International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 8 Issue 6, Nov-Dec 2024 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Big Data in Pharmaceutical Industry

Matthew N. O. Sadiku[1], Matthias Oteniya[2], Janet O. Sadiku[3], Susan Abunene[4]

[1,2,4]Roy G. Perry College of Engineering, Prairie View A&M University, Prairie View, TX, USA
[3]Juliana King University, Houston, TX, USA

ABSTRACT

Big data is the datasets that are so large and complex that conventional methods and hardware for collecting, sharing, and analyzing them are impossible. The use of big data is a trend that results in amassing massive amounts of information and data from digital platforms and applications generated in a wide variety of industries such as healthcare, pharmaceuticals, business analytics, and advertising. By adopting big data, the pharma sector can drive improvements at each step of the drug development process. In pharmaceuticals, the importance of big data has boomed over the past decade due to the incorporation of high performing automation processes. Big data is changing the way drugs are developed. This paper provides an overview of recent developments in big data-related deals within the pharmaceutical industry.

KEYWORDS: *big data, big data analytics, pharmaceutical industry*

How to cite this paper: Matthew N. O. Sadiku | Matthias Oteniya | Janet O. Sadiku | Susan Abunene "Big Data in Pharmaceutical Industry" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-8 | Issue-6, December 2024, pp.922-931, URL: www.ijtsrd.com/papers/ijtsrd72727.pdf

IJTSRD72727

INTRODUCTION

Advances in data science and digitalization are transforming the world, including the pharmaceutical industry. In this technology era, like every other industry, things are changing very fast in the pharma sector. We live in an age where there is too much information for one single person to analyze. Big data is a term for analyzing massive data sets. It refers to the massive and varied datasets generated by recording digital touchpoints everywhere. The use of big data has been on the rise in recent years because they offer a better way for companies to analyze customer needs and wants than ever before.

The pharma sector generates immense amounts of information. Clinical trial data, electronic health records, genomics information, real-world evidence, and patient-reported outcomes; all these data entries combined can be referred to as big data. Multiple sensor-equipped manufacturing processes and laboratory analysis are the main sources of primary data. Big data is voluminous and diverse information of any format and from any source that can be converted into insights via analytics. With the hope of the world pinned on the pharmaceutical industry more than ever before, big data analytics plays a crucial role in drug and vaccine development. Big data in the pharmaceutical industry have made positive changes throughout the years because they make the process of drug discovery more successful.

WHAT IS BIG DATA?

Big data applies to data sets of extreme size (e.g. exabytes, zettabytes) which are beyond the capability of the commonly used software tools. It involves situation where very large data sets are big in volume, velocity, veracity, and variability [1]. The data is too big, too fast, or does not fit the regular database architecture. It may require different strategies and tools for profiling, measurement, assessment, and processing.

Big Data is essentially classified into three types [2]:

- *Structured Data:* This is highly organized and is the easiest to work with. Any data that can be stored, accessed, and processed in the form of fixed format is known as a structured data. It may be stored in tabular format. Due to their nature, it is easy for programs to sort through and collect data. Structured data has quantitative data such as

age, contact, address, billing, expenses, credit card numbers, etc. Data that is stored in a relational database management system is an example of structured data.

- *Unstructured Data:* This refers to unorganized data such as video files, log files, audio files, and image files. Any data with unknown form or the structure is classified as unstructured data. Almost everything generated by a computer is unstructured data. It takes a lot of time and effort required to make unstructured data readable. Examples of unstructured data include Metadata, Twitter tweets, and other social media posts.
- *Semi-structured Data:* This falls somewhere between structured data and unstructured data, i.e., both forms of data are present. Semi-structured data can be inherited such as location, time, email address, or device ID stamp.

The different types of big data are depicted in Figure 1 [3].

The process of examining big data is often referred to big data analytics. It is an emerging field since massive computing capabilities have been made available by e-infrastructures [4]. Big data analytics is the application of advanced analytic techniques to large, heterogeneous data sets that comprise structured, semi-structured, and unstructured data from many sources with sizes ranging from terabytes to zettabytes.

Analytics include statistical models and other methods that are aimed at creating empirical predictions. Data-driven organizations use analytics to guide decisions at all levels. Several techniques have been proposed for analyzing big data. These include the HACE theorem, cloud computing, Hadoop, and MapReduce [5].

CHARACTERISTICS OF BIG DATA

Big data is growing rapidly and expanding in all science and engineering, including physical, biological, and medical services. Different companies use different means to maintain their big data. As shown in Figure 2 [6], big data is characterized by 42 Vs. The first five Vs are volume, velocity, variety, veracity, and value.

