SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

HealthCare Data Analytics

Vaibhav P. Vasani

Assistant Professor

Department of Computer Engineering

K. J. Somaiya College of Engineering

Somaiya Vidyavihar University

vaibhav.vasani@gmail.com

Somaiya TRUST

Benefits of EHR

- EHRs are transformational tools. The scope of paper-based systems is severely limited.
- We need EHRs to improve the quality of patient care and increase productivity and efficiency.
- In terms of the overall management and costs, EHRs are a better choice.
- They also help in complying with government regulations and other legal issues.

Enhanced Revenue

- An EHR system can capture the charges and bills for clinical services provided, laboratory tests, and medications more accurately.
- Utilization of electronic systems decrease billing errors. They also provide a better documentation opportunity for these services that can be used to resolve financial disputes. Better management of information yield more accurate evaluation and increase reimbursements.
- According to experts, due to inaccurate coding systems, 3%–15% of a healthcare provider's total revenue is lost
- An EHR system can be programmed or configured to generate alerts for both patients and doctors when a healthcare service is due. This can aid better management of collecting revenue. It can be used to garner more revenues by incorporating services like telemedicine, e-visits, virtual office visits, etc. ***It is true that all kinds of services are not possible over the Internet or telephone network, but not all diseases will require extensive diagnosis and laboratory testing.*** Diseases commonly treated through telemedicine include acne, allergies, cold and flu, constipation, diabetes, fever, gout, headache, joint aches and pains, nausea and vomiting, pink eye, rashes, sinus infection, sore throat, sunburn and urinary tract infections, anxiety and depression, etc.

Averted Costs

- After adopting electronic systems, some costs associated with the previous way of operating a business are eliminated.
- The Center for Information Technology leadership suggested that the use of EHRs will save a total of $44 billion each year.
- Adopting EHR has the following averted costs
  - **Reduced paper and supply cost**: To maintain paper-based health records an organization will require a lot of paper, printing materials, and other supplies. Adopting EHR will reduce these costs. After adopting EHRs, one organization estimated a reduction of 90% of paper usage within a few months.
  - **Improved utilization of tests**
  - **Reduced transcription costs**
  - A study of fourteen solo or small-group primary care practices in twelve U.S. states reports the median transcription cost saving to be $10,800, where a minimum saving was $8,500 and a maximum was $12,000 for the year 2004–2005 . Other related research work also describes saving $1,000–$3,000 per physician, per month.
  - **Improved productivity**
  - **Better availability of information and elimination of chart**
  - **Improved clinician satisfaction**.

Averted Costs – for students

- After adopting electronic systems, some costs associated with the previous way of operating a business are eliminated.
- The Center for Information Technology leadership suggested that the use of EHRs will save a total of $44 billion each year.
- Adopting EHR has the following averted costs
  - **Reduced paper and supply cost**: To maintain paper-based health records an organization will require a lot of paper, printing materials, and other supplies. Adopting EHR will reduce these costs. After adopting EHRs, one organization estimated a reduction of 90% of paper usage within a few months.
  - **Improved utilization of tests**: In electronic systems, test results are better organized. A healthcare staff no longer needs to carry the reports from one place to another. Identifying redundancy or unnecessary tests is easier. This can reduce the loss of information and ensure improved utilization of tests. A study by Wang et al. reports better utilization of radiology tests after adopting EHRs.
  - **Reduced transcription costs**: An EHR can reduce transcription costs for manual administrative processes. It utilizes structured flow sheets, clinical templates, and point-of-care documentation. In a typical outpatient setting, physicians generate about 40 lines of transcription per encounter. For a group of three practicing physicians, treating 12,000 patients annually at the cost of $0.11 for each transcription line results in over $50,000 per year.
  - A study of fourteen solo or small-group primary care practices in twelve U.S. states reports the median transcription cost saving to be $10,800, where a minimum saving was $8,500 and a maximum was $12,000 for the year 2004–2005 [47]. Other related research work also describes saving $1,000–$3,000 per physician, per month.
  - **Improved productivity**: EHR helps to improve workflows by utilizing resources more efficiently and reducing redundancies. As a result, the overall productivity of individuals increases.
  - **Better availability** of information and elimination of chart: In EHR, all the charts are in digital format. It eliminates the need to pull, route, and re-file paper charts A significant amount of effort is spent on creating, filing, searching, and transporting paper charts. A study estimated that the elimination of paper charts can save $5 per chart pull. It is also comparatively easier to manage digital charts.
  - **Improved clinician satisfaction**: Electronic technology can save time by reducing the paperwork burden, which can create additional time for patient encounters and delivery of care . A study reports the use of EHR has reduced the physician's office visit time by 13% and a nurse's pre-exam interview time by 1 minute . This can improve satisfaction for professionals, which can indirectly enhance revenue.

