Surgeons, American Academy of Orthopedic Surgeons, American Academy of Pediatrics, American College of Obstetricians and Gynecologists, American Urology Institution, and National Association of Children hospitals and other related institutions. In 1999, ICD-10 was implemented in the United States for mortality reporting. Death statistics and data regarding leading causes of death for the years 1999 and 2000 were published using ICD-10 [16]. In October 2002, ICD-10 was published in 42 languages. In June/July 2003, the American Health Information Management Association (AHIMA) and American Hospital Association (AHA) jointly conducted a pilot study to test ICD-10-CM. In their study, they have compared ICD-9-CM and ICD-10-CM and the initial results indicated ICD-10-CM is an improvement over ICD-9-CM; and ICD-10-CM is more applicable in non-hospital environments compared to ICD-9-CM. Canada, Australia, Germany, and others countries have their own revision of ICD-10 by adding country specific codes. The revisions are ICD-10-CA, ICD-10-AM, ICD-10-GM, and so on. The standard for procedure codes ICD-10-PCS was also developed during the same time frame to replace the Volume 3 of ICD-9-CM. The first revision of it was released in 1998.

ICD-9-CM is around thirty years old. Many of its categories are full, and there have been changes in technology. Some of them are also not descriptive enough. A newer coding system is needed, which would enhance reimbursement, better facilitate evaluation of medical processes and outcomes, and be flexible enough to incorporate emerging diagnoses and procedures. For example, in a scenario where a patient had a fractured left wrist and, after a month a fractured right wrist, ICD-9-CM cannot identify left versus right; additional information is required. However, ICD-10-CM can report distinguishing left from right. It can also characterize initial and subsequent encounters. Further, it can describe routine healing, delayed healing, nonunion, or malunion.

The major differences between ICD-10 and ICD-9-CM are [17]

- ICD-10 has 21 categories of diseases; while ICD-9-CM has only 19 categories.
- ICD-10 codes are alphanumeric; while ICD-9-CM codes are only numeric.
- ICD-9-CM diagnoses codes are 3–5 digits in length, while ICD-10-CM codes are 3–7 characters in length.
- Total diagnoses codes in ICD-9-CM is over 14,000; while ICD-10-CM has 68,000.
- ICD-10-PCS procedure codes are 7 characters in length; while ICD-9-CM procedure codes are 3–4 numbers in length.
- ICD-10-PCS total number of codes is approximately 87,000. The number of procedure codes in ICD-9-CM is approximately 4,400.

The Center for Medicare and Medicaid Services (CMS) guidelines mandated a conversion from ICD-9-CM to ICD-10-CM by October 1, 2014 in the United States. Adopting a new coding system will have the following benefits:

- Improve patient care. The increased detail in the coding system will improve the measurement of quality, safety, and efficacy of care, which will ultimately lead to improved patient care.
- Determine the severity of illness and prove medical necessity. ICD-10 codes are more granular and provide option to input the level of sickness along with complexity of disease of a patient in a code-based system.
- Improve research. The better and more accurate organization of code will be able to more precisely classify diseases and injuries, and correlate them with the cause, treatment, and outcome. The collected data will be less ambiguous and such a better-defined structure of the information will make data analysis easier. Information processing will be easier with newer coding system and it will open new opportunities for developing an intelligent prediction system. It will also allow the United States. to conduct comparative research with other countries that are already using ICD-10.
- Lend insight to the setting of health policy. With improved data analytics made possible through ICD-10, policy makers will be able to make informed policy decisions.
- Facilitate improved public health reporting and tracking. The comprehensive coding structure will allow concerned agencies to track public health risks and trends in greater detail.
- Improve clinical, financial, and administrative performance and resource allocation. The quality of data can reveal essential insights. It will allow the administrators to track time and workforce spent for procedures. This will help administrators to allocate resources more efficiently and achieve positive financial and managerial outcomes.
- Increase the accuracy of payment and reduce the risk that claims will be rejected for incorrect coding. Reduced number of claim denials is expected due to higher specificity of ICD-10. It will also create a better electronic record of evidence to receive proper payment from government payers, insurers, hospitals, health systems, and others.
- Make room for new procedures and techniques. The adaptation ability of ICD-9-CM is limited, where all the codes are already utilized and has no more room for new codes. The expanded coding of ICD-10 will be able to accommodate new procedures.
- It will have other facilities like reduced hassle of audits, help preventing and detecting healthcare fraud and abuse.

2.4.1.3 ICD-11

The World Health Organization is currently working on the eleventh revision of ICD. The final publication of ICD-11 is expected by 2017 [18]. The beta draft [19] was made public online for initial comments and feedback in May 2012. This development of ICD-11 revisions is taking place in a web-based platform called iCAT, where all the concerned parties collaborate. For interested groups or people, there are options to give structured input and field testing of revised editions. It will be available in multiple languages and free to download for personal use. In ICD-11, disease entries will have definitions and descriptions of the entry and category in human readable forms. The current version ICD-10 has only the title headings. There are 2,400 codes in ICD-11 that are different in the ICD-10 code set, where 1,100 codes are related to external causes and injury [20].

Although the beta version does not support any social network platforms, the support of websites such as Wikipedia, Facebook, Social Reader, LinkedIn, etc. is in the plan. The structure of definitions and other contents related to diseases and procedures will be defined more accurately. It will be more compatible with EHRs and other technologies.

2.4.2 Current Procedural Terminology (CPT)

Current Procedural Terminology (CPT) is a set of medical codes developed, maintained, and copyrighted by the American Medical Association (AMA). CPT codes are a list of descriptive terms, guidelines, and identifying codes of medical, surgical, and diagnostic services designed to provide uniform communication language among physicians, coders, patients, accreditation organizations, and payers for administrative, financial, and analytic purposes.

