Data Mining in Health Informatics

Abstract

In this paper we present an overview of the applications of data mining in administrative, clinical, research, and educational aspects of Health Informatics. The current or potential applications of various data mining techniques in Health Informatics are illustrated through a series of case studies from published literature. The paper also provides a detailed discussion of how clinical data warehousing in combination with data mining can improve various aspects of Health Informatics. Finally, we point out a number of unique challenges of data mining in Health informatics.

1. Introduction

Health Informatics is a rapidly growing field that is concerned with applying Computer Science and Information Technology to medical and health data. With the aging population on the rise in developed countries and the increasing cost of healthcare, governments and large health organizations are becoming very interested in the potential of Health Informatics to save time, money, and human lives.

Human errors cause the death of between 44000 to 98000 American patients annually [30 as cited in 9]. Furthermore, in Unites States alone, drug-related morbidity and mortality costs more than $136 billion per year [26 as cited in 9]. Electronic patient records, computer based alerting, reminder, and predictive systems, and adaptive training tools for healthcare professionals can help reduce both the human and financial costs of healthcare.

As a relatively new field, Health Informatics does not yet have a universally accepted definition. The American Medical Informatics Association defined health Informatics as "all aspects of understanding and promoting the effective organization, analysis, management, and use of information in health care"[1]. Similarly, the Canada's Health Informatics Association definition of Health Informatics is "Intersection of clinical, IM/IT and management practices to achieve better health"[2]. These are both broad definitions that cover a wide range of technologies, from developing electronic patient record data warehouses to installing wireless networks in hospitals. A more specific definition is provided by the National Library of Medicine, which defines Health Informatics as "the field of information science concerned with the analysis and dissemination of medical data through the application of computers to various aspects of health care and medicine"[3]. Note that here, Health Informatics is limited to "analysis and dissemination of medical data", and would not cover pure IT practices such as installing a network in a hospital. Zaiane provides an even more specific definition, which divides Health Informatics into four subfields:

> *“Health Informatics is the computerization of health information to support and optimize (1) administration of health services; (2) clinical care; (3) medical research; and (4) training. It is the application of computing and communication technologies to optimize health information processing by collection, storage, effective retrieval (in due time and place), analysis and decision support for administrators, clinicians, researchers, and educators of medicine.”*

In this survey, we present an overview of the applications of data mining in various subfields of Health Informatics. For each subfield of Health Informatics, we provide a number of published papers as case studies of the current and potential applications of data mining. We also present how clinical data warehousing in combination with data mining can help administrative, clinical, research and educational aspects of Health Informatics. Finally, we discuss a number of unique challenges of data mining in Health Informatics.

2. An Overview of Health Informatics and Applications of Data Mining

As mentioned in the introduction, Health Informatics can be divided into four main subfields:

1. Administration of health services
2. Clinical care
3. Medical research
4. Training.

The following subsections present an overview of each subfield of health Informatics, and how data mining is, or can be, applied to extend and improve each subfield.

2.1 Clinical Care

Physicians and nurse practitioners make diagnostic decisions and treatment recommendations based on history, medical imaging, lab results and other text or multimedia records of patients. Health informatics allows doctors to have faster access to more relevant information, and thus make more optimal decisions. For instance, a centralized patient record database will allow a physician in a local clinic to have access to all the relevant medical records of the patient, anywhere in the country. Furthermore, applying data mining techniques on the centralized database will give doctors analytical and predictive tools that go beyond what is apparent from the surface of the data. For instance, a new practitioner can query for all the decisions that previous practitioners have made on a similar case. Similarly, a predictive model can advise doctors whether a certain case would be better treated as an outpatient or an inpatient.

2.1.1 Clinical Decision Support Systems

The applications of Health Informatics in clinical care decision-making are known as (Computer based) Clinical Decision Support System (CDSS)[1] Shortliffe defines a decision support system as "any computer program that is designed to help health professionals to make clinical decisions" [44 as cited in 34]. Applications of Clinical Decision Support Systems can be categorized into:

- *Information retrieval*: CDDS can offer search capabilities for medical queries. For instance the "antibiotic assistant" of HELP system (introduced in section 2.1.1.1) allows doctors to query the hospital experience with previous infections through the last five years [9].
- *Alerting systems*: A useful application of CDSS is to monitor inputs and check them for predetermined triggers [21]. These alert systems can be simple, like predefined drug-drug or drug-allergy conflicts, or complex, such as alerts based on analysis of various lab results and comparison with expected result protocols.
- *Reminders*: unlike alerts that are triggered by a specific change in input data, reminders are triggered by passage of time and are used for periodic tasks such as immunization or diabetes tests [21].
- *Suggestion Systems*: Unlike alerts, which indicate predetermined conditions in input data, suggestion systems are interactive processes that suggest action oriented messages based on their medical knowledge base.
- *Prediction Models*: CDSS prediction models can be categorized into diagnosis (defined as "aiding in the determination of the existence or nature of a disease" [4] and prognosis (defined as “the forecast of the probable outcome of an illness" [4]) [21]. An example of a diagnosis predictor is a model that detects *nosocomial* hospital infections based on information from Microbiology

[1] As we will discuss in section 2.1.3, this is not a universal view of CDSS. Some experts believe that CDSS include other aspects of Health Informatics, like administrative decision support, research, and training.

laboratory, nurse charting, and other sources. APACHE, introduced in section 2.1.1.2, is an example of a prognosis predictor which predicts ICU mortality based on a number of physiological variables.

The following subsections describe a number of Clinical Decision Support Systems currently in use in clinics and hospitals.

2.1.1.1 Case Study: HELP system

Health Evaluation through Logic Processing (HELP) system is an example of a Clinical Decision Support System that includes alerting systems, suggestion systems, and prediction models [9]. An example of an alerting system used in HELP is a model that monitors patient laboratory results, and has simple rule-based triggered to detect anomalies. A suggestion system included in HELP is a set of computerized protocols for managing care of Adult Respiratory Distress Syndrome (ARDS) patients. Both alerting and suggestion systems in HELP are rule-based models, developed by physicians, nurses, and specialists in medical informatics.

HELP includes two types of prediction models. One of these models is rule-based models, such as the one used in the Adverse Drug Events (ADE) detection system. The ADE detection system predicts the possibility of a drug reaction based on patient history and a set of predefined protocols. Aside from rule-based models, some prediction models in HELP use logistic regression, *e.g.* the model that predicts nosocomial hospital infections based on a number of risk factors.

HELP system has been developed and tested for more than 25 years and it is currently in use in many of the 20 hospitals operated by Intermountain Healthcare (IHC) [31 as cited in 9]

2.1.1.2 Case Study: APACHE series of models

The Acute Physiology and Chronic Health Evaluation (APACHE) series of models are developed to predict the individual patient's risk of hospital death in ICU, based on a number of physiological variables. The original APACHE model was developed in 1981 as an export-based scoring system. The later versions are based on logistic regression models. The models were trained on 17000 of cases in more than 40 hospitals [21].

2.1.1.3 Case Study: Pneumonia severity of illness index

The Pneumonia Severity of Illness Index is another logistic regression model that predicts the risk of death within 30 days for adult patients with pneumonia. The model was developed by the Pneumonia Patient Outcome Research Team (PORT) in 1997 and was validated over 50000 patients in 275 hospitals in US and Canada. The developers claim that by using this model, up to 30% of pneumonia patients can be treated safely as outpatients, resulting in an annual savings of 1.2 billion dollars [21].

2.1.2 Data Mining in Clinical Decision Support Systems

Aside from some use of logistic regression in predictive models, there is currently limited or no applications of data mining in Clinical Decision Support Systems. Most of the current systems are rule-based and are developed manually by experts. Data mining can extend and improve all categories of CDSS, as illustrated by the following examples.

In information retrieval systems, data mining can be applied to query multimedia records. Image and video mining, along with applications of natural language processing techniques will allow physicians to effectively search through patients' medical imagery, laboratory results, and other medical records.

