**Module3: DATA-DRIVEN APPROCHES OF HEALTH CARE DATA**

**Topic 1: OVERVIEW OF HEALTH CARE DATA**

RESOURCE LINK:

<https://www.healthcatalyst.com/insights/5-reasons-healthcare-data-is-difficult-to-measure>

Healthcare data is not linear. It is a complex, diverse beast unlike the data of any other industry. Five ways in particular make healthcare data unique:

1. Much of the data is in multiple places.   
   2. The data is structured and unstructured.   
   3. It has inconsistent and variable definitions; evidence-based practice and new research are coming out every day.   
   4. The data is complex.   
   5. Changing regulatory requirements.

Healthcare data tends to reside in multiple places. From different source systems, like EMRs or HR software, to different departments, like radiology or pharmacy. The data comes from all over the organization. Aggregating this data into a single, central system, such as an enterprise data warehouse (EDW), makes this data accessible and actionable.

Healthcare data also occurs in different formats (e.g., text, numeric, paper, digital, pictures, videos, multimedia, etc.). Radiology uses images, old medical records exist in paper format, and today’s EMRs can hold hundreds of rows of textual and numerical data.

Sometimes the same data exists in different systems and different formats. Such is the case with claims data versus clinical data. A patient’s broken arm looks like an image in the medical record but appears as ICD-9 code 813.8 in the claims data.

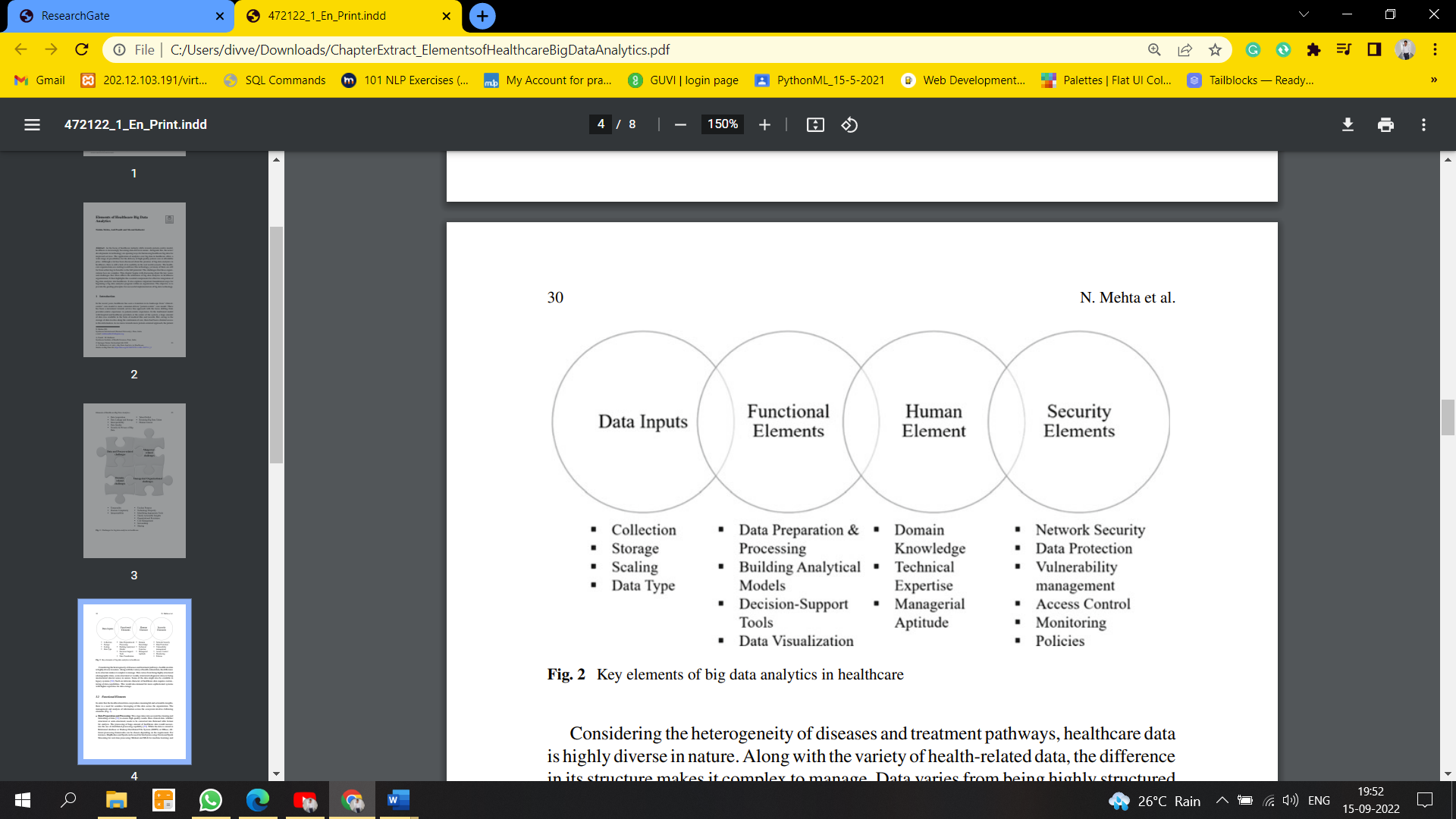
And it looks like the future holds even more sources of data, like patient-generated tracking from devices like fitness monitors and blood pressure sensors.

