# *Unit-1 - An Introduction to Healthcare Data Analytics:*
## *Healthcare Data Sources and Basic Analytics, Advanced Data Analytics for Healthcare*

Dr.K.Dhanasekaran

# What is data analytics in healthcare?

- Healthcare data analytics is the process of examining healthcare-related data to identify trends, patterns, and insights that can improve patient care, operational efficiency, and inform strategic decisions.

# *Healthcare Data Sources and Basic Analytics*

- Healthcare analytics uses statistical methods, machine learning (ML) and artificial intelligence (AI) to examine data from various sources.
  - **claims and cost data, pharmaceutical research and development data, clinical data, patient behaviours and preference data, electronic medical records, medical imaging and financial data.**
- This analysis helps healthcare professionals make more informed decisions, optimize resource allocation, and ultimately enhance the overall quality of healthcare delivery.
- Healthcare data analytics involves using data to improve healthcare outcomes and operations.

# Data sources

- Data sources:
  - **claims and cost data, pharmaceutical research and development data, clinical data, patient behaviours and preference data, electronic medical records, medical imaging and financial data.**
  - Websites.
  - Books.
  - Journal articles.
  - Newspapers.
  - Encyclopedias.

# Data sources

- Electronic Health Records (EHRs):
- These digital versions of patient history contain data relevant to a patient's care such as demographics, problems, medications, physician's observations, vital signs, medical history, laboratory data, radiology reports, progress notes, and billing data.
- An important property of EHRs is that they provide an effective and efficient way for healthcare providers and organizations to share with one another.
  - For example, a hospital or specialist may wish to access the medical records of the primary provider. An electronic health record streamlines the workflow by allowing direct access to the updated records in real time.

# EHRs

- It can generate a complete record of a patient's clinical encounter, and support other care-related activities such as evidence-based decision support, quality management, and outcomes reporting.

- The storage and retrieval of health-related data is more efficient using EHRs.
  - It helps to improve quality and convenience of patient care, increase patient participation in the healthcare process, improve accuracy of diagnoses and health outcomes, and improve care coordination.

# Data sources

- Claims Data:
- Information from insurance claims provides insights into costs, utilization patterns, and diagnoses associated with specific treatments and procedures.
- Patient Registries:
- These are organized systems for collecting and maintaining information on patients with specific diseases or conditions.
- Public Health Data:
- Includes data from government agencies and organizations on disease outbreaks, immunization rates, and other health-related statistics.

# Data sources

- Clinical Trials Data:
- Provides valuable information from research studies on new treatments and therapies.
- Most of the information about patients is encoded in the form of clinical notes. These notes are typically stored in an unstructured data format and are the backbone of much of healthcare data.
- These contain the clinical information from the transcription of dictations, direct entry by providers, or use of speech recognition applications.
- The processing of clinical text using NLP methods is more challenging due to the ungrammatical nature of short and telegraphic phrases, dictations, shorthand lexicons such as abbreviations and acronyms,and often misspelled clinical terms.
  - All these problems are making the clinical text processing highly challenging in the various standard NLP tasks such as shallow or full parsing, sentence segmentation, text categorization, etc.,

# Data sources

- The manual encoding of this free-text form on a broad range of clinical information is too costly and time-consuming.
- Clinical notes are challenging to analyze automatically due to the complexity involved in converting clinical text that is available in free-text to a structured format.
  - It becomes hard mainly because of their unstructured nature, heterogeneity, diverse formats, and varying context across different patients and practitioners.
- Natural language processing (NLP) and entity extraction play an important part in inferring useful knowledge from large volumes of clinical text to automatically encoding clinical information in a timely manner.

# Data sources

- Genomic Data:
- Information about an individual's genes, which can be used to personalize treatments and understand disease risk.
- The nature of the causality between the genetic markers and the diseases has not been fully established.
  - For example, diabetes is well known to be a genetic disease.
  - However, the full set of genetic markers that make an individual prone to diabetes are unknown.
  - In some other cases, such as the blindness caused by Stargardt disease, the relevant genes are known but all the possible mutations have not been exhaustively isolated.

# Data sources

- Clearly, a broader understanding of the relationships between various genetic markers, mutations, and disease conditions has significant potential in assisting the development of various gene therapies to cure these conditions.
- Moreover, Translating genetic discoveries into personalized medicine practice is a highly non-trivial task with a lot of unresolved challenges.
  - For example, the genomic landscapes in complex diseases such as cancers are overwhelmingly complicated, revealing a high order of heterogeneity among different individuals.
- Researchers are generating new insights into the biology of human disease to predict the personalized response of the individual to a particular treatment.
- Also, genetic data are often modeled either as sequences or as networks.

# Data sources

- Medical Imaging:
- Includes data from X-rays, computed tomography (CT) scans, magnetic resonance imaging (MRIs), ultrasound (U/S) scans, and other imaging technologies.
- It allows physicians to better understand the cause of an illness or other adverse conditions without cutting.
- It can aid disease monitoring, treatment planning, and prognosis.

