CHAPTER 2

Precision Medicine: Decoding the Biology of Health and Disease

James M. Snyder with case scenario by Joseph Tan

LEARNING OBJECTIVES

- Define Precision Medicine (PM) and showcase how PM transforms traditional healthcare thinking and services
- Articulate underlying concepts and principles of PM, especially in cancer care
- Highlight key discoveries and events leading to PM
- Detail key barriers and challenges in implementing PM

CHAPTER OUTLINE

Scenario: Origo—Crafting a Precision Medicine Platform for Cancer Patients on a Global Scale[1]

In 2017, Genotech Matrix, a New Haven, Connecticut, biotech company, partnered with Vishuo Biomedical, a Singapore healthcare technology company, to advance cancer care on a global scale. Together, these companies have developed an award-winning Precision Medicine Platform to resolve challenges for the need of personalized cancer care faced by a leading Manhattan hospital group that wanted to implement their own precision medicine initiative. This hospital group was not prepared to have their patient database management and cancer genetic sequencing process, including secured information sharing and management, outsourced when the viability of their *Origo* Clinical Cancer Genome (CCG) Platform solution was successfully demonstrated.

According to cancer researchers working in these biotechnology companies,[2(p. 1)] "advancements in genetic testing have allowed clinicians and researchers to better characterize types of tumor, specific mutations and then tailor treatment regimens to save time and money. Simultaneously, genetic sequencing costs have decreased significantly, while the speed of analysis and generating results has increased. With this alignment, academic and healthcare organizations have been migrating quickly into the arena of **precision medicine**… Cancer continues to bewilder even the best clinicians. While there are more than 100 types of cancer, we are learning that there are many more genetic mutations that cause one person's tumor to be unique and respond differently to treatments that may work for others with the same type of cancer."

As shown in **FIGURE 2-1**, the proposed solution involves implementing the Genotech Matrix *Precision Medicine Platform* in

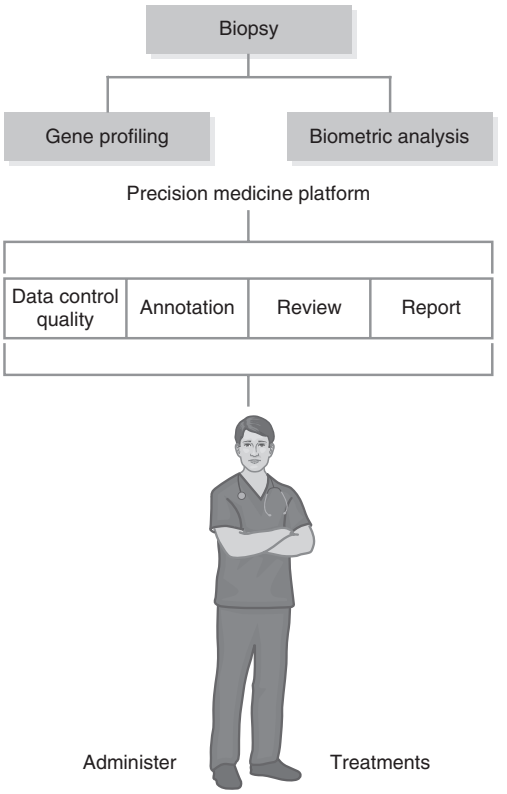

FIGURE 2-1 Precision medicine workflow pipeline.[2(p. 2)]
Data from genotechmatrix.com/wp-content/uploads/2017/05/Case-Study.pdf

combination with *iCMDB® (intelligence in Clinical Medicine for Decision-Making and Best Practices)*, a core product of Vishuo Biomedical. This combination advances the manipulation of a globally enriched knowledge base for automating sequencing and analysis of variants of gene expressions to support personalized treatment by clinicians based on the genetic and histologic pathology profiles of patients.

Aside from integrating *iCMDB* into the Manhattan hospital group's database, a personalized workflow pipeline that supports report generation to meet the specific needs of caring oncologists within the organization's setting has also been developed. For secured patient data management, an on-site server installed with the *Origo* platform has been deployed, and the analysis pipelines have been customized to protect and encrypt any confidential

personal information used in generating automated personalized reports that support clinical decision-making. In aiming to provide a more efficient and effective (personalized) approach to cancer care, it is purported that to date over 2000 oncology patient samples have been sequenced and their respective reports have been generated via the *Origo* platform since its 2016 implementation.

Watch the YouTube video about the power of the *Origo* platform.[3] Think about how this new Genotech–Vishuo collaboration will impact the future of global cancer care delivery in light of the rapidly increasing mobility of patients. Additionally, reflect upon why and how precision medicine may now be redefining health care and fulfilling the potential to minimize side effects from traditional cancer treatments as cancer care moves toward more personalized treatment. What other innovative biotechnological services might potentially be supported on the *Origo* platform, and how will such collaborative efforts usher in a new era of precision medicine?

I. Introduction

This chapter introduces the readers to an emerging healthcare model known as Precision Medicine (PM). PM is an approach to health care that is largely dependent on the *digitization of health data and bioinformatics*, a field of study intersecting key areas of computer science and biology. PM attempts to improve health outcomes by refining diagnosis, treatment, and disease prevention through understanding of the many factors that can contribute to the intrinsic biology of disease.

A core principle of PM is that by decoding the genetic and molecular changes that lead to the development of disease, we can alter the course of disease and preserve health. The technology used to analyze someone's genetic sequences and other molecular events is now readily available and may be clinically utilized. Molecular data provide perspective as to how a disease develops or how someone may respond to treatment. New technologies are creating a wealth of health-related data that will offer additional insight into behavioral and environmental influences on molecular biology and genetic changes that cause specific disease(s). Harnessing the power of computer science to create big data knowledge networks that connect the intrinsic biology of disease with other health-related factors, such as behavior, exposures, and environment, is central to PM.[4]

In this chapter, we review advances in the scientific understanding of the intrinsic biology of disease with an emphasis on genomics, introduce the utilization of large-scale molecular testing in health care, survey barriers to implementing PM, and discuss the future of this new approach.

"And that's the promise of precision medicine – delivering the right treatments, at the right time, every time to the right person. And for a small but growing number of patients, that future is already here." - President Barack Obama[5]

II. Background: What Is PM?

PM is an emergent healthcare perspective focused on the prevention, diagnosis, and treatment of disease based on an individual's unique health features, with an emphasis on the molecular underpinnings of health and disease.[6,7]

In the last several years, there has been a tremendous increase in available healthcare information, including genetic analysis, environment, and behavior. Molecular data, which include information about an individual's genes, gene activity, proteins, epigenome, and cellular activity, have entered everyday clinical care and disease management. There is great optimism that this influx of molecular and other health data into medical management (specifically, PM) will accelerate our understanding

of disease and dramatically improve treatment outcomes and disease prevention. This is a progression from Western medicine's current approach of guideline-modeled care where therapeutic regimens are intended to be applicable to large groups of people.

Importantly, advances in computer science and bioinformatics have ushered in PM through the ability to process large amounts of data from many data sources so as to identify new factors in the development, prevention, and treatment of disease. PM attempts to answer why some people with similar risk factors develop an illness and others do not and why a therapeutic strategy is curative in only select cohorts of people, and, ultimately, illustrate how an illness can be prevented from occurring in the first place. The PM community is confident that the answer to these questions is hidden in the subcellular molecular data that we are beginning to understand. Most cancers, for example, are thought to occur due to genetic instability. Three common explanations for the genetic instability include inherited mutations that we are born with, somatic mutations that occur in cells during development or throughout life, and deviations in the regulatory mechanisms that maintain genetic integrity.[8] Science has made great strides in understanding the genetic and genomic features of cancer and noncancerous conditions. Efforts are underway to create networks of knowledge that connect the molecular and genetic building blocks of disease with other health data at a population and individual level.

As an example, we will review a standard patient presentation and physician evaluation in our current model of care. A 52-year-old woman presents to her doctor with symptoms of weight loss and a progressive cough over the last 3 months. The doctor asks a litany of symptom-based questions, performs a clinical exam, and orders additional testing to help make a diagnosis, such as chest X-ray and blood work. The chest X-ray reveals a mass. The patient undergoes a biopsy of the mass, which shows a type of lung cancer called adenocarcinoma. The cancer is identified by histologic diagnosis (how the cells look under a microscope), and the patient is treated per the national guidelines for that type of cancer. In recent years, we have seen tremendous developments in our understanding of the molecular drivers of disease and have another layer of clinically relevant diagnostic and therapeutic information to add to this patient's diagnosis and treatment decision-making. Through molecular information, we are finding similarities between cancer types that were previously thought to be unrelated. In this example, the patient was identified to have a gene mutation called anaplastic lymphoma kinase (ALK), which is also seen in a type of brain tumor called neuroblastoma. As the network of knowledge matures, advances in treating ALK-mutant lung cancer may shed light on neuroblastoma brain tumors, two diseases that present very differently in the body. The lung cancer patient is started on an approved drug that targets this ALK pathway. Specific treatment recommendations based on molecular information is slowly being integrated into guideline-based care recommendations.

Considerable overlap exists between PM and other approaches to health care, including our current symptom-driven model. P4 medicine stands for *predictive, preventive, personalized, and participatory* health care.[9] P4 medicine approaches health care with a broad view spanning population health to subcellular science and molecular medicine. P4 medicine, developed by the Institute for Systems Biology in Seattle, Washington, attempts to bring together system-based biology with patient-provided data generation and advanced technology through the use of digital tools that aggregate multidimensional patient-health-experience data, which they call the "networks of networks" (see **FIGURE 2-2**).[10]

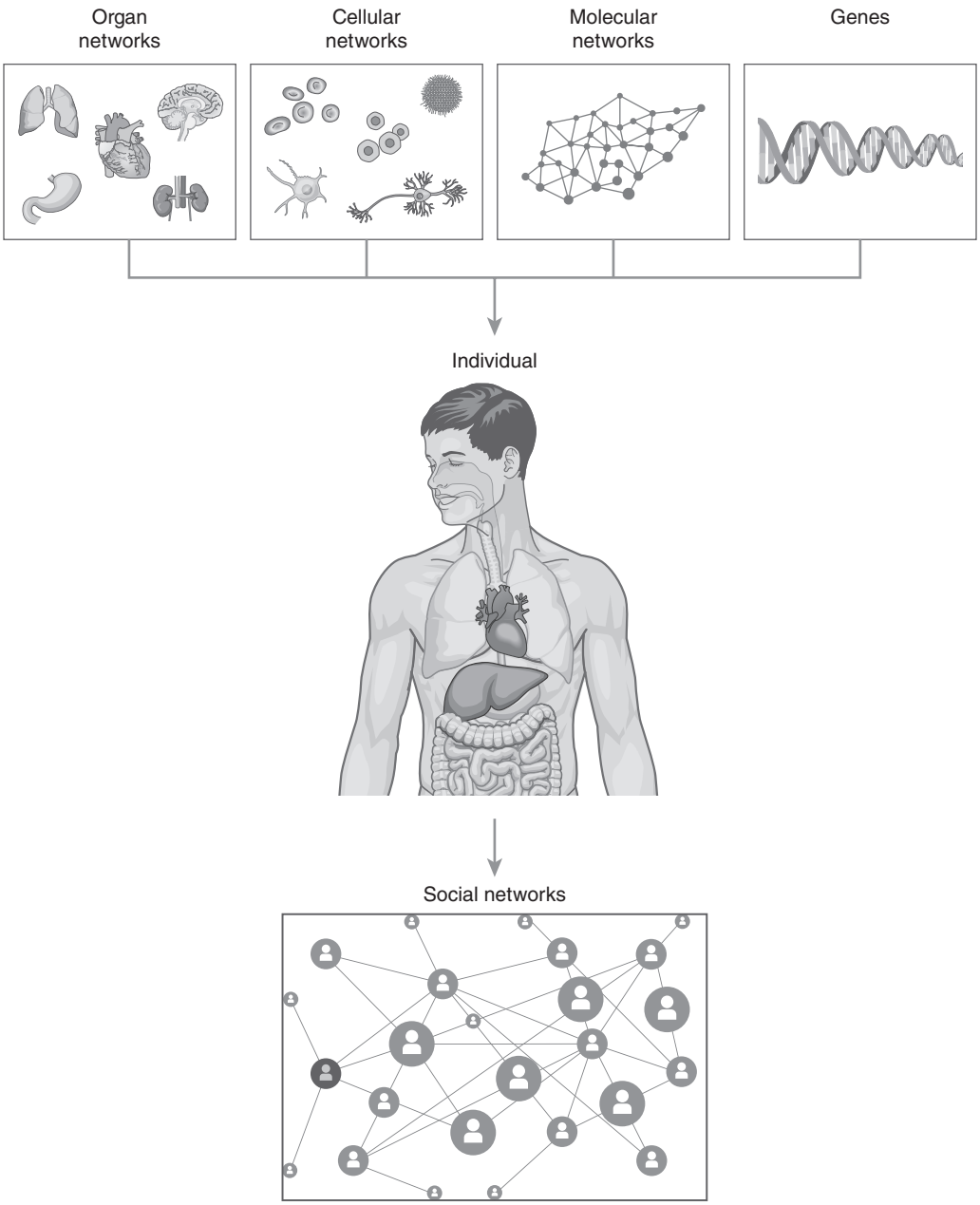

FIGURE 2-2 The "Network of Networks" aggregated health data foundation of the P4 medicine approach.

From http://future.psjhealth.org/scientific-wellness/about-the-institute-for-systems-biology?utm_source=TWITTER&utm_medium=social_organic&utm_term=--&utm_content=psjh-1773051757-&utm_campaign=evergreenContent+Type+%28Secondary%29?

III. Key Events in the History of PM

The opportunity to implement PM has occurred due to simultaneous milestones in our understanding and access to molecular data, advances in computer science that can curate and analyze large multivariate datasets, and routine use of molecular testing, which has dramatically reduced cost while increasing efficiency. This has ignited an interdisciplinary effort with cross-training in many disciplines, the most notable being bioinformatics, which is the combination of biology and computer science. In the "omics" research lab, I observe and participate in many biologists and computer scientists work side by side with overlapping educational paths and research skills. "Omics" refers to scientific disciplines that end in the suffix "-ome," which implies large-scale study of the subject field, such as gene(omics).[11] Leading omics disciplines include genomics, proteomics, transcriptomics, epigenomics, metabolomics, radiomics, and others.

Survey of Human Genetics in Health Care

Early understanding of genetics is attributed to Gregor Mendel's work in the 1860s describing the inheritance of traits in pea plants. In the 1950s, the structure of deoxyribonucleic acid (DNA) was identified. DNA is the hereditary chemical code made up of adenine, guanine, cytosine, and thymine base pairs that create a structure called a double helix. The sequence of DNA base pair combinations directs how cells are formed and maintained. A practical method of DNA sequencing was published in 1977, ushering in an era of genomic medicine.[12]

Conceptualized in 1990 and completed in 2002, the Human Genome Project courageously mapped the complete human genome, providing the world with "the structure, organization and function of the complete set of human genes."[13] In 2006, the National Cancer Institute (NCI) launched The Cancer Genome Atlas (TCGA) characterizing the genetic and molecular features of 33 tumor types. Massively Parallel Sequencing (MPS), which is also called Next Generation Sequencing, performs many sequencing tests at once. MPS was introduced in the academic literature in 2008 and dramatically changed the landscape of genetic testing by reducing costs 100-fold and reducing the time to completion of sequencing to just 8 weeks (it can now be done in matter of few days).[14]

Prior to MPS, each exon, which is a segment of DNA that codes for a corresponding segment of ribonucleic acid (RNA) and, ultimately, a protein, had to be sequenced and amplified individually, requiring considerable time and resources. The 2011 report by the National Research Council (NRC) laid out the framework for a molecular taxonomy of disease and implementation of PM, and in the United States in 2015 President Barack Obama committed $215 million to funding a national PM effort. Only recently has genetic sequencing entered routine clinical care, as previously this testing was cost prohibitive.

