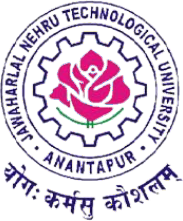
**THE ROLE OF BIG DATA ANALYTICS IN PHARMACEUTICAL SCIENCE**

**Practice School submitted to the**

**JAWAHARLALNEHRU TECHNOLOGICAL UNIVERSITY,** **ANANTAPUR, ANDHRA PRADESH**



**SUBMITTED BY**

## R.JESWANTH

**21ER1R0091**

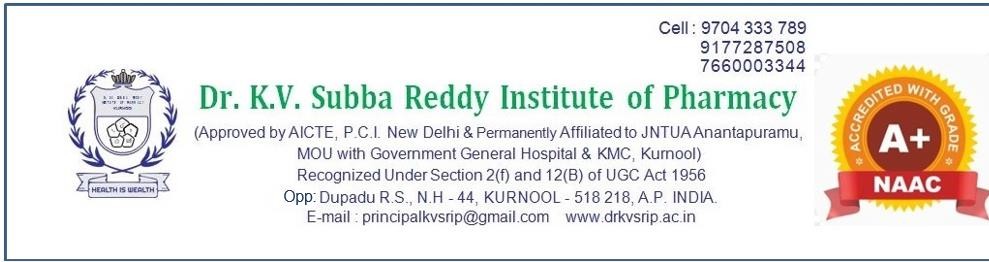
###### Under the Supervision of

**DORNALA CHAITHANYA DIXIT** M. Pharm (Ph. d)

Associate Professor Department of Pharmaceutical Analysis

# Dr. K. V. Subba Reddy Institute of Pharmacy, Dupadu, Kurnool–518218, A.P

## NOVEMBER -2024



**CERTIFICATE BY SUPERVISOR**

This is to certify that the work contained in the Practice School Report **“Big Data Analytics in pharmaceutical science Report ”**, submitted by R.JESWANTH **Regd.No:21ER1R0091** to the **Dr. K. V. Subba Reddy Institute of Pharmacy**, is a record of bonafide practice work carried out by him under my direct supervision and guidance.

I considered that he/She work has reached the standards and fulfilling the requirements of the rules and regulations relating to the nature of the Practice School Report.

**Date:**

**Place:**

**Signature of Supervisor**

## CERTIFICATE BY HEAD OF THE INSTITUTE

This is to certify that the Practice school Report **“Big Data analytics in pharmaceutical science Report ”** was done for the partial fulfillment of B.Pharm IV Year I sem and has been carried out by **R.JESWANTH Regd. No.: 21ER1R009** under the guidance and supervision of MR.DORNALA **CHAITHANYA DIXIT (PHARMACEUTICAL**

**ANALYSIS)** at **Dr. K. V. Subba Reddy Institute of Pharmacy**, Dupadu, Kurnool, during the period of 2024-2025.

Date:

Place:

Signature of the Principal:

## ACKNOWLEDGEMENTS

I would like to extend my heartfelt gratitude to all those who have contributed to the completion of this review article. I am immensely grateful to Mr. DORNALA CHAITANYA DIXIT M. Pharm (Ph.D.) for his guidance, invaluable insights, and constructive feedback throughout the review and writing process. His expertise and encouragement have been crucial in shaping this work.

I am especially grateful for his patience and willingness to guide me at each step, whether it was in the formulation of ideas, selection of relevant literature, or refining the final drafts. Mr. Chaitanya Dixit has always been generous with his time, offering guidance and support despite his numerous commitments, and his unwavering belief in my potential gave me the confidence to pursue this work diligently.

His profound knowledge in [pharmaceutical analysis] and his dedication to advancing this field have inspired and motivated me throughout the process. His keen insights and constructive suggestions helped shape this work and allowed me to gain a deeper understanding of the nuances and complexities involved.

His mentorship extended beyond the academic realm, instilling in me a sense of professionalism, ethical responsibility, and curiosity that will serve as guiding principles in my future endeavours. This article would not have been possible without his invaluable assistance, and I am truly honored to have had the privilege of working under his guidance.

I would also like to express my appreciation to Dr. K.V. SUBBA REDDY INSTITUTE of pharmacy for providing the facilities, resources, and intellectual environment that made this study possible. Special thanks to Dr. B V Ramana, Principal, for his support, thoughtful discussions, and technical assistance that have greatly enriched the depth and scope of this review.

I am deeply thankful for his encouragement to think critically, his commitment to excellence, and the countless hours he dedicated to helping me achieve the highest standard in this review. For all of this and more, I express my sincere appreciation and look forward to applying the insights and skills I have gained through this experience.

Furthermore, I acknowledge the contributions of authors of the studies reviewed in this article, whose pioneering work has paved the way for our understanding of [THE ROLE OF BIG DATA ANALYTICS IN PHARMACEUTICAL SCIENCE]. Their dedication and

innovation are truly inspiring and have provided a solid foundation for this comprehensive analysis.

R.JESWANTH

## LIST OF ABBREVATIONS

|  |  |
| --- | --- |
| **ABBREVITIONS** | **DESCRIPTION** |
| BDA | Big Data Analytics |
| IDC | International Data Corporation |
| WHO | World Health Organization |
| OMR | Optical Mark Recognition |
| DBMS | Data Based Management System |
| ACOS | Accountable Care Organizations |
| HER | Electronic Health Record |
| DBSI | Drug Based Similarity Interference |
| TBSC | Target Based Similarity Interference |
| NBI | Network Based Interference |
| SVM | Support Vector Machine |
| LMS | Learning Management System |
| BW | Business Ware Housing |
| MK | Microsoft Kinetics |

|  |  |
| --- | --- |
| **Figure** | **Page No** |
| Figure:1:Big Data Source | 02 |
| Figure:2:The Drug Development Process | 04 |
| Figure:3:Network Representation Example of Drugs, Targets and Diseases. | 08 |
| Figure:4:Overview of Yamanishi method | 11 |
| Figure:5: Predicting New Drug-Target Interactions Based on Grug-drug and Target-Target Similaritie | 12 |
| Figure:6:Overview of the Cross Validation Framework | 16 |
| Figure:7: ***.*** Overview of Big Data Analysis in Animal Science Using\ Machine Learning and Data Mining Tools. | 17 |

|  |  |
| --- | --- |
| **TablE** | **Page No** |
| Table:1.1 Data Bases Containing Chemical and Biological Source | 07 |

|  |  |  |
| --- | --- | --- |
|  | CONTENT | PAGE NO |
| 1. | Introduction | 01 |
| 2. | History   * 1. Pre-marketing stage   2. Post-marketing stage   3. Data sources and other applications   4. Chemical and Biological Data   5. Constructing a network representation | 02-07 |
| 3. | Evolution of Big Data Technology | 08 |
| 4. | Big Data and Health Care   * 1. Single similarity based methods | 09-11 |
| 5. | Big Data in Pharmacy Education | 12 |
| 6. | Policy Implication in the Era of Big  Data | 13-14 |
| 7. | Machine Learning Framework   * 1. *Genomic Prediction*   2. Phenotype Fraud Detection   3. Mastitis Detection | 15-21 |
| 8. | Big data analytics and Internet of Things | 22 |
| 9. | Future Research Directions | 23 |
| 10. | Healthcare Stakeholders & Big Data Sources   * 1. Hospital   2. Patient   3. Medical Practioner | 24-35 |
| 11. | Conclusion | 36 |
| 12. | Biplography | 37-40 |

**THE ANALYTICS IN PHARMACEUTICAL SCIENCE ROLE OF BIG DATA**

## ABSTRACT:

The pharmaceutical industry is critical in modern healthcare, focusing on the discovery, development, and delivery of medications. However, it faces significant challenges, such as the high costs of clinical research, long development timelines, and high failure rates of drugs. The advent of Big Data analytics in the 21st century has been transformative in addressing these issues, optimizing drug discovery, development, clinical trials, personalized medicine, pharmacovigilance, and commercial strategies.

Big Data leverages advanced tools like predictive analytics, machine learning, and AI to enhance target identification, accelerate clinical trials, develop tailored therapies, and ensure drug safety. Nonetheless, challenges such as data privacy, security, interoperability, cost, and quality management persist, requiring careful navigation to fully harness Big Data's potential in pharmaceuticals. This essay explores the multifaceted applications of Big Data analytics in the pharmaceutical industry and discusses the challenges and limitations of its implementation.

Keywords: Bid data analytics, drug discovery, clinical trails, personalized medicine, pharmaco vigilance.

**Role of Big Data Analytics in Pharmaceutical Science:**

### Introduction:

Inter disciplinary computational approaches that combine statistics, computer science, medicine, and biology are becoming highly valuable for drug discovery and development. Data mining and machine learning methods are being more commonly used to properly analyze the emerging high volumes of structured and unstructured biomedical and biological data from several sources including hospitals, laboratories, pharmaceutical companies, and even social media. These data may include sequencing and gene expression, drug molecular structures, protein and drug interaction networks, clinical trial and electronic patient records, patient behavior,and self- reporting data in social media, regulatory monitoring data, and biomedical literature. Data mining methods can be used in several stages of drug discovery and development to achieve different goals.

Most new compounds fail during this approval process in clinical trials or cause adverse side effects. The cost of successful novel chemistry-based drug development often reaches millions of dollars, and the time to introduce the drug to market often comes close to a decade [1]. The high failure rate of drugs during this process, make the trial phases known as the “valley of death” [2]. Similar to many other domains, pharmaceutical data mining algorithms aim to limit the search space and provide recommendations to domain experts for hypothesis generation and further analysis and experiments. One way to categorize data mining and machine learning approaches is based on their application to pre-marketing and post-marketing stages. In pre-marketing stage, data mining methods focus on discovery activities, including but not limited to, finding signals that indicate relations between drugs and targets, drugs and drugs, genes and diseases, protein and diseases, and finding bio-markers. In this stage potential interactions that could cause therapeutic or adverse effects are studied. Most of the chemical compounds under study at this stage have not been through clinical trails and the *in silico* experiments serve as a basis for further explorations for them. In the post marketing stage an important application of data analytics is in finding indications of adverse side effects for approved drugs. These algorithms provide a list of potential drug side effect associations that can be used for further studies.

