Chapter 3

MEDICAL KNOWLEDGE EXTRACTION: PARTICULAR DIFFICULTIES AND OBLIGATIONS

Constantinos Koutsojannis, PhD*

Health Physics & Computational Intelligence Lab,
University of Patras, Patras, Greece

ABSTRACT

In this chapter, the special difficulties of data mining concerning healthcare or medical big data are discussed. Human health care data are the most difficult of all data to collect, deal and analyze. The first important issue is that the mathematical representation and understanding statistical approach in medical data are internally different than those from other data mining activities. As in medicine the primary direction is the care work for a specific patient, and secondarily dealing with him as a research resource; almost the only acceptable reason for collecting medical data is to benefit against the aforementioned individual disease. Software engineers or researchers that are working in other fields will never face the same

* Corresponding Author address
Email: ckoutsog1 @gmail.com

constraints of primarily privacy-sensitive, greatly heterogeneous, and massive medical data. The second important issue is the serious ethical and legal aspects of data mining in this field, including data ownership and special data treatment issues. Consequently, researchers should be aware when are dealing with medical databases they may face the possibility that their work will never be accepted or even used from health care professionals if all these obligations will not be correctly addressed from the early beginning. The last important issue is that most of the health care data have particular obligations concerning their openness to other people that are not dealing with the specific patient; their emergency (often dealing with life-and-death); and the ethical commitment to be used only for salutary. Common language, good communication with end-users and mined knowledge supervising only from the field experts, by well-established health care guidelines, are also important aspects that are undetectable issues of an effective data mining engineer. Transparency of medical knowledge extracted from big medical databases is the *holy grail*, resulting in the appropriate acceptance from end-users as medical community members.

Keywords: data mining, medical applications, ethics, legal, expert supervising

INTRODUCTION

"*Data mining*" tools are based on data classification methods in accordance with their final output features, which allow us to discover patterns that we call "*information*" in order to predict future events or discover new "*knowledge*", both resulted from discovered connections between different data "*classes*". To perform data mining, we need the data and the computing algorithms capable to deal with the data. The more organized the data is, the easier it is to extract "*information*" for further analysis using effective tools from other disciplines as Statistics, Artificial Intelligence (A.I.) and Database management (Figure 1). Data mining is commonly used for medical data mining is becoming increasingly essential [1]. All types of data cannot be processed and analyzed using traditional methods because of the complexity and volume of the data.

Data mining provides the methodology and technology for healthcare organizations to:

- *manage healthcare services*
- *improve treatment effectiveness*
- *save and formulate experts' experience that is not included in textbooks yet*
- *provide new directions for basic or clinical research*
- *save or improve the lives of a specific group of patients*

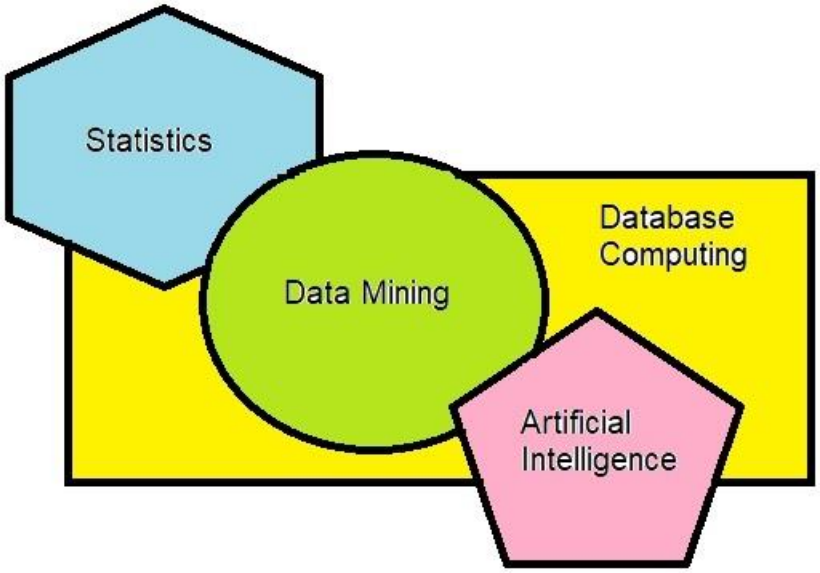

Figure 1: *Data Mining is a convergence of many relevant disciplines*

To work with the collected medical data and to extract knowledge hidden under large and different types of numbers, symbols, and words, data mining includes the following steps:

- Data collection of a target data set from the original data, on which knowledge extraction has to be performed. This means that a large amount of these contains a lot of "*noise*" that could cover the clear signal that should be uncovered

- Data cleaning and interpretation for dealing with missing data, accounting for time-sequence information and interpretation of the reliability of the rest fields

- Data transformation using methods to find invariant aspects of the data coming from individuals' biological and health diversity

- Data mining means extracting patterns of important relations by choosing the proper and more efficient algorithms and presents the output appropriately.
- Data evaluation by the developer in order to extract knowledge from the mined patterns of established classes (Figure 2).

