Clinical Data Mining: a Review

J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, A. Geissbuhler
University and Hospitals of Geneva, Switzerland

Summary

Objective: Clinical data mining is the application of data mining techniques using clinical data. We review the literature in order to provide a general overview by identifying the status-of-practice and the challenges ahead.

Methods: The nine data mining steps proposed by Fayyad in 1996 [4] were used as the main themes of the review. MEDLINE was used as primary source and 84 papers were retained based on our inclusion criteria.

Results: Clinical data mining has three objectives: understanding the clinical data, assist healthcare professionals, and develop a data analysis methodology suitable for medical data. Classification is the most frequently used data mining function with a predominance of the implementation of Bayesian classifiers, neural networks, and SVMs (Support Vector Machines). A myriad of quantitative performance measures were proposed with a predominance of accuracy, sensitivity, specificity, and ROC curves. The latter are usually associated with qualitative evaluation.

Conclusion: Clinical data mining respects its commitment to extracting new and previously unknown knowledge from clinical databases. More efforts are still needed to obtain a wider acceptance from the healthcare professionals and for generalization of the knowledge and reproducibility of its extraction process: better description of variables, systematic report of algorithm parameters including the method to obtain them, use of easy-to-understand models and comparisons of the efficiency of clinical data mining with traditional statistical analyses. More and more data will be available for data miners and they have to develop new methodologies and infrastructures to analyze the increasingly complex medical data.

Keywords

Medical records systems, computerized; data mining;

Geissbuhler A, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2009. Methods Inf Med 2009; 48 Suppl 1: xxx

Introduction

The term data mining becomes ever more popular in the biomedical research community. Within the last 10 years, the number of papers having the term "data mining" in their title and referenced in MEDLINE has increased by a 10-fold. The activities of interdisciplinary researchers to promote the clinical data mining techniques are likely to be one of the reasons of this explosion. The Intelligent Data Analysis and Data Mining (IDA-DM[1]) workgroup of the International Medical Informatics Association (IMIA), for example, has organized a yearly workshop on intelligent data analysis and data mining in biomedicine and pharmacology since 1996. This workshop, called IDAMAP[2], is an opportunity for researchers and practitioners to meet and discuss data analysis techniques for the biomedical domain.

The purpose of data mining is a methodic extraction of knowledge, patterns, useful information, or trends from retrospective, massive, and multidimensional data. However, there are also some studies using real-time data [1]. Three main categories of data mining strategies are reported in the literature: supervised, unsupervised, and semi-supervised learning [2, 3]. In a supervised learning setting, a set of input variables is used to predict a target or dependant variable. In an unsupervised learning setting, the target variable does not exist and data mining techniques are used to discover patterns, clusters, or relationships in the data set. In semi-supervised learning, the target variable exists but the value is only provided for a small amount of the examples and data mining techniques are used to predict the values of missing target values or extract patterns, clusters, or relationships in the data set. The final objective of data mining can be grouped into five categories: prediction (for supervised and semi-supervised learning only), regression, classification, exploration, and affinity (unsupervised and semi-supervised learning only). The reader could refer to [2] for an overview of data mining algorithms for each data mining objective, both for supervised and unsupervised learning, and [3] for semi-supervised learning.

Data mining is a multidisciplinary field at the intersection of database technology, statistics, machine learning, and pattern recognition. It profits from all these disciplines. These multiple roots that are characteristic of data mining introduce differences in the terminology used and the process for data analysis. In a medical setting, for example, the patients' characteristics (sex, age, risk factors, etc) may be called columns or fields for a data miner having a strong database knowledge; independent variables for a statistician; features for a machine learning specialist. Data mining is the application of specific algorithms for extracting patterns from massive data and is only a step in the knowledge discovery in databases (KDD) process [4] but, in practical settings, it is broadly assimilated as the whole KDD process [5].

The KDD process, as described in [4] is an iterative and interactive process having nine steps to extract latent information in data: 1) the business un-

[1] http://magix.fri.uni-lj.si/idadm/ (available online on the 31/01/2009)
[2] http://www.idamap.org/ (available online on the 31/01/2009)

derstanding, 2) data set selection, 3) data cleaning and preprocessing, 4) data reduction and projection, 5) matching the objective defined in step 1 into a data mining method (classification, clustering, regression, etc.), 6) choice of the algorithm and search for data patterns, 7) pattern extraction, 8) data interpretation and 9) use of the discovered knowledge. Some of these steps (3, 4, 9) may be optional in a data mining process based on the quality of the data at hand, algorithm at hand, and the final objective. It is important to note the industries' initiatives to standardize the KDD process. Many processing models were proposed to achieve successful data mining: the five A's (Assess, Access, Analyze, Act and Automate) from SPSS, the SEMMA[3] (Sample, Explore, Modify, Model, Assess) from SAS, the CRISP-DM[4] (Cross-Industry Standard Process for Data Mining) from a consortium of vendors and users (Clementine® for example), and the two-crows model[5]. A detailed description and comparison of the SEMMA and CRISP-DM processing models can be found in [6].

Objective

The objective of this survey is to provide a general overview of clinical data mining by identifying the state-of-the-art practices and the challenges ahead. The term "clinical data mining" indicates the data mining application on clinical problems. In the following, we use the broad definition of data mining and papers will be reviewed and discussed according to these steps. These steps are important for the reproducibility of the process to obtain the knowledge, which is the main concern of medical experts with respect to the concept of knowledge [7].

During the past ten years, several reviews on clinical data mining were produced in the biomedical domain. This review will study the actual practice of data mining on clinical data and their main differences: it is not limited to the review of algorithms used in medical domain according to the data mining objectives such as in [8]; it is not focused on a specific data mining algorithm such as in [9]; it studies the effective application of data mining on clinical problems not focused on a particular type of data such as the textual ones already addressed in [10], or the temporal aspect of the data such as in [11, 12], nor on a particular disease or specific task like pharmacoepidemiology as in [13]. A review of existing open-source data mining tools can be seen in [14]. A methodological review of the predictive data mining in clinical medicine is proposed by Bellazzi et al. with a focus on current issues coupled with practical guidelines for a better construction, assessment and exploitation of data mining models [15].

Methods

The main source of information for this review is MEDLINE. Google Scholar® was used as a complementary information source. For MEDLINE searches, we used a combination of MeSH terms related to data mining because the term "data mining" is not referenced in MeSH. Narrower terms of data mining such as "artificial intelligence", "pattern recognition, automated", "cluster analysis", "bayes theorem", "classification", "data interpretation, statistical", "information storage and retrieval/methods", "models, statistical", "Decision Trees" as major topic and "medical records systems, computerized". The existence of an abstract, written in English, published between 1998 and December 2008 was our inclusion criteria. Papers having one of the MeSH keywords cited above but not based on a data mining methodology (pure statistical data analysis, papers dealing using other methodology of artificial intelligence, etc.), or not accessible in full format were excluded from the study.

Results

A total of 84 papers were included in this review and from these we extracted the status of the nine data mining processes and the challenges for the domain of clinical data mining.

1. Learning the Application Domain

The objective of this first step is to determine 1) the relevant prior knowledge of the domain and 2) the goal of the data mining application.

The prior knowledge of the domain can be seen as the problem that requires a response or solution by applying data mining. The problem can be related to the clinical data. In this case data mining may provide a better understanding of the data related to the clinical care, in which data mining may provide assistance to medical staff or related to the data mining methodology for its improvement or adaptation in the medical domain. Bayat et al., for example, wanted to obtain from a renal transplantation database the most important risk factors for patient incorporation in a renal transplantation waiting list [16]. García-Gómez et al. investigated the contribution of data mining in the differential diagnosis of musculoskeletal soft tissue tumors for the radiology department [17]. Juhola and Laurikkala developed a methodology to measure the similarity of clinical cases, having mixed-type variables (quantitative and nominal), in order to classify similar elements [18]. The latter was applied on female urinary incontinence and otoneurological data.