# Data sources

- Medical imaging:
  - The final goal of biomedical image analysis is to be able to generate quantitative information and make inferences from the images that can provide far more insights into a medical condition.
    - It includes many challenges since the images are varied, complex, and can contain irregular shapes with noisy values.
  - A number of general categories of research problems that arise in analyzing images are object detection, image segmentation, image registration, and feature extraction.
    - All these challenges, when resolved, will enable the generation of meaningful analytic measurements that can serve as inputs to other areas of healthcare data analytics.

# Data sources

- Biomedical Signals:
- Data from devices like electrocardiogram (ECGs), electroencephalogram (EEGs), and other sensors measuring physiological signals.
- Social Media Data:
- Information from social media platforms can be used to monitor public health trends and understand patient sentiment.
- Administrative Data:
- Includes data on hospital operations, staffing, and resource allocation.

# Basic Healthcare Data Analytics

# Basic Healthcare Data Analytics

- **Diagnostic Analytics:**
- Aims to understand the reasons behind observed trends or events.
  - For example, investigating why there was a sudden increase in a particular disease in a specific geographic area.
- **Descriptive Analytics:**
- Focuses on summarizing past data to understand trends and patterns.
  - For example, calculating the average length of stay for patients with a specific condition or identifying the most common diagnoses in a hospital.

# Basic Healthcare Data Analytics

- **Prescriptive Analytics:**
- Recommends actions to optimize outcomes based on data analysis.
  - For example, suggesting the most effective treatment plan for a patient based on their individual characteristics and the available evidence.
- **Predictive Analytics:**
- Uses historical data to forecast future outcomes.
  - For example, predicting which patients are at high risk of readmission or identifying patients who are likely to develop a specific disease.
- These methods help in optimizing patient care, managing public health, and improving hospital efficiency.

# Key Points to Remember

- Diagnostic analytics tells you why it happened.
  - Example: Finding out why infections increased in one hospital ward.
  - Tools: Data mining, deep data checks.

- Descriptive analytics tells you what happened.
  - Example: Reports on patient visits or age groups.
  - Tools: Dashboards, Excel sheets, reports.

# Key Points to Remember

- Prescriptive analytics helps you decide what to do.
  - Example: Helping doctors choose the best treatment for long-term illnesses.
  - Tools: AI tools, test simulations.

- Predictive analytics shows what might happen.
  - Example: Predicting flu season cases using older data.
  - Tools: Machine learning, trend models.

# Types of data analytics: Comparison

| Type | Focus | Example | Tools used |
|---|---|---|---|
| Diagnostic | Reasons for outcomes | Checking Readmission problem | SQL, Python |
| Descriptive | Past data | Patient admission reports | Excel, BI dashboards |
| Prescriptive | What to do next | Advice on Medicine dose | AI Decision support systems |
| Predictive | Future guesses | Patient health risk prediction | Machine learning, deep learning |

# *Unit-1 - An Introduction to Healthcare Data Analytics*

Dr.K.Dhanasekaran

# *Advanced Data Analytics for Healthcare*

# Advanced healthcare analytics

- Advanced healthcare analytics refers to the use of sophisticated data analysis techniques and tools to extract valuable insights from healthcare data.
- It enables better decision-making, improved patient outcomes, and more efficient operations.
- This includes using methods like predictive modeling, machine learning, deep learning, and data mining to analyze vast amounts of data from various sources, including electronic health records, medical imaging, and claims data.

# Key Concepts

- **Generating Data-Driven Insights:**
  - Advanced analytics moves beyond basic reporting and dashboards to provide deeper, actionable insights that can drive meaningful change.
- **Predictive Modeling:**
  - Using historical data and machine learning, advanced analytics can predict future events, such as patient readmissions, disease outbreaks, or resource needs.
- **Personalized Care:**
  - By analyzing patient data, healthcare providers can tailor treatment plans to individual needs and preferences.

# Key concepts

- **Improved Efficiency:**
  - Advanced analytics can identify areas for operational improvement, such as streamlining workflows, optimizing staffing, or reducing waste.
- **Cost Reduction:**
  - By improving efficiency and identifying high-risk patients, advanced analytics can help reduce healthcare costs.
- **Enhanced Diagnostics:**
  - Machine learning algorithms can assist in faster and more accurate diagnoses by analyzing medical images and other clinical data.

# Privacy preservation

- Data privacy preservation in advanced healthcare analytics ensures patient information remains confidential during data analysis.
- Techniques like data anonymization, encryption, and federated learning are used to protect sensitive data.
  - For example, hospitals can analyze patient trends across institutions without sharing raw patient records.
  - Privacy-preserving models allow AI to learn from data while keeping it locally stored.
  - This approach balances innovation in healthcare with compliance to regulations like HIPAA or GDPR.

# HIPAA and GDPR

- HIPAA and GDPR are data protection laws designed to safeguard individuals' personal information:
  - HIPAA (Health Insurance Portability and Accountability Act) is a U.S. law that protects the privacy and security of patients' medical records and other health information. It applies to healthcare providers, insurers, and related organizations.
  - GDPR (General Data Protection Regulation) is a European Union law that protects personal data of all EU citizens. It applies to any organization, anywhere in the world, that processes EU residents' data.
  - Both laws require organizations to ensure data is collected, stored, and used securely and only for legitimate purposes, with individuals' consent where applicable.

