ELSEVIER

Artificial Intelligence in Medicine 26 (2002) 1–24

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Uniqueness of medical data mining

Krzysztof J. Cios[a,b,c,d,*], G. William Moore[e,f,g]

[a]*Department of Computer Science and Engineering, University of Colorado at Denver, Campus Box 109, 1200 Larimer Street, Denver, CO 80217-3364, USA*
[b]*University of Colorado at Boulder, Boulder, CO, USA*
[c]*University of Colorado Health Sciences Center, Denver, CO, USA*
[d]*4cData LLC, Golden, CO, USA*
[e]*Baltimore Veterans Affairs Medical Center, Baltimore, MD, USA*
[f]*University of Maryland School of Medicine, Baltimore, MD, USA*
[g]*The Johns Hopkins University School of Medicine, Baltimore, MD, USA*

Received 5 March 2002; accepted 11 March 2002

Abstract

This article addresses the special features of data mining with medical data. Researchers in other fields may not be aware of the particular constraints and difficulties of the privacy-sensitive, heterogeneous, but voluminous data of medicine. Ethical and legal aspects of medical data mining are discussed, including data ownership, fear of lawsuits, expected benefits, and special administrative issues. The mathematical understanding of estimation and hypothesis formation in medical data may be fundamentally different than those from other data collection activities. Medicine is primarily directed at patient-care activity, and only secondarily as a research resource; almost the only justification for collecting medical data is to benefit the individual patient. Finally, medical data have a special status based upon their applicability to all people; their urgency (including life-or-death); and a moral obligation to be used for beneficial purposes.

Keywords: Medical data mining; Unique features of medical data mining and knowledge discovery; Ethical; Security and legal aspects of medical data mining

1. Introduction

This article emphasizes the uniqueness of medical data mining. This is a position paper, in which the authors' intent, based on their medical and data mining experience, is to alert the data mining community to the unique features of medical data mining. The reason for

* Corresponding author. Tel.: +1-303-556-4314; fax: +1-303-556-8369.
E-mail address: krys.cios@cudenver.edu (K.J. Cios).

0933-3657/02/$ – see front matter
PII: S0933-3657(02)00049-0

writing the paper is that researchers who perform data mining in other fields may not be aware of the constraints and difficulties of mining the privacy-sensitive, heterogeneous data of medicine. We discuss ethical, security and legal aspects of medical data mining. In addition, we pose several questions that must be answered by the community, so that both the patients on whom the data are collected, as well as the data miners, can benefit [15].

Human medical data are at once the most rewarding and difficult of all biological data to mine and analyze. Humans are the most closely watched species on earth. Human subjects can provide observations that cannot easily be gained from animal studies, such as visual and auditory sensations, the perception of pain, discomfort, hallucinations, and recollection of possibly relevant prior traumas and exposures. Most animal studies are short-term, and therefore cannot track long-term disease processes of medical interest, such as preneoplasia or atherosclerosis. With human data, there is no issue of having to extrapolate animal observations to the human species.

Some three-quarter billions of persons living in North America, Europe, and Asia have at least some of their medical information collected in electronic form, at least transiently. These subjects generate volumes of data that an animal experimentalist can only dream of. On the other hand, there are ethical, legal, and social constraints on data collection and distribution, that do not apply to non-human species, and that limit the scientific conclusions that may be drawn.

The major points of uniqueness of medical data may be organized under four general headings:

- Heterogeneity of medical data
- Ethical, legal, and social issues
- Statistical philosophy
- Special status of medicine

2. Heterogeneity of medical data

Raw medical data are voluminous and heterogeneous. Medical data may be collected from various images, interviews with the patient, laboratory data, and the physician's observations and interpretations. All these components may bear upon the diagnosis, prognosis, and treatment of the patient, and cannot be ignored. The major areas of heterogeneity of medical data may be organized under these headings:

- Volume and complexity of medical data
- Physician's interpretation
- Sensitivity and specificity analysis
- Poor mathematical characterization
- Canonical form

2.1. Volume and complexity of medical data

Raw medical data are voluminous and heterogeneous. Medical data may be collected from various images, interviews with the patient, and physician's notes and interpretations.

