An eBook on Cardiology

Chapter 1

Cardiology in the World of Informatics, Big Data and Machine Learning

Thomas Kingsley; Robert Kirchoff; Rahul Chaudhary*

Mayo Clinic, Rochester, MN, USA

****Correspondence to: FRahul Chaudhary,** DAssistant Professor of Medicine, Division of Hospital Internal Medicine/ Department of Internal Medicine, Mayo Clinic 200 First Street SW, Rochester MN 55905*

Phone: 507-255-8043; Fax: 507-255-9189, Email: Chaudhary.rahul@mayo.edu

1. Introduction

1.1. Healthcare from Industrial to Information Revolution

In the pre-industrial era healthcare for the most part was delivered in patient's homes or small clinics. It was infeasible for most patients to travel outside the small proximity of their town given the logistical constraints of traveling, even distances that today would be considered insignificant. The industrial revolution brought larger roads, train lines and mass production of automobiles, which made travel much easier. As a result, academic centers and healthcare organizations started to sprout and grow throughout the United States. From the industrial revolution came an age of scientific discovery and technical advances that would shape current healthcare. The majority of today's population was not alive during the first industrial revolution, but we are living through another time period just as profound to human experience and healthcare, the information age (1975-2020) [1,2].

The information age started in the 1970's and is defined by rise of information technology (transistor, personal computer, internet, etc). Healthcare was initially behind most industries in utilizing information technology. However, in the early 2000's with the publication of landmark reports like the Institute of Medicine's (now the National Academy of Medicine) "To Err is Human" and "Crossing the Quality Chasm" the US healthcare system was exposed for being too expensive, having too many medical errors, and being too inaccessible. Major quality improvements were needed.

One focus was improved utilization of information technology. Law makers passed several major health policy reforms. For example, the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, was passed into law and created incentives for healthcare organization to adopt electronic health records (EHRs). The result has been major adoption of HIT in the past decade. In 2008, only 9.4% of hospitals in the US had a basic electronic health system (EHR), but by 2015 adoption had increased to 94% [3]. Data from clinical documentation, lab and image tests, therapeutic orders, claims, and wearable devices - to name a few of many examples - are being digitalized and stored. This has led to an unimaginable wealth of available health data.

For example, Datasets used for analytics can often contain a zeta bye (10^21 bytes) of data, and have led to the term "big data". Clinical researchers who used to rely on small-specialized datasets maintained in their healthcare organization, or disease registries in departments of health, or who would spend months searching through paper charts by hand to create their own dataset, now have access to a vast amount of data with a few clicks of a mouse [4].

One result of increased data has been the increased efficiency of clinical research. Figure 1 shows that from 2000 to 2019 the number of registered clinical trials increased from thousands to a quarter of a million annually (clinical trials did not have to be registered until 2005 but most of the growth has occurred after this date) [5]. Medical knowledge and evidence-based guidelines are shaped from clinical research. Consequently, as HIT has advanced, medical knowledge has exponentially increased. The doubling rate of all medical literature was estimated to be 50 years in 1950, 7 years in 1980, 3.5 years in 2010, and 73 days in 2020 [6]. This means for Cardiologists by the time they get through medical school, residency and fellowship training all the knowledge that existed when they started their educational journey will be long out of date, but even more disturbing so, will the knowledge they gained in their first year of fellowship. Paradoxically, keeping up-to-date and practicing evidence-based medicine has never been more challenging, despite the increased access to this knowledge through HIT.

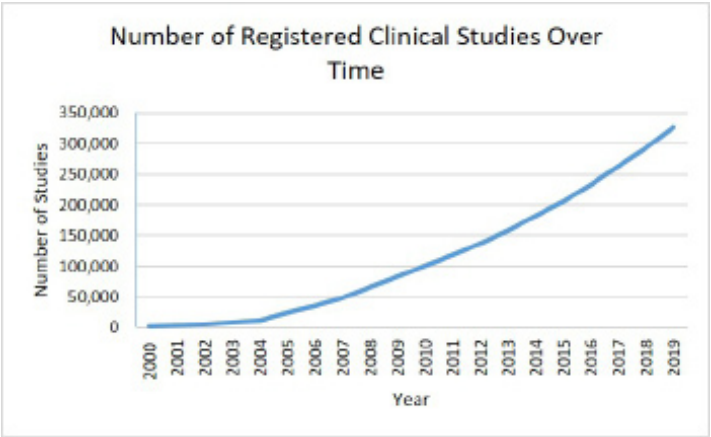

Figure 1: modified from ClinicalTrials.gov

As you will learn in this chapter, the field of informatics has been on the forefront of addressing the challenges to achieve the full potential of healthcare's evolution in the information age. Therefore, investment in informatics education for Cardiologists in practice and training will be essential for continued innovation and improved healthcare delivery in the field [7,8].

