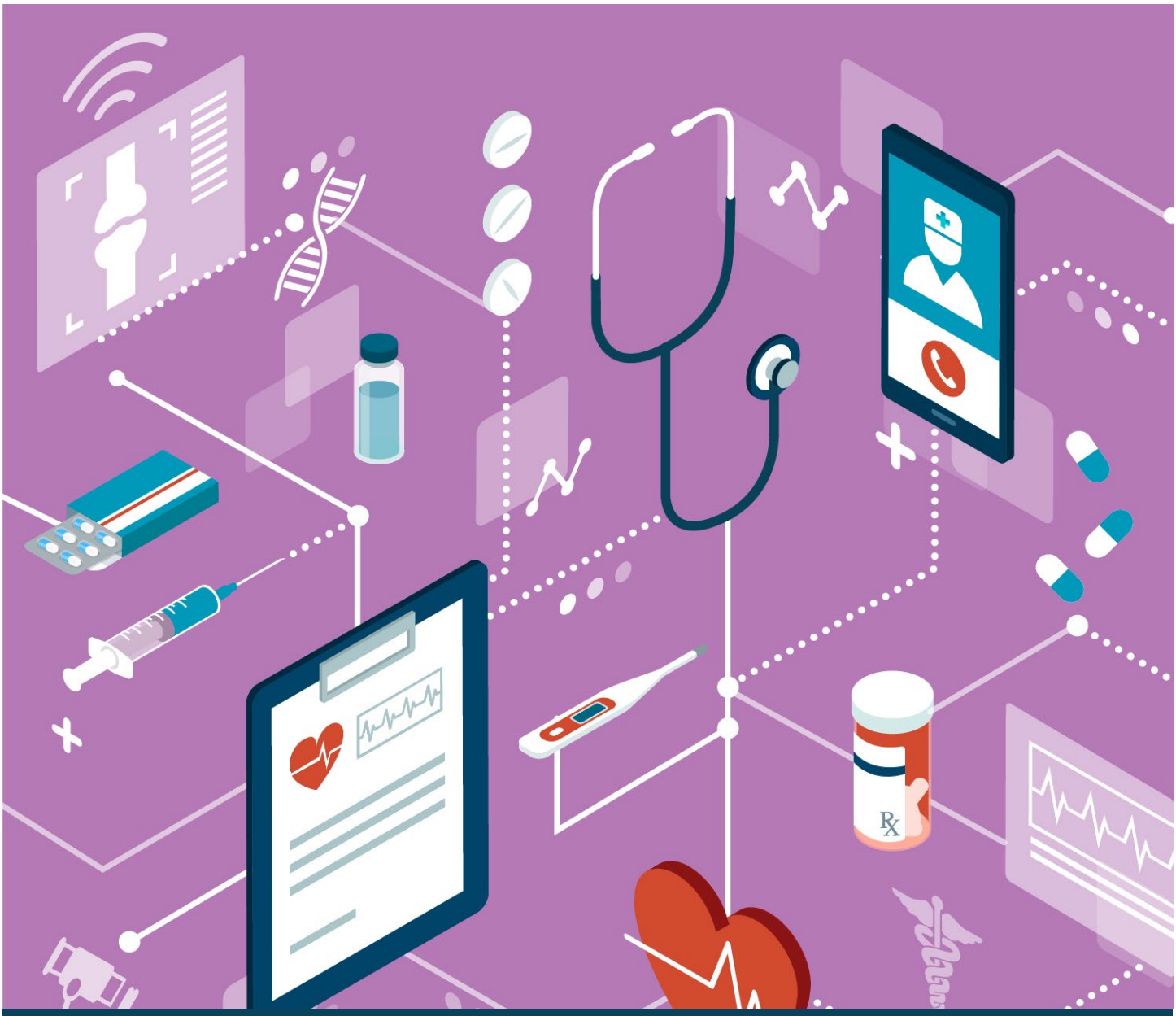


Introduction to Clinical Data

Study Guide

CONTENTS

MODULE 1 – ASKING AND ANSWERING QUESTIONS VIA CLINICAL DATA MINING

LEARNING OBJECTIVES

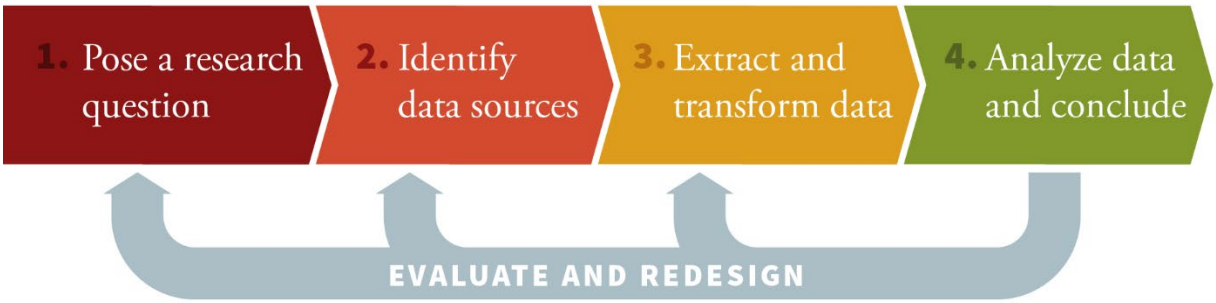
1. Explain the main steps in the data mining workflow
2. Describe the important categories of research questions
3. List properties that make a research question a useful one

THE DATA MINING WORKFLOW

Goal in this course is to explain how clinical data can be used to answer research questions to improve the health of patients and populations.

What we'll cover

1. How to choose research questions that are important
2. Structure of the healthcare system to understand how (and what) patient data are generated
3. Different kinds of data
4. Overview of the processing and analysis methods that get us answers to our questions
5. Problems and biases that can arise as well as ways to manage them



Data mining work will be referred in this course which has four steps:

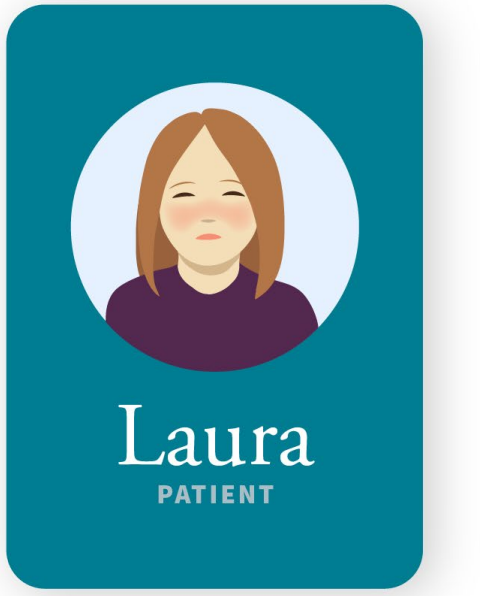
1. Pose a research question.
2. Identify one or more data sources that can answer the question.
3. Extract and transform the data into a form needed for the analysis.
4. Conduct the analysis using those data

After completing the steps, results are evaluated and repeat the process if necessary.

Two representations of healthcare data that we will focus on this course are:

1. a patient timeline
2. a patient-feature matrix.

REAL LIFE EXAMPLE



Laura

- A teenager with a chronic disease called Systemic Lupus Erythematosus (SLE).
- Has a flare up of the condition and develops proteinuria (protein in the urine), pancreatitis (inflammation of the pancreas), and has antiphospholipid antibodies in her blood.
- She is at risk for developing a blood clot.

Step 1 in the data mining workflow:

- Our clinical question:

  "Should a teenager with SLE who develops proteinuria and antiphospholipid antibodies receive an anticoagulant medication?"

  (X) Review medical literature

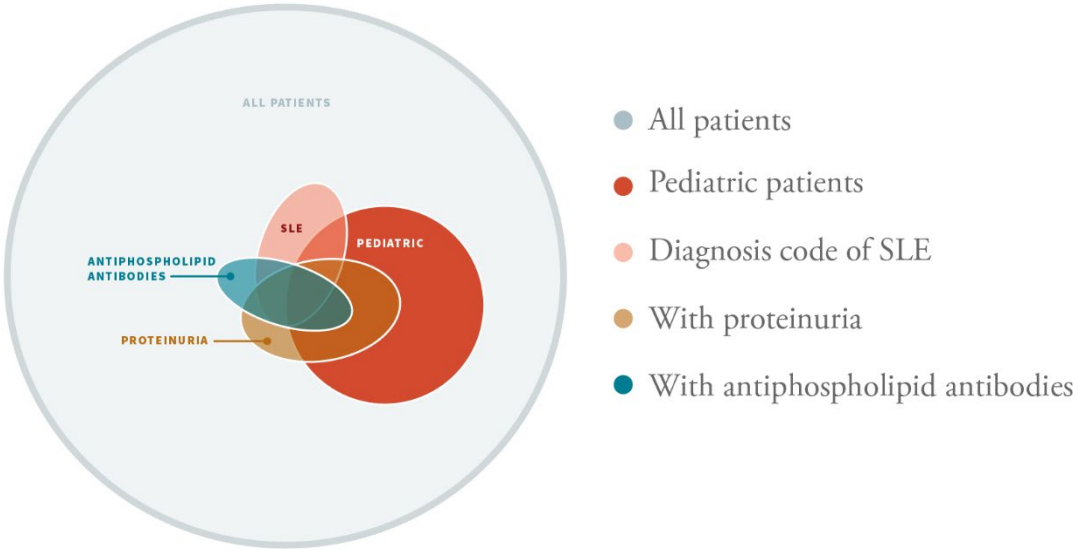
  (X) Past experience

  (X) Consult experts

- One approach is to examine what has happened to similar patients in the past, drawing on all the relevant data that appears in the electronic medical record, or EMR, of a large academic medical center.

Now we identified our data source, so we have completed Step 2 in the workflow.

EMR data are not necessarily organized in a way that makes searches straightforward. Medical expertise is needed to choose the diagnosis codes and medical terms that can identify patients that are in a similar situation.



Steps we would need to take.

1) Find all pediatric patients in the medical record system. Using a query based on patient age.
2) Diagnosis code of SLE.
3) Find patients with proteinuria by checking on the values listed in a urine test.
4) Find patients with antiphospholipid antibodies.
5) Laboratory test
    a) Recorded in numeric form
    b) Result in textual document; Use treatment with aspirin as a proxy marker
    c) Confirm the results of that search by checking laboratory results for those antibodies for those patients who have the results in a searchable form

6) Now in the midst of **Step 3**; we have defined our criteria that allows to identify the appropriate group of similar patients

7) Outcome of interest is clotting

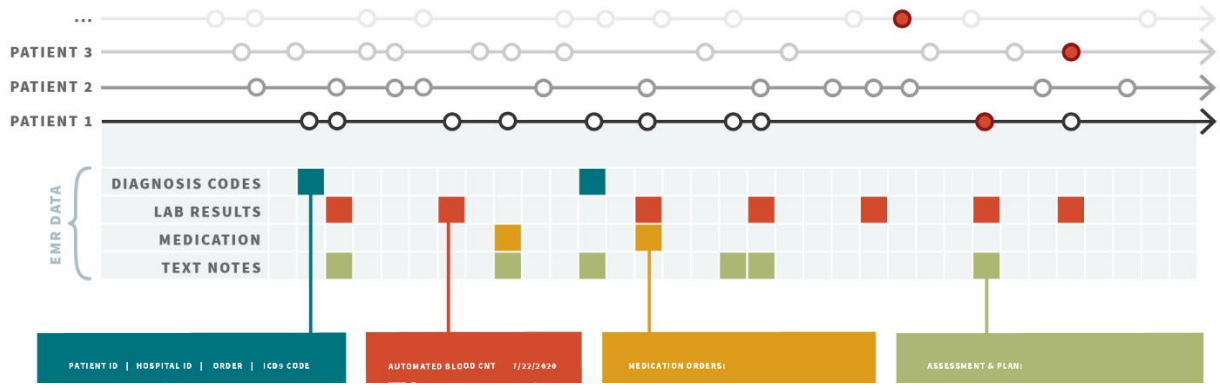
    a) Search for "thrombus", "thrombosis", and "blood clot", again relying on clinical expertise to choose those terms.

Analysis is straightforward: compare the risk of clotting in teenagers with SLE, proteinuria, and antiphospholipid antibodies, to the baseline risk of clotting in teenagers with SLE.

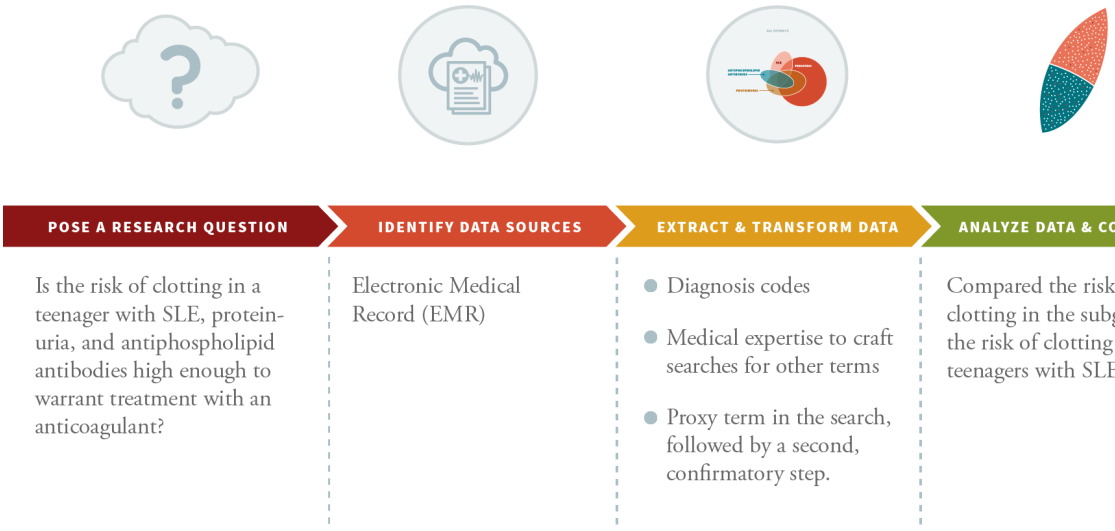
- Need to get the patient data out of the EMR system in a form that allows analysis.
- Found that the relative risk of getting a blood clot when proteinuria and antiphospholipid antibodies are present is twice as high, when compared to the baseline, thus choosing to treat with an anticoagulant.

We have now completed **Step 4** of the "data mining workflow".

Information in the form of **a patient timeline**.



- Start with data collected by the healthcare system, in this case, the electronic medical record, which includes diagnosis codes, lab results, medication orders, and text notes written by clinicians.
- Arrange these data on a timeline, one for each patient and then identify pediatric patients, and flag those with SLE, some of whom develop the comorbidities of concern (proteinuria, antiphospholipid antibodies).
- Use the timeline to count which patients experienced the outcome of clotting after they developed each clinical condition. Then compute the fraction of those with each condition who developed blood clots to arrive at the relative risk.