# Clinical Prediction Models

- Clinical prediction models models have made a tremendous impact in terms of diagnosis and treatment of diseases.
- **Three categories:**

(i) Statistical methods such as linear regression, logistic regression, and Bayesian models

(ii) Sophisticated methods in machine learning and data mining such as decision trees and artificial neural networks

(iii) Survival models that aim to predict survival outcomes.

- All of these techniques focus on discovering the underlying relationship between covariate variables, which are also known as attributes and features, and a dependent outcome variable.

# Clinical Prediction Models

- The choice of the model to be used for a particular healthcare problem primarily depends on the outcomes to be predicted.
- Some of the most common outcomes include binary and continuous forms. Other less common forms are categorical and ordinal outcomes.
- In addition, there are also different models proposed to handle survival outcomes where the goal is to predict the time of occurrence of a particular event of interest.

# Temporal Data Mining

- Healthcare data almost always contain time information and it is inconceivable to reason and mine these data without incorporating the temporal dimension.
- There are two major sources of temporal data generated in the healthcare domain.
  - The first is the electronic health records (EHR) data
  - The second is the sensor data.
- Mining the temporal dimension of EHR data is extremely promising as it may reveal patterns that enable a more precise understanding of disease manifestation, progression and response to therapy.

# Temporal Data Mining

- EHR data:
  - Some of the unique characteristics of EHR data (such as of heterogeneous, sparse, high-dimensional, irregular time intervals) makes conventional methods inadequate to handle them.
- Sensor data:
  - Unlike EHR data, sensor data are usually represented as numeric time series that are regularly measured in time at a high frequency.
- Examples of these data:
  - physiological data obtained by monitoring the patients on a regular basis and other electrical activity recordings such as electrocardiogram (ECG), electroencephalogram (EEG), etc.
  - Sensor data for a specific subject are measured over a much shorter period of time (usually several minutes to several days) compared to the longitudinal EHR data (usually collected across the entire lifespan of the patient).

# Visual Analytics

- The ability to analyze and identify meaningful patterns in multimodal clinical data must be addressed in order to provide a better understanding of diseases and to identify patterns that could be affecting the clinical workflow.
  - Visual analytics provides a way to combine the strengths of human cognition with interactive interfaces and data analytics that can facilitate the exploration of complex datasets.
  - Visual analytics is a science that involves the integration of interactive visual interfaces with analytical techniques to develop systems that facilitate reasoning over, and interpretation of, complex data.

# Visual Analytics

- Due to the rapid increase of health-related information, it becomes critical to build effective ways of analyzing large amounts of data by leveraging human–computer interaction and graphical interfaces.
- In general, providing easily understandable summaries of complex healthcare data is useful for a human in gaining novel insights.
  - In the evaluation of many diseases, clinicians are presented with datasets that often contain hundreds of clinical variables.
  - The multimodal, noisy, heterogeneous, and temporal characteristics of the clinical data pose significant challenges to the users while synthesizing the information and obtaining insights from the data.

# Clinico-Genomic Data Integration

- Human diseases are inherently complex in nature and are usually governed by a complicated interplay of several diverse underlying factors, including different genomic, clinical, behavioral, and environmental factors.
  - Clinico-pathological and genomic datasets capture the different effects of these diverse factors in a complementary manner.
  - It is essential to build integrative models considering both genomic and clinical variables simultaneously so that they can combine the vital information that is present in both clinical and genomic data.
  - Such models can help in the design of effective diagnostics, new therapeutics, and novel drugs, which will lead us one step closer to personalized medicine.

# Clinico-Genomic Data Integration

- **Clinical data** refers to a broad category of a patient's pathological, behavioral, demographic, familial, environmental and medication history.
- **Genomic data** refers to a patient's genomic information including Single Nucleotide Polymorphism (SNPs), gene expression, protein and metabolite profiles.
  - In most of the cases, the goal of the integrative study is biomarker discovery which is to find the clinical and genomic factors related to a particular disease phenotype such as cancer vs. no cancer, tumor vs. normal tissue samples, or continuous variables such as the survival time after a particular treatment.

# Information Retrieval

- Additional information for use in the healthcare data analytics includes scientific data and literature.
- The techniques most commonly used to access this data include those from the field of information retrieval (IR).
- IR is the field concerned with the acquisition, organization, and searching of knowledge-based information, which is usually defined as information derived and organized from observational or experimental research.
- Information retrieval models are closely related to the problems of clinical and biomedical text mining.
- The basic objective of using information retrieval is to find the content that a user wanted based on his requirements.
  - This typically begins with the posing of a query to the IR system. A search engine matches the query to content items through metadata.