1.2. A new age is upon us with the rise of artificial intelligence and machine learning

To make use of the vast data created from the information age, technologies in data mining and analytics were required. One such technology has been natural language processing (NLP), a sub discipline of artificial intelligence (AI), which is a powerful tool for extracting unstructured text data and transforming it into more structured data elements. NLP has been essential in utilizing vast stores of unstructured data that otherwise would be functionally useless. Machine learning (ML) has also been crucial in analyzing data. Conversely, data has also been crucial for advancing ML, which requires large amounts of data to train its algorithms. A perfect storm of technological advances such as increasing computational power, development of cloud computing, the rise of the graphical processing unit (GPU), along with big data has ushered in the age of ML and AI.

Many believe the age of AI and ML will have a larger impact on humanity and possibly healthcare than the industrial and information ages. How much is hype versus reality? It is hard to know but AI is already changing our day-to-day life and is likely here to stay. In healthcare ML and AI have shown impressive results in imaging-based technology and with rapid advances on a yearly basis, its expected to have an even broader impact in the near future. Consequently, healthcare organizations are already starting to invest in ML and AI, and high impact journals are increasingly publishing ML studies.

1.3. Cardiologist need training in informatics and data science

To realize the full potential of our new data landscape will take training the current and next generation of Cardiologists in technical fields like informatics and data science. Otherwise, the chasm between the world of clinicians and technical experts will grow, and the possibility for improved patient care diminished.

The goal of this chapter is to provide an introduction into the important topics such as big data, artificial intelligence, and knowledge engineering through the lens of informatics. Although the examples used will largely be in the field of Cardiology, most of the information is useful for practitioners in other disciplines as well. Furthermore, the information provided in the chapter is not meant to be a comprehensive overview but instead a primer on these topics.

2. Cardiology Informatics and Information Hierarchy

2.1. Introduction to Health Informatics

The American Medical Informatics Association (AMIA), one of the largest informatics organizations, defines informatics as "The science of how to use data, information, and knowledge to improve human health and the delivery of care services [9]." A broader definition of health informatics is:

"A scientific discipline that deals with the collection, storage, retrieval, communication and optimal use of health-related data, information and knowledge. The discipline utilizes the methods and technologies of the information sciences for the purpose of problem solving, decision making and assuring highest quality health care in all basic and applied areas of the biomedical sciences [10]."

2.2. Informatics and HIT are synonymous?

Informatics and HIT are terms that are often used interchangeably. However, it is important to understand their distinction. HIT is focused on technology and the management of technological systems. Informatics is focused on the systems of information and the data, information, and knowledge within them.

The focus on health informatics should always be improving and advancing patient care, and not first on the information technology.

2.3. Informatician in healthcare

Informaticians are informatics experts who often have formal training and postgraduate degrees in the field of informatics or related disciplines (see below). Informaticians have a broad range of skills and job roles within a healthcare organization. This reflects the many disciplines that contribute to the field. Clinicians who are informaticians typically have leadership roles. For example, one leadership position is the Chief Medical Information Officer (CMIO) who make high level decisions regarding HIT infrastructure. Cardiologists informaticians that support their practice are increasing and have a number of important roles which can include specialty specific EHR changes, development of clinical decision support (CDS) tools, and development of analytic dashboards to name a few of many. Another role of Cardiology informatician are those who primarily do research. They are often interested in mining, creating, or maintaining databases and performing data analytics for research purposes. Clinicians who are informatics specialist often lead teams with various technical expertise such as IT specialist, computer scientists, data scientist and system engineers (see **Figure 2**).

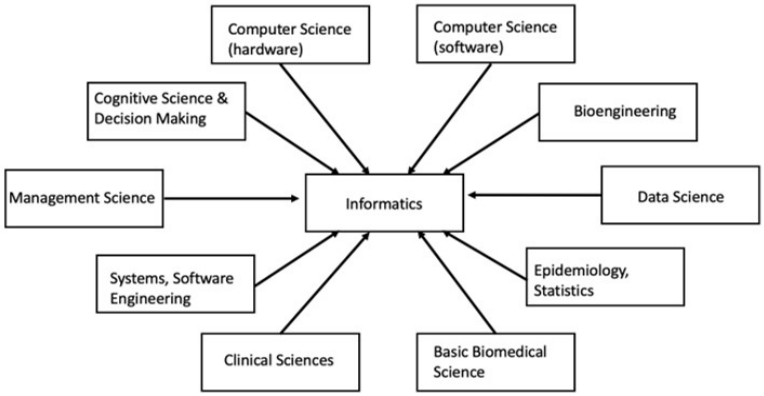

Figure 2: Modified from the American Medical Association. Informatics is a field with foundations in several other scientific disciplines

2.4. Information Hierarchy

The concepts of data, information, knowledge and wisdom are foundations of informatics. Although these concepts are regular in everyday lexicon, they are often difficult to define and used interchangeably. Defining these terms is important because they create a conceptual information hierarchy, which often is helpful in creating methods to approach informatics projects. Moreover, this basic framework is built with more detailed approaches to applied informatics, such as Dr. Lehmann of John's Hopkins Information Stack (see Informatics in practice below) (**Figure 3**).