Revisit the **data mining workflow** steps.

1. What was the clinical question? Is the risk of clotting in a teenager with SLE, proteinuria, and antiphospholipid antibodies high enough to warrant treatment with an anticoagulant?
2. What is the data source? The electronic medical record.
3. What extract/transform steps did we take? We defined how we will find teenagers with SLE, and how we will define subgroups based on clinical conditions. This involved the use of diagnosis codes in some cases and the use of medical expertise to craft searches for other terms. In one case we used a proxy term in the search, followed by a second, confirmatory step.
4. Finally, we compared the risk of clotting in the subgroups to the risk of clotting for teenagers with SLE in order to guide our decision to treat.

In this example we primarily used one data source, the EMR. Remember that there are no "perfect" ways of doing all the steps we reviewed in the example and it is best to think of the entire data mining process as something that should be done with an expert human in the loop rather than by an automated algorithm that provides answers without knowing the larger context of the situation.

TYPES OF RESEARCH QUESTIONS

- A **descriptive question** asks for a summary of the data
- An **exploratory question** attempts to find what patterns might exist in the dataset available.

- An **inferential question** looks for patterns that go beyond just the particular dataset available. The goal is to find generalizable knowledge
- A **predictive question** looks for quantitative relationships between some features and the outcome of interest.
- A **causal question** looks for the effect of changes in one variable on a second variable.
- A **Deterministic question** is directly addressing the underlying mechanism

Clinical data are best suited for answering descriptive, exploratory, inferential, and predictive questions.

We ask these questions to accomplish two primary goals:

1. Risk stratification to decide if to treat
2. Data-driven selection of how to treat

What do you think about our analysis for Laura

The question we asked was about treating Laura who was experiencing a set of clinical conditions. However, the question we answered was about the proportion of patients with a set of clinical conditions who developed a blood clot. What we answered was a descriptive question relying on counts and proportions.

We then need to make an assumption that what happened in the past to those patients is likely to happen to Laura as well. The assumption, and the resulting conclusion, provides us with a 'risk-stratification'.

If we conclude that Laura is at high risk, what treatment to offer is clear, which is to use anticoagulation. In real life we would also need to draw similar conclusions about the risks of adverse outcomes resulting from the treatment itself before making a final decision.

What makes answering a question useful:

- How many lives are affected? What is the disease burden?
- What is the chance that results will have a beneficial effect on the target community?
- What happens as a result of answering the question?
- Does knowing the answer help more than one constituent group among patients, healthcare professionals, and payers of care?

MODULE 2 – DATA AVAILABLE FROM HEALTHCARE SYSTEMS

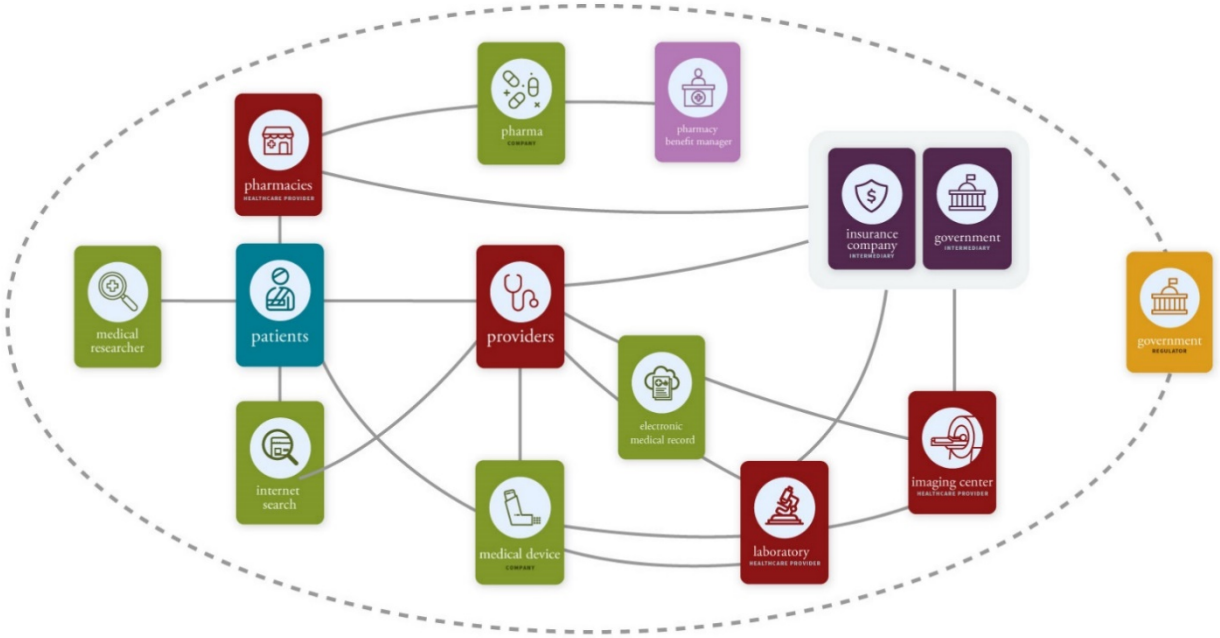
LEARNING OBJECTIVES

1. Describe the key actors in a healthcare system
2. Give examples of how different actors in the healthcare system can have different goals and interests.
3. List the different kinds of data the healthcare system produces
4. Describe the important healthcare data types
5. List pros and cons of using observational data
6. List examples of biases in observational data
7. Describe how to assess if a data source is useful

THE HEALTHCARE SYSTEM

Module goal is to show how clinical data can be used to ask and answer interesting and important research questions.

Below are the key entities in the healthcare system and how/what data each entity generates:

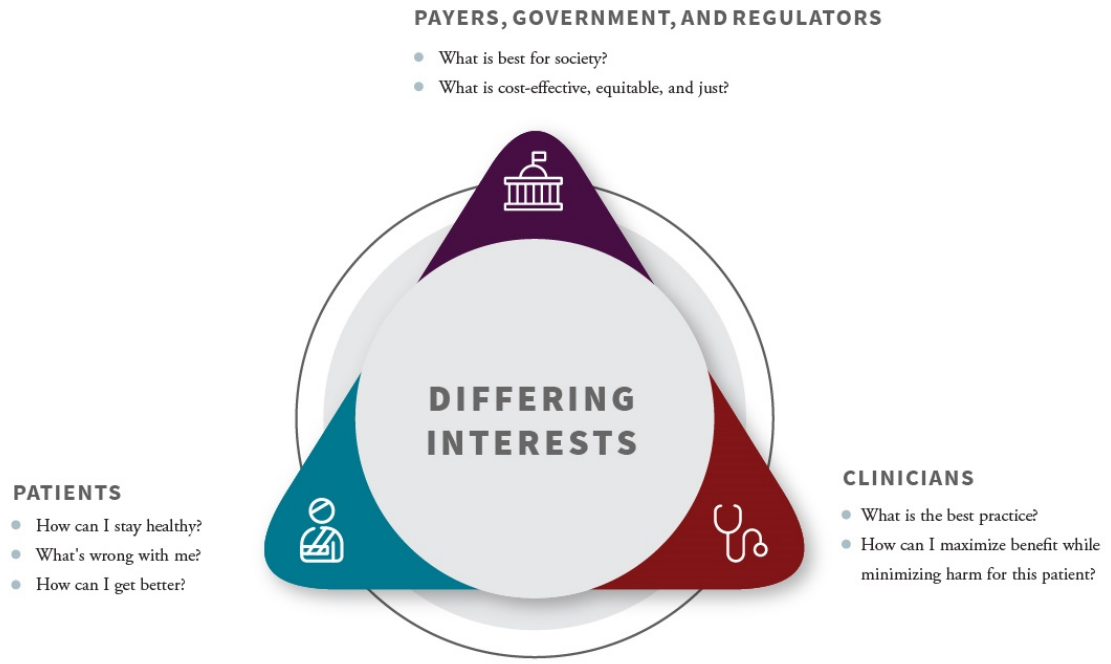


- **Patient:** If they need disease screening, or get sick, then they may seek treatment, or, more often search the internet to decide if they should seek treatment.

- **Healthcare providers:** Typically, the providers order laboratory tests or imaging procedures, make diagnoses, write prescriptions, and record their observations in the patient's electronic medical record (or EMR)
- **Pharmacies:** Provide drugs to patients
- **Pharmaceutical companies:** Design and manufacture drugs.
- **Drug distributors:** pharmacies procure drugs from this entity
- **Pharmacy benefits management companies:** manage payment for drugs
- **Medical device companies:** Design and manufacture medical equipment
- **State and federal government agencies:** Monitor and regulate the healthcare system
- **Governments:** Collect healthcare data to understand patterns of disease, which groups are underserved, and how healthcare is paid for
- **Medical researchers:** Investigate the patients, their disease conditions, and medical and surgical treatments. Publish their research results to increase scientific understanding of health and disease, and to guide the development of new drugs, devices, and other treatments. Also address policy questions on how we should best organize the provision of medical care.

The healthcare system is extremely complex, with many different types of actors, and many actors of each type. Actors can have different interests, which are not always in alignment.

The primary actors in the healthcare systems:



Different interests lead to tension in the delivery of healthcare, in the generation of data, and their use. It is important to keep these different interests in mind when defining a research question and choosing a data source to answer the question.

Think about:

- Who could benefit from answering your question
- How their interests could introduce some biases into the data sources used
- Ways to address more than one audience or interest group at the same time

HEALTHCARE DATA TYPES

Types of healthcare data:

- **Structured data:** Consistent organization; a table with rows and columns
- **Unstructured data:**
    - **Clinical text:** Quite different from ordinary written language, a haiku of acronyms
    - **Images:** Large two-dimensional arrays of intensity values, measuring the degree to which some kind of physical energy is transmitted or absorbed by tissue. Sometimes many two-dimensional images are collected together to form representations of volumes.
    - **Signals:** Measurements coming from a sensor, usually at regularly-spaced time intervals

Healthcare data vary along several dimensions: occur over different time scales, generated at different points in the patient's care journey, different possible values, different patterns of missing values

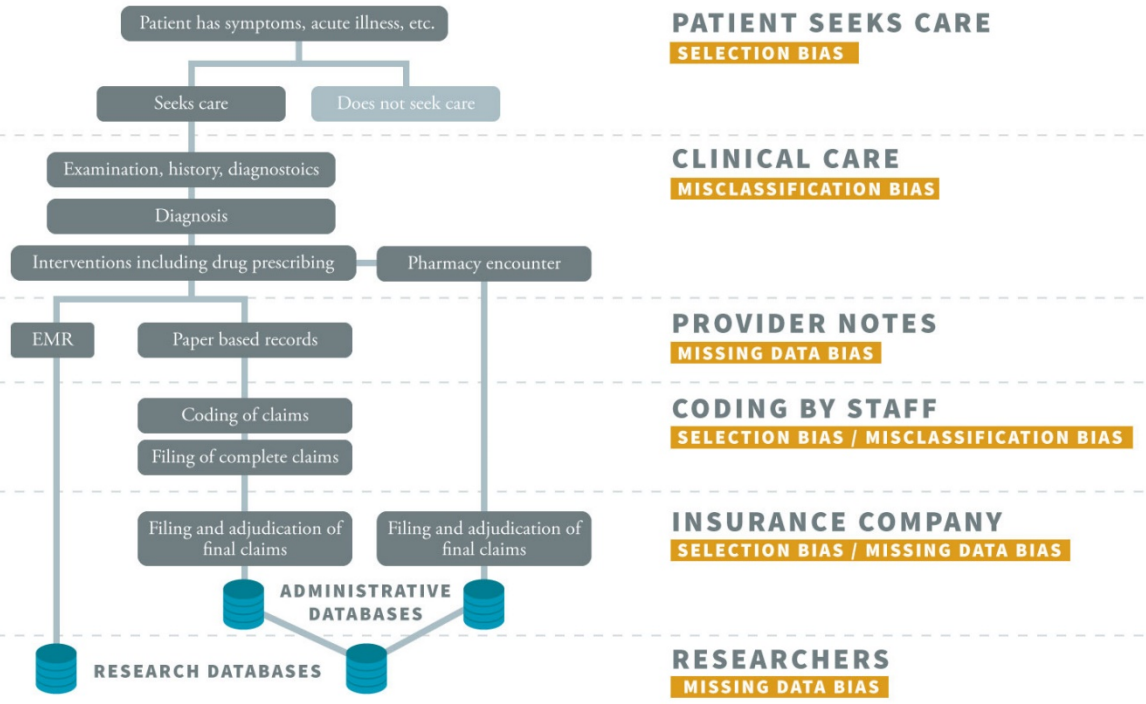
There is an **inherent time ordering** to all the different data types that are generated for a particular patient at various places in the healthcare ecosystem. Knowing this **timeline view of the data** is essential for effectively using the data.

- **Observational data:** data collected for other purposes and are collected as a byproduct of care delivery. Our use is referred to as *secondary use*.

Strengths of Observational Data	Weaknesses of Observational Data
• Very large datasets: Ability to study rare events	• Datasets are static: Difficult to obtain additional features/detail • Patient records not always linked

• Datasets are created during the routine operation of the healthcare system: Study real-world effectiveness and utilization • Available at relatively low cost without long delays: Accessible and efficient • Not subject to rapid changes in format, and the format may even be standardized	• Data creation and collection can be imperfect and biased, and may require significant cleanup • A record exists only if something (bad) happens

SOURCES OF BIASES AND ERRORS



Ways in which the data produced by each entity in the healthcare system might be inaccurate or biased:

- **Patient:** Decides to seek, or not seek, care. Many patient factors that might affect their decision, thus making this a selection bias
- **Clinical Care:** Not recording the health status of everyone, because leaving out those in their normal state of health. Not recording the health status of those who are sick but who treat themselves at home with over-the-counter medications, or those who are treated outside of

them health system for which we have records. In general, records are neither complete, nor a sample chosen at random

- **Healthcare Provider:** Might respond to financial incentives by changing how they decide whether to treat and which treatment to offer. Those incentives may also affect how they record what they did. Their records may be written at the time of service or with some delay afterwards. As a result, these records may be inaccurate or incomplete.
- **Coding:** Assign diagnosis and procedure codes to the medical documentation for the purpose of generating a bill for services rendered. A medical bill often contains only enough information to construct the bill, not a complete record of the treatment or what happened as a result of the treatment. There may be biases from systematic errors in coding. Coders may make mistakes as well as omit codes that are unlikely to be reimbursed.