It was first created by the AMA in 1966. The first edition contained mainly surgical codes. A significant development took place for the second edition, which was published in 1970. The second edition contained 5 digits instead of 4 digits, and it included lab procedures. In 1983, the Health Claim Financial Administration (HCFA), which is now known as the Center for Medicine and Medicaid Services (CMS), merged its own Common Procedure Coding System (HCPCS) with CPT and mandated CPT would be used for all Medicare billing. Every year the new version is released in October. The Healthcare Common Procedures Coding System (HCPCS, often pronounced as "hick picks") is another set of codes developed by AMA based on CPT. Although the CPT coding system is similar to ICD-9 and ICD-10, it describes the treatment and diagnostic services provided while ICD codes describe the condition or the disease being treated. CPT is used only in inpatient settings.

2.4.3 Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)

Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) is a comprehensive, computer-processible, multilingual clinical and healthcare terminology, originally created by the College of American Pathologists (CAP). SNOMED was started as Systematic Nomenclature of Pathology (SNOP) in 1965 [21]. It was enhanced further and SNOMED was created in 1974. It had two major revisions in 1979 and 1993. In 1999, SNOMED-CT was created by the merger of SNOMED Reference Terminology (SNOMED-RT) developed by the CAP and Clinical Terms Version 3 (CTV3) developed by the National Health Services of the United Kingdom. This merged version was first released in 2002. SNOMED-RT had a vast coverage of medical specialities with over 12,000 concepts. It was designed for the retrieval and aggregation of healthcare information produced by multiple organizations or professionals. The strong suit of CTV3 was its coverage of terminologies for general practice. With more than 200,000 concepts, it was used to store primary care encounter information and patient-based records [22]. Currently SNOMED has more than 311,000 concepts with logic-based definitions organized into a hierarchy. In July 2003, the National Library of Medicine (NLM) on behalf of the U.S. Department of Health and Human Services signed a contract with CAP to make SNOMED-CT available for users. Since April 2007, it has been owned, maintained, and distributed by a newly formed Denmark-based nonprofit organization named International Health Terminology Standards Development Organization (IHTSDO) [9]. CAP collaborates with IHTSDO and continues to provide support for SNOMED-CT operations. More than 50 countries use SNOMED-CT.

SNOMED-CT is a valuable part of EHR. Its main purpose is to encode medical and healthcare-related concepts and support recording of data. It provides a consistent way to store, index, retrieve, and aggregate clinical data across different sites. It also helps to organize data in a more meaningful way and reduce the variability of the data collection and management process. Its extensive coverage includes clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices, and specimens [23].

SNOMED-CT has a logical and semantic relationship between concepts. It has a multiaxial hierarchy, which allows different level of details of information. Its extensible design permits the integration of national, local, and vendor specific requirements. It primarily consists of four components.

- Concept Codes: numerical codes to identify terms

- Descriptions: textual descriptions of the concept codes
- Relationships: represents relationships between the concept codes
- Reference Sets: used for grouping concept codes or descriptions. Supports cross mapping to other classification standards.

SNOMED-CT can be mapped to other well-known terminologies like ICD-9-CM, ICD-10, and LOINC. Renowned standards like ANSI, DICOM, HL7, and ISO are supported by it. In a joint project with WHO, it is providing insights for the upcoming ICD-11.

SNOMED-CT has some fundamental differences from ICD. It is mainly a terminology system while ICD is a classification system. SNOMED-CT is designed to encode and represent data for clinical purposes [24]. Information coded with ICD is used for statistical analysis, epidemiology, reimbursement, and resource allocation. SNOMED-CT facilitates the information input into the EHR and provides standardization for primary data purposes while ICD codes enable retrieval for secondary data purposes.

2.4.4 Logical Observation Identifiers Names and Codes (LOINC)

Logical Observation Identifiers Names and Codes (LOINC) is a universal code system for identifying laboratory observations and clinical test results. In response to the demand for electronic clinical data, it was created in 1994 by Regenstrief Institute Inc., an Indianapolis-based nonprofit research organization affiliated with Indiana University. It was originally called Laboratory Observations, Identifiers, Names, and Codes and the development was sponsored by NLM and other government and private agencies. Original sources of information include the following [25]:

- Silver book for International Union of Pure and Applied Chemistry
- International Federation of Clinical Chemistry
- Textbooks of Pathology
- EuCliD (European Clinical Database)
- Expertise and work of the LOINC members

LOINC coding system helps to improve the communication of information. In January 2009, Regenstrief Institute released a Windows operating system-based mapping software called Regenstrief LOINC Mapping Assistant (RELMA) where codes can be searched and local codes can be mapped to a LOINC database. The current version of LOINC is LOINC 2.46 released in December 2013. With more than 600 new users per month, it has 27,000 users from 158 different countries. LOINC vocabulary continues to grow till today.

Each LOINC record represents a single test result. A record consists of six fields [26].

- Component: what is measured and evaluated (e.g., glucose, hemoglobin)
- Kind of property: characteristics of the component that is measured (e.g., mass, length, concentration, volume, time stamp, etc.)
- Time: observation period of the measurement
- System: the specimen or the substance, in context of which the measurement was done (e.g., blood, urine)
- Scale: the measurement scale (e.g., quantitative, nominal, ordinal, or narrative)

- Method (optional): the procedure performed for measurement

Certain parameters and descriptors related to the test are explicitly excluded in LOINC from observation name. They are made as fields of test/observation report message [25]. These fields are

- The instrument used for testing
- Fine details of the sample or the site of collection
- The priority of the test
- Who verified the result
- Size of the sample
- Place of testing

LONIC's overall organization is divided into four categories: laboratory, clinical, attachments, and surveys. The laboratory component is further divided into subcategories such as chemistry, hematology, serology, microbiology (includes parasitology and virology), and toxicology. The clinical attributes are vital signs, hemodynamics, intake/output, EKG, obstetric ultrasound, cardiac echo, urologic imaging, gastroendoscopic procedures, pulmonary ventilator management, and other clinical observations [25]. It also contains information about nursing diagnoses and nursing interventions.