Data mining "*uses mathematical analysis to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data*". Data mining involves the creation of association rules, the use of support and confidence criteria to locate the most important class relationships within the data. Other healthcare data mining parameters include [1]

- "*Clustering*" (i.e. grouping a set of objects and aggregating them based on their similarity to each other), or
- "*Classification*" (i.e. looking for new patterns and predicting variables based on the power factors the database contains), and finally
- "*Sequence*" or patterns analysis (i.e. finding paths form one situation to another maybe later, situation), or
- "*Forecasting*", in order to deal with new records that describe usually a new entity, for example, a patient or event (Figure 2).

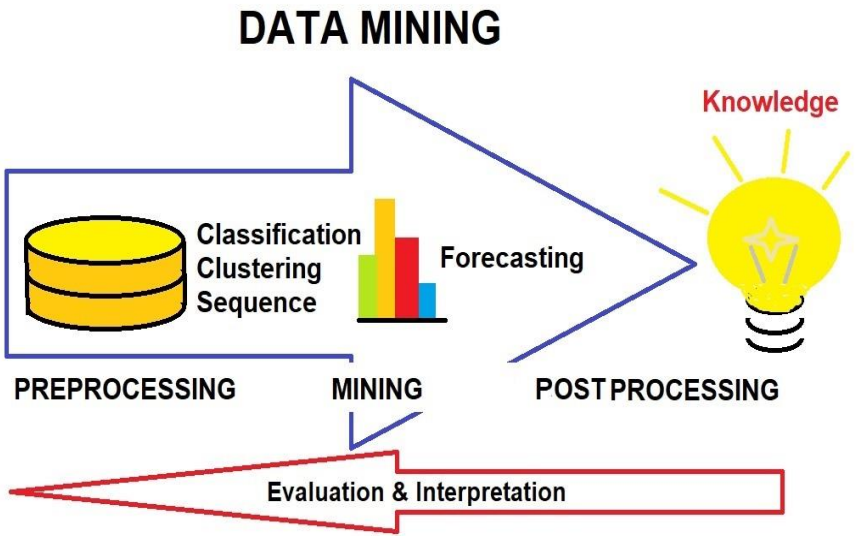

Figure 2: *Knowledge Discovery from Data collectors to new Guidelines*

Usually, mining hospital patient health records, use all data mining techniques including text mining, natural language processing, machine learning, predictive modeling, relationship, and link analysis, statistical analysis, decision trees, etc.

The healthcare organizations as hospitals or governmental health care sectors or even health insurance companies possess rich data sources, such as electronic medical records, administrative reports, and other benchmarking results. Today, data mining in healthcare is used mainly for predicting various diseases, supporting biomedical and clinical diagnosis, advising doctors in making appropriate treatment decisions and avoiding accidental risky events from medical or organizational errors. Additionally, the advantages of data mining are much more than these, as it can provide question-based answers, anomaly-based discoveries, provide more informed decisions, probability measures, predictive modeling, and finally human decision support. Using data mining, healthcare providers can be very effective in such fields as medical research, pharmaceuticals, medical devices, genetics, hospital management, and health care insurance, etc.

Till today because of the restrictions and obligations ruling medical databases, concerning their openness to other people that are not dealing with the specific patient despite the number of publications, there are only a few examples of successful data mining in use in healthcare, but their benefits for healthcare systems are very promising [2, 4, 5, 6, 7, 8]. Consequently, data mining follows a growing interest in the healthcare sector because it offers benefits to care, providers, physicians, patients, healthcare institutions, researchers, and insurers. Healthcare providers and physicians can use data mining to identify effective treatments and best practices as well as to develop new guidelines and standards of care. Even patients, especially those having chronic or high-risk diseases, can receive better, more affordable healthcare services with appropriate identification, tracking and use of appropriate interventions and treatment protocols [7, 9]. Healthcare institutions can use data mining to decrease costs and increase operating efficiency while maintaining high-quality care. This may reflect patient satisfaction and the provision of more patient-centered care. Finally, health insurance organizations can detect medical insurance fraud and abuse through data mining and eventually reduce their financial risk and losses [1, 2].

Dealing with human health care data a software engineer soon realizes that they are the most difficult of all data to collect, deal and analyze. As in medicine the primary direction is the care work for a specific patient, and secondarily dealing with him as a research resource; almost the only acceptable reason for collecting medical data is to benefit against the aforementioned individual disease [3, 4]. Software engineers or researchers that are working in other fields will never face the same constraints of primarily privacy-sensitive, greatly heterogeneous, and

massive medical data. The first thing that one should understand is that a diagnostic test that usually provides data for one patient, is one of many values used to characterize the medical condition of a patient; and the final diagnosis is the synthesis of many tests and observations, that describes a pathophysiologic process in this patient, that in western countries should be totally described from the physician. In Western type of medical practicing the doctor or health professional owes to answer to the patient or administration heads, "*why*" and "*when*" at every stage of its' diagnostic and treatment approach and possible answers like "*I don't really know*" or like "*it works, it is OK…*" are not acceptable as in Eastern medical approaches or the so-called "*traditional*". Even A.I. "*black box*" solutions are not permitted from medical community members [10, 11].

