UNIT I INTRODUCTION TO HEALTHCARE ANALYSIS

Overview - History of Healthcare Analysis Parameters on medical care systems- Health care policy- Standardized code sets – Data Formats – Machine Learning Foundations: Tree Like reasoning , Probabilistic reasoning and Bayes Theorem, Weighted sum approach.

Introduction to Healthcare Analytics

Health care analytics:

Healthcare analytics is *the use of advanced computing technology to improve medical care*.

What is meant by Health Care Analytics? ***[2M]***

Health care analytics is a subset of data analytics that uses historic and current data to produce actionable insights, improve decision-making, and optimize outcomes within the health care industry. Health care analytics is used to benefit health care organizations and improve the patient experience and health outcomes.

Healthcare analytics uses advanced computing technology ***[2M]***

--In 2020, computers and mobile phones have taken over many aspects of our lives, the healthcare industry is no exception.

-Most of our healthcare data is being migrated from paper charts to electronic ones, in many cases motivated by massive governmental incentives.

-Meanwhile, countless medical mobile applications are being written to track vital signs, including heart rates and weights, and even communicating with doctors

Reason for Healthcare analytics improves medical care

-The effectiveness of medical care is commonly measured using the so-called healthcare triple aim: improving outcomes, reducing costs, and ensuring quality

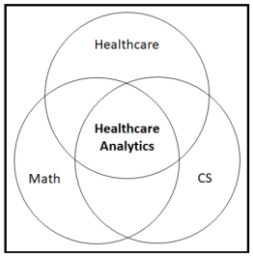
Better outcomes ***[2M]***

On a personal level, everyone can relate to better **healthcare outcomes**. We yearn for better outcomes in our own lives whenever we visit a doctor or a hospital. Specifically, here are some of the things about which we are concerned:

Accurate diagnosis
Effective treatment
No complications
An overall improved quality of life:
Ensure quality

Foundations of healthcare analytics [13 M]

Healthcare analytics can be viewed as the intersection of three fields: healthcare (**Healthcare Analytics**), mathematics (**Math**), and computer science (**CS**)



Healthcare

Healthcare is the **domain-knowledge** pillar of healthcare analytics. Here are some of the significant healthcare areas of knowledge that comprise healthcare analytics:

Healthcare delivery and policy: An understanding of how the healthcare industry is structured, who the major players in healthcare are, and where the financial incentives lie can only help us in improving healthcare analytics endeavours.

Healthcare data: Healthcare data is rich and complex, whether it is structured or unstructured. However, healthcare data collection often follows a specific template. Knowing the details of the typical **history and physical examination** (**H&P**) and how data is organized in a medical chart goes a long way in helping us turn that data into knowledge.

Clinical science: A familiarity with medical terminology and diseases helps in knowing what's important in the vast ocean of medical information. Clinical science is commonly divided into two areas:
physiology, or how the human body functions normally, and **pathology**, or how the human body functions with a disease. Some basic knowledge of both can be beneficial in doing effective healthcare analytics.

Mathematics

High school mathematics: Subjects such as algebra, linear equations, and precalculus are essential foundations for the more advanced math topics seen in healthcare analytics.

Probability and statistics: Believe it or not, every medical student takes a class in biostatistics during their training. Yes, effective medical diagnosis and treatment rely heavily on probability and statistics, including concepts such as sensitivity, specificity, and likelihood ratios.

Linear algebra: Commonly, the operations done on healthcare data while making machine learning models are vector and matrix operations. You'll effectively perform plenty of these operations as you work with NumPy and scikit-learn to make machine learning models in Python.

Calculus and optimization: These last two topics particularly apply to neural networks and deep learning, a specific type of machine learning that consists of layers of both linear and nonlinear transformations of data. Calculus and optimization are important for understanding for how these models are trained.

Computer science

Here are some of the significant computer science domains that comprise healthcare analytics:

Artificial intelligence: At the center of healthcare analytics is artificial intelligence or the study of systems that interact with their environment. Machine learning is a subarea within artificial intelligence, in

which predictions are made about future events using information from previous events. The models that we will study in the later parts of this book are machine learning models.

Databases and information management: Healthcare data is often accessed using **relational databases**, which can often be dumped by **electronic medical record** (**EMR**) systems on demand, or which are located in the cloud. **SQL** (short for **Structured Query Language**) can be used to select the specific data we are interested in and to make transformations on that data.

Programming languages: A programming language provides an interface between the human programmer and the ones and zeros inside of a computer. A programming language allows a programmer to provide instructions to the computer to make calculations on data that humans cannot practically do. In this book, we will use Python, a popular and emerging programming plan comprehensive, and features plenty of machine learning libraries.

Software engineering: Many of you are presumably learning about healthcare analytics because you are interested in deploying production-grade healthcare applications in your workplace. **Software engineering** is the study of the effective and efficient building of software systems that satisfy user and customer requirements.

Human-computer interaction:

History of healthcare analytics **[13 M]**

- The origin of healthcare analytics can be traced back to the *1950s.*

- At the time, medical records were still on paper, regression analysis was done by hand, and there were no incentives given by the government for pursuing value-based care. Nevertheless, there was a burgeoning interest in developing automated applications to diagnose and treat human disease, and this is reflected in the scientific literature of the time.
- For example, in *1959*, the journal Science published an article entitled "Reasoning Foundations of Medical Diagnosis," by Robert S. Ledley and Lee B. Lusted explains mathematically how physicians make a medical diagnosis (Ledley and Lusted, 1959). The paper explains many concepts that are central to modern biostatistics, although at times using terminology and symbols that we may not recognize today.
- In the *1970s*, as computers gained prominence and became accessible in academic research centers, there was a growing interest in developing **medical diagnostic decision support** (**MDDS**) **systems**, an umbrella term for broadly based, all-in-one computer programs that pinpoint medical diagnoses when input with patient information.
- The INTERNIST-1 system is the most well-known of these systems and was developed by a group of researchers at the University of Pittsburgh in the 1970s (Miller et al., 1982).
- Described by its inventors as "an experimental program for computer assisted diagnosis in general internal medicine," the INTERNIST system was developed over 15 person-years of work and involved extensive consultation with physicians. Its knowledge base spanned 500 individual diseases and 3,500 clinical manifestations across all medical subspecialties.
- The user starts by entering positive and negative findings for a patient, after which they can check a list of differential diagnoses and see how they change as new findings are added. The program intelligently asks for specific test results until a clear diagnosis is achieved. While it showed initial promise and captured the imagination of the medical world, it ultimately failed to enter the mainstream after its recommendations were outperformed by those made by a panel of leading physicians. Other reasons for its demise (and the demise of MDDS systems in general) may include

the lack of an inviting visual interface (Microsoft Windows had not been invented yet) and the fact that modern machine learning techniques were yet to be discovered.

- In the *1980s*, there was a rekindled interest in artificial intelligencetechniques that had largely been extinguished in the late 1960s, after the limitations of perceptrons had been explicated by Marvin Minsky and Seymour Papert in their book, *Perceptrons* (Minsky and Papert, 1969). The paper "*Learning representations by back-propagating errors*" by David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams was published in *Nature* in 1986 and marked the birth of the back-propagation-trained, nonlinear **neural network**, which today rivals humans in its performance on a variety of artificial intelligence, such as speech and digit recognition (Rumelhart et al., 1986).
- It took only a few years before such techniques were applied to the medical field. In 1990, William Baxt published a study entitled "*Use of an Artificial Neural Network for Data Analysis in Clinical Decision-Making: The Diagnosis of Acute Coronary Occlusion*" in the journal *Neural Computation* (Baxt, 1990). In the study, an artificial neural network outperformed a group of medical physicians in diagnosing heart attacks using findings from **electrocardiograms** (**EKGs**). This pioneering study helped to open the flood gates for a tsunami of biomedical machine learning research that persists even today. Indeed, searching for "machine learning" using the biomedical search engine PubMed returns only 9 results in 1990 and over 4,000 results in 2017, with the results steadily increasing in the intervening years:
- Several factors are responsible for this acceleration in biomedical machine learning research. The first is the increasing number and availability of machine learning algorithms. The neural network is just one example of this. In the 1990s, medical researchers began using a variety of alternative algorithms, including recently developed algorithms such as decision trees, random forests, and support vector machines, in addition to traditional statistical models, such as logistic and linear regression.
- The second factor is the increased availability of electronic clinical data. Prior to the 2000s, almost all medical data was on paper charts and conducting computerized machine learning studies meant hours of manually entering the data into computers. The growth and eventual spread of electronic medical records made it much simpler to use this data to make machine learning models. Additionally, more data meant more accurate models.