In this chapter we provide a brief overview of some data analytics applications in this domain, and mainly focus on two major tasks from each stage.We first summarize some of the main

methods for drug-target interaction prediction that is highly important during the pre-marketing stage. We then provide an overview of *pharmacovigilance* (or drug safety surveillance) which is an important focus in the post-marketing stage.

### History of big data analytics:

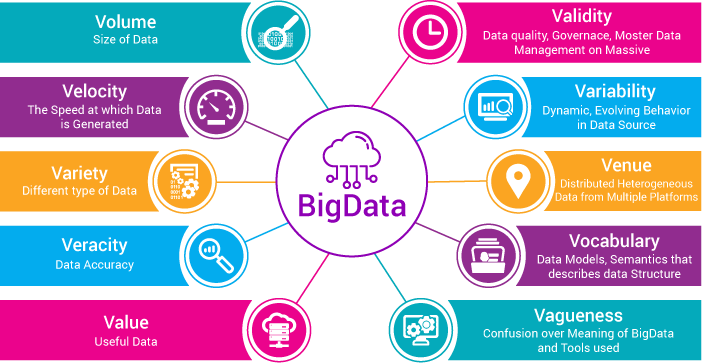
The advent of big data analytics was a response to the rise of big data that started in the 1990s. Very Long before the term “big data” was coined, the concept was applied to the dawn of the computer age when businesses used large spreadsheets to crunch numbers and find trends.

Figure:1 Big Data Source

The large amount of data created in the late 1990s and early 2000s was fueled by new data sources. The popularity of mobile devices and search engines created more data than any company knew what to do with. Speed was another factor. The faster the data was made, the more it had to be handled. In 2005, Gartner explained that these are the “3 Vs.” of big data – variety, volume, and velocity. Recent research by IDC projected that data generation would grow tenfold worldwide by 2020.Anyone who could tame the vast amount of raw, unstructured information would open up a treasure chest of never-before-seen consumer behavior, business operations, natural phenomena, and population change. Traditional data

warehouses and relational databases were not up to the task. Innovation was needed. In 2006, Hadoop was created by engineers from Yahoo and launched as an open-source Apache project. The distributed processing platform made it possible to run big data applications on a clustered platform. This is the main difference between traditional and big data analytics. At first, big companies like Google and Face book used big data analytics. In 2010, retailers, banks, manufacturers, and healthcare companies began to understand the value of being big data analytics companies as well. Initially, large organizations with on- premises data systems were best suited to collect and analyze large data sets. But Amazon Web Services and other cloud platform vendors have made it easy for businesses to use big data analytics services. The ability to handle Hadoop clusters in the cloud has given any size company the freedom to spin up and run only what they need on demand. The big data analytics ecosystem is a key component of the agility required for today’s companies to succeed. Insights can be discovered more quickly and efficiently, translating into instant trading decisions that can decide a winner.

###### Pre-marketing stage:

In the pre-marketing stage, data mining algorithms primarily focus on drug discovery and predicting potential adverse effects using characteristics of the compounds (e.g., drug targets, chemical structure) or screening data (e.g., bioassay data) [3]. One of the important challenges where data mining and machine learning methods could be very beneficial is drug-target interaction prediction. This task is also highly important for drug repurposing and drug adverse reactions prediction. [4] *In vitro* identification of drug-target associations is a labor-intensive and costly procedure. Hence, *in silico* prediction methods are promising approaches for focusing *in vitro* investigations [5].

Most drugs affect multiple targets, and *Poly pharmacology*, the study of such interactions, is an area of growing interest [6]. These multi-target interactions potentially result in adverse side effects or unintentional therapeutic effects, and is the main cause in the high failure rate of drug in clinical trials. Unacceptable toxicities resulting from these interactions account for approximately 30% of the failures [7]. Predicting these interactions during the drug developmental phase can reduce the high cost of clinical trials and can be crucial for the commercial success of new drugs.

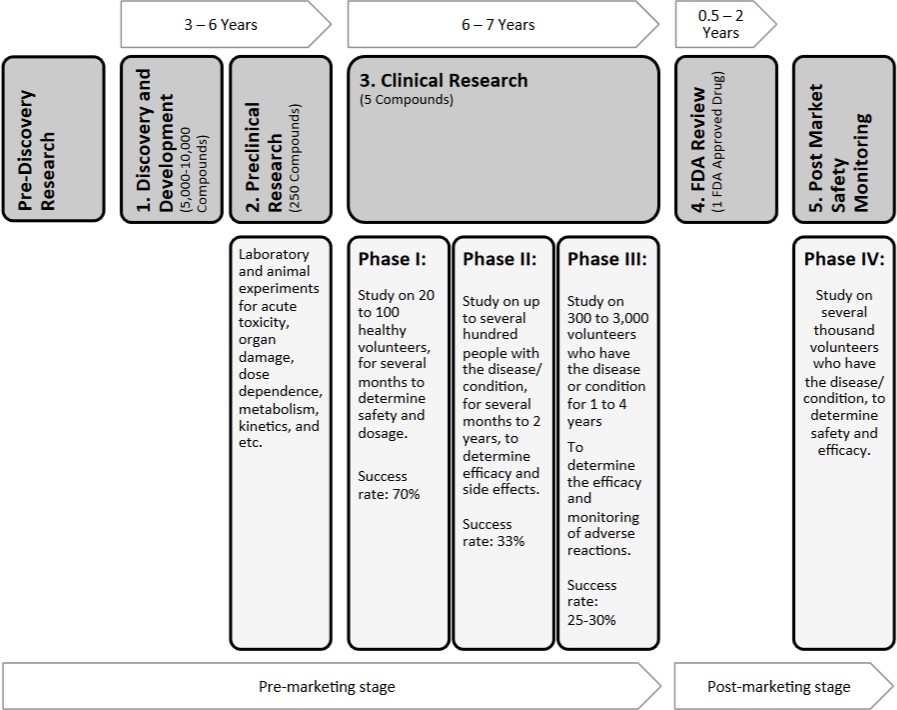


FIGURE:2 The drug development process

**Post-marketing stage:**

In the post-marketing stage an important focus of data mining methods is on finding patterns that indicate potential drug related adverse events [3]. Undiscovered severe adverse events may lead to drug withdrawals which can be financially detrimental for the manufacturers [3]. Several drugs havebeen withdrawn from the market over the years [10]. For example, *Vioxx*, which was considered a powerful anti-inflammatory drug was withdrawn due to an increased coronary risk [11,12]. Each year more than two million hospitalizations and injuries, and seven hundred thousand emergency visits in the United States have been estimated to be caused by these effects

[13, 3, 14]. They have also been estimated to cost seventy five billion dollars annually [9]. It is also estimated that each year 6–7% of hospitalized patients experience severe adverse drug related events, which can lead to a potential hundred thousand deaths, making it the fourth largest cause of death in the U.S. [14]. Since only a limited number of patient characteristics are studied in clinical trials and for a limited duration, often complex safety issues associated with a new drug cannot be fully studied with clinical trials [8]. Adverse drug effects are often defined as the following [15]: Any unintended and undesirable effects of a drug beyond its anticipated therapeutic effects occurring during clinical use.”*Pharmacovigilance* (or drug safety surveillance) is the science that concerns with the detection, assessment, understanding and prevention of adverse drug reactions [3]. Pharmacovigilance is formally defined by World Health Organization (WHO). [16]

The science and activities related to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems.” Data analysis algorithms are crucial to narrow the search space and detect the hidden patterns. Harpaz et al. [9] define data mining algorithms for Pharmacovigilance as: “Automated high-throughput methods to uncover hidden relationships of potential clinical significance to drug safety.” They report that volume of publications on data mining methods for Pharmacovigilance index in PubMed3 has grown from less than 40 in the year 2000 to about 200 a year in 2011. An important focus of data mining algorithms in post-marketing stage is on computing measures of statistical association between pairs of drugs and clinical outcomes recorded in underlying data sources.[17].

###### Data sources and other applications:

There are several important data mining applications that we do not address in this chapter. For example, another area of growing interest where data mining algorithms play a significant role is predicting individual drug responses and personalized medicine [18, 19]. Personalized medicine or *Pharmacogenomics*, is using an individual’s genetic profile to make the best therapeutic choice by facilitating predictions about whether that person will benefit from a particular medicine or will suffer serious side effects [20]. For example, *Pharmacogenomics Knowledgebase (PharmGKB)* is a resource that collects, curates, and disseminates information about the impact of human genetic variation on drug responses [21].

Data mining algorithms in different stages of drug development use different data sources.

Chemical and biological data are mainly used in pre-marketing stage for tasks such as hypothesis generation and prediction, while spontaneous reporting systems, electronic health records and administrative claims data are often used in post-marketing data mining tasks mainly to detect signals of association. Biomedical literature and patient-generated data in health-related Internet forums has also received considerable research interest in recent years [9].

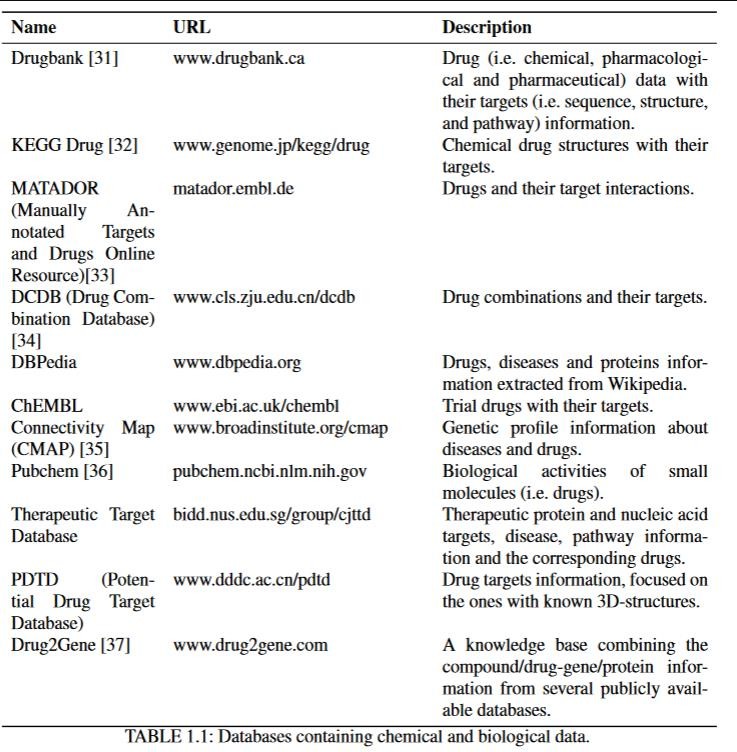
In the rest of the chapter, we highlight some of the related data mining tasks and methods based on the data resource they are applied to. First, we summarize some of the methods that use chemical and biological data focusing on approaches that predict drug-target and drug-drug interactions. We then highlight methods that detect patterns of drug related adverse events using spontaneous reports, electronic health records and patient generated data such as web search engine logs. We also mention some of the advances in application of data mining in biomedical literature that can facilitate pharmaceutical discoveries.