MEDICAL STATISTICS AND DATA MINING

The heart of statistical approximation is that the mathematical representation and understanding in medical data are deeply different from other data mining activities because all diagnostic and treatment approaches are imprecise, with serious error risks [11]. The risk here deals with human life and not with money, products, consumer behavior, technical accuracy, etc. Additionally, the variety of biomedical equipment and the aforementioned software cannot easily malfunction or "*reset*". As all experts in biostatistics know statistical approaches are affected not only from the actual data that has been collected but also from "a priori" assumptions of the researchers behind the experimental protocols. The most frequently used statistical tests are designed from the idea of a repeatable experiment, but when employed in medicine this may be subject to ambush [1].

Common problems in medical statistics are [14]:

-*variable classification:* usually undergoes some sort of conversion or analysis following initial collection (real differences or percentage of volume change of a tumor)

-*biases:* when those patients who are lost to follow-up differ in a systematic way to those who did return for assessment, subjects in different arms of the study are treated differently (other than the exposure or intervention)

-*confounding variables:* are those which are linked to the outcome but have not been accounted for in the study protocols

-*standard normal distribution:* that has particular importance in statistics, usually a small number of big deviations result in *leptokurtosis* and large amounts of small deviations result in *platykurtosis*.

-*standard error:* that should be here used to describe the *precision* in the sample parameters

-tests and p-values: in order to establish the likelihood that the association we are observing is genuine, or simply due to chance of the selected sample but as normal distribution that in medicine is somehow rare reflects on mathematics of used statistical tests [1, 13] as some statistical tests are designed, not as a search for truth, but as a search for the winning medical scientist. As a federal grant administrator recently observed, many medical researchers who seek grant funding typically formulate their ideas in terms of ‘‘*fairness for me*’’, rather than ‘‘*fairness for competing medical hypotheses.*’’ [1]

-the number needed to treat and harm: the number of patients that would have to receive the intervention in question in order to prevent one adverse event, and the number needed to harm is a similar concept only applied to interventions that have a detrimental effect on patient health.

-sensitivity, specificity, and predictive values: Sensitivity and specificity deal with the diagnostic capacity of the test and the predictive values demonstrate how likely it is that an individual does or does not have the disease based on their test results.

The target of the Test	When we use Parametric data	When we use Nonparametric data
Assess the difference between the two groups	*Two-sample t-test*	Wilcoxon rank-sum test Mann-Whitney U test Kendall's S-test
Assess the difference between more than two groups	*One-way analysis of variance (ANOVA)*	Kruskal-Wallis test
Measure the strength of an association between two variables	*Correlation coefficient*	Kendall's tau rank correlation Spearman's rank correlation
Assess the difference between paired observations	*Paired t-test*	Wilcoxon signed-rank test Sign test

Figure 3: *Common medical statistical tests*

The usual method in medicine for measuring error *is sensitivity* and *specificity* analysis. Before this one should distinguish between a test and a final diagnosis in medicine because they are different and both tests and diagnoses are subject to sensitivity/specificity analysis [3, 12].

Accuracy calculations are different in medical fields or health care administration. In the end, the most important factor may be the “*false positive*” or

"*false negative*" real numbers and not the "*positive predictive*" or "*total accuracy*" values. This happens because dealing with specific patient's life doctors don't care about the "*mean*" but the "*upper or lower limit*" values in order to decide.

All the above limitations of biostatistics, data mining strive toward discovering some structure in data, draws heavily from many other disciplines, most notably machine learning. Data mining differs from statistics in that it must deal with heterogeneous data fields in health records, not just heterogeneous numbers, as is the case of statistics. The best example of heterogeneous data is medical data that, say contains images, signals, clinical information as well as the physician's interpretation written in short plain text. Most successful examples in data mining are due to advances in database technology structure and access, rather than to advances in data mining tools because due to a subset of data is selected from a large database that most data mining tools are actually applied.

As a result, in day-to-day clinical practice, the details of the overall treatment in a given patient may vary slightly because consent and guidance data vary from patient to patient. This is called "*evidence-based*" medical practice. Another peculiarity of the mining of medical data is that the basic data structures of medicine are poorly mathematical or even logical in comparison with many fields of science. Engineers collect data that they can insert into formulas, equations, and models that adequately reflect the relationships between their measures. On the other hand, the conceptual structure of medicine consists of a description of words and images with very few formal limitations in vocabulary, in the composition of images, or in the permissible relationships between basic terms. Parameters used in medicine, such as inflammation or neoplasia, are real to the physician, such as other sciences such as length or strength. However, there is no comparable formal structure in medicine in which a data miner can organize information that can be modeled by pooling, regression models or sequence analysis [1, 12]. In order to defend it, medicine has to deal with too many different anatomical sites and pathologies. So far, the breadth of this conceptual space has been overwhelming. In addition, some suggestions suggest that the logic of medicine is significantly different from the logic of the engineering sciences [3, 13].