2.4.5 RxNorm

RxNorm is a drug vocabulary maintained and distributed by the National Library of Medicine [27]. It assigns standard names to the clinical drugs and drug delivery devices available in the United States. It is used as a basis for the capture and presentation of drug-related information in EHRs. In 2001, NLM started to develop RxNorm for modeling clinical drugs in the Unified Medical Language System (UMLS) in consultation with the HL7 vocabulary technical committee and the Veterans Administration [28]. It was developed to standardize the medication terminology that would reduce the missed synonymy in clinical drugs [29]. Additional goals were to facilitate electronic capture of related data, improve interoperability by supporting information exchange across platforms and systems, develop clinical decision support, and provide opportunity for research.

RxNorm follows a standard for naming drugs. The normalized name of a drug include the following components [28]:

- IN: Ingredient of the drug.
- DF: Dose form of the drug.
- SCDC: Semantic clinical drug component. It represents the ingredients and strength.
- SCDF: Semantic clinical drug form. It represents the ingredient and dose form.
- SCD: Semantic clinical drug. It represents the ingredient, strength, and dose form.
- BN: Brand name. This is the formal name for a group of drugs containing a specific active ingredient.
- SDBC: Semantic branded drug component. It represents the branded ingredient and strength.
- SBDF: Semantic branded drug form. It represents the branded ingredient and dose form.
- SDB: Semantic branded drug. It represents the branded ingredient, strength, and dose form.

RxNorm organizes drugs by concept. A concept is a set of names with similar meaning at a specific level of abstraction. It can distinguish similar drugs from different providers using concepts. The concepts and relationships between each other form a semantic network.

2.4.6 International Classification of Functioning, Disability, and Health (ICF)

The International Classification of Functioning, Disability, and Health, commonly known as ICF, is a classification of health-related components of function and disability. ICF concentrates on the functionality and body structure of people with a given health condition or disability rather than diagnosis or diseases. It does not account for the cause of disability. It is a unified and standard framework first developed by the World Health Organization (WHO) in 1980 [30]; initially it was known as International Classification of Impairments, Disabilities, and Handicaps (ICIDH). After years of coordinated revision, in May 2001, the 191 member states of WHO agreed to adopt ICF as the standard coding method of functioning and disability. In June 2008, the American Physical Therapy Association (APTA) joined WHO for endorsing ICF. ICF is the only method of its kind. It has been developed and tested for applicability in more than 40 countries.

Body functions and disability can be viewed as interactions between health condition and personal and environmental factors. ICF has mainly two parts: Functioning and disability, and Contextual factors. It can be categorized into further subparts. The components of ICF are listed below [31]:

- Functioning and disability
  - Body functions
    - Mental functions
    - Sensory functions and pain
    - Voice and speech functions
    - Functions of the cardiovascular, hematological, immunological, and respiratory systems
    - Genitourinary and reproductive functions
    - Neuromusculoskeletal and movement-related functions
    - Functions of the skin and related structures
  - Body structures
    - Structure of the nervous system
    - The eye, ear, and related structures
    - Structures involved in voice and speech
    - Structures related to cardiovascular, immunological, and respiratory systems
    - Structures related to digestive, metabolic, and endocrine systems
    - Structures related to genitourinary and reproductive systems
    - Structures related to movement
    - Skin and related structures
  - Activities and participation
    - Learning and applying knowledge
    - General tasks and demands
    - Communication
    - Self-care
    - Domestic life

    - Interpersonal interactions and relationships
    - Major life areas
    - Community, social, and civic life

- Contextual factors
  - Environmental factors
    - Products of technology
    - Natural environment and human-made changes to the environment
    - Support and relationships
    - Attitudes
    - Service, systems, and policies
  - Personal factors
    - Gender
    - Age
    - Coping styles
    - Social background
    - Education
    - Profession
    - Past and current experience
    - Overall behavior pattern
    - Character and other factors

ICF complements WHO's classification of disease scheme, ICD-10. ICD contains diagnosis and health condition-related information, but not functional status. Together they constitute the WHO Family of International Classifications (WHO-FIC) shown in Figure 2.2.

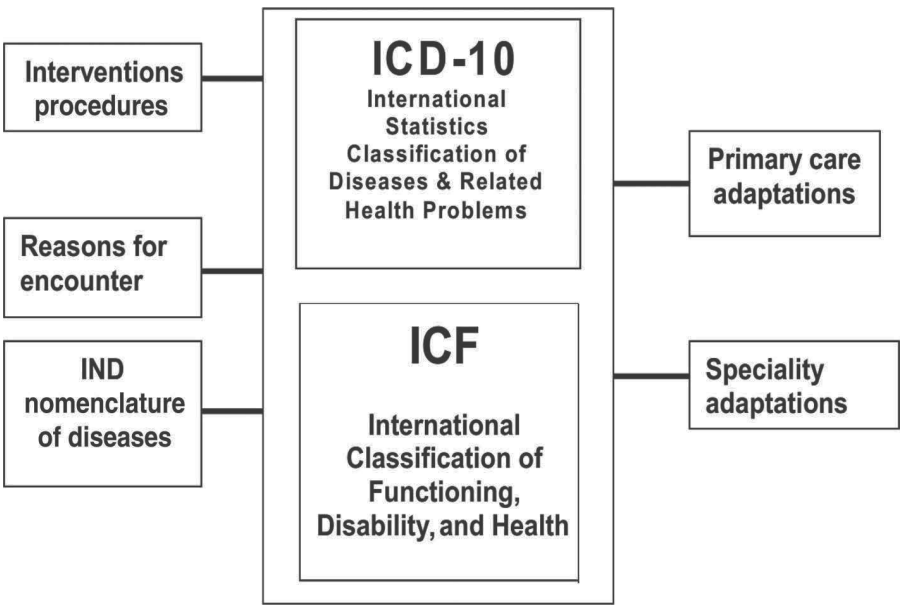

FIGURE 2.2: WHO Family of International Classifications taken from [32].

Diagnosis is used to define cause and prognosis of diseases, but by itself it does not predict service needs, length of hospitalization, or level of care of functional outcomes. Nor can it accurately provide support for disability. ICF allows incorporating all aspects of a person's life. The current ICF creates a more understandable and comprehensive profile of health forming of a person instead of focusing on a health condition [33]. It is used as a clinical, statistical, research, social policy, and educational tool. A common misconception about ICF is that it deals with only the disabled people. However, ICF has some limitations regarding the ability to classify the functional characteristics of developing children [34].