Additional Benefits

- **Improved accuracy** of diagnosis and care:
- EHR provides **comprehensive** and **accurate** patient information to physicians that can help to quickly and systematically identify the correct problem to treat.
- EHRs do not just contain the patient information; they have the capability to **perform computation and make suggestions**. They can also present **comparative results** of the standard measurements.
- A U.S. national survey of doctors demonstrates the following:
  – 94% of the providers report EHR makes records readily available at the point of care.
  – 88% report that EHR produces clinical benefits for their practice.
  – 75% report that EHR allowed them to deliver better patient care.
- The gathered information can guide a physician in the emergency department to take prudent
  and safer actions. Such services are unimaginable with paper-based systems.
- Diagnostic errors are difficult to detect and can be fatal to a patient. A new study suggests that EHR can help to identify potential diagnostic errors in primary care by using certain types of queries (triggers).

- **Improved quality and convenience of care**: EHRs have the potential to improve the quality of care by embedding options such as Clinical Decision Support (CDS), clinical alerts, reminders, etc. Research suggests that EHRs are linked to better infection control, improved prescribing practices, and improved disease management in hospitals.
- In such applications, convenience is also an important measure. EHRs greatly reduce the need for patients to fill out similar (or even sometimes the same) forms at each visit.
- Patients can have their e-prescriptions ready even before they leave the facility and can be electronically sent to a pharmacy. Physicians and staff can process claims insurance immediately.
- Following are the results of a study on the effects of e-prescribing reports.
  – 92% patients were happy with their doctor using e-prescribing.
  – 90% reported rarely or only occasionally having prescriptions not ready after going to the pharmacy.
  – 76% reported e-prescribing made obtaining medications easier.
  – 63% reported fewer medication errors.

- **Improved patient safety**: Just like improving the quality of care, clinical decision support systems (CDSS) and computerized physician order entry (CPOE) have the potential to improve patient safety. ***Medication errors are common medical mistakes and in the United States it is responsible for the death of a person every day on average as well as injuring more than a million annually.***
- Research shows that utilization of CPOE can **reduce medication errors**. Medication errors can occur at any stage of the medication administration process from a physician ordering the drug, followed by the dispensing of the drug by the pharmacist, and finally the actual administration of the drug by the nurse.
- CPOE is a technology that allows physicians to act on a computerized system that introduces structure and control.
- Along with patient information, EHR holds the medication records for a patient. Whenever a new medication is prescribed, it can check for potential conflicts and allergies related to the particular medication and alert the physician.
- The system also can provide the chemical entities present in the drug and cross-reference allergies, interactions, and other possible problems related to the specific drug. Introducing technologies such as **Barcode** Medication Administration can make the system even more accurate.
-

- **Improved patient education and participation**: In an EHR system, certain features can provide simplified patient education.
- EHRs can be used by the provider as a tool to illustrate procedures and explain a patient's conditions. It can increase a patient's participation by offering follow-up information, self-care instructions, reminders for other follow-up care, and links to necessary resources.
- Information technology affects every part of our life. In this digital era, patients may feel more comfortable with an electronic system.

- **Improved coordination of care**: EHRs are considered essential elements of care coordination. The National Quality Forum defines care coordination as the following : "*Care coordination is a function that helps ensure that the patient's needs and preferences for health services and information sharing across people, functions, and sites are met over time. Coordination maximizes the value of services delivered to patients by facilitating beneficial, efficient, safe and high-quality patient experiences and improved healthcare outcomes."*
- For a patient with multiple morbidities, a physician is responsible for providing primary care services and coordinating the actions of multiple subspecialists. According to a Gallup poll , it is a common scenario for older patients to have multiple doctors: no physician 3%, one physician 16%, two physicians 26%, three physicians 23%, four physicians 15%, five physicians 6%, and six or more physicians 11%.

- EHRs allow all clinicians to document services provided and access up-to-date information about their patient.
- It streamlines the transition process and knowledge sharing between different care settings. This facilitates an improved level of communication and coordination. Research suggests that the clinicians having 6+ months use of EHRs reported better accessing and completeness of information than clinicians without EHRs.
- Clinicians having EHRs have also reported to be in agreement on treatment goals with other involved clinicians.

