The Practice of Predictive Analytics in Healthcare

Article · April 2013

CITATIONS
5

READS
12,349

1 author:

Gopalakrishna Palem

24 PUBLICATIONS **31** CITATIONS

SEE PROFILE

The Practice of Predictive Analytics in Healthcare

Gopalakrishna Palem

Contents

Predictive Analytics & Healthcare

Falling sick is not just an individual's problem. Nations crumble when their people are not strong. History is full of events riddled with contagious diseases that brought whole societies to their knees. Kofi Annan, Secretary-General of the United Nations 1997-2006, once aptly put it as [1]:

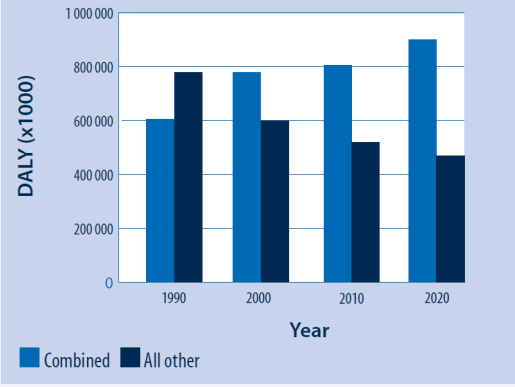
> "When we are sick, working is hard and learning is harder still. Illness blunts our creativity, cuts out opportunities. Unless the consequences of illness are prevented, or at least minimized, illness undermines people, and leads them into suffering, despair and poverty."

There is a two-way interdependent relationship between economic poverty and chronic disease. Many of the world's poor, despite regional differences in geography, culture and commerce, experience the same discouraging cycle: being healthy requires money for food, sanitation and medical care - but to earn money one must be healthy. And this problem is compounded by poor drug adherence.

A number of rigorous reviews found that, in developed countries, drug adherence among patients suffering from chronic diseases averages only 50%. It is much lower in developing countries.

Noncommunicable diseases, mental health disorders, HIV/AIDS and tuberculosis, *combined* represented 54% of the burden of all illness worldwide in 2001 and will exceed 65% of the global burden in 2020. Contrary to popular belief, noncommunicable diseases and mental health problems are also prevalent in developing countries, representing as much as 46% of the total burden of diseases for year 2001 and predicted to rise to 56% by 2020 [1].

Figure: Burden of chronic conditions worldwide



DALY: disability-adjusted life year
Combined: noncommunicable diseases + mental disorders + AIDS + TB
source: World Health Organization

There is a strong evidence that many patients with chronic illness including asthma, hypertension, diabetes and HIV/AIDS, have difficulty adhering to their recommended medical regimes. This results in less than optimal management and con-

[1] Kofi Annan, Secretary-General of the United Nations on the occasion of the release of the *Report of the Commission on Macroeconomics and Health*, in London, 20 December 2001

trol of the illness. Poor adherence is one of the primary reasons for suboptimal clinical benefit [1]. It causes medical and psychological complications of disease, reduces patient's quality of life, and wastes health care resources. Taken together, these direct consequences impair the ability of health care systems around the world to achieve population health goals.

Population-based strategies seek to change the social norm by encouraging an increase in healthy behavior and a reduction in health risk. They target risks via legislation, tax, financial incentives, health-promotion campaigns or engineering solutions. However, although the potential gains are substantial, the challenges in changing these risks are great. Population-wide strategies involve shifting the responsibility of tackling big risks from individuals to governments and health ministries, thereby acknowledging that social and economic factors strongly contribute to disease.

An important facet the socio-economic factors contribute to is *risk transition*. As a country develops, the types of diseases that affect a population shift from primarily infectious, such as diarrhoea and pneumonia, to primarily noncommunicable, such as cardiovascular disease and cancers. This shift is caused by:

- improvements in medical care, which mean that children no longer die from easily curable conditions such as diarrhoea
- ageing of the population, because noncommunicable diseases affect older adults at the highest rates
- public health interventions such as vaccinations and the provision of clean water and sanitation, which reduce the incidence of infectious diseases.

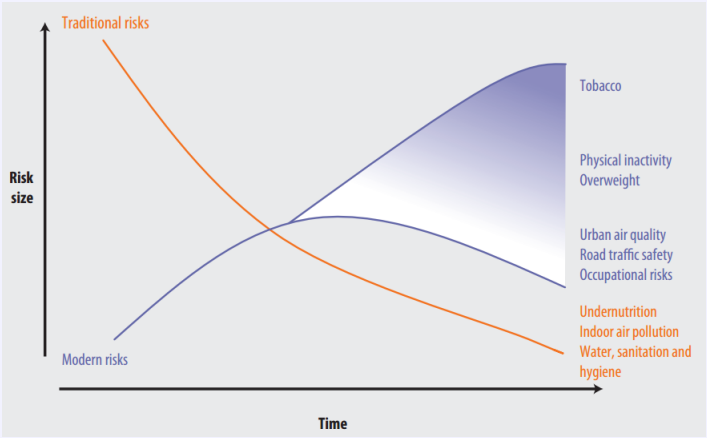


Figure: Risk transition

Socio-economic factors play important role in public health. Low-income populations are most affected by risks associated with poverty, such as under-nutrition, unsafe sex, unsafe water, poor sanitation and hygiene, and indoor smoke from solid fuels; known as the *traditional risks*. In the developed countries, on the contrary, as the major causes of death and disability shift to the chronic and noncommunicable, populations are increasingly facing modern risks due to physical inactivity; overweight and obesity, and other diet-related factors; and tobacco and alcohol-related tasks. As a result, many low- and middle-income countries now face a growing burden from the modern risks to health, while still fighting an unfinished battle with the traditional risks to health.

Modern technologies in an effort to create comfortable life, are rather endangering human race with modernized risks. As modern medicine grew, so did modern diseases with no known cure. Did you know *hypertension* is the one disease that is causing most healthcare spends with 54% share of the total healthcare spendings in the United States? [2]

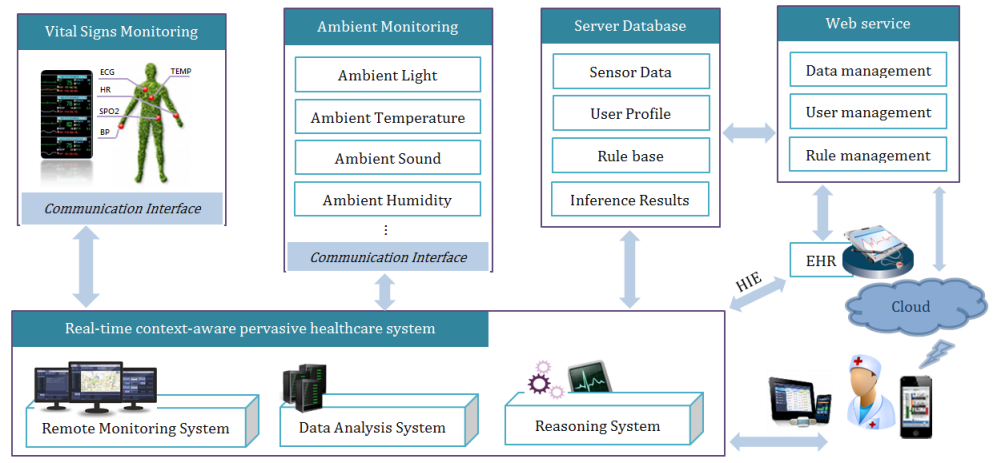
Overweight and obesity is the 5^{th} most leading risk factor in the death-due-to-diseases list, with *urban outdoor air pollution* taking 14^{th} place [3]. Sensitization rates to one or more common allergens among school children are currently approaching 40%-50%, with teens and pre-teens reporting more asthma cases year by year than middle-aged or old-aged ones [4]. Adverse drug reactions (ADR) may affect upto 1/10 of the world's population and upto 20% of all hospitalized patients; and have significant socio-economic impact on both direct costs (management of reactions and hospitalizations) and indirect costs (missed work/school days; alternate drugs) [4]. Inaccurate diagnosis (under-diagnosis or over-diagnosis) is the major contributing factor for the adverse drug reactions. All these are modern risks, attributable to modern lifestyle caused by technology advancements impacting the daily life.

However, not all technology advancements are repressive. Take big-data for example. This young technology, though it has been there for quite a long time since the early days of computation under different names, as *distributed computing*

and *high-performance computing*, in its new form, however, combined with predictive analytics, is promising to address majority of the today's healthcare cost concerns and quality concerns.

Consider the real-time health monitoring system presented in the figure below. Telemedicine and health monitoring systems such as this have been around for quite a while, and been growing slowly and steadily. However, in recent times, several factors have boosted the growth of this sector:

1. Developments in the field of electronics, especially micro-electromechanical systems (MEMS) and system-on-chip (SoC) technologies have vastly improved the size, quality and precision of medical equipment.
2. Wireless technologies have enabled remote delivery of services. Standardization and adaptation of IPV6 across industry has opened the possibility for billions of uniquely addressable IP devices, such as smart sensors.
3. Information technology, especially the cloud and big-data, has improved the ways in which medical information is shared, stored and processed.
4. Availability of large-scale data processing on commodity hardware is enabling radical reimplementations of clinical decision support systems and unstructured medical record text processing systems.
5. Federal regulatory requirements and financial incentives are driving many to adapt the new connected systems approach.



Real-time health monitoring system

Smart devices and smart application development kits are creating new opportunities for application developers to bring telemedicine and health monitoring to the masses—literally. In fact, the use of mobile devices for keeping track of vital signs, analyzing information collected by medical equipment, communicating health status to doctors and receiving advice back from them has expanded the scope of telemedicine and health monitoring.