All these data-elements may bear upon the diagnosis, prognosis, and treatment of the patient, and must be taken into account in data mining research.

More and more medical procedures employ imaging as a preferred diagnostic tool. Thus, there is a need to develop methods for efficient mining in databases of images, which are more difficult than mining in purely numerical databases. As an example, imaging techniques like SPECT, MRI, PET, and collection of ECG or EEG signals, can generate gigabytes of data per day. A single cardiac SPECT procedure on one patient may contain dozens of two-dimensional images. In addition, an image of the patient's organ will almost always be accompanied by other clinical information, as well as the physician's interpretation (clinical impression, diagnosis). This heterogeneity requires high capacity data storage devices and new tools to analyze such data. It is obviously very difficult for an unaided human to process gigabytes of records, although dealing with images is relatively easier for humans because we are able to recognize patterns, grasp basic trends in data, and formulate rational decisions. The stored information becomes less useful if it is not available in an easily comprehensible format. Visualization techniques will play an increasing role in this setting, since images are the easiest for humans to comprehend, and they can provide a great deal of information in a single snapshot of the results.

2.2. *Importance of physician's interpretation*

The physician's interpretation of images, signals, or any other clinical data, is written in unstructured free-text English, that is very difficult to standardize and thus difficult to mine. Even specialists from the same discipline cannot agree on unambiguous terms to be used in describing a patient's condition. Not only do they use different names (synonyms) to describe the same disease, but they render the task even more daunting by using different grammatical constructions to describe relationships among medical entities.

It has been suggested that computer translation may hold part of the solution for processing the physician's interpretation [26,10,20]. Principles of computer translation may be summarized as follows [33]:

- Machine translation is typically composed of the following three steps: analysis of a source language sentence; transfer . . . from one language to another; and generation of a target language sentence.
- Natural language can be regarded as a huge set of exceptional expressions . . . as many expressions as possible must be collected in the dictionary . . . It is an endless job.
- One of the difficulties of translation . . . is that the translation of an input sentence is not unique (see Section 2.5).
- Current translation systems can analyze and translate sentences composed of less than 10 words . . . A reason for such failure is the ambiguity . . . Even a human cannot understand the meaning of a long sentence at the first reading.
- Grammatical rules in machine translation can be regarded as (artificial intelligence) production rules.

These principles, suitably customized for medical text, may be required for future medical data mining applications that depend upon the physician's free-text interpretation as part of the data mining analysis.

2.3. Sensitivity and specificity analysis

Nearly all diagnoses and treatments in medicine are imprecise, and are subject to rates of error. The usual paradigm in medicine for measuring this error is *sensitivity and specificity analysis*. One should distinguish between a *test* and a *diagnosis* in medicine. A test is one of many values used to characterize the medical condition of a patient; a diagnosis is the synthesis of many tests and observations, that describes a pathophysiologic process in that patient. Both tests and diagnoses are subject to sensitivity/specificity analysis.

In medical sensitivity and specificity analysis, there are test-results and an *independent measure of truth*, or *hypothesis*. Typically, the test-results are a proposed, inexpensive new test, whereas the hypothesis is either a more expensive test, regarded as definitive, or else a complete medical workup of the patient.

The *accuracy of a test*, on the other hand, compares how close a new test value is to a value predicted by if . . . then rules. To classify a test example, the rule that matches it best determines the example's class membership. An accuracy test is defined as:

$$\text{accuracy} = \frac{\text{TP}}{\text{total}} 100\%$$

where TP stands for *true positive*, and indicates the number of correctly recognized test examples, and total is the *total number of test examples*. This measure is very popular in the machine learning and pattern recognition communities, but is not acceptable in medicine because it hides essential details of the achieved results, as illustrated in the following example. In an accuracy test, a new case is checked against the rules describing all classes, row-wise. Let us examine the hypothetical data shown in Table 1. The numbers in Table 1 represent the degree of matching of a test example with the rules generated for the three classes. The matching can be understood as the degree to which the if . . . parts of the rule's conditions (there may be many of them) are satisfied by a new case. For instance, if out of 10 conditions only 8 are satisfied, then the degree of matching is 0.8. Since the decision is made based on the highest degree of matching, we see that all cases for the first two classes