- **Improved legal and regulatory compliance**: As organizations develop their systems, it is important to understand and comply with many federal, state, accreditation, and other regulatory requirements.
- A health record is the most important legal and business record for a healthcare organization. The use of an EHR system will provide more security and confidentiality of a patient's information and thus, comply with regulations like HIPAA, Consumer Credit Act, etc. Moreover, the Center for Medicare and Medicaid Services (CMS) has financial incentive programs for hospitals regarding the meaningful use of health information technology.
- To receive the financial reimbursement, professionals have to meet a certain criteria and can get up to $44,000 through Medicare EHR Incentive Program and up to $63,750 through the Medicaid EHR Incentive Program. Adaptation of certified EHR can help providers get reimbursed.
- **Improved ability to conduct research and surveillance**: In conjunction with the direct use of EHR in primary patient care, there is an increasing recognition that secondary use of EHR data can provide significant insights.
- Using quantitative analysis of functional values, it has the potential to identify abnormalities and predict phenotypes. Pakhomov et al. demonstrated the use of text processing and NLP to identify heart failure patients. EHR data can be used to predict survival time of patients. Data from different EHRs can be integrated into a larger database and geo-location specific surveillance

- **Improved aggregation of data and interoperability**: Standards play a crucial role in data aggregation and interoperability between different systems.
- EHRs maintain standard procedure and follow defined coding system while collecting data. This accommodates easier aggregation of data and greater interoperability, which offer the following benefits.
  – Manage increasingly complex clinical care
  – Connect multiple locations of care delivery

- – Support team-based care
  – Deliver evidence-based care
  – Reduce errors, duplications, and delay
  – Support ubiquitous care
  – Empower and involve citizens
  – Enable the move to the Personal Health Paradigm
  – Underpin population health and research
  – Protect patient privacy

Barriers to Adopting EHR

- Despite of having great potential of EHRs in medical practice, the adoption rate is quite slow and faces a range of various obstacles.
- Many other developed countries are doing far better than the United States. Four nations (United Kingdom, the Netherlands, Australia, and New Zealand) have almost universal use (each ~90%) of EHRs among the general practitioners. In contrast, the United States and Canada have only around 10–30% of the ambulatory care physicians using EHRs.
- Health informatics has been a high priority in other developed nations, while until recently, the degree of involvement and investment by the U.S. government in EHRs has not been significant.

- **Financial barriers**: Although there are studies that demonstrate financial savings after adopting EHRs, the reality is that the EHR systems are expensive.
- Several surveys report that the monetary aspect is one of the major barriers of adopting EHRs.
- There are mainly two types of financial costs, start-up and ongoing. A 2005 study suggests that the average initial cost of setting up an EHR is $44,000 (ranging from a minimum of $14,000 to a maximum of $63,000) and ongoing costs average about $8,500 per provider per year. Major start-up costs include purchasing hardware and software.
- In addition, a significant amount of money is also required for system administration, control, maintenance, and support. Long-term costs include monitoring, modifying, and upgrading the system as well as storage and maintenance of health records. Besides, after the substantial amount of investment, physicians are worried that it could take up to several years for the return on the investment.

- An EHR is not the only electronic system that exists in any healthcare provider like practice management. There might be other old systems that also need integration into the new system. It is important that an EHR system is integrated into other systems, and this integration can sometimes be very expensive. Surveys show that due to the high financial investment required, EHR adaptation was far higher in large physician practices and hospitals.
- **Physician's resistance:** To adopt EHRs, physicians have to be shown that new technology can return financial profits, saves time, and is good for their patients' well-being. Although research-based evidence is available, it is difficult to provide concrete proof of those benefits. As given in a report by Kemper et al., 58% of physicians are without any doubt that EHR can improve patient care or clinical outcomes. Finally, adopting EHRs in a medical practice will significantly change the work processes that physicians have developed for years. Besides, physicians and staffs might have insufficient technical knowledge to deal with EHRs, which leads them to think EHR systems are overly complex. Many physicians complain about poor follow-up services regarding technical issues and a general lack of training and support from EHR system vendors.
- A study reports that two-thirds of physicians expressed inadequate technical support as a barrier to adopting EHRs . Some physicians are also concerned about the limitation of EHR capabilities.
- Under certain circumstances or as time passes, the system may no longer be useful. Besides, all physicians do not perform the same operations. EHR systems have to be customizable to best serve each purpose. Surveys suggest that one of the reasons for not adopting EHRs is that the physicians cannot find a system that meets their special requirements.
- However, an increased effort and support from vendors may play a role in motivating physicians towards adopting EHRs.