Today's modern neural networks (commonly referred to as *deep learning* networks) are commonly outperforming humans in tasks that are more complex than EKG interpretation, such as cancer recognition from x-ray images and predicting sequences of future medical events in patients. Deep learning often achieves this using millions of patient records, coupled together with parallel computing technology that makes it possible to train large models in shorter time spans, as well as newly developed techniques for tuning, regularizing, and optimizing machine learning models. Another exciting occurrence in present healthcare analytics is the introduction of governmental incentives to eliminate excessive spending and misdiagnosis in healthcare. Such incentives have led to an interest in healthcare analytics not just from academic researchers, but also from industrial players and companies looking to save money for healthcare organizations (and to make themselves some money as well).

While healthcare analytics and machine algorithms aren't redefining medical care just yet, the future for healthcare analytics looks bright. I like to imagine a day when hospitals, equipped with cameras, privately and securely record every aspect of patient care, including conversations between patients and physicians and patient facial expressions as they hear the results of their medical tests. These words and images could then be passed to machine learning algorithms to predict how patients will react to future results, and what

those results will be in the first place. But we are getting ahead of ourselves; before we arrive at that day, there is much work to be done!

Examples of healthcare analytics [13 M]

Using visualizations to elucidate patient care

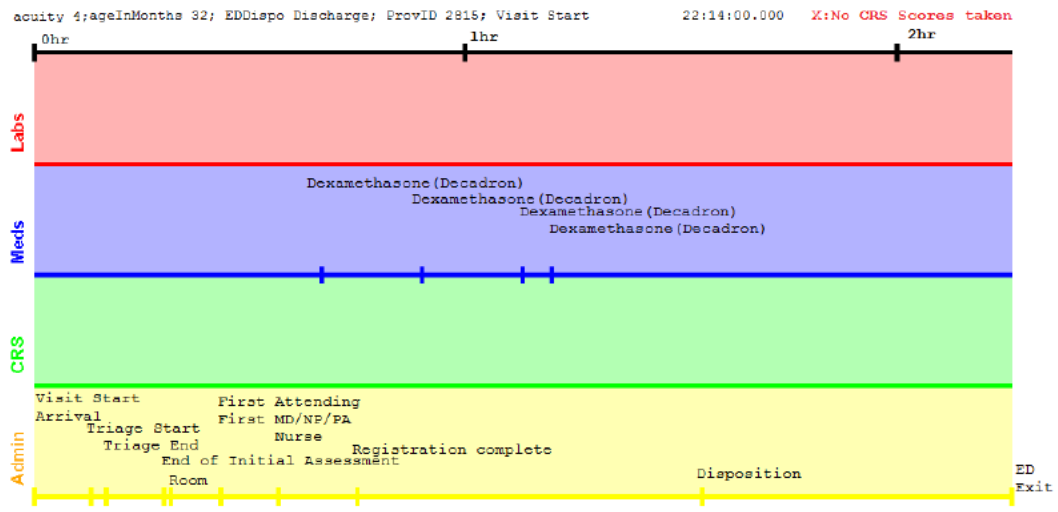
Analytics is often divided into three subcomponents–
Descriptive analytics,
Predictive analytics, and
Prescriptive analytics.

Descriptive analytics:

Descriptive analytics uses the previously discussed analytic techniques to better describe or summarize the process under study. Understanding how care is delivered is one process that stands to benefit from descriptive analytics.

Example of a visualization of a toddler's **emergency department** (**ED**) care record is when they presented complaining of an asthma exacerbation (Basole et al., 2015). It uses structured clinical data commonly found in EMR systems to summarize the temporal relationships of the care events they experienced in the ED. The visualization consists of four types of activities–administrative (yellow), diagnostic (green), medications (blue), and lab tests (red). These are encoded by color and by *y*-position. Along the *x*-axis is time. The black bar on top is divided by vertical tick marks into hour-long blocks. This patient's visit lasted a little over two hours. Information about the patient is displayed before the black time bar.

While descriptive analytical studies such as these may not directly impact costs or medical care recommendations, they serve as a starting point for exploring and understanding the patient care and often paving the way for more specific and actionable analytical methods to be launched:



Predictive analytics:

Predicting future diagnostic and treatment events

A central problem in medicine is identifying patients who are at risk of developing a certain disease. By identifying high-risk patients, steps can betaken to hinder or delay the onset of the disease or prevent it altogether.

This is an example of predictive analytics at work–using information fromprevious events to make predictions about the future. There are certain diseases that are particularly popular for prediction research: congestive heart failure, myocardial infarction, pneumonia, and chronic obstructive pulmonary disease are just a few examples of high-mortality, high-cost diseases that benefit from early identification of high-risk patients.

Not only do we care about what diseases will occur in the future, we are also interested in identifying patients who are at risk of requiring high-costtreatments, such as hospital readmissions and doctor visits. By identifying these patients, we can take money-saving steps proactively to reduce the risk of these high-risk treatments, and we can also reward healthcare organizations that do a good job.

This is a broad example with several unknowns to consider. First: what specific event (or disease) are we interested in predicting? Second: what data will we use to make our predictions? Structured clinical data (data organized as tables) drawn from electronic medical records is currently the most popular data source; other possibilities include unstructured data (medical text), medical or x-ray images, biosignals (EEG, EKG), data recorded from devices, or even data from social media.

Measuring provider quality and Performance

While making visualizations or predictions represents the sexier aspectsof healthcare analytics, there are other types of analytics that are also important. Sometimes, it boils down to good, old number crunching.

Monitoring the performance of physicians and healthcare organizations using healthcare measures is a good example of this analytical technique. Healthcare measures provide a mechanism by which individualscan measure and compare the compliance of participating providers on evidence-based medical recommendations. For example, it is a widely accepted recommendation that patients with diabetes receive foot exams todetect diabetic foot ulcers every three months by a physician.

A state-sponsored healthcare measure might specify guidelines for calculating the number of diabetic patients receiving care at an institution,and then determine the percentage of those patients that received appropriate foot care. Similar measures would exist for the common heart,lung, and joint diseases, among many others. This provides a way to identify the providers that provide the highest quality care, and these recommendations can be downloaded for public consumption.

Prescriptive analytics:

Patient-facing treatments for Disease

In rare cases, healthcare analytics comprise medical technologies that are used to actually treat diseases, not just perform research on them. An example of this is **neuroprosthetics** **[2M]**. Neuroprosthetics can be defined as the enhancement of nervous system function using man-made devices.

Neuroprosthetics research has enabled patients with disabilities such as blindness or paraplegia to recover some of their lost function. For example,a paralyzed patient may be able to move a computer cursor on a screen not with their hand, but by using their brain signals! In this specific application,

recordings of the electrical activity of specific neurons are obtained, and a machine learning model is used to determine in which direction the cursor should move given the firing of the neurons. Similar analytics can be used for visual impairments, or for visualizing what a human is seeing. A second example includes implanting devices in the body that detect seizures beforethey occur and proactively administer preventive medication. Certainly, the sky is the limit for analytic-driven treatments.

Healthcare policy [13M]

Healthcare reform needs support from legislators to succeed. Certain legislation has paved the way for patients' rights and privacy, due to the rise of EMRs, value-based care, and the advancement of big data in healthcare.

1.6.1 Protecting patient privacy and patient rights

Many countries around the world have enacted legislation for the protection of patient privacy. In the United States, legislation for protecting patient privacy was first signed into law in 1996 and is known as the **Health Insurance Portability and Accountability Act (HIPAA)**. It has been revised and updated several times since then.
Two of HIPAA's main components are the **Privacy Rule** and the **Security Rule.**

The **Privacy Rule** states the specific situations for which healthcare data can be used. In particular, any information that can be used to identify the patient (known as **Protected Health Information (PHI)**) can be freely used for the purposes of medical treatment, bill payments, or other certain healthcare operations.
Any other uses of the data require written authorization from the patient. A covered entity is an organization that is required to comply with HIPAA law.
Eg: In 2013, the **Final Omnibus Rule** extended the jurisdiction of HIPAA to include business associates or independent contractors of the covered entities .
Therefore, if you work with healthcare data in the United States, you must protect your patients' data or face the risk of fines and/or imprisonment.