- *Volume*: This refers to the size of the data being generated both inside and outside organizations and is increasing annually. Some regard big data as data over one petabyte in volume.
- *Velocity*: This depicts the unprecedented speed at which data are generated by Internet users, mobile users, social media, etc. Data are generated and processed in a fast way to extract useful, relevant information. Big data could be analyzed in real time, and it has movement and velocity.
- *Variety*: This refers to the data types since big data may originate from heterogeneous sources and is in different formats (e.g., videos, images, audio, text, logs). BD comprises of structured, semi-structured or unstructured data.
- *Veracity*: By this, we mean the truthfulness of data, i.e. weather the data comes from a reputable, trustworthy, authentic, and accountable source. It suggests the inconsistency in the quality of different sources of big data. The data may not be 100% correct.
- *Value*: This is the most important aspect of the big data. It is the desired outcome of big data processing. It refers to the process of discovering hidden values from large datasets. It denotes the value derived from the analysis of the existing data. If one cannot extract some business value from the data, there is no use managing and storing it.

On this basis, small data can be regarded as having low volume, low velocity, low variety, low veracity, and low value. Additional five Vs has been added [7]:

- *Validity:* This refers to the accuracy and correctness of data. It also indicates how up to date it is.
- *Viability:* This identifies the relevancy of data for each use case. Relevancy of data is required to maintain the desired and accurate outcome through analytical and predictive measures.
- *Volatility:* Since data are generated and change at a rapid rate, volatility determines how quickly data change.
- *Vulnerability:* The vulnerability of data is essential because privacy and security are of utmost importance for personal data.
- *Visualization:* Data needs to be presented unambiguously and attractively to the user. Proper visualization of large and complex clinical reports helps in finding valuable insights.

Instead of the 10V's above, some suggest the following 5V's: Venue, Variability, Vocabulary, Vagueness, and Validity) [8].

Industries that benefit from big data include the healthcare, financial, airline, travel, restaurants, automobile, sports, agriculture, and hospitality industries. Big data technologies are playing an essential role in farming: machines are equipped with sensors that measure data in their environment. Structured and unstructured data are generated in various types [9-11].

PHARMACEUTICALS

Medicines have evolved from crude herbal and botanical preparations into more complex manufacturing of sophisticated drug products and dosage forms. Figure 3 shows a typical drug [12]. The pharmaceutical industry is really complex, and this complexity has translated to the pharma needing better management of their data. Constantly growing, sheer amount of data generated by the pharma companies has also made data management and interpretation a daunting task. The largest pharmaceutical companies in North America are shown in Figure 4 [13].

The pharma sector has access to a lot of data: electronic health records, genomic information, real-world evidence, and more. The emergence of big data is providing pharmaceutical companies with an opportunity to gain novel insights that can enhance and accelerate drug development. Pharmaceutical companies are increasingly leveraging big data technologies to enhance innovation and operational efficiency. The proper application of big data to business problems requires employing data scientists who have a cross-disciplinary ability to translate domain-specific needs into analytical solutions.

Data can be unstructured, semi-structured, or structured, and each format has its place in the pharmaceutical industry. But real-world data comes in a variety of different formats, is often highly unstructured. Unstructured data includes physicians' notes, scans and images, and pathology reports. Unstructured data can also alert drug manufacturers about potential safety issues culled from social media posts and Google searches that report possible adverse reactions. Supervised learning can include methods such as ANN or multivariate regression and classification analysis, which learn from and connect input data and outcomes. Supervised learning methods are commonly associated with process design and controls. Semi-structured data is a hybrid of both the unstructured and structured data.

APPLICATIONS OF BIG DATA IN PHARMACEUTICALS

In recent years, the pharma industry has invested heavily in "data lake" style technologies. Real-world data is used extensively for crafting deterministic models to measure the incidences of adverse drug reactions in patients. Here, we want to analyze some decisive applications of big data in the pharmaceutical sector, with their advantages. Such applications include the following [14,15]:

- *Drug Development:* Drug development is a long and risky road. Very few drug candidates make it to the market. The process of drug development is lengthy and complex combined with several processes, applications, and approvals. At every step of the drug development process, pharmaceutical big data can come in handy. Drug discovery starts with researchers understanding the process behind a disease at a cellular or molecular level. With potential targets identified, the process follows by searching for compounds that can interact with the target and interfere with its activity. In drug discovery, gene expression is one of the most widely used molecular features that has been used to inform target selection. Traditionally, researchers used plant or animal compounds to test candidate drugs. With big data in the pharmaceutical industry, targets can also be directly discovered by analyzing public big data. Figure 5 illustrates drug development [16].
- *Pharmaceutical Manufacturing:* Along with the evolution of medicines, the manufacturing practices for their production have advanced from small-scale manual processing with simple tools to large-scale production as part of a trillion-dollar pharmaceutical industry. Today's pharmaceutical manufacturing technologies continue to evolve as the Internet of things, artificial intelligence, robotics, and advanced computing begin to challenge the traditional approaches, practices, and business models for the manufacture of pharmaceuticals. The application of these technologies has the potential to dramatically increase the agility, efficiency, flexibility, and quality of the industrial production of medicines.
- *Precision Medicine*: It can be defined as an approach aiming to provide the right treatment to the right person at the right time. It is a data-driven approach to disease prevention and treatment that takes into account a person's medical history, genome, lifestyle, and other information. It supports the development of unique drugs that target specific patients. Traditionally, precision strategies remained mostly aspirational for most clinical problems. Patients with a similar cancer subtype often respond differently when they receive the same chemotherapeutics. Using big data is becoming a popular way to study the complex relationship between genomics and chemotherapeutic resistance, toxicity, and sensitivity. Big data can be a key factor in precision medicine, where a disease is diagnosed and treated using appropriate data on a patient's genetic makeup, environmental factors, and behaviors.

- *Personalized Medicine:* Medical practitioners and facilities are leveraging big data to determine individualized treatment for specific patients. They do this by looking at a patient's genetics and using their information to prescribe medications working for them. If a person's DNA is found to be highly responsive to a drug, the doctor can prescribe the medicine to them as this may have better odds of success. On the other hand, if the DNA reacts adversely or none at all, then doctors can look for alternatives.
- *Digital Health:* This is the use of digital technologies for the health and well-being of individuals. It is a field that straddles the line between pharma and fitness. It is a sector that is booming and feeds on big data. For example, Takeda Pharmaceuticals is designing an app with Apple Watch to fight depressive disorders. An industry giant like Roche has developed a sensor to be implanted under the skin, which constantly monitors the blood glucose level of diabetic patients.
- *Clinical Trials:* Clinical trials are crucial in the pharmaceutical and life sciences world as it is used to test whether a specific treatment is effective and safe for human subjects. Clinical trials are costly and time-consuming to run and several clinical trials fail as recruiting the right patient for the trial is quite difficult. The goal of a clinical trial is to tell whether a treatment is safe and effective for humans. Usually, it follows in three sequential stages: (1) a drug is tested on a small group of healthy individuals, (2) the drug is tested on a larger group of people showing a specific condition being targeted, (3) that involves a larger number of patients. The process has always been time consuming and tedious. However, with the wider adoption of big data in pharma, clinical trials are changing. Using big data in pharma can change the way clinical trials are designed and managed. Now researchers can track and detect drug exposure levels, the immunity provided by the medicine, the tolerability and safety of the treatment, and other factors that are crucial for recruits' safety in real time. Analyzing pharmaceutical big data can facilitate adaptive trial design and let researchers change trial parameters based on interim results. Clinical trial information grouped by demographics and genetic factors can be accessed and used to create more personalized treatment options.
- *Monitoring Prescribed Drugs*: When doctors prescribe medication to patients, there is a likelihood not all of them may follow the prescription. This situation used to be almost impossible to track until big data analytics came. Experts can now analyze data from pharmacies and pharmaceutical insurance on the number of patients not taking their prescribed medication. These results are then used in formulating policies and practices to resolve these issues. Also, prescription drug abuse is also one of the trends and problems identified with the help of big data. Records from a pharmacy, insurance claims and many more can be used to track patterns of healthcare professionals overprescribing. They may also be used to assess how many patients consume more than their advisable dosage of medication.