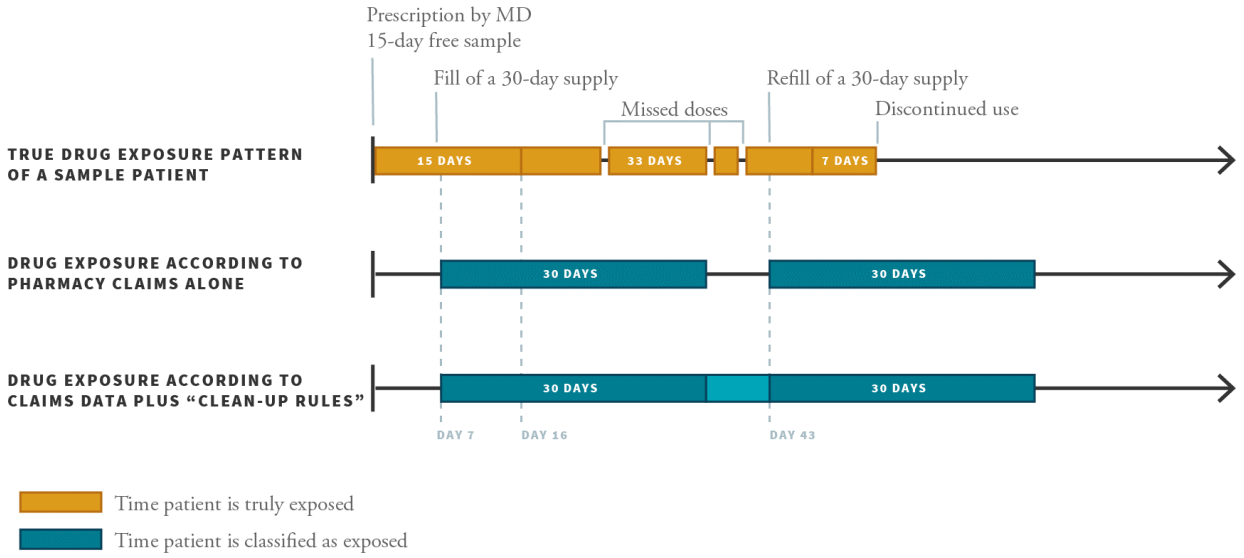
One way to overcome some of these issues is to combine data sources. However, linking records across sources can be technically challenging, and may be inaccurate or incomplete, and may itself introduce some additional biases. Also, there is a need to protect patient privacy.

Research questions require relating exposures to outcomes, which gives us a framework to identify sources of biases and possible errors in the data.

- **Exposure**: Something that could happen to the patient - diseases, drugs, or medical procedures
- **Outcome:** A condition of interest that is assessed as having occurred in the patient timeline, at some point after the exposure

Exposure Misclassified Example:

Suppose a doctor writes a prescription. The prescription is for a 30-day supply of the medication with one refill. The doctor gives the patient a 15-day free sample to cover this gap. The patient starts taking the free sample, and gets the prescription filled at their local pharmacy. Suppose that happens 7 days after the doctor visit. The patient starts the medication from the pharmacy on day 16 after the sample runs out, and continues taking it. Shortly before it runs out, the patient orders a refill, picks that up at the pharmacy, and takes it until the refill runs out. Let's say this sequence of events is what actually happened.



The time the patient is actually exposed to the drug is different from what can be inferred by classifying exposure on the basis of claims data.

We can construct "cleanup rules" that try to produce more accurate estimates by filling in the gaps. A rule might, for example, fill in small gaps, such as less than 7 days, in the record. However, these rules are not perfect.

Outcome Misclassified Example:

A common situation where a patient's outcome could be misclassified is the assignment of codes when the clinical status of a patient is not yet certain.

What can we do to reduce misclassification biases?

- Require multiple mentions of a diagnosis code for us to believe that the patient indeed has the condition
- Require that along with a diagnostic code for the condition of interest, there also must be an occurrence of a disease-specific procedure code
- Require the medication that is very specific to a disease be present in the record to conclude if the patient had that disease

Terminology

- **Electronic phenotyping**: Strategies used to determine if the outcome indeed occurred in the patient's timeline

We can validate the error rate (or misclassification rate) of these strategies for electronic phenotyping by:

- Comparing to a manual chart review of the patient's primary medical record
- Putting bounds on the net-effect of misclassification by constructing a computer simulation that adds random misclassification to the exposure or outcome measurements to see the effect on the answer to the question at hand

When using clinical data, use multiple sources because one source may protect from a weakness in another source or may help in estimating the error rate in another source.

HEALTHCARE DATA SOURCES

Sources of Data:

- Medical record of a patient: Referred to as "charts", stored in computer systems as electronic medical records (EMR) or sometimes also called the electronic health record (EHR)
- Newer kinds of EMR data available in some systems come from genetic tests as well as from consumer devices such as wireless blood pressure cuffs and activity monitors
- Delivery of clinical care produces documentation about the care which includes:
    - Progress notes written by doctors, nurses, and other clinical providers
    - Any orders written by a doctor
    - Results of laboratory tests or imaging studies

EMR Datasets:

- Medical Information Mart for Intensive Care (MIMIC)
- Cerner Health Facts

Claims Data

- Bill: Healthcare professionals collect payment for a medical service. Contains identifying information about the patient, some description of their insurance status, diagnosis and procedure codes as well as requested charges
- Insurance data include everything that insurance paid for, so these records can follow a single patient across multiple providers
- Datasets of Claims:
    - Truven Marketscan Commercial Claims and Encounters
    - Optum Clinformatics Data Mart

- The Centers for Medicare & Medicaid Services (CMS)

Pharmacies Record:

- The written prescriptions
- When prescriptions were filled
- How the prescriptions were paid for

It is evidence that goes one step beyond the act of writing the prescription. The "drug record" for a single person can be spread out over multiple pharmacy datasets.

Post-marketing surveillance: Governments and other monitoring agencies maintain databases of reported effects and side effects

- Goal: identify serious problems as soon as possible, and then possibly restrict use of the drug or device, or recall it from the market entirely

Sources of **Surveillance Data**:

- The US Food and Drug Administration (FDA)
- FDA Adverse Events Reporting System (FAERS)
- Manufacturer and User Facility Device Experience (MAUDE)
- Centers for Disease Control and Prevention (CDC)
- Registries run by professional societies, governments
  - Examples of societies: the American Board of Family Medicine, the American Society for Clinical Oncology, the American Academy of Ophthalmology
  - Examples of registries: the PRIME registry, CancerLinq, the Intelligent Research in Sight (IRIS) registry
- Registries for particular diseases or conditions:

Population Health Data: Record expenditures by treatment type, medical condition, geographic area.

- Agency for Healthcare Research and Quality (AHRQ)
- National Inpatient Sample (NIS)
- The Medical Expenditure Panel Survey (MEPS)
- The National Health and Nutritional Examination Survey (NHANES)

Patient-generated: Patients can record their own health states and they can choose to provide this information to their doctor, or make the information publicly available for study by others. Patients can report data to centralized databases about their diagnoses, symptoms, treatments, and outcomes.

Examples of patient-generated data sets include:

- Comments made on social networks organized around medical conditions and diseases
- Online portals for patient-reported information about conditions, symptoms and treatments

Researcher-generated: Research is the systematic investigation of biomedical phenomena in order to find valid and generalizable results.

- Example: Researchers can recruit patients into randomized clinical trials in order to systematically study the effectiveness of a treatment. Typically the treatment is compared to either the best existing treatment, or no treatment.
- Clinical trials are usually quite expensive, rigorously analyzed, and their results influence medical practice for a long time. However, by using random assignment to the treatment or the control groups, ***clinical trials are the most reliable source of data to answer questions about causality***.

Clinical trials in the United States have to be registered at www.Clinicaltrials.gov. The *data* from these clinical trials are increasingly becoming available for re-analysis and re-use after the completion of the trial.

There are multiple sources of healthcare data; each source capturing some aspect of what happened to a patient in their care timeline. No one source has the complete picture of the patient timeline, and using multiple sources to answer the same question will lead to more reliable answers.

Questions to ask when considering a particular data source:

1. Is there a well-documented data model?
2. Where are the data from?
3. Are the data accessible?
4. What are the known errors in the data?

Additional questions you should consider:

1. Does this dataset have the data elements corresponding to the patient characteristics that you need to observe?
2. If not, can you use other data elements as a proxy for the characteristics you really want?
3. If you are studying rare conditions, is the dataset large enough to observe those conditions in sufficient numbers?

MODULE 3 – REPRESENTING TIME, AND TIMING OF EVENTS, FOR CLINICAL DATA MINING

LEARNING OBJECTIVES

1. Explain why timelines are useful for healthcare data
2. Identify the timescales of interest in a clinical research question.
3. Identify which timescales are represented in which kinds of data
4. Explain the difference between explicit and implicit representations of time
5. Describe at least one problem arising from temporal classification of exposures and outcomes
6. Describe non-stationary and why it is important

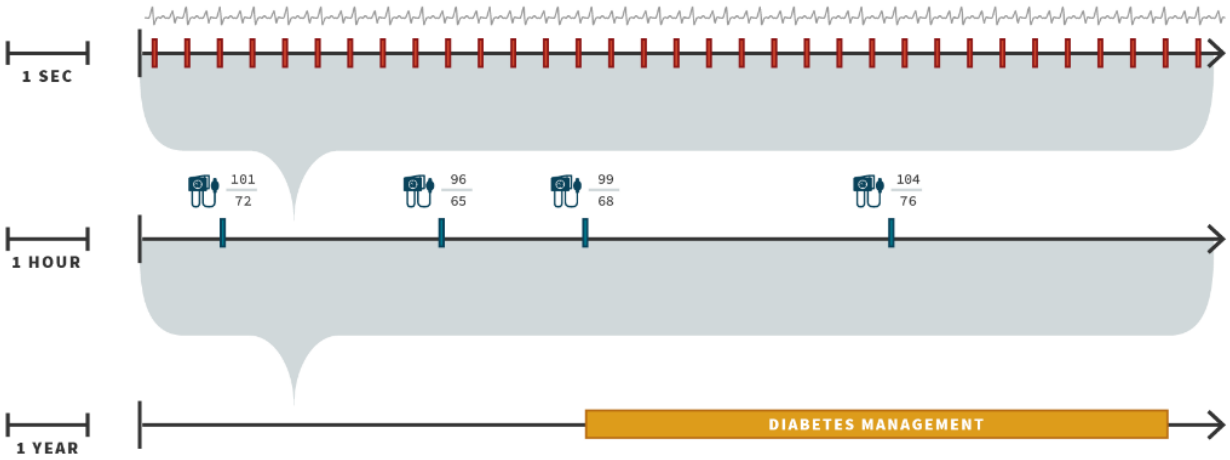
HEALTHCARE HAPPENS OVER TIME

The goal of this module is to discuss how **the patient timeline** relates to **timescales of the questions** we seek to answer as well as the issues that matter when **representing time** and working with healthcare **data that changes over time**.

A good way to integrate all the different sources and types of data for a patient is to place all the events on a timeline. Each patient will have their own timeline. The timeline explicitly captures *when* the patient experienced each event.

We care about **time** for two main reasons.

- A patient's age is a key factor in making medical care decisions about them
- We will often examine the order in which events occur



The **timescales** involved in answering medical questions span many orders of magnitude. We may need to consider intervals of well under one second, all the way to a patient's entire lifetime.

The **choice of the timescale** and the way **we choose to represent time** have to deal with the fact that, not only do patients change over time, but the healthcare system itself is evolving over time, so the meaning of the data elements we capture as a by-product of routine care can change without our noticing.

- A **stationary process** is one that generates data that looks similar over time
- A **non-stationary process** is one that generates data whose distribution of values changes over time

In healthcare, the relevant units of time can span many orders of magnitude, from fractions of a second to a century. The relevant interval of time is directly influenced by the disease process, the measurements that current technology can make, and how the healthcare system is organized.

The relevant timescale depends on the question we want to answer. In addition to being influenced by the question, the relevant timescale is also often determined by the kind of data.

Some diseases are chronic, meaning that a cure is not possible, so the disease must be managed indefinitely, perhaps for the rest of the patient's life. Diabetes is an example of a chronic disease.

These two considerations -- **the question**, and **the kind of data** – interact and inform strategy.

Strategy on what features you are going to use:

- How accurately should they be ascertained?
- How many different kinds of features you are going to use?
- How will you infer whether a patient has a condition of interest?

REPRESENTATION OF TIME

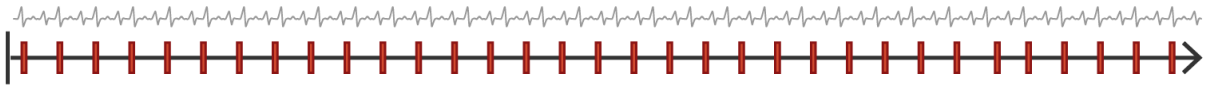
How we would capture (or encode) information about time in a computer-based representation -- that is how we *represent time*.

- **Patient-feature matrix**: A rectangular data frame in which each row is a patient and each column is a characteristic about them

In a patient-feature matrix, each row is a patient and each column is a characteristic about them, such as an identifier.

A **time series** is a set of measurements that are sampled at regularly-spaced intervals, with each measurement being of the same type.

It is important to know that an important area of healthcare data that uses time series data and related analysis methods is in the ICU. A patient in the ICU is typically extremely ill, and their physiological status needs to be monitored continuously via sensors attached to the patient's body. These streams of data have regular sampling intervals and methods from the field of signal processing can handle these data well.



TIME SERIES DATA

NON-TIME SERIES DATA

Most medical data are not acquired on regular clock ticks. They are sampled asynchronously as determined by necessity. For example, the EKG is a continuous measurement, but blood pressure is measured as needed.

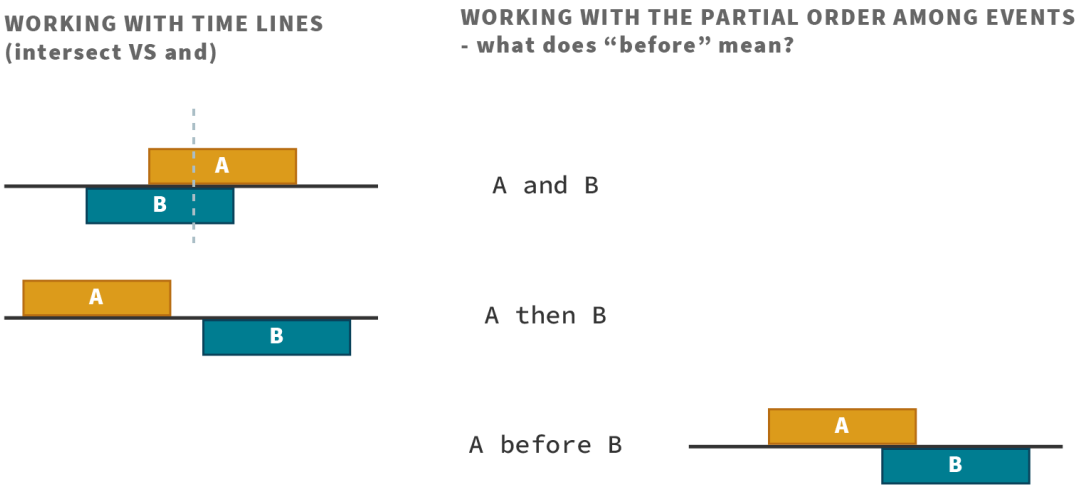
Many medical measurements are acquired in two stages:

1. The clinician orders the test
2. The test is actually performed

An **indicator variable** marks *what* test was ordered and *when* it was ordered, but does not record the result of the test. The indicator variable is separate from the test value, which records the actual value and the units in which it was measured.