Protection of electronic patient health information (e-PHI):

The Security Rule breaks down the safeguarding methods into three categories: administrative, physical, and technical. Specifically, according to the website of the US Department of Health and Human Services, healthcare data scientists should:

"ensure the confidentiality, integrity, and availability of all e-PHI" in their possession; protect against "reasonably anticipated threats" to the security of the information and impermissible uses or disclosures; and "ensure compliance by their workforce"

Guidelines listed by the HHS website for safeguarding techniques by the US Dept:
Covered entities and business associates should designate a privacy officer in charge of HIPAA enforcement and maintain training programs for employees who have access to e-PHI.

- Access to hardware and software containing e-PHI should be carefully controlled, regulated, and limited to authorized individuals
- e-PHI sent over open networks (for example, via email) must be encrypted
- Covered entities and business associates are required to report any breaches of security to affected individuals and the Department of Health and Human Services

1.6.2 Advancing the adoption of electronic medical records

EMRs, together with healthcare analytics, are seen as a possible remedy to escalating healthcare costs. In the United States, the major piece of legislation that has promoted the use of EMRs is the **Health Information Technology for Economic and Clinical Health (HITECH)** Act, passed in 2009 as part of the American Recovery and Reinvestment Act (Braunstein, 2014). The HITECH Act provides incentive payments to healthcare organizations that do two things:

1. Adopt the use of "certified" **electronic health records (EHRs)**
2. Use the EHRs in a meaningful fashion. Starting in 2015, healthcare providers who did not use EHRs were subject to penalization from their Medicare reimbursement

In order for an EHR to be certified, it must meet several dozen criteria. Examples of such criteria include those that support clinical practice, such as allowing for computerized physician order entry and recording demographic and clinical information about patients, such as medication lists, allergy lists, and smoking statuses.
Other criteria focus on maintaining the privacy and security of medical information and they call for secure access, emergency access, and access timeout after a period of inactivity. The EHR should also be able to submit clinical quality measures to the appropriate authorities. Full lists of such criteria are available at www.healthit.gov.

It is not enough for providers to have access to a certified EHR; in order to receive an incentive payments, providers must use the EHR in a meaningful fashion, as stipulated by the meaningful use requirements. Again, dozens of requirements exist, some of which are mandatory, and some optional. These requirements are distributed across the following five domains:

- Improving care coordination Reducing health
- disparities Engaging patients and their families
- Improving population and public health Ensuring
- adequate privacy and security

The rise of EHRs will lead to an unprecedented volume of clinical information becoming available for subsequent analysis in efforts to cut costs and improve outcomes.

1.6.3 Promoting value-based care

The **Patient Protection and Affordable Care Act (PPACA)**, also known as the **Affordable Care Act (ACA**), was passed in 2010. It is a mammoth piece of legislation that is most well- known for its attempt to reduce the uninsured population and to provide health insurance subsidies for the majority of citizens. Some of its lesser publicized provisions, however, added new value-based reimbursement models discussed earlier in the chapter (namely, bundled payments and accountable care organizations), and created the four original value-based programs:

- Hospital Value-Based Purchasing Program (HVBP)
- Hospital Readmission Reduction Program (HRRP)
- Hospital Acquired Conditions Reduction Program (HAC)
- Value Modifier Program (VM)

The Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) initiated the Quality Payment Program, composed of both the **Alternative Payment Models (APM)** program and the **Merit-Based Incentive Payments System (MIPS)**. The US healthcare system further away from FFS reimbursement toward value-based reimbursement.

1.6.4 Advancing analytics in healthcare

There are a handful of legal initiatives that are related to advancing analytics in healthcare. The most relevant of these is the **All of Us** initiative (formerly known as the **Precision Medicine Initiative)**, which was enacted in 2015, and aims to collect health and genetic data from one million people by 2022 in an effort to advance precision medicine and medicine tailored to individuals.

Additionally, the following three initiatives, while not directly related to analytics, may indirectly increase funding for analytics research in healthcare.
The Brain Initiative, passed in 2013, has the goal of radically improving our understanding of brain-related and neurological diseases such as Alzheimer's and Parkinson's disease.
Cancer Breakthroughs 2020, passed in 2016, is focused on finding vaccines and immunotherapies against cancer.
The 21st Century **Cures Act of 2016** streamlines the **Food and Drug Administration** (FDA) drug approval process, among other provisions.

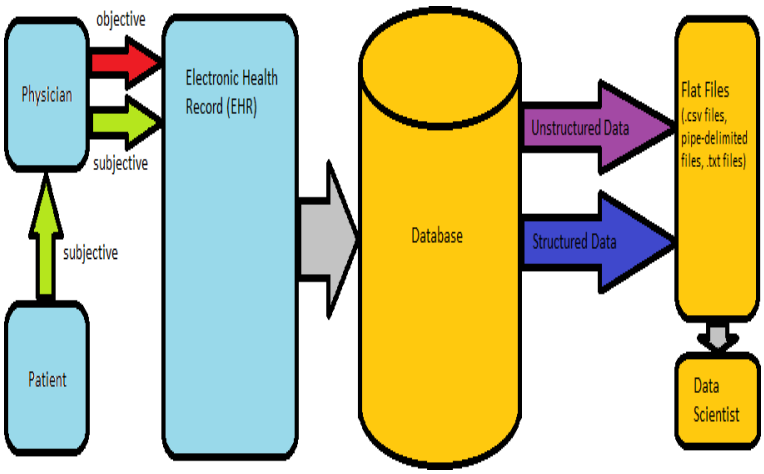
Patient data – the journey frompatient to computer [13 M]

The clinical data collection process starts when a patient starts telling a physician about his or her condition. This is known as the **patient history**, and since it is not observed directly by the physician, but instead recountedby the patient, the patient's story is known as **subjective information**. In contrast, **objective information** comes from the physician and consists of the physician's own observations about the patient, from the physical examination, lab tests, and imaging studies, to other diagnostic procedures.Together, the subjective and objective information makes up the clinical note.

There are several types of clinical notes used in healthcare. The **history and physical** (**H&P**) is the most thorough and comprehensive clinical note. It is usually obtained when an **outpatient** physician sees a patient for the first time, or when a patient is first admitted to the hospital. Collecting all the data from the patient and typing up the H&P on the hospital computer may take a total of 1-2 hours for a single patient. Usually, an H&P is only done once per physician/hospital admission. For successive outpatient visits, or an **inpatient** admission lasting several days, briefer clinical notes are compiled. These are termed **progress notes**, or **SOAP notes** (SOAP stands for subjective, objective, assessment, and plan). In these notes, the focus is on events that have occurred since the initial H&P or the previous progress note.

Before patient data appears in your database, it makes a long journey, starting from the patient history as interpreted by the physician team. Thepatient story is combined with other pieces of information from different clinical departments (for example, laboratory, and imaging) to form the **electronic health record** (**EHR**). When the hospital wants to make the data available to a third party for further analysis, it typically releases the data to the cloud in a database format.

Once the data is captured in a database system, the analytics professional can then use a variety of tools to visualize, pivot, analyze, and build predictive models:



In the following subsections, we will describe the important aspects of thesetwo types of clinical notes.

The history and physical (H&P)

The history and physical is the most comprehensive type of documentation available for patients and is usuallyconducted upon their admission to the hospital and/or when seeing new outpatient physicians. The standard sections of the H&P clinical note are discussed in the following sections.

Metadata and chief complaint

The metadata includes basic information about the patient's visit, such as the patient's name, date of birth, date/time of admission, and the name ofthe admitting hospital and attending physician.

The chief complaint is the reason for the patient's visit/hospitalization, usually in the patient's own words. **Example**: "I'm having some chest discomfort." This chief complaint may, or may not, be translated by thehistory taker into the corresponding medical terminology, for example,"chest pain."

History of the present illness (HPI)

The HPI includes details surrounding the chief complaint. This section is often split into two paragraphs as follows:

The first paragraph provides the immediate details surrounding the chief complaint, usually using information obtained from the patient. The first sentence often provides important demographic details about the patient andany relevant details about the past medical history, in addition to the chief complaint. For example:

"Mr. Smith is a 53-year-old Caucasian male with a history of hypertension, hyperlipidemia, diabetes,and smoking, who presents to the emergency room complaining of chest pain."