# Information Retrieval

- The two key components of IR are:
  - Indexing, which is the process of assigning metadata to the content, and retrieval, which is the process of the user entering the query and retrieving relevant content.
  - The most well-known data structure used for efficient information retrieval is the inverted index where each document is associated with an identifier.
    - Each word then points to a list of document identifiers. This kind of representation is particularly useful for a keyword search.
  - Furthermore, once a search has been conducted, mechanisms are required to rank the possibly large number of results, which might have been retrieved.
    - A number of user-oriented evaluations have been performed over the years looking at users of biomedical information and measuring the search performance in clinical settings.

# *Unit-1 - An Introduction to Healthcare Data Analytics*

Dr.K.Dhanasekaran

# Applications and practical systems

# Applications and practical systems

- Disease Surveillance:
  - Tracking disease outbreaks and developing strategies for prevention and control.
- Patient Risk Stratification:
  - Identifying high-risk patients who may require more intensive care or monitoring.
- Personalized Treatment Plans:
  - Developing tailored treatment plans based on individual patient characteristics and preferences.
- Operational Efficiency:
  - Optimizing staffing levels, managing inventory, and streamlining workflows.
- Fraud Detection:
  - Identifying and preventing fraudulent activities in healthcare billing and claims processing.

# Drug discovery

- Drug discovery in healthcare analytics uses AI and big data to identify potential drug candidates faster and more cost-effectively.
  - For example, machine learning models can predict how a compound will interact with disease-related proteins.
- Decision support systems (DSS) assist clinicians by providing data-driven recommendations for diagnosis and treatment.
  - For instance, a DSS might analyze patient symptoms and history to suggest possible disease prevention and the best treatment options.
- These tools improve accuracy, reduce trial-and-error, and accelerate medical advancements.

# Data Analytics for Pervasive Health

- Pervasive health refers to the process of tracking medical well-being and providing long-term medical care with the use of advanced technologies such as wearable sensors.
  - For example, wearable monitors are often used for measuring the long-term effectiveness of various treatment mechanisms.
- These methods face a number of challenges, such as knowledge extraction from the large volumes of data collected and real-time processing.
  - However, recent advances in both hardware and software technologies (data analytics in particular) have made such systems a reality.
  - These advances have made low cost intelligent health systems embedded within the home and living environments a reality.

# Data Analytics for Pervasive Health

- A wide variety of sensor modalities can be used when developing intelligent health systems, including wearable and ambient sensors.
- In the case of wearable sensors, sensors are attached to the body or woven into garments.
  - For example, 3-axis accelerometers distributed over an individual's body can provide information about the orientation and movement of the corresponding body part.
- In addition to these advancements in sensing modalities, there has been an increasing interest in applying analytics techniques to data collected from such equipment.
  - Some examples include cognitive health monitoring systems based on activity recognition, persuasive systems for motivating users to change their health and wellness habits, and abnormal health condition detection systems.

# Healthcare Fraud Detection

- Healthcare fraud has been one of the biggest problems faced by the United States and costs several billions of dollars every year.
- With growing healthcare costs, the threat of healthcare fraud is increasing at an alarming pace.
- Given the recent scrutiny of the inefficiencies in the US healthcare system, identifying fraud has been on the forefront of the efforts towards reducing the healthcare costs.
  - One could analyze the healthcare claims data along different dimensions to identify fraud.

# Healthcare Fraud Detection

- The complexity of the healthcare domain, which includes multiple sets of participants, including healthcare providers, beneficiaries (patients), and insurance companies, makes the problem of detecting healthcare fraud equally challenging and makes it different from other domains such as credit card fraud detection and auto insurance fraud detection.
- In these other domains, the methods rely on constructing profiles for the users based on the historical data and they typically monitor deviations in the behavior of the user from the profile.
  - However, in healthcare fraud, such approaches are not usually applicable, because the users in the healthcare setting are the beneficiaries, who typically are not the fraud perpetrators.
  - Hence, more sophisticated analysis is required in the healthcare sector to identify fraud.

# Healthcare Fraud Detection

- Several solutions based on data analytics have been investigated for solving the problem of healthcare fraud.
- The primary advantages of data-driven fraud detection are automatic extraction of fraud patterns and prioritization of suspicious cases.
- Most of such analysis is performed with respect to an episode of care, which is essentially a collection of healthcare provided to a patient under the same health issue.

# Data Analytics for Pharmaceutical Discoveries

- 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.
  - The high failure rate of drugs during this process, make the trial phases known as the "valley of death."
  - Most new compounds fail during the FDA approval process (U.S Food and Drug Administration) in clinical trials or cause adverse side effects.
- Interdisciplinary computational approaches that combine statistics, computer science, medicine, chemoinformatics, and biology are becoming highly valuable for drug discovery and development.
- In the context of pharmaceutical discoveries, data analytics can potentially limit the search space and provide recommendations to the domain experts for hypothesis generation and further analysis and experiments.

# Data Analytics for Pharmaceutical Discoveries

- Data analytics can be used in several stages of drug discovery and development to achieve different goals.
  - In this domain, one way to categorize data analytical approaches is based on their application to pre-marketing and post-marketing stages of the drug discovery and development process.
- In the pre-marketing stage, data analytics focus on discovery activities such as finding signals that indicate relations between drugs and targets, drugs and drugs, genes and diseases, protein and diseases, and finding biomarkers.
- In the post-marketing stage an important application of data analytics is to find indications of adverse side effects for approved drugs.
- These methods provide a list of potential drug side effect associations that can be used for further studies.