Order of events

Knowing what happened when is a great start, but in many cases, we want to be able to reason about the order of events such as: finding patients with A and B, or patients with A *then* B.



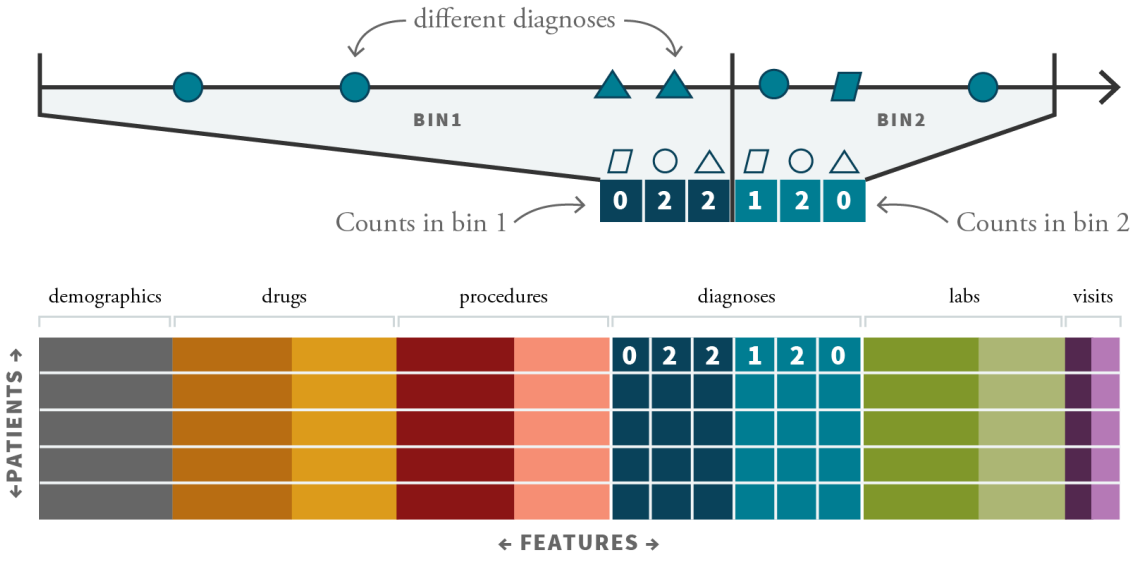
If A and B are instantaneous points in time, then answering this question is straightforward. However, if A and B refer to intervals of having diseases, such as "pneumonia" and "rheumatoid arthritis", then the reasoning becomes more complicated.

What does patients with "pneumonia" and "rheumatoid arthritis" mean? In one interpretation, event A finishes before event B starts. In the second interpretation event A finishes after event B starts.

Most general-purpose databases are not structured to make it easy to work with these distinctions—but a timeline representation does.

Represent time information in the patient-feature matrix

- Binning: Record the number of times that the relevant events occur during specified intervals



Let us look at it in detail: We start with a patient timeline. We define time intervals, called bins, that are relevant to the analysis, and count the number of events of each type that occur in each bin. The bin counts become features in the patient-feature matrix.

Choices in time-binning:

- How many bins?
- What granularities of time?
- How does this relate to the timescales at play in your research question?
- How do you aggregate the data within each bin?

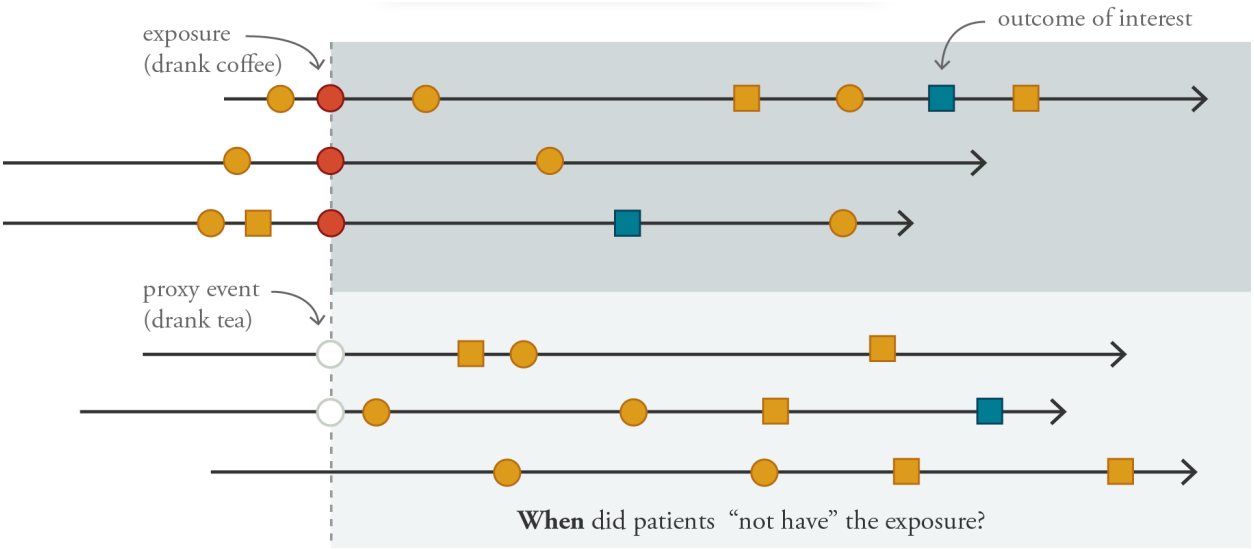
Possibilities for making a decision on aggregating or summarizing data within each bin:

1. It might make sense to use the count of the number of events that occur in each bin as the feature. You could record a count of zero vs a positive count. This would mark absence versus presence.
2. Use the average of all the values in the bin, or the maximum value of all the values in the bin, or the most recent value of all the values in the bin, or the variance of the values in the bin. This choice would be governed by the research question and the clinical item of interest.
3. Add a second feature that records the rate of change of a feature. Creating new features is called feature engineering, and a little medical knowledge goes a long way in crafting such useful features that encode time.

Timing of exposures and outcomes

- **Cohort**: A set of patients that satisfies some inclusion criterion, typically an exposure of some sort
- **Exposure**: Something that could happen to the patient
- **Outcome**: A condition of interest that is assessed as having happened to the patient, usually at some point after the exposure

In general, we are interested in whether there is an association between the exposure and the outcome for patients in the cohort. Therefore, we need to identify those that are exposed and those not exposed as well as those that had the outcome and those who did not. In addition, it matters *when* the exposure and outcome events occurred.



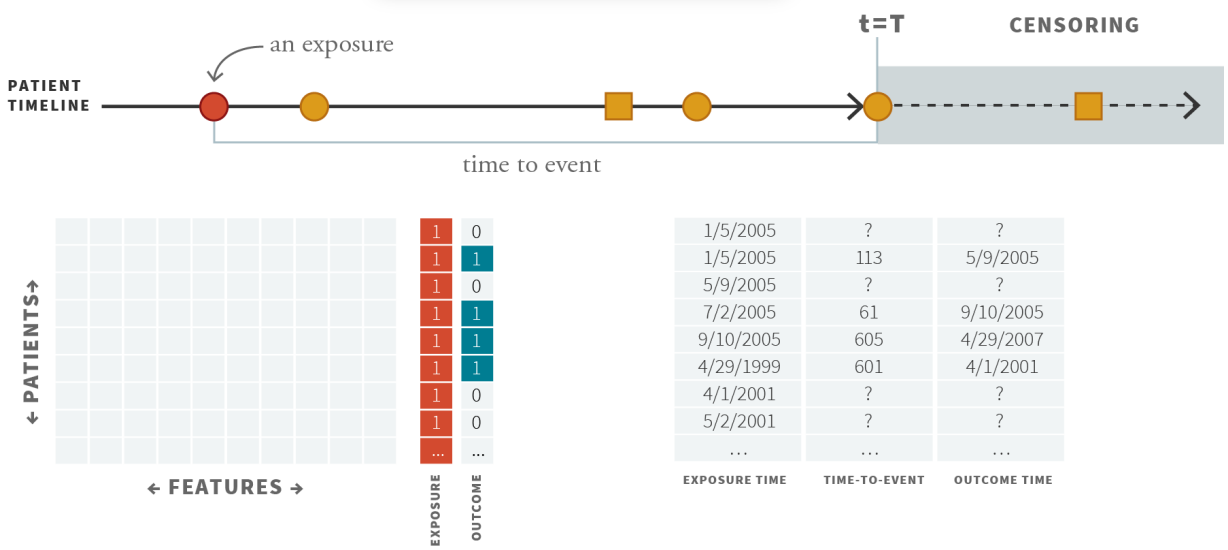
For those in the exposed group, such as those who had a coffee, the start time is straightforward to determine. It is when the exposure (i.e. having a coffee) appears on the patient timeline. The start time is also referred to as the index time.

For those non-exposed, meaning those that did not have a coffee, it is straightforward to determine that they are not exposed, because you won't find any exposure or coffee drinking, events on the patient timeline. However, what time should we use for when the non-exposure to coffee happens? This is a very important issue, and not always easy to solve.

A possible solution is to introduce a proxy event (such as drinking tea) whose presence is taken to mark the start of non-exposure. However, this may change who enters the control group and introduces a selection bias. For example, it excludes everyone who did not drink tea or coffee.

Remember: Even if you know *who* is not exposed, it is hard to decide *as of when* they should be counted as not exposed.

We also need the time when the outcome occurs. This will allow us to compute the time-to-event, which is the difference between the time of the outcome and the time of the exposure.



Without a time-stamp, we cannot compute the time difference. In epidemiology, this condition is called "right censoring". At a given point in time T, there will be some patients who have not yet developed the outcome.

- One possible solution is to use a special code, which we will write as ">T", which means "the outcome has not happened yet."
- Another solution is to record the time of the outcome, if known, or the time when the patient was last observed along with an indicator variable to mark whether the recorded time is for the outcome or for the last observation.

In summary, you need to diligently determine exposure and outcome times correctly.

DATA CHANGE OVER TIME

Along with the choice of the **timescale** and the **representation of time**, we have to deal with the fact that the healthcare system itself is evolving over time. Collectively, these changes make our **data generation "nonstationary"**.

- **Stationary process**: One that generates data that looks similar over time
- **Non-stationary process**: One that generates data whose distribution of values changes over time

Non-stationarity is usually a problem, but is often ignored.

One clever way to use machine learning to *detect* non-stationarity is to remove time as a feature, and then try to predict time from the remaining variables. If we can do that accurately, then there is some pattern in the predictor variables that correlates with time -- meaning there is strong non-stationarity.

The point of the discussion is for you to know that this phenomenon exists, and that you have to test for its presence. This is particularly true for **datasets that span long time intervals** and **studies that have large timescales**.

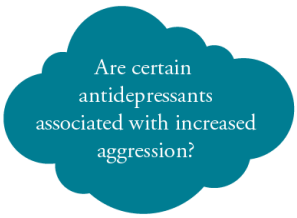
MODULE 4 – CREATING ANALYSIS READY DATASETS FROM PATIENT TIMELINES

LEARNING OBJECTIVES

8. Identify the key steps in converting messy clinical data into the tabular shape used in machine learning
9. List some factors that guide the decision to include or exclude a feature
10. Describe what is meant by missing data and explain ways to address it
11. Describe the goal of feature engineering
12. Define what a knowledge graph is, and give an example of one
13. Explain how a knowledge graph can help in analyzing clinical data

CREATING FEATURES TO ANALYZE

In this conversation, we'll dive into the construction of a patient-feature matrix that is the foundation for all subsequent analyses. The **patient-feature matrix** contains data about patients in a tabular format. The data for a given patient occupies a single row. Each column is a different measurement or feature.



PATIENT	MOST RECENT BLOOD SUGAR	SMOKER?	SEX	AGE
1	101	1	F	55
2	120	0	M	12
...	...	...	...	...

PATIENT	TAKING SEROQUEL?	TAKING PROZAC?	AGGRESSIVE?
1	1	1	0
2	0	1	1
...	...	...	...

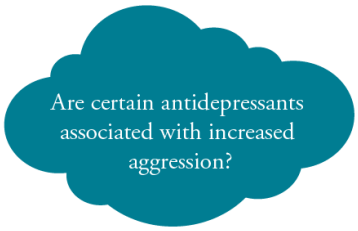
In clinical studies, the unit of observation and analysis is almost always the patient. In the data frame, each row contains all the data for a single patient. Each column contains a different feature.

Once we have decided the unit of observation, and analysis, we turn to determining which features to extract for the data.

It is best to start with all the features. If there are constraints on computing resources, then you may need to remove features to reduce the size of the dataset. Some modern machine learning methods can automatically remove those features that contribute the least to the accuracy of the model.

Important: You may need to remove some sensitive features for patient privacy.

We can use subtle information implicit in the data to help us craft features. Metadata, data that refers to other data, can be quite informative.



(D) Client spent most of the visit talking about her relationship with her boyfriends. She mentioned that he yells at her from time to time for no apparent reason. She said she gets her feelings hurt when he does that, and that sometimes she fears his yelling will escalate to violence.

(A) She appears stressed about her relationships with her boyfriend. She spoke little about her daughter and appears to be more preoccupied with her relationship during this visit. Not much improvement from her last visit.

(P) Will follow up with healthy relationship material and talk when we convene again next month. Continue to work with participant on self-esteem issues and focusing her energies on her daughter.

Example: We need to determine who is a diabetic. Ideally we would look at the *results* of a test such as the HbA1C. However, we can also use the counts of orders of laboratory tests that have something to do with measuring glucose instead of the actual results of those tests. So, we used the metadata -- the number of times a test related to measuring glucose is ordered -- and some prior knowledge that diabetes mellitus is a disease where glucose levels get messed up to craft a **feature (percentage of tests that are about glucose)** which informs us whether someone is a diabetic or not.

Usually such features are **created**, or **engineered**, by using some prior knowledge. It is also possible to learn such features via computation.

Healthcare data can be structured or can be unstructured.

Making datasets from structured sources:

1. Accessing structured data
2. Standardizing features
3. Dealing with too many features
4. Dealing with missing data
5. Constructing new features

Structured data, generally reside in database tables. Databases are queried using SQL (Structured Query Language) and the results can be loaded into systems for analysis. Data in different tables may be linked using a database operation called a "join". The data may need to be reshaped into a useable format.

It is common to standardize features, which transforms all features into a common numerical range. The process of standardizing is sometimes called **normalizing**. Standardization facilitates later analysis by reducing the effect of values that are extremely large or extremely small relative to other values in the dataset.

PATIENT	DX 993.4	BLOOD SUGAR	AGE IN DAYS	...
1	0	120	11315	
2	2	120	32000	
30	0	110	6003	
46	0	120	13500	
54	0	130	522	

Different scales will throw off many analyses.