2.4.7 Diagnosis-Related Groups (DRG)

Diagnosis-Related Groups (DRG) are a patient classification scheme that group related patients and relate these groups with the costs incurred by the hospital. DRGs divide diagnosis and illness into 467 categories identified in ICD-9-CM [35]. The 467th group is "ungroupable." The classification is based on a patient's principal diagnosis, ICD diagnoses, gender, age, sex, treatment procedure, discharge status, and the presence of complications or comorbidities. The goals of developing DRGs were to reduce healthcare cost, and improve quality of care and efficiency of the hospitals. DRGs are by far the most important cost control and quality improvement tool developed [36].

It was first created at Yale University with the support from the Health Care Financing Administration, now known as the Center for Medicine and Medicaid Service (CMS). In 1980, it was first implemented in a small number of hospitals in New Jersey [37]. It is used to define the reimbursement amount of hospitals from Medicare. Medicare pays hospitals per patient and efficient hospitals receive better incentives. DRGs help to decide the efficiency of the hospital.

2.4.8 Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) is a collection of comprehensive biomedical concepts and ontologies. It was developed by the U.S. National Library of Medicine (NLM) in 1986. It provides the development of computer-based systems that can behave as through they understand the biomedical and health concepts [38]. It is intended to be mainly used by medical informatics professionals. NLM maintains and distributes UMLS knowledge sources (database) and related software tools for developers to build enhanced electronic information system that can create process, retrieve, integrate, and/or aggregate health and biomedical-related information. The knowledge sources of UMLS are as follows [39]:

- Metathesaurus
  - Source Vocabularies
  - Concepts
- Relationships, Attributes
  - Semantic Network
  - Semantic Types (categories)
  - Semantic Relationships
- Lexical Resources
  - SPECIALIST Lexicon
  - Lexical Tools

Metathesaurus is a very large, multipurpose, and multilingual vocabulary database. It contains health and biomedical-related concepts of their various names and the relationships among them. It has 126 vocabularies in 17 languages [27]. It clusters similar terms into a concept. The semantic network provides consistent categorization of concepts defined in Metathesaurus. The network contains information regarding basic semantic types/categories that may be assigned to concepts and relationships between semantic types. In the semantic network, the semantic types are nodes and the relationships are links between them. In the current version of semantic network, there are 135 semantic types and 54 relationships [38]. The SPECIALIST Lexicon provides the lexical information needed for the SPECIALIST natural language processing tool.

2.4.9 Digital Imaging and Communications in Medicine (DICOM)

The Digital Imaging and Communications in Medicine (DICOM) is a medical imaging standard. It determines the data exchange protocol, digital image format, and file structure for biomedical images and related information [40]. DICOM was developed by the American College of Radiology (ACR) and National Electric Manufacturers Association (NEMA). The first version ACR/NEMA 300 was released in 1985. DICOM is generally used in the following application areas [40]

- Network image management
- Network image interpretation management
- Network print management
- Imaging procedure management
- Offline storage media management

DICOM allows the integration of scanners, servers, workstations, printers, and network hardware into a Picture Archiving and Communication Systems (PACS). It has been extensively used by the hospitals and other organizations. It provides a widely accepted foundation for medical imaging standards. It promotes interoperability between radiology systems.

2.5 Benefits of EHR

EHRs are transformational tools. The scope of paper-based systems is severely limited. We need EHRs to improve the quality of patient care and increase productivity and efficiency. In terms of the overall management and costs, EHRs are a better choice. They also help in complying with government regulations and other legal issues. The benefits of EHRs are described in this section.

2.5.1 Enhanced Revenue

An EHR system can capture the charges and bills for clinical services provided, laboratory tests, and medications more accurately. Utilization of electronic systems decrease billing errors [41]. They also provide a better documentation opportunity for these services that can be used to resolve financial disputes. Better management of information yield more accurate evaluation and increase reimbursements. According to experts, due to inaccurate coding systems, 3%–15% of a healthcare provider's total revenue is lost [42]. An EHR system can be programmed or configured to generate alerts for both patients and doctors when a healthcare service is due. This can aid better management of collecting revenue. It can be used to garner more revenues by incorporating services like

telemedicine, e-visits, virtual office visits, etc. *It is true that all kinds of services are not possible over the Internet or telephone network, but not all diseases will require extensive diagnosis and laboratory testing*. Diseases commonly treated through telemedicine include acne, allergies, cold and flu, constipation, diabetes, fever, gout, headache, joint aches and pains, nausea and vomiting, pink eye, rashes, sinus infection, sore throat, sunburn and urinary tract infections, anxiety and depression, etc.

2.5.2 Averted Costs

After adopting electronic systems, some costs associated with the previous way of operating a business are eliminated. The Center for Information Technology leadership suggested that the use of EHRs will save a total of $44 billion each year [43]. Adopting EHR has the following averted costs [44].

- **Reduced paper and supply cost:** To maintain paper-based health records an organization will require a lot of paper, printing materials, and other supplies. Adopting EHR will reduce these costs. After adopting EHRs, one organization estimated a reduction of 90% of paper usage within a few months [45].

- **Improved utilization of tests:** In electronic systems, test results are better organized. A healthcare staff no longer needs to carry the reports from one place to another. Identifying redundancy or unnecessary tests is easier. This can reduce the loss of information and ensure improved utilization of tests. A study by Wang et al. [41] reports better utilization of radiology tests after adopting EHRs.

- **Reduced transcription costs:** An EHR can reduce transcription costs for manual administrative processes [46, 47]. It utilizes structured flow sheets, clinical templates, and point-of-care documentation. In a typical outpatient setting, physicians generate about 40 lines of transcription per encounter. For a group of three practicing physicians, treating 12,000 patients annually at the cost of $0.11 for each transcription line results in over $50,000 per year [46]. A study of fourteen solo or small-group primary care practices in twelve U.S. states reports the median transcription cost saving to be $10,800, where a minimum saving was $8,500 and a maximum was $12,000 for the year 2004–2005 [47]. Other related research work also describes saving $1,000–$3,000 per physician, per month [48].