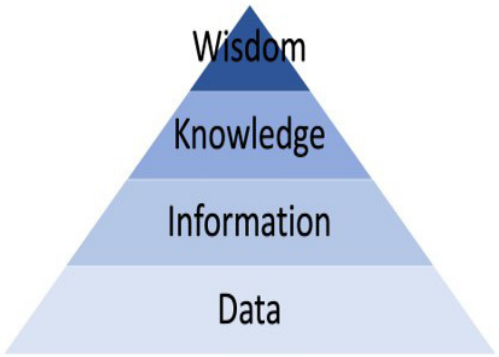

Figure 3: Information Hierarchy

2.5. Data

Data are symbols that represent observations, but without larger meaning. For example, the number 110 put into a database has no intrinsic meaning. It could mean systolic blood pressure or heart rate, or the age is 110. Data reduced to its indivisible element is datum. This is what databases store and computers process.

The digitalization of patient data has created an exponential increase in its volume. Understanding the characteristics of the data you are working with is key to data mining and analysis. One important distinction is data structure. Structured data is highly organized often in a relational database and often where the data has both an underlying known architecture and metadata (data about data) is present. Contrarily, an example of unstructured data is often text written in clinical notes, which does not have a defined organization.

3. Big Data

What is Big Data and does this mean there is little data? The debate over defining big data continues, but big data is typically characterized by five V's [11].

Volume: There is n't an exact volume of data that creates a line between small-or-medium size data and big. However, the volume in big data is usually considered more than could be stored on a single server, and often requires hundreds of servers for storage.

Variety: The heterogeneity of data types, sources, function and fidelity are all key characteristics of big data.

Velocity: The rate of data querying (mining) and processing for functional purposes.

Value: The usefulness of the data for its intended purpose

Veracity: Veracity is basically the data quality. Is the data accurately capturing what it supposed to?

3.1. Data warehouse

Healthcare organizations are increasingly developing clinical data warehouses (CDWS) to store various health data with the hope of it being accessible for analytics and research. Older legacy EHR systems typically were de-centralized, meaning multiple different applications such as computer physician order entry (CPOE) and clinical documentation could run on different applications. This made older EHR systems notoriously slow, prone to system errors (often causing frequent downtimes), and difficult to extract data to a centralized database. The evolution of centralized EHRs that run on a cloud framework are now becoming the standard in most organizations. Epic is the most common EHR now in academic organizations.

CDWS typically use structured data from places like the EHR, digital imaging and communication in medicine (DICOM), and administrative databases. However, with the development of NLP, various unstructured data are transformed to structured data in the CDWS. Typically, the step before entering the data in CDWS is a staging database where the data from multiple sources is cleaned and mapped into a Meta database. The data from CDWS is often then used to support specialty specific databases called data marts that are often formed to support research purposes. In the case below is the development of a Cardiology Heart Failure Data Mart (**Figure 4**). The data in these specialty data marts has typically gone through a validation process, which helps reassure researchers of its quality. They can also be linked to larger disease specific registries. In the case of heart failure there is active work in linking these institutional data marts to large national registries [12,13].

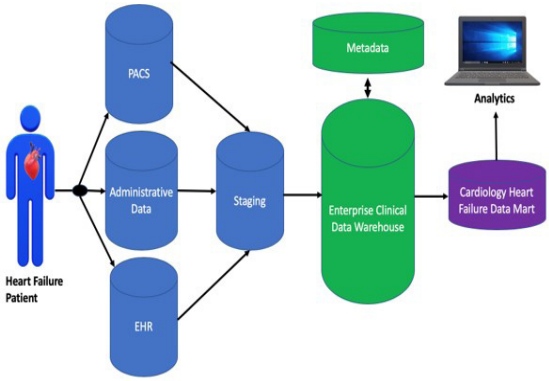

Figure 4: CDWS and Cardiology Heart Failure Data Mart for Analytics

3.2. Information

3.3. Defining and measuring information

The most basic way of describing information is data + meaning = information. However, this is basic conceptual approach to information. In reality, quantifying information can be much more complicated. A simple example is trying to calculate information over a simple information channel, which be calculated by the equation 1.

$$H=n\ [(\log)]\ \_2\ S$$

In this equation H is Shannon's entropy or information measured in bits, n is the number of symbols (data), and S is the number of possible symbols.

3.4. Information Systems

In section data Figure 4 showed how data is can be organized within a healthcare institution. The architecture of a health information systems is focused on the systems transmitting patient health information. This can be very complex.

3.5. Interoperability

The concept of interoperability (see definition below) helps define how various system elements communicate information and interface together at various levels within a part or the entire healthcare organization, public health system, or nationally.

Interoperability is the ability of different information systems, devices and applications ('systems') to access, exchange, integrate and cooperatively use data in a coordinated manner, within and across organizational, regional and national boundaries, to provide timely and seamless portability of information and optimize the health of individuals and populations globally [14].

The Healthcare Information and Management Society (HIMSS), one of the largest

informatics societies, defines four different levels of interoperability.

HIMMS Four Levels of Interoperability [14]:

1. Foundational (Level 1) – establishes the inter-connectivity requirements needed for one system or application to securely communicate data to and receive data from another.