There is a plethora of high quality research recently published related to data analytics in pharmaceutical discoveries which we could not cover in this chapter. We did not aim to provide a complete or comprehensive survey; our goal was to provide highlights of some of the important data analytics methods in this domain.

###### Chemical and Biological Data:

It is one of the important goals of data mining methods that use chemical and biological data is predicting interactions between chemical compounds (e.g., drugs) and biological targets (e.g., proteins) which could cause therapeutic or adverse effects, or interactions between two or more chemical compounds that could cause potential adverse effects. Openly available databases, including multiple resources available on the Internet that include drug related data and information about their targets are highly used for this task. These databases are used to study properties of drugs for several purposes, including drug-target and drug-drug interaction elucidation. Table 1.1 summarizes some of the more commonly used databases that contain information about drugs, their targets and interactions between them.

There are several methods to model the drug-target interaction prediction task [29]. They can be separated into two categories based on their explicit emphasis on the graph or network representation of drugs and targets interactions. The first category constructs a network structure to predict interactions [30], while others make predictions based on other factors. In this section, we mainly cover network based approaches.



###### Constructing a network representation:

A number of research publications study network structures to predict interactions. Cockell

et al. [8] described how to integrate drugs, targets, genes, proteins, and pathways into a network for different tasks. Nodes in this network usually include drugs, proteins and diseases, and edges include their interactions and similarities, where similarities could be extracted from several sources such as chemical structure of the compounds [3]. Figure 1.2 shows an example of a schematic overview of such networks. Lee et al. [31] described drug repurposing, multi-agent drug development, and estimation of drug effects on target perturbations via network-based solutions. Yildirim et al. [30] explained trends in the drug-discovery industry over time using a

network-based analysis and showed that sequencing the genome is changing the traditional trends of drug development. They also discussed different structural aspects of this network including preferential attachment and cluster formation.

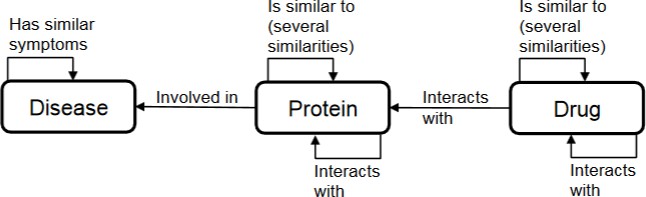


Figure:3 Network representation example of drugs, targets and diseases.

# Evolution of Big Data Technology:

From the ancient times human knows how to store and process data, the initial storage being on stones of caves and tamarapatra (leaves), etc., which was referenced by next generations to interpret the previous generation’s wisdom and heritage. After invention of Paper around 220AD the data was stored in bunch of papers, which was nothing but a book. However, the ability to analyse the data is limited to one’s own brain and knowledge inferred by an individual, which was stepping stone for inventing an alternate for human brain that is computer evolution. This era began with punch cards and OMR for data storage.

Since the computer became integral part of every business there was a need to communicate among computers in many geographical locations, which led to invention of Internet, which triggered massive improvement in data storage, processing and analysing techniques that started digital electronics era. In order to manage and process the data, files systems and database management systems have been developed. File systems enable to store and access hard disk and DBMS enables to store and access data. Data generation rate has been tremendously increased

from last two decades and it is the origin for new horizons in data storing and processing requirements for wide variety, huge volumes of data from kilobytes to zeta bytes.

## BIG DATA AND HEALTH CARE:

Big data” has been defined as “the linking of disparate large data sets to provide insight at the individual level.”6 In the area of population health, big data is not new. However, the sources are multiplied and often nontraditional, and the volume of data generated is huge compared to traditional methods .The Institute of Medicine has promoted the concept of a “learning healthcare system” in which “each patient care experience naturally reflects the best available evidence from basic, translational, comparative effectiveness, and health service research, and, in turn, adds seamlessly to learning what works best in different circumstances.”7 It is estimated that in 2005, approximately 75,000 deaths could have been prevented if all states delivered the same quality of care as the highest performing state. In addition, inefficiency in the healthcare system due to unneeded services, fraud, administrative costs, and other issues cost in excess of

$750 billion in 2009, about 30% of health spending that year.8 Healthcare leads many industries in recognizing and utilizing scientific inquiry over expert opinion and testimonials to guide rational decision-making. However, the evidence-based techniques used in healthcare lag behind many industries in utilizing information technology and analytic techniques to optimally utilize big data to guide personal and population health decisions. The torrent of health-related information, from genomics and payer provider data to web-based social sites and mobile wearable data sources, must be responsibly gathered, analyzed, and utilized. The rich information contained in EHRs has the potential to improve quality and efficiency in healthcare. Elements of the Affordable Care Act that provide incentives to Accountable Care Organizations (ACOs) to collect, analyze, and exchange health care information may accelerate this process. The information provided by EHRs can be combined with other sources to transform medical practice and improve quality of care. Murdoch and Detsky9 describe four ways in which big data may improve quality and efficiency in health care Big data increases the capacity to generate new knowledge by using computational techniques to analyze unstructured data within EHRs including natural language processing to extract medical concepts from free-text. These techniques can create an observational evidence base for clinical questions that was not previously possible. Lin et al10 have identified clinically relevant, accurate symptom-disease- treatment associations for seven diseases from cancers to chronic and infectious diseases from

mining of approximately 2.1 million EHRs. Miller11 has utilized health-related social media sites such as Patients Like Me to research healthcare decision support and patient empowerment for chronic diseases.

1. Personalized medicine initiatives may be integrated into clinical practice by analytical methods that integrate genomics information with EHR data. The Electronic Medical Records and Genomics Network uses natural language processing to phenotype patients.
2. Knowledge dissemination can be optimized. Physicians struggle to stay current with the latest evidence guiding clinical practice due to the volume of information.
3. Big data analytics may allow delivery of information directly to patients, thereby encouraging a more active role in their healthcare. In the future, medical records may reside with patients rather than healthcare providers. Big data could link traditional health-related data found in the EHR to personal data from other sites without the need to gather the information through patient interview.

### Single similarity based methods:

Network-based approaches integrate drug-drug and target-target similarities extracted via different methods (e.g. SEA and CMap) with the drug-target interactions network [32]. The following methods proposed a single similarity measure for drugs and targets to predict interactions.Cheng et al. [33] predicted potential interactions using drug-drug and target-target similarities and a bipartite interaction graph. Using SIMCOMP [34], they computed the 2D chemical drug similarities and sequence similarities for targets via the Smith-Waterman score. They used the following.

three link-prediction methods:

1. Drug-based similarity inference (DBSI) where they only considered similarities between drugs for prediction.
2. Target-based similarity inference (TBSI) where they only considered target similarities for prediction.
3. Network-based inference (NBI) where they combined both drug-drug and targettarget similarities.

Alaimo et al. [35] extended this approach by proposing a hybrid drug-target method that integrated prior domain knowledge.Yamanishi et al. [36] proposed three methods for interaction prediction, including a nearest neighbor approach, a weighted *k*-nearest neighbors approach, and

a space integration. In their space integration method, they described a genomic space, using the Smith-Waterman score, and a pharmaceutical space, using the SIMCOMP score. They proposed a method to integrate drugs and targets in a unified latent *pharmacological space*, and they predicted interactions based on the proximity of drugs and targets in that space. Figure 1.3 shows an overview of their method. They separated out four categories of targets, namely enzymes, ion channels, GPCR, and nuclear receptors for their experiments which was adopted by most subsequent drug-target interaction prediction methods [29].

Overall steps in their method include:

1. Embed compounds and proteins on the interaction network into a unified space called “pharmacological space”.
2. Learn a model between the chemical/genomic space and the pharmacological space, and map any compounds/proteins onto the pharmacological space.
3. Predict interacting compound–protein pairs by connecting compounds and proteins which are closer than a threshold in the pharmacological space.

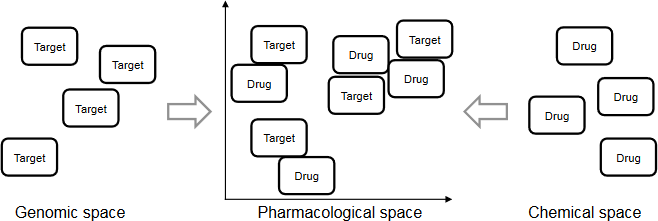


Figure:4 Overview of Yamanishi [50] method.

Bleakley and Yamanishi [37] extended this method by constructing local models for graph infer- ence. They classified each interaction twice and combined the results to provide predictions. First, they built a classifier based on drugs and then based on targets. They used the similarities as the *support vector machine* (SVM) kernels. Further extending this method, Mei et al. [38] proposed an approach to infer training data from neighbors’ interaction profiles to make predictions for new drug or target candidates that do not have any interactions in the network. Wang and Zeng [39] proposed a method based on restricted Boltzmann machines.