From being used only for delivering medical services remotely to rural areas initially, these technologies are now also being used to provide better services to patients in hospitals, take care of specially-challenged or aged people, monitor foetal health and even improve the fitness level of people.

There is a wide array of telemedicine and health monitoring equipment available today. Broadly they can be classified as diagnostic and therapeutic equipment like electrocardiographs, peacemakers; imaging equipment like X-ray and magnetic resonance imaging (MRI); medical instruments like blood analysis systems and dialysis systems, patient monitoring and consumer products like remote monitoring tools, insulin pumps and heart-rate monitors, as well as wellness equipment like pedometers and cardio trainers.

Such equipment can be thought of as having a layer of information and communication technology (ICT) atop a medical device. These usually comprise a combination of sensors, MEMS, low-power high-performance microcontrollers, wireless modules, signal conditioners, analogue-to-digital converts, all of which work alongside a user-friendly software. The infrastructure required includes uninterrupted power supply and network connectivity.

The goal of predictive analytics is to help companies transform data into actionable insights that can improve business decisions. Increased global competition and the need for sustainable growth are pushing more and more companies to adapt analytical approaches for business insights. Healthcare organizations more than ever are being seen using analytics to consume, identify and apply new insights from information. Innovative analytical methods are being used to drive

clinical and operational improvements to meet business challenges. The below summarizes how predictive analytics is benefiting different segments of the healthcare industry:

Life-sciences: Aid in clinical research and drug discovery

Healthcare providers: Aid in clinical decision support and diagnostic assistance

Insurance providers: Aid in optimizing healthcare costs and preventing fraud

Public health: Aid in monitoring public health status and identifying epidemic outbreaks

Individuals: Aid in real-time health monitoring and critical care intervention

From being extensive users of descriptive analytics, using reporting based tools and applications descriptively to understand what happened in the past and classify, categorize historical, structured data, healthcare organizations are moving towards predictive analytics techniques that take an understanding of the past to predict future activities and model scenarios using simulation and forecasting, supporting advanced capabilities such as enterprise analytics, evidence-based medicine and clinical decision support systems. A case in point being, analytics enabling the compilation of information about trends, patterns, deviations, anomalies and relations in the patients' medical history data across cultures and regions to detect any epidemic breakouts in real-time. Some of the use-case scenarios are as below.

Critical care intervention: Clinical surveillance for real time bedside and remote monitoring solutions provide clinicians with proactive alerts for critical new values, reducing patient's risk of infection, adverse drug events, and other potential complications.

Diagnostic assistance: Advanced voice and natural language processing methodologies help physicians recommend the right list of diagnosis routines based on patient symptoms to quickly arrive at the illness root-cause with optimal diagnostic test costs.

Clinical decision support: Analyzing the diagnosis results by cross-referencing patient medical history notes, prior cases, clinical trials and reference materials, all in real-time to suggest the possible course of treatment actions, intelligent decision support systems help clinicians decide the best treatment plan personalized for each patient.

Disease management: Clinical events and treatment interventions from disparate health providers are notified for central immunity records and epidemic control centers tracking public health safety in real-time, watching for any emerging trends and potential epidemic outbreaks.

Optimized healthcare costs: Analytics models built to identify the risk of patients based on their age, genealogy, diet habits, daily routines, previous medical history and which diseases they are most susceptible to, yields right protection and treatment plans for individual patients, optimizing the healthcare and insurance costs. For example, the data collected by a recent Medical Expenditure Panel Survey (MEPS) reveals that people in the age group 35 and above are suffering from *hypertension* more than any other disease, while teens and young-adults below 30 are suffering more from *depression*. *Asthma* on the other hand is found to be reported more in pre-teens and teens (ref figure .1).

Fraud detection and prevention: Instead of relying on reactive measures, healthcare organizations and insurance providers are taking more hands-on approaches for fraud detection and prevention. Based on key risk indicators, point-in-time or ad hoc testing is helping the insurance providers identify the transactions to be investigated. If the testing reveals indicators of fraud, recurring testing or continuous analysis will be considered. Unlike retrospective analyses, continuous transaction monitoring allows an organization to identify potentially fraudulent transactions on real-time or near real-time basis, such as daily or weekly. Organizations are increasingly using continuous monitoring efforts to focus on narrow bands of transactions or areas that pose particularly strong risks.

Personal healthcare: Pervasive and context-aware applications have been widely recognized as promising solutions for improving the quality of life for both patients suffering from chronic conditions and their relatives, as well as for reducing long-term health care costs and improving quality of care. Telemedicine, assisted living through supportive sensor aids and safety guides, medication alerts, and wearable body monitoring sensors are few personal healthcare initiatives that are gaining increased attention in this area.

Readmission prevention: By gathering patient-specific information, organizations personalize care plans, proactively reduce readmission rates, compare treatment efficacy against industry-wide outcomes, and focus resources on most effective treatments. Effectively targeting and managing patient discharge and follow-up care significantly reduces the re-admission costs and prevents repeat admissions.

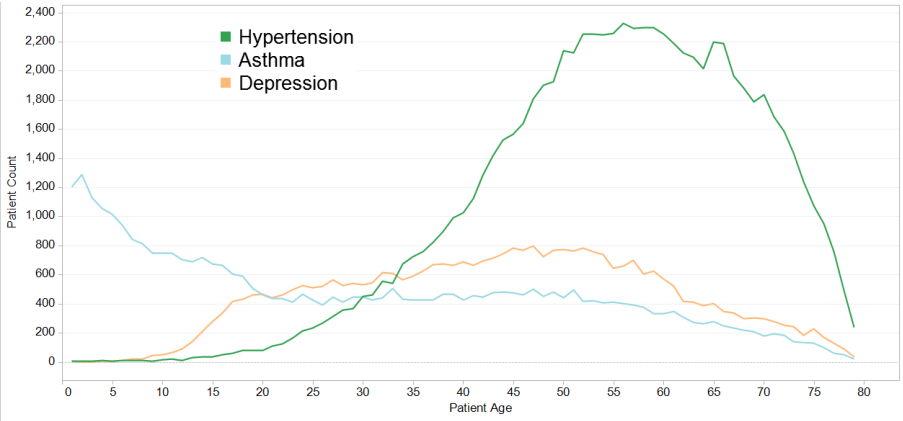


Figure .1: Disease distribution by age. The above data highlights that teenagers and young-adults suffer more from depression, while middle-aged people suffer more from hypertension. Asthma patients are more in pre-teens and teens. source: [2]

Predictive analytics provides significant operational level advantages for healthcare providers. Enterprises rely on predictive analytics to further their understanding of the effectiveness of clinical treatments. The ability to quickly and accurately diagnosing and personalizing patient treatment the first time they are admitted, boosts the patient confidence in the healthcare system, fosters the patient-doctor relationship and reduces the incidence of costly readmissions.

This paper reveals the practice of such predictive analytics in healthcare segment, touching upon the concepts of electronic health records, the meaningful use incentives, natural language processing techniques used in expert decision systems and so on, presenting detailed use-case scenarios relevant for each. Many of these concepts explained in this paper are currently in active use in the industry and the author can be contacted for more information on how your organization can benefit from them and start adapting them.

✧ ✧ ✧

Electronic Health Records

Electronic health record (EHR) is an emerging concept of shared, comprehensive computerized healthcare records in enterprise-wide systems. It is defined to be a systematic collection of electronic health information about individual patients or populations [5]. Primarily it will be a mechanism for integrating healthcare information currently collected in both paper and electronic medical records (EMR) for the purpose of improving the quality of care. Some of the goals of EHR are:

- ✓ To interconnect with and enhance other error-reducing and cost-saving technologies
- ✓ To streamline healthcare data flow using an interoperable and standardized nomenclature
- ✓ To improve quality by encouraging accurate and legible communication among providers
- ✓ To automate adverse event and medical error disclosure
- ✓ To facilitate reliable and reproducible outcomes research and reporting [6].

In a nutshell, Electronic health records (EHRs) are real-time digital versions of patient records that make information available instantly, "whenever and wherever it is needed", bringing together in one place everything about a patient's health. EHRs typically:

- ➪ Contain information about a patient's medical history, diagnoses, medications, immunization dates, allergies, radiology images, and lab, test results
- ➪ Offer access to evidence-based tools that providers can use in making decisions about a patient's care
- ➪ Automate and streamline providers' workflow

A health record is essentially temporal in structure, with time dependent observations of different types i.e. controlled vocabularies, numbers, time series, audio, images etc. and to some extent time dependent opinions and conclusions made by health care providers. As per federal regulations, an electronic health record system should at the minimum facilitate:

- Patient summary record: Being able to share lists of problems, medications, and allergies along with diagnostic test results among different providers and patients
- Electronic prescribing: Route prescriptions to mail order and retail pharmacies electronically
- Submission to public health agencies: Send lab, syndromic surveillance, and immunization data to public health agencies in real-time or near real-time electronically
- Quality reporting: Provide ambulatory and hospital quality measurement data to Centers for Medicare and Medicaid Services
- Problem list (diagnoses): Enter health problems using uniform and consistent terminology
- Lab test results: Codify lab results and observations in a consistent terminology
- Encryption and decryption: Ensure data are secured during transport

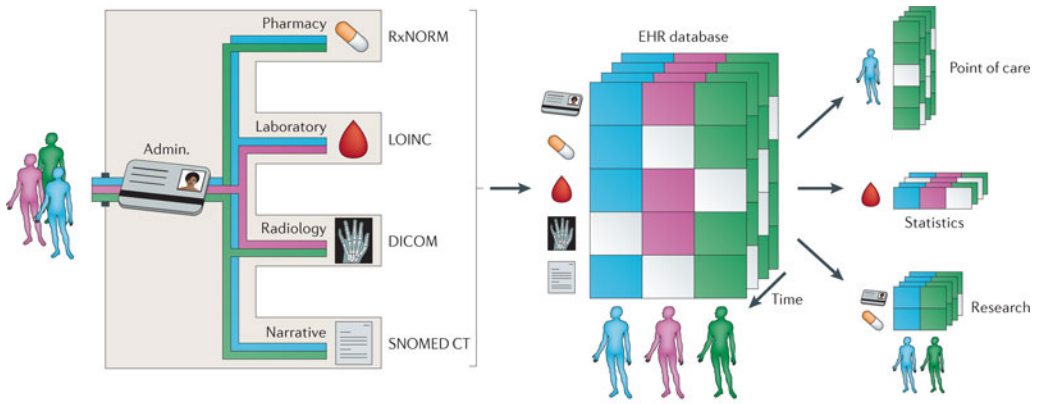


Figure .1: EHR system overview (source: Nature Reviews)

Also few standards are established around EHR technology to support standardized communication, continuity of care and information sharing.