2. Structural (Level 2) – defines the format, syntax, and organization of data exchange including at the data field level for interpretation.

3. Semantic (Level 3) – provides for common underlying models and codification of the data including the use of data elements with standardized definitions from publicly available value sets and coding vocabularies, providing shared understanding and meaning to the user.

4. "New" Organizational (Level 4) – includes governance, policy, social, legal and organizational considerations to facilitate the secure, seamless and timely communication and use of data both within and between organizations, entities and individuals. These components enable shared consent, trust and integrated end-user processes and workflows.

An essential component of achieving interoperability is having standards at the various levels of interoperability. For instance, to achieve semantic interoperability, HIT systems commonly use **HL7**, which is an international text standard for clinical and text data.

3.6. Knowledge

Knowledge is often obtained by accumulation and validation of information, generally accepted to be true.

The exponential growth of medical literature has created a massive amount of medical knowledge. Medical knowledge guides our individual clinical behavior, informs our practice guidelines, creates new inferences for future research studies, and helps educate the next generation of physicians and Cardiologists. There are several relevant fields of study that directly work on knowledge and its management. Two important ones to mention are epistemology that works on the theories of knowledge, and knowledge engineering is a sub discipline of AI, which tries to create technological systems to reflect expert decision making. Knowledge is an important ingredient to optimal clinical decision making.

Prior to the information age, memorization of medical knowledge was a necessity for practicing physicians. This made sense in an environment where medical knowledge doubling rate was decades or longer, and where the accessibility of medical knowledge often required going to the nearest library – not ideal when needing to make quick life and death decision about a patient's clinical management. Today, a clinician's smartphone can carry more medical knowledge than any physical medical library could accommodate. Furthermore, the internet

has readily accessible resources like Ask Mayo Expert, Up-To-Date, etc. Physicians have access to quick guidelines and reviews about numerous pathologies. However, medical training continues to be antiquated in its approach to medical practice. Medical students, residents and fellows are still often graded on their memorization of a list of facts rather than their ability to quickly acquire up-to-date medical knowledge and use it correctly.

3.7. Clinical Decision Support

As the medical knowledge base grows the use of technology can help in clinical decision making. Clinical decision support is a method to resolve this challenge. Robert Greenes, an author of the popular book Clinical Decision Support, defines computer based CDS as "the use of information and communication technologies to bring relevant knowledge to bear on the health care and well-being of a patient [15]."

Figure 5 shows a broad conceptual framework for steps of knowledge creation, knowledge extraction, knowledge storage, and knowledge deployment. Each of these steps are important to consider when developing a CDS tool.

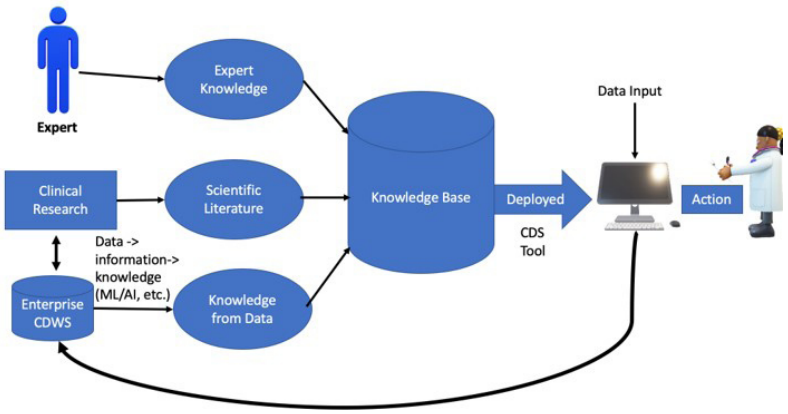

Figure 5: CDS Overview- Knowledge Creation to Deployment with HIT.

CDS is now becoming widely used in EHRs around the country, especially in large academic centers. One of the most frequent uses is medication alerts where a physician is alerted at the time of ordering a medication that is marked as an allergy or contraindicated due to drug-drug interactions. However, specialty practices such as Cardiology are making CDS tools for their practice and healthcare organization. For instance, CDS tool that help identify acute coronary syndrome in ED patients or the use of CDS tool to help guide management in a known ACS case based on risk factors [15,17-19].

As knowledge, creation continues at a remarkable pace, the goal for evidence-based care to improve the value of healthcare delivery, and the rise of machine learning makes CDS tools increasingly powerful.

3.8. Wisdom

Wisdom is characterized by the ability to use knowledge to make intelligent decisions. Intelligence and its definition, especially with the possibility of the age of AI, has become a debated topic. There is certainly a fear that AI will take over physicians' jobs in the future. The jobs that are most affected by AI and other technology, typically have a high degree of automation. Providers are responsible for the lives of their patients, and work in field requiring high level of expertise. Therefore, it is unlikely that AI will replacing physicians see **Table 1**. However, a more likely reality is that AI will support and improve efficiency of the daily work done by physicians.

Table 1: Probability of automation based on task.