# Clinical Decision Support Systems

- Clinical Decision Support Systems (CDSS) are computer systems designed to assist clinicians with patient-related decision making, such as diagnosis and treatment.
- CDSS have become a crucial component in the evaluation and improvement of patient treatment since they have shown to improve both patient outcomes and cost of care.
  - They can help in minimizing analytical errors by notifying the physician of potentially harmful drug interactions, and their diagnostic procedures have been shown to enable more accurate diagnoses.
- Some of the main advantages of CDSS are
  - Ability in decision making and determining optimal treatment strategies
  - Aiding general health policies by estimating the clinical and economic outcomes of different treatment methods and
  - Estimating treatment outcomes under certain conditions.

# Clinical Decision Support Systems

- The main reason for the success of CDSS are their electronic nature, seemless integration with clinical workflows, providing decision support at the appropriate time/location.
- Two particular fields of healthcare where CDSS have been extremely influential are pharmacy and billing.
  - CDSS can help pharmacies to look for negative drug interactions and then report them to the corresponding patient's ordering professional.
  - In the billing departments, CDSS have been used to devise treatment plans that provide an optimal balance of patient care and financial expense.

# Computer-Aided Diagnosis

- Computer-aided diagnosis/detection (CAD) is a procedure in radiology that supports radiologists in reading medical images.
  - CAD tools in general refer to fully automated second reader tools designed to assist the radiologist in the detection of lesions.
  - There is a growing consensus among clinical experts that the use of CAD tools can improve the performance of the radiologist.
- The radiologist first performs an interpretation of the images as usual, while the CAD algorithms is running in the background or has already been precomputed.
- Structures identified by the CAD algorithm are then highlighted as regions of interest to the radiologist.
  - The principal value of CAD tools is determined not by its stand-alone performance, but rather by carefully measuring the incremental value of CAD in normal clinical practice, such as the number of additional lesions detected using CAD.
  - Secondly, CAD systems must not have a negative impact on patient management (for instance, false positives that cause the radiologist to recommend unnecessary biopsies and follow ups).

# Computer-Aided Diagnosis

- From the data analytics perspective, new CAD algorithms aim at extracting key quantitative features, summarizing vast volumes of data, and/or enhancing the visualization of potentially malignant nodules, tumors, or lesions in medical images.
- The three important stages in the CAD data processing:
  - candidate generation (identifying suspicious regions of interest),
  - feature extraction (computing descriptive morphological or texture features), and
  - classification (differentiating candidates that are true lesions from the rest of the candidates based on candidate feature vectors).

# Mobile Imaging for Biomedical Applications

- Mobile imaging refers to the application of portable computers such as smartphones or tablet computers to store, visualize, and process images with and without connections to servers, the Internet, or the cloud.
- Today, portable devices provide sufficient computational power for biomedical image processing and smart devices have been introduced in the operation theater.
  - While many techniques for biomedical image acquisition will always require special equipment, the regular camera is one of the most widely used imaging modality in hospitals.
  - Mobile technology and smart devices, especially smartphones, allows new ways of easier imaging at the patient's bedside and possess the possibility to be made into a diagnostic tool that can be used by medical professionals.
- Smartphones usually contain at least one high-resolution camera that can be used for image formation. Several challenges arise during the acquisition, visualization, analysis, and management of images in mobile environments.

# Resources for Healthcare Data Analytics

Dr.K.Dhanasekaran

# Resources

- There are a few popular organizations that are primarily involved with medical informatics research.
  - They are American Medical Informatics Association (AMIA) [49], International Medical Informatics Association (IMIA) [50], and the European Federation for Medical Informatics (EFMI).
- These organizations usually conduct annual conferences and meetings that are well attended by researchers working in healthcare informatics.
- The meetings typically discuss new technologies for capturing, processing, and analyzing medical data. It is a good meeting place for new researchers who would like to start research in this area.

# International Journals

- The following are some of the well-reputed journals that publish top-quality research works in healthcare data analytics:
  - Journal of the American Medical Informatics Association (JAMIA)
  - Journal of Biomedical Informatics (JBI)
  - Journal of Medical Internet Research
  - IEEE Journal of Biomedical and Health Informatics, Medical Decision Making
  - International Journal of Medical Informatics (IJMI), and Artificial Intelligence in Medicine.

# Blogs

- **Health Affairs Blog**
- **Health IT Analytics**: https://healthitanalytics.com/
- **The Healthcare AI Blog** (by Health Catalyst)

# Specific health data repositories

- Due to the privacy of the medical data that typically contains highly sensitive patient information, the research work in the healthcare data analytics has been fragmented into various places.
- Many researchers work with a specific hospital or a healthcare facility that are usually not willing to share their data due to obvious privacy concerns.
  - However, there are a wide variety of public repositories available for researchers to design and apply their own models and algorithms.
- Due to the diversity in healthcare research, it will be a cumbersome task to compile all the healthcare repositories at a single location.