There are a variety of tools and systems used to collect, store, share and analyze health data gathered through various means. These tools include:

* Electronic Health Records (EHRs)
* Personal Health Records (PHRs)
* Electronic Prescription Services (E-prescribing)
* Patient Portals
* Master Patient Indexes (MPI)
* Health-Related Smart Phone Apps and more

**Topic 2: Basic elements of Health Care Data**

Resource link - https://www.researchgate.net/publication/336219246\_Elements\_of\_Healthcare\_Big\_Data\_Analytics/link/5e535b2a458515072db7a0ba/download



In order that the health-related data can produce meaningful and actionable insights, there is a need for seamless leveraging of this data across the organization. The management and analysis of information across the ecosystem involves following elements (Fig. 3).

**• Data Preparation and Processing:** This stage considers the cleaning and formatting of data [31] to ensure high quality results. Raw clinical data, whether structured or semi-structured, needs to be converted into flattened table format for analysis. The processing of huge amount of healthcare data would necessitate the use of distributed processing capability [35]. While the data is stored in Relational database or Hadoop Distributed File System (HDFS) or HBase, different processing frameworks can be chosen depending on the requirement. For instance, MapReduce and Spark can be used for batch processing; Storm and Spark Streaming for real-time processing; Mahout and MLib for machine learning; and GraphX for graph processing capabilities [36].

A picture containing table

Description automatically generated

Data processing layer thus forms the interface between data storage and analytics.

**• Building Analytical Models:** Analytics is of paramount significance for deriving value from health-related data. To facilitate data-driven optimization and for real-time decision-making in emergency situations, advanced analytical models are essential [34]. These models should explicitly identify the areas where additional value can be created by their use; the users who will make use of them; and the measures to evade inconsistencies in scaling the model.

Healthcare industry is also taking interest in emerging technologies including artificial intelligence and machine learning. With the potential for using descriptive data to predict outcomes, these technologies are expected to drive evidence-based practice and improve clinical decision support. Use of predictive analytics models will assist in accurate diagnoses and treatment. Moreover, it will facilitate early identification of at-risk patients for chronic diseases or for readmission. Making use of operational analytical algorithms will thereby help reduce the cost of care. Advanced data analytics tools will not just assist in predicting outcomes but will move a step beyond to prescribe best possible solution for the problem. Prescriptive analytics will provide healthcare professionals with evaluation of all the possible actions and suggest most appropriate course of action. It will assist healthcare practitioners to prescribe a particular treatment procedure from amongst the probable treatment procedures indicated by analytical model. Also, prescriptive analytics will promote quicker decision-making in the areas related to resource allocation, inventory management, manpower requirement etc.

**Topic 3: Sources of Health care data**

Resource link - https://emerj.com/ai-sector-overviews/where-healthcares-big-data-actually-comes-from/

**1 – The Internet of Things (IoT)** – The IoT has already made waves in the energy and utilities, home monitoring, and transportation industries, and the number of connected things in healthcare is growing. IoT platforms are also relatively cheap and a cost-effective option for building and marketing apps at scale. Data sources at present and in the future include (but are not limited to):

Wearables now allow people to track their heart rate, blood pressure, weight, activity levels, stress levels (Examples: FitBit, PIP, Muse headband, etc.)

