Big data analytics
Lab practical 1

Submitted by:
Divya Mahur
18BCE106

Topic: Big Data in Healthcare

- ## Abstract

"Big data" is a wealth of information that can create miracles. Due to the enormous potential it hides, it has become a topic of special interest in the last two decades. Various industries in the public and private sectors generate, store and analyze big data to improve the services they provide. In the healthcare industry, various sources of big data include hospital records, patient medical records, physical exam results, and devices that are part of the Internet of Things. Biomedical research has also generated a large part of the big data related to public health. These data require proper management and analysis to obtain meaningful information. Otherwise, finding a solution quickly through big data analysis is like finding a needle in a haystack. There are several challenges at every step of big data processing, and only the use of high-end computing solutions for big data analytics can overcome these challenges. That is why to provide related solutions to improve public health, healthcare providers must be equipped with the appropriate infrastructure to systematically generate and analyze big data. Effective management, analysis, and interpretation of big data can be a game-changer by opening up new avenues for modern healthcare. That is why various industries, including the healthcare industry, are taking active steps to transform this potential into better services and financial benefits. Through the powerful integration of healthcare and biomedical data, modern healthcare organizations can revolutionize medical therapies and personalized medicine.

- ## What is big data

In recent years, the term "big data" has become very popular around the world. Almost all fields of research, whether related to industry or academia, generate and analyze big data for various purposes. With regard to this huge pile of data that can be organized and disorganized, the most challenging task is managing it. In view of the fact that big data cannot be managed with traditional software, we need technologically advanced applications and software that can use fast and cost-effective high-end computing power to complete such tasks. The implementation of novel artificial intelligence (AI) algorithms and fusion algorithms is necessary to understand such a large amount of data. In fact, it would be quite a feat to implement automated decision-making by implementing machine learning (ML) methods such as neural networks and other artificial intelligence technologies. However, in the absence of proper hardware and software support, big data can be very vague. We need to develop better technologies to process this "infinite ocean" data and smart grid applications in order to perform effective analytics for actionable insights. With the right storage and analytics tools, information and insights gained from big data can make critical components and services of social infrastructure (such as healthcare, security, or transportation) more comprehensive, interactive, and efficient [3]. Furthermore, visualizing Big Data in a user-friendly way will be a key factor in social development.