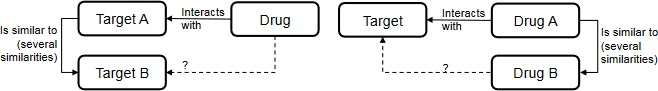


Figure:5 Predicting new drug-target interactions based on drug-drug and target-target similaritie

## BIG DATA IN PHARMACY EDUCATION:

The Bill and Melinda Gates Foundation published its Postsecondary Success advocacy plan. The first priority identified was educational data and information. Their recommendation is “(c)reate a national data infrastructure that enables consistent collection and reporting of key performance metrics for all students in all institutions that are essential to promoting the change needed to reform the higher education system.”31 The magnitude of the data sets used in education tend to be much smaller than those described previously for health care and research, yet educational datasets may well represent big data. The 2013-14 AACP Academic Affairs Committee32 produced a good overview of the use of big data and analytics in pharmacy education, specifically delineating the differences between “academic analytics ”and “learning analytics.” Academic analytics are those that help administration, typically at the full university level or higher (e.g., consortial level) to identify effectiveness of units and/or processes. These analyses may help make decisions on resource allocation, changing programs to better meet their desired goals, etc. Fain33 described an effort currently underway at the consortial level (the University Innovation Alliance) where innovations may be assessed and shared. This group is also looking at pooling the same data from the member schools to identify trends or gaps in data. Colleges/schools often use academic analytics to look at their student populations to identify admissions criteria from predictors of programmatic success as well as to identify predictors of success within their student bodies. This latter area often transitions to “learning analytics” when the results are used to aid the success of an individual student. Learning analytics is the “micro” equivalent.

Learning analytics is the “micro” equivalent of the “macro” academic analytics focusing on assessing factors influencing an individual student’s performance or learning. Kolowich34

reported on several learning management systems (LMS) that collect data(e.g.participation in various on-line activities, access of online resources, activity during lecture). Analyzing these LMS data against student performance allows the identification of certain predictors individual success in a course. Programs cited were Student Explorer\_, ECoach\_, and Lecture Tools\_.Online testing software American Journal of Pharmaceutical Education (40]. can capture student performance on assessments and match these to desired course and programmatic outcomes .Most of these software packages also have tools to increase faculty efficiency and ease with identifying and communicating with students exhibiting behaviors that affect their performance. Often, these analytics can: occur in real time (e.g., classroom programs showing student visits to course site tools); produce both individual student andaggregated data; indicate student progress or lack thereof over time; identify the preferred learning habits/styles of individual students or a class; and assist in identifying at-risk students through more than grades and before it is too late to recover.

## POLICY IMPLICATIONS IN THE ERA OF BIG DATA:

Clearly, colleges and schools of pharmacy and their faculty have found and will continue to find that their work both creates and utilizes big data of numerous types from many sources. The ability access, analysis, and apply such data to our work is essential. The Argus Commission felt it was important to co inside policy implications related to their charge and accessed a report to President Obama from his Council Advisors on Science and Technology. “Big Data and Privacy: A Technological Perspective”39 offers severe recommendations germane to academic pharmacy. Certainly the collection and analysis of data have the potential to impact the privacy of individuals and groups in ways unimaginable at American Journal of Pharmaceutical Education [41]. times when current laws and regulations were promulgated. It is essential that those engaged in the generation and use of big data remain knowledgeable of applicable and emerging laws and regulations and contribute their insights into policies and procedures that will insure the protection of our students, our faculty, our patients, and those who are involved with our research and assessment activities. As stated in the recommendations to the President, policies at every level should focus more on the actual uses of big data and intended outcomes of that use rather than on data collection, storage and analysis, or specific technological solutions.

##### Recommendation 1:

AACP should assist member colleges and schools with the identification and use of relevant data analytic technologies, facilitate the development of consortia to advance collaboration, and serve as a secure repository for de-identified data that could be used to identify innovative practices in teaching, learning and assessment.

##### Recommendation 2:

AACP should AA provide guidance to member colleges and schools to assure that any student’s, employee’s, or other individual’s information included in big data reporting and analyses be limited to that information that can be legally and ethically divulged.

##### Recommendation 3:

AACP and member colleges and schools, in partnership with individuals with big data expertise, should assess methods of using social media and other potential marketing tools to identify, engage, recruit, and retain students to programs offered by the pharmacy academy and provide the academy with resources to apply such processes.

##### Recommendation 4:

AACP and member colleges, in partnership with individuals with big data expertise, should develop data-driven systems and processes to guide member colleges in areas such as the recruitment and selection of the most appropriate candidates for admission based on pooled student performance data; assessing and modifying pre-pharmacy and pharmacy curricula; and directing student professionalism development and career advising for pharmacy students.

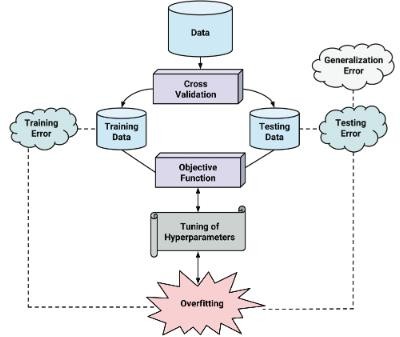
##### Recommendation 5:

AACP and member colleges and schools should facilitate the use of data analytics to increase the health of individuals and communities through education, research, and service, and such efforts should contribute to the promotion and tenure process for faculty and professional staff. New Recommendation 6: AACP should lead the planning and development of an interprofessional summit of health professions educators to fully explore the impact of big data on education, research, and practice.

## MACHINE LEARNING FRAMEWORK:

Machine learning, also known as statistical learning, is a subfield of artificial intelligent dedicated to the study of algorithms for prediction and inference. Learning from data is at the core of machine learning. Data mining shares a similar spirit with machine learning and is often discussed in the same context. If we are more stringent in definition, data mining encompasses the study of database systems, which becomes crucial in dealing with extremely large datasets. In most practical cases machine learning ultimately aims to learn, or choose from, a pool of candidate probability models that can best predict unobserved data. Technically, the selection is called the “training process.” However, how can we measure the prediction ability of the selected function? Suppose, for example, that our task is to predict a phenotype of an animal from a set of genotypes and that we have a dataset consisting of pairs of phenotypes and corresponding genotypes. In machine learning, this type of task is called supervised learning, with the target of prediction (phenotype) referred to as the supervisory signal. If the phenotypes If the phenotypes are quantitative, it is known as a regression task. In contrast, when the dataset is incomplete and only genotypes are available for the selected individuals (no phenotypes), the task is called unsupervised learning. To choose a probability model with good prediction ability in supervised learning, we begin by splitting the dataset into two sets, a training and a testing dataset, where the latter of which playing the role of the dataset that are not available to us at the moment. When we select a probability model, we use the information from the training dataset exclusively. In particular, we construct an objective function based exclusively on the training dataset to represent the user’s choice of desirable properties for the function. We then choose from the pool of probability models the one that maximizes the objective function. One naive property used in this specific example is the likelihood of the probability model observing phenotypes in the training dataset given the corresponding genotypes in the training dataset. The deviation in the model’s prediction of the testing dataset based on the content of a real testing dataset is called testing error and serves as the measure of prediction ability. This process iscalled cross-validation. By construction, the selected probability model is good at reproducing phenotypes from genotypes on the training dataset, at least better than on the testing dataset. This is to say that the training error, or the error in the predictions of the probability model on the training dataset, is bound to be smaller than that on the testing dataset. Thus, we see that the

training error is not a good measure of the prediction ability of the probability modelbecause there is no point in predicting what we have already observed. Ideally, we look at an error as a random variable that measures the deviation of the prediction from the random sample from the true underlying distribution.



**Figure:6** Overview of the cross-validation framework.

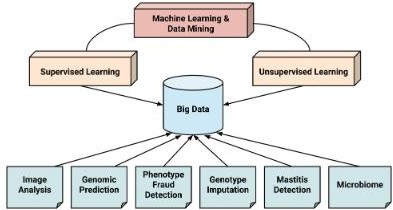
## EXAMPLES FROM ANIMAL SCIENCES:

We now introduce examples of predictive big data analysis using machine learning in animal science. An overview of how these examples are related to big data analysis is provided in, ***Genomic Prediction***

Genetics has arguably made the earliest use of machine learning and data mining among the myriad of animal science fields, in the context of genome-enabled prediction of phenotypes using big data dating back to work by. Big data were referred to here as routine genetic evaluation at national- or company-level involving millions of animals with massive amounts of molecular information, such as SNPs. This continues to be a popular topic in genetics and has been extensively reviewed elsewhere.

***Phenotype Fraud Detection:***

Outlier detection aims to identify profiles that may differ from all other members of a particular group. Genetic evaluation models used to compare animals and identify genetically superior ones can be affected by animals that are outliers in the dataset. Tested the use of the Mahalanobis distance on a dataset consisting of observations of the Jersey dairy cow using routine Nordic genetic evaluation. They reported increased accuracy of predicted breeding values for animals with one or more edited records, in addition to bias reduction for animals from the same contemporary group. Similarly, data electronically submitted by producers to genetic evaluation programs around the world may contain errors incurred during data-capture events.



***Figure:7. Overview of big data analysis in animal science using\ machine learning and data mining tools.***

#### Image Analysis:

Although animal behavior has been at the center of digital image analysis in animal sciences, BW determination in livestock is an emerging area forimage analysis. Livestock body weight is critical for nutritional and breeding management because it is a direct indicator of animal growth, health status, and readiness for market. Therefore, accurate BW estimation is essential to livestock research. This domain separates itself from the traditional method to record BW using ground scales, which is a more laborious and less accurate practice. The application of image analysis for BW determination is a suitable technique to minimize these limitations, given that it is possible to automatically measure the dimensions of an animal’s images and use prediction equations to establish the relationship between them and live BW. Recently, machine vision systems have been successfully used under the above framework. In general, studies have

reported the feasibility of biometric index analysis based on digital images. Infrared light-based depth sensors, such as a Microsoft Kinetic (**MK**) device (Microsoft Corporation, Redmond, WA), are an appropriate vision system for this purpose. The system minimizes the steps of interferences in the captured images owing to ambient light and the animal’s hide color using depth mapping image technology. Images generated from an MK camera are analyzed through specific computational tools, such as the Image Acquisition Toolbox in MATLAB. In this tool, a depth map channel must be specified to ensure that good images can be acquired during the measurement process. For example, assumed depth maps of 50 and 20 frames per acquisition, respectively, in BW studies on pigs and beef cattle. The images composed by these frames were stored and used to close the measurement session for a particular animal.Depending on the aims of research, different sections of images can be utilized. For instance, used section images of the top view of animals provided by the chest width, thorax width, abdomen width, body length, and to predict animal phenotypes using the micro bio me investigated a dairy cattle population of 78 animals representing the extremes of feed efficiency and showed that both the species and the gene composition of the rumen microbes can be used to predict the feed efficiency phenotype with an accuracy of up to 91%. The species composition recorded an accuracy of 80%, whereas the gene composition was 91% accurate. These results underscore the importance of investigation beyond species’ composition and exploration of the functional features of the microbes as such features are better predictors of host phenotypes. Moreover, this study reported that features of the microbes were highly predictive of physiological features, such as milk lactate and milk yield Similarly reported the ability to predict the methane phenotype in dairy cattle populations. They reported an accuracy ranging from 0.163 to 0.553. The authors showed that training dataset size and training dataset variation have a significant effect on prediction accuracy Furthermore, this study compared predictive models and reported that linear mixed models outperform random forests on meta genomic datasets. Such studies demonstrate the value of investigating large datasets for patterns in conversation to predict phenotypes. Developing such meta genomic prediction tools can yield global applications for disease prediction and diagnosis, trace-back, functional phenol typing, and selective breeding. Due to advancements in DNA sequencing technology, DNA sequence information can be generated at high rate, but tools to harness such rich datasets are lacking. For example, the ability to annotate the functional relevance of microbes in the gut is in its infancy. Furthermore, most studies identify correlations

between shifts in the microbes and host phenotypes but fail to identify causality. With the narrow ability of predicting how the microbes reacts to changes and manipulations of the gut ecosystem in livestock species, the opportunities for microbes manipulations are limited and require a multidisciplinary approach as well as novel data mining and machine learning approaches[42].