PATIENT	DX 993.4	BLOOD SUGAR	AGE	...
1	0	0.92	0.34	
2	1	0.92	1	
30	0	0.84	0.18	
46	0	0.92	0.36	
54	0	1	0.01	

SCALE

$$X'_j = \frac{X_j - \min(X_j)}{\max(X_{j)} - \min(X_j)}$$

A commonly used transformation rescales each column so it spans the same range, often 0 to 1. Another transformation is such that the column has an arithmetic mean of 0 with a standard deviation of 1.

Reasons why you might not want to use all of the features:

- Some features might be useless
- Missingness: A feature might be missing for most patients
- Sparsity: A large number of features are missing for a given patient
- Redundancy: Feature1 might be highly correlated with Feature2
- Speed: Large number of features can slow the analysis
- Privacy: Saving more features increases the chance of violating patient privacy

Low-prevalence, low-variance features are good candidates for removal.

Another way to reduce features is to combine features using domain knowledge. A benefit of such aggregation is that it may remove some idiosyncrasies of how individual clinical sites code features, making cross-site comparison easier. The aggregation step requires accurate representations of domain knowledge in the form knowledge graphs.

It is also possible to use mathematical operations to detect and use patterns in the data. Many of these methods, such as principal components analysis (PCA), use linear algebra. The benefits of using such techniques is that they are domain-independent and do not hinge on specific medical knowledge.

The main drawback of mathematically combining existing features, is that it makes the derived feature difficult to interpret. The new derived feature set may enable accurate predictions, but the features that contribute to the prediction may not have understandable clinical interpretation.

When you are considering reducing the number of features:

- Think about whether the distinctions reflected in a feature are relevant to your question
- The more flexible the model you are using in the later analysis stage, the less benefit aggregation will provide. Regressions will benefit more than gradient boosted trees
- Reducing the required computational resources provides benefits independent of the choice of model

MISSING VALUES

When we convert a patient timeline view of the data into a patient-feature matrix, naturally some entries in this matrix will be missing.

Missing Data: In prospective studies, when a value for a data element is missing, it is reasonable to assume that it should have been recorded but was not. However, in data that are a byproduct of routine care, just because there is a place in a user constructed patient-feature matrix, does not mean that the value **should** have been recorded

The absence of a specific value in a column of patient feature matrix could mean three things:

1. The value should have existed, but does not (the usual meaning of missing data)
2. The value not being present is an artifact of adopting a tabular view of the data
3. The value could have existed, but was deemed unnecessary to collect

Absent values create problems for analysis in two ways:

1) From how they are reported
2) If they were truly 'missing data', then from how they are imputed

Some systems allow for a special data element, often written as "NA" or "null", to represent a missing value. Other systems might use numerical values that are outside the range of possible values for that feature, such as 0 or -999, to denote a missing value.

Removing patient records with missing values a tempting and simple solution, but it often creates problems. Removing patients with a missing value would tend to remove under represented patients from dataset. This would bias our analysis.

Dealing with missing values

One widely used method is to impute the missing values. **Imputation** is a kind of prediction that fills in the missing values based on other information in the dataset.

A simple imputation method, called **column mean imputation**, replaces the missing value with the mean of the known values in the same column. This assumes that the variable's values in the other rows of that column have information about the missing value, which is often not true in medicine.

We can use **values in other columns of the same patient** to improve the imputation procedure. This is usually better than column imputation which considers only values in one column at a time. We are using expected correlations among different features of the same patient to infer the value of the missing feature.

A procedure called **k-nearest neighbors imputation** fills in a missing cell by looking for patients who are similar to the current patient on the basis of other features, and then uses those known values to impute the missing one.

A modern method called **multiple imputation**, repeatedly invokes imputation to create multiple versions of the data, which can be analyzed to provide an estimate of the variance in the imputed values.

Making a decision on whether **to remove missing values or to impute:**

- If a variable is mostly measured with only a few missing values, then you should consider imputation
- If a particular variable has mostly missing values, then you should consider dropping the variable. That variable does not contribute useful information for most patients, and we would have to impute the values for most patients.

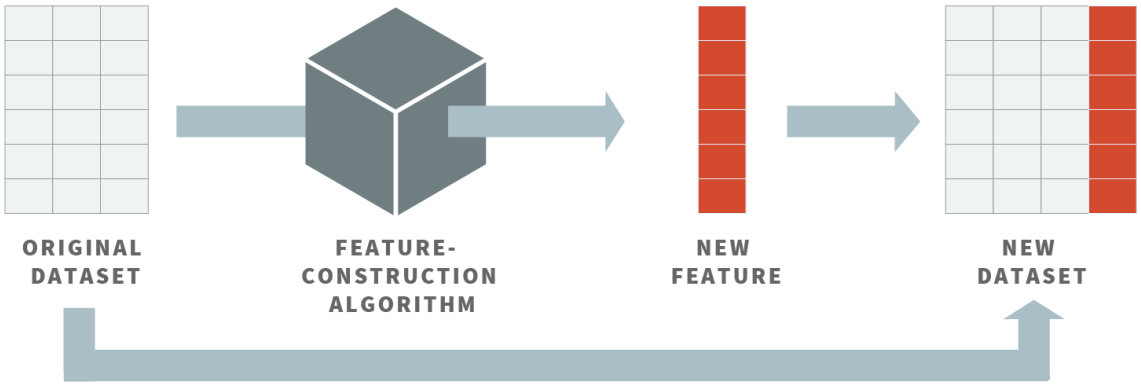
If the distribution of missing values is in middle-zone between these alternatives, some experts have advised adding indicator variables to mark which values have been imputed, but this recommendation has been disputed by others.

From a practical perspective, it may be better to use an analysis method that naturally handles missingness rather than imputing the missing values.

Finally, think about how important the feature with missing data is to your question. Could you avoid imputation by answering the question without considering that variable?

CREATING NEW FEATURES

As we convert a patient timeline view of the data into a patient-feature matrix, we can also perform simple operations on the source data **to create new features**. Such construction of new features is called "**feature engineering**".



Constructed features are transformations of the original features or their combinations. Simple models with well-engineered features can perform better than fancy models with original features.

Examples of engineered features

Clinical scoring systems, simple formulas that combine values found in the EMR, are great examples of engineered features. The body mass index is a relatively simple example of a scoring system that allows us to estimate the severity of how over or underweight someone is.

Other scoring systems quantify the overall burden of multiple diseases; often called the **comorbidity burden**. They are typically used to account for overall patient illness in analyses and avoid comparing sick people to healthy people.

Among other examples: Create proxy features for a patient's socioeconomic status from their zip code, and the number of EMR records they have, scaled by a measure of their overall health discussed above. Infer unrecorded conditions, such as smoking status, by looking for the presence of keywords in text, such as “cigarette”. In other cases, could look for specific combinations of drugs and procedures. Can also use clinical knowledge to guide feature engineering.

General Advice for Feature Engineering:

- Think about what features might be important but are not directly measured
- Take advantage of pre-validated clinical scoring systems

- When creating a new feature, consider including counts, differences, change over time, and ratios of existing measurements
- Lean towards creating new features using some clinical knowledge and creativity
- Balance the benefit from building new features against the effort used to create them

Deep learning: A new method in machine learning that uses more than conventional amounts of raw data and builds the needed features without domain knowledge

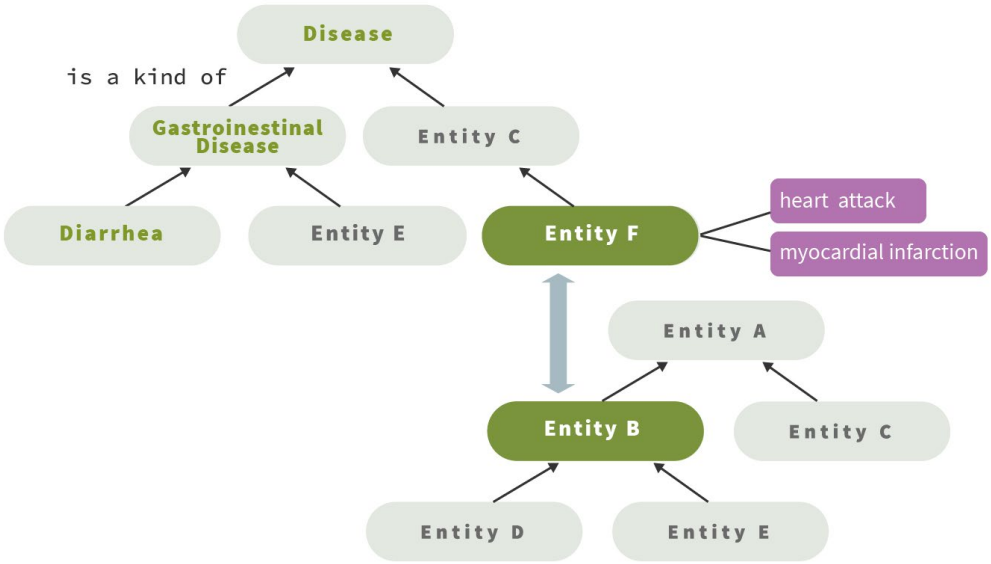
Creating analysis ready datasets from patient timelines

- Structured data in database tables can be transformed into analysis ready datasets called patient feature matrix
- The number of features in the patient feature matrix can be reduced by aggregating features using domain knowledge, or by using mathematical techniques such as principal component analysis
- Missing data can be removed, or imputed with different levels of methodological sophistication
- Consider creating additional features from the original data

KNOWLEDGE GRAPHS

A **knowledge graph** is a declaration of what entities exist in a domain and the relationships among them. It is also referred to as an ontology. Ideally represented in digital form.

A common problem when working with clinical data is that there are multiple ways to express the same concept. Knowledge graphs can help because they **explicitly represent synonyms of entities** and the **relations among different entities**.



So what exactly is in a knowledge graph?

1. They contain **entities**.
2. There are sets of **equivalent names** for those entities, or synonyms.
3. There are **relations** between the entities. A very common relation represents "is a kind of"
    - This 'kind of' relationship in Knowledge graphs is the most important because it codifies what is a kind of what in the medical domain. Entities will inherit properties from the entities they are a kind of.
4. Contain **links to other knowledge graphs**. This can provide a consistent way to refer to entities across different data sources

Knowledge graphs are extremely useful when querying clinical data. They help identify different terms with the same meaning.

There are hundreds of knowledge graphs available in medicine and in biological research. If you want to explore the many knowledge graphs available, check out the BioPortal from the National Center for Biomedical Ontology at Stanford.

Important Knowledge Graphs:

- International Classification of Diseases (ICD-10. ICD-9)
- The Current Procedural Terminology (CPT)
- RxNorm and RxNav provided by the US National Library of Medicine (NLM)
- Anatomic Therapeutic Chemical (ATC) Classification System
- The Logical Observation Identifiers Names and Codes (LOINC)

The Unified Medical Language System's metathesaurus, or the UMLS metathesaurus, is a union of over 140 knowledge graphs. It contains all of the **knowledge graphs** we have just mentioned, along with declarations of **relationships *between* these knowledge graphs**.

Questions to evaluate a knowledge graph:

1. What are the **entities** the knowledge graph has, and what is the basis of classification?
2. What **words** are used to name the entities in the graph? Are there synonyms and alternative names?
3. Is it **mapped** to other knowledge graphs? How 'connected' is a knowledge graph with other knowledge graphs?

Aside from these three principled questions, there are some practical approaches to assessing a knowledge graph. For example, given data from an EMR, count the number of terms from each knowledge graph that are mentioned in EMR text documents. In addition, if the knowledge graph is too big, with millions of terms, you can use the counts of term occurrences to help decide which terms to keep.

In summary, knowledge graphs are large, highly curated collections of medical **entities**, alternative **names** of those entities, and **relationships** between them. They are an extremely important source of medical knowledge for creating features and processing clinical text.

MODULE 5 - HANDLING UNSTRUCTURED HEALTHCARE DATA: TEXT, IMAGES, SIGNALS

LEARNING OBJECTIVES

- Describe the ways in which clinical text is different from natural language
- Describe how simple text mining strategies can be as effective at NLP when applied to clinical text
- What are negation and context detection, and why are they important?
- Explain how de-identification is different from anonymization
- Describe the important properties of healthcare images
- Describe the important properties of healthcare signals

UNSTRUCTURED DATA

This module focuses on text, images, and signals. These are all called **unstructured data** in contrast to **structured data**, the rectangular tables found in databases.

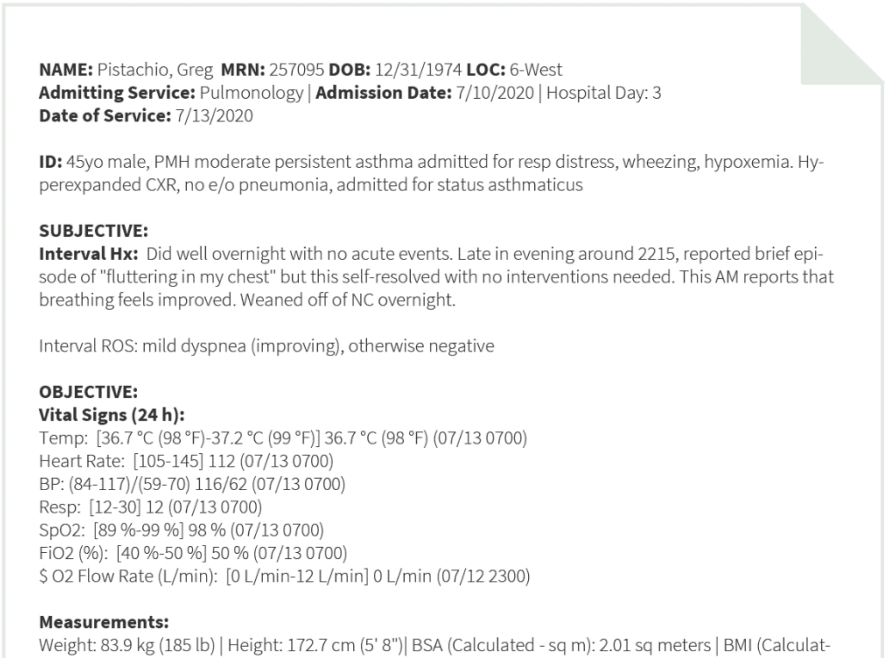
- Unstructured data: text, images, and signals
- Structured data: the rectangular tables found in databases

CLINICAL TEXT

Clinical text is different from ordinary natural language and the language used in scientific publication.