Cognitive Task	Level of Automation
Skill-based	High probability of automation with the assumption reliable feedback loop and error feedback
Rule-based	Moderate probability of automation. If rules are well established and validated.
Knowledge-based	Low probability of complete automation, but possible contribution of automation to help synthesize data
Expertise	No probability of currently being able replace humans, but possible support in tasks.

Judea Pearl who won the Turning Award (prestigious award for contribution in computer field) for his work in AI describes what goes into intelligence. Although there have been tremendous advances in machine learning, these algorithms are still mostly used for association. Judea Pearl describes a ladder of intelligence with three rungs (**Figure 6**). The first rung is the ability to make associations between objects that have correlation to some outcome of interest. All the advances and successes of machine learning to date are still on this fist rung, below the intelligence of human baby. The next rung of intelligence is starting to understand causal reasoning by performing/doing interventions on the surrounding world and watching the outcome. The top rung is imagination and explores the ability to project what-if scenarios with interventions or lack thereof to the surrounding environment– counterfactual. This level of intelligence is a defining human trait, and no machine or algorithm has come close to achieving this level of intelligence.

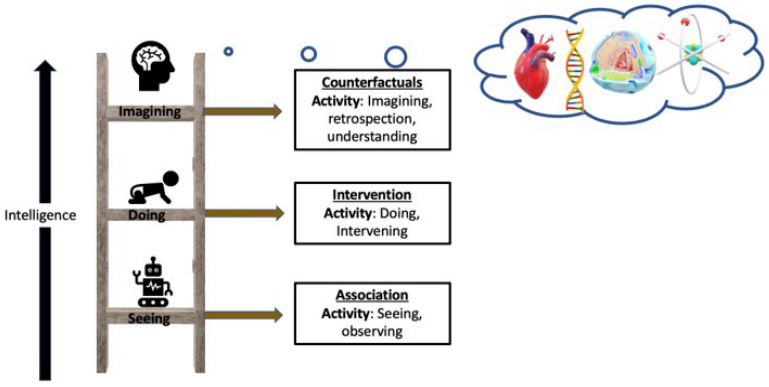

Figure 6: The Ladder of Causation by Judea Pearl

1. Artificial Intelligence and Machine Learning

In 1950 Alan Turing a famous British Mathematician postulated that computers could have similar intelligence to humans in the future. Many believe this marked the beginning of artificial intelligence (AI) as a field - defined by the pursuit of creating non-living intelligence. Over the decades, mathematicians, computer scientists, physicists and other fields of science have contributed to its growth. The evolution of AI, however, has not been linear. It has been characterized by incremental discoveries that would move the field forward and increase its hype, but then would be followed by "AI winters," characterized by long periods of minimal gains.

Today, AI is hyped as ever and has regained the public's imagination, but the hype versus reality remains to be determined. During this period, AI has proliferated into consumer products. Smartphones and applications are using speech and facial recognition. AI has beat the best Jeopardy players and the world's best Go player, in a five-game match. In healthcare, AI has achieved remarkable success. It performed at the level or better than radiologists and dermatologist in making diagnoses from chest x-rays and identifying a malignant mole, respectively. In Cardiology there has been a wave of AI uses and related publications. For example, a deep neural network (DNN) (type of machine learning algorithm, see below) outperformed most Cardiologist's sensitivity at detecting ECG arrhythmias. (see **Figure 7**) [20].

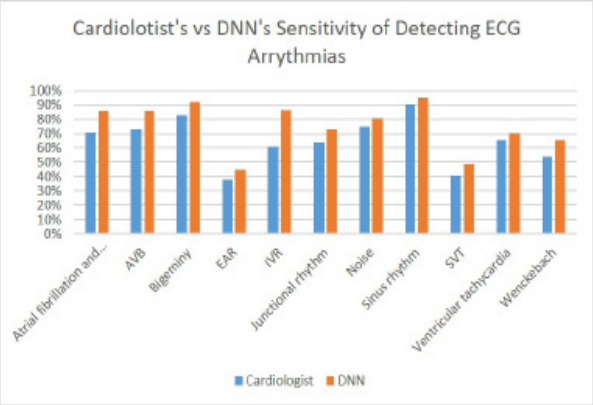

Figure 7: Cardiologist vs DNN in sensitivity of detecting arrhythmia on ECG.

The goal of the following section is to provide a summary of the major concepts. The concepts described have, in and of themselves, entire books dedicated to them. Here, we will focus on concepts that help with practical considerations when using ML in healthcare applications and the approach to interpreting medical literature where ML/AI was used.

Artificial Intelligence Versus Machine Learning

The terms AI and ML often are used interchangeably, but ML is a subfield of the broader concept of AI. ML algorithms learn from data - giving it both its name and defining characteristic. AI is focused on the concept of non-living intelligence, learning from data being

one aspect.

3.9. Machine Learning Versus Statistics

The machine learning algorithms have a basis in statistics. The differentiating feature is the objective of ML to improve with iterations of examples from a dataset. A strong foundation in statistics is valuable in learning ML.

4. Machine Learning in Research

The paradigm in medical research has been to use deductive reasoning to guide epidemiologic study design and inference testing with traditional statistical techniques. Deductive reasoning starts with a hypothesis founded on theory from concepts that already exist.