BENEFITS

Big data has emerged as a powerful tool for solving some of the most pressing scientific research and drug discovery challenges. Big data can improve the patient's safety on medication by assessing the risk and side effects of drugs accurately according to the population health statistics on medication. One of the greatest benefits of big data and advanced analytics is that they enable physicians to better match patients with treatments. The pharmaceutical and healthcare industry has experienced a watershed moment in its evolution for treatment, as more and more hospitals look towards big data methods to resolve issues. Other benefits include the following [15,19]:

- *Marketing:* By using pharmaceutical big data, companies can predict industry trends and anticipate the sales of specific medicine based on demographic factors. This can help tailor pharma marketing campaigns to customer behavior. With the help of algorithms, information can be catalogued and structured, providing a baseline for future formulations.
- *Compliance Management:* Big data in pharma plays an important role in facilitating regulatory compliance. Companies in the pharmaceutical industry are subject to a complex web of regulations as well as strict data privacy laws. Pharma companies can minimize regulatory risks by detecting anomalies, deviations, and non-compliant activities early on by using automated monitoring systems and big data analytics solutions.
- *Personalized Medicine:* With the advent of personalized medicine, the patient is moving more and more into the spotlight. Biomarkers are essential for personalized medicine. In recent years, it has become evident that developing new medicines cannot rely on the "one size fits all"

approach. Patient stratification is becoming a prerequisite not only in the real world but also in the design of successful development programs.

- *Predictive Modelling:* This enables researchers to predict drug interactions, toxicity, and inhibition and thus speeds up the whole process. With the help of data analytics, researchers can utilize predictive modelling for drug discovery.

Figure 6 shows the advantages of big data in pharma R&D [17].

CHALLENGES

Adopting big data in the pharma sector is a challenging enterprise that will require companies to overcome organizational silos, seamlessly integrate disparate data sources, and ensure regulatory compliance. The main challenges the industry is facing are associated with the variety of data. While big data has been around for some time, data sets in the pharmaceutical industry have always been complex. Other challenges include the following [15,19]:

- *Privacy Concerns:* Despite its significant usage, big data can still negatively affect pharmacy practice in different ways. This includes privacy and data validity. The problem is that big data can collect information about any person without their consent or knowledge. It means there could be a public access to a person's private information such as their health status and what illnesses they have.
- *Integrating Sources:* Having all data sources well linked is one of the key challenges for the pharma sector to overcome to reap the benefits of big data. Effectively using big data in the pharmaceutical sector requires integrating data generated at all stages of the drug development process, from discovery to regulatory approval to real-world application.
- *Crisis Management:* Pharma companies can no longer afford reactive crisis management. New paradigms should emerge that can help the industry battle long-overdue issues. Industry players may invest in emerging markets and diversify their product portfolios to battle those issues.
- *Regulatory Compliance:* Adopting big data in pharma and rolling out centralized data management systems, you must make sure the data is handled safely and securely. The FDA requires software used in the sector to meet a number of requirements, including access control procedures, user identity verification, tracking of performed actions, and more. When planning your project, make sure to carefully study relevant compliance requirements and incorporate them into the design of your data management solution.
- *Lack of Talent:* The biggest challenge by far has been talent: upgrading skill sets from those sufficient to analyze relatively small amounts of clinical trial data to those required to gain insights from the vast amount of real-world data. The pharma sector has traditionally been a slow adopter of technology, so many companies are still lacking the needed talent to realize their ambitious plans. Pharma industry players must think of an appropriate way to close the knowledge gap, be it breeding in-house talent or turning to external teams.
- *Productivity:* Over the last 20 years, productivity in the pharmaceutical industry has been diminishing because of constantly increasing costs while output has overall been stagnant. Despite many efforts, productivity remains a challenge within the industry. With the implementation of big data initiatives trying to integrate data from disparate data sources and disciplines that are available in life science, the industry has identified a new frontier that might provide the insights needed to turn the ship around and allow the industry to return to sustainable growth.
- *Collaboration:* No single organization or company has all of the data available. It is therefore important for companies, the healthcare sector, and also the academic community to work together. This has been recognized, and many pre-competitive or non-competitive collaborations are taking shape. Big data enabled collaboration among different internal and external healthcare stakeholders will benefit pharma companies by breaking the silos that separate internal functions and enhance integrated, consistent research, and care management.
- *Data Corruption:* This can completely spoil AI models and prediction results, costing pharma companies millions of dollars and making their AI and machine learning unreliable. Bad data produces bad results. Getting good data begins by eliminating these factors and keeping a close vigilance on the effects that matter most. It is therefore essential to have clean curated data before any analytics and insight engines can interpret the data. This data problem underscores the need for data observability from the very earliest stages in creating a data lake. Ensuring data integrity will help the company to acquire a

smaller number of warning letters. Figure 7 shows data integrity [19].