DICOM: Digital imaging and communications in Medicine, is an international file format and communications standard for handling, storing and transmitting radiology and other medical imagery data.

HL7: Health-level 7, provides a framework and related standards for the exchange, integration, sharing and retrieval of electronic health information. Claims attachments, structured product labeling etc. are some of the prominent features of HL7.

ANSI X12(EDI): Specifies transaction protocols used for transmitting patient data. Widely used in the United States for transmission of billing data.

HISA (EN 12967): Health information service architecture is a standard aimed at enabling the design and development of modular open systems to support healthcare. Widely used as a service standard for inter-system communication in a clinical information environment.

CONTSYS (EN 13940): Defines a system of concepts to support *continuity of care*, an organizational principle based on semantic interoperability representing the important aspect of quality and safety in healthcare.

Apart from these, there are content structure related HITSP c62, vocabulary related RxNorm, SNOMED-CT, LOINC, UMLS, ICD and security related LDAP, x509, standards that are prominently being used in the industry in relation with implementing EHR.

A key feature of EHR is that it can be created, managed and consulted by authorized providers and staff across multiple healthcare organizations. A single EHR can bring together information from current and past doctors, emergency facilities, school and workplace clinics, pharmacies, laboratories and medical imaging and diagnosing facilities.

> “Most EHR initiatives are national in scope and frequently government initiated or funded. EMR initiatives are typically hospital-or-system-wide. A personal EHR model is quite different in concept. It assumes that individual patients will aggregate their diverse records and then make them selectively available to new or emergency providers . . . ”

Although there has been debate among providers about the feasibility and safety of having all patient information computerized and available across institutions, many are starting to realize that EHR implementation is inevitable because of the support for the idea from healthcare regulators, third party players, hospital administrators, and physician advocacy groups such as American Medical Association. More efforts to increase EHR adaptation are coming through federal guidelines, regulations, and financial incentives. For example,

- Administrative Simplification provisions of the United States in Title II of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) mandates streamlining the back-end administrative functions by standardizing identifiers and formats for transactions and code sets.

- The Healthcare component of American Recovery and Reinvestment Act of 2009 (ARRA) brings broad range of provisions that range from amendments for HIPAA's privacy and security rules to construction, into law. One of the provisions, Health Information Technology for Economic and Clinical Health Act (HITECH) provides financial incentives to encourage the adoption and meaningful use of certified EHR technology (CEHRT) [7].
- Australia government's national health information network proposed in 2004, named *HealthConnect*, extracts summary records from locally collected patient data and then aggregated to create centralized *HealthConnect* record that may then be shared among participating authorized providers [8].
- The European Union directive 2011/24/EU enacts patients' rights in cross-border healthcare suggesting to boost the digital economy by enabling all Europeans to have access to online records anywhere in Europe by 2020 [9].

Meaningful Use (MU)

Among the drivers of EHR, incentives for *Meaningful use of certified EHR technology* (or simplify *meaningful use*) by HITECH became quite popular in the US region. Any healthcare provider implementing an EHR system that is capable of demonstrating the below listed core EHR functionality gets qualified as using the certified EHR technology and becomes eligible for MU incentives:

- Patient demographic and clinical health information
- Electronic order entry, including e-prescribing
- Health information exchange (HIE)
- Clinical decision support (CDS)
- Data collection and query to support quality initiatives

HITECH envisioned a three-stage process for *meaningful use* with the below overarching goals:

- Improve quality, safety and efficiency
- Engage patients and their families
- Improve care coordination
- Improve population and public health
- Ensure privacy and security protections

Each stage has a set of objectives developed by the Dept. of Health and Human Services (HHS) and published in the Federal Register [7].

HITECH proposes meaningful use of interoperable electronic health records throughout the United States health care delivery system as a critical national goal. In fact, interoperability of personal health records is one of the core expected functionalities of *meaningful use* in stage2, tracked as Health Information Exchange.

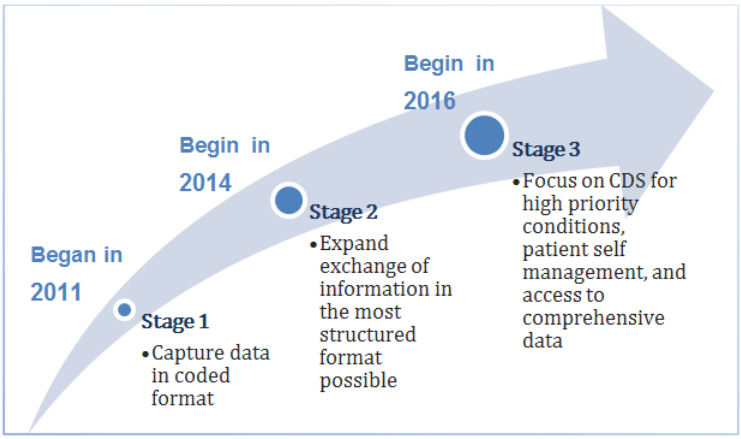


Figure .2: Meaningful use stages

Health Information Exchange (HIE)

The demand for electronic health information exchange from one healthcare professional to another is growing alongside the nationwide efforts to improve quality, safety and efficiency of healthcare delivery. Meaningful use requirements, new payment approaches that stress care coordination, and federal financial incentives all are driving the interest and demand for health information exchange.

HIE allows healthcare professionals and patients to appropriately access and securely share a patient's vital medical information electronically. Sharing updated, electronic patient information with other providers enables one to:

- Access and confidentially share patients' vital medical history, no matter where the patients are receiving the care – specialists' offices, labs or emergency rooms
- Provide safer, more effective care tailored to the patients' unique medical needs

However, a big challenge in effectively using health information exchange in mainstream is, a large proportion of the clinical data that is contained within clinical documents, such as radiology reports, surgical pathology reports, operative reports, discharge summaries and clinic notes, does not use consistent format. Much of this document-based data is represented as unstructured narrative text with little, if any, standardization of the language used to represent important information such as diagnosis, therapy or test results. The data contained within these documents is therefore difficult to integrate into clinical or research databases that need to support efficient standards-based retrieval.

Data normalization

Clinical data comes in all different forms even for the same piece of information. For example, in a medical record, age could be reported as 40 years for an adult, 18 months for a toddler or 3 days for an infant. In such notation, these records are not directly comparable with each other without normalization. Database normalization of clinical data fields in general fosters a design that allows for efficient storage avoiding duplication or repetition of data; data querying becomes easier. Without normalization, data cannot be exchanged meaningfully.

Un-normalized	Normalized (days)	Normalized (months)
40 years	14360	478
18 months	543	18
3 days	3	0.1

Detailed clinical models [10] are the basis for retaining meaning when data is exchanged between heterogeneous systems. Clinical models are also the basis for shared computable meaning when clinical data is referenced in decision support logic. There are multiple healthcare delivery scenarios driving the technology behind the different forms of HIE available today. Some of the models and business approaches emerging to support electronic health information exchange are:

- Development of regional, local or state non-profit or government-sponsored exchange networks
- Local models advanced by newly formed accountable care organizations
- Exchange options offered by electronic health record vendors
- Services provided by national exchange networks [11]

HIE provides the capability to electronically move clinical information among disparate health care information systems while maintaining the meaning of information being exchanged. HIE systems facilitate the efforts of physicians and clinicians to meet high standards of patient care through electronic participation in patients' continuity of care with multiple providers. Research has demonstrated that where access to medical specialties is scarce, telehealth technologies can improve the access to specialty care in underserved urban and rural areas and among underserved populations who are institutionalized, such as inmates and nursing home residents [12].

Apart from providing the basic level of interoperability among EHRs maintained by individual physicians and organizations, HIE benefits include:

- ✓ Increases efficiency by eliminating paper work
- ✓ Eliminates redundant or unnecessary testing
- ✓ Improves public health reporting and monitoring
- ✓ Creates a potential loop for feedback between clinical research and actual practice
- ✓ Provides a vehicle for improving quality and safety of patient care by reducing medication and medical errors
- ✓ Stimulates consumer education and patients' involvement in their own health care
- ✓ Provides caregivers with clinical decision support tools for more effective care treatment
- ✓ Facilitates efficient deployment of emerging technology and healthcare services
- ✓ Provides backbone of technical infrastructure for leverage by nation and state-level initiatives

Despite these potential benefits, there is relatively slow rate of adoption for HIE technologies in the overall healthcare field, primarily due to the high capital and maintenance costs involved in implementing HIE solutions, lack of available staff with adequate expertise in Health IT, security and confidentiality concerns about patient data.