Table 1
Hypothetical results on 10 test examples

Correct classification	Classification (using best matching with rules for class . . .)		
	Rules for class 1	Rules for class 2	Rules for class 3
Test example of class 1	**0.91**	0.37	0.97
	0.92	0.42	0.97
	0.94	0.14	0.99
	0.93	0.31	0.98
Test example of class 2	0.21	**0.95**	0.98
	0.13	**0.93**	0.97
	0.34	**0.90**	0.98
	0.53	**0.96**	0.99
Test example of class 3	0.36	0.17	**0.93**

Table 2
Possible outcomes of a test

	Test result positive	Test result negative
Hypothesis positive	TP	FN
Hypothesis negative	FP	TN

are incorrectly classified as belonging to class 3. Thus, the overall accuracy is only 20% (only 2 out of 10 test examples are correctly classified). True positives are in bold in Table 1.

When only two outcomes (positive and negative) of a test are possible, three evaluation criteria can be used for measuring the effectiveness of the generated rules. There are four possibilities, as shown in Table 2.

Where *true positive* (TP) indicates the number of correct positive predictions (classifications); *true negative* (TN) is the number of correct negative predictions; *false positive* (FP) is the number of incorrect positive predictions; and *false negative* (FN) is the number of incorrect negative predictions.

The three measures are:

$$\text{sensitivity} = \frac{\text{TP}}{\text{hypothesis positive}} 100\% = \frac{\text{TP}}{\text{TP} + \text{FN}} 100\%$$

$$\text{specificity} = \frac{\text{TN}}{\text{hypothesis negative}} 100\% = \frac{\text{TN}}{\text{FP} + \text{TN}} 100\%$$

$$\text{predictive accuracy} = \frac{\text{TP} + \text{TN}}{\text{total}} 100\% = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} 100\%$$

Sensitivity measures the ability of a test to be positive when the condition is actually present, or how many of the positive test examples are recognized. In other words, the sensitivity measures how often you find what you are looking for. It goes under a variety of near-synonyms: false-negative rate, recall, Type II error, β error, error of omission, or alternative hypothesis.

Specificity measures the ability of a test to be negative when the condition is actually not present, or how many of the negative test examples are excluded. In other words the specificity measures how often what you find is what you are looking for. It goes under a variety of near-synonyms: false-positive rate, precision, Type I error, α error, error of commission, or null hypothesis. Predictive accuracy gives an overall evaluation. *A high level of confidence can be placed only for results that give high values for all three measures*.

Sensitivity/specificity results are illustrated on data from Table 1. For the first two classes, the sensitivity, specificity, and predictive accuracy are all 100% (if we use the threshold of acceptance at 0.9). For class 1 the sensitivity is 4/4, specificity is 6/6, and predictive accuracy is 10/10. For class 3 the sensitivity is 100% (2/2), but specificity is 0% (0/8) and predictive accuracy is 20% (2/10). The results suggest that the rules generated from the first two classes are correct (since all three values are high) in recognizing test

Table 3
Hypothetical results on 27 test examples

Correct diagnosis	Classified as A	Classified as B	Classified as C	Classified as D
A	4	1	1	1
B	0	7	0	0
C	0	0	2	1
D	0	0	0	10

examples, but the rules for class 3 are not. In other words, if the rules generated for one class fail to describe a class this failure has no effect on recognition of other classes, as opposed to the accuracy test.

As another example, let us suppose that we have the results shown in Table 3, which is known as a confusion matrix. Table 4 shows how the sensitivity, specificity, and accuracy are calculated [12]. For comparison, the accuracy test would give just one, quite misleading, result: $23/27 = 85\%$.