After the definition of the problem, the goal of data mining applications has to be stated. According to the descrip-

[3] http://www.sas.com (available online on the 31/01/2009)
[4] http://www.crisp-dm.org (available online on the 31/01/2009)
[5] http://www.twocrows.com (available online on the 31/01/2009)

tion of the problem, the potential uses of the data mining processes in medical application are described in the following paragraphs. Table 1 provides examples for each type of problem.

i) Understanding Clinical Data

Clinical information systems are mainly used to manage a patient-by-patient electronic health record. The clinical data warehouse is the source for analysis at a patient population level in order to better understand the care practice and its effectiveness [19]. Data mining can be used to explore the insight of these data. There are three main approaches for this purpose: data visualization, data exploration, and data quality assessment.

Data visualization approaches tend to provide quick and understandable access to information i.e. converts data into information [1, 20, 21]. Grant et al. proposed a dashboard system in clinical practice to visualize (retrospectively) the resource allocation in the emergency department and to evaluate the quality of the service provided by the biochemistry department [20]. The content of an electronic health record is a succession of events collected over time. There are situations where the time of each event is of high importance. Visualizing time-oriented data of a patient population is a challenging task and it was proposed by Klimov and Shahar in [21]. While most of data mining applications analyze retrospective data, Chen et al. proposed a real-time data summarization by parsing hospital communication messages [1].

Data exploration provides the tacit relationship in the clinical data [5, 16, 22, 23, 24, 25, 26]. These relationships are expressed as rules in the form of IF conditions THEN conclusion [24, 25, 26]; risk factors identification for a specific disease [5, 16]; and subgroup characterization [22]. In the risk factor identification, the most important variables providing a high correlation with a target or dependant variable are selected while the subgroup discovery provides the characteristics of the (independent) variables concerning a specific value of a dependant variable of interest (most of the time over or under represented).

The clinical data are collected to support clinicians in the care delivery and it contains noise, contradiction, missing values [27], and important information (signs, symptoms, clinical reasoning, and so on) may be stored in an unstructured way. The missing values, for example, may be due to a neglect of the clinicians because they thought that the variable is of less importance for a particular patient [22, 28]. Some clinical parameters are considered as clinically normal and not confirmed or simply omitted due to time pressure [29]. The third data mining application in relation to clinical data problems is the assessment of clinical data quality [30, 31, 32, 33, 34]. The goals of this application are multiple such as detection of medical errors, assessment of the data coding quality, provision of a structure from unstructured data, etc. Jannin and Morandi, for example, use data mining to assess the quality of a surgical procedure model [31]. Spangler et al. investigated the adequacy and effectiveness of two coding classifications (ICD-9 for diagnostic and CPT for procedure) in two hospitals [34]. Chapman et al. proposed a method for extracting a clinical concept from emergency department reports [35] and Goldstein et al. classified automatically radiology reports with ICD-9-CM classification [36].

Table 1 Papers classified according to the data mining purpose in medical applications

Data mining goal	Papers
Data exploration	[5, 16, 22, 23, 24, 25, 26, 33, 64]
Data visualization	[1, 20, 21, 97]
Data quality assessment and evaluation	[30, 31, 32, 33, 34]
Prognostic evaluation	[39, 40, 41, 42, 43, 45, 46, 77]
Diagnostic assistance	[17, 32, 39, 89]
Quality of Care	[24, 31, 48, 49, 51, 52, 53, 54, 66, 76, 78, 81, 86, 93]
Information retrieval	[56]
NLP	[35, 36, 55, 57, 70, 82, 90, 98]
Image analysis	[39, 45]
DM Methodology evaluation	[5, 18, 19, 45, 62, 99]

Table 2 Papers classified according to the data type

Data type	Papers
Structured	[1, 5, 16, 18, 20, 23, 25, 26, 29, 30, 32, 40, 42, 43, 50, 53, 54, 63, 64, 69, 73, 75, 84, 85, 100]
Structured + temporal dimension	[51, 52, 59, 60, 86]
Unstructured	[19, 33, 35, 36, 39, 44, 45, 46, 55, 56, 57, 66, 68, 70, 82, 83, 98]
Mixed type	[19, 22, 41]
Benchmark datasets	[39, 57, 70]
Data from multiple sources	[16, 26, 32, 42, 43, 54, 63, 64]

ii) Assistance in Clinical Care

In the 70s and 80s, many expert systems such as MYCIN [37] or INTERNIST [38] were developed based on knowledge provided by medical experts. However the acquisition of the knowledge used by these systems has a high cost and artificial intelligence was used to extract knowledge in the clinical

data. These systems were designed for one of the following purposes: prognostic assistance, diagnostic assistance, quality of care improvement, and support to access clinical data.

The prognostic assistance took a form of an early detection of high risk patients in a patient population [39, 40, 41, 42, 43, 44, 45, 46]. Gellerstedt et al., for example, provided a support for pre-hospitalized patients suffering from acute myocardial infarction at the emergency department dispatch center using subjective information provided by phone [43] and Goletsis et al. proposed a system for early detection of high risk patients suffering from myocardial ischemia using electrocardiographic data [44]. Daemen et al. compared the breast cancer prognosis model provided by clinical versus genetic data [41].

Data mining may also be used to provide diagnostic assistance for rare events like the otoneurological diseases [28, 47]; for difficult diagnosis such as the musculoskeletal soft tissue tumors [17, 48]; for automatic digital image reading [39].

The improvement of the quality of care concerns mainly support for patient safety like in [29, 31, 49, 50, 51, 52]. It can also take the form of an improvement of the administrative work of healthcare staff as in [24, 48, 53]. Data mining can also provide a measure for the population's access to healthcare resources such as suggested in [54].

In order to better deal with clinical data, data mining techniques have developed solutions to support the retrieval of narrative or imaging documents [35, 55, 56], to classify medical reports automatically [36], and to preserve the patient privacy and confidentiality in medical reports for a secondary usage [57].

iii) Data Mining Methodology Development

Harper compared different classification algorithms for decision making in [58] and concluded that there was no single best classification tool but the best performing algorithm depends on the features of the data at hand as well as any preference of the end-user. However, due to the complexity of medical data, it is sometimes necessary to adapt existing algorithms or optimize their use to obtain better results. Hripcsak et al., for example, proposed and evaluated a new distance metrics for narrative clinical data for clustering [19]. Juhola and Laurikkala proposed a new distance measure in the case of mixed-type variables (quantitative and nominal) for classification [18]. Ramoni and Sebastianini proposed a new version of the Naïve Bayes classifier to handle missing values automatically [29]. In essence, data mining exploits a large number of variables and measurements. Computational efficiency and scalability are important issues. To address this problem, Huang et al. investigated a new feature selection algorithm to reduce the computational complexity of data mining [5]. The heterogeneity of the medical data prompts medical data miners to develop new approaches to analyze data. Jesneck et al., for example, investigated decision fusion as a strategy for the classification of imaging data from multiple modalities, multiple sources and having various types of features [45]. The analysis relationships of time-stamped or time series clinical data exploits the temporal abstraction mechanism (identification of time interval in which a specific data pattern occurs) [51]. The introduction of a knowledge base represented in an ontology was introduced by [59] and [60] in order to improve the mining of temporal associations in clinical data.