The first two challenges, namely the high capital costs and IT skilled staff requirement, can be somewhat mitigated by the use of cloud-based SaaS EHR deployments, as opposed to localized or client-server deployment within premises. As care providers move to EHR, web-based SaaS systems offer attractive price alternatives, particularly for physician practices in small offices. But, even for many of such small office practitioners these cost advantages are not enough to overcome the remaining major concern about the security of their patients' personal health data when stored off-site.

It is exactly to address these concerns, the security and privacy regulations such as HIPAA, are being put in place by federal governments enforcing strict compliance.

Health Information Privacy and Security

The need for privacy and confidentiality is at the forefront of the health informatics movement. Ensuring privacy and security of health information, including information in electronic health records, is a key component for building the trust required to realize the full benefits of health information exchange (HIE). If individuals and other participants in a network were to lack trust in electronic exchange of information due to perceived or actual risks of theft or information accuracy, it may as well result in affecting their willingness to disclose even the necessary health information required for diagnoses. To prevent this, regulations were enacted by federal governments to help protect patient information through:

- ✓ Access controls to make sure only those who are authorized can access health information
- ✓ Audit functions that track who has accessed what pieces of information
- ✓ Internet-based portals that allow patients to access their own health records, see who else has viewed their records and check the accuracy of records

Further, stringent requirements were put in place to notify data breaches and minimize the consequences of information misuse of patient health records. For example, the HITECH act requires HIPAA covered entities to report data breaches affecting 500 or more individuals to HHS and the media, in addition to notifying the affected individuals [13]. European data protections Directive 1995, New Zealand's Privacy Act 2000, Canada's Personal Health Information Protect Act 2004, etc. similar regulations in other countries have similar enforcements in place.

Regulations such as these are legal, standard and official rules for healthcare privacy and security. They require security services that support implementation features including access control, audit controls, authorization control, data authentication and entity authentication. For example, audit controls of HIPAA aim to assess compliance with a secure domain's policies, to detect any instances of non-complaint behavior, and to facilitate the detection of improper creation, access, modification and deletion of Protected Health Information.

The strength of these regulations derives from their mandatory and punitive nature. Hence they present an exceptional significant advance in securing e-Healthcare information. Also, numerous other industry security standards have been established in addition that assure patients' healthcare information privacy and security. The Health Level 7 (HL7) standard, for example, that is based on ISO-OSI network model focuses on e-Healthcare standards at the 7^{th} layer of the OSI model. The key enabler for the HL7 is the Clinical Document Architecture (CDA), which is a document architecture standard for representing medical and legal healthcare encounter documents in a standardized format. The HL7 CDA security takes a *transactional security* approach. This approach is similar to the internet-based SSL when applied to a document that is sent from a webserver to a client or vice versa.

In HL7 CDA security, e-Healthcare information in standard format has a known healthcare sender and known healthcare recipient. The sender and recipient would have agreed to the security mechanism needed for that exchange. The goal of HL7 CDA security, then, is to secure the resulting transaction.

In the area of Role Based Access Control (RBAC), the HL7 security technical committee has formalized a set of permissions. In this model, the elements of the permissions contain only one object and at least one operation. The permission set was then come up with scenario-based engineering process models.

> "Good patient care means safe record-keeping practices. Do not forget that an EHR represents a unique and valuable human being: it is not just a collection of data that you are guarding – it's a life !!"
>
> -Office of National Coordinator for Health Information Technology

For implementing confidentiality of patient information, the technological platforms are required to provide functions that limit the right to view or transfer selected data to users with specific kinds of authorization and auditing access. The HL7 CDA standard, for example, contains the necessary data objects, attributes and transaction contents to support conveying the necessary information from one healthcare application system to another, enabling the systems perform confidentiality functions.

With all these standards and federal regulations in place, the EHR information is heavily protected from unauthenticated use and misplacement. In today's world, it can be said that these security rules do not just play the mere role of protecting EHR data, but also are to be perceived as a critical integral component of EHR itself for enabling its massive adoption in large scale on industry wide.

EHR use-case scenarios

Electronic health records are paving way for many advanced patient health care information exchange and query use-case scenarios that would not have been possible without it. Few of them are as listed below.

Transitioning to a new provider of care: An emergency department physician who is attending a patient on vacation in a remote place, can access any previous medical records of the patient (for example any allergies the patient might have for a particular drug) and treat him. After the emergency treatment completion, the emergency care provider can submit back continuity of care document to the primary care physician, so that the patient when he goes back to his location, can seamlessly resume the treatment.

Sending a patient follow-up care instructions: Rehabilitation patients, while transiting from emergency care providers or specialized care providers to their primary care provider, can access their follow-up instructions electronically in real-time from their previous provider without having to physically be present

Delivering lab results: When a patient visits a reference lab for blood work prior to an annual health checkup with his primary care physician, the lab technician compiles the results and enters them directly into the EHR system, which is then electronically delivered to the primary care physician. Patient does not need to wait for the lab technician – instead he can directly set an appointment with his primary care provider once he is made aware that the lab results have reached the physician.

Sending reminders to patients: A patient's primary care physician can recommend that all his patients over the age 65 get annual influenza vaccines in time. As a convenience, the primary care physician's EHR system automatically sends the patient an electronic message reminder to make the vaccination appointment.

Submitting immunization data to public health: Patients take their toddlers to their primary care provider to be immunized in accordance with Centers for Disease Control and Prevention (CDC) recommendations. Upon completion of vaccination, primary care physician updates the patients' EHR and accesses the state's immunization information system (IIS) interface from within the EHR system and sends the immunization records to IIS.

Reporting notifiable diseases and conditions to public health agency: When a patient is diagnosed with any infections that are highly contagious, Centers for Disease Control and Prevention (CDC) requires all such cases to be reported to a public health registry in order to track and prevent nationwide epidemics. Physicians will be using the EHR system to report any such incidents, as per regulations.

Clinical Decision Support Systems

Of all the modern technological quests, the search for a true Artificial Intelligence (AI) computer system has been one of the most ambitious and, not surprisingly, controversial endeavors undertook by researchers. With intelligent computers able to store and process vast stores of knowledge, the hope is that they would become perfect *expert systems* capable of assisting or surpassing human beings in the tasks of their own expertise. For example, intelligent systems today are found supporting medication prescribing, in clinical laboratories and educational settings, for clinical surveillance, or in data-rich areas like the intensive care monitoring. When a patient's case is complex or rare or the person making a diagnosis of the case is simply inexperienced, an *expert system* can help in the formulation of likely diagnoses based on the patient data presented to it and its understanding of the illness as stored in its knowledge base. Such an expert AI system that is trained to be capable of assisting human clinicians in their decision making process is referred to as clinical decision support system (CDSS).

Clinical decision support systems' functionality essentially revolves around two concepts: 1. acquiring domain knowledge and 2. Making inferences based on the acquired knowledge.

The first step of acquiring domain knowledge can happen either through ontological structures embedded inside the systems programmatically during their design time, or through epistemological models of reasoning derived from the data supplied to them during their runtime or a combination of both.

Inference on the other hand has multiple approaches that are continually evolving even as we are speaking, ranging from context-based reasoning, to causal reasoning to task-based reasoning etc. However, these inference methods are tightly coupled with the particular scheme of underlying knowledge representation mechanism used by the system. Following sections discusses these two in more detail.

Knowledge acquisition & organization

Knowledge acquisition has traditionally been considered as the process of extracting and transcribing the expert knowledge into a computer-usable form. However, current trends are to consider knowledge acquisition as an integral part of the designing process, centered on the notion of a model. Domain knowledge essentially is an ensemble of names, properties, relations, facts, observations, hypotheses and results that extracted from different cases. It can be broadly categorized as constituting two parts: 1. Ontological and 2. Epistemological;

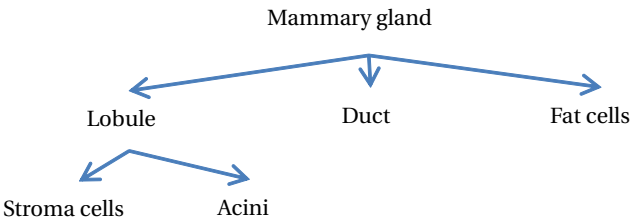


Figure: "Part of" Hierarchy example

Ontological knowledge refers to the body of the knowledge that does not vary on case-by-case basis and is independent of the observation point. Epistemological knowledge on the other hand is purely derived from observations, and as such varies from case to case. It can be understood in simple terms as, ontology being the abstraction, where epistemology is its instantiation.

Taxonomy ontologies have been widely used in medicine due to their ability to classify and organize diseases into hierarchical structures. *Part of* hierarchies make it possible to describe the compositional characteristics of concepts, and are particularly useful for describing anatomical properties. However, current research in emphasizing the notion of causal ontologies where a deep understanding of the process leading to a disease is sought.

An epistemological model of diagnostic reasoning on the other hand resolves around the concepts of abstraction, abduction, deduction and eliminative induction. This approach uses the process of handling facts, observations and hypotheses to reach a final understanding of a situation.