Regarding the rest of the paragraph, a first HPI paragraph usually contains the seven standard elements listed here. These seven elements tend to assume that the chief complaint is some type of pain; some chief complaints(for example, amenorrhea) require different sets of questions. The seven elements are summarized in the following table:

HPI element	Corresponding question	Example answer
Location	Where is the pain located?	The pain is left-sided and radiates to the left arm andback.
Quality	What does the pain feel like?	Patient reports a shooting,stabbing pain.
Severity	On a scale of 1-10, how bad is the pain?	Severity is 8/10.
Timing	**Onset**: When did the pain first start? **Frequency**: How often does the pain occur? **Duration**: How long are the pain episodes?	The current episode began half an hour ago. Episodes have occurred for a few months, following exercise, and for periods of up to 15-20 minutes.
Exacerbating Factors	What makes the pain worse?	Pain is exacerbated byexercise.
Alleviating Factors	What relieves the pain?	Pain is relieved by rest andweight loss.
Associated Symptoms	Do you notice any other symptoms whenthe pain is present?	Patient reports symptomsassociated with dyspnea.

The second paragraph should contain all the previous medical care the patient has already received for their ailment. Typical questions include: Have they seen a physician already or been hospitalized previously? What labs and tests were performed? How well controlled are the patient's medical conditions relevant to the chief complt? Which treatments werepreviously tried? Is there a copy of the x-ray?

Past medical history

This part of the H&P lists all current and previous medical conditions that affect the patient, including, but not limited to, hospitalizations (whether for medical, surgical, or psychiatric reasons).

Medications

Current prescription and **over-the-counter** (**OTC**) medications are provided in this section, usually with the following details: medication name, dose, route of administration, and frequency. Every medication listedshould correspond to one of the patient's current medical conditions given in the past medical history. The route of administration and frequency are often written using abbreviations; refer to the following table for a list of common abbreviations.

Family history

The family history includes a disease history for family members up to twogenerations preceding the patient, with an emphasis on chronic diseases as well as diseases relevant to the chief complaint and the organ systems affected.

Social history

The social history provides details of social and risk factor information notobtained in the HPI. Included in this section are demographic factors not previously mentioned, occupation (and any occupational

exposures to dangerous substances if applicable), social support (marriage, children, dependents), and substance use/abuse (tobacco, alcohol, recreational/illicitdrugs).

Allergies

The allergies section commonly includes substances to which the patient ishypersensitive, including drugs, and the corresponding reaction. If the patient has no known drug allergies, it is often abbreviated using the acronym NKDA.

Review of systems

The **review of systems** (**ROS**) serves as a final screening for significant symptoms after the other parts of the history have been obtained. In this section, the patient is asked about experiencing symptoms relevant to different functional organ systems of the body (for example, gastrointestinal, cardiovascular, and pulmonary). An emphasis is placed on organ systems and symptoms relevant to the chief complaint. Symptoms foras many as 14 different organ systems may be touched upon

Physical examination

The physician proceeds to examine the patient and records the findings inthis section. The description usually starts with general patient well-beingand appearance, followed by pertinent vital signs (see table for additionaldetails), before proceeding with the **head, eyes, ears, nose, and throat** (**HEENT**), and continuing down the body with specific organs/organ systems

Additional objective data (lab tests,imaging, and other diagnostic tests)

The physical examination marks the beginning of what is called objective data, or data about the patient that is observed, interpreted, and recorded bythe physician. This is in contrast to subjective data, which is information provided to the physician by the patient first-hand, and which includes the patient history. After the physical examination, all additional objective dataabout the patient is provided. This includes the results of any lab tests, imaging studies if applicable, and any other tests specific to the present illness that may have been performed. Common imaging studies include **x-rays** (**XR**), **computed tomography** (**CT**) scans, and **magnetic resonance imaging** (**MRI**) scans of the body region of interest.

Assessment and plan

This is the final part of the H&P. In the assessment, the physician consolidates all of the subjective and objective data of the previous sectionsto make a concise summary of the chief complaint, along with significant findings from the history, physical examination, and additional tests. The physician lists the most likely causes of the patient's condition, in an itemized manner for each distinct group of complaints/findings. In the plan,the physician discusses the blueprint for treating the patient, again in an itemized fashion.

The progress (SOAP) clinical note [2 M]

The SOAP note, as stated previously, is typically done on a daily basis for patients admitted to a hospital and includes one section for every letter in itsacronym: **subjective, objective, assessment, and plan** (**SOAP**). The subjective section focuses on any new complaints the patient is having, or had, on the previous night. The objective section includes the daily and focused physical examination and lab, imaging, and test results from the previous day. The assessment and plan are similar to those of the H&P, updated from previous notes with all of the day's events taken into consideration.

Standardized clinical codesets [13M]

Being philosophical for a moment, every known object that has a significant importance attributed to it has a name. The organs that are used to read these words are known as eyes. The words are written on pieces of paper called pages. To turn the pages, you use your hands. These are all objects that we have named so that we can identify them easily.

In healthcare, important entities—diseases, procedures, lab tests, drugs, symptoms, bacteria species, for example, have names and identities too. For example, the failure of the heart valves to pump blood to the rest of the body is known as heart failure. ACE inhibitors are a class of drugs used to treat heart failure. A problem arises, however, when healthcare industry workers associate the same entity with different identities. For example, one physician may refer to "heart failure" as "congestive heart failure", while another may refer to it as "CHF." Also, there are varying levels of specificity: a third doctor may call it "**systolic heart failure**" to indicate that the dysfunction occurs during the systolic phase of the heartbeat. In medicine, accuracy and specificity are of the utmost importance. To ensure that all members of the healthcare team are talking and thinking about the same thing, use Clinical codes.

Clinical codes can be thought of as unique identities for medical concepts. Each code is typically comprised of a pair of objects: an alphanumeric code and a verbal description of the entity that the code represents. For example, In the ICD10-CM coding system, the code I50.9 represents "Heart failure, unspecified." There are additional, more specific codes to represent more specific heart failure diagnoses when they are known.
Some of the more important standardized coding systems include **the International Classification of Disease (ICD)** for medical diagnoses, **Current Procedural Terminology (CPT)** for medical procedures, **Logical Observation Identifiers Names and Codes (LOINC)** for laboratory tests, **National Drug Code (NDC)** for drug therapies, and **Systematized Nomenclature of Medicine (SNOMED)** for all of the preceding and more.

1.7.1 International Classification of Disease (ICD)

Diseases and conditions are usually coded using the ICD coding system. ICD was started in1899 and is revised (every 10 years) and maintained by the **World Health Organization** (WHO). As of 2016, the tenth revision (ICD-10) is the most recent and consists of more than 68,000 unique diagnostic codes, more than any previous revision.

ICD-10 codes may consist of up to eight alphanumeric characters. The first three characters indicate the major disease category; for example, "N18" specifies chronic kidney disease. These characters are followed by a period and then the remaining characters, which can provide an extraordinary amount of clinical detail (Braunstein, 2014). For example, code "C50.211" specifies "malignant neoplasm of the upper-inner quadrant of the right female breast." With all of its precision, ICD-10 facilitates the application of analytics in healthcare.

1.7.2 Current Procedural Terminology (CPT)

Medical, surgical, diagnostic, and therapeutic procedures are coded using the CPT coding system. Developed by the **American Medical Association (AMA),** CPT codes consist of four numeric characters followed by a fifth alphanumeric character. Commonly used CPT codes include those for outpatient visits, surgical procedures, radiological tests, anaesthetic procedures, history and physical examination, and emerging technologies. Unlike the ICD, the CPT is not a hierarchical coding system. Some concepts, however, have multiple codes depending on factors such as the visit length (for outpatient visits) or the amount of tissue removed (for surgical procedures).

1.7.3 Logical Observation Identifiers Names and Codes (LOINC)

Laboratory tests and observations are coded using the LOINC coding system. Written and maintained by the Regenstrief Institute, there are over 70,000 codes, each of which is a six-digit number, the last number being

separated by the other numbers with a hyphen. Like CPT codes, a specific type of laboratory test (for example, white blood cell (WBC) count) often has multiple codes that vary, depending on the timing of the sample, the measurement units, the measurement method, and so on. While each code contains a large amount of information, this may pose a problem when trying to find a code for a lab test such as a WBC count when not all relevant information is known.

1.7.4 National Drug Code (NDC)

The NDC is maintained by the US FDA. Each code is 10 digits long and has three sub components:

- A labeler component, which identifies the manufacturer/distributor of the drug.
- A product component, which identifies the actual drug from the labeler, including strength, dosage, and formulation.
- A package code, which identifies the specific package shape and size.

1.7.5 Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)

SNOMED-CT is a huge coding system that uniquely identifies over 300,000 clinical concepts.
These concepts may be diseases, procedures, labs, drugs, organs, infectious agents, infections, symptoms, clinical findings, and more.
Additionally, SNOMED-CT defines over 1.3 million relationships between these concepts. SNOMED-CT is maintained by the National **Institutes of Health (NIH)** and is a subset of an even larger coding system, SNOMED, which includes concepts not relevant to clinical practice. The NIH has a software program called MetaMap (https://metamap.nlm.nih.gov/), which can be used to tag clinical concepts appearing in text, making it useful for natural language processing in healthcare.