2. Creating a Target Dataset

i) Structured Data Creation

The second step in a data mining project is the creation of the dataset for analysis. This can be performed by consulting medical knowledge sources to identify relevant variables for the analysis [61] or with the objective help of experts of the domain. Multiple experts are needed if the acquisition of domain knowledge has a high cost [26]. The data used in a data mining study is a two-dimensional data matrix (dataset) where the columns represent the variables and each row a case. In this paragraph, we will use two terms to characterize a data matrix: the width to characterize the number of variables and the height for the number of cases as shown in figure 1. A wider (respectively long) dataset is a dataset having a high number of variables (respectively cases). Usually, the clinical data warehouse organized in a relational database is the principal source of data as in [5, 17, 18, 19, 20, 22, 23, 25, 28, 30, 31, 48, 49, 61, 62]. Most of the studies use only a subset of the warehouse data, i.e., a subset of data agreeing with the problem described in the first step. Alvarez et al., for example, selected the data of pediatric emergency patients undergoing an ultrasound for pyloric stenosis to evaluate the predictive power of a Bayesian classifier and had a dataset of 118 patients [49]. The complexity of the query to create the dataset induces short datasets (the number of cases rarely goes beyond 1'000) and supposes the use of a prior hypothesis/knowledge of the domain (e.g. the variables to use, cases to incorporate in the dataset, etc.).

As the objective of data mining is to extract hidden knowledge in a database, one step to obtain an objectively valid

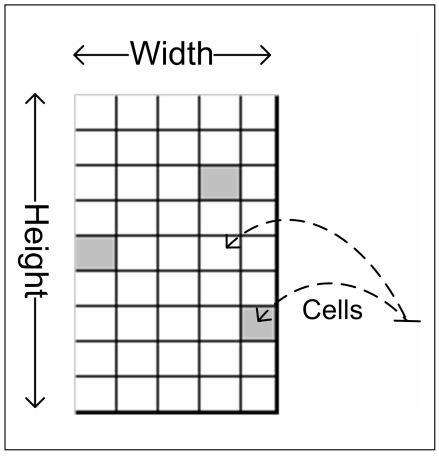

Fig. 1 Characteristics of a data matrix. Blank cell (resp. grey cell) corresponds to a filled (resp. missing) value of a particular variable for a particular case.

knowledge is to have the longest dataset possible. Two main solutions are used to obtain a higher number of cases: extract data accumulated over a long period of time as in [5, 19, 25, 30, 40, 48, 61] or integrate data from multiple sources [16, 42, 54, 63]. The latter may also be used to obtain a wider dataset [26, 43, 64]. Two studies [25, 30] analyze the whole clinical data warehouse for exploration and to extract new knowledge inherent in the data. Pregnant women data collected from a 20 year-long perinatal database was used in [30] and the whole clinical data of 13 years were analyzed in [25].

In essence a data warehouse should contain clean and coherent data. However, these data were not collected to support a knowledge extraction but to support patient-by-patient clinical care. Due to inter-patient variability, missing values are frequent in datasets [5, 18, 26, 28, 48]. However, the main advantage of the use of the clinical data warehouse is its organizational structure and coherence of the data with respect to information coding.

ii) Unstructured Data

Sittig et al. highlighted that 50% of clinical information describing the patient's condition during the therapy is stored in an unstructured way in free-text reports [65]. Analyzing free-text in order to extract information and presenting them in a structured form is a text mining task. In some cases, data mining is used to analyze reports like physical examination letters or discharge letters from various medical units as in [19, 36, 46, 56, 66]. The unstructured data is converted into a (structured) data matrix by means of natural language processing. A column here represents a concept, a line a report of a specific patient, and a cell the presence or absence or the occurrence of the concept in the patient's report. Data mining algorithms can be applied to the dataset for a specific objective. Sometimes, the reports do not contain important information such as diagnoses and these can be manually labeled by medical experts such as in [66] and [56], or associated with structured data to improve the knowledge extraction [32]. Reports are not the unique kind of unstructured information sources in a healthcare setting. Indeed, images [39, 45, 67, 68] and signal data such as those from electrocardiographic (ECG) data [44, 67] or microarray data [41] must also be processed to extract structured features. The main characteristic of unstructured data such as text and images is their multiplicity for a single patient. Another possible challenge is the heterogeneity of these multiple unstructured data instances [45].

iii) Data Availability

Creating a target dataset may be time consuming especially if the important variables to study have to be inferred from multiple variables from multiple sources [69]. Patient privacy and confidentiality is also an important issue when the data are used out of their initial collection purpose [25, 30, 39, 70]. Some research use publicly available datasets as those provided by universities like the UCI[6] repository or those provided by learned societies such as the Mammographic Image Analysis Society [71] or ImageCLEF[7]. These publicly available datasets serve as benchmarks to evaluate the performance of newly developed algorithms.

3. Data Cleaning and Preprocessing

The goal of the data cleaning and preprocessing step is to transform the dataset in order to remove inconsistencies, noise, incoherence, bias and redundancies characterizing medical data [67, 72] to avoid the "garbage in, garbage out" phenomenon. These transformations fall into two categories: a vertical transformation concerning the variables and their values and a horizontal transformation concerning the cases. This step needs more effort than the data mining algorithm application itself because it requires an analysis of the dataset, which is not always possible to automate. Moreover these analyses are not well documented in papers although being crucial for the success of a data mining project. This step is guided by the issues inherent in the dataset created in the second step described above.

i) Horizontal Transformation

During the dataset creation, all variables related to the domain are collected according to knowledge of the domain. The number of these variables varies from 4 as in [42] and [73] to 11'118 like in [46].

Data type issue

The variables may have different types: binary, nominal, ordinal or numerical. Some data mining algorithm can handle all these data types but sometimes there is a need to transform the type of the variables due to algorithmic limitation or to ease the final interpretability of the results. The transformation can be done manually or supported by the data mining software such as in [25] because the algorithm can only handle nominal values.

Variable domination issue

Variables with numerical values may introduce bias into the dataset when some variables dominate the others. The bias may influence the data mining algorithm decision attributing too much importance to the dominating variable. A variable dominates another one if it has large values compared to the second. Normalization of the numerical values into the interval [0, 1] is the solution in this situation [73]. Another approach to resolve the variable domination is the discretization of numerical variables as in [61].

[6] http://archive.ics.uci.edu/ml/ (available online on the 31/01/2009)

[7] http://imageclef.org (available online on the 31/01/2009)

Abstraction issue
It is also possible to have a set of variables expressing the same concept and aggregating them into a single variable removes redundancies and may prevent errors in the dataset. Lin and Haug, for example, added metadata describing the structured data variables and domain knowledge to aggregate variables automatically [61].

Temporality issue
The temporality of the clinical data may also introduce a bias in a dataset. Some clinical parameters are recorded in the electronic health record of the patient multiple times during its stay in the hospital and the representation of these multiple-valued variables in a data mining matrix may raise inconsistencies in the dataset. There are two approaches to represent the dataset for a classification purpose: use a single value or an aggregated (e.g. mean, mode) value if the temporality concerns a small subset of variables [61] or consider each entry as a case if the temporality concerns most or the variables [5]. For a temporal association analysis, the time series data have to be transformed into an interval-based representation by the mean of the temporal abstraction mechanism [51, 59, 60].