#### Genotype Imputation:

Another demand for machine learning methods is related to the statistical inference of unobserved genotypes, a technique defined as imputation. Imputation accuracy, measured by the ratio of correct calls compared with the overall call rate, can only be determined by validation strategies that use masked genotypes from a high-density genotype panel, and not necessarily on commercially targeted animals. The prediction of imputation accuracy, based uniquely on the relatedness of low-density genotypes to those in a reference dataset using a high-density panel, was investigated by These results introduced a method for determining the imputed animals to be used for further genomic studies using imputed genotypes with sufficient accuracy without causing bias in the future analysis. This method was based on a single parameter and can be improved upon by machine learning models that contain other information (e.g., the number of animals genotyped in both marker densities [low and high numbers of SNP markers], density of each panel, and breed composition of each animal from the reference and imputed set).

#### Mastitis Detection:

According to mastitis is a major disease in dairy cattle that affects production and udder health in the first and subsequent lactations. This significant disease in dairy herds is associated with a complex set of events triggered by various biological causes and followed by bacterial infection that promotes certain physiological and behavioral effects Milking data such as electrical conductivity, milk yield, lactate dehydroganese and somatic cell scores are usually obtained over time by automatic milking machines and periodic lab tests as well as veterinarian diagnostic tests to determine the incidence of mastitis. A type of NN trained using unsupervised learning can be used to detect mastitis and provide farmers with diagnostic tools for managing mastitis. For instance. Applied an N to detect mastitis, with high accuracy, and to monitor the health status of a herd, especially for early intervention.

#### Image Analysis:

Although animal behavior has been at the center of digital image analysis in animal sciences BW determination in live tock is an emerging area for image analysis. Livestock body weight is critical for nutritional and breeding management because it is a direct indicator of animal growth, health status, and readiness for market. Therefore, accurate BW estimation is essential to livestock research. This domain separates itself from the traditional method to record BW using ground scales, which is a more laborious and less accurate practice. The application of image analysis for BW determination is a suitable technique to minimize these limitations, given that it is possible to automatically measure the dimensions of an animal’s images and use prediction equations to establish the relationship between them and live BW. Recently, machine vision systems have been successfully used under the above framework. In general, studies have reported the feasibility of biometric index analysis based on digital images. Infrared light-based depth sensors, such as a Microsoft Kinetic (**MK**) device (Microsoft Corporation, Redmond, WA), are an appropriate vision system for this purpose. The system minimizes the steps of interferences in the captured images owing to ambient light and the animal’s hide color using depth mapping image technology. Images generated from an MK camera are analyzed through specific computational tools, such as the Image Acquisition Toolbox in MATLAB. In this tool, a depth map channel must be specified to ensure that good images can be acquired during the measurement process. For example, assumed depth maps of 50 and 20 frames per acquisition, respectively, in BW studies on pigs and beef cattle. The images composed by these frames were stored and used to close the measurement session for a particular animal. Depending on the aims of research, different sections of images can be utilized. For instance used section images of the top view of animals provided by the chest width,thorax width, abdomen width, body length, and dorsal height. They found that the chest width section correlated well (0.85) with BW. used selected image sections to estimate pig volume, which was posteriorly correlated with BW. They reported a small average error in BW prediction using pigs of different sizes and breeds. Although the aforementioned studies indicate that digital images taken through the MK system have potential for use in BW estimation in livestock research, some challenges still exist. These include the automation of image data storage and statistical analysis. Along these lines, NN might be a feasible solution due to its flexibility and efficiency in terms of image recognition and prediction performance.

#### Microbiome:

With advancements in next-generation sequencing methods, many opportunities have emerged for developments in animal agriculture. These include investigating complex traits, such as microbiome. Metagenomic investigations on species of livestock have shed light on the importance of the microbiome to feed efficiency, animal health, performance, and productivity. However, although such metagenomic investigations have led to a better understanding of the microbiome in livestock health and productivity, a majority of the microbial genetic information generated is uncharacterized and underutilized. As such, the increasing number of metagenomic studies published has thus far failed to uncover the critical role of the microbiome and harness its metabolic capacity to increase animal productivity. This is mainly due to limitations in current bioinformatics-based approaches to identifying patterns of gene covariation to predict microbiome function. Novel data mining and machine learning approaches are critical for future investigations’ on the microbiome to improve animal production and phenotype prediction in animal agriculture. At present, a number of statistical approaches have been described to understand mechanistic relationships between the host microbiome and the environment. Such approaches have enabled the investigation of the association between the host and environmental factors in the context of microbiome composition. However, few studies to date have attempted of data but also by the complex and dynamic nature of its collection in real time. With the support of data-intensive technologies, we can monitor animals continuously during production, and this information can be used to improve health, welfare, performance, and environmental load. The animal science community today often lacks the infrastructure and tools to make full use of these new types of data. When combined with molecsular information, such as genomics, transcriptomics, and microbiota on individual animal basis, novel machine learning and data mining techniques can advance the implementation of precision animal agriculture to extract critical information and predict future observations from big data. To address such knowledge gaps, we have pointed to the availability of data mining and machine learning tools for analyzing big data, outlined their statistical framework, and illustrated examples from animal sciences. The cyberinfrastructure to host, secure, and share data can also be utilized to exploit big data. It is expected that predictive big data analysis will become increasingly common across all animal science disciplines. We contend that the first steps along this path involve grasping the advantages and pitfalls of these tools when applied to animal science-specific domains.

Furthermore, close collaboration among transdisciplinary fields with complementary backgrounds, such as computer science, economics, engineering, mathematics, and statistics, along with industry, is indispensable to efficiently develop cutting-edge approaches to analyze high-throughput and heterogeneous data. As once argued, predictive modeling is oftentimes more relevant than making inferences about the data-generating mechanism in practical scenarios. Precision animal agriculture allows farmers to formulate prompt management practices, and a predictive machine learning approach for big data-driven agriculture can prove invaluable for addressing challenges lying.

### Big data analytics and Internet of Things (IOT):

Internet of things (IOT) put focus on creating an intelligent environment in which things socialize with each other by sensing, processing, communicating, and actuating activities. As IOT sensors gathered a huge amount of raw data, which is needed to be processed and analyzed, powerful tools will enforce the analytics process. This will stimulate to deploy BDA and its methods on IOT-based data. Ref. [43] proposed a four-layer model to show how BDA can help IOT-based system to work better. This model comprised of data generation, sensor communication, data processing, and data interpretation [43]. It is cited that beyond 2020 cognitive processing and optimization will be considered on IOT data processing . In IOT-based systems, acquired signals from sensors are gathered and deployed for processing in frame-by- frame or batch mode. Also, gathered data in IOT system will be deployed in feature extraction, which is followed by classification stage. Machine learning algorithms will be used in data classifying [44]. Machine learning classification can be deployed on three types of data, which are supervised, semi supervised, and unsupervised [44]. In decision-making level, which is comprised of pattern recognition, deep learning methods, namely, RNN, DNN, CNN, and ANN can be used for discovering knowledge. Optimization process in IOT can be used to create an optimized cluster in IOT data.

In the process of IOT is shown. Data is gathered from sensors. Data enters the filtering process. In this level, de noising and data cleansing happen. Also, in this level, feature extraction is considered for classification phase. After preprocessing, decision making happens on the basis of deep learning methodology. Deep learning and machine learning algorithms can be used in analyzing of data generated through IOT device, especially in the classification and decision- making phase. Both supervised and unsupervised techniques can be used in classification phase

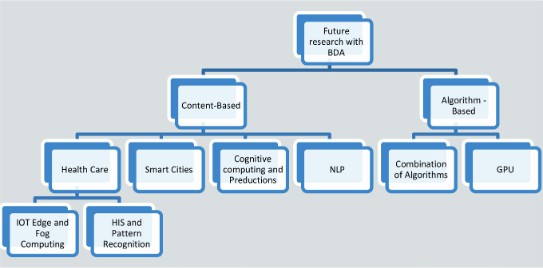
considering the data type. However, both deep learning and machine learning algorithms are eligible in deploying in decision-making.

### Future research directions:

For feature endeavors, it is proposed to work on application of big data analytics methods on IOT fog and edge computing. It is useful to extract patterns from hidden knowledge of data gathered from sensors deploying powerful analytical tools. Fog computing is defined as a technology that is implemented in near distance to end user, which provides local processing and storage to support different devices and sensors. Health care systems gain advantage from IOT for fog computing, which supports mobility and reliability in such systems. Health care data acquisition, processing, and storage of real-time data are done in edge, cloud, and fog layer [45]. In future research, the area that machine learning algorithms can provide techniques for fog computing can be on the focus. IOT data captured from smart houses needs analytical algorithms to overcome the complexity of offline and online data gathered in processing, classification, and also next best action, or even pattern recognition [46]. Hospital information system creates “life sciences data,” “clinical data,” “administrative data,” and “social network data.” These data sources are overwhelmed with illness predictions, medical research, or even management and control of disease [47]. Big data analytics can be a future subject by helping HIS to cover data processing and disease pattern recognition. Smart house creates ground for real-time data with high complexity, which entitles big data analytics to overcome such sophistication. Classical methods of data analyzing lost their ability in front of evolutionary methods of classification and clustering. So graphic processing unit (GPU) for machine learning and data mining purposes bring advantage for large scale dataset [48], which leads the applications into lower cost of data analytics. Another way to create future research is to work over different frameworks like Spark, which is an in-memory computation, and with the help of big data analytics, optimization problems can be solved [49]. Deployment of natural language processing (NLP) in text classification can be accompanied by different methods like CNN and RNN. These methods can gain the result with higher accuracy and lower time (Li et al., 2018).