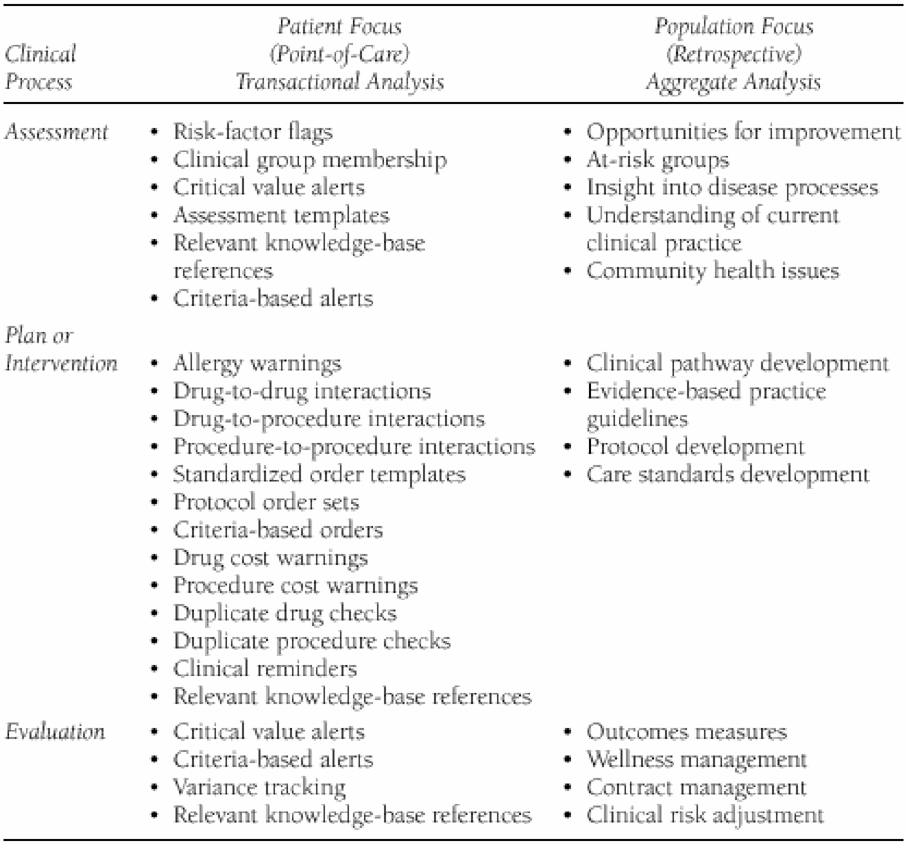
Data mining can be used to automatically discover and update thresholds used in alerting and reminder systems. Maintaining and updating the underlying knowledge of rules is one of the important challenges that limit the adoption of CDSS by health organizations [21]. In the most basic form, data mining algorithms can be applied to monitor the thresholds used in alerting and reminder systems, and either automatically update them or alert human experts that the current thresholds should be reconsidered.

In suggestion systems, instead of depending on experts to manually develop the underlying protocols, data mining approaches can be applied to automatically generate these protocols based on historic data. A team of human experts can then review the generated protocols before deploying them in the final suggestion systems.

Prediction models are the most evident and straightforward targets for applying data mining algorithms. There has been extensive research on the applications of supervised and unsupervised learning algorithms on medical data in machine learning and data mining communities. However, as we will discuss in section 4, most of these algorithms are not well understood or accepted in the medical community. The more advanced prediction models developed in the data mining community have the potential to increase the accuracy of the current models used in CDSS.

2.1.3 A broader view of Clinical Decision Support Systems

Some experts present a broader view of CDSS that it is not limited to the clinical care subfield of Health Informatics. Ledbetter and Morgan state that the CDSS capabilities are useful in all phases of the clinical process: (a) assessment, (b) planning, (c) intervention, and (d) evaluation [32]. Table 1, taken from their article, describes the potential applications of CDSS for the cases of a patient-specific focus as well as a population-specific (or aggregation based) focus.



Clinical *Process*	*Patient Focus* *(Point-of-Care)* *Transactional Analysis*	*Population Focus* *(Retrospective)* *Aggregate Analysis*
Assessment	• Risk-factor flags • Clinical group membership • Critical value alerts • Assessment templates • Relevant knowledge-base references • Criteria-based alerts	• Opportunities for improvement • At-risk groups • Insight into disease processes • Understanding of current clinical practice • Community health issues
Plan or Intervention	• Allergy warnings • Drug-to-drug interactions • Drug-to-procedure interactions • Procedure-to-procedure interactions • Standardized order templates • Protocol order sets • Criteria-based orders • Drug cost warnings • Procedure cost warnings • Duplicate drug checks • Duplicate procedure checks • Clinical reminders • Relevant knowledge-base references	• Clinical pathway development • Evidence-based practice guidelines • Protocol development • Care standards development
Evaluation	• Critical value alerts • Criteria-based alerts • Variance tracking • Relevant knowledge-base references	• Outcomes measures • Wellness management • Contract management • Clinical risk adjustment

Table 1: Potential applications of CDSS (taken from [32])

Courtright *et al.* have developed a list of core requirements for CDSS tools and the following comprise the major requirements discussed in their article [14]. The CDSS tools need to:

- Have enhanced networking and distributive features
- Be used at all decision making levels in an organization

- Be used in both real time and retrospective modes
- Enabled predictive capabilities using classical statistics
- Utilize “white box” (openly disclosed but protected) methodologies for prediction and detailed support to provide the kind of accuracy rates required for health care decision making

Note that these requirements are not limited to the clinical care subfield of health informatics. In addition to the above, we feel that in the broader view, CDSS tools also need to:

- Apply AI techniques for disease prediction
- Use other techniques such as spatial data mining and spatio-temporal data mining to assist in health care decision-making
- Be able to provide a feedback to the decision makers regarding the efficiency of the system
- Have Graphics and graphing capabilities so as to be able to present the data in several formats such as tables, bar charts, pie charts, graphs *etc.*
- Have tighter security, and access controls in order to avoid personal data falling into malicious hands.

In the longer term, it is expected that the clinical data can be used to assess “episodes of risk” [14] wherein CDS systems will help in early identification of risk factors such as diet, exercise, travel, and air and water standards. It is also expected that in the future CDS systems will also help in performance benchmarking, continuing medical education of the clinicians by the use of their own data, identification of best practices, creation and utilization of standard terminology etc. [14].

2.2 Administration of Health Services

Administrators of health care organizations make hundreds of critical decisions on daily basis. As in any administrative position, the quality of these decisions directly depends on the quality of the information that the decisions are based on. For example, the administrators in a hospital need to decide on the amount of supplies and number of staff and free beds required for an upcoming month. To make this decision, the administrators require an accurate prediction of the number of patients to expect during the coming month, and an approximation of how long each patient will remain in the hospital. As another example, the federal and provincial health administrators need to decide whether a disease outbreak is in progress, and if so, what preventive measures will be most effective against it. To make these decisions, the administration requires a system that can accurately predict a disease outbreak, and also model the cost and benefit of different preventive measures.

The following case study illustrates the applications of data mining techniques on epidemic detection. More examples of administrative decision support will be discussed in Section 4, where electronic patient records and various data warehousing techniques are introduced.

2.2.1 Case Study: detecting disease outbreaks

In "Decision Theoretic Analysis of Improving Epidemic Detection", Izadi and Buckeridge introduce a method to improve existing threshold-based epidemic detection methods by using POMDPs (Partially Observable Markov Decision Processes) [24]. The main idea is that the potential costs and effects of intervention can be quantified and be used to optimize the alarm function. Furthermore, the intermediate investigation steps, such as asking for more systematic studies, or more investigation done by human expert, can also be quantified in terms of cost and effect. Based on these cost and effects, the system can learn to recommend the optimal action. While the paper concludes that POMDPs can improve the accuracy of the current outbreak detection methods, the current level of false alarms (3 false alarms in every 100 days) seems to be unacceptable for practical use.

Similarly, Cooper et al. investigates the use of Bayesian Networks for outbreak detection, focusing on modeling non-contagious outbreak diseases, such as airborne anthrax [13]. The Bayesian network is divided into 3 groups: global (G), interface (I) and people (P). Furthermore, in order to make the algorithm scalable, people with the same attributes are grouped in the same class. The network is evaluated based on data generated by a simulator. Given weather conditions from Historical meteorological conditions for a region, parameters for location and amount of airborne anthrax, a Gaussian plume model derives the concentration of anthrax spores that are estimated to exist in each zip code. The authors compare a non-spatial model with a spatial model and conclude that with spatial data they can get better results based on false positive rate.

2.3 Medical Research

Most current successful applications of data mining in Health Informatics are in the subfield of medical research. The reason is that most of the current health related data are stored in small datasets scattered through various clinics, hospitals, and research centers. However, most applications of data mining in clinical and administrative decision support systems require homogeneous and centralized data warehouses (see section 3). On the other hand, data mining methods can still be successfully applied on small and scattered datasets, and help researchers extract insightful patterns, cause and effect relationships, and predictive scoring systems from currently available data.

The following subsections introduce a number of examples of data mining techniques applied on small datasets for medical research.