ML research often starts with inductive reasoning that begins with observations and pattern recognition and then builds theories. Inductive reasoning has been used less commonly in medical research. Before data was widely abundant, correlative patterns in a smaller dataset were often noise and not signal. Moreover, without advanced analytic techniques, like machine learning, isolating the signal was difficult. The lack of causal inferencing has been another reason for its lack of acceptance.

The factors often limiting inductive reasoning are changing with the availability of big data, improved computer power and speed, and improved ML analytic techniques. Therefore, the future will see a large paradigm shift in medical research. This emphasizes the importance to start training Cardiologists on how to interpret and conduct studies using inductive reasoning as the basis [8,21].

4.1. Machine learning basics:

4.2. ML algorithms learn from data, but what does learn in this context mean?

Imagine what goes into your own ability to learn from experience. What are the elements required? There are tasks T (objectives) that you're accomplishing with some degree of success, termed performance P, and there are multiple examples that build your experience (E) at performing these tasks. If your performance P of doing tasks T improves through your experience E then you have learned – right? The saying practice (E) makes perfect (improved P at T) has a lot meaning in ML/AI. We will revisit these elements of learning later in this section.

Here are few terms to add into our ML lexicon: examples and features. First experience E from above was made up of many examples of doing a task T. Each example has features, which the ML algorithm will use to learn from. An analogy from real life would help explain

the concept better. In an excel spreadsheet, each example would be represented by a row. The features would be columns. Researchers developed a ML acute coronary syndrome (ACS) mortality prediction model. The goal was to predict mortality in patients presenting to the ED with ACS. Examples would represent each patient arriving to the ED with ACS. Features in this case would be variables like smoking status, cholesterol, prior MI, etc.

4.3. Everything in ML starts with data

The following datasets are important to know in ML algorithm development:

1. Development dataset: the data selected at the beginning of the process to develop the ML algorithm. This will be further divided into the training and tuning data set.

2. Training dataset: Originating from the development dataset, this dataset will be used for developing the ML algorithm through modification of parameters over iteration of examples.

3. Tuning dataset: This dataset also originates from the development dataset and is used to modify hyper parameters (see concepts of ML algorithm development below) for improvement in the model.

4. Validation dataset: This dataset is used to validate the model developed from the development dataset.

4.4. Broad Categories of Learning

There are three broad categories of ML algorithms: 1) supervised learning, 2) unsupervised learning, and 3) reinforcement learning.

4.5. Supervised Learning

Supervised learning algorithms experience numerous features of a dataset (shared with unsupervised learning as well). The difference between unsupervised and supervised learning is that in supervised learning each example is linked to a label [22].

In the ML ACS mortality prediction model above. In supervised learning each example or patient arriving to the ED with ACS would then be associated with a label (outcome) mortality or not in specified timeframe. Through experience (E) in many examples of patients the ML algorithm would start to learn which features predicted mortality, and therefore its performance (P) in the prediction of task (T) would improve.

4.6. Unsupervised Learning

Unsupervised learning, as mentioned above, experiences (E) a dataset but is not linked to a label. Instead it often will analyze the probability distribution between features. One

common use is to cluster the examples based on unique attributes of the feature's probability distribution thus providing insight into these features. The insight developed can be used to develop clinical phenotypes.

Unsupervised learning has had less application in clinical medicine to date but promises to be important in future discovery of novel relationships between clinical variables not yet discovered.

4.7. Reinforcement Learning

Reinforcement learning learns from its environment by interacting with it and through feedback loops develops its experience. Reinforcement learning tries to elevate ML to the second rung, a baby's intelligence (**Figure 6**). It has little application in healthcare, yet, and we will not be discussed further in this chapter.

Majority applications of ML in healthcare occur on supervised learning and the remaining chapter will only focus on this aspect of ML.

4.8. Revisiting learning: Tasks, Experience, and Performance

Task (T)

Learning is not directly associated with doing a task. In the ACS mortality prediction model the task was prediction of mortality. If after experience (E) the performance (P) has not changed at this task (T) then learning did not occur.

There are a number of tasks (T) supervised ML algorithms can perform as shown in **Table 2**.

Table 2: Examples of tasks for ML

Task (T)	Comments
Classification	The algorithm selects a category from a set of possibilities based on inputs. **Examples in Cardiology literature:** Reading ECGs and categorizing based arrhythmia [20], categorizing patients into coronary heart disease versus not based on number of clinical features[23], categorizing restrictive versus constrictive pericarditis based on ECHO findings [24]
Regression	Regression produces a numerical value
Structured output (example of classification)	Involves a task where data is unstructured and annotates it for mapping into an organization and thus structuring the data. This can be an important tool when managing a large dataset.
Forecasting	Using time-series data as features for predicting an outcome. **Examples in Cardiology literature:** Predicting the risk of heart failure with EHR sequential data modeling [25]

Performance (P)

Performance (P) is often measured as the accuracy or conversely the error rate. Other measures are used depending on the application, but will not be discussed here.