It is:

- Written by clinicians or other healthcare providers for clinicians and those documenting the services provided
- Used to describe patients, their pathologies, their personal, social and medical histories, and findings made during interviews and procedures
- Not written for publication and may not use full sentences



NAME: Pistachio, Greg **MRN:** 257095 **DOB:** 12/31/1974 **LOC:** 6-West
Admitting Service: Pulmonology | **Admission Date:** 7/10/2020 | Hospital Day: 3
Date of Service: 7/13/2020

ID: 45yo male, PMH moderate persistent asthma admitted for resp distress, wheezing, hypoxemia. Hyperexpanded CXR, no e/o pneumonia, admitted for status asthmaticus

SUBJECTIVE:
Interval Hx: Did well overnight with no acute events. Late in evening around 2215, reported brief episode of "fluttering in my chest" but this self-resolved with no interventions needed. This AM reports that breathing feels improved. Weaned off of NC overnight.

Interval ROS: mild dyspnea (improving), otherwise negative

OBJECTIVE:
Vital Signs (24 h):
Temp: [36.7 °C (98 °F)-37.2 °C (99 °F)] 36.7 °C (98 °F) (07/13 0700)
Heart Rate: [105-145] 112 (07/13 0700)
BP: (84-117)/(59-70) 116/62 (07/13 0700)
Resp: [12-30] 12 (07/13 0700)
SpO2: [89 %-99 %] 98 % (07/13 0700)
FiO2 (%): [40 %-50 %] 50 % (07/13 0700)
$ O2 Flow Rate (L/min): [0 L/min-12 L/min] 0 L/min (07/12 2300)

Measurements:
Weight: 83.9 kg (185 lb) | Height: 172.7 cm (5' 8")| BSA (Calculated - sq m): 2.01 sq meters | BMI (Calculat-

What makes clinical text valuable:

- It can augment the billing codes that have been assigned to medical records
- It contains a description of what happened to the patient
- It is used for biosurveillance to detect and monitor disease outbreaks
- It is used for improving the set of words that are used to refer to diseases
- It is used in clinical decision support
- It is used to add codes to the patient's medical record for querying and reporting

The value derived from clinical text may depend on the medical condition under consideration. Some conditions are accurately represented by codes. In other cases, a significant fraction of the patients would only be identified through the analysis of the text in the clinical notes.

Clinical text can enable clinical research by facilitating the construction of study cohorts. It has even been used as part of an automated procedure to discover new knowledge by mining the text for particular patterns.

There are many important challenges to using clinical text:

- Be ungrammatical.
- Have misspellings and concatenations
- Contain short telegraphic phrases, acronyms, and abbreviations
- The quality of the sources can vary widely

A very important problem in analyzing clinical text is the problem of **negation**.

- **Negation**: Refers to the use of a phrase to indicate that the patient does not have a condition
- The problem of negation: the analysis of clinical text needs to detect when a term mentioned is inside a negation. Roughly 40% of the content of clinical text is stated as a negation.

A related problem is called **context**.

- **Context:** Refer to a condition that a patient had before, or that a patient's family member has or had
- The analysis needs to detect the context of a term mentioned

Finally, there is pervasive fear, misunderstanding, and confusion around security, privacy, anonymization, and de-identification. This has artificially increased the burden to obtaining data access to clinical text.

PRIVACY AND DE-IDENTIFICATION

Textual data can contain Protected Health Information (PHI), which cannot be shared without the patient's permission.

Methods for de-identification:

1. Safe Harbor: Requires the removal of 18 specific items
2. Statistical Method: A statistician validates and documents that the statistical risk of re-identification is very small
3. "Hiding in plain sight": Replaces text that looks like identifiers with realistic-appearing surrogate information.

Visit the US Government's Health and Human Services website for more detail on statistical method and the safe harbor method.

NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is an area of artificial intelligence that develops methods to allow computers to process human language.

NLP consists of the following steps:

1. Tokenization: the purpose of this step is to detect words and identify where sentences begin

Terminology

- **Anonymization:** Data cannot be linked to a specific person
- **De-identification**: Removal of identifiers that constitute protected health information

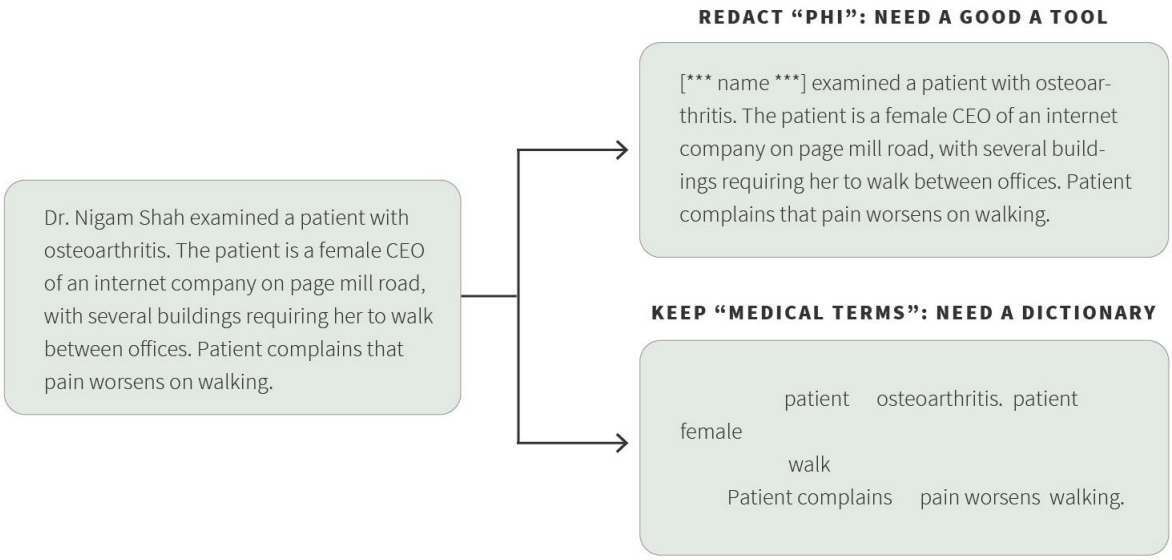
and end

2. Parsing: determines the grammatical structure of each sentence and tags each word with its part of speech, such as noun or verb
3. Named entity recognition: identifies words or phrases of interest, and assigns appropriate category labels
4. Section detection: identifies boundaries between different sections of a document

When NLP tools are used on clinical text, they typically require some adaptation of the usual steps in an NLP system in order to work. The NLP tools need to address negation and context problems.

PRACTICAL APPROACH TO PROCESSING CLINICAL TEXT

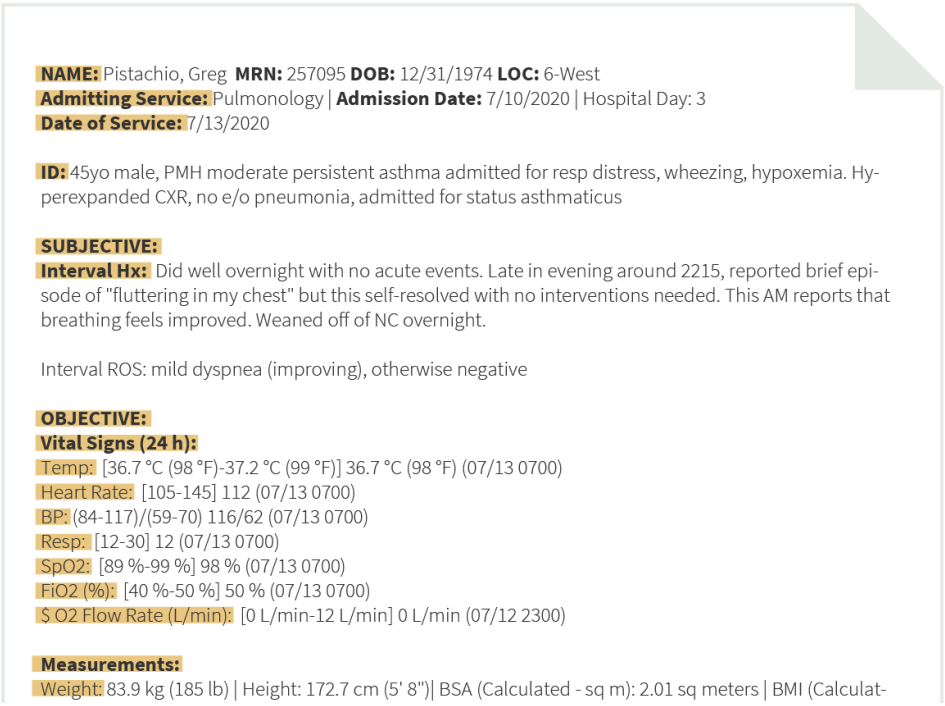
The ultimate goal is not to understand the contents of the document but to identify features in the text that will enable the downstream analyses that answer questions.



There are two main approaches for considering the issue of PHI:

1. Remove the PHI. This requires locating all the PHI in the clinical text. If you can do this accurately, then the PHI can be removed or skipped, and the desired features can be extracted from the remaining text.
2. Keep only medical terms that are useful for analysis downstream; all other terms will be passively filtered out. Use a **knowledge graph**, a dictionary of terms from a computed-based representation of medical information, to extract the features.
    a. A knowledge graph can also help you decide which terms are ambiguous
    b. An **ambiguous term** is a term with multiple meanings
    c. Another text processing problem is how to find similar terms that might be missing from your dictionary.

We described the important problem of detecting negation and context in clinical text. To detect negation and context in clinical text, there are software packages that have been designed, such as Negex and Context.



NAME: Pistachio, Greg **MRN:** 257095 **DOB:** 12/31/1974 **LOC:** 6-West
Admitting Service: Pulmonology | **Admission Date:** 7/10/2020 | Hospital Day: 3
Date of Service: 7/13/2020

ID: 45yo male, PMH moderate persistent asthma admitted for resp distress, wheezing, hypoxemia. Hyperexpanded CXR, no e/o pneumonia, admitted for status asthmaticus

SUBJECTIVE:
Interval Hx: Did well overnight with no acute events. Late in evening around 2215, reported brief episode of "fluttering in my chest" but this self-resolved with no interventions needed. This AM reports that breathing feels improved. Weaned off of NC overnight.

Interval ROS: mild dyspnea (improving), otherwise negative

OBJECTIVE:
Vital Signs (24 h):
Temp: [36.7 °C (98 °F)-37.2 °C (99 °F)] 36.7 °C (98 °F) (07/13 0700)
Heart Rate: [105-145] 112 (07/13 0700)
BP: (84-117)/(59-70) 116/62 (07/13 0700)
Resp: [12-30] 12 (07/13 0700)
SpO2: [89 %-99 %] 98 % (07/13 0700)
FiO2 (%): [40 %-50 %] 50 % (07/13 0700)
$ O2 Flow Rate (L/min): [0 L/min-12 L/min] 0 L/min (07/12 2300)

Measurements:
Weight: 83.9 kg (185 lb) | Height: 172.7 cm (5' 8")| BSA (Calculated - sq m): 2.01 sq meters | BMI (Calculat-

Many types of clinical notes have section headings that can be useful in processing the text. One way to find section headings is to look for everything that occurs from the beginning of a line up to the first colon character.

Mining clinical text:

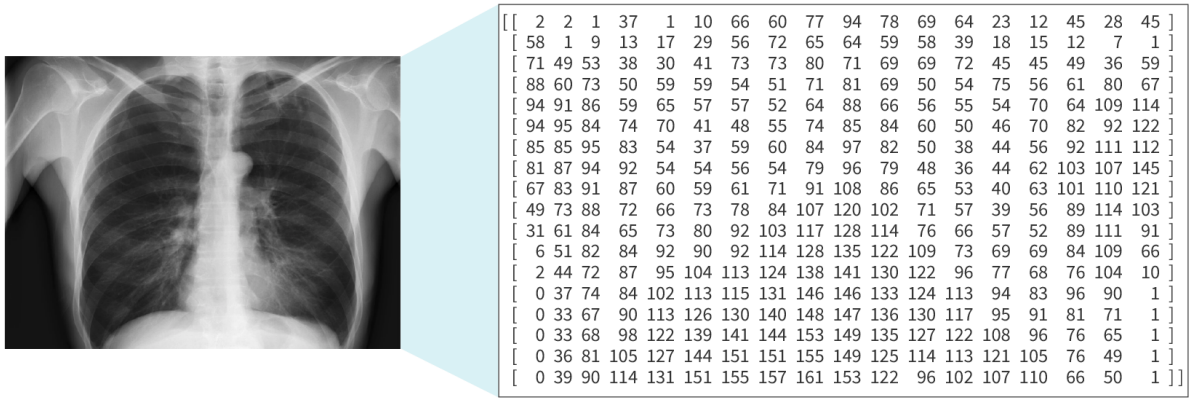
- Pre-process the entire collection of documents containing clinical text
    - Use knowledge graph and negation/context detection to find important terms
    - The output of this step is indexed positive, present mentions of diseases, drugs, devices and procedures
        - Indexed: a record of what string was mentioned where
        - Positive: negations are omitted
        - Present: personal and family history of the condition are omitted
- Answer a clinical research question: Use the knowledge graph to find synonyms of terms relevant to that question; use the timeline to help resolve ambiguous terms
- Count present, positive mentions about the patient: Use the temporal information, the aggregated event and drug mentions, and contextual filters to create a patient-feature matrix and construct patient cohorts for further statistical analysis. Using the temporal information is crucial.

IMAGES

Medical images serve several important goals:

1. Diagnosis
2. Disease staging and response to treatment
3. Guiding surgical interventions

Images capture important details about anatomic structures and physiological processes. However, images are large and require interpretation by human or machine to turn the low-level image elements into meaningful features for downstream prediction tasks.



```
[[  2   2   1  37   1  10  66  60  77  94  78  69  64  23  12  45  28  45 ]
 [ 58   1   9  13  17  29  56  72  65  64  59  58  39  18  15  12   7   1 ]
 [ 71  49  53  38  30  41  73  73  80  71  69  69  72  45  45  49  36  59 ]
 [ 88  60  73  50  59  59  54  51  71  81  69  50  54  75  56  61  80  67 ]
 [ 94  91  86  59  65  57  57  52  64  88  66  56  55  54  70  64 109 114 ]
 [ 94  95  84  74  70  41  48  55  74  85  84  60  50  46  70  82  92 122 ]
 [ 85  85  95  83  54  37  59  60  84  97  82  50  38  44  56  92 111 112 ]
 [ 81  87  94  92  54  54  56  54  79  96  79  48  36  44  62 103 107 145 ]
 [ 67  83  91  87  60  59  61  71  91 108  86  65  53  40  63 101 110 121 ]
 [ 49  73  88  72  66  73  78  84 107 120 102  71  57  39  56  89 114 103 ]
 [ 31  61  84  65  73  80  92 103 117 128 114  76  66  57  52  89 111  91 ]
 [  6  51  82  84  92  90  92 114 128 135 122 109  73  69  69  84 109  66 ]
 [  2  44  72  87  95 104 113 124 138 141 130 122  96  77  68  76 104  10 ]
 [  0  37  74  84 102 113 115 131 146 146 133 124 113  94  83  96  90   1 ]
 [  0  33  67  90 113 126 130 140 148 147 136 130 117  95  91  81  71   1 ]
 [  0  33  68  98 122 139 141 144 153 149 135 127 122 108  96  76  65   1 ]
 [  0  36  81 105 127 144 151 151 155 149 125 114 113 121 105  76  49   1 ]
 [  0  39  90 114 131 151 155 157 161 153 122  96 102 107 110  66  50   1 ]]
```

Images are produced by the transduction of some kind of physical energy.