2.3.1 Case Study: drug exposure side effects from mining pregnancy data

Chen et al. investigate the possible effects of multiple drug exposures at different stages of pregnancy on preterm birth, using SmartRule, a data mining technique for generating associative rules [11]. In this work, two subsets of Danish National Birth Cohort (DNBC) dataset are used. The first subset contains 4454 records including 1000 women who were depressed and/or exposed to various active drugs. This set is used for finding the side effects of anti-depression drugs. The second subset contains 6231 records, including 414 preterm cases. This set is used for finding side effects of multiple types of drugs. The authors develop a tree hierarchical model for organizing the generated rules, in order to ease the recognition of interesting rules by human experts. Using this system, the authors claim that they are able to find novel and interesting rules.

2.3.2 Case Study: Automatic in vivo microscopy video mining for leukocytes

Zhang et al. introduce a framework for video mining in vivo microscopy images [47]. The goal is to track leukocytes in order to predict inflammatory response. In vivo microscopy allows researchers to capture images of the cellular and molecular processes in a living organism. However, automatic mining of the imagery is challenging due to severe noise, background movement of the living organism, and change of contrast in different frames. Zhang *et al.* first apply a frame alignment technique, using RANSAC, to correct the camera-subject movement, and then apply a number of probabilistic methods to detect moving leukocytes. Adherent leukocytes are detected, after the moving ones are removed, by finding thresholds for contrast values. The experimental results show 1% false positives and 50% recall on detecting moving leukocytes, and 2% false positives and 95% recall on detecting adherent leukocytes.

2.3.3 Case Study: Knowledge-based analysis of microarray gene expression data using Support Vector Machines

Brown et al. apply Support Vector Machines on gene expression data to classify genes based on functionality [10]. This is based on previous experiments suggesting that genes with similar functionality have similar patterns in microarray data. The authors claim that SVMs are well suited to the problem of microarray gene classification, because they perform well in extremely high-dimensional feature space. A training set is generated by combining the DNA microarray data of a set of genes that have certain functionality (i.e. positive labels) and a set of genes known not to be a member of this functional class (i.e. negative labels). Once SVM is trained on this training set, it can determine whether a new gene belongs to the certain functional class, or not. The authors apply SVM, with a number of different kernels, on gene expression data from the budding yeast Saccharomyses cerevisiae, with 5 predefined functional classes. The prediction performance of SVM is compared to predictions by a number of other classification methods, including decision trees, Fisher's linear discriminates, and Parzen Windows. The authors claim that SVM outperforms all the other classification methods.

2.3.4 Case Study: Association rules and decision trees for disease prediction

Ordonez applies different classifiers, associative classifier and decision trees, for predicting the percentage of vessel narrowing (LDA, RCA, LCX and LM) compare to a healthy artery [35]. The dataset contains 655 patient records with 25 medical attributes. Three main issues about mining associative rules in medical datasets are mentioned in this work. A significant fraction of association rules are irrelevant and most relevant rules with high quality metrics appear only at low support. On the other hand, the number of discovered rules becomes extremely large at low support. Hence, association rules are used with constraints. Each item corresponds to the presence or absence of one categorical value or one numeric interval. First constraint is that there is a limit on the maximum item-set size. Second, the items are grouped and in each association, there is at most one from each group. The third constraint is that each item can only appear in antecedent or consequent. The result from associative classifier is compared with two decision tree algorithms: CN4.5 and CART. The authors demonstrate that associative rules can do better than decision trees for predicting diseased arteries.

2.4 Education and Training

The fourth subfield of health informatics is related to educating new healthcare professionals and retraining and keeping the current staff up-to-date with recent advances in technology. The education and training subfield of Health Informatics can be viewed as an instance of the rapidly growing field of e-learning. An increasing interest in applying data mining techniques to e-learning has emerged in recent years, and some of the early applications show promising results [38].

Data mining techniques can benefit all three groups of people who are in contact with a learning system: students, educators, and administrators [38]. Data mining techniques can monitor the success of students at various learning tasks, and recommend relevant resources, materials, and learning paths to achieve a more successful learning experience. For educators, data mining techniques can provide objective feedback of the structure and the content of a course, discover the learning patterns of the students, and cluster learners into smaller groups that have similar educational habits and needs. Administrators benefit from data mining techniques by learning about the behavior of their users, so they can optimize the servers, distribute network traffic, and learn about the overall effectiveness of the offered educational programs.

The following two case studies present an overview of a relatively new Health Informatics e-learning tool called HOMER, and a data mining technique to find relevant articles for a particular gene.

2.4.2 Case Study: Homer, an online learning community

Homer is a centralized e-learning system and an Internet community, developed for the medical students of the University of Alberta [5]. Homer provides online access to a variety of learning materials, including medical dictionaries, demonstration videos, and faculty presentations. One important feature of Homer is the lifetime membership, which grants medical students continued access to learning materials after graduation [18].

2.4.2 Case Study: Finding relevant references to genes and proteins in Medline using a Bayesian approach

Leonard *et al.* apply a Bayesian approach to find cross-references between the symbol of genes and proteins and Medline articles [33]. The authors extract gene and protein symbols from article titles and abstracts, using a dictionary of gene and protein symbols and a dictionary of English words along with a set of rules. A different set of rules is used to find new gene and protein symbols that are not included in the gene and protein symbol dictionary. After assigning articles to identified genes and proteins, a Bayesian estimated probability (EP) based on word frequency is used to find the relevancy of each assigned article to each gene or protein. Hence, only the relevant articles are chosen for each gene or protein and the result will be a set of relevant references for each gene or protein.

3. Data Warehousing in Health Informatics

This section demonstrates how clinical data warehousing in combination with data mining can help each of the four subfields in Health Informatics discussed in section 2. In particular, we will focus on how clinical data warehouses support the following:

- Improvement in Clinical Care
- Better administration of health services
- Aiding medical research, and enhancing its quality
- Cheaper and more effective training

In the present times Electronic Patient Record (EPR) has become a buzzword in the field of E-health. Ledbetter [32] defines EPR as an electronically maintained (computerized) patient record system with point-of-care tools that support clinical care. According to Ledbetter, in an ideal situation an EPR should "support all episodes of care to create a complete longitudinal patient record". Kim *et al.* define EPR as an electronic collection of diagnostic reports of an individual patient's entire medical history. These reports can have varied formats such as text, multimedia, etc. where multimedia itself would encompass Digital Image and Communication (DICOM), 3D Image set, Voice recording, Health level 7 (HL7) types [**27**].

EPR based records hold several advantages over the paper-based records that are currently being phased out. Some of these features are: (a) Simultaneous access by multiple users (b) on-line information processing for clinical and administrative decision (c) access to data from multiple sources (d) cost-effectiveness/apart from the initial investment (e) data representation and richness of the content of data (f) reliability and ease of distribution of data and (g) security. It is worth emphasizing that all of the above would not have been possible without the great strides made in the field of Information Technology, Computing, Data mining, Information Security, and also the advent and proliferation of the World Wide Web (WWW).

The use and storage data in the electronic form has created opportunities for applying data mining techniques to extract the hidden knowledge in the data. Frawley et al. define data mining as the "nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [17]. Unfortunately the electronic data resides on different and heterogeneous systems with the result that integration becomes a challenging task. Data warehouses allow us to perform this complex task of integrating the heterogeneous data; simultaneously they act as central repositories for the data. The data warehouses used in health Informatics are somewhat different in nature (more complex), hence they are called clinical data warehouses (as discussed later).

3.1 Data Warehouses vs. Real-time Databases

The real time decision-making processes rely on the use of Online Transaction Processing (OLTP) systems that are patient specific while the Online Analytical Processing (OLAP) systems carry out an aggregate analysis based on data for a group of people. The OLTP and the OLAP systems together contribute to the success of a CDS system. OLTP systems need to handle a large volume of transactions required by patient-care system such as patient registration, clinical documentation, order entry, results review and clinical alerting. Ledbetter [32] argues that for this reason the performance of an OLTP system may suffer if the system is used for aggregate analysis. OLAP systems, on the other hand, do not have any data of their own, and rely on OLTP systems for data feed. These systems are always off-line as they lag behind the OLTP systems sometimes by a day or so, and sometimes months altogether. Systems that employ OLAP techniques are called Data Warehouses (DW). In this section we look at Data Warehouses from the standpoint of their theoretical foundation, and their functionality; the issues related to design and construction are dealt later.

Inmon [23] defines a Data Warehouse as a repository for keeping data in a "subject oriented, integrated, time variant and non-volatile manner that facilitates decision support". A Data Warehouse transforms the

OLTP data in a way that facilitates mining, the information from that data much easier—the standard data structure is a multi-dimensional cube (figure 1) that allows the user to rapidly change the dimensions by which a report is filtered, sorted or grouped [32]. Of course, a researcher could “drill “to the patient record level if they so desire, however, it would be much easier and faster to get the same information by querying on OLTP system. From the point of design, a data warehouse consists of fact tables and dimensions tables, sometimes called a STAR schema. The fact tables could include measurements, orders, and observations along with events such as admissions, discharges and transfers while the dimension tables could include patients, diagnosis, medications, supplies, clinical units etc.