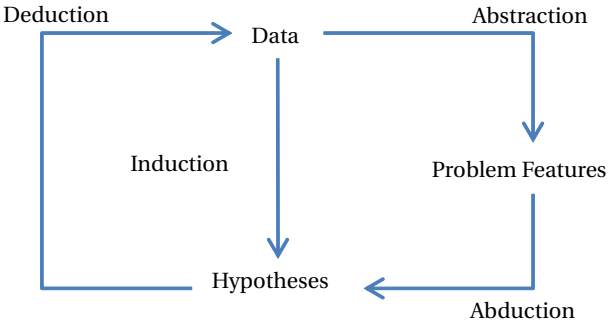


Figure: Epistemological model of diagnostic reasoning

Knowledge discovery, data mining, and machine learning techniques have recently attracted considerable attention, due to the growing amount of data available, and to the growing necessity to base the reasoning on evidence taken from physical measurements. Data-driven approaches to knowledge extraction have been developed as a consequence, complementing the more traditional human-centered approaches, by enabling systems to create new knowledge, update existing knowledge, and improve their performance without intervention and reprogramming.

The organization (or representation) of medical knowledge is a very active research field, characterized by a wide range of tools, models and languages, which, together with the availability of increasing computer abilities, allows one to specify and emulate systems of growing complexity. Frame representations, semantic networks, conceptual graph representations are few popular basic representational schemes heavily used in the industry today.

A *frame* defines the prototypical description of concepts that share similar properties and behavior. It is defined as a set of slots describing the concept attributes and their values – similar to object oriented classes in programming languages. The frames are organized in hierarchies, tying classes and instances. The basic inference mechanism in this model is, *instantiation*, in which the attribute values of the new instance are obtained either by inheritance, by computation or by default.

For example, in a knowledge-based system to diagnose HIV-pneumonias, frames are used for describing the physiopathological states of the disease for each particular pneumonia, in terms of several slots describing the epidemiological data, clinical picture, laboratory tests and diagnoses [14].

Semantic networks are a flexible formalism to represent the semantics of concepts (as nodes in a network) and their relations (the arcs) in the framework of graph-like structure. The reasoning proceeds by unification between an unknown fact and a known concept or sub-graph in the network. The reasoning strategy in these systems is of event-driven type: initial data triggers a number of hypotheses, which are then to be confirmed. The sub-graph of confirmed or undetermined hypotheses constitutes a patient-specific model. Various weights and scores are usually introduced to render the reasoning strategy more flexible.

Conceptual graphs were designed to be an extension of the afore-mentioned semantic networks formalism, in which an explicit representation of the links between concepts is sought. A conceptual graph is a directed graph comprising two kinds of nodes: concept nodes and conceptual relationship nodes. The representation is grounded on first-order logic and may also be used as a formalism that supports both expressiveness and classification purposes.

Reasoning

Reasoning in decision support systems happens through inference engines (computer programs built based on logic: such as propositional logic, predicate logic, modal logic, temporal logic, epistemic logic, fuzzy logic etc.). Typically these engines run in two modes: batch or conversational. In batch mode, the system has all the necessary data to process from the beginning. User provides the data upfront and gets back the results immediately, the reasoning being invisible. The conversational method, on the other hand becomes necessary when the user cannot supply all the necessary data up front. The system should identify the missing parts of the data and invent ways of collecting that data through probing questions. To guide this probing, the engine may use several levels of sophistication such as: forward chaining, backward chaining and mixed chaining.

Based on the level of sophistication an inference engine uses, wide variety of reasoning methodologies have been implemented in the decision support systems, some of which are presented in the following.

Rule-based reasoning

The use of *if-then* rules (also known as production rules or condition-action rules) is popular and straight-forward way to represent the expert know-how. This type of knowledge-modeling is often said to be shallow or heuristic, since it is largely empirical and non-formalized representation mechanism.

Given a set of facts, the reasoning process is then modeled as the successive activation of rules, which in turn produce new facts to be considered, until no applicable rule is found. Each applicable rule will have certain pre-conditions to

be satisfied (stated in the form of if-clauses). For a system in any given condition, only certain set of rules whose pre-conditions match the current state of the system can be applied. Each application of the rule, then potentially changes the state of the system, enabling other rules whose pre-conditions match the new state become applicable.

Rules are handled by means of inference engine, which may work under two inference modes – data driven or goal driven. In the data-driven mode, the reasoning is modeled as a deductive process proceeding from a given premise to some conclusion. In the goal-driven mode, on the contrary, the reasoning is modeled as an inductive process proceeding backwards from hypothesized conclusion to the conditions to be verified. Both inference schemes are usually combined to implement complex hypotheses and test strategies. In addition, Meta rules are often used to structure the reasoning process by constraining the application of rules.

MYCIN, a system for diagnosing and treating infectious blood diseases, is one of the earliest and most widely known expert system in the medicine built on rule-based reasoning [15].

Uncertain and imprecise reasoning

Since human body is not fully understood and it is impossible to measure all parameters, health records are characterized by uncertainty, imprecision and incompleteness. Observations are uncertain by nature, since they depend on the context of the observation. Usually observations are registered and treated as facts. Likewise most opinions about a patient must be accompanied with some kind of uncertainty. In health records, uncertainty is stated when it comes to hypothesis, whereas conclusions are registered as certain. In opinions, uncertainty is often traded for imprecision. For example, the opinion "it might be a snake bite from a rattle snake", can be converted into "it's almost certainly a snake bite", which has higher certainty but lower precision.

Reasoning about uncertainty has as its goal determining how much is known about something. It can be performed in a context-free manner. The result of reasoning about uncertainty is a characterization of the uncertainty, for example a probability distribution over all possible outcomes of a random variable.

In *reasoning under uncertainty*, characterization of uncertainty is not a goal in itself, but is rather a means to achieve some other goal.

Bayesian model has been the primary numerical approach to reasoning under uncertainty for a long time. It is based on the principle of assigning a probability distribution to each of the variables representing the problem at hand. These probabilities express the uncertainty and likelihood of occurrence of symptoms as well as diseases. Given $P(D_i)$, the a-priori probability of disease D_i, and given $P(S|D_i)$, the probability that symptom S may occur in the context of disease D_i, Bayes' rule allows one to compute $P(D_i|S)$, the probability that disease D_i occurs when S is observed, according to the following formula:

$$P(D_i|S) = \frac{P(S|D_i) * P(D_i)}{\sum_j P(S|D_j) * P(D_j)} \tag{.1}$$

The most important assumption underlying the use of Bayes' rule is that all disease hypotheses must be mutually exclusive and exhaustive. It means that only one disease is assumed to be present. If this condition is met, Bayes has a completely predictable performance and guarantees a conclusion with minimum overall error. Some other theories, such as the theory of possibility and fuzzy reasoning are also some of the other commonly used techniques under the uncertainty reasoning apart from Bayes.

Case-based reasoning

Case-based reasoning is based on the hypothesis that new situations are analyzed by reference to past experience, i.e. by considering their similarity to well-experimented situations, a notion that is made central in this kind of reasoning. New problem-solution pairs may then be learned as a consequence: this inherent combination of problem solving with learning through problem solving experience gives a particular strength to case-based reasoning over most other methods.

Case-based reasoning essentially is comprised of four stages, retrieve the most similar case, reuse the retrieved information to solve the problem by analogical reasoning, revise the proposed solution, and finally retain the part of the experience likely to be useful for future problem solving.

The advantage of this approach is two-fold: First of all, it is based on the description of experiences, rather than on modeling of generic or abstract knowledge; second, such a system may evolve in an incremental way, as cases grow. However, the choice of similarity measure turns out to be critical with respect to the reasoning accuracy. The structurization of the case base also has to be carefully considered in order to avoid combinatorial search.

To be able to retrieve similar cases, a similarity measure for both observations and reach goals need to be defined. This measure must be based on semantic distances. A semantic similarity between one patient case and another indicates the degree of similarity between the two. An approach would be to describe distance measures between classifications

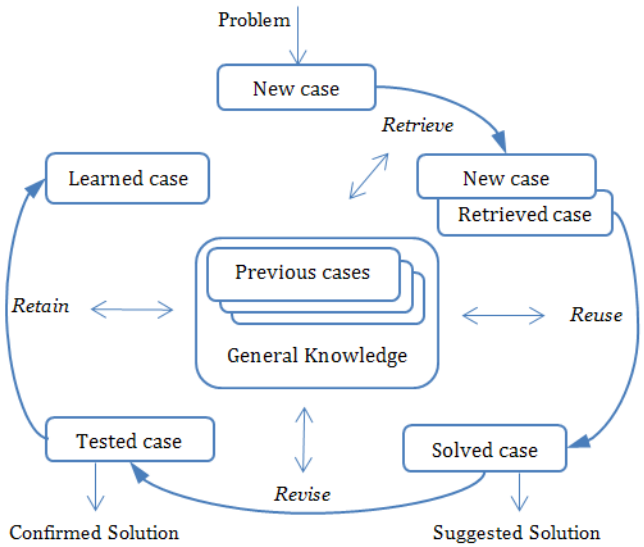


Figure .1: Overview of case-based reasoning model

within each dimension, by using functions, hierarchies or graphs. This, approach, however, has the problem that semantic similarity has to be estimated based on general defined semantic distances of many independent dimensions. Another approach for semantic similarity is to base it on the reasoning context, defining it as the neighborhood of the patient case, either in reached task goals or task dependent observations, where another case is considered to be similar. This idea of establishing a reasoning context leads to the context-based reasoning methodology described in the next section.