- **Improved productivity:** EHR helps to improve workflows by utilizing resources more efficiently and reducing redundancies. As a result, the overall productivity of individuals increases.

- **Better availability of information and elimination of chart:** In EHR, all the charts are in digital format. It eliminates the need to pull, route, and re-file paper charts [46]. A significant amount of effort is spent on creating, filing, searching, and transporting paper charts [49]. A study estimated that the elimination of paper charts can save $5 per chart pull [41]. It is also comparatively easier to manage digital charts.

- **Improved clinician satisfaction:** Electronic technology can save time by reducing the paperwork burden, which can create additional time for patient encounters and delivery of care [3]. A study reports the use of EHR has reduced the physician's office visit time by 13% and a nurse's pre-exam interview time by 1 minute [50]. This can improve satisfaction for professionals, which can indirectly enhance revenue.

2.5.3 Additional Benefits

EHR offers many additional benefits that are discussed in more detail below.

- **Improved accuracy of diagnosis and care:** EHR provides comprehensive and accurate patient information to physicians that can help to quickly and systematically identify the correct problem to treat. EHRs do not just contain the patient information; they have the capability to perform computation and make suggestions. They can also present comparative results of the standard measurements. A U.S. national survey of doctors demonstrates the following [51]:
  - 94% of the providers report EHR makes records readily available at the point of care.
  - 88% report that EHR produces clinical benefits for their practice.
  - 75% report that EHR allowed them to deliver better patient care.

  The gathered information can guide a physician in the emergency department to take prudent and safer actions. Such services are unimaginable with paper-based systems. Diagnostic errors are difficult to detect and can be fatal to a patient. A new study suggests that EHR can help to identify potential diagnostic errors in primary care by using certain types of queries (triggers) [52].
- **Improved quality and convenience of care:** EHRs have the potential to improve the quality of care by embedding options such as Clinical Decision Support (CDS), clinical alerts, reminders, etc. Research suggests that EHRs are linked to better infection control [53], improved prescribing practices [12], and improved disease management [42] in hospitals. In such applications, convenience is also an important measure. EHRs greatly reduce the need for patients to fill out similar (or even sometimes the same) forms at each visit. Patients can have their e-prescriptions ready even before they leave the facility and can be electronically sent to a pharmacy. Physicians and staff can process claims insurance immediately. Following are the results of a study on the effects of e-prescribing reports [54].
  - 92% patients were happy with their doctor using e-prescribing.
  - 90% reported rarely or only occasionally having prescriptions not ready after going to the pharmacy.
  - 76% reported e-prescribing made obtaining medications easier.
  - 63% reported fewer medication errors.
- **Improved patient safety:** Just like improving the quality of care, clinical decision support systems (CDSS) and computerized physician order entry (CPOE) have the potential to improve patient safety. Medication errors are common medical mistakes and in the United States it is responsible for the death of a person every day on average as well as injuring more than a million annually [55]. Research shows that utilization of CPOE can reduce medication errors [56, 57]. Medication errors can occur at any stage of the medication administration process from a physician ordering the drug, followed by the dispensing of the drug by the pharmacist, and finally the actual administration of the drug by the nurse. CPOE is a technology that allows physicians to act on a computerized system that introduces structure and control. Along with patient information, EHR holds the medication records for a patient. Whenever a new medication is prescribed, it can check for potential conflicts and allergies related to the particular medication and alert the physician. The system also can provide the chemical entities present in the drug and cross-reference allergies, interactions, and other possible problems related to the specific drug. Introducing technologies such as Barcode Medication Administration can make the system even more accurate. The Institute of Medicine (IOM) recommends CPOE and CDS as main information technology mechanisms for increasing patient safety in the future [58].

- **Improved patient education and participation:** In an EHR system, certain features can provide simplified patient education [42]. EHRs can be used by the provider as a tool to illustrate procedures and explain a patient's conditions. It can increase a patient's participation by offering follow-up information, self-care instructions, reminders for other follow-up care, and links to necessary resources. Information technology affects every part of our life. In this digital era, patients may feel more comfortable with an electronic system.

- **Improved coordination of care:** EHRs are considered essential elements of care coordination. The National Quality Forum defines care coordination as the following [59]: "Care coordination is a function that helps ensure that the patient's needs and preferences for health services and information sharing across people, functions, and sites are met over time. Coordination maximizes the value of services delivered to patients by facilitating beneficial, efficient, safe and high-quality patient experiences and improved healthcare outcomes." For a patient with multiple morbidities, a physician is responsible for providing primary care services and coordinating the actions of multiple subspecialists [60]. According to a Gallup poll [61], it is a common scenario for older patients to have multiple doctors: no physician 3%, one physician 16%, two physicians 26%, three physicians 23%, four physicians 15%, five physicians 6%, and six or more physicians 11%. EHRs allow all clinicians to document services provided and access up-to-date information about their patient. It streamlines the transition process and knowledge sharing between different care settings. This facilitates an improved level of communication and coordination [62]. Research suggests that the clinicians having 6+ months use of EHRs reported better accessing and completeness of information than clinicians without EHRs. Clinicians having EHRs have also reported to be in agreement on treatment goals with other involved clinicians [63].

- **Improved legal and regulatory compliance:** As organizations develop their systems, it is important to understand and comply with many federal, state, accreditation, and other regulatory requirements. A health record is the most important legal and business record for a healthcare organization. The use of an EHR system will provide more security and confidentiality of a patient's information and thus, comply with regulations like HIPAA, Consumer Credit Act, etc. Moreover, the Center for Medicare and Medicaid Services (CMS) has financial incentive programs for hospitals regarding the meaningful use of health information technology. To receive the financial reimbursement, professionals have to meet a certain criteria and can get up to $44,000 through Medicare EHR Incentive Program and up to $63,750 through the Medicaid EHR Incentive Program [64]. Adaptation of certified EHR can help providers get reimbursed.