Finally, the details of performing a complete medical workup may be slightly different for each patient, since details of consent and ethical management vary from patient to patient. Therefore, the analysis may be considered subjective by some standards. Furthermore, a sensitivity/specificity analysis must be formulated as an appropriate yes–no question, which is sometimes fairly challenging in medical investigations. A carelessly formulated yes–no question may become a self-fulfilling prophecy, such as the ''have you stopped beating your spouse'' question, in which either a yes or no answer automatically implies a history of spousal abuse. Reluctance to use the sensitivity/specificity measures of error analysis in medical data mining may be due to many factors: the expectation that the results will not appear very convincing, and will therefore be not publishable or grant-fundable; and the burdensome, expensive, and sometimes imprecise process of evaluating each case in the study.

2.4. Poor mathematical characterization of medical data

Another unique feature of medical data mining is that the underlying data structures of medicine are poorly characterized mathematically, as compared to many areas of the physical sciences. Physical scientists collect data which they can put into formulas, equations, and models that reasonably reflect the relationships among their data. On the other hand, the conceptual structure of medicine consists of word descriptions and images, with very few formal constraints on the vocabulary, the composition of images, or

Table 4
Calculation of sensitivity/specificity test

	A	B	C	D
Sensitivity	57% (4/7)	100% (7/7)	67% (2/3)	100% (10/10)
Specificity	100% (20/20)	95% (19/20)	96% (23/24)	88% (15/17)
Accuracy	89% (24/27)	96% (26/27)	93% (25/27)	93% (25/27)

the allowable relationships among basic concepts. The fundamental entities of medicine, such as inflammation, ischemia, or neoplasia, are just as real to a physician as entities such as mass, length, or force are to a physical scientist; but medicine has no comparable formal structure into which a data miner can organize information, such as might be modeled by clustering, regression models, or sequence analysis. In its defense, medicine must contend with hundreds of distinct anatomic locations and thousands of diseases. Until now, the sheer magnitude of this concept space was insurmountable. Furthermore, there is some suggestion that the logic of medicine may be fundamentally different from the logic of the physical sciences [7,29,30,55]. However, it may now happen that faster computers and the newer tools of data mining and knowledge discovery (DMKD) may overcome this prior obstacle.

2.5. Canonical form

In mathematics, a *canonical form* is a preferred notation that encapsulates all equivalent forms of the same concept. For example, the canonical form for one-half is 1/2, and there is an algorithm for reducing the infinity of equivalent expressions, or *aliases*, namely 2/4, 3/6, 4/8, 5/10, . . . , down to 1/2. Agreement upon a canonical form is one of the features of any mature intellectual discipline. For example, the importance of a canonical form became apparent to the dictionary writers of the 18th century, who realized that one could not prepare a dictionary without consistent orthography. The variable orthography, say, of 14th century poet, Geoffrey Chaucer, could not be supported by the need to have each individual English word appear and be defined in only one place in the dictionary. Investigators working with medical text have reached the same conclusion [47].

Unfortunately, in biomedicine, even elementary concepts have no canonical form. For example, the canonical form for even a simple idea, such as: ‘‘adenocarcinoma of colon, metastatic to liver’’, has no consistent form of expression. The individual medical words, of course, all have a unique spelling and meaning; but the following distinct expressions (and many others, easy to imagine) are all medically equivalent:

Colon adenocarcinoma, metastatic to liver;
Colonic adenocarcinoma, metastatic to liver;
Large bowel adenocarcinoma, metastatic to liver;
Large intestine adenocarcinoma, metastatic to liver;
Large intestinal adenocarcinoma, metastatic to liver;
Colon’s adenocarcinoma, metastatic to liver;
Adenocarcinoma of colon, with metastasis to liver;
Adenocarcinoma of colon, with liver metastasis;
Adenocarcinoma of colon, with hepatic metastasis.

What about even more complex ideas? What about size-quantifiers (e.g. 2.5 cm metastasis to the liver), logical-quantifiers (for some, for every, etc.), cardinality (three metastases to the liver), ordinality (the third-largest metastasis in the liver), conditionals (if there is a liver metastasis . . .), logical-not, logical-and, logical-or, etc.? If there is no canonical form for equivalent ideas in biomedicine, then how are indexes and statistical tables constructed, when these data mining methods depend upon equivalent concepts

being tabulated together? Some of these canonical form issues will be addressed in the emerging XML standards for biomedical data in Section 3.3.