Context-based reasoning

It is now well recognized that advanced medical decisions systems should situate their reasoning with respect to the context in which a problem is considered, including the patient clinical context, and also the chronology of events together with their causal relationships. Context-based reasoning is an intuitive and effective means to model tactical behavior in either simulated or real-life scenarios. The concept of context can provide a model to partition the operation of a complex system into *scenarios*, where knowledge, strategies, parameters and objectives are organized. Awareness of the context permits to define which knowledge should be considered, what are the conditions of activation and limits of validity and when to use it at any given time.

Two subtasks of diagnosis where a clinician extensively uses context elements are, generating hypotheses about the cause of failure, and planning information gathering actions. Hypothesis generation is often achieved by retrieving past episodes where similar features have been experienced. The other task of information gathering utilizes other types of context elements than when generating hypotheses.

A context is a set of environmental and physical conditions that may suggest a specific behavior or action. Within a context-based reasoning model, however, a context is a functional state induced as a result of these conditions. The main challenge in designing a context-based reasoning system is: given the raw data, how to model the context? One approach is to adopt fuzzy logic model to represent the relevant variables and to build low level and high level context models. For example, the low level context can be structured according to the physiological context, personal context and environmental context and generate high level context consisting of activity event and medical condition. These contexts will successfully model a ubiquitous context-aware environment.

Contexts are inserted within a model to represent all possible conditions that can arise during the course of the transitions within the model. This ensures that a model can exhibit intelligent behavior no matter what occurs during the execution.

Context based reasoning models are constructed such that a single context is active at any one point during a scenario. The knowledge engineer responsible for creating the model is in charge of defining and creating each context. Because of this, contexts themselves are often intuitive subsets of the behavior to be modeled. When encoding the knowledge for these contexts, the idea is to achieve a model that can take same actions that an expert might take when put in the same situation.

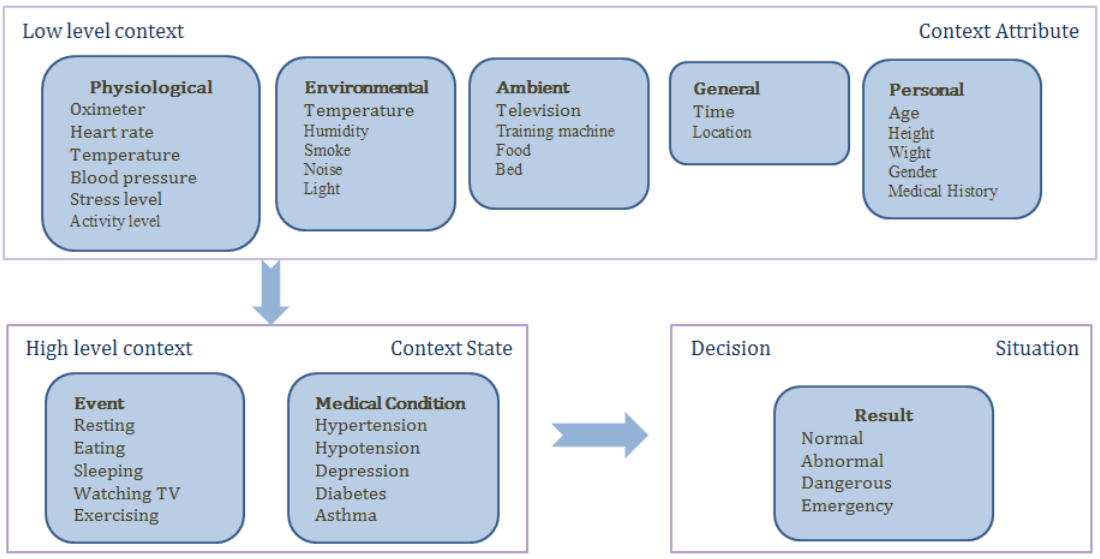


Figure .2: Overview of context-based reasoning model

Causal reasoning

Developments in the solid foundations of diagnostic systems employing logic, probability theory, and set theory in the recent decades has provided richer and robust modeling tools to the developer, allowing one to cope with the increasing necessity of basing diagnostics on the models of disease process and structure and functions of human body. There are at least two alternative ways to represent causal knowledge. First method is based on a network representation, and the second implies the representation of both the structure and behavior of a physiological system.

Network representations, called *causal probabilistic networks* or *Bayesian belief networks*, are in the form of a direct acyclic graph in which nodes represent stochastic variables and arcs represent conditional dependency among the variables. These representations make explicit the dependency and independency assumptions among variables. They are based on sound semantics and easily extended to compact representations called *influence diagrams*.

MUNIN [16], an early expert system in electromyography, is a representative of this approach. The domain knowledge is embedded in a causal probabilistic network and is further divided into three levels, representing diseases, pathophysiological features, and findings. These levels are linked by causal relations: diseases cause certain affections in muscles. These affections, in turn, cause expectations for certain findings. Some expectations regarding, for example, the muscle force and atrophy or the presence of spontaneous myotonic distrophy are finally obtained. Probabilities are used to characterize each state in the network and propagate through causal links.

Despite its effective representation power, the approach still bears strong limitations, due to the large amount of information required: current propagation algorithms require that all the conditional probabilities defining a conditional dependency be known, together with all the a priori probabilities attached to the root. MUNIN, for example, consists of about 1100 discrete random variables linked together, posing memory and calculation time challenges.

An alternative way is to represent both the structure and behavior of physiological system. The system structure is simply given in terms of variables and relations. The system behavior then is modeled by a set of mathematical equations tying related variables, which represent potential perturbations of the system state. Compartmental systems theory is often used in this respect because it provides a robust modeling framework based on differential calculus. It implies, however, that a precise quantitative modeling of the pathophysiological phenomena is possible. Qualitative models, on the other hand allow one to cope with the fuzziness and incompleteness of pathophysiological knowledge. Among these QSIM is the most widely applied formalism in medicine [17]. This formalism provides a descriptive language to represent the structure of physiologic system and simulation algorithm to infer its qualitative behavior. The language consists in qualitative constraints that abstract the relationships in a different equation.

It should be noted that representing pathophysiological knowledge requires a special emphasis on the notion of time. Time constitutes an integral and important aspect of medical concepts, and its explicit modeling is increasingly considered as central to the design of advanced medical systems. However, such modeling still remains challenging, due to the necessity to consider compound objects (e.g., disorders, treatments, and patient states) exhibiting different temporal existences and complex interactions, through mechanisms that are not completely understood. The definition of adapted temporal ontologies conveying a clear semantic is currently an important subject of debate among the community.

Use-case scenarios

There are numerous reasons why clinical decision support systems (CDSSs) have not been in prime use in the earlier years in the healthcare industry – the primary of them being that they require the existence of electronic patient record system to supply their data, and till now most institutions and practices do not have all their working data available in electronic form. The other reasons being, suffering from poor human interface designs, lack of standards, non-interoperability etc. – all of which are currently being addressed through EHR. Progressively more and more EHR systems are being built around the usability aspects of clinician users, with integrated clinical decision support capability, conforming to norms and HIE interoperability features, leading way to CDSS mainstream availability. CDSSs offer a number of important benefits, including:

- ✓ Increased quality of care and enhanced health outcomes
- ✓ Avoidance of errors and adverse events
- ✓ Improved efficiency, cost-benefit, and provider and patient satisfaction

There are many clinical tasks and scenarios where these systems can be applied:

Alerts and reminders: In real-time situations, expert system attached to a patient monitoring device like ECG or pulse oximeter can warn of changes in patient condition. Reminder systems notify clinicians about important tasks that need attention. For example, an out-patient clinic reminder system may generate list of immunizations that each patient requires on the daily schedule.

Diagnostic assistance: Diagnostic assistance is often needed with complex data, such as ECG, where most clinicians can make straightforward diagnoses, but miss rare presentations of common illness like myocardial infarction, or may struggle in formulating diagnoses that require specialized expertise. In such cases clinical decision support systems can help with diagnoses.

Therapy critiquing and planning: Critiquing systems can look for inconsistencies, errors and omissions in existing treatment plans. For example, on entering an order for blood transfusion a clinician may receive a message stating that the patient's hemoglobin level is above the transfusion threshold, and the clinician may justify the order by stating an indication, say active bleeding.

Prescription decision support: Assisting in medication prescription by checking for drug-drug interactions, dosage errors, and contraindications such as allergy. Such systems have also been used to suggest cheaper generic drug alternatives when a more expensive drug was initially ordered, resulting in cost-effective drug selection.

Image recognition and interpretation: Many clinical images can now be automatically interpreted, from plain X-rays to more complex images like angiogram, CT and MRI scans. This is of value in mass-screenings, for example, where the system can flag potentially abnormal images for detailed human attention.

Text mining medical records

Increased use of electronic health records is leading way to the widespread adoption of data-mining efforts for uncovering hidden trends in health, disease, and treatment response data. However, since a significant portion of EHR content is stored as narrative text, unsuitable for the conventional data mining techniques, demand for special text mining procedures based on Natural Language Processing techniques is growing.

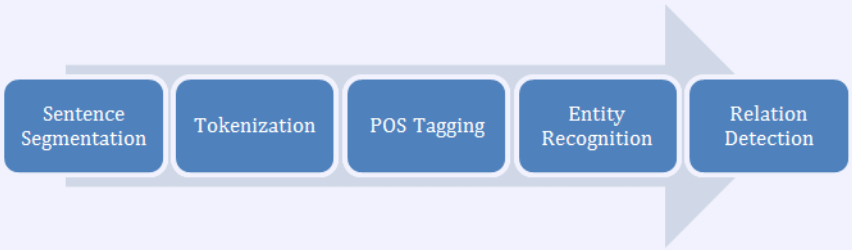
Natural language processing, or NLP in short, is a research field dedicated to empowering the computers with the right knowledge for understanding natural language text and facilitate different types of language interactions between humans and machines.