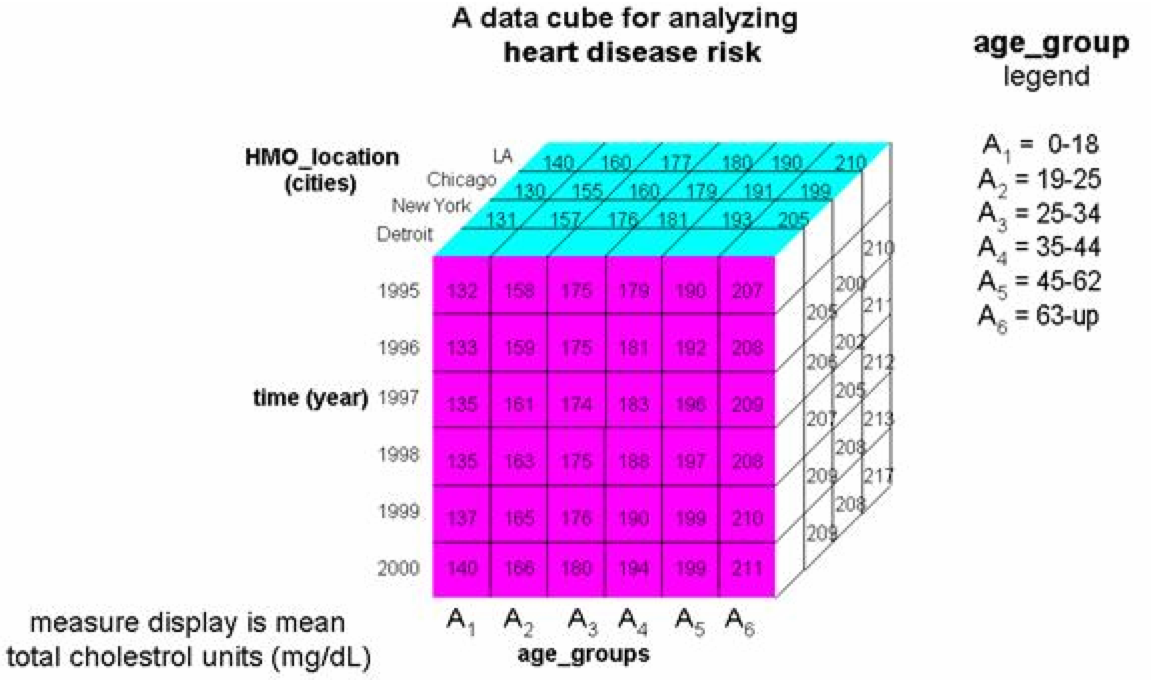


Figure 1: Pictorial view of a typical cube in data warehousing (taken from [6])

Courtright et al. believe that CDS systems need to go beyond “simple flags and alarms” at the point of care [14]. They state that the OLTP systems should aim for what is really required by the clinicians at the point of care, such as a likely “clinical outcome trajectory” for a patient, the optimal course of treatment in the short and the long term. Also the clinicians should be able to make mid-course corrections, and also they should be able to “model the impacts of different clinical decisions as the patient’s clinical course changes”. The information that clinicians really require for making informed decisions are factors such as the patient’s risk characteristics, diagnostic and therapeutic interventions, and clinical outcomes.

On the other hand, OLAP techniques can be used to analyze those business rules and clinical action that affect the profitability, resource-planning and productivity of the healthcare institution. Courtright et al. are of the view that OLAP techniques should provide an easily interpretable single value indexed score for any assessment, and this score should incorporate measures such as cost, health status etc. [14]. The chosen health care pathway should be the one that maximizes the above-mentioned score.

It has been well established that the aggregation analysis helps in improving the quality of healthcare delivered to the patients; however what has not been observed is its impact on the clinicians. Centralizing the databases provides the clinicians with a broad insight into the actual clinical practices. Furthermore aggregation analysis can be very useful in the case of performance benchmarking for clinicians by utilizing the same clinical data that is used for making healthcare decisions. Further such systems can help clinicians determine the statistical impact of individual clinical decisions, which, in the long run, can help them come up with their own clinical pathways, and to be able to compare those with the standard practices, thus fuelling and aiding medical research with greater ease. The ease of comparison of different pathways helps in improving the quality of medical research.

3.2 Chalk and Cheese: Data Warehousing and Clinical Data Warehousing

A Clinical Data Warehouse (CDW), as defined by Gray [20], is a "place where healthcare providers gain access to clinical data gathered in the patient care process". An appropriate question at this point would be: "How is a CDW different from other Data Warehouses?" The answer lies in the fact that everything from the planning process for building a data warehouse to its design components, the software employed in the ETL (Extraction Transformation Loading) phase, the extent of the essential background knowledge of the architect is vastly different between the two kinds of Data Warehouses. A CDW is immensely complex to build, and maintain when compared to other Data warehouses. Herein, we discuss some of the differences and the complexities of a CDW.

In order to speed up the time-consuming queries DW architects employ a very common practice –building materialized views based on aggregate values. However, according to Gray [20], a lot of data going into the CDW is not additive at all e.g. vital signs of patients such as blood pressure, heart rate measurements etc. These form a large volume of the patient data. As a result no aggregation can be done even if only one such non-additive column is present in a table, thus precluding the possibility of speeding up queries by using materialized views.

The process that moves the data from the source to the CDW should have a minimum impact on the operations of the transactional system (OLTP). Also, the time taken to transform and store the data on the CDW should be as less as possible. For ordinary data warehouses the transformation step is carried out in the evening when there is little or no activity; however the CDW being operational 24×7 the assigned budget is never more than one hour. Since the transformation is highly CPU intensive that can affect the query performance of CDW itself, the real transformation budget is only 10 minutes. A CDW needs to integrate data from multiple sources, and hence synchronization and consistency of this data is very important. If a database loading a part of the data goes down such that the rest of the data will force the CDW to be in an inconsistent state then the decision whether to proceed or wait becomes difficult in the light of the fact that the time window allowed for loading transformation is very small.

While understanding how an organization operates—their business rules and business logic is not an easy task, it all the more difficult in case of hospitals. It is compounded by the fact that individual hospitals can follow different practices, and tend to have different terminologies for the same task. As a result not only off-the-shelf (generic) CDW's cannot work, architects who design these DW have to be conversant with the terminologies and the practices. Finding such architects is not an easy job.

3.3 Evidence-based medicine: Data Warehouses for Healthcare Decision-making

Evidence-based medicine is the use of the latest, most respected, and well judged piece of evidence for making informed choices about the diagnosis and treatment of a diseased patient. Stolba and Tjoa define the task of evidence-based medicine as one that "complement[s] the existing clinical decision-making process with the most accurate and the most efficient research evidence" [45]. Another definition given by Sackett et al. describes eidence-based medicine as the "conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients" [39]. In their article, Stolba and Tjoa [45] present an example of a diabetic patient suffering from progressive liver disease, such that the clinician will need to find the most effective therapy for the patient's condition that does not conflict with their diabetic treatment. This is achieved by the clinician searching through the evidence-based guidelines for finding the most recent and most effective treatment for the liver diseases, and then using another query to make sure that the treatment for the liver disease does not conflict with the one for diabetes.

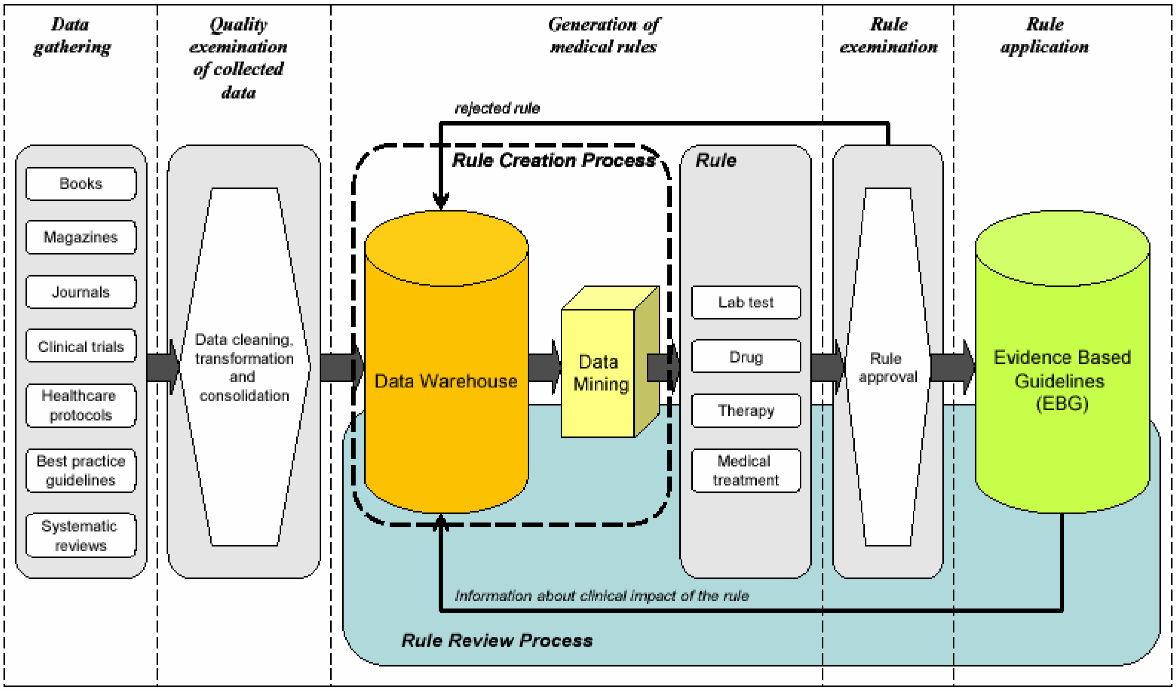


Figure 2: Generation of evidence-based guidelines (taken from [45])