While coding systems cannot uniquely identify every clinical concept with all of their variations and nuances, they come reasonably close, and, in so doing, make certain activities in medicine (particularly billing and analytics) much easier.

1.8 Breaking Down Health Care Analytics [13 M]

In healthcare, the problems to solve can be broken down into four categories:

Population
Medical task
- Screening
- Diagnosis
- Outcome/Prognosis
- Response to treatment

Data format
- Structured
- Unstructured
- Imaging
- Other data format

Disease
- Acute versus chronic diseases
- Cancer
- Other diseases

1.8.1 Population

In healthcare, patient **populations** are what make groups of patients, and therefore, their data anddisease characteristics—homogeneous.
Examples of patient populations include inpatients, outpatients, emergency room patients, children, adults, and US citizens. Geographically, populations can even be defined at the state, municipal, or local levels.

1.8.2 Medical task

In healthcare practice, the evaluation and treatment of patients can be broken down into different cognitive subtasks. Each of these tasks can potentially be aided by using analytics.
Screening, diagnosis, prognosis measurement, outcome measurement, and response to treatment are some ofthese basic tasks, and we will look at each one in turn.

Screening

Screening can be defined as the identification of a disease in a patient before the onset of signs and symptoms. This is important because in manydiseases, particularly chronic diseases, early detection coincides with early treatment, better outcomes, and lower costs to the healthcare provider.

Screening for some diseases has greater potential benefits than screeningfor others. In order for disease screening to be worthwhile, several conditions, as listed here, must be met (Martin *et al.*, 2005):

- The outcome must be alterable at the time of identifying the diseaseThe screening technique should be cost-effective
- The test should have high accuracy
- The disease should carry a large burden on the population

An example of a popular screening problem and solution is using the **Pap smear** to screen for cervical cancer; women are recommended to undergo this cost-effective test every 1-3 years throughout most of their lives. **Lungcancer** screening is an example of a problem that has yet to find an ideal solution; while using x-rays to screen for lung cancer may be accurate andmay lead to earlier detection in some cases, x-rays are costly and expose patients to radiation, and there is no strong evidence that early detection influences the prognosis or outcome (Martin *et al.*, 2005). Increasingly, machine learning models are being developed in lieu of medical tests to screen for diseases including cancer, heart disease, and strokes
.

Diagnosis

Diagnosis can be defined as the identification of a disease in an individual.In contrast to screening, diagnosis may happen at any time during the course of the disease. Diagnosis is important for almost every disease because it dictates how the signs or symptoms (and the underlying disease)should be treated. The exception occurs when diseases have no effective treatment, or when differentiating between diseases does not change the treatment.

A common use of machine learning in diagnosis problems is to identify potential causes of underlying disease in the face of a mysterious symptom, **for example, abdominal pain**. In contrast, building a machine learning model to differentiate between different types of psychiatric personality disorders may be of limited efficacy, since personality disorders are difficultto treat effectively.

Outcome/Prognosis

Healthcare is primarily concerned with producing better outcomes at a lower cost. Often, we try to determine which patients are at a high risk of a poor outcome directly, without necessarily focusing on the specific cause of their signs and symptoms. Popular outcomes for which machine learning solutions are being applied include predicting which patients will likely be readmitted to a hospital, which patients will suffer death, and which patients will be admitted to the hospital from the emergency room. Many of these outcomes are actively monitored by governments and healthcare organizations and, in some cases, governmentseven provide financial incentives to improve specific outcomes.

Often, instead of dividing outcomes into two classes (**for example, readmission versus non-readmission**), we can attempt to quantify a patient'schances of survival in terms of a specific time period, given the characteristics of the patient's disease. For example, in cancer and heart failure patients, you can attempt to predict for how many years the patient islikely to survive. This is referred to as **prognosis**, and it is also a popular machine learning problem in healthcare.

Response to treatment

In healthcare, diseases often have a variety of treatments, and predicting which treatment a patient will respond to is a problem in itself. **For example, cancer patients** can undergo a variety of chemotherapy regimens, and depressed patients have dozens of pharmacological treatments to choose from. Although this is a machine learning problem that is still in its infancy, it is gaining popularity and is also known as personalized medicine.

1.8.3 Data format

Machine learning use cases in healthcare also vary, depending on the formatof the available data. The data format often dictates what methods and algorithms can be used to solve the problem, and therefore plays an important part in determining the use case.

1) Structured

Structured data is data that can be organized into rows and columns having discrete values. Much of the patient data in an electronic health record may be stored in, or converted to, this format. In healthcare, individual patients or encounters often form the rows (or observations), and various features (**example**, demographic variables, clinical characteristics, lab observations) of the patient/encounter form the columns. Such a format is particularly conducive to performing machine learning analyses using various algorithms.

2) Unstructured

Unfortunately, much of the data in an EHR (such as that in a clinical note) consists of free-form text; this is known as **unstructured data**. Provider notes generated as part of health care delivery provide extensive information regarding the patient and the progress of a hospital visit. Depending on the complexity of the diagnoses, radiology reports, pathology reports, and otherdiagnoses, notes would also include unstructured information.

While unstructured data is capable of communicating far more extensive and valuable information about the patient, analysis of such data poses much more of a challenge than that of structured data.

3) Imaging

In certain specialities, such as radiology and pathology, data is collected using photographs and images of disease, using either photographs of lesions, pathological slides, or x-ray images. An emerging area is the automated analysis of this image data to screen, diagnose, and measure theprognosis of various diseases using these images, including benign and malignant cancers, heart disease, and strokes.

4) Other data format

The **Electrophysiological signal collection** is yet another data modality inhealthcare; collection and analysis of such signals, be it **electroencephalographic** (**EEG**) signals in epilepsy patients, or **electrocardiographic** (**EKG**) signals in heart attack patients, can be valuable for disease diagnosis and prognosis measurement. In 2014, the popular data science competition website, Kaggle, offered a $10,000 prizefor the data science team that could most effectively predict seizures in epilepsy patients using EEG data.

1.8.4 Disease

A fourth way in which use cases are permuted in healthcare is according tothe disease. Thousands of medical diseases are actively being studied in medical research, and each one represents a potential target for machine learning models. However, in machine learning, not all diseases are createdequal; some promise better potential rewards and opportunities than others.

Example: Acute versus chronic diseases

In healthcare, diseases are often classified as being acute or chronic (Braunstein, 2014). Both types of disease are important targets for predictive modeling.

Acute diseases are characterized by a sudden onset, are usually self-limited, and patients often experience a full recovery afterthe appropriate treatment. Also, risk factors for acute conditions are oftennot determined by patient behavior.

Examples of acute diseases include influenza, kidney stones, and appendicitis.

Chronic diseases, in contrast, typically have a progressive onset and last for the lifetime of the patient. They are influenced by patient behavior, such as smoking and obesity, and also by genetic factors.

Examples of chronic diseases include hypertension, atherosclerosis, diabetes, and chronic kidney disease. Chronic diseases are particularly dangerous because they tend to belinked and cause other serious chronic and acute diseases. Chronic diseases are also costly to society; billions of dollars are spent annually on preventing and treating common chronic conditions.

Acute-on-chronic diseases are particularly popular in health care predictive modelling. These are acute, sudden-onset diseases that are caused by chronic conditions.

For example, stroke and myocardial infarction are acute conditions that are by-products of the chronic conditions hypertension and diabetes. Acute-on-chronic disease modeling is popular because it allows usto filter the population to a high-risk group that has the corresponding chronic condition, increasing the yield of predictive models.

For example, To predict the onset of **congestive heart failure** (**CHF**), a useful starting place would be patients who have hypertension, which is a major risk factor. This would lead to a model with a higher percentage of true positives than if you were to randomly sample the population. In other words, if we were trying to predict CHF onset, it wouldn't be very useful to include healthy 20-year-old males in our model.

Cancer

There are several reasons why predictive modeling for cancer has become an important use case. For one thing, cancer is the second leading cause of death among medical diseases, just behind heart attacks. It's insidious onsetand course make cancer diagnosis just that bit more surprising and devastating. No one can dispute the importance of fighting cancer with every tool in our arsenal, including machine learning methods.

Second, within cancer machine learning, there are a variety of use cases thatare well-suited to being solved by machine learning.