Therefore, we theoretically design a clinical study with a specified null hypothesis and a specified sample size and carry out the study until the agreed sample size is reached. A unit may not be interrupted (proven) by a study if a numerical value for statistical significance has been reached until the predetermined sample size is reached. The reason for this is that mathematical thinking explains the experimental results based on the original experimental design and the expectations that this design brings with it, the so-called a priori thinking. The basic assumptions of the statistical test cannot be restored during the examination. The dilemma created by this paradigm is that there may be compelling evidence that a priori assumptions

are wrong before a predetermined sample size is reached, and that these wrong assumptions are harmful to patients.

There is an ambush similar to that of data mining tools, such as *artificial neural networks*, where the paradigm of the training/test set exists. If someone has exhausted the network training observations, they can no longer take the test for those observations: they are fraudulent. Instead, we have to use elements from a different set, that is, the amount of data. Unlike animal testing, the problem with medical data is that we can no longer recruit people, repeat the experiment, and create a different training set. You have to make the same observations ethically over and over again.

A new doctrine is emerging, according to which the methods of data processing, in particular statistics, and the underlying assumptions on which these methods are based can differ significantly with regard to medical data. Human medicine is primarily a patient care activity and only serves as a secondary research area. In general, the only justification is to collect medical data or to refuse to collect certain data for the benefit of a single patient. Some patients may agree to participate in research projects that they do not directly benefit from. However, this data collection is usually very small, narrowly focused and is subject to strict legal and ethical scrutiny. Another problem with the paradigm used in artificial neural networks is that they may want to examine the entire data set for reasons other than network training, e.g. In this case, hiding a subset of examples from the training set is self-sufficient. On the other hand, if we look at all cases, we contaminate them a priori reasoning required for the training course - the test [1, 3].

Consequently, a data miner should be aware of the differences between statistics and data mining and their limitations in order to reach the proper conclusion of his /her research approach from the early beginning [15]:

-*Data implementation* is the beginning of data science and covers the entire process of data analysis, *statistics* form the basis and core of the section on data mining algorithms.

-*Data mining is an exploratory analysis process* in which we first examine and collect data and build a model on it that can recognize a pattern and build theories on it to predict future results or solve problems. *Statistics* is a validation process, theories are first created and then validated to test the data sets. Uses only numeric data types for probability and mathematical calculation and prediction.

-*Data execution is an inductive process* and uses an algorithm such as a decision tree, a clustering algorithm to derive a data partition and hypothesize from the data, *statistics* are a deductive process and not predictions for acquisition and verification of knowledge contains hypotheses.

-*Data mining is not a big problem for data acquisition* or acquisition because it is an exploratory data analysis that is mostly a software and calculation process for pattern recognition in large databases, *statistics* are more required for data

collection. All data collected can be quantitative or qualitative, primary or secondary data.

Data Mining	Statistics
Explore and gather data first, builds a model to detect patterns and make assumptions	It provides theories to test using statistical methods
Inductive Process (Generation of new theory from data)	Deductive Process (Does not involve making any predictions)
Data Cleaning is part of data mining.	Clean data before applying the statistical methods.
Needs less user interaction to validate model, easy to automate.	Needs user interaction to validate model, difficult to automate
Suitable for big data sets	Suitable for small data sets
It's an algorithm that learns from data without using any programming rule.	Formalization of relationship in data in the form of the mathematical equation
Use heuristics thinking (rules used to form judgments)	No heuristic thinking
Classification, Clustering, Neural network, Association, Sequence-based analysis, Visualization	Descriptive Statistical, Inferential Statistical
Biological Data Analysis, Medical Applications etc.	Biostatistics, Quality Control, etc.

Figure 4: *The differences between Data Mining vs Statistics in health care*

-*Data cleanup when retrieving data is the first step*, as it helps understand and correct the quality of the data to get a precise final analysis. During data cleansing, the user has the option of deleting incorrect or incomplete data. After collecting data from various sources, *statistical* methods are used for confirmatory analysis.
-*Data mining deeply digs previously available unknown* but useful information from large databases to help you make critical decisions. A number of methods are used to find patterns and relationships within the available data. This is the convergence of various processes, including statistics, machine learning, database management, artificial intelligence (AI) and data recognition, etc., because *statistics* are an important component of data mining, which provides efficient analysis techniques and tools for processing large amounts of data.
-*Data mining is primarily used for commercial applications* such as financial data analysis, retail, telecommunications, biology, and other scientific discoveries. *Statistics* are used in every data sample to create a range of new information. It describes the character of the data to be analyzed and examines the relationship between them. Uses predictive analytics to run scenarios to determine future actions. On the other hand, statistics breathe in lifeless data. Some of the popular trends in data mining are application research, visual data

mining, biological data mining, online mining, software mining, data distribution, real data mining, and more. *Statistics* help identify new patterns of available unstructured data.

Due to the advent of big data with a large amount and different data speeds, the decision plays an important role in every field and the prediction of the results is crucial for data mining and statistics. *Data mining always uses statistical thinking to access the results.* In the end, different results of a statistical breakdown are not completely random but are limited by the fact that some combinations of health events are normal or rare. Expert doctors usually recognize these events as common or rare, but the exact probabilities are still unknown [15].