Task:

- List out the Specific health data repositories dealing with a particular healthcare problem and data sources.

# Online courses

- Coursera
- Health Informatics on FHIR
- Data Analytics in Healthcare
- edX
- Big Data Analytics in Healthcare
- Johns Hopkins Data Science Specialization (on Coursera)
  - Includes tools like R, regression modeling, machine learning—highly applicable to healthcare.
- Udacity – Data Analyst Nanodegree
  - Offers real-world projects that can be tailored to healthcare datasets.

# Tools & Technologies

- Programming Languages: Python, R, SQL
- Libraries:
  - Python: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
  - R: tidyverse, caret, ggplot2, dplyr
  - BI & Visualization: Tableau, Power BI, Qlik
  - Databases: PostgreSQL, MySQL, MongoDB, Google BigQuery
  - Healthcare-specific Tools:
    - FHIR (Fast Healthcare Interoperability Resources)
    - SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms)(SNOMED CT), LOINC, ICD-10 (for coding and classification)
    - Note:
    - SNOMED CT is a comprehensive, multilingual clinical healthcare terminology. It provides a standardized way to represent clinical information in electronic health records (EHRs).
    - **LOINC (Logical Observation Identifiers Names and Codes)**
      - LOINC is a universal standard for identifying health measurements, observations, and laboratory tests.
    - **ICD-10 (International Classification of Diseases, 10th Revision)**
      - ICD-10 is a diagnostic coding system created by the World Health Organization (WHO) to classify diseases and a wide variety of signs, symptoms, abnormal findings, and external causes of injury or diseases.

# Public Datasets

- **Centers for Medicare & Medicaid Services**(CMS) Medicare Data:
  - https://data.cms.gov/
- Medical Information Mart for Intensive Care (MIMIC)-III and MIMIC-IV
  - ICU data from Beth Israel Deaconess Medical Center
  - https://physionet.org/about/database/
- HealthData.gov
  - Government healthcare data including hospital performance, community health, etc.
  - https://healthdata.gov/
- **Surveillance, Epidemiology, and End Results** Program (SEER) Cancer Statistics, a program of the **U.S. National Cancer Institute (NCI)**
  - https://seer.cancer.gov/data/

# Communities & Forums

- LinkedIn Groups: Health Informatics, Healthcare Data Analytics
- Reddit: r/HealthIT, r/datascience
- Stack Overflow: For technical questions (tag: healthcare, biostatistics)
- GitHub: Search repositories related to health data science
- AMIA (American Medical Informatics Association): https://amia.org/
- HIMSS (Healthcare Information and Management Systems Society)
- https://wonder.cdc.gov/
- Kaggle Healthcare Datasets
- https://www.kaggle.com/datasets?search=healthcare

# Thank you

# Electronic Health Records-Components of EHR

Dr.K.Dhanasekaran

# What is an Electronic Health Record (EHR)?

- EHRs are real-time, patient-centred records that provide instant and secure access to authorised users.

- Unlike traditional paper records, EHRs allow for broader data access, automation, and integration with other digital systems (e.g., labs, pharmacies, imaging centres).

# Key Features of EHR

- Stores comprehensive patient history.
- Enables real-time data access for multiple users.
- Integrates with lab systems, imaging, pharmacy, etc.
- Includes tools for clinical decision support.
- Helps in billing, scheduling, and reporting.
- Improves coordination among healthcare providers.

# Components of EHR

- Below is a breakdown of the core components of an Electronic Health Record system:

1. Patient Demographics
    - Basic patient information:
        - Name, age, sex
        - Contact info
        - Insurance details
        - Emergency contacts

- Example: John Smith, 45, Male, Phone: 123-456-7890, Insurance: ABC Health

# Components of EHR

2. Medical History
   - Past diagnoses
   - Surgeries
   - Allergies
   - Family history
   - Social history (e.g., smoking, alcohol)

- Example: Diagnosed with diabetes in 2017; allergic to penicillin

# Components of EHR

3. Medication and Prescriptions
   - Current and past medications
   - Dosage, frequency, and prescribing provider
   - Drug allergy alerts

- Example: Metformin 500 mg, 2x daily, prescribed by Dr. Lee

# Components of EHR

4. Laboratory and Test Results
   - Blood tests
   - Urinalysis
   - Imaging (X-rays, MRI, CT scans)
   - Automatic integration from lab systems

- Example: CBC done on 01/01/2025 shows WBC = 8.3 $x10^9/L$

# Components of EHR

5. Progress Notes and Clinical Documentation
    - Notes written by healthcare providers during visits
    - SOAP format (Subjective, Objective, Assessment, Plan)
    - Example: “Patient reports chest pain; ECG ordered; diagnosis: Angina”
6. Vital Signs Monitoring
    - Blood pressure
    - Heart rate
    - Temperature
    - Respiratory rate

- Example: BP = 130/85 mmHg, Temp = 98.6°F

# Components of EHR

7. Immunization Records
    - Vaccination history
    - Dates administered
    - Providers who administered them
- Example: COVID-19 vaccine, 2nd dose on 03/15/2025