Experience (E)

A general principle is ML algorithms will have improved performance with more experience or data. There is no rule-of-thumb on the right amount of data because it depends on the context of its use and the characteristics of the data features.

4.9. Concepts about machine learning algorithm development

Machine learning has several models or algorithm it can use. The best choice often depends on the data and comparison between different models. To understand these essential principles, we will use the simplest ML algorithm and one in healthcare we are most familiar with, linear regression.

$$\hat{y}=w^{T} x+b$$

The purpose of the regression model above will be to take a vector x with several input variables and predict a scalar value $\hat{y}$. The value $\hat{y}$ represents the prediction from the model of the actual y. Each example will have a number of features that will impact the outcome. To adjust for these features based on their importance to the prediction, we weight their importance on the prediction with a set of weights w^{T}. The entire set of weights can be considered its **parameters**, which is a key concept in ML. As can be seen from this simple example these parameters would change as the algorithm learns which features improve performance. More sophisticated algorithms can also have **hyper parameters**, which are fixed and do not change with experience (E) but may help improve performance. Lastly, b is a bias term which just reflects the intercept (unlike the bias in statistics).

For the machine learning algorithm to learn, it will need a performance measure. The performance measure in this case would be mean squared error. This function optimizes the weights to improve performance (P), and is the key step in learning.

4.10. Generalization

Another important concept for either supervised or unsupervised ML models is **generalization** - which is how well ML algorithms respond to new data outside of the training environment which is the ultimate goal of any ML algorithm. In the training dataset, the **training error** is measured and the test error is measured afterwards (sometimes referred to as generalization error). The **test error** provides external validity of the ML model.

The goal of training the data is to make training error small, and subsequently the gap between the training error and gap error is also small. This can be realized by appropriate fitting of the model.

4.11. ML model fitting

The ML models fit the data based on several factors. Fitting involves graphical mathematical models (Figure 8). Over-or under-fitting are one of the biggest issues facing the effectiveness of a ML algorithm. Fitting the model in training will then determine its generalization during testing on data outside of the training dataset– remember generalization is the ultimate goal of ML models because it reflects the performance on applied data.

Under fitting a model will cause high training error because it does not fit closely to the training data; however this could mean better performance than an over fitted model when it comes to data outside the environment. Overfitting a model, contrarily, will cause low training error but oftentimes results in poor performance when used outside of the training dataset. However, if the training dataset closely reflects the data where the ML model will be applied, a highly fitted models performs well. However, generally the best fitted model is often between a low-or-highly fitted model.

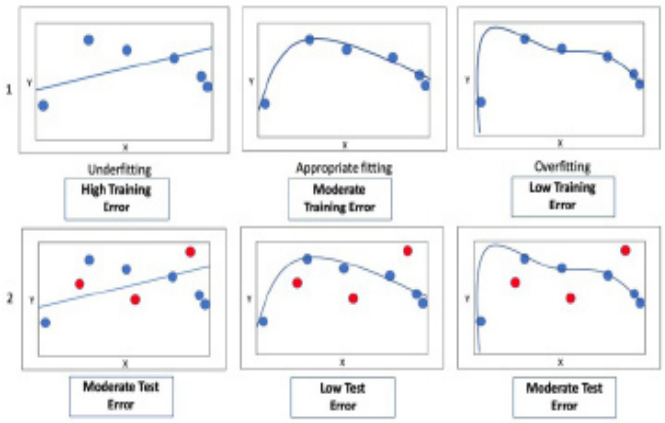

Figure 8: Modeling fitting: The graphs on row 1 are the results of validity testing the training data (represented by blue dots) and show the difference of between under, appropriate, and overfitted models. In row 2 the graphs represent how the selected models in row 1 perform with new data outside the training model (red dots).

Several ways to avoid overfitting (aka **regularization**):

- Reducing parameters and dimensionality reduction
- Data augmentation: Modifying input data to achieve optimal size
- Parameter regularization: ways to avoid parameters increasing in size

4.13. Machine Learning Models

There are hundreds of models. No model is better than another. Instead it is important to

understand the data to define the most appropriate model suited for the purpose.

4.14. Artificial Neural Networks

The development of artificial neural networks (ANN) has proven to be a large success in ML. ANN were designed from the concept of human neurons; they resemble their function more than their biology. In general, ANN have an input layer, which are connected to an output layer sometimes through a hidden layer. **Deep neural networks (DNN)**, are part of **deep learning** and a subcategory of ANN and has had some the most profound results. In this model there is a substantial hidden layer (typically greater than two hidden layers) between the input and output layers, where abstraction of features of the input layer occur. Each layer selects more complex features of the input variables, and place mathematical parameters on information from previous layers. The final layer, or output layer, then makes a prediction based on the patterns in features within each hidden layer.