CONCLUSION

Big data in pharma presents vast opportunities for innovation, efficiency, and improved patient outcomes. The pharmaceutical industry is only starting to implement big data initiatives, and a long road still lies ahead. While many pharma manufacturers have both the right tools and growing access to data, relatively few thus far have developed the capabilities to leverage big data fully. The use of big data is still an untapped asset in the pharma industry. Reaping the rewards is still a matter of clear goal setting, strategy, and execution. When pharmaceutical companies collect large amounts of data at different stages of the value chain, they can leverage big data analytics to generate actionable insights for research and development. More information about big data in the pharmaceutical industry can be found in the books in [20,21] and the following related journals:

- *Intelligent Pharmacy*
- *International Journal of Pharmaceutics*
- *International Journal of Research in Pharmaceutical Sciences*

REFERENCES

[1] M. N.O. Sadiku, M. Tembely, and S.M. Musa, "Big data: An introduction for engineers," *Journal of Scientific and Engineering Research,* vol. 3, no. 2, 2016, pp. 106-108.

[2] "The complete overview of big data," https://intellipaat.com/blog/tutorial/hadoop-tutorial/big-data-overview/

[3] R. Allen, "Types of big data | Understanding & Interacting with key types (2024)," https://investguiding-com.custommapposter.com/article/types-of-big-data-understanding-amp-interacting-with-key-types

[4] P. Baumann et al., "Big data analytics for earth sciences: The earthserver approach," *International Journal of Digital Earth*, vol. 19, no. 1, 2016, pp.3-29.

[5] X. Wu et al., "Knowledge engineering with big data," *IEEE Intelligent Systems*, September/October 2015, pp.46-55.

[6] "The 42 V's of big data and data science," https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

[7] P. K. D. Pramanik, S. Pal, and M. Mukhopadhyay, "Healthcare big data: A comprehensive overview," in N. Bouchemal (ed.), *Intelligent Systems for Healthcare Management and Delivery.* IGI Global, chapter 4, 2019, pp. 72-100.

[8] J. Moorthy et al., "Big data: Prospects and challenges," *The Journal for Decision Makers*, vol. 40, no. 1, 2015, pp. 74–96. https://www.grandviewresearch.com/industry-analysis/industrial-wireless-sensor-networks-iwsn-market

[9] A. K. Tiwari, H. Chaudhary, and S. Yadav, "A review on big data and its security," *Proceedings of IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems*, 2015.

[10] M. B. Hoy, "Big data: An introduction for librarians," *Medical Reference Services Quarterly*, vol. 33, no 3. 2014, pp. 320-326.

[11] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: Big data for personalized healthcare," *IEEE Journal of Medical and Health Informatics*, vol. 19, no. 4, July 2015, pp. 1209-1215.

[12] J. Alea, "The benefits of big data in drug development," https://www.contractpharma.com/issues/2024-01-02/view_features/the-benefits-of-big-data-in-drug-development/

[13] A. Deshmukh, "Visualizing the world's biggest pharmaceutical companies," September 2021, https://www.visualcapitalist.com/worlds-biggest-pharmaceutical-companies/#google_vignette

[14] "How big data is affecting pharmacy practice," May 2021, https://bigdataanalyticsnews.com/big-data-affecting-pharmacy-practice/#google_vignette

[15] V. Shashkina, "From drug development to marketing: the potential of big data in pharma," June 2023, https://itrexgroup.com/blog/big-data-in-pharma-definition-use-cases/

[16] "How big data is affecting pharmacy practice," May 2021, https://bigdataanalyticsnews.com/big-data-affecting-pharmacy-practice/#google_vignette

[17] "The big data magic for pharma," https://www.wipro.com/pharmaceutical-and-life-sciences/the-big-data-magic-for-pharma/https://www.wipro.com/pharmaceutical

-and-life-sciences/the-big-data-magic-for-pharma/

[18] P. Sahoo and R. Kamaraj, "Review of data integrity in pharmaceutical industry," *International Journal of Research in Pharmaceutical Sciences*, vol. 11. No. SPL4, 2020.

[19] P. Tormay, "Big data in pharmaceutical R&D: Creating a sustainable R&D engine," *Pharmacist Med,* vol. 29, no. 2, March 2015, pp. 87-92.

[20] M. N. O. Sadiku, U. C. Chukwu, and P. O. Adebo, *Big Data and Its Applications.* Moldova, Europe: Lambert Academic Publishing, 2024.

[21] M. Johannes, *Big Data for Big Pharma: An Accelerator for The Research and Development Engine?* ibidem, 2016.