8. Billing and Insurance Information
    - Insurance providers
    - Billing codes (ICD-10, CPT)
    - Claims processing
    - Payment tracking

# Components of EHR

9. Clinical Decision Support (CDS)
   - AI/logic-based tools to support decision-making
   - Alerts for drug interactions
   - Diagnostic recommendations

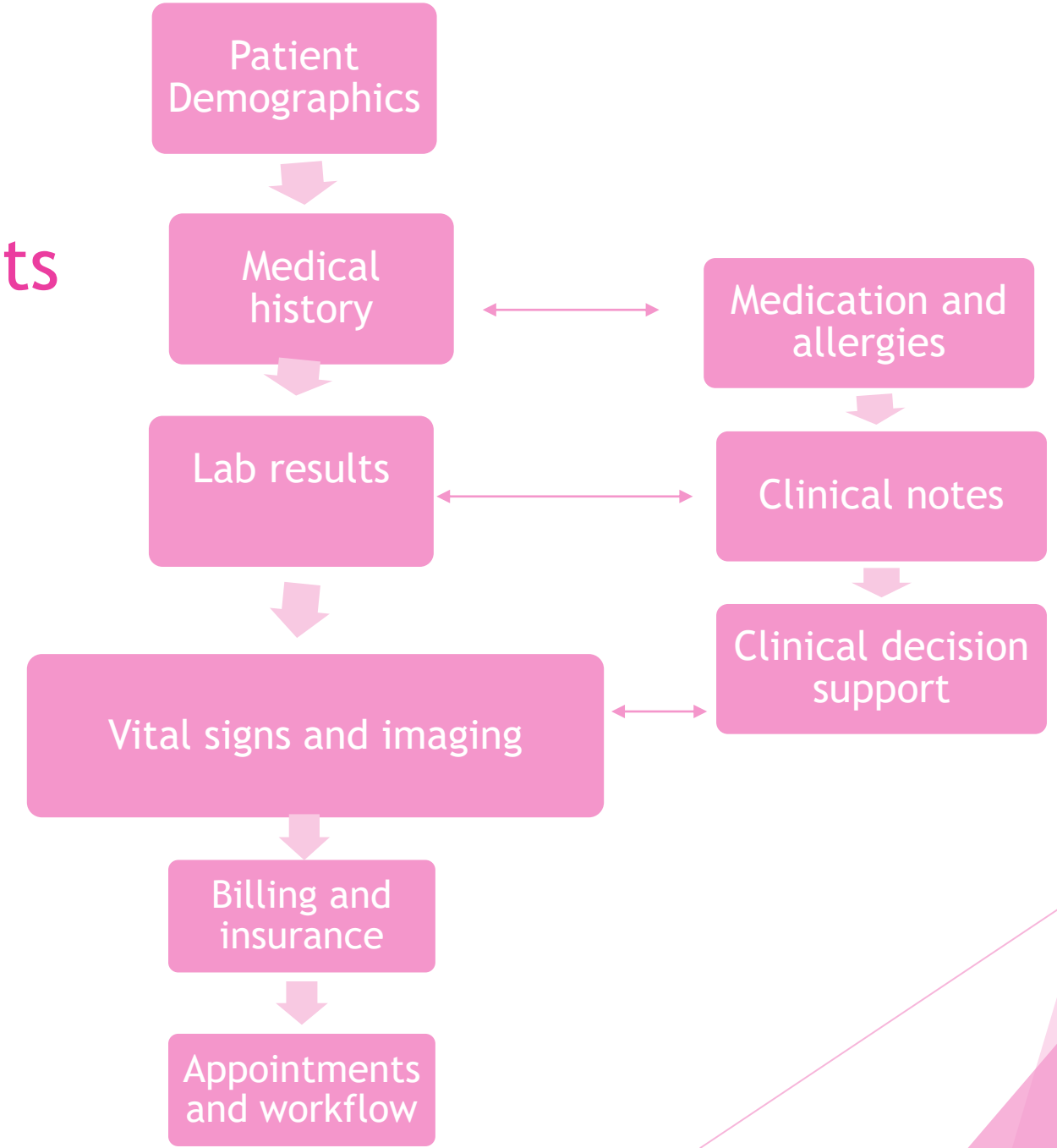
- Example: Alert: “Potential interaction between Warfarin and Aspirin”

# Components of EHR

10. Appointment Scheduling and Workflow Tools
    - Appointment booking
    - Staff workload tracking
    - Follow-up reminders

11. Communication Tools
    - Secure messaging between patient and providers
    - Internal communication among healthcare staff

# Diagram of EHR System Components



# Example Scenario

- A 55-year-old woman, Mary Taylor, visits her doctor for a check-up. The EHR system:
  - Pulls her history – diabetes, hypertension
  - Shows medication – Lisinopril and Metformin
  - Displays last lab results – HbA1c = 7.5%
  - Prompts alert – annual eye exam due
  - Doctor updates note – blood pressure slightly elevated
  - Orders a lab test – lipid profile
  - Schedules a follow-up in 1 month
- All of this happens seamlessly within the EHR interface.