In 2001, the cost to sequence one entire human genome was estimated at $95 million.[15] Sequencing costs continued to drop, and in 2007 the estimated cost for the sequencing of a single human genome was approximately $10 million.[16] In 2011, the cost was $21,000, and in 2018 whole genome sequencing was obtained for less than $3000.[17] This reduction in cost coupled with the availability of results in less than 2 weeks has brought molecular medicine into clinical practice for a growing list of health conditions.

Now that molecular data are approaching a cost-effective and actionable timeline, efforts are underway to include these data into routine clinical practice. The point of obtaining genetic data is to identify the cellular blueprints

of health and disease. In understanding the building blocks of disease, we can better categorize conditions and hopefully reveal why some patients respond to treatment and others do not. The goal is to prevent disease before it occurs, but to do so we must also understand how and why deviations from health develop. Cancer is particularly ripe for a PM approach as most cancers are thought to occur due to a complex relationship between genetic instability and environment at the cellular, individual, and population level.

A brief discussion of genetics and molecular anatomy is helpful to understand the molecular aspects of PM. DNA is the building block of proteins in our body. DNA is the cellular template that undergoes a process called transcription to create precise sequences of ribonucleic acid (RNA). During translation, this code of RNA is the blueprint used to build defined amino acids with which proteins are formed. The epigenome refers to the chemical compounds and proteins that package and control access to the genetic code and cellular functions, controlling how the genetic code is implemented.[18] As one can imagine, environmental factors such as smoking, age, or disease can impact this process. There are many additional factors that contribute to the development of disease and disruption of "normal" molecular pathways. Phenotype refers to the observable characteristics or expression from the genetic code,[19] which can be thought of as the manifestation of the genetic code. The relationship between the genetic code and an individual's phenotype is also complex, with each interconnected layer of biologic information harboring tremendous potential insight into health and disease.

"Variant calling" is the identification of molecular deviations from the expected genetic code, which are also called mutations. There is no consensus definition of the "normal" genetic code in humans, as existing data are based on small sample sizes that may not reflect the general population. Prospective efforts are underway to characterize the genomic and health data of large groups of people, numbering from 500,000 to more than 1 million people. Many existing molecularly characterized datasets are based on retrospective data, or data obtained in isolation that may not include other relevant attributes such as someone's activities, quality of life, geographic location, comorbidities, environmental exposures, or family history. Large prospective studies that track participants longitudinally over a number of years with cross platform data collection, including Electronic Health Records (EHR), and personal health data like the All of Us project and others have the potential to illuminate the complex factors that contribute to disease development and, ultimately, prevention.[20]

Research to characterize the nearly 3 billion units of DNA across 23,000 DNA base pairs has made tremendous progress over the last several years. Similar research in other omics disciplines has also shown progress. The magnitude, specificity, and types of testing available in health care are evolving rapidly as is our insight into the relationships between this data and health. Some diseases may reveal direct variants in DNA or RNA that can be successfully targeted with PM therapies; however, it is more likely that there is a complex relationship between many factors, including environment, molecular events, and other attributes that contribute to the development of disease.

In 2011, the NRC laid out the framework for a molecular taxonomy of disease summarizing the state of molecular medicine and presenting an action plan to implement PM. Implementing PM requires a profound change to clinical practice, including how data are recorded, what data are recorded, how clinicians analyze and process vast quantities of data, how patients participate in healthcare data, the medical decision-making process, access to molecular testing, access to targeted therapies, approval and safety process for drug

development, and many other aspects of our current care model, in addition to new technologies that will develop.

IV. Current Perspective

Over the last few years, we have seen the availability and commercialization of high-throughput sequencing flood oncology clinics with real-time genetic data, including hundreds (and soon thousands) of anticipated genetic variants that may have clinical significance, as well as new biomarkers that can be used to measure health or disease. These genetic data are typically reported within 2 weeks of sending out the bio-specimen and are therefore clinically actionable. Molecular testing is usually requested to subclassify a diagnosis, refine therapeutic options, or fulfill claims of PM. Oncology centers are racing to utilize these data through molecular tumor boards (MTBs) and third-party data navigation platforms provided by academic institutions, non-profit organizations, and for-profit companies.

In addition, new realms of data such as digital phenotyping, wearable devices, Internet of Things (IoT) medical devices, smart phones, patient self-reporting, social media inputs, and beyond are being generated for the healthcare sector at a rapid pace, leaving clinical teams scrambling to digest all of these data. Curating medical data for security and meaningful use is necessary to maximize this opportunity and ensure the safety of potentially vulnerable patient health information (PHI). At the time of this publication, there is no publicly available standard tool to view these data in concert and leverage their collective value. To accomplish this task, advancements in healthcare genetics education, health data curation, and technology implementation in clinical practice will have to occur.

This volume of genetic and molecular testing available in clinical care and the commercialization of this data production process through large third-party providers that aggregate test results into large privately held or public datasets is a paradigm shift in medicine. Our current use of molecular data is built upon the groundbreaking work of expansive genome atlases, such as TCGA and others. Whereas participation in foundational genomics studies like the TCGA was primarily performed at large academic medical institutions, the utilization of MPS through commercial platforms and regional labs is available through any provider with access to a tissue sample. This testing, which typically investigates a panel of known genetic variants, is now commonly performed in academic and community centers alike. This is a critical shift—placing volumes of genetic medical data into public registries and private companies. Both private and academic bodies offer genetic testing using MPS methods, retaining data in functional databases that hold great intellectual and financial value. Increasingly, these groups are partnering with private and for-profit entities to harvest the data for research and discovery such as drug development or academic consortia. Prior collaborative academic efforts (like TCGA) have generated this data for public utility, fueling countless research efforts, and are available in searchable formats through web interfaces such as the NCI genomic data commons portal, the open source Clinical Interpretation of Variants in Cancer (CIViC) dataset, Cbioportal, and others.[21–23] Of note, most federally funded research efforts are required to provide the molecular data in a publicly accessible format after publication. Owing to the escalating financial and intellectual value of healthcare data, new paradigms of data protection and monetization have evolved.

Molecular data are increasingly integrated with routine cancer care. Many cancer types use some molecular or genetic data for diagnosis or to define disease subtypes. Multiple cancers have guidelines that use molecular data

for treatment decisions that are oftentimes also supported by dramatically improved patient outcomes, such as those seen in melanoma and lung cancer. Many major cancer centers are in the early phases or have recently created dedicated molecular tumor boards. A few cancer centers have had dedicated MTBs for several years. A MTB or PM tumor board typically refers to a multidisciplinary team of healthcare professionals who prospectively review a patient's molecular testing in the context of their disease and treatment plan. An MTB often discusses treatment options when: (1) a possible drug target is present; (2) associated conditions require further testing (i.e., concern for a germline mutation); and (3) the molecular variant may impact health outcomes. Most MTBs are restricted to cases with actionable genetic variants but are not restrictive to any disease or histologic type. At larger tertiary centers, specialized tumor boards, which are also called prospective multidisciplinary cancer meetings, are organized by disease site such as a lung cancer tumor board or a nervous system tumor board. Tumor boards at dedicated cancer centers typically consist of a medical oncologist, disease-specific surgeon, radiation oncologist, radiologist, pathologist, nurses, and other healthcare providers, such as genetic counselors, social workers, and clinical trial experts.[24]

There are several barriers to implementing a PM recommendation, such as identification of a targeted therapy that disrupts a critical tumor growth pathway, access to the desired drug or therapy, and safety of administration. A PM-targeted therapy must disrupt the identified molecular pathway, and the tumor must be dependent on the specific pathway.[25] Some variants that are identified may not be the primary driver of the disease and are less likely to have a clinical impact if targeted. The goal is to identify molecular variants that are thought of as driver mutations or master regulators that may have great impact on disease development. The process to bring a new drug into clinical practice requires tremendous regulatory oversight that should be followed in the interests of patient safety and scientific advancement. Investigational therapies should be administered through a clinical trial with extensive safety monitoring.

Clinical trials are historically organized by disease, organ involved, and histology (the cellular features seen under a microscope). Only recently have molecularly driven clinical trials become available. Scientists have postulated for many years that the specific molecular features of a disease likely contribute to therapeutic response but have lacked tools of scale to prospectively investigate and quantify these features. PM and the addition of molecular taxonomy to histologic diagnosis have changed the way diseases are categorized and also impacted the way clinical trials are designed. Clinical trials are historically described in phases, with different questions being asked at each phase. Phase 1 trials primarily research safety and the tolerated dosing of a new therapy. Phase 2 and 3 trials investigate if the intervention has an impact on the disease as well as associated adverse events.[26]

Two newer types of clinical trials designed for PM are "basket" trials (treatment cohorts are based on a shared mechanism of action across multiple histologic tumor types) and "umbrella" trials (which may include multiple molecular pathways and corresponding investigational drugs in the same trial designed for one histologic cancer type).[27] Several other innovative clinical trial designs are being explored.

In addition to refining histologic diagnosis with molecular subclassification, we must also decipher the variation of molecular and genetic expression both spatially within a tumor and as time progresses. This variation likely contributes to treatment resistance. Some healthcare clinics are attempting to obtain longitudinal genetic testing at multiple

points in a person's disease course or samples from multiple locations within a tumor. As the cost of testing decreases and the insights provided by the test increase, we will undoubtedly see the utilization of molecular testing for cancer throughout a disease course as a monitoring tool. As the knowledge networks mature, we anticipate increased use of PM-driven informatics in disease prevention programs. If a person is identified to have known risk factors for disease, they may undergo periodic molecular screening tests to measure their risk of developing the disease and to hopefully reduce known risk factors. To accomplish this, the medical community must understand the many factors that contribute to disease development, identify a way to measure these factors, and then implement an intervention process.

Several technology platforms exist with the purpose of curating molecular and healthcare data. Major U.S. hospitals often use some form of EHR to curate healthcare data. In the PM space, there are specific technologies designed to organize and help clinical teams interpret genetic and other molecular data, facilitate PM MTBs, aggregate clinical data, and navigate clinical trial opportunities. Some of these platforms are open source, while others are proprietary. Through multisite PM applications, large databases are created that hold immense monetary and intellectual value. The opportunity to impact health care through big data analysis with machine learning and other analytic approaches in medicine is currently underutilized. The bulk of existing healthcare systems have not capitalized on big data; however, many are showing greater interest, which may be in response to the avidity with which private companies are trying to accomplish this task. To maximize this opportunity, healthcare data will need to evolve and solve concerns over semantic heterogeneity, technical heterogeneity, patient data security, financial limitations, and the resulting impact on clinician workflows.

Genetic Data

In the clinical practice of oncology, physicians use genetic panels to look for known cancer variants to fuel PM. Many of these panels test for changes at specific areas of DNA but may also include other molecular investigations, such as RNA or whole genome sequencing. Clinically available sequencing tests look at DNA for base pair substitutions, deletions, insertions, and fusions that have been shown to be relevant in cancer.[28] In some cases, investigators will look for only a few specific pertinent mutations or genetic aberrations that are important for a type of disease. For example, in neuro-oncology, which is the field of medicine focused on cancer and the nervous system, the molecular features of brain tumors have only been included in the World Health Organization's pathologic diagnosis recommendations since 2016.[29]

Prior to 2016, a brain tumor diagnosis was made solely on histologic review despite the availability, clinical utility, and known importance of molecular data in the classification of brain tumors. When someone is diagnosed with a brain tumor, it is standard practice for the pathologist to report select mutations such as IDH1 mutation, which conveys information on tumor development and prognosis; MGMT promoter hypermethylation, which provides insight into treatment response with certain types of chemotherapy; and genetic deletions on the 1p arm and the 19q arm, which are a diagnostic requirement for a tumor type called oligodendroglioma.[30] This is an example of PM in current practice. The clinical team may elect to investigate these tests as part of a broader panel or they may test individually. There are many ways to perform these tests; however, some centers may not have the needed equipment or expertise and elect to send the tissue to a qualified testing center.

In this chapter, we have focused on known variants of DNA used in the clinical management of someone diagnosed with

cancer, although there is a wealth of additional molecular testing available. Many other pertinent investigations into molecular data are evolving at a rapid pace. RNA, protein expression, genetic fragments found in blood, whole genome sequencing, tests that investigate the accessibility of DNA, and the characterization of the tumor microenvironment are other areas of research. Nonetheless, a detailed discussion is beyond the scope of this chapter. Some molecular tests are only used in preclinical research, while others have met requirements that permit use in clinical care. Historically, research and clinical testing have been performed and analyzed separately, but with the development of PM and the reduced cost of testing, we hope to see hybrid research and clinical molecular investigations.

In summary, genetics is the study of how specific traits are inherited. This differs from genomics, which is the study of large-scale genetic data, such as the entirety of the human genome. "Omics" refers to scientific disciplines in biology that end in the suffix "-ome," which implies large-scale study of the subject field, such as gene(omics). Genomics is generally considered the first of the omics disciplines—but now there are many. The omics disciplines were ignited by advancements in computer processing that now allow scientist to analyze large quantities of biologic data. A driving message of PM, and the root of the omics disciplines, is that through processing large multivariate datasets, we can unlock the keys of health and disease. The emphasis on connecting data from many different sources by removing barriers across research disciplines and data silos is a tremendous undertaking and a rate-limiting factor for PM, as well as health informatics.

V. Future Trends

In 2013, the NRC laid out the framework for a molecular taxonomy of disease summarizing the state of molecular medicine and presenting an action plan to implement PM.[31] Most of the issues, objectives, and solutions outlined in this landmark publication are still relevant and can be applied toward curating and accessing the data and implementation of findings. As discussed, PM has existed in health care for many years and has made significant progress toward a mechanism-based classification of disease and treatment decision-making. Large-scale efforts to understand human genomics across populations (a) in the pre-disease state, (b) as disease develops, and (c) during treatment, are now underway. In cancer and other disease states, molecular-derived classification and treatment protocols are becoming routine clinical practice; still, much work is needed to fully support a paradigm shift toward PM.

Curating the Data

Only recently has western medicine recorded healthcare data in digital formats through adoption of an EHR. While EHRs are relatively similar in concept to Electronic Medical Records (EMRs), an EHR is meant to encompass more data and extend beyond the health system or an individual doctor's office. In the United States, a handful of large commercial EHR services dominate the market. It is possible that separate institutions or hospitals that use the same EHR software can link medical records for an individual patient. By reducing institutional barriers and promotion of data aggregation, PM efforts may also be strengthened; however, this may require aggregating the data to some level of uniform reporting and analysis.

The power in PM stems from connecting many data types into a larger knowledge network so that subgroups and patterns can be identified. For an individual healthcare dataset to contribute to a larger knowledge network, the descriptors and attributes that describe the same value must use the same language. In medicine, there are many ways to say the same thing. A common data model should be

implemented across healthcare sectors to permit data aggregation and sharing. A common data model uses a defined dictionary of terms. If care teams are not recording data using a common language, then a solution is required to convert the recorded information into the common data model while maintaining the integrity or value of the data. For example, a devastating brain tumor that affects people of all ages may be referred to as a glioblastoma, glioblastoma multiforme, astrocytoma grade IV, or GBM; despite these four names, the clinical diagnosis is the same. If the cohort is reduced by "fragmented naming," then the power of the sample size may not be sufficient to reveal disease subtypes and associations needed to power PM.