- **Improved ability to conduct research and surveillance:** In conjunction with the direct use of EHR in primary patient care, there is an increasing recognition that secondary use of EHR data can provide significant insights [65]. Using quantitative analysis of functional values, it has the potential to identify abnormalities and predict phenotypes. Pakhomov et al. demonstrated the use of text processing and NLP to identify heart failure patients [66]. EHR data can be used to predict survival time of patients [67]. Data from different EHRs can be integrated into a larger database and geo-location specific surveillance is also possible.

- **Improved aggregation of data and interoperability:** Standards play a crucial role in data aggregation and interoperability between different systems. EHRs maintain standard procedure and follow defined coding system while collecting data. This accommodates easier aggregation of data and greater interoperability, which offer the following benefits [68].

  - Manage increasingly complex clinical care
  - Connect multiple locations of care delivery

- Support team-based care
- Deliver evidence-based care
- Reduce errors, duplications, and delay
- Support ubiquitous care
- Empower and involve citizens
- Enable the move to the Personal Health Paradigm
- Underpin population health and research
- Protect patient privacy

We need high-quality aggregated data from multiple sources in order to make evidence-based decisions. The level of achievable interoperability using EHRs is unthinkable from paper-based systems. The American Medical Association recognizes that enhanced interoperability of EHRs will further help to attain the nation's goal of a high-performing healthcare system.

- **Improved business relationships:** A healthcare provider organization equipped a with superior EHR system can be in a better bargaining position with insurers and payers compared with less equipped ones. The next generation of business professionals will expect and demand a state-of-the-art information healthcare technology system.
- **Improved reliability:** Data is more reliable in a digital format. Due to the reduction of storage costs, having multiple copies of data is possible.

2.6 Barriers to Adopting EHR

Despite of having great potential of EHRs in medical practice, the adoption rate is quite slow and faces a range of various obstacles. Many other developed countries are doing far better than the United States. Four nations (United Kingdom, the Netherlands, Australia, and New Zealand) have almost universal use (each $\sim$90%) of EHRs among the general practitioners. In contrast, the United States and Canada have only around 10–30% of the ambulatory care physicians using EHRs [69]. Health informatics has been a high priority in other developed nations, while until recently, the degree of involvement and investment by the U.S. government in EHRs has not been significant. Major barriers to adopting EHRs are discussed below.

- **Financial barriers:** Although there are studies that demonstrate financial savings after adopting EHRs, the reality is that the EHR systems are expensive. Several surveys report that the monetary aspect is one of the major barriers of adopting EHRs [70, 71, 72, 73, 74, 75, 76]. There are mainly two types of financial costs, start-up and ongoing. A 2005 study suggests that the average initial cost of setting up an EHR is $44,000 (ranging from a minimum of $14,000 to a maximum of $63,000) and ongoing costs average about $8,500 per provider per year [47]. Major start-up costs include purchasing hardware and software. In addition, a significant amount of money is also required for system administration, control, maintenance, and support. Long-term costs include monitoring, modifying, and upgrading the system as well as storage and maintenance of health records. Besides, after the substantial amount of investment, physicians are worried that it could take up to several years for the return on the investment.

An EHR is not the only electronic system that exists in any healthcare provider like practice management. There might be other old systems that also need integration into the new system. It is important that an EHR system is integrated into other systems, and this integration can sometimes be very expensive. Surveys show that due to the high financial investment required, EHR adaptation was far higher in large physician practices and hospitals [77].

- **Physician's resistance:** To adopt EHRs, physicians have to be shown that new technology can return financial profits, saves time, and is good for their patients' well-being. Although research-based evidence is available, it is difficult to provide concrete proof of those benefits. As given in a report by Kemper et al. [76], 58% of physicians are without any doubt that EHR can improve patient care or clinical outcomes. Finally, adopting EHRs in a medical practice will significantly change the work processes that physicians have developed for years.

  Besides, physicians and staffs might have insufficient technical knowledge to deal with EHRs, which leads them to think EHR systems are overly complex. Many physicians complain about poor follow-up services regarding technical issues and a general lack of training and support from EHR system vendors [72]. A study reports that two-thirds of physicians expressed inadequate technical support as a barrier to adopting EHRs [75]. Some physicians are also concerned about the limitation of EHR capabilities. Under certain circumstances or as time passes, the system may no longer be useful [71, 74]. Besides, all physicians do not perform the same operations. EHR systems have to be customizable to best serve each purpose. Surveys suggest that one of the reasons for not adopting EHRs is that the physicians cannot find a system that meets their special requirements [71, 72, 73, 75, 78, 76]. However, an increased effort and support from vendors may play a role in motivating physicians towards adopting EHRs.

- **Loss of productivity:** Adoption of an EHR system is a time-consuming process. It requires a notable amount of time to select, purchase, and implement the system into clinical practice. During this period physicians have to work at a reduced capacity. Also, a significant amount of time has to be spent on learning the system. The improvement will depend on the quality of training, aptitude, etc. The fluent workflow will be disrupted during the transition period, and there will be a temporary loss of productivity [79].

- **Usability issues:** EHR software needs to be user-friendly. The contents of the software must be well-organized so that a user can perform a necessary operation with a minimal number of mouse clicks or keyboard actions. The interface of software workflow has to be intuitive enough. In terms of usability, a comprehensive EHR system may be more complex than expected. It has to support all the functionalities in a provider's setting. There might be a number of modules and submodules, so the user might get lost and not find what he is looking for. This has the potential to hamper clinical productivity as well as to increase user fatigue, error rate, and user dissatisfaction. Usability and intuitiveness in the system do not necessarily correlate to the amount of money spent. The Healthcare Information and Management Systems Society (HIMSS) has an EHR usability task force. A 2009 survey by the task force reported 1,237 usability problems, and the severity of 80% of them was rated "High" or "Medium" [80]. Apart from the workflow usability issue, other related issues are configuration, integration, presentation, data integrity, and performance. The task force defined the following principles to follow for effective usability [81]: simplicity, naturalness, consistency, minimizing cognitive load, efficient interactions, forgiveness and feedback, effective use of language, effective information presentation, and preservation of context.