- **Loss of productivity**: Adoption of an EHR system is **a time-consuming** process. It requires a notable amount of time to select, purchase, and implement the system into clinical practice. During this period physicians have to work at a reduced capacity. Also, a significant amount of time has to be spent on learning the system. The improvement will depend on the quality of training, aptitude, etc. The fluent workflow will be disrupted during the transition period, and there will be a temporary loss of productivity.
- **Usability issues**: EHR software needs to be user-friendly. The contents of the software must be well-organized so that a user can perform a necessary operation with a minimal number of mouse clicks or keyboard actions. The interface of software workflow has to be intuitive enough. In terms of usability, a comprehensive EHR system may be more complex than expected. It has to support all the functionalities in a provider's setting. There might be a number of modules and submodules, so the user might get lost and not find what he is looking for.
- This has the potential to hamper clinical productivity as well as to increase user fatigue, error rate and user dissatisfaction. Usability and intuitiveness in the system do not necessarily correlate to the amount of money spent. The Healthcare Information and Management Systems Society (HIMSS) has an EHR usability task force. A 2009 survey by the task force reported 1,237 usability problems, and the severity of 80% of them was rated "High" or "Medium".
- Apart from the workflow usability issue, other related issues are configuration, integration, presentation, data integrity, and performance. The task force defined the following principles to follow for effective usability: simplicity, naturalness, consistency, minimizing cognitive load, efficient interactions, forgiveness and feedback, effective use of language, effective information presentation, and preservation of context.

- **Lack of standards**: Lack of uniform and consistent standards hinders the EHR adoption. Standards play an integral role in enabling interoperability. CMS reimbursement for meaningful use requires EHR systems to demonstrate the ability to exchange information. Many of the currently used systems have utility only for certain specific circumstances. Different vendors have developed systems in different programming languages and database systems.
- They do not have any defined best practice or design patterns. This makes the data exchange difficult or impossible between the systems. This lack of standardization limits the proliferation of EHRs.
- While large hospital systems have moved to EHRs, many others are skeptical about the available systems. They fear that the EHR software they buy now might not work with standards adopted by the healthcare industry or mandated by the government later on.

- **Privacy and security concerns:** Health records contain personal, diagnostics, procedures, and other healthcare related sensitive information.
- Due to the immense importance of this information, an EHR system may be subjected to attack. Some of the medical diagnoses are considered socially stigmatized, like sexually transmitted disease. Some information relates to direct life threats, like allergies. Employers as well as insurance companies may be interested to know more about a patient to make unethical decisions whether to cover a patient and/or his specific diagnosis. It can also influence some of the hiring decisions.
- EHRs contain information like social security numbers, credit card numbers, telephone numbers, home addresses, etc., which makes EHRs attractive target for attackers and hackers. A patient might even be motivated to alter his or her medical records to get worker's compensation or to obtain access to narcotics. Therefore, it is important that the privacy and security of EHRs are well maintained.
- The most used certification for privacy and security is given by the Certification Commission for Healthcare Information Technology (CCHIT). The CCHIT website claims that by mid-2009, 75% of EHR products in the marketplace were certified.

Student Section

- In addition to that, the Health Information Technology for Economic and Clinical Health (HITECH) Act introduced a new certification process sponsored by the Office of the National Coordination for Health Information Technology (ONC) in 2009. In January 2010, the ONC released the interim final rule that provides an initial set of standards, implementation specifications, and certification criteria of EHR technology. Its requirement includes database encryption, encryption of transmitted data, authentication, data integrity, audit logs, automatic log off, emergency access, access control, and account of HIPPA release of information. Physicians doubt the level of security of patients' information and records. According to Simon et al., physicians are more concerned about this issue than patients.
- The inappropriate disclosure of information might lead to legal consequences. Testing the security of HER products, a group of researchers showed that they were able to exploit a range of common code-level and design-level vulnerabilities of a proprietary and an open source EHR.
- These common vulnerabilities could not be detected by 2011 security certification test scripts used by CCHIT. EHRs pose new challenges and threats to the privacy and security of patient data. This is a considerable barrier to EHRs proliferation.
- However, this risk can be mitigated by proper technology, and maintaining certified standards with the software and hardware components.