Apps are available on smart phones that track a user’s exercise regimen and intensity, amount and quality of sleep, (Examples: Pebble Time, AliveCor Heart Monitor, MyFitnessPal, etc.). Company Medtronic recently partnered with IBM to crowdsource medical data from apps and other devices.

Medical devices and sensors that can also send data into the cloud: pulse oximeters, glucose monitors, electronic scales, blood pressure monitors, SpO2 sensors, proximity sensors (like iBeacon), as well as future sensors that will provide data from millions of patients on a continual basis (a suggested read is our interview with Tracy Ingram, founder of BioscanR, a personal health monitor that collects a range of vital signs and that competed for the Qualcomm X-Prize)

**2 – Electronic Medical Records/Electronic Health Records (EMR/EHR)** – Though they sound identical, the two are different. EMRs contain data from a particular physician’s office, while EHRs are designed to provide a more holistic picture of a patient’s health records over time (birth to death); EHRs can be used as collaborative tools among different medical practitioners and move with a patient from one location to another.

**3 – Insurance Providers** – Including private payer and plan claims, government health plan claims, and pharmacy claims.

**4 – Other Clinical Data** – Data from computerized physician order entry (CPOE) and clinical decision support systems (physician’s written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and other administrative data).

**5 – Opt-in Genome and Research Registries** – GenomeConnect, ClinicalTrials.gov, and NIH Clinical Research Studies are just a handful of organizations and institutions that run genomic research studies and collect participant data.

**6 – Social Media** – Social media posts, including Twitter feeds, blogs, status updates on Facebook and other platforms, and web pages can reflect and provide evidence of a person’s health, mood, state of mind, etc.

**7 – Web Knowledge** – Less patient-specific information, such as emergency care data, news feeds, and articles in medical journals

**Putting Data to Good Use**

Companies and organizations who want to put data to good use will need a comprehensive plan for health data collection, optimization and analysis. Applying predictive analytics, modeling and gleaning pattern-based and intelligent insights has the potential to benefit all ‘players’ in the system, including individuals, health care practitioners, public health facilities, life science organizations, health insurance companies, and medical and pharmaceutical manufacturers. Reduced aggregate medical costs seem likely to be one of the major benefits associated with all of the following advantages:

**1 – Preventative Healthcare and Patient Empowerment**

Using health data and other variables like socioeconomics can help organizations predict missed appointments, noncompliance with medications, and also predict patient trajectory over time

**2 – Combatting Fraud**

Implementing advanced analytic systems for fraud detection and checking the accuracy and consistency of claims will help minimize fraud, as will nearer to real-time claim authorization

**3- Increased Sources of Revenue**

Creating new revenue streams may be available in companies that are able to provide aggregated and synthesized data – patient clinical records and claims data, for example – to third parties who are looking to make advances in the field of medicine and pharmaceuticals (for example, licensing data to assist pharmaceutical companies in identifying patients who have opted-in for participation in clinical trials), which in turn will help drive the creation of new medical devices and pharmaceuticals

**Topic 4: Machine Learning Modeling on Healthcare Data**

**Q1.** Discuss the role of Machine Learning Modeling on Healthcare Data.

**Problem : Prediction of heart and liver Disease**

Diagnosis and prediction of heart, liver disease is an important issue. Due to this reason, a lot of work has been done in this area. We categorize this process into two steps, they are:

The first introduces the steps of selecting the appropriate patient with a selection of features, secondly to test those learning algorithms which provide high accuracy.

**Dataset Collection**

In this research paper we have used the dataset of people who are already tested for the tests related to heart disease in any hospitals. This dataset is a dataframe which can be represented by a table which is having rows and columns to represent the features of dataset which is helpful to train the ML model.

**Manual Exploration**

In this step of Manual exploration we have manually checked the data that is, is the patient having any disease or not by checking the target data column. To give a target data to the algorithm we have used this process.