For example, given a healthy patient, how likely is that patient to develop a particular type of cancer? Given a patient just diagnosed with cancer, can we inexpensively predict whether the cancer is benign or malignant? How longcan the patient be expected to survive? Will they likely be alive in 5 years? 10 years? To which, chemotherapy/radiotherapy regimen is the patient mostlikely to respond? What is the chance of cancer recurring once it is successfully treated? Questions like these benefit from mathematical answers that may be beyond the capabilities of a single doctor's reasoning or even that of a panel of doctors.

Model frameworks for medical decision making

Tree-like reasoning

In tree-like reasoning we closely examine the machine learning counterparts: the decision tree and the random forest.

Categorical reasoning withalgorithms and trees

In one medical decision making paradigm, the clinical problem can be approached as a **tree** or an **algorithm**. Here, an algorithm does not refer to a "machine learning algorithm" in the computer science sense; it can be thought of as a structured, ordered set of rules to reach a decision.

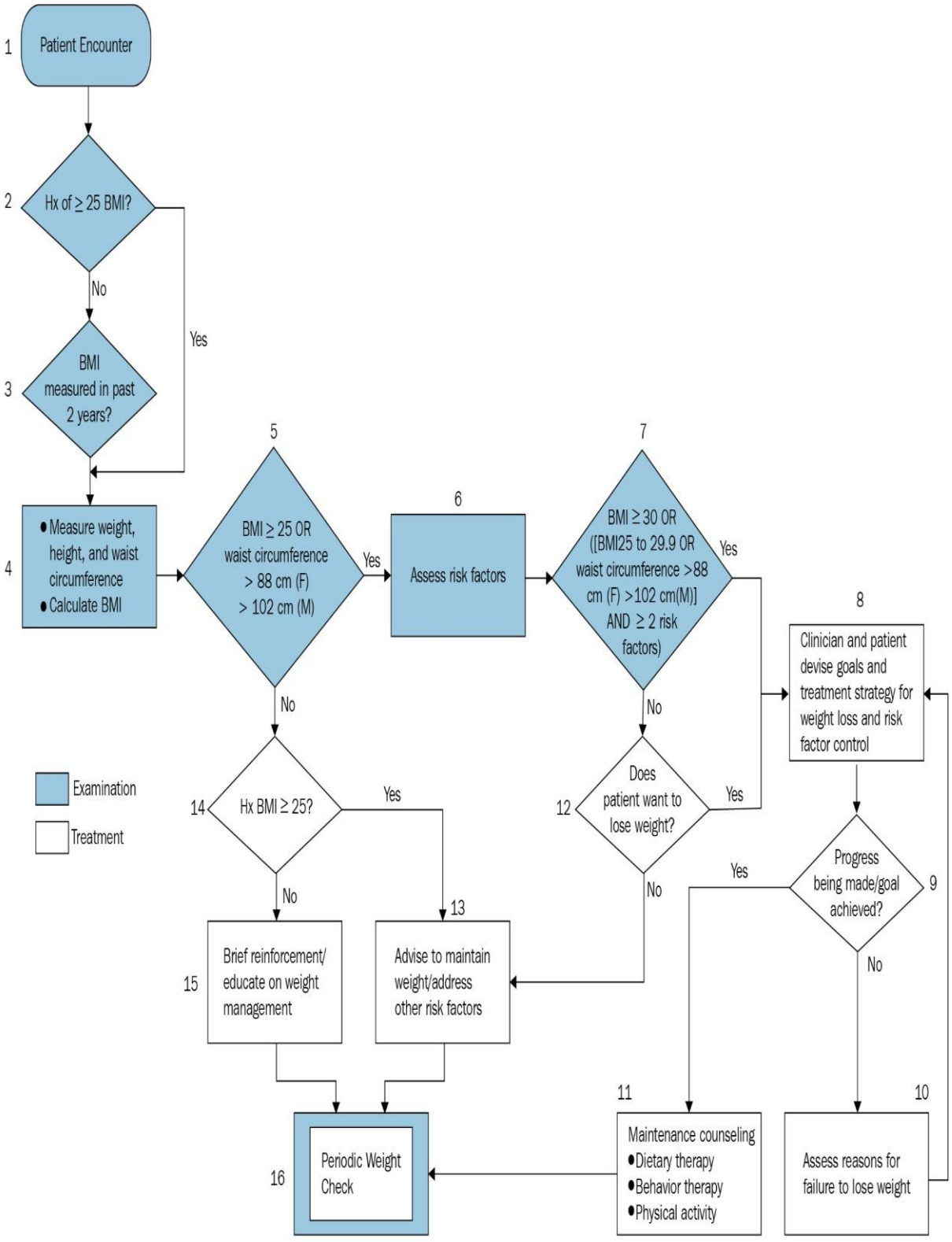
Decision Tree: [13M]

Decision trees evolved in the 1990s and used principles of information theory to optimize the branching variables/points of the tree to maximize the classification accuracy. The most common and simple algorithm for training a decision tree proceeds using what is known as a **greedy** approach. Starting at the first node, we take the training set of our data and **split** it based on each **variable**, using a variety of **cutpoints** foreach variable. After each split, we calculate the entropy or information gain from the resulting split. The measurement of how much information is gained from the split, which correlates with how even the split is.

Here the **root of the tree** represents the initiation of the **patient encounter**. As the physician learns more information while asking questions, they come to various branch or **decision points** where the physician can proceed in more than one route. These routes represent different clinical testsor alternate lines of questioning. The physician will repeatedly make decisions and pick the next branch, reaching a terminal node at which there are no more branches. The **terminal node** represents a definitive diagnosis or a treatment plan.

Example 1: of a clinical management algorithm for weight andobesity management (National Heart, Lung, and Blood Institute, 2010). Eachdecision point (most of which are binary) is a diamond, while management plans are rectangles.

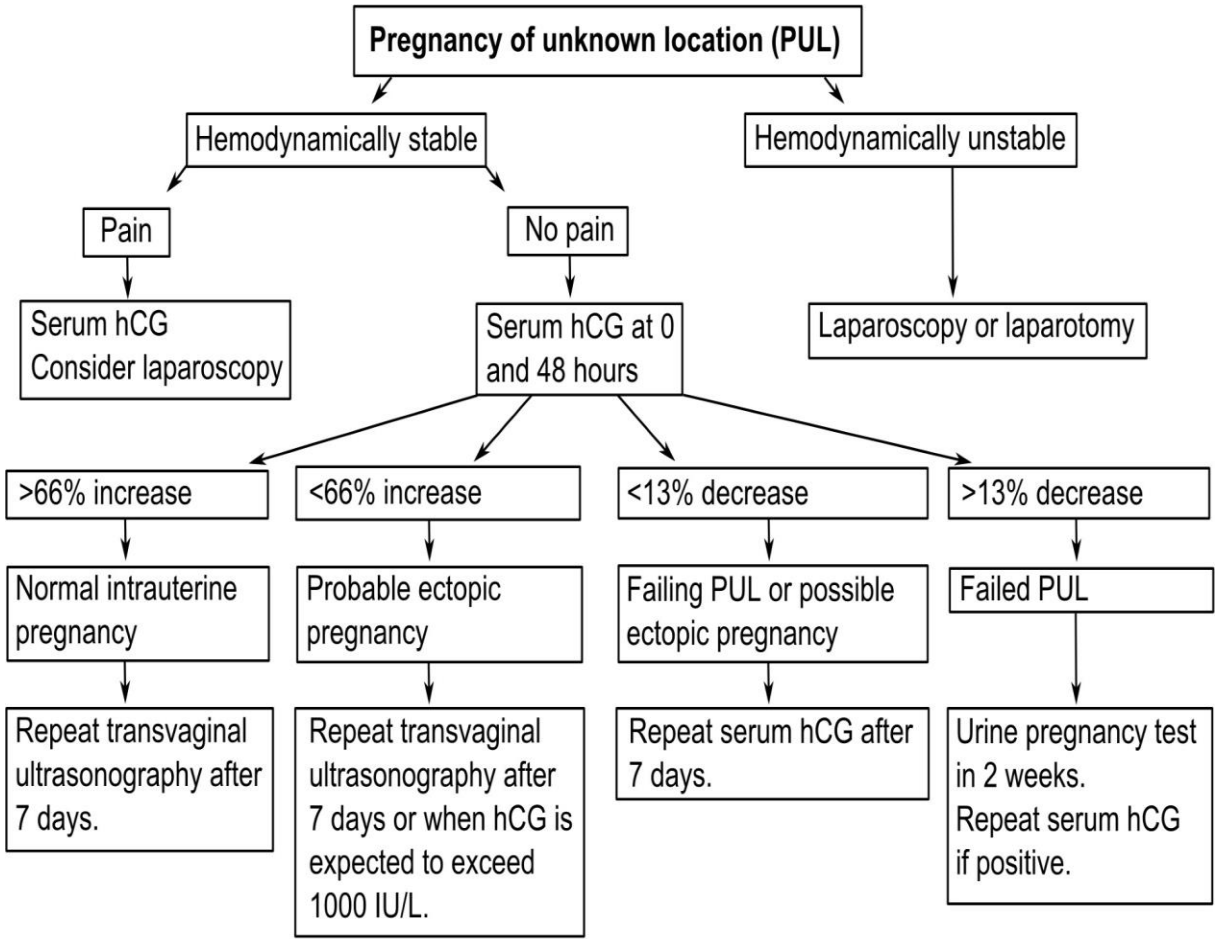
For example, suppose we have a female patient with several clinical variables that are measured: BMI = 27, waist circumference = 90 cm, andthe number of cardiac risk factors = 3. Starting at node #1, we skip from Node #2 directly to Node #4, since the BMI $\geq$ 25. At Node #5, again the answer is "Yes." At Node #7, again the answer is "Yes," taking us to the management plan outlined in Node #8:



• This algorithm applies only to the assessment for overweight and obesity and subsequent decisions based on that assessment. It does not include and initial overall assessment for cardiov ascular risk factors or diseases that are indicated

Example 2 : Algorithm that combines both diagnosis and treatment is shown as follows. In this algorithm for the **diagnosis/treatment** of pregnancy of an unknown location,a hemodynamically stable patient with no pain (a patient with stable heart and blood vessel function) is routed to have serum hCG drawn at 0 and 48 hours after presenting to the physician. Depending on the results, several possible diagnoses are given, along with corresponding management plans.

Note that in the clinical world, it is perfectly possible for these trees to be wrong; those cases are referred to as predictive errors. The goal in constructing any tree is to choose the best variables/**cutpoints** that minimizethe error:



Algorithms have a number of advantages. For one, they model humandiagnostic reasoning as sequences of hierarchical decisions or determinations. Also, their goal is to eliminate **uncertainty** by forcing the caretaker to provide a binary answer at each decision point. Algorithms havebeen shown to improve standardization of care in medical practice and are inwidespread use for many medical conditions today not only in outpatient/inpatient practice but also prior to hospital arrival by **emergency medical technicians** (**EMTs**).

However, algorithms are often overly simplistic and don't consider the factthat medical symptoms, findings, or test results may not indicate 100% certainty. They are insufficient when multiple pieces of evidence must be weighed to arrive at a decision.

In **Example 1** tree most likely uses *subjectively* determined cutpoints in deciding which route to follow. For example, Diamond #5 uses a BMI cutoff of 25, and Diamond #7uses a BMI cutoff of 30. Nice, round numbers! In the decision analysis field, trees are usually constructed based on human inference and discussion. What if we could *objectively* determine the best variables to cut (and the corresponding cutpoints at which to cut) in order to minimize the error of the algorithm?

In **Example 2**: With the PUL algorithm, a split that results in eight normal intrauterine pregnancies and seven ectopic pregnancies would be favored over a split that results in 15 normal intrauterine pregnancies andzero ectopic pregnancies. Once we have the variable and cutpoint for the best split, we proceed and then repeat the method, using the remaining variables. To prevent **overfitting** the model to the training data, we stop splitting the tree when certain criteria are reached, or alternatively, we could train a big tree with many nodes and then remove (**prune**) some of thenodes.

Limitations of Decision trees.

Decision trees must split the decision space linearly at each step based on a single variable. Another problem is that decision trees are prone to overfitting. Because of these issues, decision trees typically aren't competitive with most state-of-the-art machine learning algorithms in terms of minimizing errors.

However, the **random forest,** which is basically an ensemble of de- correlated decision trees, is currently among the most popular and accurate machine learning methods in medicine.

Probabilistic reasoning and Bayes theorem [13 M]

The mathematical way of approaching the patient involves initializing the baseline probability of a disease for a patient and updating the probability of the disease with every new clinical finding discovered about the patient. The probability is updated using Bayes theorem.

Using Bayes theorem for calculating clinical probabilities

Briefly, Bayes theorem allows for the calculation of the post-test probabilityof a disease, given a pretest probability of disease, a test result, and the 2 x 2 contingency table of the test. In this context, a "test" result does not have to be a lab test; it can be the presence or absence of any clinical finding as ascertained during the history and physical examination.

For example, the presence of chest pain, whether the chest pain is substernal, the result of an exercise stress test, and the troponin result all qualify as clinical findings upon which post-test probabilities can be calculated. Although Bayes theorem can be extended to include continuously valued results, it is most convenient to binarize the test result before calculating the probabilities.

To illustrate the use of Bayes theorem, a primary care physician and that a 55-year-old patient approaches you and says, "I'm having chest pain." When you hear the words "chest pain," the first life- threatening condition you are concerned about is a myocardial infarction. You can ask the question, "What is the likelihood that this patient is havinga myocardial infarction?" In this case, the presence or absence of chest painis the test (which is positive in this patient), and the presence or absence of myocardial infarction is what we're trying to calculate.

Calculating the baseline MIprobability

To calculate the probability that the chest-pain patient is having a **myocardial infarction** (**MI**), we must know three things:

- The pretest probability
- The 2 x 2 contingency table of the clinical finding for the disease inquestion (MI, in this case)
- The result of this test (in this case, the patient is positive for chest pain)

Because the presence or absence of other findings is not yet known in the patient, we can take the pretest probability to be the baseline prevalence ofMI in the population. Let's pretend that in your clinic's region, the baselineprevalence of MI in any given year is 5% for a 55-year-old person.

Therefore, the pretest probability of MI in this patient is 5%. Later the post-test probability of disease in this patient is the pretest probability multiplied by the likelihood ratio for positive chest pain (LR+).To get LR+, we need the 2 x 2 contingency table.

2 x 2 contingency table for chestpain and myocardial infarction

Suppose the following table is the breakdown of chest pain and myocardialinfarction in 400 patients who visited your clinic:

	Myocardial Infarction present (D+)	**Myocardial Infarction absent(D-)**	**Total**
Chest pain present (T+)	15 (TP)	100 (FP)	115
Chest pain absent (T-)	5 (FN)	280 (TN)	285
Total	20	380	400

Interpreting the contingency tableand calculating sensitivity and specificity

In the preceding table, there are four numerical cells, labeled **TP**, **FP**, **FN**, and **TN**. These abbreviations stand for **true positives**, **false positives**, **falsenegatives**, and **true negatives**, respectively. The first word (true/false) indicates whether or not the test result matched the presence of disease as measured by the gold standard. The second word (positive/negative) indicates what the test result was. True positives and true negatives are desirable; this means that the test result is correct and the higher these numbers, the better the test is. On the other hand, false positives and false negatives are undesirable.

Two important quantities that can be calculated from the true/false positives/negatives include the **sensitivity** and the **specificity**. The sensitivity is a measure of how powerful the test is in detecting disease. It isexpressed as the ratio of positive test results over the number of total patients who had the disease:

$$Sn = TP/(TP + FN)$$

On the other hand, the specificity is a measure of how good the test is atidentifying patients who do not have the disease. It is expressed as the following:

$$Sp = TN/(TN + FP)$$

These concepts can be confusing initially, so it may take some time and iterations before you get used to them, but sensitivity and specificity areimportant concepts in biostatistics and machine learning.

Calculating likelihood ratios forchest pain (+ and -)

The **likelihood ratio** is a measure of how much a test changes the likelihood of having a condition. It is often split into two quantities: the likelihood ratio of a positive test (LR+), and the likelihood ratio of a negative test (LR-).