- **Legal aspects**: Electronic records of medical information should be treated as private and confidential. Various legal and ethical questions obstruct adoption and use of EHRs. The legal system that relies on the paper-era regulations does not offer proper guidance regarding the transition to EHRs. EHRs may increase the physicians' legal responsibility and accountability.
- With computer-based sophisticated auditing, it is easy to track what individuals have done. The documentation is comprehensive and detailed in EHRs. It can both defend and expose physicians regarding malpractice. According to a *Health Affairs* article, malpractice costs around $55 billion in the United States, which is 2.4% of total healthcare spending.
- A 2010 research reveals that it was unable to determine whether the use of EHR increases or decreases malpractice liability overall. HIPAA's privacy standards also present reasonable barriers to EHR adaptation.

Challenges of Using EHR Data

- A vast amount of data is being collected every day, the secondary use of EHR data is gaining **increased attention in research** community to discover new knowledge.
- The main areas of use are clinical and transitional research, public health, and quality measurement and improvement.
- Using the EHR data, we can conduct both patient-oriented and public health research. EHR data can be used for the early detection of epidemics and spread of diseases, environmental hazards, promotes healthy behaviors, and policy development. The integration of genetic data with EHRs can open even wider horizons.
- But the data does not automatically provide us the knowledge. The quality and accuracy of the data is an issue to be taken care of. Beyley et al. presents an excellent survey of the challenges posed by the data quality.

- Incompleteness: Data incompleteness or missingness is a widespread problem while using EHR data for secondary purpose.
- Missing data can limit the outcomes to be studied, the number of explanatory factors to be considered, and even the size of population included. Incompleteness can occur due to a lack of collection or lack of documentation .The following reasons for inaccurate reporting by professionals.
  – Unaware of legal requirements
  – Lack of knowledge of which diseases are reportable
  – Do not understand how to report
  – Assumption that someone else will report
  – Intentional failure for privacy reasons
- A pancreatic malignancies study at the Columbia University Medical Center found that 48% of the patients had corresponding diagnoses or disease documentation missing in their pathology reports.
- Also report a significant amount of key variables missing (see Table 2.1).
- Patients' irregularity of communicating with the health system can also produce incompleteness.
- Based on the application in hand, type of data and proportion of data that is missing, certain strategies can be followed to reduce the missingness of data.

- TABLE: Percentage of Incompleteness of Variables in a Pancreatic Malignancies Study

Variables	Endocrine
Necrosis	20%
Number of Mitoses	21%
Lymph Node Metastasis	28%
Perineural/Lymphovascula Invasion	15%
Differentiation	38%
Size	6%
Chronic Pancreatitis	14%
Smoking—Alcohol	27%–29%
History of Other Cancer	35%
Family History of Cancer	39%
Tumor Markers	46%

Source: Taken from Botsis et al. [93].

- Erroneous Data: EHR data can be erroneous as well. Data is collected from different service areas, conditions, and geographic locations. Data is collected by busy practitioners and staff.
- Therefore, the data can be erroneous due to human errors. Faulty equipment can also produce erroneous data. Validation techniques should be used to both identify and correct erroneous data.
- Both internal and external validation measures can be applied. Internal validation is a way to check the believability of the data, e.g., unrealistic blood pressure, BMI values, etc.
- Dates can be used to check whether the result generated before a test has taken place.
- External validation includes comparing the data with other patients or historical values.

- Uninterpretable Data: The captured EHR data might be uninterpretable to a certain extent.
- It is closely related with data incompleteness. It may occur when some part of the data is captured but the rest is missing.
- For example, if a specific quantitative or qualitative measurement unit is not provided with the result value, it will be difficult to interpret.
- Inconsistency: Data inconsistency can heavily affect the analysis or result. Data collection technologies, coding rules, and standards may change over time and across institutions, which may contribute to inconsistency. For multi-institutional studies this issue might be common, especially because different healthcare centers use different vendors for providing apparatus, softwares, and other technologies.
- A study in Massachusetts of 3.7 million patients found that 31% of patients have visited two or more hospitals in the course of five years.

- Unstructured Text: In spite of having many defined structures for collecting the data, a large portion of the EHR data contain unstructured text.
- These data are present in the form of documentation and explanation.
- It is easy to understand them for humans, but in terms of automatic computational methods, detecting the right information is difficult.
- Sophisticated data extraction techniques like Natural Language Processing (NLP) are being used to identify information from text notes.
- Selection Bias: In any hospital, the patient group will mostly be a random collection.
- It varies depending on the nature of practice, care unit, and the geographical location of the institution.
- It will not contain the diversity of demography.