Missing values issue
As we have seen in the above paragraphs, having missing values is one of the properties of medical data and it reflects the situations of clinical care due to the patients' variability. Lin and Haug proposed four approaches to deal with missing values: use the data as it is i.e. the algorithm used should support missing values like those proposed by Ramoni et al. [29], impute the missing values, add an explicit value indicating a missing value (for categorical variables only) and add a new column [48]. The treatment of variables having missing values depends also on the final objective of the data mining application. Some studies remove variables having a high rate of missing values (>50% in [5]), other studies infer the missing values [28, 48, 73]. Richards et al., in a rule extraction problem, highlighted that keeping the features with an important rate of missing values may not be necessary because they might be important for rare cases [26].

Privacy and confidentiality issue
The patients' data are collected to primarily support clinical care for this single patient [67]. When the data is used for secondary purposes information permitting to directly identify the patient has to be removed from the dataset [67, 74]. Few studies talk about the manner they handle the issue. Mullins et al. de-identify the data according to the US Health Insurance Portability and Accountability Act (HIPAA) regulation [25]. Goldstein et al. de-identified the free-text reports during the pre-processing step [36]. It is important to highlight that there is a possibility to re-identify individual patients, if required, with a de-identified data and in opposition to anonymized data [30].

ii) Vertical Transformation

As the clinical data mining objective is to extract inherent knowledge in clinical databases, case selection varies according to the final objective of the study and may need knowledge of the domain such as in [5, 16, 17, 19, 25, 31, 33, 35, 40, 49, 53, 54, 64, 75, 76, 77]. The success of clinical data mining depends on a rigorous and documented sampling strategy out of the concern of having a reproducible process of the knowledge extraction and its validity when applied on new cases. García-Gómez et al., for example, selected all patients having confirmed musculoskeletal soft tissue tumors in order to build a model grading them into benign/malignant cases [17]. Chapman et al. selected reports of patients admitted at the emergency department randomly [35].

Missing value issues
The problem related to missing values can also be solved by removing cases with missing values especially if the data mining algorithm cannot handle them [73]. When the data mining algorithm can be applied even if missing values are present, Bilska-Wolak and Floyd showed that it is more beneficial to leave them out [78]. Hripcsak et al. highlighted that reducing the analysis to only accurate and complete records may also introduce bias in the dataset [19]. However, missing values can also be inferred with an analysis of the data at hand like in [26] or filled by the data mining algorithm during its application [32].

Outlier issues
Another characteristic of medical data is the presence of outliers due to errors [61] or simply rare cases. Even if the cases selected for analysis should represent the real world situation, outliers should be removed because they introduce noise in the dataset and may break the patient privacy and confidentiality [30].

4. Data Reduction and Projection

The number of variables in a dataset ranges from 4 to more than 10'000. Table 3 provides an overview of 25 datasets with respect to the number of variables, cases, their type and their origin. The variables in a created target dataset are based on the domain knowledge. Computational efficiency is one of the concerns of data mining and data mining researchers are influenced by Occam's Razor principle, which can be interpreted as "simpler is better". If m variables can provide equal or better results than n variables, where m<<n, it is straightforward to use the reduced dataset with m variables to reduce computational cost. Variable reduction may be needed by the data mining algorithm: Autio et al., for example, highlighted that the "basic" multilayer perceptron provides worse results when the number of variables was above a certain threshold [73]. Data reduction permits also to reduce ambiguous variables from the dataset [36]. For medical professionals, reducing the dataset

to the most important variables permits to identify the most important factors for a problem [5] and in their daily practice, they perform ad-hoc variable selection due to time pressure [5, 79].

Four approaches were proposed to reduce the dimensionality of the dataset. The first approach is to use an existing knowledge source such as an ontology or terminology such as SNOMED [19, 22]. The background knowledge source provides, among others, weight of the variables, their similarity for an aggregation, constraints on the variables, etc. The second approach is a subjective method where the domain experts perform the feature selection [20, 42]. This approach is applicable if the number of initial variables is not high. Five experts, for example, were interviewed to obtain their judgment of the most important variables [42]. The third approach uses data mining algorithms to evaluate the predictive power of each attribute or a combination of them [5, 69, 73]. In [73], the algorithmic results were also validated by the domain expert. The fourth approach uses information measures like the information gain to select relevant features [26, 69]. This last approach measures the importance of each variable with respect to a specific target but it cannot detect the combination effect of a subset of variables.

Table 3 Example of database characteristics showing the number of variables (initial and after reduction), cases (initial and after preprocessing, database type (Structured S or Unstructured U or Mixed M), Data from multiple sources or not

Variables		Cases		Type	Multiple Sources	Papers
Initial	After reduction	Initial	After Pre-processing			
4	2	279		S	YES	[42]
4		7738		U		[82]
5		76989		U		[44]
5		17974		S		[58]
7		582		S		[58]
7		153619		S		[54]
8		503		S	YES	[43]
8		493		S	YES	[43]
9		699		S		[62]
12		530		S		[23]
13		270		S		[62]
14		4056		S		[58]
16		671		S		[78]
16		2402		S		[58]
18		208		U		[39]
21		101		S		[23]
38		815		S		[28]
38		815		S		[18]
112		667000		S		[25]
166	55	21000	3971	S	YES	[26]
208		667000		S		[25]
410	47	3857		S		[5]
781	45	4358		M		[22]
4000	1622	71753	19970	S		[30]
11118		7620		U		[46]

5. Choosing the Function of Data Mining

The objective of this step is to map the initial objective onto a data mining method. If the final goal of the data mining application is variable selection, this step is not included in the data mining process except for the algorithmic-based variable selection. Fayyad et al. provided a list of data mining methods: classification, regression, clustering, summarization, dependency modeling, change, and deviation detection [80]. The following paragraphs discuss the data mining function used with clinical data. An overview of selected data mining functions found in the literature is presented in table 4.

Table 4 Overview of data mining function found in the papers

Data mining function	Papers
Classification	[5, 17, 18, 19, 28, 29, 31, 32, 34, 35, 36, 39, 41, 43, 44, 45, 48, 49, 50, 55, 56, 58, 62, 68, 69, 70, 73, 77, 78, 81, 82, 97]
Clustering	[1, 31, 33, 76, 83]
Temporal association rules	[51, 59, 60, 86]
Summarization	
• Association rules	[22, 24, 36, 93, 101]
• Rule extraction	[25, 26, 27, 32]
• Data visualization	[20]
• Feature selection	[83]
Dependency modeling	[5, 16, 27, 28, 35, 46, 48, 49, 56, 61, 62, 64, 82, 84]
Subgroup discovery	[22, 42, 52]

Classification is a supervised learning whose goal is to find a mapping function from a set of variables to a target discrete variable. It is the most popular application of data mining found in the papers especially for assistance in clinical care. It was, for example, used to improve diagnostic accuracy [17, 28, 49, 78, 81], to detect high-risk patients early [41, 43, 44, 45, 77], and to extract concepts in unstructured free-text data [35, 48, 82]. The classification may be a binary classification (the target variable has two values) or a multi-class classification such as in [19, 24, 28, 39, 55, 56] (613 classes in [19]). In a supervised learning, when the target variable is continuous, the data mining method is called regression (not to confound with classification using logistic regression). There were no papers dealing with regression in our results set analyzed for this article. There are classification algorithms based on numerical distance measures among cases and an issue may rise with categorical data and a dataset having missing values. Juhola and Laurikkala proposed a new distance measure to fix the issue for mixed-type (quantitative and categorical) datasets having missing values [18].

Clustering is unsupervised learning of which the goal is to find interesting structures in data to support clinical data understanding or for assistance in clinical care. Bernstein et al. use this data mining function to reduce proprietary antibiotic prescriptions in an emergency department [76]. Chen et al. use this function to categorize radiology information to simplify data visualization [1]. Jannin and Morandi use the clustering function to group similar surgical procedures based on the patients' characteristics [31].