**Data Pre-Processing**

Data pre-processing is a very important step, because the data we collected had null values, we had to understand that in which column the null value was present and how many null values were coming. So that if null values are very less, then we could remove those data, otherwise we have to replace those null values with any analysis like mean of the column. So we have to Pre-Process the data before sending it to the Machine Learning Model.

**Features Selection**

The Feature Selection is one of the significant difficulties to prepare the Machine Learning algorithms with the given information. So as per a few examinations and explorations in this issue for grouping the Heart disease, it utilizes 13 features. They are Age, Sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal. And for Liver disease classification we are going to use only 10 features. They are Age, Gender, Total\_Bilirubin, Direct\_Bilirubin, Alkaline\_Phosphotase, Alamine\_Aminotransferase, Aspartate\_Aminotransferase, Total\_Protiens, Albumin, Albumin\_and\_Globulin\_Ratio . According to John Peter et al., I.S Jenzi, S.Radhimeenakshi if the no. of attributes is 13 then the accuracey will be more. According to Chaitrali S et al., C.Kalaiselvi, the no. of attributes should be 10 for better accuracy.

**Splitting Data set**

Prior to preparing the Machine Learning model, we gave the objective segment in informational index then we split the information involving train test split capacity in scikit-learn bundle. We partitioned the information into training and testing datasets, later, that training set was consequently isolated into training and validation.

Graphical user interface, application

Description automatically generated

**Machine Learning Algorithms**

For the good accuracy prediction, an efficient Machine Learning algorithm is required. Prediction of the Heart, Liver disease is very complex and critical, it is included with so many procedures, and simple mistake in this procedure can lead to the patient death also. So we have to use the highest accuracy algorithm. In this step we have evaluated the algorithms of machine learning using accuracy score method and we have selected the algorithm with high accuracy.

**Model Evaluation**

In this step we executed the algorithms of machine learning such as Logistic Regression, KNN, Decision Tree Classifier, and SVM on the prediction of Heart disease and we appled models like Logistic Regression, Random Forest Classifier, Decision Tree Classifier on the classification of Liver disease. Here our methodology was to apply various algorithms on those two Heart and Liver illness datasets to approve and assess the accuracy of each model.

We used this score of accuracy to test the algorithms. Accuracy score is the no. of correct predictions to the total no. of predictions.

Graphical user interface, application

Description automatically generated

**Q2.** Write short note on Electronic Medical Records/Electronic Health Records.

An electronic health record (EHR) is a digital version of a patient’s paper chart. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. While an EHR does contain the medical and treatment histories of patients, an EHR system is built to go beyond standard clinical data collected in a provider’s office and can be inclusive of a broader view of a patient’s care. EHRs are a vital part of health IT and can:

1. Contain a patient’s medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results
2. Allow access to evidence-based tools that providers can use to make decisions about a patient’s care
3. Automate and streamline provider workflow

One of the key features of an EHR is that health information can be created and managed by authorized providers in a digital format capable of being shared with other providers across more than one health care organization. EHRs are built to share information with other health care providers and organizations – such as laboratories, specialists, medical imaging facilities, pharmacies, emergency facilities, and school and workplace clinics – so they contain information from all clinicians involved in a patient’s care.

**Benefits of EHRs**

**An electronic health record (EHR) is more than a digital version of a patient’s paper chart.**

EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. While an EHR does contain the medical and treatment histories of patients, an EHR system is built to go beyond standard clinical data collected in a provider’s office and can be inclusive of a broader view of a patient’s care. EHRs can:

iconContain a patient’s medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results

iconAllow access to evidence-based tools that providers can use to make decisions about a patient’s care

iconAutomate and streamline provider workflow

**Q3.** Discuss the issues in Collecting and maintaining EHR

Implementation of EHR in a healthcare system is not as easy as it sounds. There are numerous potential challenges in implementing an electronic health records system. Being aware of the EHR challenges in advance will help the providers to avoid unnecessary data sharing and make the transition from traditional to digital storage of the health data process smoother. Here goes a list of major hurdles that providers should be aware of while implementing EHR.