- Interoperability: Lack of EHR interoperability is a major impediment towards improved healthcare, innovation, and lowering costs.
- There are various reasons behind it. EHR software from commercial vendors are proprietary and closed systems.
- Most software were not built to support communication with a third party and developing new interfaces for that purpose might be a costly undertaking. Absence of standard also contributes to the problem. Many patients are not lenient towards sharing their information.
- Besides EHR systems must comply with the HIPAA Act to ensure the security and privacy of the data.

Phenotyping Algorithms

- Phenotyping algorithms are combinations of multiple types of data and their logical relations to accurately identify cases (disease samples) and controls (non-disease samples) from EHR as illustrated in Figure 2.3.
- Based on the structure, EHR data can be broadly divided into two parts, structured and unstructured data.
- Structured data exists in a name–value pair while unstructured data contains narrative and semi-narrative texts regarding descriptions, explanation, comments, etc.
- Structured data include billing data, lab values, vital signs, and medication information.
- Billing and diagnosis-related data are collected using various coding systems like ICD, CPT, and SNOMEDCT.
- These codes are important parts of the phenotyping process. ICD codes generally have high specificity but low sensitivity. Table 2.2 lists different characteristics of EHR data.

- The primary purpose of EHR data is to support healthcare and administrative services.
- Information is produced as a byproduct of routine clinical services. They are not a suitable format for performing research tasks. They often require further processing to be used for phenotyping algorithms.
- Within existing EHR systems, querying for a particular diagnosis or lab test across all patients can be a not-trivial task. An EHR can quickly pull the information related to a patient's current medications, and easily find any test results. But combining different data with a temporal relationship might require manual processing of data.
- From clinical operational settings, data are often extracted and reformatted to make them more convenient and suitable for doing research, typically storing them in relational databases.
- Researchers have created a number of Enterprise Data Warehouses (EDWs) for EHR data. Examples include Informatics for Integrating Biology and the Bedside (i2b2) , the Utah Population Database, Vanderbilt's Synthetic Derivative, etc.
- Commercial EHR vendors are also developing research repositories. For example, EPIC users can add the "Clarity" module to their system, which will convert the EHR data into SQL-based database for research purposes.

- To build a phenotype algorithm, first we need to select the phenotype of interest, followed by the identification of key clinical elements that define the phenotype.
- It may contain billing codes, laboratory and test results, radiology reports, medication history, and NLP-extracted information. The gathered information may be combined with a machine learning method.
- For example, in, the authors have applied Support Vector Machine (SVM) to a both naive and well-defined collection of EHR features to identify rheumatoid arthritis cases.
- A medication record can be used to increase the accuracy of case and control identification of phenotyping algorithms. Patients who are believed to be controls must be having a different medication profile. They may not even have any medications prescribed to them at all.
- Sufficient dosage of a particular medication serves the confirmation that a person is having the disease of interest. For example, a patient treated with either oral or injectable hypoglycemic agents will be having diabetes. These medications are highly sensitive and specific for treating diabetes.

- Studies have shown that CPT codes can accurately predict an occurrence of a given procedure.
- The standard terminology codes for lab tests are LOINC. On the other hand, clinical notes are in free-text format. To be used for phenotyping algorithms, it has to undergo subsequent text processing.
- Certain procedures and test results may also exist in a combination of structured and unstructured form.
- For example, an electrocardiogram report typically contains structured interval durations, heart rates, and overall categorization, along with a narrative text of cardiologist's interpretation of the result.
- Recently, researchers have been linking EHR data with biological databanks (biobanks).
- The most popular biobanks are the collection of DNA samples.

- Hospitals and clinics can collect DNA samples from a patient's blood sample that is used in routine tests. The Personalized Medicine Research Population (PMRP) project in Marshfield Clinic has a biobank of 20,000 individuals.
- Similar DNA biobanks exist at eMERGE Network sites, Northwestern University, Geisinger Health System, Mount Sinai School of Medicine, and at other places. The eMERGE network is funded and organized by the National Human Genome Research Institute (NHGRI) and until today it has created and validated twenty-one EHR-derived phenotyping algorithms (see Table 2.3).