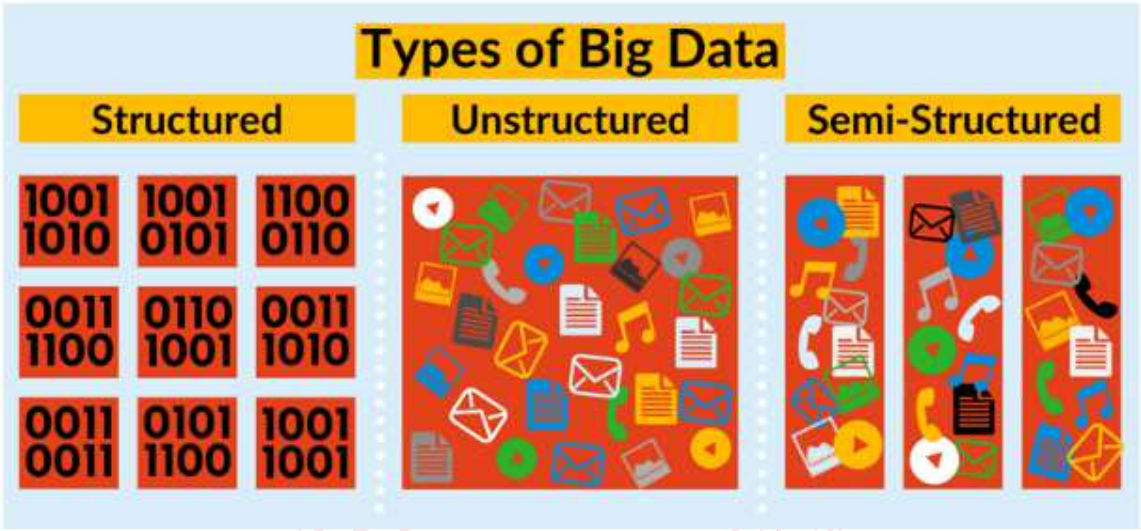

Figure 1 Types of big data [3].

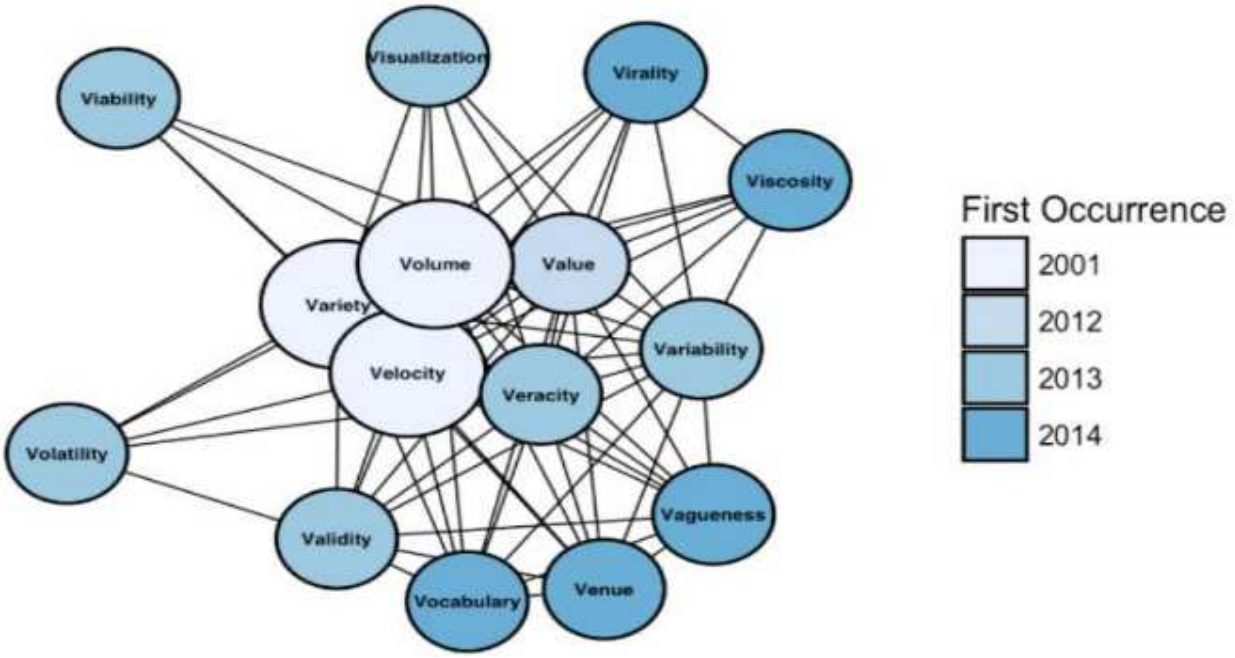

Figure 2 The 42 V's of big data [6].