# Coding Systems in EHR, Barriers to Adopting EHR, Challenges of Using EHR Data

Dr.K.Dhanasekaran

# 1. Coding Systems in EHR (Electronic Health Records)

Step 1: Understand what coding systems are

- Coding systems are standardized ways to record medical information (diagnoses, procedures, medications) so that everyone—doctors, hospitals, insurers—can understand and share data clearly.

Step 2: Common coding systems in healthcare

- ICD (International Classification of Diseases): Used to code diagnoses.
- CPT (Current Procedural Terminology): Codes medical procedures and services.
- SNOMED CT: A comprehensive clinical terminology for detailed patient data.
- LOINC (Logical Observation Identifiers Names and Codes): Codes lab tests and clinical observations.

# Example

- Imagine a patient comes in with chest pain. The doctor diagnoses acute myocardial infarction (heart attack) and orders an ECG test.
  - The diagnosis is coded using ICD-10: e.g., I21.9 (acute myocardial infarction, unspecified).
  - The ECG procedure is coded using CPT: e.g., 93000 for electrocardiogram.
  - Lab tests related to cardiac enzymes are coded in LOINC.
  - Additional detailed clinical terms (symptoms, findings) could be coded with SNOMED CT.

- This standardized coding allows the hospital to communicate the patient's status to insurers for billing and to public health agencies for tracking heart attack trends.

# Barriers to Adopting EHR

Step 1: Recognize the benefits of EHR

- EHRs improve documentation, patient safety, data sharing, and analytics. Yet, adoption can be slow or difficult.

# Barriers to Adopting EHR

Step 2: Identify common barriers

- Cost:
  - Buying, implementing, and maintaining EHR systems is expensive.
- Training and Usability:
  - Healthcare staff need time and training to use new software, which may have poor user interfaces.
- Workflow Disruption:
  - EHRs can change how clinicians work, sometimes slowing them down initially.
- Data Privacy Concerns:
  - Fear of patient data breaches or misuse.
- Interoperability Issues:
  - Different EHR systems may not “talk” well with each other, limiting data sharing.

# Example

- A small rural clinic wants to switch from paper records to an EHR system. However,
  - The upfront cost is high, and they don't have a dedicated IT budget.
  - Staff worry about learning a complex system and how it will affect their daily patient visits.
  - The clinic's existing lab uses a different EHR software that doesn't easily exchange data, so lab results have to be manually uploaded.
  - The clinic's management is also worried about keeping sensitive patient information safe.

- Because of these barriers, the clinic delays adopting EHR despite its long-term benefits.

# Challenges of Using EHR Data

Step 1: Understand the nature of EHR data

- EHR data is rich but can be messy, incomplete, and inconsistent because it's primarily collected for clinical care, not analytics.

# Challenges of Using EHR Data

Step 2: Common challenges

- Data Quality Issues:
  - Missing, incorrect, or inconsistent entries (e.g., typos, wrong units).
- Data Complexity:
  - Different coding standards, unstructured text notes mixed with structured fields.
- Privacy and Security:
  - Regulations limit data access and sharing.
- Data Integration:
  - Combining EHR data with other sources (wearables, claims, genomics) is hard.
- Bias and Representativeness:
  - Data may reflect biases in who seeks care or how clinicians document.

# Example

- A hospital tries to use EHR data to predict patients at risk of readmission within 30 days.
- Many patients' discharge instructions are recorded in free-text notes, making it hard to extract actionable data automatically.
- Some key lab results are missing or recorded under different units, complicating analysis.

# Example

- The dataset mostly includes urban patients, so the model may not perform well for rural populations.
- Data privacy rules restrict sharing patient information with external researchers.
- The hospital's data science team spends a lot of time cleaning, standardizing, and interpreting the data before they can build useful models.

# MCQs

Which of the following techniques in healthcare data analytics is most effective for predicting patient outcomes based on historical data and patient features?

A) Descriptive Analytics

B) Predictive Analytics

C) Prescriptive Analytics

D) Diagnostic Analytics

# Answer:

B) Predictive Analytics

# MCQs

Which of the following challenges most significantly impacts the quality of data used for predictive analytics in healthcare EHR systems?

A) Data duplication across different systems

B) Lack of real-time data processing capabilities

C) Inconsistent coding practices and unstandardized data entry

D) Excessive patient consent documentation

# Answer:

C) Inconsistent coding practices and unstandardized data entry

# MCQs

Which of the following is a practical system used to analyze healthcare data for patient care improvement?

A) Decision Support Systems (DSS)

B) Social Media Tools

C) Personal Finance Apps

D) Content Management Systems

# Answer:

A) Decision Support Systems (DSS)

# Benefits of EHR

Dr.K.Dhanasekaran

# Benefits of EHR

| Benefit | Description |
| --- | --- |
| Improves Patient Care | Accurate, up-to-date records |
| Enhances Coordination | Shared access across providers |
| Reduces Errors | Alerts and standardized data |
| Saves Time | Automated workflows |
| Enables Data Analytics | For population health and research |
| Supports Telemedicine | Remote access to records |