- Its mission is to develop, disseminate, and apply methods to combine DNA biorepositories and EHR systems for large scale and high throughput genetic research.
- But the phenotype information extracted from EHRs may be challenging. Validation of phenotypes is important before integration of EHRs into genetic studies. By validating EHR-derived phenotypes from eMERGE network, Newton et al. report the following points:
  • Multisite validation improves phenotype algorithm accuracy
  • Targets for validation should be carefully considered and defined
  • Specifying time frames for review of variables eases validation time and improves accuracy
  • Using repeated measures requires defining the relevant time period and specifying the most
  meaningful value to be studied
  • Patient movement in and out of the health plan (transience) can result in incomplete or fragmented data
  • The review scope should be defined carefully
  • Particular care is required in combining EMR and research data
  • Medication data can be assessed using claims, medications dispensed, or medications prescribed
  • Algorithm development and validation will work best as an iterative process
  • Validation by content experts or structured chart review can provide accurate results

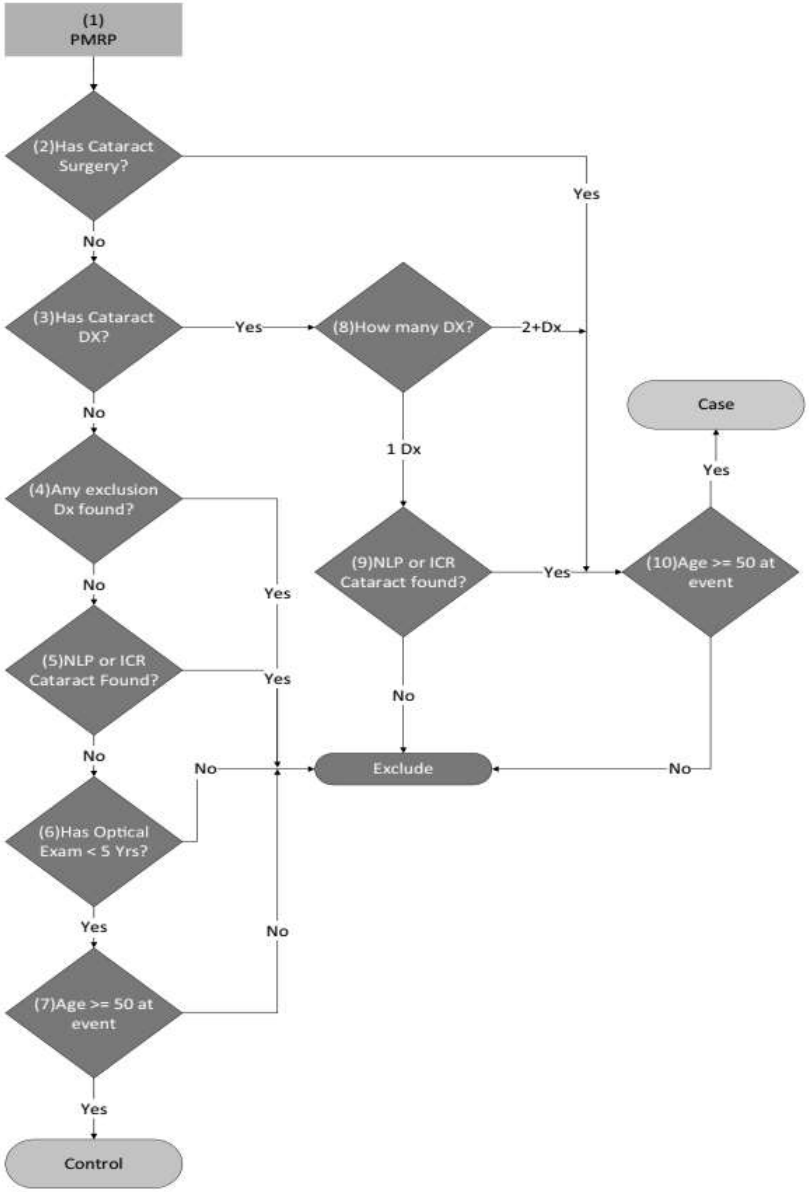

FIGURE 2.3: Flowchart for cataracts phenotyping algorithm taken from [98].

TABLE 2.2: Characteristics of Different EHR Data

	ICD	CPT	Lab	Medication	Clinical notes
Availability	High	High	High	Medium	Medium
Recall	Medium	Poor	Medium	Inpatient: High Outpatient: Variable	Medium
Precision	Medium	High	High	Inpatient: High Outpatient: Variable	Medium/High
Format	Structured	Structured	Mostly	Structured	Structured
Pros	Easy to work with, good approximation of disease status	Easy to work with, high precision	High data validity	High data validity	More details about the doctors' thoughts
Cons	Disease code often used for screening, therefore disease might not be there	Missing data	Data normalization and ranges	Prescribed not necessarily taken	Difficult to process

Source: Taken from Denny [106].