The combination of data mining and evidence-based medicine is propelling Health Informatics into exciting and novel avenues. The key to practising evidence-based medicine lies in creating rules that are based on evidence from aggregate analysis, and are throroughly studied and well researched by experts. Also, these rules need to be delivered to the clinicians at the point of care in the form of alerts. The relevant data-sources for evidence-based medicine as outlined by Stolba and Tjoa [45] are:

- Evidence-based guidelines (in the form of rules)
- Clinical data (Pharmaceutical data, patient data, medical treatments, length of stay)
- Administrative data (Staff skills, Nursing care hours, staff leaves, overtime)
- Financial data (Drug costs, treatment costs, staff salaries, accounting)
- Organisational data (Facilities, Equipment, Room occupancy)

The following is after [45]. Figure 2 is a schematic diagram of the steps followed by the rule generation process wherein data is gathered from varied sources in the first step, and it is cleaned and transformed in the second step. Next, medical associations are found by applying data mining in a data warehouse environment where the data warehouse itself contains patient data, pharmaceutical and clinical data. The associations thus found are sieved through used by knowledge workers to isolate those that represent hidden knowledge in the data. All such associations are collected, and rules based on these are created in the form of laboratory tests, therapies, recommended drugs, or medical treatments; the rules thus created need to be examined by a committee of experts who can either approve or reject a rule; the approved rules are added to the Database along with the evidence-based guidelines. The clinical impact of each new automated rule needs to evaluate after a certain period of time, typically six months after it is introduced. Also, all the rules need to be evaluated at least on an annual basis and those that are found to be not valid anymore in the light of the current research need to be discarded.

The schematic diagram (figure 3) shows the role of data warehousing in facilitating evidence-based medicine, at the point of care. The following is after [45]. Initially the clinician defines a clinical question based on the patient’s disease. After that he/she uses standard reports, queries etc. to query the data warehouse. The evidence-based guidelines outputs results in the form of medical treatments, drugs etc.

which then need to be matched with a patient's health history, existing clinical equipment, and availability of the staff. Based on all of the analysis the best fitting rule is chosen, and presented to the clinician.

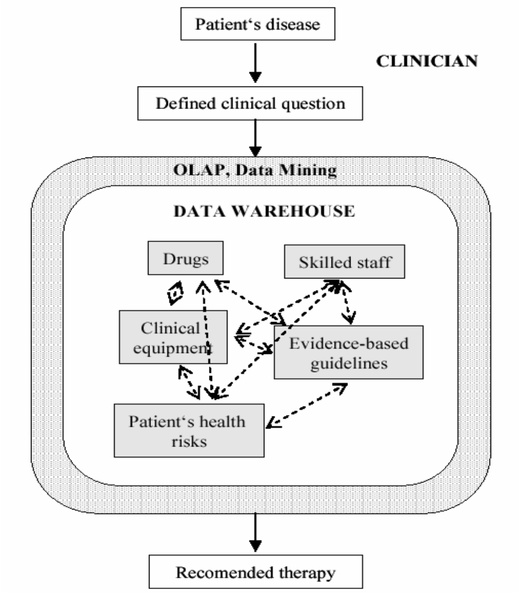


Figure 3: Determination of line of treatment (taken from [45])

In the recent times the cost of providing quality healthcare has increased tremendously. At the same time due to higher longevity, and the afflictions thereof, the total number of patients who need healthcare (as a fraction of the total population) has also increased. As a result healthcare institutions and medical insurance companies are being forced to adopt cost-cutting measures. Also the healthcare institutions are being forced to enlarge their facilities to cater towards the rising volumes. Apart from financial considerations, staffing problems, best utilization of available resources, and maintaining the quality of healthcare under such conditions are the major issues. Most case-studies, some of them even tracking projects from their inception to the end, have found that in the longer run data mining in combination with data warehousing has resulted in faster healing rates, reduction in treatment costs, better quality of care, and very few clinical mistakes (if any) on the part of the nursing staff [14], [20], [36] [47], [46]. These case-studies justify the initial investment made in constructing these systems. Stolba and Tjoa mention that the use of Data warehouse results in "avoiding the duplication of examinations, time saved through automation of routine tasks and the simplification of accounting and administrative procedures" [45].

3.4 Early Intervention: Data Warehousing in Disease Management

It has been observed that data mining, when applied in combination with a clinical data warehouse, has been very successful at extracting the early predictors of some diseases such as: Asthma, diabetes, cardiovascular diseases etc. Once the early predictors of a disease are extracted and the patients at risk are identified, they can be: invited to join awareness campaigns, signed up for disease management programs etc. Disease Management programs have been shown to improve patient care, lower the disease occurrence rates, and also lower the healthcare costs.

Ramick recommends data marts (smaller sized data warehouses specific to a department) for disease management programs because it reduces the maintenance issues, and requires less financial resources for deployment. It is important to note that whether the healthcare institution uses a data mart or a clinical data warehouse the data cleansing process in case of disease management programs is very different as compared to normal data warehouses. Ramick also points out that in case of disease management data the data in the CDW performs two functions— stratifying the patients by risk level for targeted medical conditions and tracking patient's progress through the disease management program [37]; hence, one needs to proceed with caution when eliminating data during the data cleaning phase because a patient's address, their occupation etc. could be significant. For example an asthmatic patient’s address could reveal the environmental hazards in the area that they live in.

Ramick discusses case studies where disease management programs are being implemented based on data mining in clinical data warehouses. In one such case U.S. Quality Algorithms (USQA) collects administrative data from pharmacies, laboratory claims etc. Certain ailments that can be controlled by disease management are flagged in the CDW, and patients at risk are identified based on the data related to diagnosis, procedures, laboratory tests, and drug prescriptions in the CDW for each patient. Another company, Blue Cross and Blue Shield, is in a unique position–their data warehouse contains not only the information about the lab test ordered, but also the test results. The time lag associated with this data is only a couple of days so that information can be analyzed in time, and have greatest impact when it is needed.

In our opinion, apart from cube querying other data mining techniques can be very useful for the data in the CDW. Spatial data mining and Spatio-temporal data mining could reveal great insights into a patient’s condition based on their geographic location. Once a cause is identified deeper studies are needed for all the member patients residing in that geographic location. Machine learning techniques such as clustering, building a classifier, contrast-set mining based on demographics etc. are some of the other data mining techniques that can extract interesting and previously hidden patterns in the CDW data. Data mining on the Disease Management Program data that resides in the CDW not only helps in early detection and prevention of diseases, and efficient targeting of resources, it can also aid in the current medical research by identifying the most common diseases affecting the general population that are the result of society-lifestyle, environmental factors or personal choices. At the same time it can also act as an early warning system for the health administrators as well as the general public alike.

3.5 Monitoring the Monitor: Data Warehouses for Supervising Feedback Integration

In Health Informatics one often tends to wonder if the clinical practices are being followed properly by the clinicians, or if the practices can further be improved. In most other fields the intention of application of the data mining techniques is to extract previously hidden nuggets of information, however in case of Health Informatics one of the aspects of data mining in such a case is to monitor the changes made to the processes based on the “hidden information” obtained from the data i.e. if the changes made to a previously faulty practice produce positive results. The EPR data can and should be used for providing feedback for process improvement as well as for finding the deficiencies in the system [7]. Grant *et al.* define feedback as a “source of objective information of the process and outcome of patient care” [19]. In their point of view feedback should enable itemized review by a patient care team, critique with respect to best evidence, be a primary source of information for consensual practice improvement and support education for students and the team; they provide an example wherein a system was developed to provide feedback to the clinical teams before the installation of their clinical Data Warehouse. The system was used while conducting a clinical study for investigating whether there was an excess use of blood gas measurements in the ICU. Two physicians and two nurses formed a committee of experts for studying the above-mentioned problem. During the evidence phase (figure 4) the committee considered evidence in the form of data to find the relation between blood-gas requests and special events like surgery, time of the day etc.