Predictive analytics offered by big data analytics works on developing predictive models to analyze large volume data both structured and unstructured with the goal of identifying hidden patterns and relations between variables in near future [50]. Big data analytics can help cognitive computing, and behavior pattern recognition deploys deep learning technique to predict future

action as it is used to predict cancer in health care system [51]. It also leads organizations to understand their problems [51]



*Figure8:uture research on big data analytics (BDA).*

So, future research can be focused on both the new area for application of different machine learning or deep learning algorithm for censored data gathered and also mixture of techniques that can create globally optimal solution with higher accuracy and lower cost. Researchers can put focus on existing problems of industries through mixed application of machine learning and deep learning techniques, which may results in optimize solution with lower cost and higher speed. They also can take identified algorithms in new area of industries to solve problems, create insight, and identify hidden patterns.

### Healthcare Stakeholders & Big Data sources:

This section elaborates how the diversified sources of big data cater the requirements of prominent stakeholders (patients, medical practitioners, hospital operators, pharma and clinical researchers, healthcare insurers) in attaining novel healthcare solutions. In addition, it offers appropriate analytical techniques to extract meaningful healthcare patterns of interest from the specified data sources.

###### Patients:

The patient community always expects to avail a broad range of healthcare services at affordable cost with personalized recommendations(Mancini,2014). In addition to physician’s clinical diagnosis, they have an opportunity to gain more medical knowledge through digital platforms such as social media networks, clinical forums etc. These big data sources enable the patients to connect with similar people for gaining information such as disease symptoms, side-effects, hospitalization, drug information, feedback about clinical reports and post effect scenarios with improved privacy. Those patients who are unable to visit hospitals can avail telemedicine services for their healthcare needs. The platform could act as a big data repository and capture vital health signs such as temperature, heart rate, blood pressure and stream those data into a centralized repository for triggering periodical health alerts.

###### Medical Practitioners:

The massive amount of data generated from various phases of diagnosis and treatment plans of patients helps healthcare providers to identify the real insight about the progress of the treatments that are offered by them. There are many sources of big data that are generated by the healthcare system during the execution of treatment plans. It includes classification codes for various diseases and clinical services, laboratory results, clinical notes, medical imaging data, and sensor devices that capture patient’s behaviour under different scenarios. When such big data sources are considered for constructing the Clinical Disease Repository (CDR), it improves public health surveillance and offers faster response through effective analysis of disease patterns. Besides, integration of data from wearable devices into healthcare applications also provides significant benefits such as facilitating physicians to track the usage of drugs, monitor the patient’s health condition at any point in time.

###### Hospital Operators:

To effectively manage the patient’s experiences and for optimizing the available resources

,hospital operators highly rely upon the outcome of big data sources. Models based on predictive and prescriptive analytics are developed with the expertise of data scientists for measuring the strength of relationships between patient satisfaction indexes and availed services. Besides, resource allocation and optimization techniques can be successfully deployed on the basis of available big data towards fulfilling the manpower requirements for different sections of the hospital. The strategic operators of the hospitals can also utilize the location awareness data to decide the co-location of various departments to optimize the use of expensive healthcare

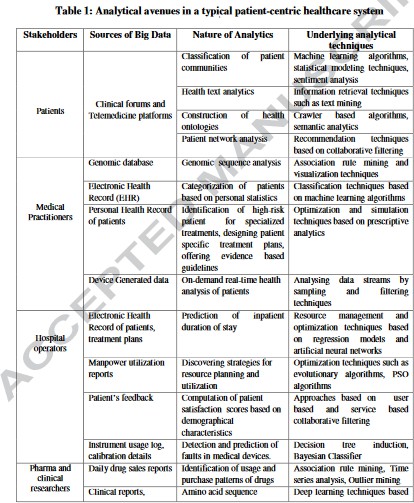
equipment. Development of descriptive models on the basis of posttreatment data generated from follow-up phone calls, email communications, text messages would also facilitate to improve the offered services.

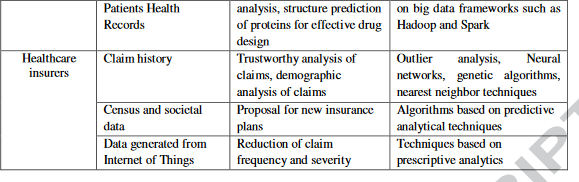
###### Pharma and clinical researchers:

The impact of big data reflects a healthcare reformation in the domain of pharmaceutical and clinical research. The usage of omics and clinical big data (Wu et al., 2017) helps to build predictive models for understanding the biological and drug processes that attribute to the high success rate in attaining effective drug designs. Effective analysis of health data from diversified big data sources helps pharma companies to measure the outcome of designed drugs with smaller and shorter trials (Merelli et al., 2014). By coupling in-memory computing technologies with automated systems in drug manufacturing units, pharmaceutical companies can effectively integrate and analyze various forms of data to build end-to-end product solutions. Inputs from other stakeholders such as drug recommendation by a physician for a particular disease, quantum of consumption by patients, sales history from the drug shops enable the pharmaceutical organizations to evaluate and visualize their current market position for arriving strategic business decisions.

###### Healthcare insurers:

The emergence of healthcare big data opens new analytical avenues for the benefit of healthcare insurers. Accordingly, novel health plans for frequently occurring diseases based on the geographical regions can be introduced with minimal premium cost. Advocating appropriate health plans for customers on the basis of various features such as age, gender, family history, income, nature of job enables benefits for both the insurer and customer. Analyzing unstructured data from claim history through predictive modeling techniques enables the insurance organization to predict patterns of authentic claims and unusual outliers for minimizing the cost of abuse. Big data coupled with Internet-of-Things (IoT) facilitate the insurers to introduce new and innovative business models such as *usage-based insurance* by analyzing the customer behaviour data captured in real-time. Mobile IoT (Internet of Things) plays a crucial role in transforming healthcare by allowing new business models to emerge and enables changes in work processes, productivity improvements and customer experiences(Dimitrov, 2016). By analyzing the influence of big data on the above stakeholders, we have identified the potential big data sources in a typical patient-centric healthcare system.





### Big Data Frameworks for healthcare:

Recent research attempts advocate various healthcare frameworks for handling large volume of diversified data from disparate data sources to churn out significant patterns and trends. This section reviews those big data frameworks and highlights the contribution in the domain of healthcare.An applied architectural framework for healthcare system using big data analytics has been proposed by Raghupathi and Raghupathi (2014). The framework consists of layers such as Data Source layer, Transformation Layer, Big Data platform layer and Analytical layer. The data source layer mainly focuses on internal and external data sources of healthcare found in multiple locations under various formats. The transformation layer is responsible for operations such as extraction, transformation, and loading of data into big data platform through various data staging techniques such as middleware and data warehousing operations. The layer of big data platform comprises of various Hadoop ecosystem tools for performing specific operations on Hadoop Distributed File System (HDFS) using Map- Reduce programming model. The analytical layer performs operations such as querying, reporting, online analytical processing and data mining techniques. Besides, the authors have outlined various tools and platforms for analyzing healthcare big data. Though the proposed architectural framework is a pioneering one in the context of big data for the domain of healthcare, it only emphasizes the theoretical aspects. No experimental evaluations have been conceived based on the proposed framework. Chawla and Davis (2013) proposed a patient centric personalized healthcare framework based on

collaborative filtering approach. It captures patient similarities and produces personalized disease risk profiles for individuals. Collaborative filtering is a data analysis technique designed to predict user’s opinion about an item or service based on the known preferences of a large group of users. In the proposed framework, individual patient’s healthcare history has been compared with all other available patients’ medical histories on the basis of defined similarity constraints such as occupation, symptom, lab result, family history and demographic data. Based on the similarity computation, a pool of similar patients is selected and prediction of diseases has been carried out. With the increasing use of electronic healthcare records, the proposed framework provides a proactive healthcare solution in the context of big data. Besides the advantage of offering patient centric personalized healthcare framework for physicians to assess the disease risk of patients, it handles only the diagnosis codes that confirm to ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) standards. A big data analytical framework that utilizes ubiquitous healthcare system has been attempted by Kim et al. (2014). The framework analyses vital signs extracted from accelerometers to provide healthcare services. Vital signs are continuous time series data that are unstructured in nature and having inadequacy to store in the traditional databases. Electro Cardiogram signals (ECG), respiration and the motion data have been accounted as vital signs. The framework employed open standard platform to support interoperability among data and different devices. Hadoop platform has been extended by adding algorithms to extract feature values from raw data of vital signs and store them for real-time analysis. Despite the novelty of proposing Hadoop based platform for extracting and processing bio signals such as ECG, their work lacks in delivering substantial analytical models on the top of Map-Reduce programming model. The implication of computational aspects in medical and health big data informatics have been extensively surveyed by Fang et al. (2016). They advocated a framework called ‘Health informatics processing pipeline framework’ that combines a sequence of steps to reap meaningful patterns from healthcare big data. The framework consists of process pipeline such as data capturing (identifying data sources such as electronic healthcare data, clinical support data sources, and laboratory results), storing (identifying cost effective storage infrastructure for analysing healthcare data), analysing (performing tasks such as data preprocessing, feature selection and machine learning) searching (extracting meaningful patterns of interest from the outcome of analysis), decision support ( utilizing the pattern base for effective decision making in the health