With developing technology comes renewed questions of ethics and barriers to implementation. PM is limited by a litany of regulatory hurdles that are designed to protect patient safety and security. Smartphone navigation platforms and wearable devices provide a wealth of available environmental and behavioral data that until recently was too complicated to record and aggregate. Collecting healthcare data while maintaining privacy and adherence to strict regulatory policy as required by the Health Insurance Portability and Accountability Act (HIPAA) remains a challenge. Digital phenotyping, as defined by Dr. John Touros et al., is the use of digital devices to provide health data through "moment-by-moment quantification of the individual-level human phenotype in-situ".[32] Digital phenotyping holds tremendous potential to identify modifiable disease risk factors and environmental association with molecular data and health outcomes. Best practice in aligning these data sources while respecting privacy is unclear. One solution is that patient groups opt in and provide their own digital phenotyping data and connect this with their health history or molecular testing. But can this solution provide the volume required to power analysis and what bias does this introduce? How can science be representative when segments of the population do not have access to molecular data or digital phenotyping devices? Implementing PM requires thoughtful review of regulatory and ethical concerns as each new technology enters the knowledge network and clinical arena.

PM is dependent on the merger of clinical and research data from across the spectrum of health and disease. Clinical medicine in cancer requires multidisciplinary collaborative teams of clinicians to adequately care for patients with complex disease. Cancer-based PM research may benefit from a similar approach of multidisciplinary teams to connect disparate data sources and fuel collaborative research. Open data networks can compromise intellectual property housed in the data and devalue individual or institutional contributions, which are critical aspects of research funding. A shift in how research is being organized and approached is needed. Only laboratory tests that meet strict requirements can be used in patient care. There is clearly opportunity to learn from preclinical work and research-level investigations. New policies of preclinical research investigation of molecular mechanisms and targeted therapies that support clinically approved molecular testing and PM treatment access are needed. PM and the enormity of new health data sources pose unique challenges to health informatics, requiring collaboration and data fluidity across traditionally isolated clinical and research efforts. Research and clinical care should be a connected, closed-loop system in the interest of delivering on PM for improved health outcomes.

The data commons needed for promoting PM will have data inputs from many different sources. A knowledge network will require data inputs from technically heterogeneous sources into a shared data commons. The days of a medical record coming solely from the doctor's scribbles and notes are gone. A patient may have a histology report from proprietary software, DNA methylation analysis from

another software type, nutrition data from a phone app, and environmental data from a smartwatch that must all connect in a central repository. The opportunity for new streams of healthcare data is endless. A PM solution will need to address how to connect technical heterogeneous information so that the data commons can access and support data from many sources.

Informatics and computational science play a huge role in processing this vast quantity of information so that it can be digested and harvested for scientific discovery and improvement in human health. With this explosion in innovation and opportunity comes a never-ending stream of questions. How are the individual patient's rights protected in this age of mass data collection? The patient always "owns" their health record, but when this data is monetized who is the beneficiary? How can healthcare systems pay for the considerable resources needed to execute PM? How do we safeguard this data against irresponsible use and prevent harm to those who agree to share their personal health data?

Access to the Data

In a PM-optimized healthcare environment, all data would be uploaded into a large, publicly accessible, international, pan socioeconomic, anonymized knowledge network that shares the common data model with uniform definitions and testing assays connected to a wealth of multi-dimensional omics data, powered to decode the molecular mechanisms of disease and therapeutic discovery. In the United States, private companies, hospitals, government organizations, consortiums, and other healthcare groups are racing to develop large interdisciplinary datasets to power PM discovery. Such datasets are extremely valuable and require considerable resources to manage.

In oncology, a basic knowledge network includes patient demographics, genetic or molecular test results, chemotherapy history, and imaging, with the opportunity for so much more. PM informatics platforms require access to health data, analytics to process the data, regular updating to reflect advancing knowledge, and at least one user interface to facilitate use of the data by clinicians and researchers. There are public and private PM platforms that can provide a user or health system with the tools needed to participate in PM. It is possible that individual repositories fragment the data to the point that discoveries in rare diseases or infrequent health attributes no longer have the sample size to be found. On the other hand, curation of an information commons requires considerable resources that could potentially be funded through monetizing the healthcare data that many companies are racing to obtain. A mechanism is required to support this infrastructure for a large cross-sectional PM network and provide the resources needed to achieve data aggregation across health systems and populations. These datasets harbor valuable intellectual property, which may need to be protected so researchers can invest in new discoveries that enrich the information commons. The bottom line is that these data belong to the patients, something institutions and companies often lose sight of.

Rare diseases and molecular aberrations may require extremely large datasets to achieve the volume required to draw conclusions. Others have identified this concern and started independent repositories of information, either for rare disease types or for rare mutations. One such effort is the NCI-backed Rare Diseases Registry (RaDaR) Program that connects investigators of rare diseases to a common data management center utilizing shared data practices and resources harnessed from public–private partnerships to collectively progress our understanding of rare diseases.[33]

Although there are strict rules and regulations safeguarding patient data, new solutions are needed to protect patient privacy in the age of PM and digital phenotyping.

As our healthcare data evolve, so must the conversation regarding data privacy and security. Individual health data now include second-by-second data points from smart devices that track behaviors, actions, and locations—information that could help determine modifiable risk factors. This level of monitoring, however, comes with additional ethical and data management concerns. Increasingly, genetic testing and healthcare data are bypassing the clinical team and being delivered directly to the patient. Two examples include My Family Health Portrait by the Surgeon General, where individuals can input family history to learn about disease risk factors, and home genetic testing kits that provide a window into genetic risk factors that predispose people to different diseases.[34,35]

Largely due to opportunities to improve care through advancing technology, physicians and health systems have loosened their grip on healthcare data, increasingly using third-party services, such as genetic testing that harbor large aggregate patient-derived datasets and patient-centered health registries. Applications exist that store useable personal EMR data on cell phones and provide ways to connect with your own EHR. Patients with rare diseases are coming together through social media and creating dedicated data repositories, tissue banks of pathologic specimens, and clinical trials for their rare conditions.[36] Patient advocacy groups are creating apps to chronicle patient-reported symptoms and outcomes. In the United States, the Patient Centered Outcome Research Institute (PCORI) is driving structured inclusion of the patient perspective into healthcare delivery models and research implementation to enhance value and improve patient trust.[37] Investment by patients in personal health data when applied to molecular and clinical information may lead to new discoveries of environmental or behavioral influence on health and individualized quality of life metrics. Patient-led participation in healthcare research beyond traditional institutional or geographic boundaries has also led to increased enrollment in clinical trials, as well as other advances that are yet unseen.[38]

Implementation of PM Findings

The impetus for PM is anchored in the hope of drastic improvements in health outcomes for all people through continuing advances in biology and computer science. The addition of a molecular taxonomy and subclassification of disease is the first stage of PM implementation; it is already well underway. PM has helped to identify subsets within a histologic diagnosis that harbor distinct subcellular molecular disease development pathways that result in an altered response to therapy and outcomes when compared with the general disease cohort. These molecular subsets help explain response variability among people who carry the same diagnosis and eventually will decode why some people respond to a drug and others do not. For some conditions, targetable genetic variants have been identified that respond to pathway-disrupting therapies, whereas other subclassifications reveal less malignant conditions that may not require as aggressive therapies. These discoveries have ushered in new perspectives into clinical care and research. To fully implement PM, considerable work is required to change how health information is recorded, collected, aggregated, and analyzed. Incorporating molecular data into the current diagnostic process and treatment algorithms has proved to be a challenging albeit solvable problem. Connecting data in a useable way across health systems or from sources that are not routinely integrated into clinical care (diet and activity) will be an important milestone toward delivering on PM.

Many contemporary clinical trials are designed to disrupt the molecular and genetic events that drive disease development. For the last few decades, clinical trials have used

agents that target critical molecular pathways, but the investigators may not have had the data or computational power to prospectively stratify treatment groups based on a molecular feature or genetic biomarker. Examples of contemporary clinical trials that assign treatment cohorts based on shared biomarkers or key genetic variants as a key to trial design include basket and umbrella trials. Completed biomarker-driven clinical trials have shown feasibility and promise with this scientific approach.[39,40]

A developing treatment design referred to as N-of-1 trials asks clinical trial questions of efficacy; side effect profiles are at an individual level, using a person's own genetic and health data.[41] N-of-1 trials in PM oncology require considerable data resources and have many design concerns; however, the meta-analysis from individualized studies, if done in a controlled and reproducible manner, may reveal generalizable data and identify new disease or treatment subsets. Another nuanced trial design often referred to as "personalized medicine" is when an individual's genetic profile or other features are used to determine the optimal dose of a medication or predict an individual response to an intervention. Of note, there is ambiguity in the terms "precision" and "personalized" medicine in the medical community. Increasingly, health care is seeing the use of commercially provided screening panels as a navigational tool for participation in a clinical trial or as a deciding factor in assigning a specific therapy in a multi-arm trial. This is largely because such panels are common in clinical care and provide verified central testing sites and assays, which are important to maintain research quality. The molecular investigations of interest that define disease subgroups or treatment regimens are also evolving. Science is identifying ways that global changes to DNA, modification of tumor suppressor signals, cellular access to the genetic code, and other molecular events impact cancer and clinical outcomes in addition to the identified genetic variants that are commonly investigated. Next-generation clinical trials and other means of using PM to deliver treatments and impact patient care will need to adapt and capitalize on evolving technology and scientific discovery.

Developing a high-quality integrated multivariate knowledge network was the foundation of the U.S. NRC's landmark 2011 publication that outlined a vision for PM.[42] Delivering on PM requires a large comprehensive knowledge network capable of fueling big data analysis with sufficient volume and quality to tease out molecular subgroups and associations with other health features. National, academic, and commercial PM efforts are underway, with nearly all participants utilizing a cross-sectional data repository connected to molecular and genetic testing. This knowledge network is anchored to the intrinsic biology of disease but must also continue to evolve and add new pertinent data sources that may shed light on modifiable factors in disease development.[43]

The success of PM will depend on the quality and volume of data that is aggregated in the knowledge network. Transitioning health data into a searchable structured format is a critical step that many health systems are finding difficult. To maximize adoption, PM platforms should be designed to augment existing healthcare operations and understanding of disease. PM can only change the healthcare landscape if it is adopted into clinical care and research. The transition to indexed healthcare data where clinical and other data sources are easily aggregated in a uniform and structured way may be the Achilles heel of implementing PM.

PM principles are already integrated into the clinical care of cancer patients. Many cancer types have established molecular subclassifications that are used to subtype a diagnosis or refine disease-specific treatment options. In some settings when a patient has failed standard therapies or if no standard therapy exists, a provider may pursue a molecularly targeted

PM approach. In this scenario, the patient's cancer specimen is investigated with a cancer genetic variant panel, or test of specific actionable variants, in hopes of identifying a growth-dependent tumor-driving mutation that can be targeted by an approved drug that is typically used in other conditions or cancer types. Ideally, this patient would qualify for, and have access to, a clinical trial that targets the pathway of interest. Unfortunately, in many cases, a clinical trial or standard option is not available. Outside of a clinical trial, a molecularly targeted approach is best facilitated through an MTB with multidisciplinary review by a dedicated team of experts, including oncologists, pharmacists, other specialized providers, geneticists, and drug procurement specialists.

For patients who fall outside of the structure of a clinical trial, a dedicated and systematic process should occur that emphasizes patient safety. The first step is to identify if the genetic variant of interest is a driver of the cancer type and not a passenger mutation. Hopefully, safety data exist for use with this agent in the organ system being treated and information is available as to whether or not the drug reaches the cancer site. Typically, this information would come from a previously completed early phase clinical trial, which in many instances may not have been performed with knowledge of whether patients harbored the mutation of interest. Once these requirements have been satisfied, then an effort can be made with the support of the MTB to procure the drug. Obtaining and paying for the drug is often met by resistance from insurance providers, as there is often not an approved indication to use this agent for the patient's disease. When a treatment plan is derived in this way, a rigorous standardized process emulating an early phase clinical trial is advised. Contemporary clinical trials, such as basket and umbrella trials, often include multiple treatment arms enrolling patients in parallel based on molecular targeting of specific pathways identified in the tumor. Clinical trials for rare conditions, such as cancers that harbor rare genetic variants, are often delivered across many institutions to achieve recruitment goals needed to statistically power research questions and justify the resources needed to implement the trial. Clinical trials are the best method to deliver PM when safety, efficacy, and side effects of a treatment plan are not known. Medicine marches forward through clinical trials that validate new treatments and interventions in a rigorous and scientifically reproducible manner.

This chapter attempts to show how advances in the life sciences and informatics have ushered in PM and the molecular taxonomy of diseases and how this information is currently used in clinical care of patients with cancer. The next steps needed are: (a) to implement this change at a larger scale and (b) the creation of a robust knowledge network that has the potential to decode contributors to health and disease. It is my hope that large-scale, population-based genetic and environmental research efforts to catalog millions of people representative of the population at large will reveal modifiable risk factors and intricate associations that lead to disease development—so that we can change these factors and prevent illness. Imagine if you could identify when an individual's modifiable risk factors for a disease, such as smoking, diet, or toxin exposure, are approaching a critical threshold that escalates risk for genetic instability and a resultant cancer or disease. Technology exists that can edit the intrinsic biologic processes that occur during disease formation. It is only a matter of time before this technology is refined to the point that editing a biologic process in humans is a realistic opportunity. Herein lies the excitement behind CRISPR, an acronym that stands for Clustered Regulatory Interspersed Short Palindromic Repeats. CRISPR technology has the power to edit DNA in a precise manner.[44]

At the time of this publication, CRISPR is in its infancy, but the implications of this technology are tremendous as are the associated ethical concerns.[45]

VI. Conclusion

The availability and routine application of vast new realms of health information has ushered in an era in health care referred to as PM. This paradigm shift has occurred due to advancements in molecular analysis using high-throughput sequencing, new ways of recording biologic and environmental data across populations, and large-scale data repositories of genetic and health data, coupled with advancement in informatics to support interdisciplinary health data aggregation and real-time analyses. Understanding disease at a genetic level has become the standard of care for many conditions and has changed how we view disease. Molecular profiling has provided insight into disease development and resulted in new treatment approaches that were previously limited in histology and symptom-based diagnosis. This molecular taxonomy of PM has revealed similarities across conditions previously thought unrelated and identified profound distinctions within conditions that share a histologic diagnosis. To deliver on PM, health care must utilize an interdisciplinary approach, develop systematic ways of recording information, and change current policy to support a new healthcare perspective. Moving forward with PM requires system-wide adjustments in how health data are recorded and delivered. The medical community is only beginning to embrace PM concepts and initiate the changes required to deliver PM, which has the capacity to dramatically improve health outcomes and prevent disease.