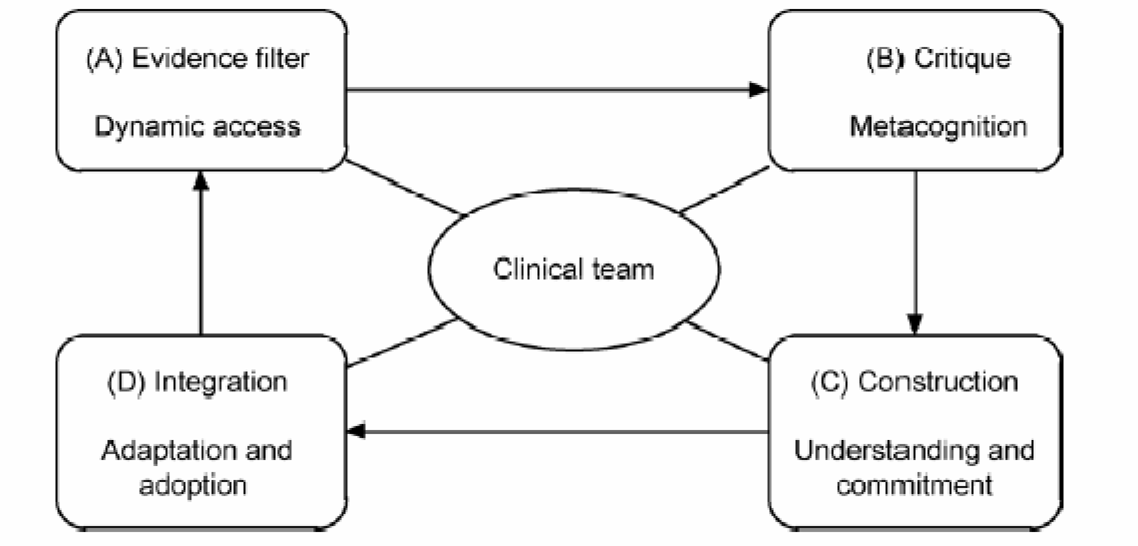


Figure 4: The autocontrol practice change methodology (taken from [19])

In the critique phase the model was improved upon and solutions and causes were discussed. In the construction phase a plan was laid out for a change in the practice. Finally, in the integration phase the changes were adopted and the practice data evaluated to measure the failure of the change that was brought into force. While the above does not require the use of a CDW, Grant et al. are of the view that the use of a CDW would optimize the previous task by using various tools and approaches for using practice data as feedback for practice change, optimization, and innovation [19].

Grant *et al.* [19] discuss the use of the dashboard concept for data enhancing, critiquing, and understanding the data. The term dashboard is usually used to describe a system that provides a human computer interface to the user by employing a set of windows can be used for dynamic querying of data within certain partitions, ranges, and combinations such that the results of the queries are portrayed in the form of graphs, tables, charts etc. Since the queries can be constructed and run dynamically, it allows the decision maker to run a set of such queries, some of them based on the results of the previous queries, in order to assess a certain situation from all angles. The dashboard concept is similar to that of performance indicators and benchmarks that can provide feedback at the same time for practice evaluation and change.

The use of audit and feedback as a tool for quality assurance has been studied widely [25]. However, the use of feedback as a tool for bringing about change in faulty practices has been severly limited because of several reasons--one of the biggest reasons is resistance to change. It has also been reported that some individuals misconstrue the whole purpose of feedback as something that might be used against them. A review study regarding the unnecessary use of lab tests found no evidence of the success of implementation of feedback.

3.6 Data Warehouse: Construction and Design

The task of designing and construction of a Data warehouse is very complex – it involves many technical issues related to a number of fields and subfields. Sen and Sinha [43] discuss about fifteen methodologies for this purpose. Sahama and Croll [40] explore the pros and cons of some of these methodologies for the purpose of designing their own Data warehouse. Batini et al. [8] discuss various strategies such as top-down, bottom-up, inside-out, and mixed strategy. For the purposes of a broad classification Hackney [22] classifies the design philosophies into two categories viz. Enterprise-wide Data Warehouse design and Data-Mart design. Sen and Sinha do an exhaustive comparison of various such Infrastructure-based philosophies in table 2 of their article. The data mart design philosophy was first discussed by Kimball [28] wherein a combination of the top-down approach and the bottom-up approach is presented, and the union of all such data marts forms a data warehouse. The metadata part of a data warehouse is much more voluminous than that of OLTP systems. Sen and Sinha advise the use of a metadata management for this purpose.

For the comparison of the different data warehousing methodologies the reader is referred to the article by Sen and Sinha [43]. Herein we discuss two important points from their article. The first point is related to change management. Company diversification, merger, acquisition etc. may lead to a redefinition of business objectives, priorities, and business rules. Also, the design of the data warehouse needs to incorporate the inherent dynamicity in the data such as new products, new sales regions, customers address changes etc. The authors stress on the fact that change management is an important issue that is often neglected by the vendors.

The final point that we would like to talk about regarding the design and construction of a data warehouse is the Extraction, Transformation and Load (ETL) step. Ironically, while the more complex technological tasks have been solved the simple task of extracting the data from different sources (that may involve different platforms, file types etc.), cleaning and integrating it together before loading it in the CDW turns out to be the most challenging task. Sahama and Croll report that as much as 90% of the effort could be spent in this step alone, however a better modeling process can save a lot of time and effort during this phase.

Unlike other decision systems in cases of a CDW coming up with all the business requirements at the beginning of the project is not feasible, as the users are not aware of the hidden knowledge in the data. At the same time they are also not aware of the capabilities of the CDW. Sen and Sinha, as well as Inmon [23] advise against using a Software Development Life Cycle strategy for the implementation purposes. Other techniques such as Spiral development approaches have also been proposed.

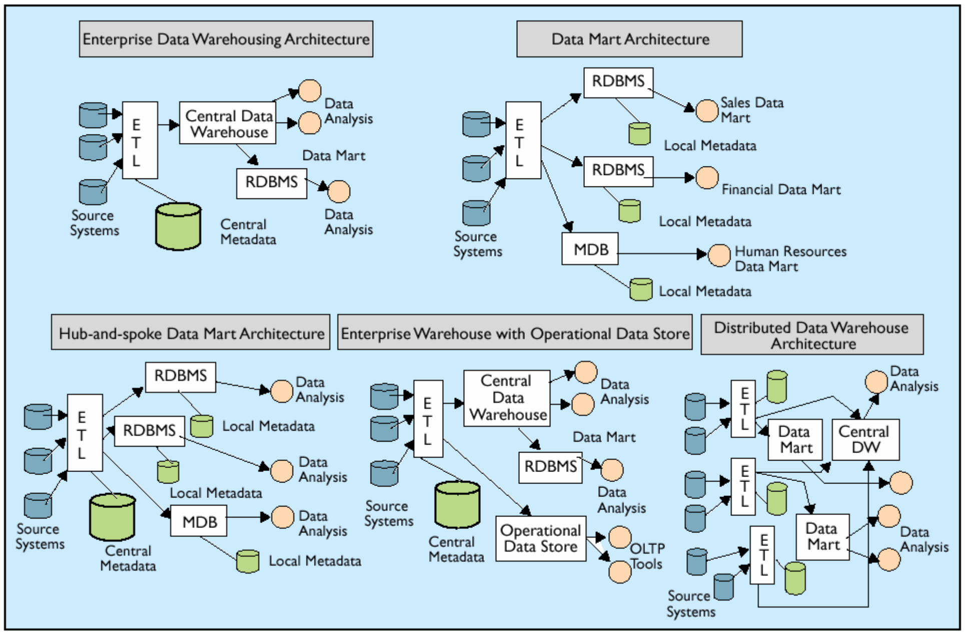


Figure 5: Different warehousing methodologies (taken from [43])

3.7 Case Studies

Here we present case studies from research papers to support the already established view that data mining, in combination with clinical data warehousing can play an important role in the administrative, clinical, research and training aspects of Health Informatics.