Each image is a two-dimensional rectangular array of values, where the value is a measure of signal intensity. Images can also be three-dimensional, with the third dimension corresponding either to time or to space.

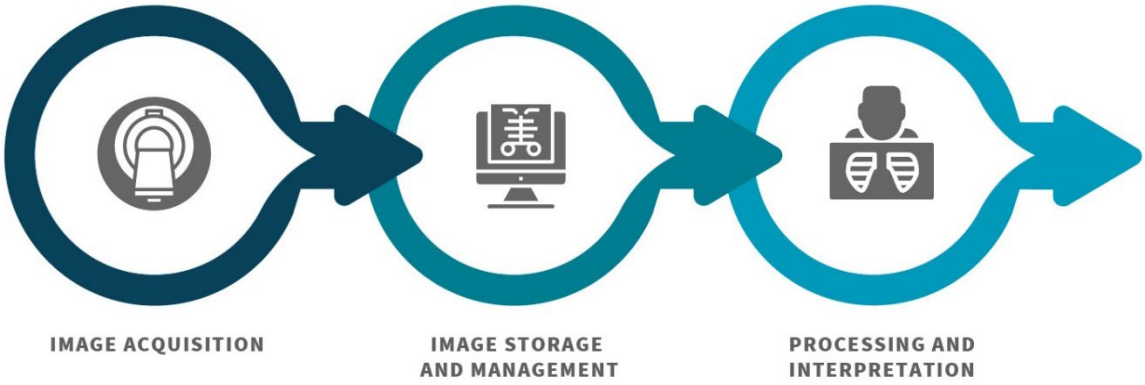
There are several ways to categorize medical images:

1. By the imaging modality--typically based on the kind of physical energy that is being detected:
    - **Visible light**: Photographs or videos
    - **X-ray**: High-frequency electromagnetic radiation, which is differentially absorbed by bones, air cavities, fluids, and tissues. X-rays, CTs, and related imaging techniques use ionizing radiation, which can cause injury to the exposed tissue in a dose-dependent way.
    - **Ultrasound**: Uses sound waves with frequencies higher than can be detected by the human ear. These waves propagate through, and are reflected from, tissue depending on the tissue density. Ultrasound has an advantage that it does not harm the tissue.

- **Magnetic resonance and nuclear medicine**: Involve the use of magnetic resonance, in which the magnetic properties of atomic nuclei are measured when brief electromagnetic pulses are applied. The density of the tissue can be calculated. The location of specific chemical tracers can also be found. Magnetic resonance also does not harm the tissue.

2. By structural versus functional:
   - **Structural images** capture the spatial location and organization of anatomic structures.
   - **Functional images** identify activity or change over time.

The most common images accessible for data mining are radiology images.



The life cycle of radiology images is:

1. Image acquisition
2. Image storage and management: The images are moved from the image acquisition device to a system that stores and organizes the images for retrieval. The storage system also records metadata.
   a. Digital Imaging and Communications in Medicine (DICOM), is a very widely used standard for the storage and transmission of medical images. DICOM applies to all the imaging modalities described earlier.
   b. Medical images are stored in Picture Archiving and Communication Systems (PACS). These representation standards and storage systems allow for image compression to reduce the amount of space occupied.

Terminology

- **Wearables**: Electronic devices that use improved sensor technology to allow individuals to monitor physical activity, exercise and sleep

3. Processing and interpretation
    a. Traditionally, the content of a medical image is interpreted by a specialist and the text of this report is stored in the EMR.
    b. Research these days concerns the processing of medical images by computer with the goals to:
        i. Automate routine tasks, and highlight important features to ease the task of the radiologist or pathologists
        ii. Automatically produce an expert-level textual description of the contents of the image

SIGNALS

A signal is created by biomedical equipment that measures some physiological value, and transduces it into an electrical signal, usually a voltage. Voltage is then converted into a numerical value in digital format for computer analysis.

Signals are important because they provide continuous, real-time physiological. They are a time series of regularly-spaced values, converted into a digital format.

Signals are:

- composed of measured values at regularly-spaced intervals as determined by the sampling frequency of the sensor
- typically captured in clinical contexts in which continuous monitoring is important, such as in the intensive care unit
- important for wearable devices, such as those that track heart rate

The research goals in using signals are similar to that for images: automatic **feature detection** and automatic construction of an **interpretation**.

Most commercially available devices use proprietary algorithms for feature detection and interpretation of the signal

What are the major issues with using signals?

- We do not know how accurate algorithms are, or which features are crucial for their interpretations to be correct
- Accuracy of consumer devices is uncertain; there are risks associated with missing real health problems, and falsely reporting problems that do not exist
- The true value of these devices and apps to health has not yet been demonstrated

- Privacy concerns

MODULE 6 - PUTTING THE PIECES TOGETHER: ELECTRONIC PHENOTYPING

LEARNING OBJECTIVES

1. Define electronic phenotyping
2. Describe the difference between a feature and a phenotype
3. Explain the purpose of electronic phenotyping
4. Describe the two main approaches to electronic phenotyping, and their strengths and weaknesses
5. Describe imperfect labelling and how it can be used in phenotyping

ELECTRONIC PHENOTYPING

Phenotype is a certain characteristic or condition of a patient that may be present or absent. The most common example is a disease.

Electronic phenotyping is a computational procedure for determining whether a patient does have or does not have the condition of interest based on electronic medical record.

The process of electronic phenotyping is found throughout clinical research.

Electronic phenotyping is useful for:

- Using observational data for research
- Recruiting into clinical trials
- Calculating quality metrics for healthcare systems
- Finding similar patients
- Sharing definitions to facilitate cross-site research

Difficulties of Phenotyping:

1. The codes might not be accurate
2. Even if the codes are correct, they might be assigned after the patient was first known to have the condition

The input to electronic phenotyping procedure will be a timeline of the patient's features. We need to choose which data to use, and which portion of the timeline to consider. We need to distinguish a feature from a phenotype.

- A **feature** is something that is directly measured
- A **phenotype** is the result of inference applied to one or more features

An **electronic phenotype** should contain the necessary and sufficient conditions of the features that should be present or absent and their required values to determine if an exposure or outcome of interest happened to a patient. It should also contain the criteria for identifying the start and end times. Locating the start time may be straightforward, it can be more difficult to determine when a condition ends.

When specifying a phenotype, it is very important to be clear about the intended meaning of that phenotype.

Evaluating a Phenotype Definition:

- Figure out exposures and outcomes
- Decide on risk thresholds
- Estimate the effects of treatments

Electronic phenotyping: Declaring the necessary and sufficient conditions of the features that should be present or absent and their required values to determine if an exposure or outcome of interest happened

It is important to realize that the accuracy of the phenotype definition depends on the research question you are trying to answer.

The last issue to consider for evaluation is how well a phenotype definition based on data from one clinical site will work at a second clinical site, often referred to as the portability of the electronic phenotype.

TWO APPROACHES TO PHENOTYPING

Approaches to Electronic Phenotyping:

1. Rule-based Phenotyping: Using rules comprised of explicit inclusion and exclusion criteria that were constructed by experts who reach consensus on the criteria in an iterative fashion

2. Probabilistic Phenotyping: Using machine learning instead of expert consensus to learn a function that assigns a probability to a patient's record for having the exposure or outcome of interest

RULE-BASED ELECTRONIC PHENOTYPING

We will use examples from the Phenotype Knowledge Base, or PheKB. This is a publicly available repository of phenotype definitions.

[**Site**: What is the Phenotype KnowledgeBase? | PheKB]

The phenotype has inclusion criteria and exclusion criteria. The inclusion criteria are features that must be present, while the exclusion criteria are features that must not be present.

INCLUSION CRITERIA	CODE	DESCRIPTION
1. If a qualifying diagnosis of sickle cell disease (see ICD-9 codes and descriptions) has been made in the problem list, medical history, as a primary diagnosis at encounter, non-primary diagnosis at encounter, or as a discharge diagnosis	282.41	Sickle cell thalassemia without crisis
	282.42	Sickle cell thalassemia with crisis
	282.61	HbSS disease without crisis
	282.62	HbSS disease with crisis
	282.63	Sickle cell/HbC disease without crisis
	282.64	Sickle cell/HbC disease with crisis
	282.68	Other sickle cell disease without crisis
	282.69	Other sickle cell disease with crisis
2. AND two outpatient visits at least 30 days apart or one hospitalization in the electronic medical record	–	–
EXCLUSION CRITERIA		
1. If number of diagnoses for sickle cell trait diagnoses > qualifying sickle cell disease diagnoses	282.5	Sickle cell trait

Example of sickle cell disease: The inclusion criteria has two parts: a set of ICD-9 codes and the requirement for one hospitalization or two clinic visits for this condition. The exclusion criteria has one part, which says that if the patient has more diagnoses for sickle cell trait than for sickle cell disease, then do not determine the patient to have sickle cell disease.

Example of type 2 diabetes mellitus: This phenotype is expressed as a flowchart, which they call an "algorithm". Starting with the records for a patient stored in the EMR; following "Yes" and "No" on whether the patient has diagnosed of Type 1 diabetes and check Type 2 diabetes diagnosis, prescription, or relevant abnormal lab value. The point here is that diagnosis codes are not accurate enough to be used on their own, but need to be augmented by prescriptions of medications, specific lab values, and consideration of the relative timing of events, stressing the need to have codified "prior knowledge" and adopting a timeline view of the patient record.

It is important to remember that each of the sources of data can be useful, that combinations of the sources are more accurate, and that the relative value of a source varies with the disease of interest.

Constructing a rule-based phenotype definition:

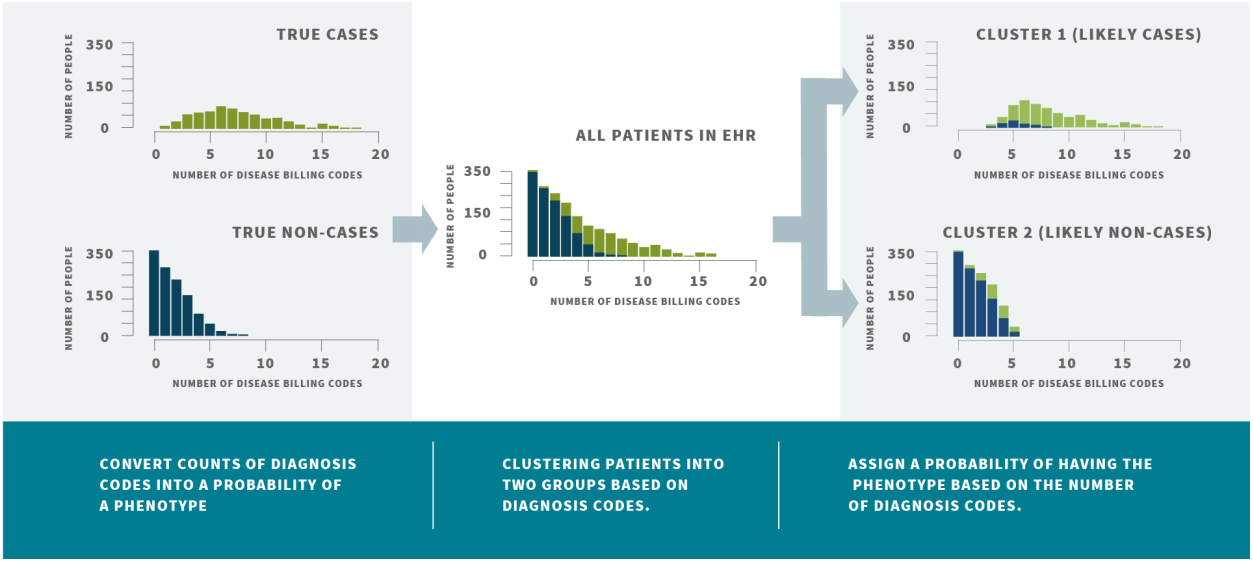
1. Identify the data elements that should appear in a medical record
2. Use a relevant knowledge graph to convert those data elements into specific identifiers
3. Create the phenotype definition by specifying the criteria
4. Iterate by comparison to some reference standard, usually clinician review of the full chart

PROBABILISTIC PHENOTYPING

Problem: Suppose you are the Chief Data Scientist at a healthcare startup. You need to identify patients who have one of fifty different conditions. The work needs to be completed within one week. What could you do?

Solution: Use Supervised Machine Learning with a training dataset, with each patient explicitly labeled as having or not having the condition of interest. The training dataset is used to build a computational model that can classify whether a previously unseen patient has the condition.

Approaches for Probabilistic phenotyping definition:



Computes the number of times billing codes applied to each patient, and interprets that count as a probability of having the phenotype. Then using that estimated probability, cluster the patients into those likely to have the disease and those not likely.

A mathematical formula:

The expansion of the sample size = 1/(1 - 2t)^2, where t = error rate

Using imperfect data:

1. Start with keywords for the condition of interest
2. Expand the set of keywords to include related terms using a knowledge graph
3. Include all patients who have those keywords as non-negated mentions somewhere on their timeline
4. Then train a classifier using all the other features to classify the phenotype

Another imperfect labeling system uses a concept called "anchors". An **anchor** is a reliable indicator of the presence of the phenotype. The anchor is used as a labeling function to obtain large amounts of training data for a classifier. We can also track how much does our classifier get better with adding more anchors.

Software for Probabilistic Phenotype Definition: Aphrodite (Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation). It uses standardized, open-source representations: OHDSI CDMv5 and Vocabulary 5, and it implements the labeling. The code is freely available on github: http://github.com/OHDSI/Aphrodite

MODULE 7 –CLINICAL DATA ETHICS

INTRODUCTION TO RESEARCH ETHICS AND AI

Any discussion of the role of artificial intelligence in healthcare must carefully distinguish the *development* of models, algorithms, predictive tools, and so on, from the routine uses of those tools in medical *practice*, whether as a decision aid, or a screening tool.

- **Research**
  - Development of new AI tools, models and algorithms
  - An activity that is usually structured and has as its primary goal, the production of generalizable knowledge
- **Clinical practice**
  - Application of validated or accepted AI tools, models and algorithms into different aspects of standard clinical practice or operations
  - An activity that aims to benefit individual patients. The purpose of the activity is to help make a diagnosis, prevent disease, or provide treatment to individual patients

History of Research Ethics

- 1947 Nuremberg Code in response to Nazi research atrocities

- 1932-1972 Tuskegee Syphilis Study
    - African American men were denied antibiotics and misled into thinking that lumbar punctures and other data collection measures were treatments for "bad blood"
- 1966 Henry Beecher NEJM article
    - Demonstrated that most published clinical trials had made no attempt to obtain informed consent from research participants
- 1964 Declaration of Helsinki
    - First major update since Nuremberg Code and included a requirement for independent review of formal research protocols
- 1978-9 The Belmont Report
    - 1974, US National Research Act created a National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, the final report becoming "the Belmont Report"
    - Significant overlap with Helsinki, though the Belmont Report included a very influential ethical framework that provided a foundation for those recommendations
- 1981 45 CFR 46 "The Common Rule" adopted

Belmont Report Ethical Framework

1. **Respect for Persons**: To respect the autonomy of agents who have the capacity to make their own decisions. Protect individuals with diminished capacity
2. **Beneficence**: A fundamental principle of medical ethics that has been widely seen as applicable to biomedical research: primum non nocere or "do no harm". Possible benefits should be maximized and possible harms should be minimized as much as possible
3. **Justice**: A fair distribution of the benefits and burdens of research.