For a clinical decision support system to be compatible with *meaningful use* stage 3 or to fully utilize the EHR capabilities, it should possess text analysis functionality that is capable of extracting text data out of EHRs and mine the text documents for new insights into the patient health behavior. Such in-depth analysis requires Natural Language Processing capabilities.

A large proportion of the clinical data that is captured in EHRs today, such as radiology reports, surgical pathology reports, operative reports, discharge summaries and clinic notes etc., is document-based data that is mostly represented as unstructured narrative text. Consider discharge summaries, for example. Discharge summaries contain much useful information about the patient, useful not only for subsequent visits but also for a variety of other tasks - documented as text. To turn a collection of discharge summaries into an effective tool for decision support or public health research, it

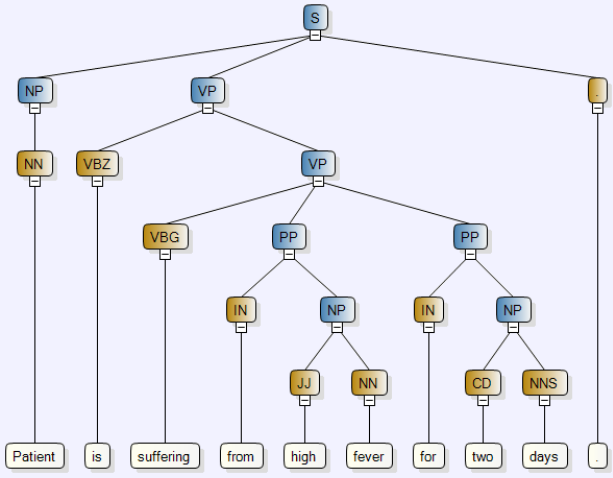
Natural Language Processing (NLP) workflow

Decision support systems search large bodies of unstructured text for specific types of entities and relations, and use them to populate well-organized databases. These databases can then be used to find answers for specific questions. The typical architecture for such a system begins by segmenting, tokenizing, and part-of-speech (POS) tagging the text. The resulting data is then searched for specific types of entity.



	Input	Output
Sentence Segmentation	Raw Text	List of strings
Tokenization	List of strings	List of list of strings
POS Tagging	List of list of strings	List of list of tuples
Entity detection	List of list of tuples	List of trees
Relation detection	List of trees	List of tuples

Shown below is an example sentence, "Patient is suffering from high fever for two days.", in its part-of-speech tagged tree form. Note how the adjective *high* followed by the noun *fever* got combined into a noun-phrase (NP).



CD	Cardinal number
IN	Preposition/subord. conjunction
JJ	Adjective
NN	Noun singular
NP	Noun phrase
NNS	Noun, plural
PP	Proposition phrase
S	Sentence
VBG	Verb, present participle
VBZ	Verb, 3rd person singular present
VP	Verb phrase

is necessary to extract the disease and procedure information from the narrative text, index it for query capability and generate numerical measures out of it for analytics.

However, this is not an easy task and there are numerous challenges associated with it. Unlike structure data, where the data model is clearly present for the automation, clinical records are highly unstructured in their organization. For example, diseases are found in different sections with a variety of names, separated in several ways in a discharge summary document. A disease statement may have additional descriptive text that is not expected, or the diseases themselves may be more specific than is codable in a single code or there may be no code at all for a particular disease being described.

While it is natural and easy for a human being to understand and act upon text, for a machine, however, natural language is not a natural choice of communication and hard to act upon. It needs to be told in explicit instructions (algorithm) how to derive meaning out of text that is unstructured. Most of these algorithms depend on pre-processing phases such as sentence splitting, word sense disambiguation etc. to extract useful information out of the supplied text. In addition, contextual features like negation, temporality, and subject identification are also crucial for accurate interpretation of the extracted information.

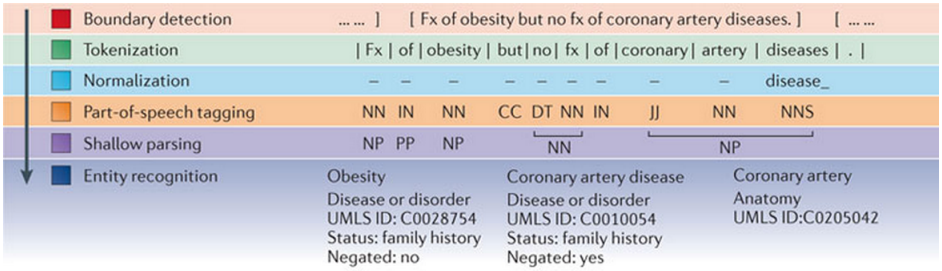


Figure .3: Text analysis flow on medical records (source: Nature Reviews)

Supplied with a text document that contains a list of sentences, text mining using the NLP starts with identifying sentence boundaries first. A sentence segmentation mechanism splits the input text into units of individual sentences. These sentences are then forwarded for tokenization, a mechanism that splits each sentence into a group of individual tokens (typically individual words), using space and other punctuation markers as a guide (with rules for handling special cases such as dates). The result of tokenization is a set of tokens that are reduced to a base form by normalizing, for example, case, inflection or spelling variants.

The next step assigns part-of-speech tags to each token to identify its grammatical category in the context (for example, *NN* for noun, *IN* for preposition or *JJ* for adjective). This is not a trivial task as many words have ambiguous meaning.

After the tokens have been tagged, the shallow parsing step identifies syntactic units, most commonly noun phrases (NPs), which are grammatical units built from a noun with optional modifiers such as adjectives. An example of this noun phrase groupings is presented in the NLP workflow sidebar. In that example it can be observed that the noun *fever*, following the adjective *high*, got grouped together to form a noun-phrase *high-fever*.

In the next step of entity recognition, NPs and various lexical permutations are mapped to controlled vocabularies using tools such as MetaMap [18]. Importantly, such systems also identify the presence of negating terms, such as 'no' or 'never', near identified entities. Named entities (NE) are definite noun phrases that refer to specific types of entities, such as organizations, persons, diseases, medications and so on. The goal of *named entity recognition* system is to identify all textual mentions of the named entities. This can be broken down into two subtasks: identifying the boundaries of the NE, and identifying its type. Named entity recognition is a crucial task for implementing Question-Answering functionality in a decision support system.

Once named entities have been identified in a text, it is then required to extract relations between them. For example, in a sentence like "patient is suffering from high-fever" the two noun phrases *patient* and *high-fever* are related to each other as one *causing suffering* to other.

All these various steps are typically implemented using combinations of logical rules (and their exceptions) and machine learning methods. For example, a full stop (period) followed by a space and a capital letter indicates a sentence boundary. Figure .3 demonstrates these rules as applied to a medical record, identifying two disorders and one anatomical structure. Both disorders are tagged as relating to family history (Fx), and note how in the case of coronary artery disease, the preceding word 'no' tagged the term as *negated*.

Use-case scenarios

Text mining medical records can provide several potential benefits. By combining structured and unstructured data together, one could get richer patient health information and build intelligent systems hitherto non-existent. In addition, text mining can help differentiate patients, enabling greater patient stratification and improved targeting of medicines, which could in turn have profound implications for the design of clinical trials. Few use-case scenarios popular in the industry for the clinical text mining are as follows.

Medical outbreak Alerts In many cases health professionals require alerts of any unusual health events based on aggregated data from multiple records. For instance, text from emergency room records are collected in real-time by a central computer, which monitors them to track the reasons for first-aid visits. If a pattern emerges and the frequency of these patterned events exceeds the expected level, a meaningful alert is issued for a suspected medical outbreak. This allows the receiving public health officials to respond immediately. An adequate response is likely to require additional information from a variety of data sources, many of which are, again, expected to be in text format.

> “By extracting data from clinicians' notes and combining the results with protein and genetic information, Danish scientist Francisco Roque and his colleagues at Technical University of Denmark discovered hidden linkages between health problems that were believed to be unrelated, such as migraines and hair loss, or glaucoma and a hunching back . . . [19]”

Disease correlations Text mining aids the discovery of unknown disease correlations and the identification of previously unknown drug side effects. For example, by extracting data from clinicians' notes and combining the results with protein and genetic information, Danish scientist Francisco Roque and his colleagues at Technical University of Denmark discovered hidden linkages between health problems that were believed to be unrelated, such as migraines and hair loss, or glaucoma and a hunching back [19]. Similarly, discovery of unknown disease correlations was reported by William Knaus and his colleagues at the University of Virginia Health System, who detected a strong correlation between a peptic ulcer disease and renal failure using a combination of text- and data-mining techniques. And researchers at three major universities—Stanford, Vanderbilt, and Harvard—used data and text mining of EHRs to discover a dangerous side effect from a combination of drugs that caused a significant spike in patients' blood glucose levels. These kinds of discoveries were made possible with text analysis of clinical records.

Question-answering systems Question-Answering (QA) systems are the next generation of search engines. They combine traditional information retrieval with natural language processing and knowledge engineering techniques to provide shorter, more precise and accurate answers to natural language questions.

The scenario involves scanning journal articles and other knowledge sources that contain clinical case studies, randomized controlled trails, breakthrough procedures etc. and extracting relevant information applicable to a patient record to stratify cohorts, and answer questions related to clinical decision support.

Often referred to as Evidence-based medicine (EBM), this is a new paradigm for medical practice that involves explicit use of current best evidence, that is, high-quality patient-centered clinical research such as reports from randomized controlled trials, in making decisions about patient care. Naturally, such evidence, as reported in the primary medical literature, must be suitably integrated with the physician's own expertise and patient-specific factors.