Summarization is mainly used for dataset understanding and comprises association rules detection, rule extraction, data visualization and variable selection (or feature selection). The association rules find all frequent relationship among variables and the pattern is represented in the form of (A,B,C) where A, B and C are variables while formal rules extract associations of conditions and a conclusion and the pattern takes the form of "IF A and/or B THEN C" [25, 26, 27, 32]. In the rule pattern, each variable is evaluated as a target or conclusion. The data visualization summarizes information using statistical properties of the data [20]. The variable selection permits to rank and select the most important variables by measuring their predictive power (classification problem) with respect to an outcome [83].

Dependency modeling consists of finding relationships in the variables. Most of the studies implementing this function have the objective to compare the performance with other classification functions [5, 27, 28, 35, 46, 49, 64, 82]. The main reason motivating comparisons is to improve results obtained in a previous study. Bayat et al. used a dependency function to rank and select important features associated with a renal transplantation waiting list [16]. Richardson and Domingos tested the incorporation of expert knowledge prior to application of the function [84].

Change and deviation detection finds new significant changes in data. Very few studies were found in this category. Cohen et al., for example, approach the classification of rare cases as a deviation from a normal situation [85].

New data mining functions appear in-between clustering and summarization. Atzmueller integrated background knowledge to improve the association rule extraction and called its function subgroup discovery [22]. Dahlström et al. also use expert knowledge but have chosen a clustering function for subgroup discovery [42]. Nannings et al. use this function in order to identify risk factors and determinants of hyperglycemia in the intensive care unit from a subgroup of high risk patients [52].

Temporal data mining is a powerful technique to analysis time-stamped data. It can be used, among others, for anomaly detection such as in [59, 86], it can provide deeper insight to clinical phenomenons because it can model temporal associations between clinical variables [51, 60].

6. Choosing the Data Mining Algorithm and Search for Data Patterns

Harper highlighted that there is no best data mining algorithm, the choice of the algorithm depends on the variables and the preference of the end-user [58]. Decision trees with C4.5, for example, are preferred by some users due to the interpretability of the extracted knowledge [5]. The choice of the algorithm to use depends on the characteristics (width, height, and quality of the variable values) of the cleaned and preprocessed dataset, and the chosen data mining function. Both reflect the principal goal of the study and the complexity of the algorithm as detailed in the next section. Neural networks, for example, cannot handle large datasets, which is not the case with SVM.

As shown in table 5, Bayesian classifiers, neural networks, SVM, C4.5 and K-nearest neighbor (KNN) are the most frequently used algorithms for classification due to their high performance. The K-means algorithm is popular for clustering purposes. We will not detail all existing data mining algorithms. An easy starting point to understand the functionality of major algorithms can be found in [87] and they can be tested using the WEKA[8] data mining environment. As the final consumers of the extracted knowledge are the medical professionals, their main concern is the understandability of the extracted knowledge and not really the performance metrics. Development of algorithms providing understandability of the results and good performance is really needed in the medical field. An example of such an algorithm was proposed by Bennett et al. when they combine SVM and decision tree algorithm [88].

7. Data Mining (Pattern Extraction)

This step concerns the application of the data mining algorithm to the target

[8] http://www.cs.waikato.ac.nz/ml/weka/ (available online on the 31/01/2009)

Table 5 Overview of data mining algorithms found in the literature

DM algorithm	Papers
Algorithm based on likelihood ratio	[78, 81]
Association Mining (AM)	[25]
Association rules	[23, 100]
Bayesian classifier	[16, 27, 29, 48, 49, 61, 64]
Naive Bayes	[5, 56, 62, 82, 100]
Beam search	[22]
Biased minimax probability machine (BMPM)	[62]
Boosting	[36, 57]
C4.5	[5, 24, 56, 57, 62, 82]
Decision-fusion algorithms	[45]
Hierarchical clustering	[31, 33]
K-means	[28, 31, 42]
Clustering	[31, 83]
Decision rule	[27, 28, 58, 77]
Discriminant analysis	[28, 58, 97]
Linear discriminant	[45]
Frequent itemset mining	[83]
Genetic algorithm to adjust MCDA model	[44]
Inductive logic programming	[27]
Instance-based learner IB1	[5, 27]
KNN	[17, 18, 28, 56, 61]
MORE (association rule algorithm)	[24]
Neural networks	[17, 28, 45, 58, 66, 73, 97]
Pattern Discovery (PD)	[25]
Predictive Analysis (PA)	[25]
Regression logistic	[43, 58]
Rule induction	[27]
Simulated annealing	[26]
SVM	[17, 41, 55, 56, 66, 70, 82]

dataset. Roughly, it has three optional steps (parameter search, knowledge extraction and evaluation of the knowledge on a new dataset) depending of the data mining function. Another concern of the final consumer of the extracted knowledge is the generalization of the extracted knowledge and the application of the data mining algorithm should be carried out in a rigorous manner. The parameter of the search process is detailed in the following paragraphs, as it is not always documented in data mining papers.

Association rules, rule extraction and data visualization algorithms do not have parameters and this step does not concern these functions. This is also valid for subgroup discovery based on association rules. For the clustering and clustering-based subgroup discovery functions, the parameters are estimated on the whole data set. For the other functions, the parameters are estimated on a training set, which is a subsample of the target dataset (the remaining examples will be used to validate the extracted knowledge). The parameter search is typically computationally intensive because a range of possible values of the parameters is scanned and the algorithm is evaluated on each value of the parameters such as in [42] and [73]. Most often, the best parameters are the ones providing the smallest error. The best parameters should not only provide a smaller error on the training set but also a smaller error when it will be applied on new cases (generalization).

The training sample used during the parameter search needs to be representative of the whole dataset for generalization. Multiple training sets (noted p) should be drawn randomly from the target dataset [25, 28, 46, 73]. If the number of cases is high, it is also interesting to train and test the parameters on subsamples of the training set, as using only one subsample of the training data to train and another one to test the parameters prevents training on some examples (test subsample of the training set). An optimized approach to overcome this limitation is cross-validation: the training set is divided into k subsamples and iteratively the parameter is trained on k-1 subsamples and tested on the remaining subsample [5, 36, 46, 66, 73].

There are cases where the interesting cases are underrepresented in a dataset and where splitting the training set may create a subsample without any representation of the interesting cases. To avoid this effect, each subsample of the k folds should have the same ratio of a target variable as in the training set and this process is called stratified cross-validation [28, 50].

When the value of k is equal to the number of cases in the training set, the process is called leave-one-out. It has been shown that a repeated random selection with replacement (bootstrapping) may also provide better results, and it was used in [61]. Once the best parameter found, it should be evaluated on the validation set to obtain its generalization performance.

The extracted knowledge is what is obtained when applying the algorithm (with the best parameters if applicable) on the whole dataset, and can be used

in routine practice. It can be used to classify, predict, or rule out new clinical cases.

8. Evaluation and Interpretation

To interpret the extracted knowledge, two types of approaches were identified: an objective evaluation (quantitative measure) and a subjective (qualitative) evaluation.

A multitude of quantitative evaluation strategies were used, based on the results obtained with the application of the algorithm on p test sets (as defined in the previous section). Two papers evaluated the results quantitatively with respect to the deployment of the results in routine [75, 76]. The most commonly used metrics are accuracy, sensitivity, specificity, the receiver operating curve characteristics (ROC curves), precision, recall, f-measure, the number of positive prediction, and the number of false positives. A description of some of these measures can be found in [27]. The application of these evaluation metrics depends on the nature of the data and most of the time multiple measures are used to interpret the data mining results. Pakhomov et al., for example, use simultaneously accuracy, recall and precision as metrics to evaluate and interpret the results of an early detection of patients with a high risk of acute congestive heart failure [46]. However, recall was selected as the most important indicator as it provides the ratio of real high risk patients detected by the data mining application. Table 6 provides a list of various quantitative evaluation metrics. The definition of most of them can be found in the listed papers.