Figure 3 A typical drug [12].

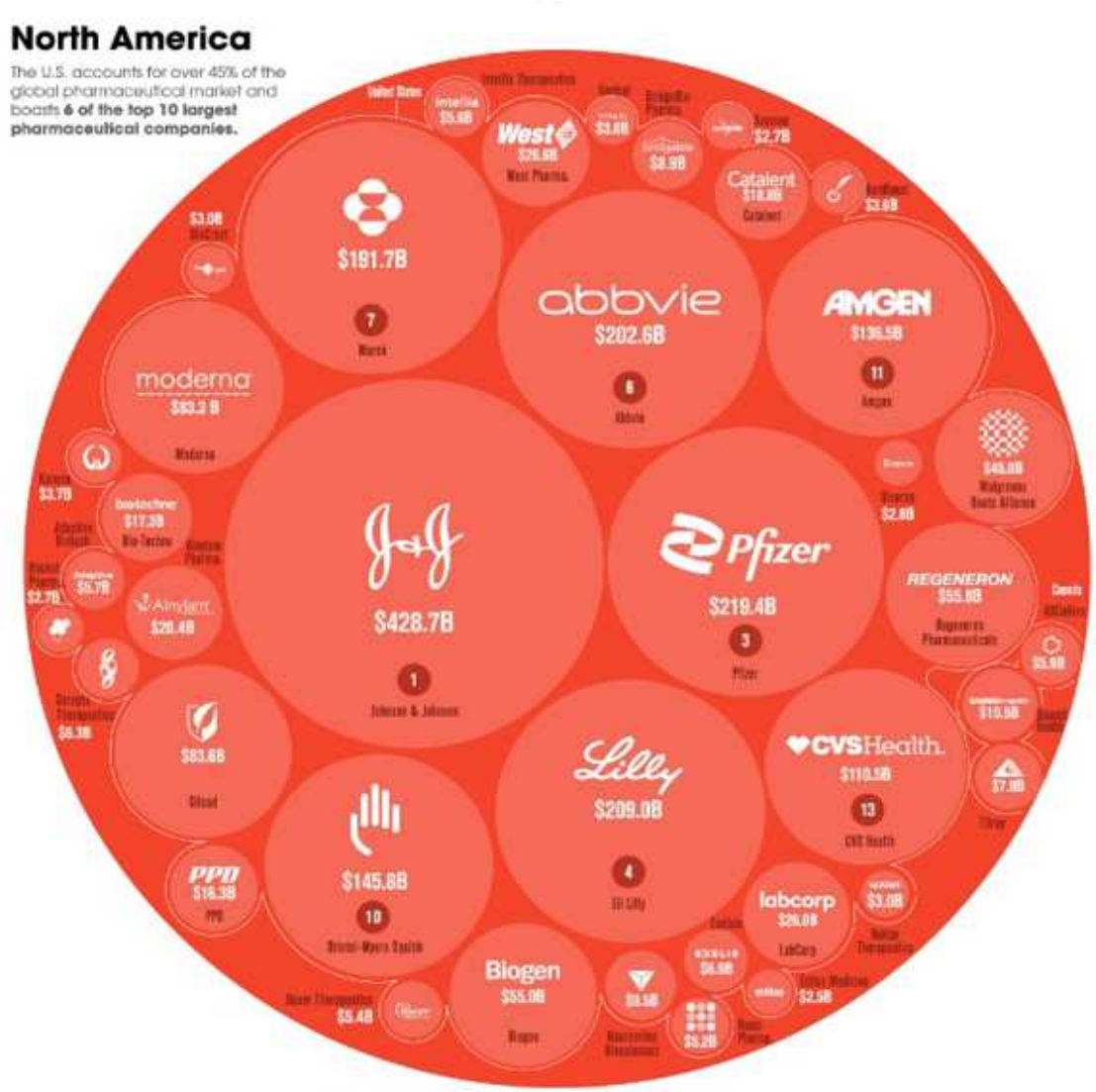

Figure 4 The largest pharmaceutical companies in North America [13].