4.15. Common Basic Workflow in developing a predictive ML algorithm

1. Data pre-processing (**Figure 9**)

- Feature selection and extraction
- Identify and deal with missing data
- Data normalization (example: put data in same magnitude of measurement scale)
- Noise reduction

2. Model selection

3. Development of ML algorithm

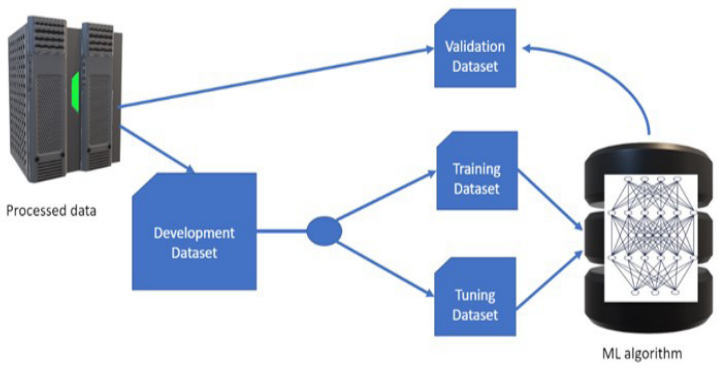

Figure 9: Machine learning model development workflow

5. Validation

There are several ways of evaluating the internal validation. The test error rate and ROC curves similar to other predictive models are analyzed. The measures are then compared

to a reference standard. Understanding how the methods of creating a reference standard is important because this can have a large impact on the results of validation studies.

6. Summary

To summarize, the accumulation and integration of data from multiple streams within EHR and the widespread use of digital technology like smartwatches has resulted in development of big data. The development of novel data mining techniques and advances in ML methodology open a plethora of its applications both in the field of Cardiology and in Medicine at large. The applications of ML and AI will continue penetrating and expanding within clinical practice. It is quintessential for Cardiologists to at least have an understanding of the basic principles of informatics to continue advancing the frontier of medical research and practice of evidence-based medicine.

7. References

1. Starr, P., The Social Transformation of American Medicine, The Rise of a Sovereign Profession & the Making of a Vast Industry. 1982.

2. Gleick, J., The Information, A History, A Theory, A Flood. 2011: Random House, INC.

3. Technology, T.O.o.t.N.C.f.H.I., Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. 2016. p. https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php.

4. Krumholz, H.M., Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. Health Affairs, 2014. 33(7): p. 1163-1170.

5. Clinical Trials Available from: https://clinicaltrials.gov/.

6. P, D., Challenges and opportunities facing medical education. Transactions of American Clinical and Climatological Association 2011.

7. Pageler, N.M., C.P. Friedman, and C.A. Longhurst, Refocusing Medical Education in the EMR Era. JAMA, 2013. 310(21): p. 2249.

8. Kim, J., Big Data, Health Informatics, and the Future of Cardiovascular Medicine. Journal of the American College of Cardiology, 2017. 69(7): p. 899-902.

9. What's Informatics 1/22/2019]; Available from: https://www.amia.org/fact-sheets/what-informatics.

10. Informatics Definition Available from: http://progenomix.com/area_of_expertise.html.

11. Natarajan, P., J.C. Frenzel, and D.H. Smaltz, Demystifying Big Data and Machine Learning for Healthcare. 2017: CRC Press.

12. Xian, Y., B.G. Hammill, and L.H. Curtis, Data Sources for Heart Failure Comparative Effectiveness Research. Heart Failure Clinics, 2013. 9(1): p. 1-13.

13. Marinelli, M., et al., A modular informatics platform for effective support of collaborative and multicenter studies in cardiology. 2015.

14. What is Interoperability? [cited 2020; Available from: https://www.himss.org/what-interoperability.

15. Greenes, R., Clinical Decision Support. 2014: Academic Press. 930.

16. Harold Lehmann, M.D.P.D. Introduction to Decision Support Systems: Motivation and Examples. 2018. Johns Hopkins University

17. Shortliffe, E.H. and M.J. Sepúlveda, Clinical Decision Support in the Era of Artificial Intelligence. JAMA, 2018. 320(21): p. 2199.

18. Bennett, P. and N.R. Hardiker, The use of computerized clinical decision support systems in emergency care: a substantive review of the literature. Journal of the American Medical Informatics Association, 2016: p. ocw151.

19. Kawamoto, K., et al., Key principles for a national clinical decision support knowledge sharing framework: synthesis of insights from leading subject matter experts. Journal of the American Medical Informatics Association, 2013. 20(1): p. 199-207.

20. Hannun, A.Y., et al., Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 2019. 25(1): p. 65-69.

21. Bizopoulos, P. and D. Koutsouris, Deep Learning in Cardiology. IEEE Reviews in Biomedical Engineering, 2019. 12: p. 168-193.

22. Ian Goodfellow, Y.B., Aaron Courville, Deep Learning Adaptive computation and machine learning series 2016, Cambridge, MA: MIT Press

23. Atkov, O.Y., et al., Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. Journal of Cardiology, 2012. 59(2): p. 190-194.

24. Sengupta, P.P., et al., Cognitive Machine-Learning Algorithm for Cardiac Imaging. Circulation: Cardiovascular Imaging, 2016. 9(6): p. e004330.

25. Jin, B., Predicting the Risk of Heart Failure with EHR Sequential Data Modeling. IEEE Access 2017. 6