- **Lack of standards:** Lack of uniform and consistent standards hinders the EHR adoption. Standards play an integral role in enabling interoperability. CMS reimbursement for meaningful use requires EHR systems to demonstrate the ability to exchange information. Many

of the currently used systems have utility only for certain specific circumstances. Different vendors have developed systems in different programming languages and database systems. They do not have any defined best practice or design patterns. This makes the data exchange difficult or impossible between the systems [73, 74, 78]. This lack of standardization limits the proliferation of EHRs [82]. While large hospital systems have moved to EHRs, many others are skeptical about the available systems. They fear that the EHR software they buy now might not work with standards adopted by the healthcare industry or mandated by the government later on.

- **Privacy and security concerns:** Health records contain personal, diagnostics, procedures, and other healthcare related sensitive information. Due to the immense importance of this information, an EHR system may be subjected to attack. Some of the medical diagnoses are considered socially stigmatized, like sexually transmitted disease. Some information relates to direct life threats, like allergies. Employers as well as insurance companies may be interested to know more about a patient to make unethical decisions whether to cover a patient and/or his specific diagnosis. It can also influence some of the hiring decisions. EHRs contain information like social security numbers, credit card numbers, telephone numbers, home addresses, etc., which makes EHRs attractive target for attackers and hackers. A patient might even be motivated to alter his or her medical records to get worker's compensation or to obtain access to narcotics. Therefore, it is important that the privacy and security of EHRs are well maintained. The most used certification for privacy and security is given by the Certification Commission for Healthcare Information Technology (CCHIT). The CCHIT website claims that by mid-2009, 75% of EHR products in the marketplace were certified [83]. In addition to that, the Health Information Technology for Economic and Clinical Health (HITECH) Act introduced a new certification process sponsored by the Office of the National Coordination for Health Information Technology (ONC) in 2009. In January 2010, the ONC released the interim final rule that provides an initial set of standards, implementation specifications, and certification criteria of EHR technology. Its requirement includes database encryption, encryption of transmitted data, authentication, data integrity, audit logs, automatic log off, emergency access, access control, and account of HIPPA release of information [84]. Physicians doubt the level of security of patients' information and records. According to Simon et al. [74], physicians are more concerned about this issue than patients. The inappropriate disclosure of information might lead to legal consequences. Testing the security of EHR products, a group of researchers showed that they were able to exploit a range of common code-level and design-level vulnerabilities of a proprietary and an open source EHR [85]. These common vulnerabilities could not be detected by 2011 security certification test scripts used by CCHIT. EHRs pose new challenges and threats to the privacy and security of patient data. This is a considerable barrier to EHRs proliferation. However, this risk can be mitigated by proper technology, and maintaining certified standards with the software and hardware components.

- **Legal aspects:** Electronic records of medical information should be treated as private and confidential. Various legal and ethical questions obstruct adoption and use of EHRs. The legal system that relies on the paper-era regulations does not offer proper guidance regarding the transition to EHRs. EHRs may increase the physicians' legal responsibility and accountability [86]. With computer-based sophisticated auditing, it is easy to track what individuals have done. The documentation is comprehensive and detailed in EHRs. It can both defend and expose physicians regarding malpractice. According to a *Health Affairs* article, malpractice costs around $55 billion in the United States, which is 2.4% of total healthcare spending [87]. A 2010 research reveals that it was unable to determine whether the use of EHR increases or decreases malpractice liability overall [86]. HIPAA's privacy standards also present reasonable barriers to EHR adaptation.

2.7 Challenges of Using EHR Data

The primary purpose of EHR data is to support healthcare-related functionalities. As a vast amount of data is being collected every day, the secondary use of EHR data is gaining increased attention in research community to discover new knowledge. The main areas of use are clinical and transitional research, public health, and quality measurement and improvement. Using the EHR data, we can conduct both patient-oriented and public health research. EHR data can be used for the early detection of epidemics and spread of diseases, environmental hazards, promotes healthy behaviors, and policy development. The integration of genetic data with EHRs can open even wider horizons. But the data does not automatically provide us the knowledge. The quality and accuracy of the data is an issue to be taken care of. Beyley et al. [88] presents an excellent survey of the challenges posed by the data quality.

- **Incompleteness:** Data incompleteness or missingness is a widespread problem while using EHR data for secondary purpose [88, 89, 90]. Missing data can limit the outcomes to be studied, the number of explanatory factors to be considered, and even the size of population included [88]. Incompleteness can occur due to a lack of collection or lack of documentation [91]. Hersh [92] reports the following reasons for inaccurate reporting by professionals.
  - Unaware of legal requirements
  - Lack of knowledge of which diseases are reportable
  - Do not understand how to report
  - Assumption that someone else will report
  - Intentional failure for privacy reasons

  A pancreatic malignancies study using ICD-9-CM code at the Columbia University Medical Center found that 48% of the patients had corresponding diagnoses or disease documentation missing in their pathology reports [93]. Authors also report a significant amount of key variables missing (see Table 2.1).

  Patients' irregularity of communicating with the health system can also produce incompleteness. Based on the application in hand, type of data and proportion of data that is missing, certain strategies can be followed to reduce the missingness of data [91].

TABLE 2.1: Percentage of Incompleteness of Variables in a Pancreatic Malignancies Study

Variables	Endocrine
Necrosis	20%
Number of Mitoses	21%
Lymph Node Metastasis	28%
Perineural/Lymphovascula Invasion	15%
Differentiation	38%
Size	6%
Chronic Pancreatitis	14%
Smoking—Alcohol	27%–29%
History of Other Cancer	35%
Family History of Cancer	39%
Tumor Markers	46%

Source: Taken from Botsis et al. [93].