The subjective evaluations are most of the time carried out by the domain experts and are usually combined with the quantitative indicator to evaluate the results. Amongst others, we can cite the comprehensibility of the results as a qualitative measure [20, 58]; analysis of the quality of the output [31, 45]; comparison of results obtained with a baseline [36, 42, 44, 49, 89]; rating of the results [19, 22, 90]. When multiple experts are involved in the evaluation, their agreement with the discovered knowledge was also measured [66, 90].

There are situations where the researchers have to compare algorithms on one or more datasets in order to choose the best one. In their works, Harper et al. combined quantitative and subjective performance metrics and didn't find any best classifier when comparing multiple classifiers on four

Table 6 List of quantitative metrics

Performance metrics	Papers
Accuracy	[5, 17, 23, 25, 26, 28, 31, 32, 46, 55, 58, 64, 66, 82, 97]
Chi-square	[26]
Clarity and precision of the display	[20]
Computing time	[58]
Confidence	[100]
Correlation	[19, 79]
Fitness function (number of attributes, number of covered examples covered i-th pattern that were covered by earlier patterns)	[23]
F-Measure	[36, 56, 70]
FN	[32]
FP	[32, 66]
Hosmer-Lemeshow C-statistic	[49]
Lift	[100]
Linear logistic regression	[89]
LOO bootstrap sampling	[78, 81]
Mann-Whitney test for comparative study	[73]
Maximum sum (Max sensitivity + specificity)	[62]
Mean accuracy	[73]
Mean true positive rate	[73]
Negative predictive value	[17]
Optimal risk pattern	[24]
Overlap on errors	[17]
PGA comparison	[42]
Positive predictive value	[17, 25]
Precision	[36, 46, 70, 98]
Prescription of proprietary antibiotics reduction	[76]
Prevalence for PA	[25]
Quality function (measures the interestingness of the subgroup)	[22]
Recall	[36, 46, 70, 98]
Relative risk (from epidemiology)	[24]
ROC	[40, 45, 48, 49, 61, 78, 81]
Root mean square error	[73]
Round-robin	[81]
Sensitivity	[5, 17, 25, 26, 28, 35, 43, 44, 77]
Specificity	[5, 17, 25, 28, 35, 44, 77]
Support	[100]
Time between query and order selection	[53]
TP	[66]
T-test for comparisons	[28]

datasets according to their criterion (accuracy, run-time speed, comprehensibility and ease of use of the algorithms) [58]. In [17], accuracy and specificity were used to choose the best classifier for soft tissue tumor classification. In [56], the F-measure was used to compare classifiers for textual clinical documents. However, researchers should pay attention to the metrics to use during these experiments: Diettrich, for example, provided the best statistical tests to use when comparing classifiers on one dataset [91] and Demsar proposed the best statistical test to use when comparing classifiers on multiple datasets [92].

9. Use of the Discovered Knowledge

Most of the papers considered in this review are exploratory studies of data mining applications on clinical data. Some studies provide new hypothesis for medical research [5, 24, 25, 26, 54, 93]; the others confirm already established knowledge [5, 16]. The ranking of the attributes was one of the important contributions of the use of data mining for important variable selection [16]. The use of the discovered knowledge in clinical routine is still low and its adoption by medical professionals is not followed-up even if the results are patent [1, 75, 76]. One application was used for educational purpose [47].

Discussion and Conclusions

The results presented in this paper have their limitations especially with respect to the terms used to query MEDLINE. We may also have missed important aspects of the data mining process as we uniquely based our literature review on the framework proposed by Fayyad in 1996.

Data mining is a complex process including intensive manual and computational tasks. The main vocation of a healthcare organization is to provide individualized patient care, not to collect data fit for mining. There are therefore many challenges in order to streamline clinical data mining in order to find interesting and valid knowledge from the accumulated clinical data. Collection of medical data related to a defined goal is still difficult even in this era of data explosion. Sharing clinical data as benchmark datasets is a useful attempt for clinical methodological data mining and algorithm developers but patient privacy and confidentiality may hinder this effort. The use of unstructured data to complete the sparse clinical data is beneficial even if much effort is needed with respect to clinical data coding.

Important progress has been achieved over the last decade, since the early years of clinical data mining: classification for assistance in clinical care is now the main application of clinical data mining, and clinical data miners are not only implementing existing algorithms but use methodological approaches and develop new algorithms suitable for clinical data.

However, some problems related to the generalization and reproducibility of the knowledge extraction process still need to be fixed: characteristics of variables used in studies should be better documented as these are important to reproduce the results in other clinical centers. Similarly, the parameters used during the pattern extraction process need to be reported systematically. Interpretability of the results is important in the medical domain and an implementation of algorithms providing easily interpretable results needs to be carried out systematically before using black-box algorithms even if they are known to provide better results. A comparison with traditional statistical analyses should be carried out to show the data mining effectiveness, in order to convince final users of the knowledge who are mainly medical practitioners. As a consequence, strong statistical tests need to be carried out to draw any conclusion.

Currently, there is a trend of creating healthcare networks at regional, national, and even international levels. It is a great opportunity for clinical data miners because acquiring relevant data for a specific clinical problem may become easier. The underlying effort with respect to data interoperability permits to integrate data from different sources.

At the same time, the promising research in the genomics, proteomics and physiomics fields [41, 94] will increase the complexity of the datasets. More than the size of data to be analyzed, the relational multiplicity of the data may cause problems. Algorithmic development such as those proposed in [45] and computational infrastructures to handle the combination of these data for a data mining application will be the next challenge on clinical data mining [95].

As stated above, unstructured data such as those contained in free-text reports are rich in information. Selection of relevant unstructured data in the context of an integrated healthcare setting may be problematic. Exploitation of semi-supervised data mining functions may help. Much progress has already been achieved in this area but there are few applications in the medical domain. The temporality, dynamic, and contextual aspects of data collected in a clinical context are rarely exploited by data mining researchers [79, 96]. As Stacey and McGregor pointed out in [12], there is a lack of interaction between the data mining and data abstraction researchers. The initiatives of the IDA-DM working group to promote and disseminate scientific results from both domains will push forward on this direction.

Acknowledgement

This work has been partially funded by the European Commission Sixth Framework Programme @neurIST project (IST-027703, see http://www.aneurist.org) and the KnowARC project (IST-032691), the Swiss National Science Foundation grant 200020–118638/1, and by the Geneva University Hospitals research grants 05-I-13 and 05–9–II.

References

1. Chen R, Mongkolwat P, Channin DS. RadMonitor: radiology operations data mining in real time. J Digit Imaging 2008;21:257-68.

2. Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol 2004;25:690-5.
3. Zhu X. Semi-Supervised Learning Literature Survey. University of Wisconsin-Madison; 2007.
4. Fayyad U, Piatetsky-Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. AAAI Press; 1996. p. 82-8.
5. Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. Artif Intell Med 2007;41:251-62.
6. Olson DL, Delen D. Advanced data mining techniques. Springer; 2008.
7. Holena M, Sochorova A, Zvarova J. Increasing the diversity of medical data mining through distributed object technology. Stud Health Technol Inform 1999;68:442-7.
8. Smyth P. Data mining: data analysis on a grand scale. In: Statistical Methods in Medical Research; 2000. p. 309-327.
9. Patel JL, Goyal RK. Applications of artificial neural networks in medical science. Curr Clin Pharmacol 2007;2:217-26.
10. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;:128-44.
11. Zhou L, Hripcsak G. Temporal reasoning with medical data—a review with emphasis on medical natural language processing. J Biomed Inform 2007; 40:183-202.
12. Stacey M, McGregor C. Temporal abstraction in intelligent clinical data analysis: A survey. Artif Intell Med 2007;39(1):1-24.
13. Hennessy S. Use of health care databases in pharmacoepidemiology. Basic Clin Pharmacol Toxicol 2006;98:311-3.
14. Zupan B, Demsar J. Open-Source Tools for Data Mining. Clinics in Laboratory Medicine. 2008; 28(1):37-54.
15. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. Int J Med Inform 2008;77(2):81-97.
16. Bayat S, Cuggia M, Kessler M, Briançon S, Le Beux P, Frimat L. Modelling access to renal transplantation waiting list in a French healthcare network using a Bayesian method. Stud Health Technol Inform 2008;136:605-10.
17. Garcia-Gomez JM, Vidal C, Marti-Bonmati L, Galant J, Sans N, Robles M, et al. Benign/malignant classifier of soft tissue tumors using MR imaging. MAGMA 2004;16:194-201.
18. Juhola M, Laurikkala J. On distance computation in space of mixed-type variables in medical data mining. Stud Health Technol Inform 2002;90:425-30.
19. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton GB. Using discordance to improve classification in narrative clinical databases: an application to community-acquired pneumonia. Comput Biol Med 2007;37:296-304.
20. Grant A, Moshyk A, Diab H, Caron P, de Lorenzi F, Bisson G, et al. Integrating feedback from a clinical data warehouse into practice organisation. Int J Med Inform 2006;75:232-9.
21. Klimov D, Shahar Y. A framework for intelligent visualization of multiple time-oriented medical records. AMIA Annu Symp Proc 2005;:405-9.
22. Atzmueller M. Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In: Proc. 19th International Joint Conference on Artificial Intelligence (IJCAI-05). 2005. p. 647-52.
23. Kwasnicka H, Katejan S. Discovery of association rules from medical data - classical and evolutionary approaches. In: XXI Autumn Meeting of Polish Information Processing Society. 2005. p. 163-77.
24. Li J, Fu AW, Fahey P. Efficient discovery of risk patterns in medical data. Artif Intell Med 2009;45 (1):77-89.
25. Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller WG, et al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Comput Biol Med 2006;36:1351-77.
26. Richards G, Rayward-Smith V, Sonksen P. Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med 2001; 22:215-31.
27. Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16:3-23.
28. Juhola M. On machine learning classification of otoneurological data. Stud Health Technol Inform 2008;136:211-6.
29. Ramoni M, Sebastiani P. Robust Bayes classifiers. Artificial Intelligence 2001;125(1-2):209-26.
30. Goodwin LK, Prather JC. Protecting patient privacy in clinical data mining. J Healthc Inf Manag 2002;16:62-67.
31. Jannin P, Morandi X. Surgical models for computer-assisted neurosurgery. Neuroimage 2007; 37:783-91.
32. Rao BR, Sandilya S, Niculescu R, Germond C, Goel A. Mining time-dependent patient outcomes from hospital patient records. Proc AMIA Symp 2002;:632-6.
33. Rost TB, Edsberg O, Grimsmo A, Nytro O. Comparing medical code usage with the compression-based dissimilarity measure. Stud Health Technol Inform 2007;129:684-8.
34. Spangler WE, May JH, Strum DP, Vargas LG. A data mining approach to characterizing medical code usage patterns. J Med Syst 2002;26:255-75.
35. Chapman WW, Dowling JN, Wagner MM. Fever detection from free-text clinical records for biosurveillance. J Biomed Inform 2004;37:120-7.
36. Goldstein I, Arzrumtsyan A, Uzuner O. Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. In: AMIA Annu Symp Proc 2007. p. 279-83.
37. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res 1975;8(4):303-20.
38. Miller RA, Pople HE, Myers JD. INTERNIST-1: An experimental computer-based diagnostic consultant for general internal medecine. N Engl J Med 1982 ;307:468-76.
39. Antonie M, Zaïane O, Coman A. Application of data mining techniques for medical image classification. In: Proceedings of Second International Workshop on Multimedia Data Mining (MDM/ KDD'2001); 2001. p. 94-101.
40. Bohm N, Wales L, Dunckley M, Morgan R, Loftus I, Thompson M. Objective risk-scoring systems for repair of abdominal aortic aneurysms: applicability in endovascular repair?.Eur J Vasc Endovasc Surg 2008;36:172-7.
41. Daemen A, Gevaert O, De Moor B. Integration of clinical and microarray data with kernel methods. Conf Proc IEEE Eng Med Biol Soc 2007:5411-5.
42. Dahlstrom O, Timpka T, Hass U, Skogh T, Thyberg I. A simple method for heuristic modeling of expert knowledge in chronic disease: identification of prognostic subgroups in rheumatology. Stud Health Technol Inform 2008;136:157-62.
43. Gellerstedt M, Glymour C, Madigan D, l Pregibon D, Smyth P. Statistical inference and data mining. Communications of ACM 1996;39(11):35-41.
44. Goletsis Y, Papaloukas C, Fotiadis DI, Likas A, Michalis LK. Automated ischemic beat classification using genetic algorithms and multicriteria decision analysis. IEEE Trans Biomed Eng 2004;51:1717-25.
45. Jesneck JL, Nolte LW, Baker JA, Floyd CE, Lo JY. Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis. Med Phys 2006;33:2945-54.
46. Pakhomov SV, Buntrock J, Chute CG. Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. J Biomed Inform 2005;38:145-53.
47. Varpa K, Iltanen K, Juhola M. Machine learning method for knowledge discovery experimented with otoneurological data. Comput Methods Programs Biomed 2008;91:154-64.
48. Lin JH, Haug PJ. Data preparation framework for preprocessing clinical data in data mining. AMIA Annu Symp Proc 2006;:489-93.
49. Alvarez SM, Poelstra BA, Burd RS. Evaluation of a Bayesian decision network for diagnosing pyloric stenosis. J Pediatr Surg 200 ;41:155-61.
50. Cohen G, Hilario M, Sax H, Hugo S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. Artif Intell Med 2006; 37:7-18.
51. Bellazzi R, Larizza C, Magni P, Bellazzi R. Temporal data mining for the quality assessment of hemodialysis services. Artif Intell Med 2005; 34(1):25-39.
52. Nannings B, Bosman RJ, Abu-Hanna A. A subgroup discovery approach for scrutinizing blood glucose management guidelines by the identification of hyperglycemia determinants in ICU patients. Methods Inf Med 2008;47(6):480-8.
53. Jalloh OB, Waitman LR. Improving Computerized Provider Order Entry (CPOE) usability by data mining users' queries from access logs. AMIA Annu Symp Proc 2006;:379-83.
54. Korhonen M, Salo S, Suni J, Larmas M. Computed online determination of life-long mean index values for carious, extracted, and/or filled permanent teeth. Acta Odontol Scand 2007;65:214-8.
55. Nguyen A, Moore D, McCowan I, Courage MJ. Multi-class classification of cancer stages from free-text histology reports using support vector machines. Conf Proc IEEE Eng Med Biol Soc 2007:5140-3.
56. Spat S, Cadonna B, Rakovac I, Gütl C, Leitner H, Stark G, et al. Enhanced information retrieval from narrative German-language clinical text documents