3.7.1 OLAP for Claims Processing

Verma and Harper discuss the claims processing system for PriMed Management, Inc. which is a management services company for Hill Physicians Medical Group [46]. Until 1996 PriMed was using their transactional system for processing claims as well as for input of daily authorization requests for medical procedures. They were finding that the reporting jobs from the transactional system were slowing down other jobs such as claims processing and authorization processing, very significantly. Hence they decided to construct a data warehouse that could, apart from speeding things up, help the senior management track the trends in the data. After the data warehouse was built, data mining was applied, and it was observed that the Health Data Analysis Department became very effective at responding to typical requests for information. They were able to initiate new reporting tools in the form of standardized monthly or quarterly reports for the purposes of: (1) Physician compensation analysis, (2) Physician profiling, (3) Utilization (facility) reporting, (4) Disease state management reporting, and (5) Analysis of contract viability.

3.7.2 Clinical Data Warehouse for University Health network

Ledbetter and Morgan discuss the details of their project that involved the construction of a Clinical data warehouse (CDW) for the University Health Network [32]. University Health Network (UHN) comprises of three Toronto hospitals: Toronto General Hospital, Toronto Western Hospital, and Princess Margaret Hospital. They have a transaction processing system called Patient 1 that is used for admission-transfer-discharge (ADT) order-entry, and results-review. This system contains information for more than 300 million patients with twelve million visits spanning a period of ten years. Patient 1 is available to the clinicians 24 hours a day. In order to identify the opportunities for quality improvement, such as cutting down on unnecessary clinical testing, optimizing anti-microbial therapy etc. UHN embarked on building a CDW Decision 1. Now, Decision 1 is being used by the CDS system to issue clinical alerts to the clinicians if the requested investigation is a duplicate, or to suggest changing the antibiotics from intravenous to oral etc. The CDW is also being used to monitor the effectiveness and impact of the alerts, and also for monitoring resource utilization as quality improvement targets.

In terms of the diagnostic tests prescribed by the clinicians it was found that of the five tests – compete blood count (CBC), actuated partial thromboplastin time (APTT), prothrombin time (PT), blood film review, and fibrinogen that constitute 95% of all the hematology examinations, at least 10%, and sometimes as much as 25% of the tests were redundant based on widely accepted time periods. The reader is advised to consult the original paper for all the advantages derived from the use of data mining techniques coupled with a Clinical data warehouse. The future endeavors include (1) flagging high-risk patients based on the risk factors discovered by the CDW, (2) Preventing medication errors, (3) Offering cost advisement on antibiotics when lower cost alternatives exist, (4) Providing clinical reminders to clinicians to help them comply with standard protocols etc.

3.7.3 Feedback Integration

Sherbrooke University Hospital (CHUS) in Montreal has a transaction processing system called ARIANE, and also a Clinical Data Warehouse called CIRESSS. Prior to the installation of the CDW complex software was designed in order to provide feedback to the clinical teams. Grant et al. [19] discuss one such software that was designed to monitor the use of blood gas measurements in Intensive care units. Apart from the fact that such software can be immensely complex to build, there are other problems such as minimal code re-use – a lot of times software needs to be redesigned from scratch in order to cater to a different problem. Both the problems were overcome by the use of a CDW. Two dashboards were designed for the use of feedback – one for the emergency department, while the other one was designed for the clinical biochemistry department. The design process included suggestions from the end users. While the post-feedback results have not been published yet, however streamlining of the processes was observed from the day the software went into operation.

3.7.4 Data Warehousing for Disease Management

The following is after [37]. U.S. Quality Algorithms (USQA), uses a data warehouse to collect administrative data to collect administrative data from pharmacies and laboratory claims. Certain ailments such as diabetes, cardiovascular diseases, asthma etc. that respond well to disease management programs are flagged by using algorithms that examine diagnosis, lab procedures, and drugs used by the patients. These patients are then targeted for applicable member mailings, or are placed in disease management programs. Once these patients are in the program(s) they generate more data that can be analyzed further to single out patients for whom the disease was under control vs. those for whom this was not the case. The proper deployment of the successful data warehouses in disease management programs benefits both the organization as well as the member patient. Another company, Horizon Mercy found that the most common diagnosis amongst pediatric patients was asthma. Placing these patients in asthma management programs allowed the company to "make key patient interventions", and create educational programs, and to save on the high emergency room costs for this segment of the population.

3.7.5 Shared Patient Records: A Means for Conducting Nationwide Research

In this day and age the data may need to be transferred across geographically different locations before data mining can be applied on to it. Knaup et al. introduce an architecture called eardap for shared electronic patient records [29]. The architecture was implemented for pediatric oncology in Germany whereby about 20 clinical trial centers spread throughout Germany require data regarding the treatment of the each patient.

The authors claim that the architecture is extensible for new research questions, and as well it can reuse data for multiple purposes. eardap places special emphasis on the information systems of the EPR's source hospital, and also to the security issues. Multiple documentation, laboratory examinations etc. are avoided by sharing the data, and thus there is a huge savings in the cost. There are two main features for data use in eardap: (1) for general functions of EPR such as patient administration, reporting and analysis, and (2) for answering research questions such as those on therapy optimization or epidemiologic questions. Also, the amount of data can be enormous and it can be complex. It has been found that the use of eardap has resulted in a smooth process of data transfer between different research partners, and it works well in heterogeneous environments. None of this would have been possible without the use of Electronic Patient Records.

3.8 Summary

The use of Clinical data warehouses is on the rise, and health institutions, insurance companies alike are reaping the benefits in terms of reduced cost of operations, timely treatment of patients, streamlining of operations, better education and training opportunities. However, it is important to realize that the field is still in the development phase, and that many challenges lie ahead. As such there are a lot of exciting opportunities ahead. Ebadollahi [15] report a new development in the field of electronic health records. They present the idea of using concept-based multimedia health records to better organize the health records at the information level. Schabetsberger et al. [42] report on the secure regional healthcare network being developed in Austria. Sartipi et al. [41] report of a new architecture called Service Oriented Architecture that provides standards for sharing data and services; they model the components in the system in the form of work flows.

4. Challenges of Data Mining in Health Informatics

In this section, we overview a number of challenges faced in both research and practice of data mining in Health Informatics.

4.1 Heterogeneity of Health Data

Currently, there are limited or no centralized databases of health informatics data. A large portion of potentially relevant health information is not stored electronically. The fraction that is stored electronically is scattered in hundreds of small databases through different clinics, hospitals, and laboratories. This data can be in many different formats (e.g. text, image, video) and is collected from various sources, such as patient records, doctor comments, and laboratory test results.

Many of the potential applications of data mining in health informatics, discussed in previous sections, require centralized databases that integrate different formats of health data from various sources. While there is a recent push from the governments of Canada and United States for developing such centralized databases, these projects are still in infant status and many application of data mining in health informatics are not attainable until the centralized databases are more fully developed

4.2 Disconnect between computer science and medical communities

A main challenge in applying data mining techniques to health informatics is a disconnect between the computer science and medical communities. At the most basic level, while simple practical issues such as placing computers in sterile areas or training physicians to use various software packages are often assumed to be trivial in the computer science community, in realty they are very difficult challenges. For instance, one of the main reasons reported by health care professionals for not utilizing a Clinical Decision Support system was the extreme difficulties physicians had in electronic ordering and interacting with electronic records [20].

There is also a disconnect between types of learning algorithms utilized by the machine learning and data mining communities, and the algorithms that the medical community feel comfortable to use. In particular, while the data mining community is interested in applying the latest and most complex algorithms to medical datasets to achieve the highest accuracy possible, the non-academic clinicians prefer "simple, understandable models" [20]. Matheny and Ohno-Machado claim that the more sophisticated machine learning techniques (such as SVM, Neural Networks and decision trees) have limited or no representation in health informatics applications. On the other hand, simpler models, such as linear and logistic regression and scoring systems are popular. Matheny and Ohno-Machado explain the one reason for unpopularity of the more complex models is that these models are not well disseminated or well evaluated in the biomedical community.

4.3 Legal and Ethical Issues

Data ownership, fear of lawsuits, and privacy concerns are other challenges that currently constrain the extended use of data mining in health informatics. Bellow is a short summary of each issue as described by Cios and Moore [12].

- *Data ownership*: There is an unsettled question of ownership of patient data. In particular, it is unclear whether the patients, the physicians, the laboratories, or the insurance companies own the data collected from patients. There have been a number of lawsuits and congressional inquiries to address these issues [16], but the question of health data ownership is still unsettled.

- *Fear of lawsuits*: In medical communities, particularly in the Unites States, there is a fear of malpractice and other costly lawsuits that adds to the challenges of applying data mining in health informatics. Potential lawsuits, that may be triggered by discovering anomalies in patient medical histories, leave medical professionals unwilling to share patient data with researchers.

- *Privacy issues*: Protecting patient privacy and doctor-patient confidentially adds another sets of challenges to data mining in health informatics. Administrators and researchers should pay utmost attention to privacy and security when transferring, storing, or mining patient data. In many cases, patient records needs to be anonymous (*i.e.* patient identities are removed at the time of information collection), anonymized (that is patient identities are removed after the data is collected), or de-identified (*i.e.* patient identities are encrypted and can be restored under certain institutional policies).