ETHICAL AND LEGAL ISSUES

Patient-identifiable data refer to any personal data that can be used to identify an individual (eg, name or registry number). This also includes encrypted data if the solution for decryption is still in existence (eg, new National Health Service (NHS) number) [15]. Patient-identifiable data are critical in medical research and required for the developmental stages of disease registers. The problem is that pursuing informed consent for the use of such data is likely to result in a biased sample and is often prohibitively expensive. The issue of data ownership is an open question in the mining of medical data. In legal textbooks, ownership determines who is entitled to sell a property. Because the sale of human data or tissue does not seem to make sense, the issue of data ownership in medicine is similarly confusing. The corpus of human medical data available to obtain data is enormous. Thousands of terabytes are now generated annually in Western countries. However, this information is buried in heterogeneous databases and scattered across the healthcare facility, without the common format or principles of the organization. The issue of ownership of patient information is unresolved and subject to repeated, highly publicized lawsuits and congressional investigations. Many questions are still open [1]:

- *Do individual patients have collected information about themselves?*
- *Do their doctors have information?*
- *Do their insurance providers own the information?*
- *If insurance providers do not have their own insured information, can they refuse to pay for data collection and storage?*
- *If the ability to process and sell human health data is unreliable, then how do we replace the data controllers who organize and mine the data?*
- *Or should this incredibly rich source of potential improvement for humanity remain unchanged?* [18]

In this uncertainty, doctors and other medical data manufacturers understandably hesitate to share their data with data miners.

Data miners could search this data for adverse events. Obvious medical history abnormalities can trigger an examination. In many cases, the occurrence of incorrect practices can be an error in omitting or transferring data and not all bad medical results are necessarily the result of the negligent behavior of the provider.

Another specialty is data protection and security issues. For example, U.S. federal regulations set guidelines for hiding individual patient identifiers. It is not just a possible breach of patient confidentiality with the possibility of a lawsuit and the erosion of the doctor-patient relationship, in which the patient is extremely open to the doctor, expecting that such private information will never be published. According to some guidelines, disguising identifiers must be irreparable. A related data protection problem can be used, for example, to determine important diagnostic information about a patient's data and to treat a patient if he or she can only return and inform the patient of the diagnosis and possible cure. In some cases, this measure cannot be carried out. Another problem is data security when handling data, especially when transferring data (Figure 5).

Only authorized persons should have access to the identifiers before they are hidden. Because electronic data transmission over the Internet is unreliable, identifiers should be carefully hidden, even when being transferred from one unit to another within a healthcare facility.

On the other hand, for example, recent US federal documents have found that there are at least two legitimate research needs to re-identify undefined medical information:

- *to prevent the same patient's records from being accidentally copied from biased research results.*
- *the original (newly identified) medical records may need to be persuaded to verify the accuracy or to obtain additional information about specific patients.*

These specific requirements could be addressed by the relevant regulatory authorities, but could not be met at all if the data were completely anonymous [1].

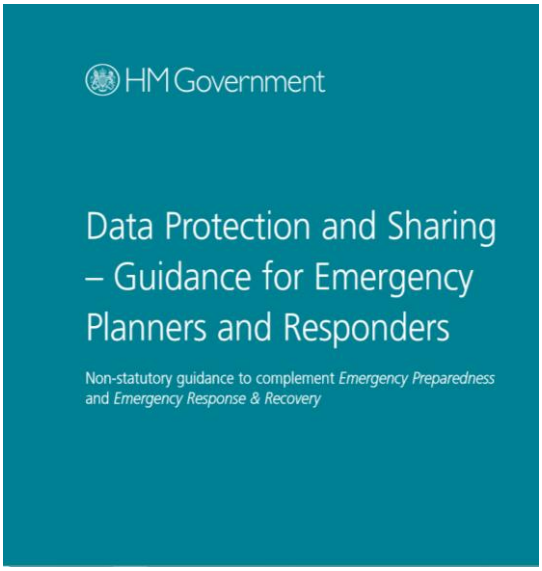

Figure 5: *A typical guide for information sharing and ensuring that necessary privacy protection is properly seen as complementary objectives and not competing ones.*

Consequently, there must be legal contracts between the organization and all external parties that have access to individually identifiable health information, and must require external parties to protect the information [17]:

- *There must be contingency plans, including a data backup plan and an emergency response plan.*
- *There must be a system for controlling access to information, which includes policies for granting, establishing and modifying access rights to data.*
- *In order to identify potential security breaches, an internal review of data access records must be kept in place.*
- *The organization must provide oversight of personnel performing technical system maintenance activities to maintain access permit records to ensure that operators and maintainers have adequate access, apply personnel safety procedures and train system users.*
- *There must be termination procedures in place when an employee leaves or loses access to data.*
- *There should be security training for all staff, including awareness training for all staff, regular security reminders, educating users on*

virus protection, educating users about the importance of monitoring login errors, managing passwords, and reporting derogations.