Notes

1. *Origo.* Retrieved from www.youtube.com/watch?v=qOhUz9FtdVE
2. *Case study: Developing a precision medicine platform solution for cancer patients at a World-Renowned Hospital.* Retrieved from http://genotechmatrix.com/wp-content/uploads/2017/05/Case-Study.pdf
3. *Origo: YouTube*, *ibid.*
4. *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease*. (2011). Washington, D.C.: National Academies Press. Retrieved from www.ucsf.edu/sites/default/files/legacy_files/documents/new-taxonomy.pdf
5. *Remarks by the President on Precision Medicine.* (2015, January 30). Retrieved from https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/remarks-president-precision-medicine
6. G. H. (n.d.). *What is DNA?* Retrieved from https://ghr.nlm.nih.gov/primer/basics/dna
7. NCI Dictionary of Cancer Terms. (n.d.) [nciAppModulePage]. Retrieved from www.cancer.gov/publications/dictionaries/genetics-dictionary
8. *The genetics and genomics of cancer | Nature Genetics* (n.d.). Retrieved from www.nature.com/articles/ng1107
9. Flores, M., Glusman, G., Brogaard, K., Price, N. D., & Hood, L. (2013). P4 medicine: How systems medicine will transform the healthcare sector and society. *Personalized Medicine*, *10*(6), 565–576. doi:10.2217/PME.13.57
10. *Network of networks.* Retrieved from http://future.psjhealth.org/scientific-wellness/about-the-institute-for-systems-biology?utm_source=TWITTER&utm_medium=social_organic&utm_term=--&utm_content=psjh-1773051757-&utm_campaign=evergreenContent+Type+%28Secondary%29
11. Yadav, S. P. (2007). The wholeness in Suffix -omics, -omes, and the Word Om. *Journal of Biomolecular Techniques: JBT*, *18*(5), 277.
12. Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. doi:10.1016/j.ygeno.2015.11.003
13. Quoted from https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/
14. Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., … Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, *452*(7189), 872–876. doi:10.1038/nature06884
15. Toward Precision Medicine, *ibid.*
16. *DNA Sequencing Costs: Data.* (n.d.). Retrieved from www.genome.gov/27541954/dna-sequencing-costs-data/
17. Toward Precision Medicine, *ibid.*

18. *Epigenomics Fact Sheet*. (n.d.). Retrieved from www.genome.gov/27532724/epigenomics-fact-sheet/
19. Definition of phenotype—NCI Dictionary of Cancer Terms. (n.d.), *ibid*.
20. National Institutes of Health (NIH)—*All of Us*. (n.d.). Retrieved from https://allofus-nih-gov.sladenlibrary.hfhs.org/
21. Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., … Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, *6*(269), pl1. doi:10.1126/scisignal.2004088
22. Griffith, M., Spies, N. C., Krysiak, K., McMichael, J. F., Coffman, A. C., Danos, A. M., … Griffith, O. L. (2017). *CIViC* is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, *49*(2), 170–174. doi:10.1038/ng.3774
23. Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *The New England Journal of Medicine*, *375*(12), 1109–1112. doi:10.1056/NEJMp1607591
24. Snyder, J., Schultz, L., & Walbert, T. (2017). The role of tumor board conferences in neuro-oncology: A nationwide provider survey. *Journal of Neuro-Oncology*, *133*(1), 1–7. doi:10.1007/s11060-017-2416-x
25. Redig, A. J., & Jänne, P. A. (2015). Basket trials and the evolution of clinical trial design in an era of genomic medicine. *Journal of Clinical Oncology*, *33*(9), 975–977. doi:10.1200/JCO.2014.59.8433.
26. *NCI Dictionary of Cancer Terms*. (n.d.). [nciAppModulePage]. Retrieved from www.cancer.gov/publications/dictionaries/cancer-terms
27. Redig & Jänne, (2015) *ibid.*
28. Frampton, G. M., Fichtenholtz, A., Otto, G. A., Wang, K., Downing, S. R., He, J., … Yelensky, R. (2013). Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature Biotechnology*, *31*(11), 1023–1031. doi:10.1038/nbt.2696.
29. Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., … Ellison, D. W. (2016). The 2016 World Health Organization classification of tumors of the central nervous system: A summary. *Acta Neuropathologica*, *131*(6), 803–820. doi:10.1007/s00401-016-1545-1.
30. Louis et al., (2016) *ibid*.
31. Toward Precision Medicine, *ibid.*
32. Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2016). New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health*, *3*(2). doi:10.2196/mental.5165
33. Groft, S. C., & Rubinstein, Y. R. (2013). New and evolving rare diseases research programs at the National Institutes of Health. *Public Health Genomics*, *16*(6), 259–267. doi:10.1159/000355929
34. Gill, J., Obley, A. J., & Prasad, V. (2018). Direct-to-consumer genetic testing: The implications of the US FDA's first marketing authorization for BRCA mutation testing. *JAMA*, *319*(23), 2377–2378. doi:10.1001/jama.2018.5330
35. My Family Health Portrait. (n.d.). Retrieved from https://familyhistory.hhs.gov/FHH/html/index.html
36. Gallin, E. K., Bond, E., Califf, R. M., Crowley, W. F. J., Davis, P., Galbraith, R., & Reece, E. A. (2013). Forging stronger partnerships between academic health centers and patient-driven organizations. *Academic Medicine*, *88*(9), 1220. doi:10.1097/ACM.0b013e31829ed2a7
37. Frank, L., Basch, E., & Selby, J. V. (2014). The PCORI perspective on patient-centered outcomes research. *JAMA*, *312*(15), 1513–1514. doi:10.1001/jama.2014.11100
38. Gallin, et al., (2013) *ibid*.
39. McNeil, C. (2015). NCI-MATCH launch highlights new trial design in precision-medicine era. *JNCI: Journal of the National Cancer Institute*, *107*(7). doi:10.1093/jnci/djv193
40. Redig & Jänne, (2015) *ibid*.
41. Lillie, E. O., Patay, B., Diamant, J., Issell, B., Topol, E. J., & Schork, N. J. (2011). The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Personalized Medicine*, *8*(2), 161–173. doi:10.2217/pme.11.7
42. Toward Precision Medicine, *ibid.*
43. Toward Precision Medicine, *ibid.*
44. Cyranoski, D. (2016). CRISPR gene-editing tested in a person for the first time. *Nature News*, *539*(7630), 479. doi:10.1038/nature.2016.20988
45. Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine*, *40*(4), 346–358.

Chapter Questions

2-1 What key events trigger PM?

2-2 What are the underlying principles of PM? Discuss the appeal and challenges of adopting PM principles for patients as well as care providers.

2-3 What drives PM's success or failure? How is PM changing the practice of traditional medicine?

2-4 What significance does PM have on personalizing cancer treatments for cancer patients?

Biography

Dr. Snyder is a board-certified Neurologist and fellowship-trained Neuro-oncologist. His practice is focused on neuro-oncologic conditions, including primary brain tumors and cancer involving the nervous system, with an emphasis on clinical trials and translational research. He received his medical degree from Michigan State University College of Osteopathic Medicine and completed post graduate education at Huron Valley-Sinai Hospital, St. John Providence Health System, and Henry Ford Hospital.

TECHNOLOGY REVIEW I

Review on Big Data Analytics in Health Care

Abir Belaala, Labib Sadek Terrissa, Noureddine Zerhouni, Christine Devalland, and Joshia Tan

Abstract

Owing to the recent digitization of medical services with tools, such as electronic health records, mobile health apps, wearable sensors, and smart fitness devices, huge amounts of medical and healthcare data have been generated and collected at an unprecedented volume, velocity, and variety. Traditional limits in handling these massive and heterogeneous datasets create the need for insights into the latest research on Big Data and Big Data techniques in health care. This review surveys the basic concepts, sources, and types of Big Data applications; the most popular analytical techniques; and tools used in the medical-Big Data field appearing in the extant literature between 2015 and 2018.

CHAPTER OUTLINE

Introduction

With massive amounts of heterogeneous data emerging from various sources,[1] such as patient information, biomarkers (e.g., genomic, proteomic, metabolomic), and diagnosis results (radiology, blood test, etc.), as well as pharmacy (e.g., prescriptions, medications), administrative (e.g., cost and claims data, population, and public health data) and behavior data (e.g., those from mobile apps, social media, sensors, wearable devices, and fitness monitors), the shift from paper-based

patient records to electronic health records (EHR) represents a necessary digitalization in today's healthcare systems. With fast growth, increased complexity, heterogeneity, and size of these accumulated data, the big challenge now is how to collect, store, analyze, and manage these Big Data in healthcare systems to improve the quality of care delivery, including the move toward personalized medicine, the sharing of real-time decisions in diagnosis and treatments, and the prediction of treatment outcomes at earlier stages, as well as the understanding of new diseases and therapies.

Traditional data analysis cannot adequately handle Big Data processing. New approaches that can analyze a wide variety of complex data and generate valuable insights are needed.[2] When applied to healthcare Big Data, these tools will have the potential to identify patterns, improve care quality, reduce costs, and enhance real-time decision-making. Big Data analytics integrate machine learning and statistical analysis. They include a set of tools and techniques, such as classification, clustering, regression, and association,[3] each serving a distinct purpose depending on the modeling objective. Often, the choice of the right technique depends on the problem at hand and how the data are represented and stored.

This review encompasses Big Data in health care. It explains the processing of Big Data in health care from collection to decision-making, citing and classifying the sources of Big Data, and illustrating the tools and technologies used to handle Big Data, for example, the *Hadoop* ecosystem. Additionally, the review covers the applied analytical techniques, such as machine learning algorithms, and sheds light on the potential benefits of Big Data to health care. Finally, it highlights some challenges of Big Data analytics and discusses potential future developments in the related areas.

Background

The review process starts with searching information databases, such as *ScienceDirect, PubMed, IEEE Xplore*, and other electronic databases, with keywords, such as "Big Data" OR "Big Data analytics" AND "Healthcare" OR "Medicine" OR "Biomedicine" OR "Medical" OR "Bioinformatics". As **FIGURE TR1-1** noted, the review covers 76 identified articles that deal with Big Data in health care published between 2015 and 2018. Six main categories emerged from an analysis of these selected

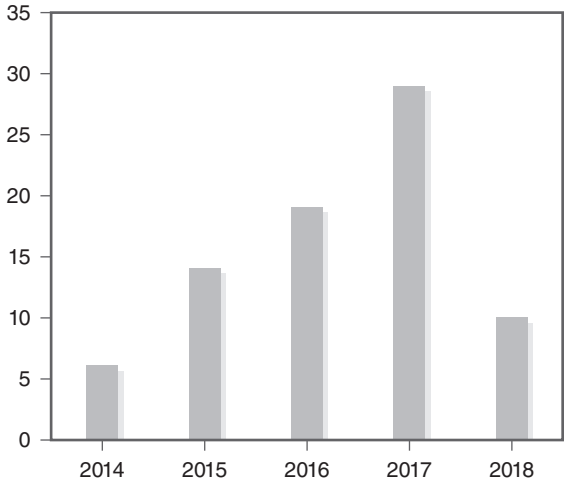

FIGURE TR1-1 Distribution of identified articles by year (76 articles).

papers: (a) Big Data definition and basic concepts; (b) Big Data sources; (c) Big Data tools and technologies; (d) Big Data analytical techniques; and, finally, (e) Big Data challenges and opportunities. These various subtopics are elaborated next.

Definition and Basic Concepts

Big Data have been defined variously in the literature. Bellazzi, et al.[4] cited Haper[5] as viewing Big Data to have "scale, diversity, and complexity," requiring "new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it." Hemingway, et al.[6] alluded to Big Data as high volume, velocity, and variety information assets demanding new forms of processing to drive enhanced decision-making, insight discovery, and process optimization. More simply, Big Data are very large datasets, structured or unstructured, static or dynamic, simple or complex, that may be gathered, stored, processed, and analyzed using different advanced techniques. **FIGURE TR1-2** contrasts between traditional versus Big Data according to *4Vs*: *Volume, Variety, Velocity,* and *Value*.

In biomedical informatics, Luo, et al.[7] and Mathew, et al.[8] define Big Data in health care according to its *Vs*: first is the exponential growth in the *Volume* of data in biomedical informatics from real-time health monitoring systems, EHRs, electronic patient records (EPRs), labs, sensor devices, and more. In fact, the U.S. healthcare system alone already reached 150 exabytes (10^{18}) of data 5 years ago.[9] Second is *Variety* of data types and structures, that is, the ecosystem of biomedical data can be structured, semi-structured, or unstructured, collected from different sources, such as wearable sensors, health community blogs, social media, and more (often in numerous formats, such as relational tables, flat files, and comma separated values or comma-separated values [CSV] files). Third is *Velocity*, which is the need to process the data in real-time, whether it is coming from streaming data, such as remote patient monitoring, from sensor devices or telemedicine servicing (e.g., the new generation of sequencing technologies that enables the production of billions of DNA sequence data each day at a relatively low cost). Fourth is *Veracity,* which deals with the quality of data being captured. Here, the truthfulness of data, or how certain we are about these data, matters. The last, and most important, *V* is *Value.* Unlike other *Vs*, this *V* is the desired outcome of processing Big Data in health care as we are

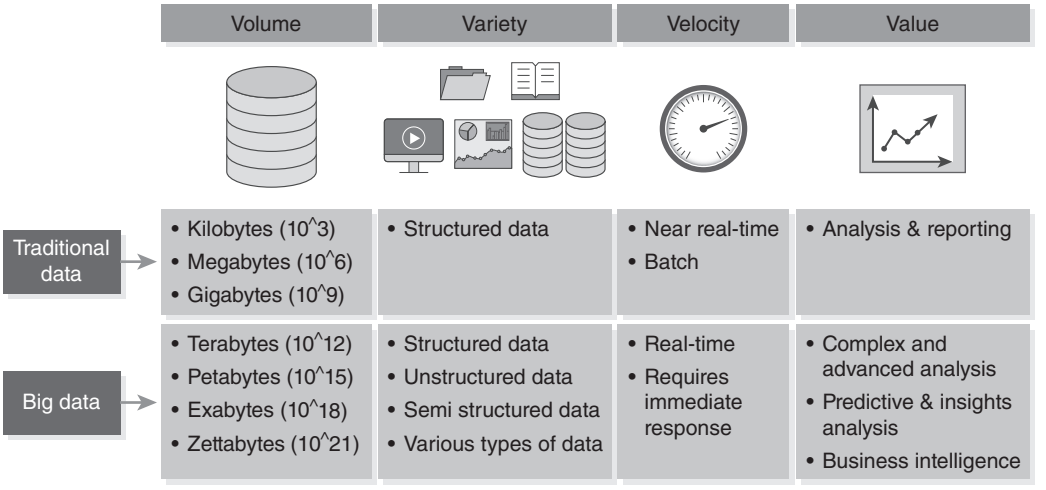

FIGURE TR1-2 Traditional Data vs. Big Data.

primarily interested in extracting maximum value and generating insights from Big Data so as to improve the quality of health care.

Sources

In health care, data heterogeneity and the variety of structured, semi-structured, and unstructured data are derived from diverse biomedical data sources. These include physiological, behavioral, molecular, clinical, environmental exposure, medical imaging, disease management, medication prescription history, nutrition, exercise parameters, and more.[10]

Big Data sources have been classified in various ways in the literature. Stokes, et al.[11] divide data sources into two general classes: Administrative (Government [CMS], National surveys [Medical Expenditure Panel Survey], commercial vendors [health plans, PBMs]) versus Clinical (Hospital EMR, Physician EMR, Integrated delivery network EMR, Clinical database). Hemingway, et al.[12] simply suggest a classification using structured versus unstructured data in clinical care: *Structured EHR* data are recorded using controlled clinical terminologies, such as Systematized Nomenclature of Medicine Clinical terms (SMOMED-CT) or statistical classification systems, such as ICD-9, ICD-9-CM, or ICD-10, while *unstructured clinical data* can be patient medical histories, discharge summaries, handover notes, and imaging reports. These data are often captured and recorded in patient's health records as raw unformatted text.