using automated document classification. Stud Health Technol Inform 2008;136:473-8.
57. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc 2007;14:574-80.
58. Harper PR. A review and comparison of classification algorithms for medical decision making. Health Policy 2005;71:315-31.
59. Tusch G, Bretl CE, Connor M, Das A. SPOT Towards Temporal Data Mining in Medicine and Bioinformatics. In: AMIA Annu Symp Proc 2008. p. 1157.
60. Raj R, O'Connor MJ, Das AK. An ontology-driven method for hierarchical mining of temporal patterns: application to HIV drug resistance research. AMIA Annu Symp Proc 2007;:614-9.
61. Lin JH, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J Biomed Inform 2008;41:1-14.
62. Huang K, Yang H, King I, Lyu MR. Maximizing sensitivity in medical diagnosis using biased minimax probability machine. IEEE Trans Biomed Eng 2006;53:821-31.
63. Barnes J, Chambers I, Piper I, Citerio G, Contant C, Enblad P, et al. Accurate data collection for head injury monitoring studies: a data validation methodology. Acta Neurochir Suppl 2005;95:39-41.
64. Le Duff F, Happe A, Burgun A, Levionnois S, Bremond M, Le Beux P. Sharing medical data for patient path analysis with data mining method. Stud Health Technol Inform 2001;84:1364-8.
65. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform 2008;41:387-92.
66. Pakhomov SV, Hanson PL, Bjornsen SS, Smith SA. Automatic classification of foot examination findings using clinical notes and machine learning. J Am Med Inform Assoc 2008;15:198-202.
67. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26:1-24.
68. Depeursinge A, Iavindrasana J, Hidki A, Cohen G, Geissbuhler A, Platon A, et al. Comparative Performance Analysis of State-of-the-Art Classification Algorithms Applied to Lung Tissue Categorization. J Digit Imaging 2008.
69. Iavindrasana J, Cohen G, Depeursinge A, Meyer R, Geissbuhler A. Minimal Set of Attributes Required to Report Hospital-Acquired Infection Cases. In: IDAMAP. 2008.
70. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc 2008;15:25-8.
71. Suckling J, Parker J, Dance DR, Astley S, Hutt I, Boggis CR, et al. The Mammographic Image Analysis Society digital mammogram database. In: Exerpta Medica. International Congress; 1994. p. 375-8.
72. Bath PA. Data mining in health and medical information. Annual Review of Information Science and Technology 2004;38(1):331-69.
73. Autio L, Juhola M, Laurikkala J. On the neural network classification of medical data and an endeavour to balance non-uniform data sets with artificial data extension. Comput Biol Med 2007; 37:388-97.
74. Berman JJ. Confidentiality issues for medical data miners. Artif Intell Med 2002;26:25-36.
75. Awaya T, Ohtaki K, Yamada T, Yamamoto K, Miyoshi T, Itagaki Y, et al. Automation in drug inventory management saves personnel time and budget. Yakugaku Zasshi 2005;125:427-32.
76. Bernstein SL, Whitaker D, Winograd J, Brennan JA. An electronic chart prompt to decrease proprietary antibiotic prescription to self-pay patients. Acad Emerg Med 2005;12:225-31.
77. Brenneman SK, Lacroix AZ, Buist DS, Chen YT, Abbott TA. Evaluation of decision rules to identify postmenopausal women for intervention related to osteoporosis. Dis Manag 2003;6:159-68.
78. Bilska-Wolak AO, Floyd CE. Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer. Phys Med Biol 2004;49:4219-37.
79. Berner ES, Moss J. Informatics challenges for the impending patient information explosion. J Am Med Inform Assoc 2005;12:614-7.
80. Fayyad U, Piatetsky-shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine 1996;17:37-54.
81. Bilska-Wolak AO, Floyd CE, Lo JY, Baker JA. Computer aid for decision to biopsy breast masses on mammography: validation on new cases. Acad Radiol 2005;12:671-80.
82. Joshi M, Pakhomov S, Pedersen T, Chute CG. A comparative study of supervised learning as applied to acronym expansion in clinical reports. AMIA Annu Symp Proc 2006;:399-403.
83. Aronsky D, Kasworm E, Jacobson JA, Haug PJ, Dean NC. Electronic screening of dictated reports to identify patients with do-not-resuscitate status. J Am Med Inform Assoc 2004;11:403-9.
84. Richardson M, Domingos P. Learning with knowledge from multiple experts. In: In ICML 20. 2003; p. 624-31.
85. Cohen G, Sax H, Geissbuhler A. Novelty detection using one-class Parzen density estimator. An application to surveillance of nosocomial infections. Stud Health Technol Inform 2008;136:21-6.
86. Jakkula V, Cook DJ. Anomaly detection using temporal data mining in a smart home environment. Methods Inf Med 2008;47(1):70-5.
87. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann; 2005.
88. Bennett KP, Blue JA. A Support Vector Machine Approach to Decision Trees. In: Department of Mathematical Sciences Math Report No. 97-100, Rensselaer Polytechnic Institute.;1997. p. 2396-401.
89. Berner ES, Maisiak RS, Heuderbert GR, Young KR. Clinician performance and prominence of diagnoses displayed by a clinical diagnostic decision support system. AMIA Annu Symp Proc 2003;:76-80.
90. Hyun S, Bakken S, Johnson SB. Markup of temporal information in electronic health records. Stud Health Technol Inform 2006;122:907-8.
91. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 1998; 10(7): 1895-923.
92. Demsar J. Statistical Comparisons of Classifiers over Multiple Data Sets. J Mach Learn Res 2006;7:1-30.
93. Jin HW, Chen J, He H, Williams GJ, Kelman C, O'Keefe CM. Mining unexpected temporal associations: applications in detecting adverse drug reactions. IEEE Trans Inf Technol Biomed 2008; 12:488-500.
94. Mitchell DR, Mitchell JA. Status of clinical gene sequencing data reporting and associated risks for information loss. J Biomed Inform 2007;40(1):47-54.
95. Pierson JM, Gossa J, Wehrle P, Cardenas Y, Cahon S, El Samad M, et al. GGM: efficient navigation and mining in distributed genomedical data. IEEE Trans Nanobioscience 2007 ;6:110-6.
96. McSherry D. Dynamic and static approaches to clinical data mining. Artif Intell Med. 1999 ;16:97-115.
97. Lee IN, Liao SC, Embrechts M. Data mining techniques applied to medical information. Med Inform Internet Med 2000;25:81-102.
98. Imai T, Aramaki E, Kajino M, Miyo K, Onogi Y, Ohe K. Finding malignant findings from radiological reports using medical attributes and syntactic information. Stud Health Technol Inform 2007;129:540-4.
99. Sauleau EA, Paumier JP, Buemi A. Medical record linkage in health information systems by approximate string matching and clustering. BMC Med Inform Decis Mak 2005;5:32.
100. Ordonez C. Association rule discovery with the train and test approach for heart disease prediction. IEEE Trans Inf Technol Biomed 2006; 10:334-43.
101. Roddick JF, Fule P, Graco WJ. Exploratory medical knowledge discovery: experiences and issues. SIGKDD Explorations 2003;5(1):94-99.

Correspondence to:
Jimison Iavindrasana
Division of Medical Informatics
University Hospitals and University of Geneva
Rue Gabrielle-Perret-Gentil 4
CH-1211Geneva 14, Switzerland
Tel: +41 22 372 88 74
Fax: +41 22 372 88 79
E-mail: jimison.iavindrasana@sim.hcuge.ch