These and many other rules impose restrictions on medical data providers that would be burdened and inhibited by other academic researchers in scientific research creativity. Researchers need to carefully weigh the perceived need for information that could allow the data to be re-identified in combination with other information [1, 17].

SUPERVISION FROM FIELD EXPERTS

According to the kind of data available and the research question, a scientist will choose to train an algorithm using a specific learning model. There are four approximations:

- *Supervised learning model*, where the algorithm learns on a labeled dataset, providing an answer key that the algorithm can use to evaluate its accuracy on training data (someone is present judging whether you're getting the right answer).

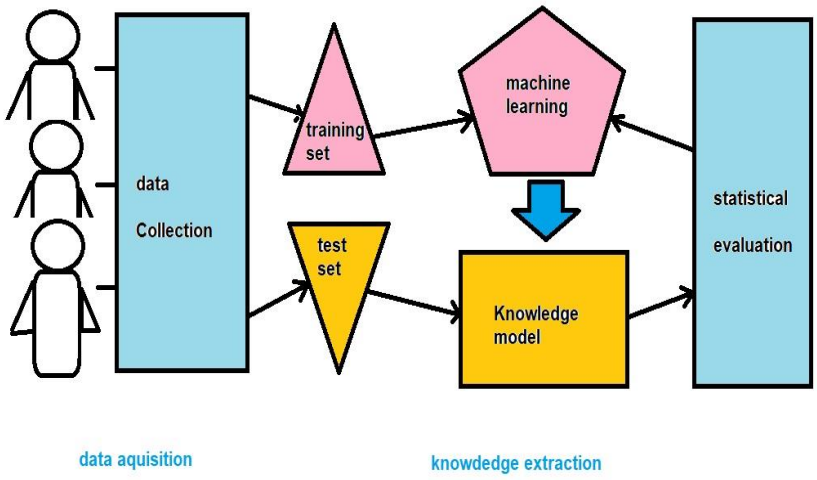

Figure 6: *Knowledge extraction from data*

- *An unsupervised learning model* that, provides unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own.

- *The Semi-supervised learning* that, uses a small amount of labeled data bolstering a larger set of unlabeled data.
- *The Reinforcement learning* finally that, trains an algorithm with a reward system, providing feedback when an artificial intelligence agent performs the best action in a particular situation.

There are two main areas where *supervised learning* is useful: classification problems and regression problems [18].

Classification problems ask the algorithm to predict a discrete value, identifying the input data as a member of a particular class, or group.

On the other hand, regression problems look at continuous data. One use case, linear regression, should sound familiar from algebra class: given a particular x value, what's the expected value of the y variable?

Supervised learning is, thus, best suited to problems where there is a set of available reference points or a ground truth with which to train the algorithm. But those aren't always available.

Clean, perfectly labeled datasets aren't easy to come by. And sometimes, researchers are asking the algorithm questions they don't know the answer to. That's where *unsupervised learning* comes in.

In unsupervised learning, a deep learning model is handed a dataset without explicit instructions on what to do with it. The training dataset is a collection of examples without a specific desired outcome or correct answer. The neural network then attempts to automatically find structure in the data by extracting useful features and analyzing its structure.

Unsupervised learning models automatically extract features and find patterns in the data.

Depending on the problem at hand, the unsupervised learning model can organize the data in different ways.

- *Clustering:* That's how the most common application for unsupervised learning, clustering, works: the deep learning model looks for training data that are similar to each other and groups them together.
- *Anomaly detection:* Unsupervised learning can be used to flag outliers in a dataset.
- *Association:* By looking at a couple of key attributes of a data point, an unsupervised learning model can predict the other attributes with which they're commonly associated.

Because there is no "*ground truth*" element to the data, it's difficult to measure the accuracy of an algorithm trained with unsupervised learning. But there are many research areas where labeled data is elusive, or too expensive, to get. In these cases, giving the deep learning model-free rein to find patterns of its own can produce high-quality results [18]. Maybe the most important issue in medical data mining and new knowledge discovery is that most of the health care data have particular obligations concerning their openness to other people that

are not dealing with the specific patient; their emergency (often dealing with life-and-death); and the ethical commitment to be used only for salutary. A.I. applications usually are not accepted by health professionals. The proper path for effective and usable mining in medicine includes several important steps:

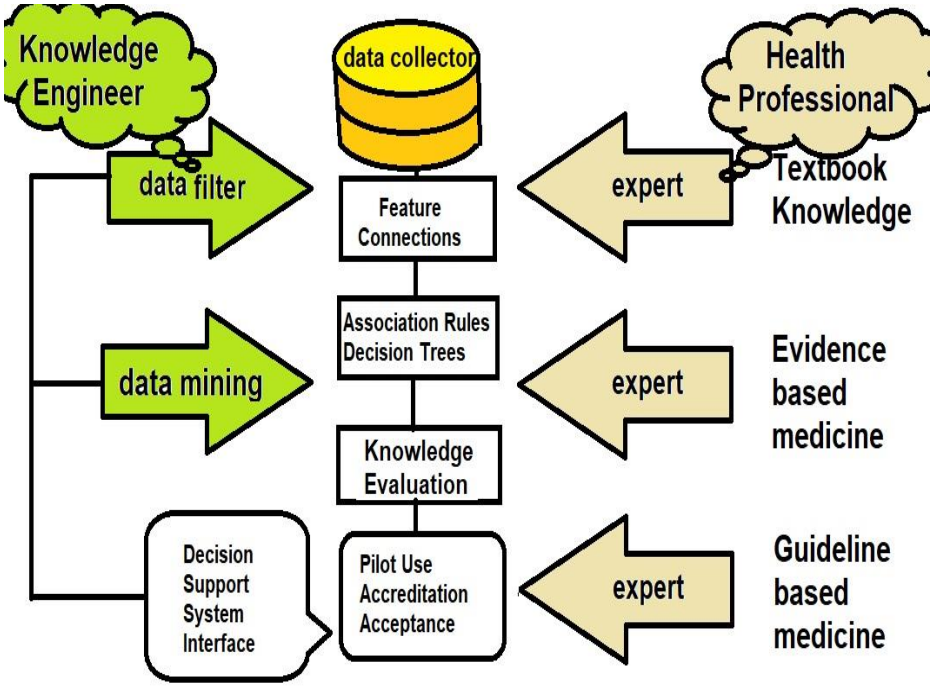

Figure 7: *Data Mining and the development of intelligent systems for medical applications*

The first step deals with data, the second with knowledge extraction and the final with acceptance test from the medical community. During previous decades all the procedure of knowledge representation through interviews with field experts has driven most of the promising applications to a "*bottleneck*" or at most to the first pilot intelligent system. Actually, when working with A.I. tools in order to develop a smart system to solve problems or to support human decisions, we describe one that can continually accept, store, evaluate and "*learn*" from new data adapting its' knowledge base after each new entry (Figure 7). Thus the following steps are looping, internally:

1. *Data collection*
2. *Knowledge Extraction*
3. *System acceptance*

One of the best ways to visualize extracted knowledge is through a *decision tree*. Decisions Trees are popular techniques for supervised classification, especially when the results are interpreted by a field expert. Medical doctors can easily accept this type of visualization because it is very close to their diagnostic trees.
The decision tree is a decision support system that uses tree-based decisions and their possible effects, including random event results, resource costs, and benefits. A decision tree or classification tree that learns the value of a dependent attribute (a variable) according to the values of the independent (input) attributes (variables) is used to teach the classification function. This confirms a problem called controlled classification because the number of dependent attributes and classes (values) is given [19]. Decision trees are the most powerful approaches to knowledge discovery and data collection. It involves the technology to examine the vast and complex majority of the data to discover useful patterns (Figure 8). This idea is very important because it enables the modeling and drawing of knowledge from most of the available data. All experts are constantly on the lookout for techniques to make the process more efficient, cost-effective and precise. The decision tree offers many advantages when retrieving data.

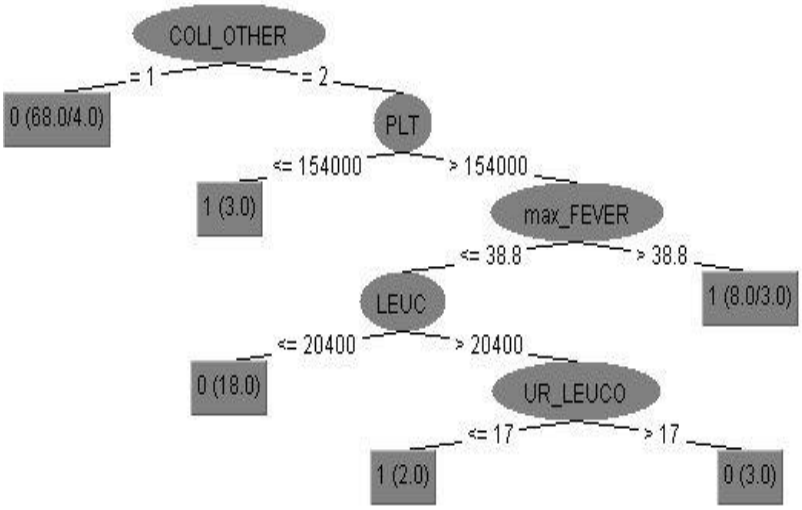

Figure 8: *Decision Tree*

Some of them are:
• *It is easy for the end-user to understand.*
• *It processes various input data: nominal, numeric and textual*
• *Ability to process incorrect records or missing values*
• *High performance with little effort* [19]

The tree contains - *root nodes, leaf nodes* that represent arbitrary classes and *internal nodes* that represent test conditions that are applied to attributes (Figure 8).

Additionally, it is up to those who create the decision tree because it is not that complicated and easy to understand. Typically, tree complexity can be measured using a metric that includes: total number of nodes, the total number of leaves, tree depth, and number of attributes used in tree construction. The size of the trees must be relatively small, which can be controlled by a technique called pruning [19]. The induction of decision trees is closely related to the induction of rules. Any path that starts from the root of the decision tree and ends at one of its permissions is a rule. These rules can be created very easily.