In Mathew & Pillai,[13] Big Data sources may come from: (a) Providers: medical data (EHRs, EPRs); (b) Payers: claims and cost data; (c) Researchers: academic or independent; (d) Consumers and Marketers: patient behavior and sentiment data; (e) Government: population and public health data; and/or (f) Developers: pharmacy and medical device research and development (R&D). Briefly, two underlying types of sources emerged here: internal sources, such as EMRs, computerized provider orders entry (CPOE), imaging data, and others versus external data sources, such as government, insurance (e.g., claims, billing), and social media. Andreu-Perez, et al.[14] also focused on two clusters: quantitative (e.g., sensor data, images, gene arrays, laboratory tests) versus qualitative (e.g., free text, demographics). However, Ma, et al.[15] identified four major sources of pharmacy Big Data: (a) Pharmaceutical research and development from pharmaceutical companies and academia, clinical trials, and high-throughput screening libraries; (b) Claims and cost data from payers and providers that contain utilization of care and cost estimates; (c) Clinical data provided by the EMR that contain patient-specific data on treatment outcomes; and (d) Patient behavior and sentiment data that come from consumers and stakeholders outside of health care (for instance, from retail exercise apparel and exercise monitoring equipment). Finally, Fang, et al.[16] classify healthcare Big Data differently with categories ranging from: (a) Human-generated data: physicians' notes, email, and paper documents; (b) Machine-generated data: readings from diverse health monitoring devices; (c) Transaction data: billing records and healthcare claims; (d) Biometric data: genomics, genetics, heart rate, blood pressure, X-ray, fingerprints; (e) Social media data: interaction data from social websites; to (f) Publications: clinical research and medical reference material.

The large variety of Big Data in health care sources and their corresponding classifications inspired from aforementioned authors are summarized in **FIGURE TR1-3**, showcasing prominent taxonomies of Big Data sources in health care.

FIGURE TR1-4 offers another proposed classification which adds more details about data types including their format (text or ASCII, image, and video) and sources (internal, external). This domain-based classification is constructed according to three specialized areas in health care: cardiology, diabetes, and oncology. **TABLE TR1-1** presents various data types used in the selected papers being reviewed.

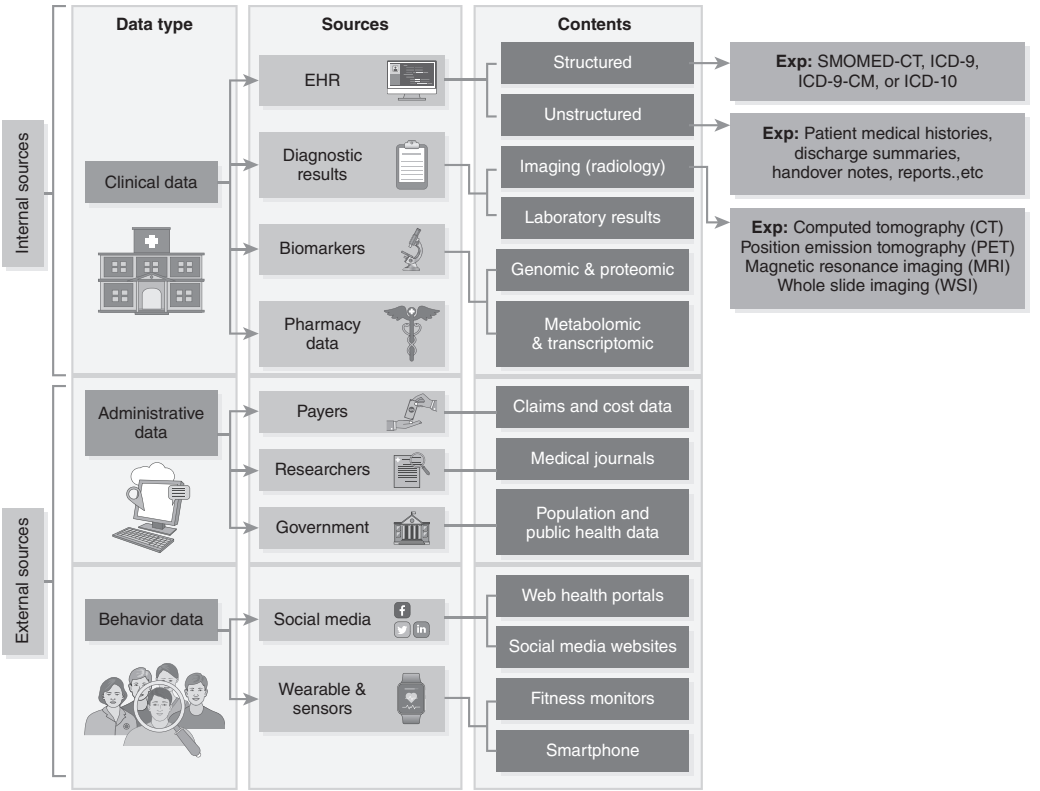

FIGURE TR1-3 Main sources of Big Data in health care.

Tools

Big Data in health care, which are difficult to store and process via traditional methods, require the use of new technological tools for their capture from different sources and systems, their transformation, storage, analysis, and visualization. Mathew & Pillai[17] classify tools of Big Data into two options: open source versus available commercial solutions. Here, some key products include Hadoop-based analytics, data warehouse for operational insights, stream computing software for real time analysis of streaming data, and NoSQL databases such as Cassandra, MongoDB, and DynamoDB.

We start with Apache Hadoop open source platform as it is among the earliest tool successfully applied in different Big Data specialized software projects. Hadoop supports the processing and storage of extremely large datasets in a distributed computing environment. Data in a Hadoop cluster is divided into small pieces and stored throughout a computer cluster with thousands of nodes. Hadoop uses two main components, MapReduce and Hadoop Distributed File System (HDFS). Closely related software tools include NoSQL databases such as MongoDB, Cassandra, and HBase, which are basically an open source version of BigTable.[18]

Vijayarani & Sharmila[19] classify Big Data tools vis-à-vis Big Data phases:

- In Big Data storage, three types of storage (in memory, in the cloud, and hard disk storage) are noted;
- In Big Data processing, we have real-time processing using Storm, Spark, S4, and more versus batch processing (Hadoop); and

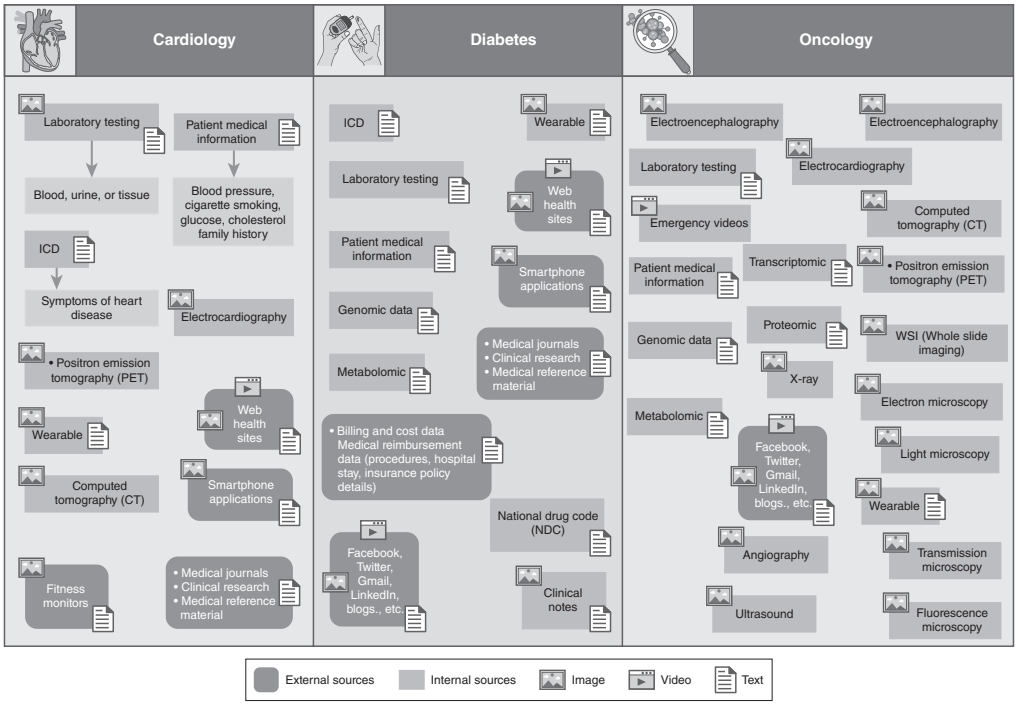

FIGURE TR1-4 Domain-based classification of Big Data in health care.

- In Big Data technologies, we have many successful applications in biomedicine,[20] including four types of tools used in bioinformatics, clinical informatics, and imaging informatics:
  - Tools used in data storage and retrieval;
  - Error identification;
  - Data analysis; and
  - Platform integration deployment.

Bellazzi, et al.[21] highlighted the main types of Big-Data tools oriented solutions in health care as comprising of: cloud computing, parallel programming, and NoSQL databases.

TABLE TR1-2 (adapted from Lourenço, et al.[22]) shows the existing Big Data platforms and tools for storing Big Data, with the advantages and disadvantages of each technique. NoSQL databases (i.e., nontraditional relational databases) are becoming the core technology for Big Data. Here, we examine the following three main NoSQL databases: key-value databases, column-oriented databases, and document-oriented databases, each based on certain data models.

Sharing and storing data over the cloud plays a key role in providing flexible, reliable, and cost-effective solutions to users.[23–25] Despite advantages of a cloud-based healthcare system, privacy of data is a major problem.[26] **TABLE TR1-3** highlights the existing Big Data platforms and tools for batch and real-time processing. As shown, there are three main types of Big data processing tools: (i) batch-only tools, (ii) stream-only tools, and (iii) hybrid tools (see also **FIGURE TR1-5**).

Analytical Techniques

The literature on Big Data techniques in health care is broad. Alonso, et al.[27] highlighted the most popular techniques of machine learning and Big Data classification (decision tree, Naïve Bayes, Artificial Neural Network

TABLE TR1-1 Types of Data Used in Literature

Data Type	Reviewed Papers
Genomic data	Turgut, et al.[89]; Xiao, et al.[96]; Su, et al.[84]; Hinkson, et al.[24]; Maia, et al.[62]; Morovvat, et al.[67]; Shah, et al.[78]; Ding, et al.[46]; Zheng & Zhang[97]
Imaging	Volynskaya, al.[92]; Nawaz, et al.[68]; Kurc, et al.[60]; Shah, et al.[78]; Margolies, et al.[65]; Albarqouni, et al.[37]; Wang, et al.[2]; Silva, et al.[81]; Panayides, et al.[69]; Kovalev, et al.[59]; Alickovic & Subasi[38]; Ivanova[56]; Hinkson, et al.[24]
Biomedical data	Asri, et al.[41]; Shen, et al.[79]
Behavior data	Asri, et al.[41]; Shen, et al.[79]
Pharmaceutical data	Choi, et al.[43]; Ma, et al.[15]; Walczak & Okuboyejo[93]
Billing data	Erekson & Iglesia[49]
Clinical notes	Forsyth, et al.[50]
Lab tests	Miranda, et al.[66]
CDI_9	Choi, et al.[43]; Forsyth, et al.[50]

[ANN]) to bundle the objects or data into groups. Clustering and search optimization are also applied as data mining strategies, such as self-organization map; vector quantization; and genetic algorithm, regression, association, and prediction. Bachiller, et al.[28] showed the various computational methods applied in health care and classified them into two clusters: machine learning (Support Vector Machines or SVM, Naïve Bayes, ANN, Auto-encoders); and deep learning (Convolutional Neural Networks, Recurrent Neural Network, Restricted Boltzmann Method). Mathew & Pillai[29] showed that analytics can be classified into three major types: predictive, descriptive, and prescriptive analytics. Mehta & Pandit[30] reviewed some of the Big Data analytical techniques across various healthcare applications including cluster analysis, data mining, graph analytics, machine learning, natural language processing (NLP), neural networks (NNs), pattern recognition, spatial analysis, and more to argue that the choice among techniques really depends on the problem at hand and the nature of the stored datasets.

As shown in Table TR1-3, there is a large variety of techniques for Big Data in health analytics. Each technique serves a different purpose depending on the modeling objective, with some techniques applicable to more than one modeling objectives (e.g., classification, regression, clustering, and more). **FIGURE TR1-6** maps out the existing analytical techniques vis-à-vis their utilizations whereas **TABLE TR1-4** defines existing computational algorithms popularly used in the medical field.

As noted previously, there are three main types of analysis: (a) Diagnostic analytics are used to answer what happened and why it happened; (b) Predictive analytics cater to

TABLE TR1-2 NoSQL Databases Comparison

Big Data Storage Platforms	Store Type	Cons	Pros
Cassandra	Column oriented data stores	Recovery Time Read Performance	Write-Performance Multi data center replication High scalability Supports rich data structure and Powerful query language (CQL). Availability Consistency
HBase		Availability Read Performance Robustness	Consistency Partition tolerance Scalability
BigTable		Availability Read Performance	Consistency Partition tolerance
MongoDB	Document data stores	Availability Scalability Write-Performance Stabilization Time	Support complex data types Consistency Partition tolerance Powerful query language High-speed access Reliability
CouchDB		Consistency Write-Performance Scalability	Flexible Availability Partition tolerance (AP)
DynamoDB	Key-value stores	Unable to do complex queries Latency in read/write	High expandability and smaller query response time Consistency Automatic data replication
Voldemort		Consistency	Availability Partition tolerance Write-Performance
Redis		Availability	Consistency Partition tolerance
OrientDB	Graph oriented data stores	Requires more schema design up front	Useful in dealing with data where relationships play an important role Easy to query Robust
Neo4j			

Data from Lourenço, J. R., Cabral, B., Carreiro, P., Vieira, M., & Bernardino, J. (2015). Choosing the right NoSQL database for the job: a quality attribute evaluation. *Journal of Big Data, 2*(1), 18.[22]

TABLE TR1-3 Big Data Processing Tools

Processing Type	Big Data Platform	Definition
Batch processing	Hadoop MapReduce	The MapReduce is a parallel programming model that enables many of the most common calculations on large scale data to be performed on computing clusters containing a large number of computing nodes efficiently using two functions: Map and Reduce (Rahim, et al.[74]).
	Oozie	Oozie is a workflow processing method that allows users to define a series of jobs written in different languages (e.g., Pig, Hive, and MapReduce) and then logically links them with each another (Raghupathi & Raghupathi[73]).
	Mahout	Mahout is another Apache project; it enables the generation of free applications of distributed and scalable machine learning algorithms that support big data analytics on the Hadoop platform (Landset, et al.[61]).
	Hive	Hive is a runtime Hadoop support architecture that supports Structure Query Language (SQL) with the Hadoop platform. It permits SQL programmers to develop Hive Query Language (HQL) statements similar to SQL statements (Raghupathi & Raghupathi[73]).
Batch processing	Pig	Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin. Pig programming language is configured to assimilate all types of data (structured/unstructured, etc.) (Singh & Reddy[82]).
Stream processing	Spark	Apache Spark is a next generation batch processing framework with stream processing capabilities. On the speed side, Spark extends the popular MapReduce model to efficiently support more types of computations, including interactive queries and stream processing (Singh & Reddy[82]).
	Storm	Apache Storm is an open-source Apache tool; its scalable and fast distributed framework has a special focus on stream processing. Storm provides a topology to control data transfers, which is a critical part of routing data where it needs to go for analytics and other operations (Fang, et al.[16]).
	Flink	Apache Flink is a tool for supporting Hadoop project structures and processing real-time data. Its stream processing framework can also handle batch tasks. As a type of batch processor, Flink contends with the traditional MapReduce and new Spark options (Gurusamy, et al.[53]).