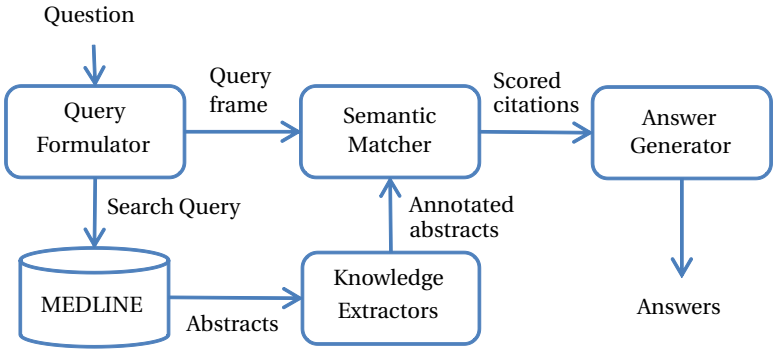


Figure: Architecture of clinical Question-Answering system

EBM offers facets that provide a framework for codifying the knowledge involved in answering clinical questions. For example, below four components, often referenced with mnemonic PICO, which stands for Patient/Problem, Intervention, Comparison, and Outcome - were identified as the key elements of a question related to patient care:

- What is the primary problem or disease? What are the characteristics of the patient (e.g. age, gender, or co-existing conditions)?
- What is the main intervention (e.g. a diagnostic test, medication, or therapeutic procedure)?
- What is the main intervention compared to (e.g. no intervention, another drug, another therapeutic procedure, or a placebo)?
- What is the desired effect of the intervention (e.g. cure a disease, relieve or eliminate symptoms, reduce size effects, or lower cost)?

Physicians are usually most interested in outcome statements that assert a patient-oriented clinical finding—for example, the relative efficacy of two drugs. Thus, outcomes serve as the basis for good answers and an entry point into the full text.

A question-answering system designed to support the practice of evidence-based medicine must be sensitive to the multifaceted considerations that go into evaluating an abstract's relevance to a clinical information need. The system might automatically evaluate the strength of evidence of the citations supplying the answer, but the decision to adopt the recommendations as suggested ultimately rests with the physician.

Question-Answer system

As a concrete example, consider the following question:

> *"In children with an acute febrile illness, what is the efficacy of single-medication therapy with acetaminophen or ibuprofen in reducing fever?"*

The information need of the question may be formally encoded using the PICO format as:

Search Task:	therapy selection
Problem/Population:	acute febrile illness/in children
Intervention:	acetaminophen
Comparison:	ibuprofen
Outcome:	reducing fever

This query representation explicitly encodes the search task and the PICO structure of the clinical question. After processing MEDLINE citations, automatically extracting PICO elements from the abstracts, and semantically matching these elements with the query, the question-answering system will be producing answer along the lines of:

> *"Ibuprofen provided greater temperature decrement and longer duration of antipyresis than acetaminophen when the two drugs were administered in approximately equal doses." PMID: 1621668*
>
> *Strength of evidence: grade A*

✧ ✧ ✧

Concluding Remarks

Problems such as inaccurate diagnoses and poor drug-adherence pose challenges to individual health and safety. These challenges are now being alleviated, if not completely eradicated, with big data analytics using personalized drug regimes, follow-up alerts and real-time diagnosis monitoring. Pervasive and context-aware monitoring solutions are improving the quality of life for both patients suffering from chronic conditions and their relatives, as well as reducing long-term health care costs and improving the quality of care.

HITECH is a powerful opportunity for public health. With a phased approach, by 2017 public health will have more and stronger partners working to improve population health. There is increased opportunity to reduce disparities, control chronic diseases, and build a health promoting healthcare system that is accountable for the health of all communities and countries.

Health Records, whether on paper or electronic, play an important role as an information source for determining the time dependent health state of, and plans for, a patient. In the paper version, they consists mostly of semi-structured information, based on local traditions. In the electronic version a lot of efforts are being put into structuring the information by defining standard ontologies for the clinical domain. HL7 and OpenEHR are examples of generic models that can be specialized into local domain models. Electronic Health Records (EHR) in standard clinical domain models hold a great potential for supporting clinical research through improved efficiency, quality and reduced cost of clinical trials. However, the clinical models may not address the needs of unstructured text that is a result of transcription of dictations, direct entry by providers, or use of speech recognition applications. This free-text form is convenient to express concepts and events, but is difficult for searching, summarization, decision-support or statistical analysis. For reasons of efficiency and scalability, automated natural language processing (NLP) approaches are currently developed as a means for obtaining various types of clinical information from such free-text.

Decision support systems search large bodies of unstructured text for specific types of entities and relations, and use them to populate well-organized databases. These databases can then be used to find answers for specific questions. The typical architecture for such a system begins by segmenting, tokenizing, and part-of-speech (POS) tagging the text. The resulting data is then searched for specific types of entity.

The acquisition and representation of knowledge in clinical decision support systems is an actively evolving research field, characterized by modeling and software engineering issues of increasing complexity. The scope of designing knowledge-based systems in medicine is continuously evolving from being a mere diagnostic task to the broader issue of patient management, leading to a better integration in hospital information systems.

✧ ✧ ✧

Abbreviations

ADR	Adverse Drug Reactions
AI	Artificial Intelligence
ARRA	American Recovery and Reinvestment Act
CDA	Clinical Document Architecture
CDSS	Clinical Decision Support System
EBM	Evidence-based Medicine
EHR	Electronic Health Record
HHS	United States Department of Health and Human Services
HIE	Health Information Exchange
HL7	Health Level 7
HITECH	Health Information Technology for Economic and Clinical Health
MEMS	Micro Electro-Mechanical Systems
MU	Meaningful Use
NLP	Natural Language Processing
PHI	Protected Health Information
POS	Part-of-Speech

Bibliography

[1] *Adherence to long-term therapies*. World Health Organization, 2003.

[2] Davis K. *Estimates for the U.S. Civillian Noninstitutionalized Population*. Agency for Healthcare Research and Quality, 2012.

[3] *Global health risks: Mortality and burden of disease attributable to selected major risks*. World Health Organization, 2009.

[4] Ruby Pawankar; Giorgio Walter Canonica; Stephen T. Holgate; Richard F. Lockey. *White Book on Allergy 2011-2012 Executive Summary*. World Health Organization, 2012.

[5] Tracy D. Gunter and Nicolas P. Terry. The emergence of national electronic health record architectures in the united states and australia: Models, costs, and questions. *J Med Internet Res*, 7(1), 2005.

[6] Open disclosure standard: A national standard for open communication in public and private hospitals, following an adverse event in health care. Technical report, Australian Council for Safety and Quality in Health Care, 2003.

[7] *EHR Incentive Programs*. United States Federal Government.

[8] *Health Connect*. Australian Government Department of Health and Ageing, 2004.

[9] *Directive 2011/24/EU*. Official Journal of the European Union, 2011.

[10] Stanley Huff; Roberto Rocha; Harold Solbrig; Wade Barnes; Steven Shrank; Matt Smith. *Linking a Medical Vocabulary to a Clinical Data Model using Abstract Syntax Notation*, chapter Methods of Information in Medicine. 1998.

[11] Farzad; Mertz Kory; Hogin Emily; Atwal Parmeeth Williams, Claudia; Mostashari. From the office of the national coordinator: The strategy for advancing the exchange of health information. *Health Affairs*, 3(31):527–536, 2012.

[12] DC Hess; S Wang; H Gross. Telestroke: Extending stroke expertise into underserved areas. *Lancet Neurol*, 5(3):275–278, 2006.

[13] *The National Law Review*, chapter HIPAA/HITECH Enforcement Action Alert. Morgan, Lewis & Bockius LLP, 2012.

[14] Sicurello; M Vigano M, Fiore; F. *Artificial Intelligence in Medicine*, chapter A knowlege-based system to classify and diagnose HIV-Pneumonias, pages 185–190. IOS Press, 1993.

[15] Shoftliffe. *Computer-Based Medical Consultations: MYCIN*. Elsevier, New York, 1976.

[16] *AIME 87: Proc. Euro. Conf. Art. Intell. Med., Lecture Notes in Medical Informatics*, Berlin, 1987.

[17] Stefanelli. European research efforts in medical knowledge-based systems. *Artificiall Intelligence in Medicine*, 5:107–124, 1993.

[18] Francois Michel Aronson, Alan R; Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[19] Peter B.; Schmock Henriette; Dalgaard Marlene; Andreatta Massimo; Hansen Thomas; Soeby Karen; Bredkjaer Soren; Juul Anders; Werge Thomas; Jensen Lars J.; Brunak Soren Roque, Francisco S.; Jensen. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Computational Biology*, 7(8), 2011.

[20] Gopalakrishna Palem. *M2M Telematics & Predictive Analytics*. Symphony Teleca Corp., 2013.

About Author

Gopalakrishna Palem is a Corporate Technology Strategist specialized in Distributed Computing technologies and Cloud operations. During his 12+ year tenure at Microsoft and Oracle, he helped many customers build their high volume transactional systems, distributed render pipelines, advanced visualization & modeling tools, real-time dataflow dependency-graph architectures, and Single-sign-on implementations for M2M telematics. When he is not busy working, he is actively engaged in driving open-source efforts and guiding researchers on Algorithmic Information Theory, Systems Control and Automata, Poincare recurrences for finite-state machines, Knowledge modeling in data-dependent systems and Natural Language Processing.

He can be reached at Gopalakrishna.Palem@Yahoo.com