informatics domain). Besides the proposed framework, certain research directions with respect to issues pertaining to data heterogeneity such as structured and unstructured healthcare data, complexity existing in the available data, privacy issues and visualization of discovered patterns are also explored in their work. Their proposed healthcare framework offers a systematic data processing pipeline for stages of big health informatics such as data gathering, storing, searching and analyzing data from diversified sources. However the focus towards technological aspects of implementation with the aid of big data tools and techniques is obliterated. With the advent of mobile devices and sensor networks, pervasive healthcare services emerged as a novel solution in the health informatics domain, since it offers healthcare services to the patient at anytime/anywhere basis. With the wide range of mobile devices, sensors and wearable applications, data generated in heterogenic formats have been utilized for the provision of on- demand healthcare services. A framework for healthcare big data analytics in mobile cloud computing environment was proposed by Youssef (2014). The framework provides high-level integration, interoperability, availability, and sharing of healthcare data among various stakeholders namely medical practitioners, patients, and drug developers through the following components: (i) Cloud component – hosts patient information and offers healthcare services (ii) EHR component – responsible for integrating distinct patient records from different sources such as pharmacy, hospital, and lab. (iii) Security component – guarantees the protection of security and privacy issues by implementing encryption and authentication techniques (iv) Data analytical component – deploys different analytical tools for discovering new kinds of patterns from the available HER (iv) Care Delivery Organization (CDO) components- represents the different healthcare organizations distributed in various locations. All such organizations can perform data sharing using HL7 protocol, which is the standard structure of communication among healthcare organizations. Though the framework elaborately discusses the security aspects for protecting patient data in a cloud environment, empirical evaluation of the proposed security policies have not been advocated. The need for self-caring services for patients under emergency situation has been focused by Lin et al. (2015) and advocated a cloud based big data healthcare framework. It consists of an off-line Hadoop cluster and an online distributed search cluster. The Hadoop cluster is responsible for off-line storage and index building of medical documents and the online cluster has been designed for processing user query in a highly concurrent and scalable fashion.

The online cluster consists of :

* 1. Search node – for retrieving medical records.
  2. Node for Data analysis – for developing disease symptom lattice
  3. Load balanced cluster – For balancing the load. of user queries.
  4. Based on the proposed framework, a prototype design for home diagnosis.

service has been attributed for testing the validity of the proposal. Also, the scalability of the cloud based framework has been achieved by dynamically adding and removing the nodes in each cluster. Though the proposed work implements a prototype of cloud-based framework for self-diagnosis service, it analyzes only the historical medical records of patients. The importance of semantic interoperability among clinical information is the prime focus of the work carried out by Legaz-Garca et al.(2016). They have argued that the lack of interoperability among clinical models and clinical record yield inefficiency in the healthcare system. To establish semantic integration of Electronic Healthcare Data (EHR), they proposed a framework based on Web Ontology Language (OWL). The patient data (EHR) obtained from the relational databases is transformed into OWL for ontology construction and the constructed ontology has been utilized for data exploration such as EHR based data classification and visualization. Though the framework offers advantages of using semantic technologies in biomedical research, adding further efforts in learning optimal set of parameters for constructing ontologies in archetypes will enhance the outcomes of framework. A cyber-physical system based healthcare framework- ‘smart healthcare framework’ that integrates sensing technologies, cloud computing, Internet of things, and big data analytics has been proposed by Sakr and Elgammal (2016). Various layers that constitute the framework are:

1. Data connection layer – for sensing, extraction and integration.
2. Data storage layer – for storing relational, non-relational and cloud oriented data.
3. Big Data processing and analytical layer – for performing various analytics such as descriptive, predictive and prescriptive analytics.
4. Presentation layer – for developing graphical dashboards and work flows. The proposed architecture holds good for various use-cases such as patient profile analytics, population management, genomic analytics and improved patient.

monitoring where integration of the above four technologies played a vital role. Their work significantly integrates various ICT advantages such as sensor technologies, cloud computing,

Internet of Things and Big Data analytics in offering Smart Healthcare services. In other side, the framework lacks analytical capability to handle complex data sources such as images and streams. A cloud based context aware framework to identify the impact of socio-economic, demographic and geographical conditions on public health has been attributed by Mahmud et al. (2016). It is an Amazon web service based cloud platform integrated with geographical information system for capturing, storing and visualizing the big data. Accordingly, contextual and healthcare data from various remote locations and regions have been captured and a predictive model based on fuzzy-rule based summarization technique for health-shock has been proposed. The fuzzy-rule based technique is used to generate interpretable linguistic rules for classifying the health shocks. The proposed cloud based model facilitates the healthcare professional to understand the impact of socio-economic, environmental and cultural norms that directly or indirectly caused the health-shocks. The novelty of the proposed work has been illustrated with real-time data sets collected from the rural and tribal areas of Pakistan. Though the paper advocates Fuzzy based predictive analytical framework, no adequate comparisons have been made with any benchmarking experimental studies. Jokonya (2014) proposed an integrated big data framework that assists for the framework mainly focuses on the Mining Industry that creates an abundant amount of silica dust. The silica dust that affects the lungs may also cause silicosis and tuberculosis. When the immune system is compromised by HIV/AIDS and silicosis, it makes easier for the tuberculosis to infect the body. The proposed framework addresses the need of Epidemiology predictive model for forecasting and controlling the above mentioned diseases. The big data component of the framework performs data capturing of different Tuberculosis risk factors and combines the data for effective analysis through predictive and descriptive models. Though the proposed framework effectively performs feature selection from different data sets found in Mining Industry for identifying Tuberculosis (TB) disease, the novelty of proposal has not been validated. It has been noted that the Radio-Frequency- Identification (RFID) technologies received significant attention in the healthcare industry by performing tasks such as tracking of medical equipment, hospital supplies, medication and patient information in an attractive proposition. Though the technological impacts are significant, it equally creates privacy concerns raised by RFID tag bearers. Rahman et al.(2017) attempted RFID based healthcare framework for addressing the privacy issues. The proposed framework consists of two components namely

1. Prisens – used as privacy preserving authentication protocol for sensing RFID tags for different identification and monitoring purposes. It includes drug usage monitoring, surgical instrument tracking, hospital personnel tracking and blood tracking.
2. HSAC – provides a privacy preserving healthcare service accessing mechanism for maintaining user’s privacy while accessing various healthcare services.
3. In addition, it follows the role based access mechanism for restricting the unauthorized access of private data that includes both structured and unstructured clinical information.

The framework holds good for cheaper tags (tags with less computational capabilities) as well. Though the motivation behind their framework is to enhance the privacy of users in RFID based healthcare system, inculcating the adherence of different privacy levels for different service requests will add promising potential to their research efforts. A cloud based distributed health information system framework that focuses on privacy and security aspects of medical information has been put forth by Sarkar (2017). The suggested framework applies set of security constraints and access control mechanisms that guarantee integrity, confidentiality, and privacy for medical data.

##### Applications and benefits of big data analytics in medical engineering and healthcare

Big data analytics has the potential in the following aspects:

1. prediction of disease outbreaks – continuous aggregation and analysis of public health data facilitate detecting and managing possible outbreaks of diseases, such as flu [53].
2. pharmaceuticals and medicine – integration of various data from clinical service, patents and public research facilitates new drug discovery [54].
3. personalised and precisive patient care – big data-derived linkages help prompt updates of patient triage, diagnostic assistance and clinical guidelines to provide more personalised and precise treatment [55].
4. tele-diagnosis and e-consultation – real-time tele-diagnosis and e-consultation (ECG) and images are feasible. Big\ data analytics can predict many deaths [56].

Big data technologies have been used in tracking patient mobility and sentiment, monitoring adverse medical events, performing disease surveillance, etc. They are being expected to improve the identification of adverse effects of drugs according to research data and patients

information online, post-exposure symptoms on Face book, Twitter or in health forums.mining of texts (unstructured data) from health data streams helps discover associations between specific drugs and adverse events efficiently [57]. The complexity of healthcare data is generally revealed in clinical trial data submissions, radiology images, population data, genomic data, financial and operational data, etc. Big data analytics can be used in handling these data, which is shown in Table 2 [58]. High-throughput “-omics” techniques bring new energy to diagnostics and provide comprehensive data resources from macro (e.g., phenomics) to micro (e.g., genomics). Big data methods for biomarker development and diagnostic medicine have been used in early detection of complex chronic diseases. In some situations, this helps deeply understand the molecular pathogenesis of diseases [59]. For precision medicine, specific accuracy is expected towards aspects, for example, disease markers, drug target challenges of how to transform unstructured healthcare data into knowledge and benefits for patients [60]. Healthcare big data are in various formats (tables, csv, flat files, etc.) and from external sources (such as laboratories and insurance companies) or internal sources (EMR, clinical decision support systems, etc.) [61]. Main sources of healthcare big data are from payers and providers and include medical images, EMR, claims and pharmacy records, etc. Data analytics will keep expanding from -omics (e.g., genomic, proteomic, epigenomic and metabolomic data). EMR only accounts for a small part of medical data; genomics produces double the amount; and patient-created sources, such as Internet activities, sensors and smartphones generate the largest volume of MBD [62]. In addition, Spatial big data have been utilised for epidemiological analysis. Spatial big data are huge in volume, velocity and variety of sources as well as additional geographic information that is precise spatial information [63].

##### CONCLUSION:

conclusion, data analytics has become a critical component of the pharmaceutical industry, transforming the way drugs are discovered, developed, and marketed. The use of data analytics in the pharma industry has enabled the development of personalized treatment plans based on individual patient data, leading to more effective and targeted treatments. It has also helped analyze genetic data to predict patient responses to specific drugs, allowing for personalized prescribing and reducing the risk of adverse drug reactions. Additionally, data analytics has been used to optimize supply chain operations, reduce costs, improve efficiency, and ensure drug availability. The pharma industry has benefited from the rapid growth of technology and constantly increased its adoption of pharma analytics and other digital technologies to become more efficient and protected against future systemic shocks like coronavirus pandemics. Overall, data analytics in the pharma industry is a powerful tool for improving patient outcomes, advancing drug development, and driving business growth.

Bibliography

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

[2].David J. Adams. The valley of death in anticancer drug development a reassessment. *Trends in Pharmacological Sciences*, 33(4):173 – 180, 2012.

[3].Mei Liu, Michael E Matheny, Yong Hu, and Hua Xu. Data mining methodologies for pharmacovigilance. *ACM SIGKDD Explorations Newsletter*, 14(1):35–42, 2012.

[4].Shobeir Fakhraei, Bert Huang, Louiqa Raschid, and Lise Getoor. Network-based drug-target interaction prediction with probabilistic soft logic. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2014.