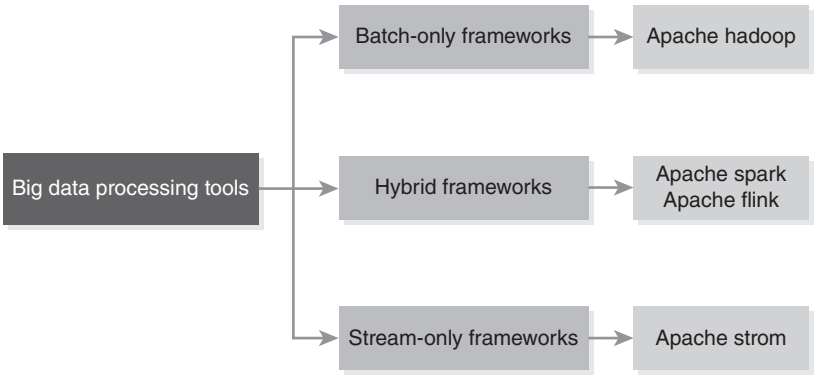

FIGURE TR1-5 Types of Big Data processing tools.

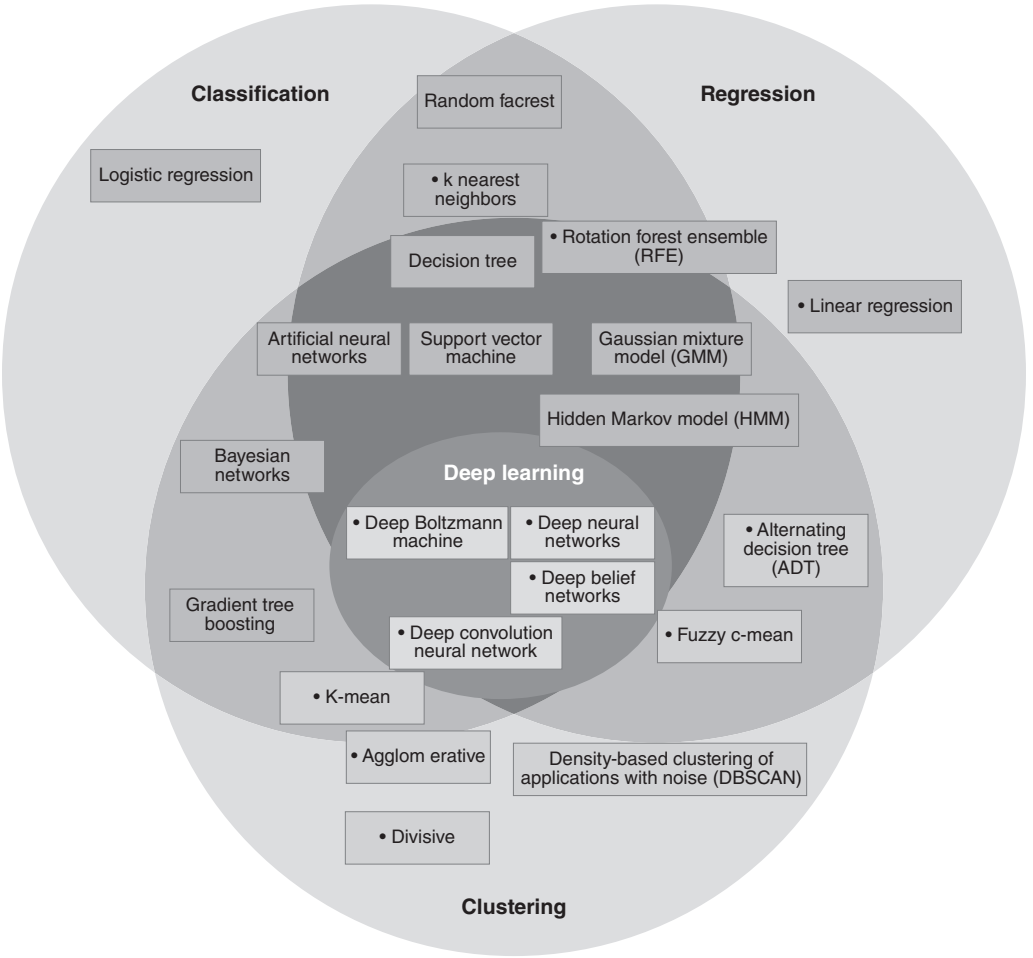

FIGURE TR1-6 Mapping out the classification of existing analytical techniques.

TABLE TR1-4 Analytical Techniques and Their Application in Health Care

Technique	Definition	Application Area	Description
Decision Tree (DT)	DT is a most popular and powerful classification technique. It classifies instances by sorting them in a tree, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. (Sivakami & Saraswathi[83]).	Oncology	This publication presents a decision tree based data mining technique for early detection of breast cancer. This is helpful because early detection of breast cancer makes it far easier to cure (Sumbaly, et al.[85]).
Naïve Bayes	Naïve Bayes are probabilistic classifiers based on applying Bayes theorem with strong independence hypothesis between the features (Prerana, et al.[71]).	Cardiology	This paper proposes a mining model using a naïve Bayes classifier that could detect cardiovascular disease and identify its risk level for adults (Miranda, et al.[66]).
Logistic regression	Logistic regression is a statistic model where the log-odds of the probability of an event are a linear combination of independent or predictor variables (Alickovic & Subasi[38]).	Oncology	This work aims to predict grad 2 acute radiation-induced dermatitis after hybrid intensity modulation radiotherapy for breast cancer using a logistic regression normal tissue complication probability model (Sung, et al.[87]).
Artificial Neural Network (ANN)	ANNs are a family of computational models based on biological neural networks, which are used to estimate complex relationships between inputs and outputs (Wu, et al.[95]).	Cardiology	They use decision support systems based on artificial neural networks to predict heart failure risks (Samuel, et al.[75]).
Support Vector Machines (SVM)	SVM is an example of supervised learning. Known labels help indicate whether the system is performing the right way or not. This information points to a desired response, either validating the accuracy of the system, or to help the system learn to act correctly (Sivakami & Saraswathi[83]).	Diabetes	The paper explores the hybrid of SVM and a system of ANN as the finest binary classification system for calculating the diabetic nature of people in comparison to SVM (Aliwadi, et al.[39]).

Random forest	Random forest is one type of ensemble learning algorithm that constructs multiple trees at training time. This algorithm overlaps the over fitting problem of decision trees by averaging multiple deep decision trees (Fang, et al.[16]).	Genomics	Identify variables correlated with a diagnosis of diabetic peripheral neuropathy (DPN) using random forest modeling applied to EHR (DuBrava, et al.[47]).
Hierarchical clustering	In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters (Ding, et al.[46]).	Genomics	This work aims, to find differentially expressed genes rather than directly de-noise the single cell data. They present a method to remove technical noise. These cells use these genes to cluster by hierarchical clustering (Ding, et al.[46]).
K-means	K-means is a known partitioning clustering algorithm. It partitions objects into k clusters, computes centroids (mean points) of the clusters, and assigns every object to the cluster that has the nearest mean in an Expectation-Maximization fashion (Fang, et al.[16]).	Cardiology	In this work they use medical terms, such as age, weight, gender, blood pressure, and cholesterol rate, for prediction. To perform grouping of various attributes, it uses a k-means algorithm and for predicting it uses the Back propagation technique in neural networks (Malav, et al.[63]).
Hidden Markov Model (HMM)	HMM is a statistical model representing probability distribution over the sequences of observations. This model uses a Markov chain to model signals in order to calculate the occurrence probability of states (Fang, et al.[16]).	Oncology	They use a Bayesian HMM with Gaussian Mixture (GM) clustering approach to model the DNA copy number change across the genome for cancer diagnosis (Manogaran, et al.[64]).
Gaussian Mixture Model (GMM)	GMM is a statistical model widely used as a classifier in pattern recognition tasks. It consists of a number of Gaussian distributions in the linear way (Fang, et al.[16]).	Oncology	This paper proposes a framework using voice pathology assessment as a case study. The machine learning algorithms in the form of a support vector machine, an extreme learning machine, and a GMM are used as the classifier (Hossain & Muhammad[55]).

(*continues*)

TABLE TR1-4 Analytical Techniques and Their Application in Health Care (*continued*)

Deep learning	The success of deep learning for big data is the use of a large number of hidden neurons and parameters such as deep neural networks, deep convolution neural network, and deep belief networks (Remadna, et al.[98]).	Oncology (Breast cancer)	This paper presents results of the use of the deep learning approach and Convolutional Neural Networks (CNN) for the problem of breast cancer diagnosis (Kovalev, et al.[59]).

knowing what will happen; and (c) Prescriptive analytics are used to find the best course of actions by providing decision support for specific scenarios or situations. **TABLES TR1-5A** and **5B** classify papers in the extant literature according to the type of analysis and summarize which machine learning techniques have been used.

By combining big data and machine learning, the knowledge and information hidden in Big Data can be uncovered to improve the quality of healthcare delivery. As shown in **FIGURE TR1-7**, key benefits are allowing diseases to be detected at earlier stages; making the right treatment decisions at the right time; identifying new diseases, new therapies, and new approaches for health care; and reducing costs.

Figure TR1-7 concludes by showing the high complexity of the big health data processing steps, which transforms the raw big health data into valuable insights. This is due to the difficulty faced in each step, with the large variety of data types and a range of competing choices in selecting the best tools and techniques to store and analyze these datasets. **TABLE TR1-6** shows the distribution of identified papers in this review (76 articles) according to the application domains.

Big Data Challenges

Despite the large potential benefits of exploring big data uses in health care, challenges and problems remain to be resolved if outcomes are to be improved. Hemingway, et al.[31] highlighted several formidable challenges: data quality; knowing what data exist; the legal–ethical dimension for their use; data sharing; building and maintaining public trust; developing standards for defining disease; developing tools for scalable, replicable science; and equipping the clinical and scientific work force with new interdisciplinary skills. Other challenges identified by Mathew & Pillai[32] include the lack of standards for representing and sharing of healthcare data, the complication in integrating heterogeneous data sources, the need for skilled resources, attention to privacy, security and infrastructure issues, the need for quality control of the acquired and input data, the demand on real-time processing, and the interpretation of the analytical results.

More challenges are identified by Cyganek, et al.[33] These include the understanding of doctors' notes (unstructured text analysis); the handling of huge volumes of medical images that are part of the EHR, which increase storage requirements; and the need to backtrack the effect of medical decisions. In pharmacy, Ma, et al.[34] noted several challenges for big data: (a) a storage challenge on the size scale of petabytes, for secure data transmission and continued development of tools to analyze the data; (b) a variety challenge and the issue of data integrity and validity; (c) a patient confidentiality challenge where Big data also raises issues regarding how to keep the information safe; and (d) a physician prescribing patterns challenge; here, the issue at hand is whether detailed information about prescriptions written by doctors (with the doctor identified) can be bought and sold.

TABLE TR1-5A The Utilization of Machine Learning Technique by Type of Analysis

	K-means	SVM	DT	ANN	CNN	RF1	RF2	LR	NB	RVM	MLP	KNN	GBM	Ada Boost	Others
Chen, et al.[25]		✓	✓	✓	✓										
Zheng & Zhang[97]						✓		✓						✓	
Sundara-sekar[86]		✓	✓									✓			
Ivanova[56]; Miranda, et al.[66]									✓						
Alickovic & Subasi[38]		✓	✓				✓	✓	✓		✓				
Asri, et al.[41]		✓	✓						✓			✓			
Wang, et al.[2]															✓
Shen, et al.[79]		✓													
Albarqouni, et al.[37]					✓										
Silva, et al.[81]	✓		✓	✓	✓				✓						
Turgut, et al.[89]		✓	✓			✓	✓				✓	✓	✓	✓	
Forsyth, et al.[50]															✓
Gambhir, et al.[51]			✓	✓						✓					

TABLE TR1-5B Prognostic

	K-means	SVM	DT	ANN	CNN	RF1	RF2	LR	NB	RVM	MLP	KNN	GBM	Ada Boost	Others
Kalyankar, et al.[57]; Prasad, et al.[70]; Brims, et al.[42]; Shah, et al.[78]															
Kourou, et al.[58]		✓	✓	✓					✓						
Sivakami & Saraswathi[83]		✓	✓												
Nawaz, et al.[68]; Hernandez, et al.[54]; Amirian, et al.[40]; Choi, et al.[43]															✓
Asri, et al.[41]		✓	✓						✓			✓			
Morovvat, et al.[67]		✓	✓						✓						
Xiao, et al.[96]		✓	✓			✓						✓			
Forsyth, et al.[50]															✓
Walczak & Okuboyejo[93]				✓											
Priyanga, et al.[72]		✓		✓								✓			

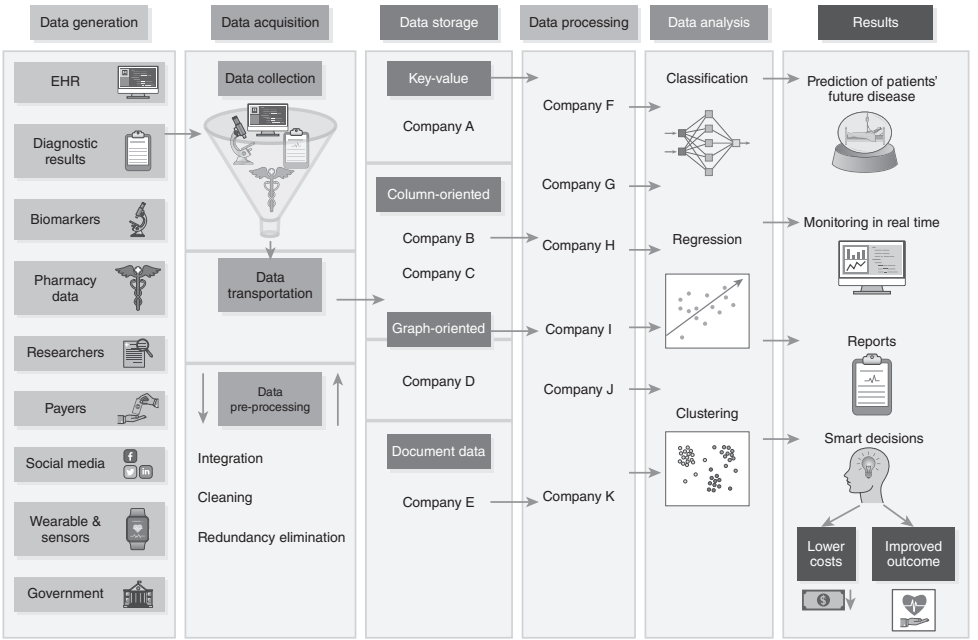

FIGURE TR1-7 Big Health Data Process.

TABLE TR1-6 Distribution of Selected Papers by Application Domains

Big Health Data Application Areas	Reviewed Papers
Ophthalmology	Clark, et al.[44]
Alzheimer	Geerts, et al.[52]; Varatharajan, et al.[90]; Aramendi, et al.[36]
Oncology	Turgut, et al.[89]; Xiao, et al.[96]; Forsyth, et al.[50]; Albarqouni, et al.[37]; Asri, et al.[41]; Margolies, et al.[65]; Shah, et al.[78]; Kurc, et al.[60]; Nawaz, et al.[68]; Brims, et al.[42]; Thiebaut, et al.[23]; Su, et al.[84]; Hinkson, et al.[24]; Ivanova[56]; Alickovic & Subasi[38]; Maia, et al.[62]; Kovalev, et al.[59]; Taglang & Jackson[88]; Wang, et al.[2]; Silva, et al.[81]; Volynskaya, et al.[92]
Pharmacy	Hernandez & Zhang[54]; Ma, et al.[15]; Geerts, et al.[52]; Taglang & Jackson[88]
Diabetes	Prasad, et al.[70]; Kalyankar, et al.[57]; Bellazzi, et al.[4]; Zheng & Zhang[97]; Chen, et al.[25]; Miranda, et al.[66]; Eljil,et al.[48]; Saravana, et al.[76]; Aliwadi, et al.[39]
Cardiology	Miranda, et al.[66]; Hemingway, et al.[6]; Choi, et al.[43]; Priyanga & Naveen[72]
Personalized medicine	Daniel, et al.[45]; Viceconti, et al.[91]

Finally, Mehta & Pandit[35] argued that major challenges include patient privacy and confidentiality; missing data and the risk of false-positive associations; security issues, such as Big Data breaches; the limitations of observational data, including data inconsistency and inaccuracy; the lack of knowledge about which data to use and for what purpose; the lack of appropriate IT infrastructure; the transition from use of paper-based records to use of distributed data processing; the lack of knowledge about the best algorithm and tool for analysis; the unavailability of trained clinical scientists and Big Data managers for interpretation of Big Data outcomes; and the need for a simple, convenient, and transparent Big Data analytics system which can be applied for real-time cases.