The likelihood ratio for MI given a positive chest pain result is given by the following formulas:

$$LR+ = (TP/\left(TP + FN\right)) \,/\, (FP/\left(FP + TN\right))$$

$$LR+ = Pr\left(Positive\,Test|Positive\,Disease\right)/Pr\left(Positive\,Test|Negative\,Disease\right)$$

$$LR+ = Sensitivity/\left(1 - Specificity\right)$$

The likelihood ratio for MI given a negative chest pain result would be given by the following formulas:

$$LR- = (FN/\left(TP + FN\right)) \,/\, (TN/\left(FP + TN\right))$$

$$LR- = Pr\left(Negative\,Test|Positive\,Disease\right)/PR\left(Negative\,Test|Negative\,Disease\right)$$

$$LR+ = Sensitivity/\left(1 - Specificity\right)$$

Since the patient is positive for the presence of chest pain, only LR+ applies in this case. To get LR+, we use the appropriate numbers:

LR+ = (TP/(TP + FN)) / (FP/(FP + TN))
= (15/(15 + 5)) / (100/(100 + 280))
= 0.750 / 0.263
= 2.85

Calculating the post-test probability of MI given thepresence of chest pain

With LR+, we multiply it by the pretest probability to get the post-test probability:

Post-Test Probability = 0.05 x 2.85 = 14.3%.

The approach for diagnosis and management of the patient seems very appealing; being able to calculate an exact probability of disease seeminglyeliminates many issues in diagnosis!

Unfortunately, Bayes theorem breaks down in clinical practice for many reasons.

First, a large amount of data is required at every step to update the probability. No physician or database has access to all the contingency tables required to update the Bayes theorem with every historical element or lab test result discovered about thepatient.

Second, this method of probabilistic reasoning is unnatural for humans to perform. The other techniques discussed are much more conducive to a performance by the human brain.

Third, while the model may work for single diseases, it doesn't work well when there are multiple diseases and comorbidities.

Finally, the assumptions of conditional independence and exhaustiveness and exclusiveness that are fundamental to the Bayes theorem don't hold in the clinical world.

The reality is that symptoms and findings are not completely independent of each other; the presence or absence of one finding can influence that of many others. Together, these facts render the probability calculated by the Bayes theorem to be inexact and even misleading in most cases, even when one succeeds in calculating it. Nevertheless, Bayes theorem is important in medicine for many subproblems when ample evidence is available (for example, using chest pain characteristics to calculate the probability of MI during the patient history).

Corresponding machine learningalgorithm – the Naive Bayes Classifier

In the preceding example, we see how to calculate a post-test probability given a pretest probability, a likelihood, and a test result. The machine learning algorithm known as the Naive Bayes Classifier does this for every feature sequentially for a given observation.

For example, in the preceding example, the post-test probability was 14.3%. Let's pretend that the patient now has a troponin drawn and it is elevated. 14.3% now becomes the pretest probability, and a new post-test probability is calculatedbased on the contingency table for troponin and MI, where the contingency tables are obtained from the training data. This is continued until all the features are exhausted. Again, the key assumption is that each feature is independent of all others. For the classifier, the category (outcome) having the highest post-test probability is assigned to the observation.

The Naive Bayes Classifier is popular for a select group of applications. Itsadvantages include high interpretability, robustness to missing data, and ease/speed for training and predicting. However, its assumptions make the model unable to compete with more state-of-the-art algorithms.

Criterion tables and the weightedsum approach

The third medical decision making paradigm is the criteriontable and its similarity to linear and logistic regression.

Criterion tables

The use of criterion tables is partially motivated by an additional shortcoming of Bayes theorem: its sequential nature of considering each finding one at a time. Sometimes, it is more convenient to consider many factors simultaneously while considering diseases. What if we imagined thediagnosis of a certain disease as an additive sum of select factors? That is, in the MI example, the patient receives a point for having positive chest pain, a point for having a history of a positive stress test, and so on. Here we establish a threshold for a point total that gives a positive diagnosis of MI.

Because some factors are more important than others, we could use a weighted sum, in which each factor is multiplied by an importance factor before adding. For example, the presence of chest pain may be worth **three points**, and a history of a positive stress test may be worth **five points.** This is how criterion tables work.

In the following table, we have given the modified wells criteria as an example. The modified wells criteria (derived from Clinical Prediction, 2017) are used to determine whether or not a patient may have a **pulmonary embolism** (**PE**): a blood clot in the lung that is life-threatening.Note that criterion tables not only provide point values for each relevant clinical finding but also give thresholds for interpreting the total score:

Clinical finding	**Score**
Clinical symptoms of deep vein thrombosis (leg swelling, painwith palpation)	3.0
Alternative diagnosis is less likely than pulmonary embolism	3.0
Heart rate > 100 beats per minute	1.5
Immobilization for ≥ 3 days or surgery in the previous 4 weeks	1.5
Previous diagnosis of deep vein thrombosis/pulmonaryembolism	1.5
Hemoptysis	1.0
Patient has cancer	1.0
Risk stratification	
Low risk for PE	< 2.0
Medium risk for PE	2.0 -6.0
High risk for PE	> 6.0

Corresponding machine learningalgorithms – linear and logistic regression [13 M]

Linear regression

Linear regression is a popular statistical technique commonly used in healthcare analytics to analyze the relationship between two or more variables. In the context of health care, it can be used to predict outcomes, understand correlations, or identify trends based on data. For instance, we could use linear regression to predict a patient's health metrics like blood pressure based on age, weight, and exercise habits. Linear regression assumes a linear relationship between variables, so it might not capture complex interactions in healthcare data.

It is used to quantify the relationship between one or more predictor variables and a response variable.

The most basic form of linear regression is known as simple linear regression, which is used to quantify the relationship between one predictor variable and one response variable.

If we have more than one predictor variable then we can use multiple linear regression, which is used to quantify the relationship between several predictor variables and a response variable

- Medical researchers often use linear regression to understand the relationship between drug dosage and blood pressure of patients.
- For example, researchers might administer various dosages of a certain drug to patients and observe how their blood pressure responds. They might fit a simple linear regression model using dosage as the predictor variable and blood pressure as the response variable. The regression model would take the following form:

  blood pressure = β_0 + β_1(dosage)
- The coefficient **β_0** would represent the expected blood pressure when dosage is zero.
- The coefficient **β_1** would represent the average change in blood pressure when dosage is increased by one unit.
- If β_1 is negative, it would mean that an increase in dosage is associated with a decrease in blood pressure.
- If β_1 is close to zero, it would mean that an increase in dosage is associated with no change in blood pressure.
- If β_1 is positive, it would mean that an increase in dosage is associated with an increase in blood pressure.
- Depending on the value of β_1, researchers may decide to change the dosage given to a patient.

Logistic regression is a popular statistical machine learning algorithm that is commonly used for binary classification tasks. It is a type of model known as a generalized linear model.

To understand logistic regression, we first understand **linear regression**. In linear regression, the i^{th} output variable (*y-hat*) is modeled as a weighted sum of the *p* individual predictor variables, x_i:

$$\hat{y}_i = b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$$

The weights (beta) (also known as **coefficients**) of the variables can be determined by the following equation:

$$\beta = \left(X^T X\right)^{-1} X^T Y$$

Logistic regression is like linear regression, except that it applies a transformation to the output variable that limits its range to be between 0 and 1. Therefore, it is well-suited to model probabilities of a positive response in classification tasks, since probabilities must also be between 0and 1.

- Logistic regression has many practical advantages.
- First of all, it is anintuitively simple model that is easy to understand and explain.
- Second, logistic regression is not computationally intensive, in terms of time or memory. The coefficients are simply a collection of numbers that is as long as the list of predictors, and its determination only involves several matrix multiplications

- Third, logistic regression does not require much preprocessing (for example, centering or scaling) of the variables (although transformationsthat move predictors toward a normal distribution can increase performance). As long as the variables are in a numeric format, that is enough to get started with logistic regression.
- Finally, logistic regression, especially when coupled with regularization techniques such as lasso regularization, can have reasonably strong performance in making predictions.

However, in today's era of fast and powerful computing, logistic regression has largely been superseded by other algorithms that are more powerful, and typically more accurate.
This is because logistic regression makes manymajor assumptions about the data and the modeling task:

- It assumes that every predictor has a linear relationship with the outcome variable. This is obviously not the case in most datasets. In other words, logistic regression is not strong at modeling nonlinearitiesin the data.
- It assumes that all of the predictors are independent of one another. Again, this is usually not the case, for example, two or more variablesmay interact to affect the prediction in a way that is more than just thelinear sum of each variable. This can be partially remedied by adding products of predictors as interaction terms in the model, but choosingwhich interactions to model is not an easy task.
- It is highly and adversely sensitive to multiply correlated predictor variables. In the presence of such data, logistic regression may causeoverfitting. To overcome this, there are variable selection methods, such as forward step-wise logistic regression, backward step-wise logistic regression, and best subset logistic regression, but these algorithms are imprecise and/or time-intensive.

Finally, logistic regression is not robust to missing data, like someclassifiers are (for example, Naive Bayes).

---**********---