TABLE 2.3: Phenotyping Algorithms Developed by eMERGE Network

Phenotype	EHR data used to characterize phenotype	Institution
Atrial Fibrillation — Demonstration Project	CPT Codes, ICD 9 Codes, Natural Language Processing	Vanderbilt University
Cardiac Conduction(QRS)	CPT Codes, ICD 9 Codes, Laboratories, Medications, Natural Language Processing	Vanderbilt University
Cataracts	CPT Codes, ICD 9 Codes, Medications, Natural Language Processing	Marshfield Clinic Research Foundation
Clopidogrel Poor Metabolizers	CPT Codes, ICD 9 Codes, Laboratories, Medications, Natural Language Processing	Denny's Group at Vanderbilt, VESPA — Vanderbilt Electronic Systems for Pharmacogenomic Assessment
Crohn's Disease — Demonstration Project	ICD 9 Codes, Medications, Natural Language Processing	Vanderbilt University
Dementia	ICD 9 Codes, Medications	Group Health Cooperative
Diabetic Retionapathy	CPT Codes, ICD 9 Codes, Laboratories, Medications, Natural Language Processing	Marshfield Clinic Research Foundation
Drug Induced Liver Injury	ICD 9 Codes, Laboratories, Medications, Natural Language Processing	Columbia University
Height	ICD 9 Codes, Laboratories, Medications	Northwestern University
High-Density Lipoproteins (HDL)	ICD 9 Codes, Laboratories, Medications, Natural Language Processing	Marshfield Clinic Research Foundation
Hypothyroidism	CPT Codes, ICD 9 Codes, Laboratories, Medications, Natural Language Processing	Vanderbilt University, Group Health Cooperative, Northwestern University
Lipids	ICD 9 Codes, Laboratories, Medications	Northwestern University
Multiple Sclerosis — Demonstration Project	ICD 9 Codes, Medications, Natural Language Processing	Vanderbilt University
Peripheral Arterial Disease	CPT Codes, ICD 9 Codes, Laboratories, Medications, Natural Language Processing	Mayo Clinic
Red Blood Cell Indices	CPT Codes, ICD 9 Codes, Laboratories, Medications, Natural	Mayo Clinic

- Before the use of a phenotyping algorithm, data has to be normalized to standard representation. Natural Language Processing (NLP) based tools have gained much popularity to extract structured information from free text.
- Several studies have shown that coded data are not sufficient or accurate to identify disease cohorts.
- Information from narrative text complements the structured data.
- There are studies that report NLP-processed notes provide more valuable data sources. For example, Penz et al. reports ICD-9 and CPT codes identified less than 11% cases in detecting adverse events related to central venous catheters, while NLP methods achieved a specificity of 0.80 and sensitivity of 0.72.
- Widely used general-purpose NLP tools include MedLEE (Medical Language Extraction and Encoding System) , cTAKES (clinical Text Analysis and Knowledge Extraction System), MetaMap, and KnowledgeMap.
- All of them have been successfully applied to phenotyping using EHR data. Task-specific NLP methods are available that aim to extract specific concepts from clinical text. The DNA sequence of a person can be huge in size (ranging from hundreds of gigabytes to terabytes) in raw format that exceeds the capability for using the current EHR systems.

- Storing, managing, and transferring a repository of such a large volume of data is difficult.
- Efficient data compression techniques can be applied to solve this problem. Genome Wide Association Study (GWAS) became the mainstay of genetic analysis over the last decade.
- In general, GWAS investigates around 500,000 genetic variants (Single Nucleotide Polymorphisms) or more to see the association of variations with observable traits. It compares the SNPs of cases versus controls to find meaningful knowledge.
- Besides traits, we can also identify SNPs that determine a particular drug response. One individual might react adversely to a particular drug while others might not.
- The genetic profile of an individual can be used for personalized medicine. One big advantage of genetic data is that the SNPs are the same for that individual and do not change based on a given/suspected disease. The same set of data can be used for different phenotype investigations as well. Researchers are working to integrate genetic information for enhanced clinical decision support. For example, researchers in Vanderbilt University are working on implementing Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment (PREDICT) . St. Jude Children's Research Hospital also has a multiplexed genotyping platform for providing decision support

Question ?