Discussion

This review highlights the role of Big Data in enhancing care quality. Specifically, it identified the latest findings on Big Data in health research between 2015 and 2018. Evidently, there is a variety of Big Data definitions largely focusing on Big Data's characterization as large volume, high velocity, huge variety, value, and veracity. In medicine, Big Data have been applied across various key domains including cardiology, diabetes, oncology, pharmacy, and more. **FIGURE TR1-8**, which displays the percentage of Big Data articles applied in specific healthcare domains, shows that oncology has the largest interest in most of the latest research work on Big Data (53%).

The oncology Big Data research includes all types of cancer, especially breast and lung cancer. Diabetes comes next (13%), with pharmacy (11%) and cardiology (9%) following. The other domains comprise only between 2% and 5% of Big Data in health applications. It may be concluded that the absence of effective cancer treatments has led Big Data researchers to focus on the oncology domain and how Big Data analytics can be used to understand these very complex diseases.

The medical field is considered among the most important sources of Big Data. In gathering Big Data in health care (see Figure TR1-3), we notice varied sources such as:

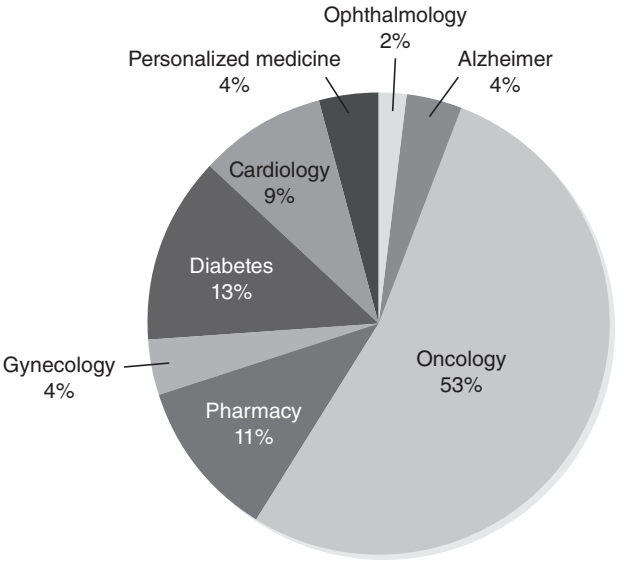

FIGURE TR1-8 Percentage of medical domains applied in big data research papers.

healthcare providers, laboratories, diagnostic companies, insurance companies, pharmaceutical firms, fitness devices and wearable sensors, government, and Web-health portals. These diverse sources generate data in various types and formats: structured, unstructured, and semi-structured data in the form of text, image, video, audio, ASCII characters, and so on. **FIGURE TR1-9** shows the results of Table TR1-1 that presents data type used in the papers we reviewed. Here, four main types were identified: genomic, behavior, imaging, and pharmaceutical data. The other EHR data involve clinical notes, International Classification of Disease or International Classification of Disease (ICD) codes, blood tests, and more. As shown in the histogram, it is clear that the majority of papers use imaging data (16 articles), including Whole Slide Imaging (WSI), Computed Tomography (CT), Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), X-ray, infrared thermographs, and more. As expected, genomic data represent the second most dominant type.

Altogether, the Big Data heterogeneity led to the issue of data integrity and validity. Vendors offer a variety of extract transform load (ETL) and data integration tools designed to make the process easier, but many researchers believe that they have yet to solve the data integration problem. The collected data from care monitoring devices vary with respect to noise, redundancy, consistency, and more. The challenge here is to improve the data quality so as to get accurate analytics (**FIGURE TR1-10**).

In the storage and processing of big health data, the extant literature on Big Data tools and techniques is broad and largely varied (see Table TR1-2). But we still have the problem of infrastructure, cost, security, corruption, scalability, user interface (UI), and accessibility. The latest research has identified that the Hadoop ecosystem is the most common adopted family of software tools used for storage and processing big health data, but since they are batch-processing tools, developers have created new tools for streaming and real-time data; for example, Spark, Storm, and GraphLab. Cloud computing has also increased our attention on accessing and storing Big Data. In health care, to share and store data over the cloud plays a key role in offering flexible, reliable, and cost-effective solution to users. Despite many advantages of a cloud-based healthcare system, security and privacy of data remain a major cause for concerns,

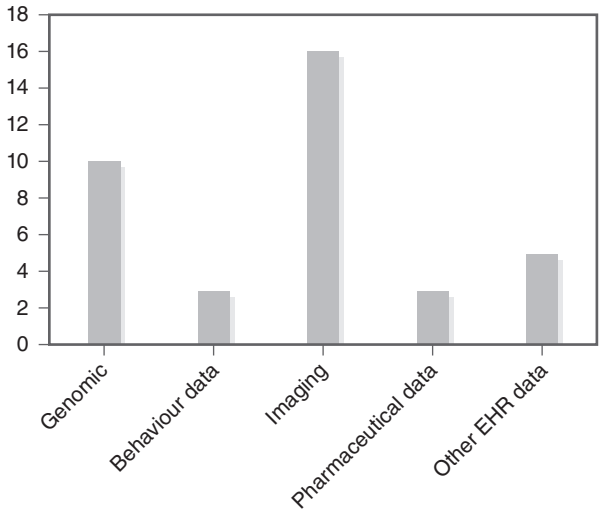

FIGURE TR1-9 Data Type used in the Extant Literature.

which have restricted the acceptance of the cloud-based model.

Various analytical techniques have been applied in health care. Currently, the most popular one is machine learning, such as NNs and SVM, decision tree, deep learning, k-means, random forest, rotation forest, conventional NN, and more. **FIGURE TR1-11** presents the machine-learning techniques most cited in the literature according to analysis type (diagnosis, predictive, prescriptive).

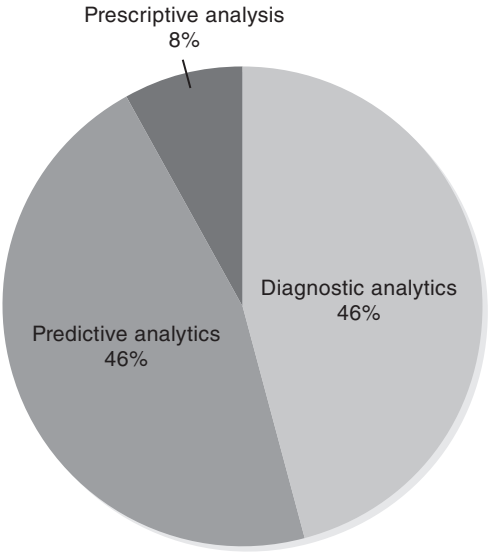

FIGURE TR1-10 Percentage of papers utilized each type of analysis.

Overall, we notice that SVM is the most used technique in diagnosis analytics. However, in predictive analysis, decision tree dominates. All in all, SVM appears to be the most accurate compared to other techniques. Despite this large advanced analytics, we still have some critical questions in this phase; for example: Does all data need to be analyzed? How does one go about finding out which data points are really important? How can the data be used to the best advantage? Which technique is more accurate? As the accuracy of medical analysis is critical, any mistake in diagnosis or prediction puts the patient's life at risk. So, the huge volume of data poses technological challenges not only for storage on the size scale of petabytes but for continued development of tools to analyze the data

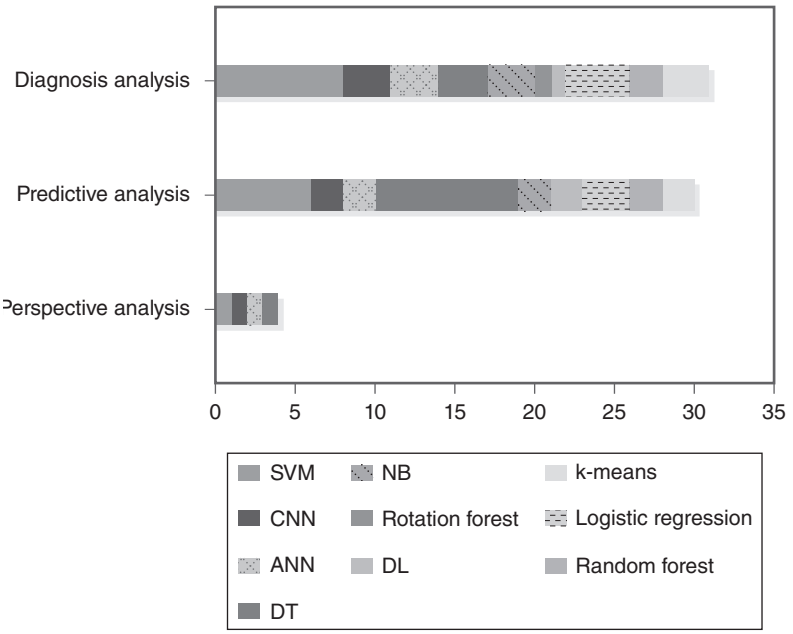

FIGURE TR1-11 Use of Machine Learning Technique vis-à-vis Major Type of Analysis.

properly, for knowing what has happened, why it happened (diagnostic), what will happen (predictive) and how we can make it happen (prescriptive). The target here is to find how to choose the right technique with the right data, to make the right decision at the right time, at the lowest cost.

Data analysis is the final and the most important phase in the processing of Big Data in health care. It has three main types:

- Diagnosis analytics is usually used to answer the question what happened and why it happened? It uses the past and current healthcare data to make quality healthcare decisions.
- Predictive analytics can be used to forecast what might happen in the future. It uses statistical approaches to search through large patient datasets and analyzes those data to predict individual patient outcomes.
- Prescriptive analytics is a type of analytics used to prescribe actions for the decision makers to act upon. In health care, prescriptive analytics is used in evidence-based medicine to improve patient care and to prescribe better business practices.

The graph in Figure TR1-10 shows results adapted from Table TR1-5, which illustrates the distributed percentages of papers that deal with diagnosis, prognosis, and perspective analyses. From the pie chart, it is clear that the majority of papers focus on diagnosis and prediction using big data in health care with equal percentage (46%); only a small minority falls in the domain of prescriptive analysis.

Conclusion

This work overviews Big Data analysis to improve health sector performance. It has focused on the newer scientific research published between 2015 and 2018 to identify the latest trends and direction of researchers in this field. The review affords a comprehensive picture of how Big Data analysis can impact medicine. Yet challenges abound, the most prominent of which is the nature and integrity of the Big Datasets serving as input to the analysis. On account of the strong relationship between quality of data and accuracy of analysis results that led to the decision taken, in addition to the sensibility of working on human lives, researchers should concentrate on this problem, as any mistake can have critical consequences.

The future of Big Data health analytics sees rapid advances in more empowering tools and technologies, incorporating greater intelligence, more user-friendliness, and other optimization features so as to ease users in making the appropriate choice when choosing among the various techniques applicable to particular dataset(s). With Big Data analytics exhibiting greater success in improving care quality, effectiveness and cost, a deeper understanding of patients, a more personalized treatment, as well as a great help for doctors to make the right decisions, there is hope for greater longevity among humankind.

Notes

1. Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics, 114*, 57–65.
2. Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718.
3. Alonso, S. G., de la Torre Díez, I., Rodrigues, J. J. P. C., Hamrioui, S., & López-Coronado, M. (2017). systematic review of techniques and sources of big data in the Healthcare Sector. *Journal of Medical Systems*, *41*(11), 183.
4. Bellazzi, R., Dagliati, A., Sacchi, L., & Segagni, D. (2015). Big data technologies: New opportunities for diabetes management. *Journal of Diabetes Science and Technology, 9*(5), 1119–1125.
5. Haper, E. (2014). Can big data transform electronic health records into learning health systems? *Studies Health Technology Informatics*, *2014*(201), 470–475.

6. Hemingway, H., Asselbergs, F. W., Danesh, J., Dobson, R., Maniadakis, N., Maggioni, A., & Anker, S. D. (2018). Big data from electronic health records for early and late translational cardiovascular research: Challenges and potential. *European Heart Journal, 39*(16), 1481–1495.
7. Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big data application in biomedical research and health care: A literature review. *Biomedical Informatics Insights, 8*, BII-S31559.
8. Mathew, P. S., & Pillai, A. S. (2015, March). *Big data solutions in healthcare: Problems and perspectives*. 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (pp. 1–6), IEEE, Coimbatore, India, 19–20 March 2015.
9. Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C., & Yang, G. Z. (2015). Big data for health. *IEEE Journal of Biomedical and Health Informatics, 19*(4), 1193–1208.
10. Mehta & Pandit, (2018), *ibid*.
11. Stokes, L. B., Rogers, J. W., Hertig, J. B., & Weber, R. J. (2016). Big data: Implications for health system pharmacy. *Hospital Pharmacy, 51*(7), 599–603.
12. Hemingway et al., (2018), *ibid*.
13. Mathew & Pillai, (2015), *ibid*.
14. Andreu-Perez et al., (2015), *ibid*.
15. Ma, C., Smith, H. W., Chu, C., & Juarez, D. T. (2015). Big data in pharmacy practice: Current use, challenges, and the future. *Integrated Pharmacy Research & Practice, 4*, 91.
16. Fang, R., Pouyanfar, S., Yang, Y., Chen, S. C., & Iyengar, S. S. (2016). Computational health informatics in the big data age: A survey. *ACM Computing Surveys (CSUR), 49*(1), 12.
17. Mathew & Pillai, (2015), *ibid*.
18. Huang, T., Lan, L., Fang, X., An, P., Min, J., & Wang, F. (2015). Promises and challenges of big data computing in health sciences. *Big Data Research, 2*(1), 2–11.
19. Vijayarani, S., & Sharmila, M. S. (2016). Research in big data—An overview. *Informatics Engineering, an International Journal (IEIJ), 4*(3), 19–23.
20. Luo et al., (2016), *ibid*.
21. Bellazzi et al., (2015), *ibid*.
22. Lourenço, J. R., Cabral, B., Carreiro, P., Vieira, M., & Bernardino, J. (2015). Choosing the right NoSQL database for the job: A quality attribute evaluation. *Journal of Big Data, 2*(1), 18.
23. Thiebaut, N., Simoulin, A., Neuberger, K., Ibnouhsein, I., Bousquet, N., Reix, N., & Mathelin, C. (2017). An innovative solution for breast cancer textual big data analysis. arXiv preprint arXiv:1712.02259.
24. Hinkson, I. V., Davidsen, T. M., Klemm, J. D., Chandramouliswaran, I., Kerlavage, A. R., & Kibbe, W. A. (2017). A comprehensive infrastructure for big data in cancer research: Accelerating cancer research and precision medicine. *Frontiers in Cell and Developmental Biology, 5*, 83.
25. Chen, M., Yang, J., Zhou, J., Hao, Y., Zhang, J., & Youn, C. (2018). 5G-Smart Diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds. *IEEE Communications Magazine, 56*, 16–23.
26. Bouzidi, Z., Terrissa, L. S., Zerhouni, N., & Ayad, S. (2018). An efficient cloud prognostic approach for aircraft engines fleet trending. *International Journal of Computers and Applications*, 1–16.
27. Alonso et al., (2017), *ibid*.
28. Bachiller, Y., & Busch, P. (2018). *Survey: Big data application in biomedical research*. ICCAE 2018 Proceedings of the 2018 10th International Conference on Computer and Automation Engineering.
29. Mathew & Pillai, (2015), *ibid*.
30. Mehta & Pandit, (2018), *ibid*.
31. Hemingway et al., (2018), *ibid*.
32. Mathew & Pillai, (2015), *ibid*.
33. Cyganek, B., Graña, M., Krawczyk, B., Kasprzak, A., Porwik, P., Walkowiak, K., & Woźniak, M. (2016). A survey of big data issues in electronic health record analysis. *Applied Artificial Intelligence, 30*(6), 497–520.
34. Ma et al., (2015), *ibid*.
35. Mehta & Pandit, (2018), *ibid*.
36. Alberdi, A. A., Weakley, A., Schmitter-Edgecombe, M., Cook, D. J., Aztiria, A., Basarab, A., & Barrenechea, M. (2018). Smart home-based prediction of multi-domain symptoms related to Alzheimer's Disease. *IEEE Journal of Biomedical and Health Informatics, 22*(6), 1720–1731.
37. Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., & Navab, N. (2016). AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging, 35*(5), 1313–1321.
38. Alickovic, E., & Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Computing and Applications, 28*(4), 753–763.
39. Aliwadi, S., Shandila, V., Gahlawat, T., Kalra, P., & Mehrotra, D. (2017, September). *Diagnosis of diabetic nature of a person using SVM and ANN approach*. 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (pp. 338–342), IEEE, Amity University Uttar Pradesh (AUUP), Noida, India.
40. Amirian, P., van Loggerenberg, F., Lang, T., Thomas, A., Peeling, R., Basiri, A., & Goodman, S. N. (2017). Using big data analytics to extract disease surveillance information from point of care diagnostic machines. *Pervasive and Mobile Computing, 42*, 470–486.
41. Asri, H., Mousannif, H., Al, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia—Procedia Computer Science, 83*(Fams), 1064–1069.