Together, these principles and their application are the keys to understanding the requirements for the ethical development of new AI tools.

THE BELMONT REPORT: A FRAMEWORK FOR RESEARCH ETHICS

The three principles that are articulated in the Belmont report are the framework for the regulatory system governing human subjects research and have been enormously influential in understanding the ethics of research on human subjects. Research to develop new AI tools or to develop new knowledge or findings from human data will fall under these regulations and should be broadly consistent with this ethical framework.

1. RESPECT FOR PERSONS

Respect the autonomy of agents who have the capacity to make their own decisions

Terminology

- **Autonomy**: The ability of individuals to self-legislate in light of reason
- **Autonomy (ethics)**: Recognition of the right of individuals to make choices that reflect their values and interests
- **Autonomy (research)**: Research participants should make an informed decision about whether they want to participate in the activity

Informed Consent:

- Requires that the individuals be capable of providing consent
- Provide enough information to make an informed choice about whether to participate
- Free of coercion

The Belmont report (and the regulations in the U.S. that were based on them) does explicitly allow research to take place without consent, even for individuals who have capacity. The Belmont report recognized that sometimes, the requirement that potential participants provide informed consent could impair the validity of the research itself.

Waiving the Normal Informed Consent Requirements:

1. Incomplete or non-disclosure is necessary for the research goals to be achieved
2. Risks of the research are minimal
3. Later plan for dissemination of information about the research
4. No other rights of the individual will be violated or other harms to the participant will take place as a result of the research

These requirements are particularly important for the data collection required to develop AI.

Respect for persons also required that individuals with diminished autonomy be respected.

2. BENEFICENCE

Minimize harms and maximize benefits to participants of research.

Risks to research participants must be reasonable in relation to anticipated benefits of the research, including the value of the generalizable knowledge to be gained.

3. JUSTICE

Fair distribution of risks and benefits of research:

- Aristotle's formal principle of justice: treat similar cases similarly
- Distributive fairness
- Procedural fairness

Fair procedures could still produce unfair distributive outcomes as a result of "social, racial, sexual, and cultural biases in society".

There are many critics of the Belmont Report, most importantly that it does not provide any guidance about how to deal with trade-offs between these principles, since there is no rank ordering among them. However, it is a critical starting point that governs how Institutional Review Boards (IRB)/Research Ethics Committees (REC) evaluate research.

ETHICAL ISSUES IN DATA SOURCES FOR AI

Example of AI Research:

- Researchers use deep learning to try to identify and understand novel pharmacogenomic variants
- Novel algorithms and models have been developed to better predict which patients will benefit from aggressive palliative radiation
- Tools have been developed to improve identification of tumors in radiological images
- Tools have been developed to improve health insurance claims processing

In order for new tools to be developed, AI requires data for it to learn from. AI has often been linked to the concepts of Big Data and Precision Health. The vision of creating more precisely tailored interventions for individual patients requires that large amounts of data from different sources be used to understand why patients respond differently to the same medications or have different side effect profiles.

Types of Data for AI Research:

1. Research repositories
    a. NIH "All of Us Research Program"
    b. UK "100,000 Genomes" project
    c. Geisinger biobank

2. Secondary uses of data collected for other purposes
    a. Electronic health records (EHR) data from patient encounters
    b. Insurance claims data
    c. Newborn blood spots
    d. Census data
3. Consumer data collected outside of healthcare
    a. Wearables
    b. Mobile health
    c. Direct to consumer (DTC) genetics
    d. Computer usage data
    e. None of the oversight or regulatory mechanisms that govern apply

Research Repositories

- Data collected through interactions with a person engaged in an activity for the purpose of research
- Consent is normally required and obtained
- Informed consent for each use of the data is not usually possible
- Broad consent is an option

Requirements for Broad Consent – Participants need to be told:

- Of any known risks
- Of any anticipated benefits
- Details about how confidentiality is maintained
- Participation is voluntary
- If genetic analysis will be done
- Whether data may be used to generate profit and whether shared

Research Repository Issues

- Informed consent
- Security efforts to ensure privacy as much as possible
- Justice
    - Research repository recruits have tended to be middle class or affluent individuals of European ancestry
    - Fair procedural process may still result in inequality in outcome as a result of historical mistrust

Creating fair access to the results of AI research will require great effort to repair trust in under-represented populations, to build relationships between under-represented communities and researchers, and to succeed in improved recruitment and engagement. There is some promising improvement in some of these efforts, but it remains an open question whether research repositories can ever be truly representative of the public.

SECONDARY USES OF DATA

A great deal of AI research makes use of data that was collected for purposes other than research; EHR data, Insurance claims data, Newborn blood spots, Census data.

Obtaining informed consent for the secondary uses of data collected for other purposes is often (though not always) impracticable, particularly if it is required to meet regulatory requirements for informed consent for research on human subjects.

Pathways to Use Without Consent:

- If the "research" activity is regarded as "Quality Assurance (QA) and Improvement"
    - Not considered research under regulations
    - QA typically does not require consent
    - Limited guidance about how to determine whether something is "QA" rather than research
        - Rule of thumb: are findings generalizable to other institutions?
    - Note: intent to publish does not make the activity research
- Secondary use of de-identified data
    - HIPAA standards of de-identification
        - Expert determination standard
        - Safe harbor standard: Strip 18 identifiers
          "does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is the subject of the information"
    - NIH sees genomic information as identifiable; researchers continue to find new ways to identify individuals in databases that were previously regarded as de-identified
    - The idea that de-identification is sufficient to remove the duty to obtain consent remains ethically controversial
- Waiver of consent

Waiver of Consent Requirements:

1. The research must present minimal risk to participants
2. The research could not be carried out successfully if consent was required
3. The waiver of consent will not adversely affect the rights or welfare of the participants
4. Afterwards, the participant will be informed of their participation if it is appropriate

"If the research involves using identifiable private information or identifiable biospecimens, the research could not practicably be carried out without using such information or biospecimens in an identifiable format"

Ethically Controversial:

- Concern by many in bioethics that unconsented data use will undermine trust and demonstrates a lack of respect
- Surveys shows that potential research participants want to provide consent for uses of their data
- Surveys also demonstrate high levels of support for research uses of their data
- Less studied is how potential participants view the trade-off between these two preferences
- Requiring consent will likely lead to fewer participants, less data, and more biased data
- Pronounced tendency for consented data in research repositories to leave out minority populations, the poor and those in rural areas
- Are there alternative (and possibly better) ways of demonstrating respect for participants

RETURN OF RESULTS

AI research on samples has the potential to discover information about individuals in the data.

Some research on data specifies that no results will be returned:

- Sharp boundary between research and clinical practice means no duty of care

Raised cautionary flags about the idea of "Therapeutic Misconception"

- "**Therapeutic Misconception**": The confusion between research and practice by participants

Ethical challenges to this approach:

1. Participants often want information
2. Ancillary duty to warn may still exist
3. Grounded in Beneficence (maximize benefit)

Permissive Approach:

If participants want any information that is discovered about them in the course of research on their samples or data, they have a right to that information and therefore any findings should be offered.

- Return all information that is discovered about them that is linked to them
    - Grounded in respect for patients and their autonomy
    - Individuals differ in what information they consider important or relevant
- Challenges
    - Findings may not be valid or even interpretable
    - Some findings will not be actionable
    - Poorly defined, unvalidated information could cause harm and fails to minimize risks
    - Potential wide range of information makes consent about risks impossible

Range of Intermediate Positions:

Intermediate positions between these two extremes may be more plausible, but there are a range of views about precisely where to draw the line.

- Duty to return at least some results - maximize benefits
    - Analytically and clinically valid
    - Significant medical impact
    - Creates opportunity to take action to treat, mitigate or avoid impact
- Findings that should not be returned - minimize harms
    - Uncertainty or lack of interpretation
    - Lack of validity
    - Risk of harm
- Grey area in between
- Variation exists in practice about what to include in each category as well as what to do about grey area

There is a tradeoff between respecting the autonomy of participants (or parents who provide permission for their children) and the obligation to minimize harm to participants, which is grounded in Beneficence. There remains a challenging balancing between these two considerations leading to variation in practice about the best place to draw the line on what should be returned.

For findings that are neither required to be returned nor impermissible to be returned, there is a grey area with variation in practice. Researchers would not be required to return these results, though it would be permissible to do so.

The issue is somewhat more complicated in the context of secondary research uses of de-identified clinical or public health data, if there has not been any consent. There is arguably a heightened duty to return actionable results because the information was collected in a clinical context, rather than a research context. The practical challenges of actually returning information when people are unaware of the research being done with their data are quite difficult. This problem would be much more manageable if patients were more aware of how their data can be used to improve quality, as well as to conduct research to improve healthcare.

AI AND THE LEARNING HEALTH SYSTEM

A potential solution to the problems presented by the regulations can be found in the concept of the Learning Health System. AI research, and in particular the data collection needed for research, presents a difficult tradeoff between fully informed consent and the ability to produce valid, representative, scientifically and medically important research.

For secondary uses of data collected for clinical or public purposes, it can sometimes be a challenge to obtain meaningful consent at all. For uses of data in research repositories, it will typically not be possible to obtain consent for each use of the data, requiring a broad consent rather than anything approximating fully informed consent. The fact that findings from this research may sometimes be returned makes the consent issues even more complicated.

Challenges to Distinction Between Research and Practice

The starting point is the distinction between research and practice. Some research is not aimed at trying to develop new drugs or interventions, but instead aims to improve different aspects of medical practice.

- Quality Improvement Research: Goal is to improve patient care

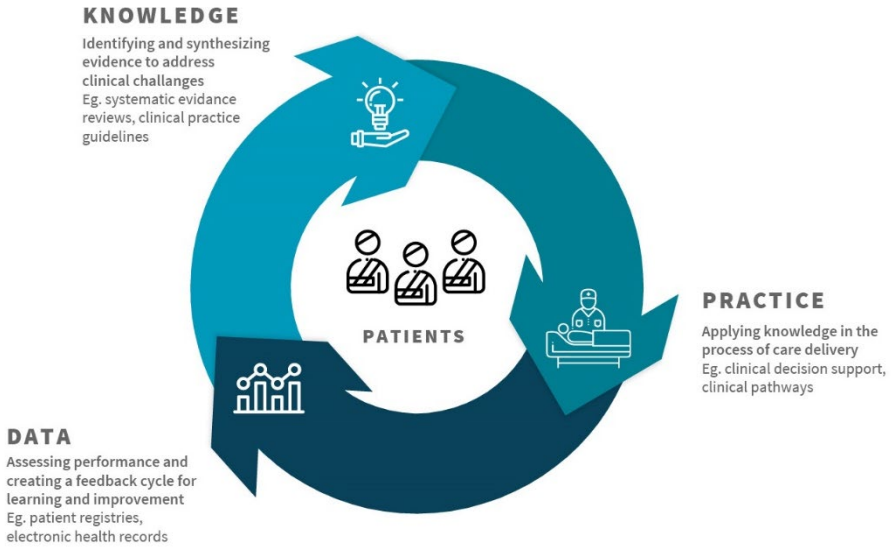
Terminology

- **ROMP:** Research on Medical Practices
- When an institution makes a commitment to systematically carry out ROMP using all of its data and clinical care practices as a way of improving care, it is sometimes called “the Learning Health Care System.”

- Comparative Effectiveness Research: Goal is to determine which interventions within the range of usual care is best
- Precision Health: Goal is to determine which interventions within the range of usual care is best for particular cohorts of patients (or individuals)

The Learning Health System:

- Institute of Medicine (now the National Academy of Medicine) defined LHS as "one in which knowledge generation is so embedded into the core of the practice of medicine that it is a natural outgrowth and product of the healthcare delivery process and leads to continual improvement in care"
- Challenges the traditional line drawn between clinical practice and research
- Every patient is a research participant whose data is an opportunity to learn
- Learning improves patient care
- Requires setting up systems to learn from patient data and to translate learning to improved patient care
- AI will be key to both feedback loops



- AI is likely to play a critical role, both in the learning process from the data and then in the development of methods to translate new knowledge into altered practice

Learning Healthcare System Ethics Framework

Critics of the current research ethics paradigm derived from the Belmont Report have developed an alternative ethical framework for the Learning Health System. Their framework is based on

recognition of a set of obligations or duties held by different stakeholders in the learning health system.

- Respect for Persons: This is captured in the LHS by the duty to respect the rights and dignity of patients
- Beneficence: This is captured in the LHS in three ways
    - Obligation to provide optimal care for each patient
    - Duty to avoid imposing nonclinical risks and burdens
    - Obligation to improve clinical care through learning
- Justice: The duty to address and reduce unjust inequalities
- Obligation to respect clinician judgment
- Patient obligation: Duty to contribute to knowledge, particularly when there is little risk or effort involved. Pass on the benefits that they themselves have received from contributions of others

Remaining challenge

This framework does not offer details about how to balance the tradeoffs that exist between the obligations:

- How to balance duties and obligations (such as consent as a way of demonstrating respect) with duties of justice and beneficence
- Is a form offering broad consent for minimal risk activities that improve patient care the best way to demonstrate consent?
- Patients and participants want to give consent, but they also want the fruits of research
- Large majority of patients and the public were willing to accept less elaborated (or no) consent if necessary for research to take place

Alternative Measures to Demonstrate Respect

Faden and Kass have suggested that there may be better ways of demonstrating respect for patient/research participants than the standard informed consent forms.

- Active engagement with patients about ongoing learning activities
- Transparency about the types of learning taking place
- Accountability in translating the results to improve patient care

It is worth noting that something like this has already taken place in regard to individual rather than systems learning:

Analogous Learning at Individual Level

- Teaching hospitals require learning for training individual clinicians
- Patients required to receive some of their care from trainees
- Benefit is care at teaching hospital
- No consent option (required)
- Fairly broad public expectation
- Transparency about fact
- Accountability to minimize risks and to ensure adequate training

The Learning Health System

The research regulations and the Belmont Report are flexible enough to accommodate this system as long as IRB's or Research Ethics Committees are willing to utilize appropriate mechanisms, such as waiver or alteration of consent, when it is appropriate.

- Analogous to learning at individual level
- Systems level learning a similar duty
- Greater potential benefit to participants than individual learning