5. Conclusion

We have provided an overview of applications of data mining in administrative, clinical, research, and educational aspects of Health Informatics. We established that while the current practical use of data mining in health related problems is limited, there exists a great potential for data mining techniques to improve various aspects of health Informatics. Furthermore, the inevitable rise of clinical data warehouses will increase the potential for data mining techniques to improve the quality and decrease the cost of healthcare.

References

[1]American Medical Informatics Association, http://www.amia.org/informatics/.

[2] Canada's Health Informatics Association, http://www.coachorg.com/.

[3] National Library of Medicine, http://www.nlm.nih.gov/tsd/acquisitions/cdm/subjects58.html.

[4] Canadian Institute of Health Research, http://www.mshri.on.ca/colorectalcancer/definitions.html, 05/25/2008

[5] Homer Learning Community, https://homer.med.ualberta.ca/.

[6] http://www.info-source.us/data_warehousing_mining/Data-Mining-and-Data-Warehousing-in-Biology-Medicine-and-Health-Care/image004.jpg

[7] Allard R.D. "The clinical laboratory data warehouse – An over-looked diamond mine". Am. J. Clin. Pathol. 817-819, 2003.

[8] Batini C., Ceri S., Navathe S. "Conceptual Database Design: An Entity-Relationship Approach". AddisonWesley. Spanish, 1991; ISBN 0-201-60120-6 .

[9] Berner E., "Clinical Decicion Support Systems". Springer Science+Business Media, 2007 .

[10] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D., "Knowledge-based Analysis of Microarray Gene Expression Data using Support Vector Machines". Proceedings of the National Academy of Sciences, 2007.

[11] Chen Y., Henning Pedersen L., Wesley W. Chu, Olsen J., "Drug Exposure Side Effects from Mining Pregnancy Data". ACM SIGKDD Explorations Newsletter, 2007.

[12] Cios K., Moore GW., "Uniqueness of Medical Data Mining". Artificial Intelligence in Medicine, 2002.

[13] Cooper G F., Dash D H., Levander J D., Wong W K, Hogan W R., Wagner M M., "Bayesian Biosurveillance of Disease Outbreaks". ACM International Conference Proceeding Series; Vol. 70, 2004.

[14] Courtright C., Crawford R. Klubert D. "Criteria for Developing Clinical Decision Support Systems". CBMS '01: Proceedings of the Fourteenth IEEE Symposium on Computer-Based Medical Systems, 2001.

[15] Ebadollahi S., Coden A. Tanenblatt M., Chang S., Syeda-Mahmood T., Amir A. "Concept-based electronic health records: opportunities and Challenges". MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia, 997—1006, 2006.

[16] Fienberg S., "Sharing Statistical Data in the Biomedical and Health Sciences: Ethical, Institutional, Legal, and Professional Dimensions". Annual Review of Public Health, Vol. 15: 1-18, 1994

[17] Frawley W., Piatetsky-Shapiro G., Matheus C. “Knowledge Discovery in Databases: An Overview”. AI Magazine. Vol 13, number 3, 57-70, 1992

[18] Geof McMaster, “Goodbye old school, hello HOMER”, Express News, University of Alberta, January 24, 2008. Available at: http://www.expressnews.ualberta.ca/article.cfm?id=9025

[19] Grant A., Moshyka A., Diaba H., Carona P., Lorenzia F., Bissona G., Menarda L., Lefebvrea R., Gauthiera P., Grondinb R., Desautelsb M. “Integrating feedback from a clinical data warehouse into practice organisation” . International Journal of Medical Informatics. Volume 75, Issues 3-4, Pages 232-239, March-April 2006

[20] Gray G. “Challenges of building clinical data analysis solutions”, Journal of Critical Care Volume 19, Issue 4, December 2004, Pages 264-270

[21] Greens R., "Clinical Decision Support". Elsevier Inc., 2007.

[22]Hackney D. “Understanding and Implementing Successful Data Marts”. Addison-Wesley Longman Publishing Co., Inc., 1997

[23]Inmon W. “What is a data Warehouse?”. Sunnyvale Calif. : Prism Solutions Inc., 1995

[24] Izadi MT, Buckeridge DL, "Decision Theoretic Analysis of Improving Epidemic Detection". American Medical Informatics Association, 2007.

[25]Jamtvedt G., Young J., Kristoffersen D., O'Brien M., Oxman A.. “Audit and feedback: effects on professional practice and health care outcomes”. Cochrane Database of Systematic Reviews 1998.

[26] Johnson JA., Bootman HL., Drug-related morbidity and mortality: a cost of illness model. Arch Intern Med 1995; 266:2847-2851.

[27] Kim J., Feng D., Cai T., Eberl S. “A solution to the distribution and standardization of multimedia medical data in E-Health”. Proc. Pan-Sydney Area Workshop on Visual information Processing - Volume 1, 2001.

[28] Kimball R., Ross M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd edition Wiley, New York, 2002.

[29] Knaup P., Garde S., Merzweiler A., Graf N., Schilling F., Weber R., Haux R. “Towards shared patient records: An architecture for using routine data for nationwide research”. International Journal of Medical Informatics, Volume 75 , Issue 3 - 4 , 191 – 2005.

[30] Kohn LT., Corrigan JM., Donaldson MS., eds. To err is human. Washington D.C.: National Academy Press: 1999.

[31] Kuperman GJ, Gardner RM, Pryor TA, "HELP: A dynamic hospital information system". Springer-Verlag, 1991

[32] Ledbetter C. Morgan, M. “Toward Best Practice: Leveraging the Electronic Patient Record as a Clinical Data Warehouse”. JOURNAL OF HEALTHCARE INFORMATION MANAGEMENT, VOL 15; PART 2, pages 119-132, 2001

[33] Leonard JE, Colombe JB, Levy JL, "Finding relevant references to genes and proteins in Medline using a Bayesian approach". Bioinformatics Vol. 18, no. 11, 2002.

[34] Nykänen P., "Decision Support Systems from a Health Informatics Perspective". Tampere, 2000.

[35] Ordonez C., " Comparing association rules and decision trees for disease prediction". Proceedings of the international workshop on Healthcare information and knowledge management, 2006.

[36]Prather J, Lobach D., Goodwin L., Hales J., hage M., Hammond W. Medical data mining: knowledge discovery in a clinical data warehouse". Proc AMIA Annu Fall Symp. 1997:101-5.

[37] Ramick, D. C. "Data Warehousing in Disease Management Programs". JOURNAL OF HEALTHCARE INFORMATION MANAGEMENT, VOL 15; PART 2, pages 99-106, 2001.

[38] Romero C., Ventura S., "Educational Data Mining: A survey from 1995 to 2005". Expert Systems With Applications, 2007.

[39] Sackett D., Rosenberg W., Gray J., Haynes R., Richardson W. "Evidence based medicine: what it is and what it isn't". BJ 312 (7023): 71-2.

[40] Sahama T., Croll P. "A data warehouse architecture for clinical data warehousing" ACSW '07: Proceedings of the fifth Australasian symposium on ACSW frontiers, 227—232, 2007.

[41] Sartipi K., Yarmand M.H., Down D.G., "Mined-Knowledge and Decision Support Services in Electronic Health". SDSOA '07: Proceedings of the International Workshop on Systems Development in SOA Environments, 2007.

[42] Schabetsberger T., Ammenwerth E., Andreatta S., Gratl G., Haux R., Lechleitner G., Schindelwig K., Stark C., Vogl R., Wilhelmy I. "From a paper-based transmission of discharge summaries to electronic communication in health care regions". International Journal of Medical Informatics, Volume 75, Issue 3 - 4 , 209 – 215.

[43] Sen A., Sinha A. "A comparison of data warehousing methodologies", Commun. ACM, Vol 48, no 3, 79—84, 2005

[44] Shortliffe EH, "Computer programs to support clinical decision making". JAMIA 258, 1987, 61-66.

[45]Stolba N, A Min Tjoa. "The Relevance of Data Warehousing and Data Mining in the Field of Evidence-Based Medicine to Support Healthcare Decision Making". Enformatika, Volume 11, 12 – 17.

[46] Verma, R. Harper, J., "Life Cycle of a Data Warehousing Project in Healthcare". JOURNAL OF HEALTHCARE INFORMATION MANAGEMENT, VOL 15, PART 2, pages 107-118, 2001

[47] Zhang C., Chen WB, Yang L., Chen X., Johnstone JK., " Automatic in vivo Microscopy Video Mining for Leukocytes". ACM SIGKDD Explorations Newsletter, 2007