- **Big data in healthcare**

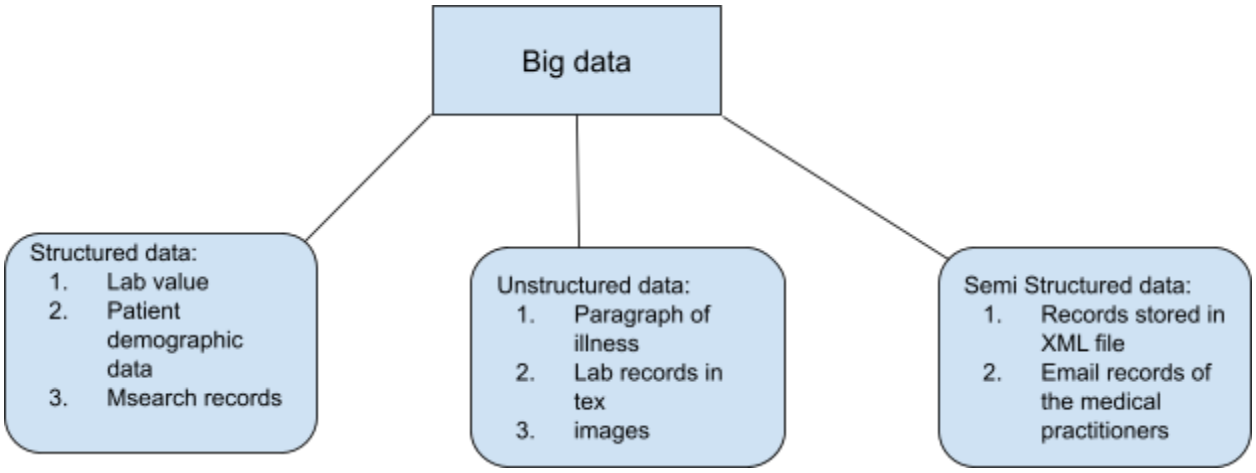
Healthcare as a big-data repository

Health care is a multidimensional system whose sole purpose is to prevent, diagnose, and treat human health-related problems or injuries. The main components of the health system are health professionals (doctors or nurses), health facilities (clinics, hospitals that provide drugs and other diagnostic or treatment technologies), and financial institutions that support the first two. Health professionals belong to various health departments, such as dentistry, medicine, midwives, nursing, psychology, physical therapy, etc. Depending on the urgency of the situation, multiple levels of medical care are required. Professionals use it as the first point of consultation (primary care), emergency care (secondary care) requiring trained professionals, advanced medical examinations and treatments (tertiary care), and very rare diagnostic or surgical procedures (fourth care). At all these levels, healthcare professionals are responsible for different types of information, such as the patient's medical history (data related to diagnosis and prescription), medical and clinical data (such as imaging data and laboratory tests), and other private or personal medical data. data. Data.

Big data is collected in healthcare as:

1. Biomedical research:
   Biological systems, such as human cells, exhibit complex interacting molecular and physical events. In order to understand the interdependencies between the various components and events of such complex systems, biomedical or biological experiments usually collect data on smaller and/or simpler components. Therefore, multiple simplified experiments are needed to generate a comprehensive map of a given biological phenomenon of interest. This shows that the more data we have, the better we understand biological processes. With this idea, modern technology develops at an extremely fast speed.
   Analysis of such big data from medical and healthcare systems can be of immense help in providing novel strategies for healthcare. The latest technological developments in data generation, collection, and analysis have raised expectations towards a revolution in the field of personalized medicine in near future.
2. Gene study
3. Medical record files
4. Medical studies
5. Mobile healthcare
   In today's digital world, everyone seems obsessed with using pedometers built into wearable and wearable devices like smartphones, smartwatches, physical activity dashboards, or tablets to keep track of their health and fitness stats. As

society becomes increasingly mobile in almost all aspects of life, healthcare infrastructure must transform to accommodate mobile devices [13]. Medical and public health practices using mobile devices (called mHealth or mobile health) extend to varying degrees of health care, especially for chronic diseases such as diabetes and cancer [14]. Healthcare organizations are increasingly using mobile healthcare and healthcare services to implement novel and innovative ways of delivering care and coordinating health and wellness. Mobile platforms can improve healthcare by accelerating interactions between patients and healthcare providers. In fact, Apple and Google have developed specialized platforms, such as Apple's ResearchKit and Google Fit, to develop research applications for health and fitness statistics [15]. These applications support seamless interaction with various consumer devices and embedded sensors for data integration. These applications help doctors directly access their general health data. Both users and their doctors know the real-time status of their bodies. These smart devices and apps also help to improve our health plans and promote a healthy lifestyle. Users or patients can become advocates for their own health.



- **Management and analysis of big data**

  Management can be done using
  1. Hadoop
  2. Apache

  Analysis can be done using
  1. Machine learning
  2. Artificial intelligence
  3. Deep learning

4. Image analytics

- ## Commercial platforms for healthcare data analytics

To address the challenges of big data and perform more fluid analysis, companies have implemented artificial intelligence to analyze published results, text data, and image data for meaningful results. IBM Corporation is one of the largest and most experienced players in this field, providing business healthcare analytics services. IBM's Watson Health is an artificial intelligence platform for sharing and analyzing health data between hospitals, providers, and researchers. Similarly, Flatiron Health provides technology-oriented healthcare analytics services with a special focus on cancer research. Other big companies, like Oracle and Google. They also focus on developing cloud-based storage and distributed computing power platforms. Interestingly, in recent years, various companies and startups have also emerged to provide healthcare-based analysis and solutions. Table 2 provides some healthcare providers. Let's take a look at some of these business solutions.

1. AYASDI
2. Linguamatics
3. IBM Watson

- ## Challenges associated with healthcare big data

1. Storage

   Storing large amounts of data is one of the main challenges, but many organizations are satisfied with storing data in their own facilities. It has multiple benefits, such as controlling security, access, and uptime. However, expanding the field server network is costly and difficult to maintain. As costs decrease and reliability increases, cloud-based storage using IT infrastructure appears to be the best option for most healthcare organizations. Organizations must choose cloud partners who understand the importance of specific compliance and security issues in healthcare. Plus, cloud storage offers lower start-up costs, flexible disaster recovery, and easier expansion. Organizations can also take a hybrid approach to their data warehousing procedures, which may be the most flexible and feasible approach for vendors with different data warehousing and access requirements.