Common language, good communication with end-users and mined knowledge supervising only from the field experts, by well-established health care guidelines, are also important aspects that are undetectable issues of an effective data mining engineer. Transparency of medical knowledge extracted from big medical databases is the *holy grail*, (as in decision trees) resulting in the appropriate acceptance from end-users as medical community members. Finally, the researcher must be ready even for the unacceptable: there are situations that even the intelligent system has revealed knowledge against national or international guidelines, for example, that uterus cancer is associated most with the number of sex partners or voluntary abortions, or that a group of the supported that entered the hospital during nights or are immigrants will be the most possible patients to die soon in an emergency unit of a hospital and consequently is not helpful for the doctors or other health professionals or counselors that are on the line of guideline-based medicine!

CONCLUSIONS

According to all previous, human health care data are the most difficult of all data to collect, deal and analyze. The first important issue is that the mathematical representation and understanding statistical approach in medical data are internally different than those from other data mining activities.

As in medicine the primary direction is the care work for a specific patient, and secondarily dealing with him as a research resource; almost the only acceptable reason for collecting medical data is to benefit against the aforementioned

individual disease. Software engineers or researchers that are working in other fields will never face the same constraints of primarily privacy-sensitive, greatly heterogeneous, and massive medical data.

The second important issue is the serious ethical and legal aspects of data mining in this field, including data ownership and special data treatment issues. Consequently, researchers should be aware when are dealing with medical databases they may face the possibility that their work will never be accepted or even used from health care professionals if all these obligations will not be correctly addressed from the early beginning.

The last important issue is that most of the health care data have particular obligations concerning their openness to other people that are not dealing with the specific patient; their emergency (often dealing with life-and-death); and the ethical commitment to be used only for salutary. Good communication with end-users and mined knowledge supervising only from the medical field experts, by the well-established health care guidelines, are also important aspects that are undetectable issues of an effective data mining engineer. Finally, only *transparency* of medical knowledge extracted from big medical databases is the most important issue, resulting in the appropriate acceptance from end-users as medical community members.

REFERENCES

1. Koh HC, Tan G. Data mining applications in healthcare, J Healthc Inf Manag. 2005 Spring; 19(2):64-72.
2. Mayer-Schönberger, V. and K. Cuckie, Big Data*: A Revolution that Will Transform How We Live, Work, and Think*. 2013: Houghton Mifflin Harcourt. Chapter 2.
3. Moore GW, Hutchins GM. *Effort and demand logic in medical decision making*. Metamedicine 1980;1:277–304.
4. Megalooikonomou, V., et al., *Data mining in brain imaging.* Stat Methods Med Res, 2000. **9**(4): p. 359-94.
5. Nayak, L., et al., *Computational neuroscience and neuroinformatics: Recent progress and resources.* J Biosci, 2018. **43**(5): p. 1037-1054.
6. Kourou, K., et al., *Machine learning applications in cancer prognosis and prediction.* Comput Struct Biotechnol J, 2015. **13**: p. 8-17.
7. Vougas, K., et al., *Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining.* Pharmacol Ther, 2019: p. 107395.

8. Dimitrov, D.V., *Medical Internet of Things and Big Data in Healthcare.* Healthc Inform Res, 2016. **22**(3): p. 156-63.
9. Ventola, C.L., *Big Data, and Pharmacovigilance: Data Mining for Adverse Drug Events and Interactions.* P T, 2018. **43**(6): p. 340-351
10. Manning CD, Schuetze H. *Foundations of statistical natural language processing*. Cambridge (MA): MIT Press, 2000.
11. Friston, K.J., *Statistical parametric mapping: the analysis of functional brain images*. 1st ed. 2007, Amsterdam; Boston: Elsevier/Academic Press. 647 p.
12. Cios KJ, Moore GW. *Medical data mining and knowledge discovery: an overview*. In: *Cios KJ, editor. Medical data mining and knowledge discovery*. Heidelberg: Springer, 2000. p. 1–16 [chapter 1].
13. *Brewka G, Dix J, Konolige K. Nonmonotonic reasoning: an overview. CSLI Lecture Notes No. 73, ISBN 1-881526-83-6, 1997. p. 179.*
14. Petrie A, Sabin C. *Medical statistics at a glance*. Chichester: John Wiley & Sons Ltd; 2009.
15. EDUCBA https://www.educba.com/data-mining-vs-statistics/ (accessed 16 January 2020)
16. Data Protection and Sharing– Guidance for EmergencyPlanners and Responders, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/_attachment_data/file/60970/dataprotection.pdf, (accessed 20 January 2020)
17. Haynes LC, Cook AG, Jones AM. *Legal and ethical considerations in processing patient-identifiable data without patient consent: lessons learned from developing a disease register*, J Med Ethics. 2007 May; 33(5): 302–307.
18. *https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/* (accessed 26 January 2020)
19. Bhargava, N, Girja S, Ritu B and Manish M. *Decision Tree Analysis on J48 Algorithm for Data Mining. (2013*), IJARCSSE *3:6:114-19.*