[5].Michael J. Keiser, Vincent Setola, John J. Irwin, Christian Laggner, Atheir I. Abbas, Sandra J.Hufeisen, Niels H. Jensen, Michael B. Kuijer, Roberto C. Matos, Thuy B. Tran, Ryan Whaley, Richard A. Glennon, J´erˆome Hert, Kelan L. H. Thomas, Douglas D. Edwards, Brian K.Shoichet, and Bryan L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462 (7270):175–181, November 2009.

[6].Aislyn D.W. Boran and Ravi Iyengar. Systems approaches to polypharmacology and drug discovery. *Current Opinion in Drug Discovery & Development*, 13(3):297, 2010.

[7].Andrew L Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4(11):682–690, 2008.

[8].S. J. Cockell, J. Weile, P. Lord, C. Wipat, D. Andriychenko, M. Pocock, D. Wilkinson,

M. Young, and A. Wipat. An integrated dataset for in silico drug discovery. *Journal of Integrative Bioinformatics*, 7(3):116, 2010.

[9].Rave Harpaz, William DuMouchel, Nigam H Shah, David Madigan, Patrick Ryan, and Carol Friedman. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics*, 91(6):1010–1021, 2012.

[10].Kathleen M Giacomini, Ronald M Krauss, Dan M Roden, Michel Eichelbaum, Michael R Hayden, and Yusuke Nakamura. When good drugs go bad. *Nature*, 446(7139):975–977, 2007. [11].Brian L Strom. How the US drug safety system should be changed. *Journal of the American Medical Association*, 295(17):2072–2075, 2006.

[12].Ren´e Amalberti, Dan Benhamou, Yves Auroy, and Laurent Degos. Adverse events in medicine: Easy to count, complicated to understand, and complex to prevent. *Journal of Biomedical Informatics*, 44(3):390–394, 2011.

[13].Adam L Cohen, Daniel S Budnitz, Kelly N Weidenbach, Daniel B Jernigan, Thomas J Schroeder, Nadine Shehab, and Daniel A Pollock. National surveillance of emergency department visits for outpatient adverse drug events in children and adolescents. *The Journal of Pediatrics*, 152(3):416–421, 2008.

[14].Jason Lazarou, Bruce H Pomeranz, and Paul N Corey. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Journal of the American Medical Association*, 279(15):1200–1205, 1998.

[15].Munir Pirmohamed, Alasdair M Breckenridge, Neil R Kitteringham, and B Kevin Park. Adverse drug reactions. *British Medical Journal*, 316(7140):1295–1298, 1998.

[16].World Health Organization. *The Importance of Pharmacovigilance-Safety Monitoring of Medicinal Products*. World Health Organization, Geneva, 2002.

[17].Rave Harpaz, William DuMouchel, Paea LePendu, and Nigam H Shah. Empirical Bayes model to combine signals of adverse drug reactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1339– 1347. ACM, 2013.

[18].Margaret A. Hamburg and Francis S. Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

[19].William E. Evans and Howard L. McLeod. Pharmacogenomics drug disposition, drug targets, and side effects. *New England Journal of Medicine*, 348(6):538–549, 2003.

[20].Jill U Adams. Pharmacogenomics and personalized medicine. *Nature Education*, 2008. [21].M Whirl-Carrillo, EM McDonagh, JM Hebert, L Gong, K Sangkuhl, CF Thorn, RB Altman,and Teri E Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, 2012.

[22].Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, Alexandra Tang, Geraldine Gabriel, Carol Ly, Sakina Adamjee, Zerihun T. Dame, Beomsoo Han, You Zhou, and David S.Wishart. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 2013.

[23].Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(suppl 1):D355–D360, 2010.

[24].Stefan G¨unther, Michael Kuhn, Mathias Dunkel, Monica Campillos, Christian Senger, Evangelia Petsalaki, Jessica Ahmed, Eduardo Garcia Urdiales, Andreas Gewiess, Lars Juhl Jensen,et al. Supertarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Research*, 36(suppl 1):D919–D922, 2008.

[25].Yanbin Liu, Bin Hu, Chengxin Fu, and Xin Chen. DCDB: Drug combination database.

*Bioinformatics*, 26(4):587–588, 2010.

[26].Justin Lamb, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N. Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A. Armstrong, Stephen J. Haggarty, Paul A Clemons, Ru Wei, Steven A. Carr, Eric S. Lander, and Todd R. Golub. The connectivity map: Using gene- expression signatures to connect small molecules, genes, and disease. *Science*,313(5795):1929– 1935, September 2006.

[27].Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant.Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37(suppl 2):W623–W633, 2009.

[28].Helge G Roider, Nadia Pavlova, Ivaylo Kirov, Stoyan Slavov, Todor Slavov, Zlatyo Uzunov, and Bertram Weiss. Drug2gene: an exhaustive resource to explore effectively the drug- target relation network. *BMC Bioinformatics*, 15(1):68, 2014.

[29].Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, and Shanfeng Zhu. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in Bioinformatics*, page bbt056, 2013.

[30].Muhammed A Yildirim, Kwang-Il Goh, Michael E Cusick, Albert-Laszlo Barabasi, and Marc Vidal. Drug–target network. *Nature Biotechnology*, 25(10):1119–1126, October 2007. [31].Soyoung Lee, Keunwan Park, and Dongsup Kim. Building a drug–target network and its applications. *Expert Opinion on Drug Discovery*, 4(11):1177–1189, November 2009. [32].Muhammed A Yildirim, Kwang-Il Goh, Michael E Cusick, Albert-Laszlo Barabasi, and

Marc Vidal. Drug– target interactions: a brief review. *Briefings in Bioinformatics*, page bbt056, 2013.

[33].Feixiong Cheng, Chuang Liu, Jing Jiang,Weiqiang Lu,Weihua Li, Guixia Liu,Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Computational Biology*, 8(5):e1002503, May 2012. [34].Masahiro Hattori, Yasushi Okuno, Susumu Goto, and Minoru Kanehisa. Heuristics for chemical compound matching. *Genome Informatics Series*, pages 144–153, 2003.

[35].Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. Drug-target interaction prediction through domain-tuned network based inference. *Bioinformatics*, 2013. [36].Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, July 2008.

[37].Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, September 2009.

[38].Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, and Jie Zheng. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, 29(2): 238–245, 2013.

[39].Yuhao Wang and Jianyang Zeng. Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics*, 29(13):i126–i134, 2013.

[40].Cain J, Conway JM, DiVall MV, et al. Report of the 2013-2014 academic affairs committee. Am J Pharm Educ. 2014;78(10):Article S23.

[41].Thakuriah P. Urban informatics, big data and community engagement: a study of enabling organizations. http:// digitalinfomediary.wordpress.com/. Accessed April 24, 2015.

[42].González-Recio, O., G. J. M. Rosa, and D. Gianola. 2014. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livest. Sci. 166:217– 231.

[43].Osipov V, Osipova M. Space–time signal binding in recurrent neural networks wit controlled elements. Neurocomputing. 2018;**308**:194-204.

[44].Peng Y et al. Orthogonal extreme learning machine for image classification. Neurocomputing. 2017;**266**:458-464.

[45].Mutlag AA et al. Enabling technologies for fog computing in health care IoT systems. Future Generation Computer Systems. 2019;**90**:62-78.

[46].Yassine A et al. IoT big data analytics for smart homes with fog and cloud computing. Future Generation Computer Systems. 2019;**91**:563-573.

[47].Li J et al. Medical big data analysis in hospital information system. In: Big Data on Real- World Applications. 2016. Chapter 4.

[48].Cano A. A survey on graphic processing unit computing for large-scale data mining. WIREs Data Mining and Knowledge Discovery. 2017;**8**(1):e1232. DOI: 10.1002/ widm.1232. [49].Gonzalez-Lopez J et al. Distributed nearest neighbor classification for large-scale multi- label data on spark. Future Generation Computer Systems. 2018;**87**:66-82.

[50].Wang Y, Hajli N. Exploring the path to big data analytics success in healthcare. Journal of Business Research. 2017;**70**:287-299.

[51].Sekaran K et al. Deep learning convolutional neural network (CNN) with Gaussian mixture model for predicting pancreatic cancer. Multimedia Tools and Applications. 2019:1-15. DOI: 10.1007/ s11042-019-7419-5.

[52].Côrte-Real N et al. Unlocking the drivers of big data analytics value in firms. Journal of Business Research. 2019;**97**:160-173.

[53].Priyanka K, Kulennavar N. A survey on big data analytics in health care. Inter J Comp Sci Inform Technol. 2014;5(4):5865–5868.

[54].Schultz T. Turning healthcare challenges into big data opportunities: a use-case review across the pharmaceutical development lifecycle. Bul Am Soc Info Sci Tech. 2013;39(5):34–40.

[55].Yang S, Njoku M, Mackenzie CF. Big data’approaches to trauma outcome prediction and autonomous resuscitation. Br J Hosp Med (Lond). 2014;75(11):637–641.

[56].Hsieh JC, Li AH, Yang CC. Mobile, cloud, and big data computing: contributions, challenges, and new directions in telecardiology. Int J Environ Res Public Health. 2013;10(11):6131–6153.

[57].Bansal S, Chowell G, Simonsen L, et al. Big data for infectious disease surveillance and modeling. J. Infect. Dis. 2016;214(suppl\_4):S375–S379.

[58].Chrimes D, Zamani H. Using distributed data over HBase in big data analytics platform for clinical.

[59].Ren G, Krawetz R. Applying computation biology and “big data” to develop multiplex diagnostics for complex chronic diseases such as osteoarthritis. Biomarkers. 2015;20(8):533– 539.

[60].Baro E, Degoul S, Beuscart R, et al. Toward a literature- driven definition of big data in healthcare. BioMed Res Inter. 2015;2015:1–9.

[61].Sabharwal S, Gupta S, Thirunavukkarasu K. Insight of big data analytics in healthcare industry. In 2016 International Conference on Computing, Communication and Automation (ICCCA), IEEE; 2016, 95–100..

[62].Monteith S, Glenn T, Geddes J, et al. Big data for bipolar disorder. Int J Bipolar Disord. 2016;4(1): 10–27.

[63].Lee EC, Asher JM, Goldlust S, et al. Mind the scales: harnessing spatial big data for infectious disease surveillance and inference. J Infect Dis. 2016; 214(suppl\_4):S409–S413.