- **Erroneous Data:** EHR data can be erroneous as well. Data is collected from different service areas, conditions, and geographic locations. Data is collected by busy practitioners and staff. Therefore, the data can be erroneous due to human errors. Faulty equipment can also produce erroneous data. Validation techniques should be used to both identify and correct erroneous data. Both internal and external validation measures can be applied. Internal validation is a way to check the believability of the data, e.g., unrealistic blood pressure, BMI values, etc. Dates can be used to check whether the result generated before a test has taken place. External validation includes comparing the data with other patients or historical values.

- **Uninterpretable Data:** The captured EHR data might be uninterpretable to a certain extent. It is closely related with data incompleteness. It may occur when some part of the data is captured but the rest is missing. For example, if a specific quantitative or qualitative measurement unit is not provided with the result value, it will be difficult to interpret.

- **Inconsistency:** Data inconsistency can heavily affect the analysis or result. Data collection technologies, coding rules, and standards may change over time and across institutions, which may contribute to inconsistency. For multi-institutional studies this issue might be common, especially because different healthcare centers use different vendors for providing apparatus, softwares, and other technologies [88]. A study in Massachusetts of 3.7 million patients found that 31% of patients have visited two or more hospitals in the course of five years [94].

- **Unstructured Text:** In spite of having many defined structures for collecting the data, a large portion of the EHR data contain unstructured text. These data are present in the form of documentation and explanation. It is easy to understand them for humans, but in terms of automatic computational methods, detecting the right information is difficult. Sophisticated data extraction techniques like Natural Language Processing (NLP) are being used to identify information from text notes [95].

- **Selection Bias:** In any hospital, the patient group will mostly be a random collection. It varies depending on the nature of practice, care unit, and the geographical location of the institution. It will not contain the diversity of demography. This is an important challenge to overcome. Therefore, EHR data mining findings will not be generalizable. This problem must be addressed while working with the secondary use of data.

- **Interoperability:** Lack of EHR interoperability is a major impediment towards improved healthcare, innovation, and lowering costs. There are various reasons behind it. EHR software from commercial vendors are proprietary and closed systems. Most software were not built to support communication with a third party and developing new interfaces for that purpose might be a costly undertaking. Absence of standard also contributes to the problem. Many patients are not lenient towards sharing their information. Besides EHR systems must comply with the HIPAA Act [11] to ensure the security and privacy of the data.

  In a recent *JAMIA* (*Journal of the American Medical Informatics Association*) article, the authors have specified 11 specific areas that present barriers to interoperability of C-CDA documents by inspecting 91 C-CDA documents from 21 technologies [96]. In June 2014, the office of the National Coordinator for Health Information Technology (ONC) unveiled a plan for robust healthcare information sharing and aggregation and interoperability increase by 2024 [97]. Its three-year agenda includes "Send, Receive, Find, and Use Health Information to Improve Health Care Quality." Its six-year agenda states "Use Information to Improve Health Care Quality and Lower Cost," and finally, its 10-year agenda proposes to achieve a "Learning Health System." The mentioned building blocks for attaining the goals are the following:

  - Core technical standards and functions
  - Certification to support adoption and optimization of health IT products and services

- Privacy and security protections for health information
- Supportive business, clinical, cultural, and regulatory environments
- Rules of engagement and governance

2.8 Phenotyping Algorithms

Phenotyping algorithms are combinations of multiple types of data and their logical relations to accurately identify cases (disease samples) and controls (non-disease samples) from EHR as illustrated in Figure 2.3 [98]. Based on the structure, EHR data can be broadly divided into two parts, structured and unstructured data. Structured data exists in a name–value pair while unstructured data contains narrative and semi-narrative texts regarding descriptions, explanation, comments, etc. Structured data include billing data, lab values, vital signs, and medication information. Billing and diagnosis-related data are collected using various coding systems like ICD, CPT, and SNOMED-CT. These codes are important parts of the phenotyping process. ICD codes generally have high specificity but low sensitivity [99]. Table 2.2 lists different characteristics of EHR data.

The primary purpose of EHR data is to support healthcare and administrative services. Information is produced as a byproduct of routine clinical services. They are not a suitable format for performing research tasks. They often require further processing to be used for phenotyping algorithms. Within existing EHR systems, querying for a particular diagnosis or lab test across all patients can be a not-trivial task. An EHR can quickly pull the information related to a patient's current medications, and easily find any test results. But combining different data with a temporal relationship might require manual processing of data. From clinical operational settings, data are often extracted and reformatted to make them more convenient and suitable for doing research, typically storing them in relational databases. Researchers have created a number of Enterprise Data Warehouses (EDWs) for EHR data. Examples include Informatics for Integrating Biology and the Bedside (i2b2) [100], the Utah Population Database [101], Vanderbilt's Synthetic Derivative [102], etc. Commercial EHR vendors are also developing research repositories. For example, EPIC users can add the "Clarity" module to their system, which will convert the EHR data into SQL-based database for research purposes.

To build a phenotype algorithm, first we need to select the phenotype of interest, followed by the identification of key clinical elements that define the phenotype. It may contain billing codes, laboratory and test results, radiology reports, medication history, and NLP-extracted information. The gathered information may be combined with a machine learning method. For example, in [103], the authors have applied Support Vector Machine (SVM) to a both naive and well-defined collection of EHR features to identify rheumatoid arthritis cases. A medication record can be used to increase the accuracy of case and control identification of phenotyping algorithms. Patients who are believed to be controls must be having a different medication profile. They may not even have any medications prescribed to them at all. Sufficient dosage of a particular medication serves the confirmation that a person is having the disease of interest. For example, a patient treated with either oral or injectable hypoglycemic agents will be having diabetes. These medications are highly sensitive and specific for treating diabetes.

Studies have shown that CPT codes can accurately predict an occurrence of a given procedure [104]. The standard terminology codes for lab tests are LOINC. On the other hand, clinical notes are in free-text format. To be used for phenotyping algorithms, it has to undergo subsequent text processing. Certain procedures and test results may also exist in a combination of structured and unstructured form. For example, an electrocardiogram report typically contains structured interval