42. Brims, F. J., Meniawy, T. M., Duffus, I., de Fonseka, D., Segal, A., Creaney, J., & Nowak, A. K. (2016). A novel clinical prediction model for prognosis in malignant pleural mesothelioma using decision tree analysis. *Journal of Thoracic Oncology, 11*(4), 573–582.
43. Choi, J. Y., Cho, E. Y., Choi, Y. J., Lee, J. H., Jung, S. P., Cho, K. R., & Park, K. H. (2018). Incidence and risk factors for congestive heart failure in patients with early breast cancer who received anthracycline and/or trastuzumab: A big data analysis of the Korean Health Insurance Review and Assessment service database. *Breast Cancer Research and Treatment, 171*(1), 181–188.
44. Clark, A., Ng, J. Q., Morlet, N., & Semmens, J. B. (2016). Big data and ophthalmic research. *Survey of Ophthalmology, 61*(4), 443–465.
45. Daniel, B., Leff, R., & Yang, G. (2015). Views & comments big data for precision medicine. *Engineering, 1*(3), 277–279.
46. Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., & Wang, W. (2015). Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics, 31*(13), 2225–2227.
47. DuBrava, S., Mardekian, J., Sadosky, A., Bienen, E. J., Parsons, B., Hopps, M., & Markman, J. (2017). Using random forest models to identify correlates of a diabetic peripheral neuropathy diagnosis from electronic health record data. *Pain Medicine, 18*(1), 107–115.
48. Eljil, K. S., Qadah, G., & Pasquier, M. (2016). Predicting hypoglycemia in diabetic patients using time-sensitive artificial neural networks. *International Journal of Healthcare Information Systems and Informatics (IJHISI), 11*(4), 70–88.
49. Erekson, E. A., & Iglesia, C. B. (2015). Improving patient outcomes in gynecology: The role of large data registries and big data analytics. *Journal of Minimally Invasive Gynecology, 22*(7), 1124–1129.
50. Forsyth, A. W., Barzilay, R., Hughes, K. S., Lui, D., Lorenz, K. A., Enzinger, A., & Lindvall, C. (2018). Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. *Journal of Pain and Symptom Management, 55*(6), 1492–1499.
51. Gambhir, S., Malik, S. K., & Kumar, Y. (2018). The diagnosis of dengue disease: An evaluation of three machine learning approaches. *International Journal of Healthcare Information Systems and Informatics (IJHISI), 13*(3), 1–19.
52. Geerts, H., Dacks, P. A., Devanarayan, V., Haas, M., Khachaturian, Z. S., Gordon, M. F., & Brain Health Modeling Initiative. (2016). Big data to smart data in Alzheimer's disease: The brain health modeling initiative to foster actionable knowledge. *Alzheimer's & Dementia, 12*(9), 1014–1021.
53. Gurusamy, V., Kannan, S., & Nandhini, K. (2017). The real time big data processing framework advantages and limitations. *International Journal of Computer Sciences and Engineering, 5*(12), 305–312.
54. Hernandez, I., & Zhang, Y. (2017). Using predictive analytics and big data to optimize pharmaceutical outcomes. *American Journal of Health-System Pharmacy, 74*(18), 1494–1500.
55. Hossain, M. S., & Muhammad, G. (2016). Healthcare big data voice pathology assessment framework. *IEEE Access, 4*, 7806–7815.
56. Ivanova, D. (2017, December). *Big data analytics for early detection of breast cancer based on machine learning*. AIP Conference Proceedings (Vol. 1910, No. 1, p. 060016), AIP Publishing.
57. Kalyankar, G. D., Poojara, S. R., & Dharwadkar, N. V. (2017). *Predictive analysis of diabetic patient data using machine learning and Hadoop*. 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (pp. 619–624), Palladam, India.
58. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17.
59. Kovalev, V., Kalinovsky, A., & Liauchuk, V. (2016, June). *Deep learning in big image data: Histology image classification for breast cancer diagnosis.* Proceedings of 2nd International Conference Big Data and Advanced Analytics (pp. 44–53), BSUIR, Minsk.
60. Kurc, T., Qi, X., Wang, D., Wang, F., Teodoro, G., Cooper, L., & Foran, D. J. (2015). Scalable analysis of big pathology image data cohorts using efficient methods and high-performance computing strategies. *BMC Bioinformatics, 16*(1), 399.
61. Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data, 2*(1), 24.
62. Maia, A., Sammut, S., Jacinta-fernandes, A., & Chin, S. (2017). ScienceDirect big data in cancer genomics. *Current Opinion in Systems Biology, 4*, 78–84.
63. Malav, A., Kadam, K., & Kamat, P. (2017). Prediction of heart disease using K-means and artificial neural network as hybrid approach to improve accuracy. *International Journal of Engineering and Technology, 9*(4), 3081–3085.
64. Manogaran, G., Vijayakumar, V., Varatharajan, R., Malarvizhi Kumar, P., Sundarasekar, R., & Hsu, C. H. (2018, October). Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. *Wireless Personal Communications, 102*(3), 2099–2116.
65. Margolies, L. R., Pandey, G., Horowitz, E. R., & Mendelson, D. S. (2016). Breast imaging in the era of big data: Structured reporting and data mining. *American Journal of Roentgenology, 206*(2), 259–264.

66. Miranda, E., Irwansyah, E., Amelga, A. Y., Maribondang, M. M., & Salim, M. (2016). Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. *Healthcare Informatics Research, 22*(3), 196–205.
67. Morovvat, M., & Osareh, A. (2016). An ensemble of filters and wrappers for microarray data classification. *Machine Learning and Applications: An International Journal, 3*(2), 01–17.
68. Nawaz, S., Heindl, A., Koelble, K., & Yuan, Y. (2015). Beyond immune density: Critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. *Modern Pathology, 28*(6), 766–777.
69. Panayides, A. S., Pattichis, C. S., & Pattichis, M. S. (2016, November). *The promise of big data technologies and challenges for image and video analytics in healthcare*. 2016 50th Asilomar Conference on Signals, Systems and Computers (pp. 1278–1282), IEEE, Pacific Grove, CA.
70. Prasad, S. T., Sangavi, S., Deepa, A., Sairabanu, F., & Ragasudha, R. (2017). *Diabetic data analysis in big data with predictive method*. International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET) (pp. 1–4), Chennai, India.
71. Prerana, T. H. M., Shivaprakash, N. C., & Swetha, N. (2015). Prediction of heart disease using machine learning algorithms-Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS. *International Journal of Science and Engineering, 3*, 90–99.
72. Priyanga, P., & Naveen, N. C. (2018). Analysis of machine learning algorithms in health care to predict heart disease. *International Journal of Healthcare Information Systems and Informatics (IJHISI), 13*(4), 82–97.
73. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems, 2*(1), 3.
74. Rahim, A., Forkan, M., Khalil, I., & Atiquzzaman, M. (2017). ViSiBiD : A learning model for early discovery and real-time prediction of severe clinical events using vital signs as big data. *Computer Networks, 113*, 244–257.
75. Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P., & Li, G. (2017). An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. *Expert Systems with Applications, 68*, 163–172.
76. Saravana, N. M., Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive methodology for diabetic data analysis in big data. *Procedia—Procedia Computer Science, 50*, 203–208.
77. Schatz, B. R. (2015). National Surveys of population health: Big data analytics for mobile health monitors. *Big Data, 3*(4), 219–229.
78. Shah, M., Wang, D., Rubadue, C., Suster, D., & Beck, A. (2017, November). *Deep learning assessment of tumor proliferation in breast cancer histological images*. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 600–603), IEEE, Kansas City, MO.
79. Shen, L., Chen, H., Yu, Z., Kang, W., Zhang, B., Li, H., & Liu, D. (2016). Evolving support vector machines using fruit fly optimization for medical data classification. *Knowledge-Based Systems, 96*, 61–75.
80. Sherri, L., & Zhangxi, C. (2016). Accepted Manuscript, 0–67. Morovvat, M. & Osareh, A. (2016). An ensemble of filters and wrappers for microarray data classification. *Machine Learning and Applications: An International Journal, 3*(2), 01–17.
81. Silva, L. F., Santos, A. A. S., Bravo, R. S., Silva, A. C., Muchaluat-Saade, D. C., & Conci, A. (2016). Hybrid analysis for indicating patients with breast cancer using temperature time series. *Computer Methods and Programs in Biomedicine, 130*, 142–141.
82. Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. *Journal of Big Data, 2*(1), 8.
83. Sivakami, K., & Saraswathi, N. (2015). Mining big data: Breast cancer prediction using DT-SVM hybrid model. *International Journal of Scientific Engineering and Applied Science (IJSEAS), 1*(5), 418–429.
84. Su, Q., Wang, Y., Jiang, X., Chen, F., & Lu, W. C. (2017). A cancer gene selection algorithm based on the K-S test and CFS. *BioMed Research International, 2017*, 1–7.
85. Sumbaly, R., Vishnusri, N., & Jeyalatha, S. (2014). Diagnosis of breast cancer using decision tree data mining technique. *International Journal of Computer Applications, 98*(10), 16–24.
86. Varatharajan, R., Gunasekaran, M., Priyan, M. K., & Sundarasekar, R. (2018, March). Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm. *Cluster Computing, 21*(1), 681–690.
87. Sung, K. C., Ting, H. M., Chao, P. J., Guo, S. S., Tran, C. K., Huang, Y. J., & Lee, T. F. (2016). Predicting grade 2 acute radiation-induced dermatitis after hybrid intensity modulation radiotherapy for breast cancer using a logistic regression normal tissue complication probability model. *European Journal of Cancer, 60*, e4.
88. Taglang, G., & Jackson, D. B. (2016). Gynecologic oncology use of "big data" in drug discovery and clinical trials. *Gynecologic Oncology, 141*(1), 17–23.
89. Turgut, M., Turgut, A.T., & Kosar, U. (2006, October). Spinal brucellosis: Turkish experience based on 452 cases published during the last century. *Acta Neurochirurgica, 148*(10), 1033–1044.
90. Varatharajan, R., Manogaran, G., Priyan, M. K., & Sundarasekar, R. (2017). Wearable sensor devices for

early detection of Alzheimer disease using dynamic time warping algorithm. *Cluster Computing*, 1–10.
91. Viceconti, M., Hunter, P. J., & Hose, R. D. (2015). Big data, big knowledge: Big data for personalized healthcare. *IEEE Journal of Biomedical and Health Informatics, 19*(4), 1209–1215.
92. Volynskaya, Z., Chow, H., Evans, A., Wolff, A., Lagmay-Traya, C., & Asa, S. L. (2017). Integrated pathology informatics enables high-quality personalized and precision medicine: Digital pathology and beyond. *Archives of Pathology & Laboratory Medicine, 142*(3), 369–382.
93. Walczak, S., & Okuboyejo, S. R. (2017). An artificial neural network classification of prescription nonadherence. *International Journal of Healthcare Information Systems and Informatics (IJHISI), 12*(1), 1–13.
94. Wang, Y., & Hajli, N. (2017). Exploring the path to big data analytics success in healthcare. *Journal of Business Research, 70*, 287–299.
95. Wu, D., Jennings, C., Terpenny, J., & Kumara, S. (2016). *Cloud-based machine learning for predictive analytics: Tool wear prediction in milling*. Proceedings—2016 IEEE International Conference on Big Data (pp. 2062–2029), Big Data, Washington, DC.
96. Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine, 153*, 1–9.
97. Zheng, T., & Zhang, Y. (2017, August). *A big data application of machine learning-based framework to identify type 2 diabetes through electronic health records*. International Conference on Knowledge Management in Organizations, Beijing, China.
98. Remadna, I., Terrissa, S. L., Zemouri, R., & Ayad, S. (2018, March). *An overview on the deep learning based prognostic*. 2018 International Conference on Advanced Systems and Electric Technologies (IC_ASET) (pp. 196–200), IEEE, Hammamet, Tunisia.

Biographies

Abir Belaala is presently a PhD student in computer science, with a specialty in Artificial Intelligence. She received a master's degree in Computer Science in 2015 from Biskra University, Algeria. Her current research interest is Big Data Analytics and Machine Learning in the medical field.

Labib Sadek Terrissa is an Associate Professor in Computer Science at Biskra University, Algeria. He is the intelligent systems and networking team head within the smart computer science Laboratory (LINFI), where he conducts his research activities. After receiving an engineering degree in electronics, he received a postgraduate degree (DEA) and a PhD in computer engineering in 2006 from LeHavre University, France. He received the first award in the national exhibition of research and development in 2017 and the best paper award in IEEE-Cist's 2016 conference. His current research interests include Cloud Computing, Cloud Robotics, Machine learning, Medical Big Data, Smart maintenance, and Prognostic and Health Management.

Zerhouni Noureddine is a full professor at École Nationale Supérieure de Mécanique et des Microtechniques. He is a member of PHM team of Automatic Control and Micro-Mechatronic Systems department within FEMTO-ST Institute. He has worked since 1999 on modeling, analysis, and control of production systems. His specializations include system modeling, artificial intelligence techniques for diagnostic and prognostics, and machine learning.

Devalland Christine is head of the Department of Pathology, specializing in breast pathology. Her current research interest is the indication of neural network in pathology.