2. Cleaning

   The data must be cleaned or debugged to ensure accuracy, correctness, consistency, relevance, and purity after collection. This cleaning process can be performed manually or automatically using logic rules to ensure high accuracy

and integrity. The most sophisticated and accurate tools use machine learning techniques to reduce time and expense and prevent corrupted data from derailing big data projects.

3. Unified format

Patients will generate a large amount of data, and the traditional EHR format is not easy to capture this data because it is complex and difficult to manage. Managing big data is too difficult, especially when healthcare providers don't have a perfect data organization. All clinically relevant information listed must be coded for claims, billing, and clinical analysis purposes. Therefore, medical coding systems such as current procedural terminology (CPT) and International Classification of Diseases (ICD) code sets have been developed to represent basic clinical concepts. However, these code sets have their own limitations.

4. Accuracy

Lack of accuracy can contribute to the quality issues for big data all along its lifecycle. The EHRs intend to improve the quality and communication of data in clinical workflows though reports indicate discrepancies in these contexts. The documentation quality might improve by using self-report questionnaires from patients for their symptoms.

5. Image pre-processing

Studies have found that various physical factors can cause changes in data quality and misunderstandings of existing medical records [30]. Medical imaging often encounters technical hurdles involving many types of noise and artifacts. Inadequate processing of medical images can also lead to image manipulation, for example, it can lead to the representation of anatomical structures not related to the real case, such as veins. Reducing noise, eliminating artifacts, adjusting the contrast of the acquired image, and adjusting the image quality after improper processing are some of the measures that can be implemented to help achieve the goal.

6. Security
7. There have been many security breaches, hacking attacks, phishing attacks, and ransomware incidents that indicate that data security is the top priority of medical institutions. After noticing a series of vulnerabilities, a series of technical protection measures for protected health information (PHI) were developed. These rules are called HIPAA security rules and can help guide organizations in storage, transmission, authentication protocols, and access control, integrity, and auditing. Common security measures such as the latest antivirus software,

firewalls, encryption of sensitive data, and multi-factor authentication can prevent many problems.

8. Meta-data

   To develop a successful data governance plan, you must have complete, accurate, and up-to-date metadata about all stored data. Metadata will consist of information such as the time of creation, purpose, and previous use (by whom, why, how, and when) of those responsible for the data, researchers, and data analysts. This will allow analysts to replicate earlier queries and help follow-up scientific research and develop accurate benchmarks. This increases the usefulness of the data and avoids the creation of "data containers" with little or no use.

9. Visualization

   Use charts, heat maps, and histograms to visualize the data clearly and attractively to illustrate the comparison chart and correctly label information to reduce possible confusion, which can greatly promote the absorption and correct use of information. Other examples include bar charts, pie charts, and scatter charts, which have their own specific ways of communicating data.

10. Data sharing

    Patients may or may not receive your care in multiple locations. In the first case, sharing data with other healthcare organizations is crucial. In this exchange process, if the data cannot be interoperable, the movement of data between different organizations may be severely restricted. This may be due to technical and organizational obstacles. This may prevent clinicians from making critical information about patient follow-up and treatment strategy decisions. Solutions like Fast Healthcare Interoperability Resource (FHIR) and public API, CommonWell (non-profit trade association), and Carequality (common interoperability framework created by consensus) are making data interoperability and sharing easy and secure. The biggest obstacle to data sharing is treating data as a commodity that can provide a competitive advantage.

- ## Conclusion

Big data analysis takes advantage of the gap between structured and unstructured data sources. Migrating to an integrated data environment is a well-known obstacle to overcome. Interestingly, the principle of big data is largely based on the idea that the more information, the more insights can be gained from this information, and future events can be predicted. Several trusted consulting companies and healthcare companies correctly predict that the big data healthcare market will grow exponentially. However, in a short period of time, we have witnessed a series of analyses currently in use that have had a significant impact on the decision-making and performance of the healthcare industry. The exponential growth of medical data from various fields has forced computer science, experts, to devise innovative strategies to analyze and interpret such a large amount of data within a given period of time. The integration of signal processing computer systems of researchers and practical medical professionals has seen growth. Therefore, developing a detailed model of the human body by combining physiological data and "omic" technology may be the next big thing. This unique idea can improve our understanding of diseases and can help develop new diagnostic tools. The continued increase in available genomic data, including hidden errors inherent in experimental and analytical practices, requires more attention. However, at every step in this broad process, there is an opportunity for systematic improvements in health research.