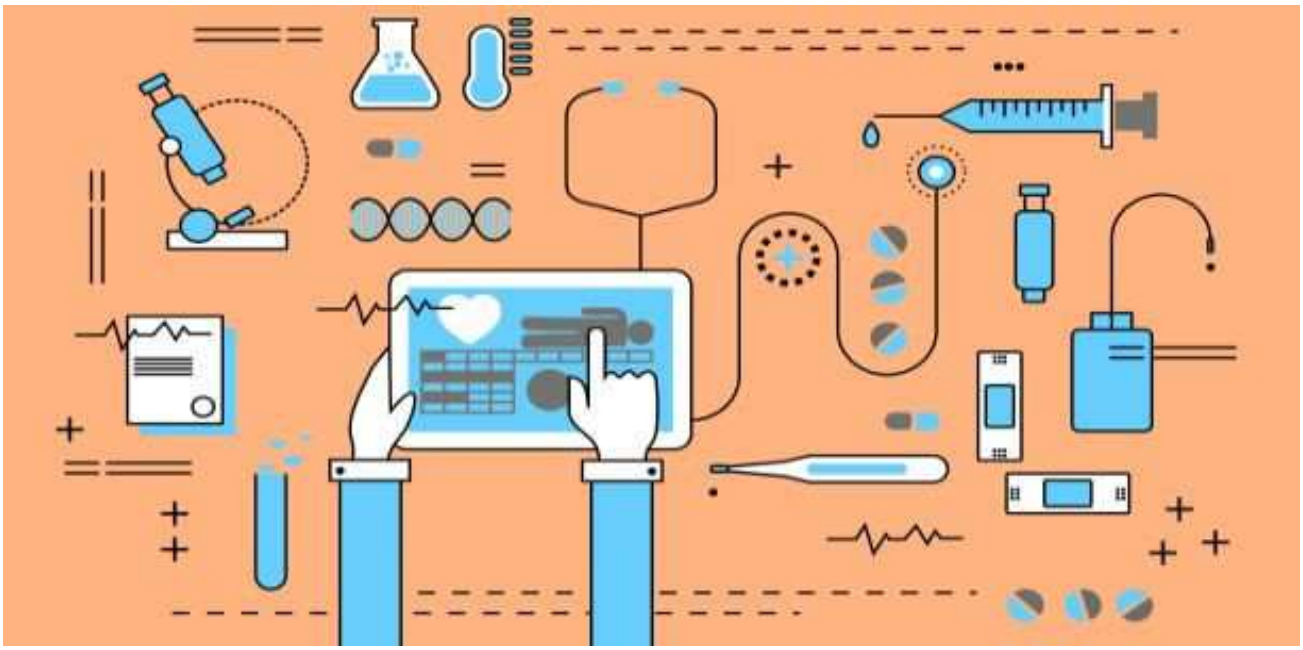

Figure 5 Illustration of drug development [16].

Figure 6 Advantages of big data in pharma R&D [17].

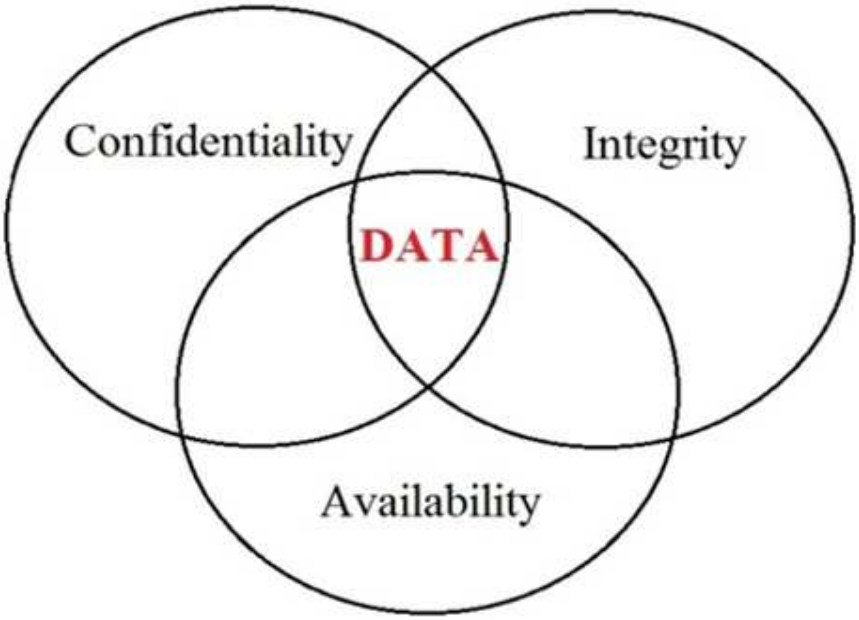

Figure 7 Data integrity [18].