**1 Cost of Implementation**

It is no surprise that EHR implementation is an expensive affair. The selection, implementation, and optimization of EHR will take away the lion’s share of the planned capital budget investment. As per a study report, the cost of purchasing and installing an EHR system ranges from $15,000 to $70,000 per provider. On a general note, the implementation process can be classified into five components viz. setting up the hardware, software costs, implementation assistance, training for the staff, ongoing network fees, and maintenance. There could be unplanned expenses as well during the implementation. Finding financial resources for EHR implementation is one of the major hurdles, especially for smaller establishments.

**2 Staff Resistance**

There continues to be resistance amongst staff members regarding the idea of EHR implementation in the establishment. In addition, there are health practitioners who are doubtful about the efficacy of electronic health records and the privacy protocols. They may show reluctance to give up the documentation process due to the lack of digital integration. In some cases, the staff lack awareness of the current technological advancements and the comprehensive benefits of EHR implementation. It leads to the delayed implementation of EHR.

**3 Training is time-consuming**

Before the implementation of EHR systems, the staff needs to be given thorough training about the new workflow. The providers and the medical team have to spend extra time and put in extra effort to understand the new system to adjust their work accordingly. It is a time-consuming process and a hassle for both the staff and the management. Small and mid-sized organizations fear the loss of business during the training phase causing barriers to implementing EHR systems in their organizations. Also, the staff may, at times, consider it an unnecessary effort.

**4 Lack of usability**

If the EHR system is inadequate to fit into the existing workflow, providers find it difficult to adapt to it. The one-size-fits-all rule does not suit the EHR system as the workflow of a therapist compared to that of a cardiologist and likewise. The flaws in the design or the inadequacy of training decrease the ease of using the EHR software. Compromised health care data can have fatal consequences on patient health outcomes.

**5 Data Privacy**

data privacy

Another major challenge of EHR is the data privacy concerns of the patient and provider community. The stakeholders often voice concerns over the risk of data leakage due to a natural disaster or a cyber attack. The federal rule has imposed a national policy to protect the confidentiality of personal health data. In case of a security breach, the organization may get into a legal hassle and have to spend millions of dollars to settle the dispute. Hence, it becomes a major responsibility of the provider to ensure the data security of the EHR systems.

**6 Data Migration**

It is a logistical nightmare for the staff to export paper-based documents to digital records. There will be large chunks of documents about the medical history of hundreds of patients and data entry might become a tedious and time-consuming task for the staff. This is one of the major EHR implementation challenges for hospitals and effort is doubled if there is no proper format in the former system.

**7 Limitation of Technical Resources**

This is one of the EHR implementation challenges often faced by small clinical establishments and private health practitioners. They rarely own an in-house team with technical expertise and provider supervision. Moreover, they might not have the required hardware to equip the EHR solution. It is a huge expense to build an in-house team with proper staff with adequate expertise and buy hardware, which is a common reason for small and mid-sized healthcare providers for delaying the EHR implementation process.

**8 Interoperability**

Interoperability is a process of integrated health data to be accessible for easier information exchange with providers and hospitals. Interoperability enables providers and healthcare practitioners to have access to patient data in a consolidated and structuralized manner to provide better treatment and care delivery services. Interoperability in EHR is a necessity to understand a patient’s complete health record, yet, it remains one of the major challenges in electronic health records for proper data transference. This lack of interoperability can hinder care coordination and health outcomes can deter from identifying the cause of medical assistance.

**9 Lack of Proper Planning**

More or less, EHR implementation brings in a cultural change in the organization rather than a mere technological upgrade. Hence, the change in management aspects of EHR implementation become a real challenge. It needs to be strategically planned and commitment is expected from all stakeholders. Without having a structuralized plan for EHR implementation can lead to data breaches and cybersecurity threats to patient information. The successful implementation and sustainability of the EHR systems can be a far-fetched dream without a great amount of planning involved.

**10 Lack of Communication**

Effective communication between the healthcare provider, IT vendor, and the patients is essential to building an EHR system that gives the desired results. It is not a one-time activity but a continuous process to ensure that the expectations of all the parties are met. The concerns and feedback of the provider should be addressed appropriately and the vendor should be able to build effective tools as per the requirement of the provider. Without proper communication, the goals cannot be met, hence creating a massive poll of EHR challenges in integrating data collection systems.