3. Ethical, legal, and social issues

Because medical data are collected on human subjects, there is an enormous ethical and legal tradition designed to prevent the abuse of patients and misuse of their data. The major points of the ethical, legal, and social issues in medicine may be organized under five headings:

- Data ownership
- Fear of lawsuits
- Privacy and security of human data
- Expected benefits
- Administrative issues

3.1. Data ownership

There is an open question of data ownership in medical data mining. In legal theory, ownership is determined by who is entitled to sell a particular item of property [32]. Since it is considered unseemly to sell human data or tissue, the question of data ownership in medicine is similarly muddled. The corpus of human medical data potentially available for data mining is enormous. Thousands of terabytes are now generated annually in North America and Europe. However, these data are buried in heterogeneous databases, and scattered throughout the medical care establishment, without any common format or principles of organization. The question of ownership of patient information is unsettled, and the object of recurrent, highly publicized lawsuits and congressional inquiries. Do individual patients own data collected on themselves? Do their physicians own the data? Do their insurance providers own the data? Some HMOs now refuse to pay for patient participation in clinical treatment protocols that are deemed experimental. If insurance providers do not own their insurees' data, can they refuse to pay for the collection and storage of the data? If the ability to process and sell human medical data is unseemly, then how should the data managers, who organize and mine the data, be compensated? Or should this incredibly rich resource for the potential betterment of humankind be left unmined?

3.2. Fear of lawsuits

Another feature of medical data mining is a fear of lawsuits directed against physicians and other health-care providers. Medical care in the USA, for those who can afford it, is the very best. However, US medical care is some 30% more expensive than that in Canada and Europe, where quality is comparable; and US medicine also has the most litigious malpractice climate in the world. Some have argued that this 30% surcharge on US medical care, about US$ 1000 per capita annually, is mostly medico-legal: either direct

legal costs, or else the overhead of ''defensive medicine'', i.e. unnecessary tests ordered by physicians to cover themselves in potential future lawsuits. In this tense climate, physicians and other medical data-producers are understandably reluctant to hand over their data to data miners. Data miners could browse these data for untoward events. Apparent anomalies in the medical history of an individual patient might trigger an investigation. In many cases, the appearance of malpractice might be a data-omission or data-transcription error; and not all bad outcomes in medicine are necessarily the result of negligent provider behavior. However, an investigation inevitably consumes the time and emotional energy of medical providers. For exposing themselves to this risk, what reward do the providers receive in return?

3.3. Privacy and security of human data

Another unique feature is privacy and security concerns. For instance, US federal rules set guidelines for concealment of individual patient identifiers. At stake is not only a potential breach of patient confidentiality, with the possibility of ensuing legal action; but also erosion of the physician–patient relationship, in which the patient is extraordinarily candid with the physician in the expectation that such private information will never be made public. By some guidelines, concealment of identifiers must be irreversible. A related privacy issue may apply if, for example, crucial diagnostic information were to be discovered on patient data, and a patient could be treated if one could only go back and inform the patient about the diagnosis and possible cure. In some cases, this action may not be taken. Another issue is data security in data handling, and particularly in data transfer. Before the identifiers are concealed, only authorized persons should have access to the data. Since transferring the data electronically via the Internet is insecure, the identifiers must be carefully concealed even for transfers within a single medical institution from one unit to another.

On the other hand, it has been noted in recent US federal documents [49–51], that there are at least two legitimate research needs for re-identification of de-identified medical data: first, there is a need to prevent accidental duplicate records on the same patient from skewing research conclusions; second, there may be a compelling need to refer to original (re-identified) medical records to verify the correctness or to obtain additional information on specific patients. These special requirements could be managed by appropriate regulatory agencies, but they could not be met at all if the data are completely anonymous. There are four forms of patient data identification:

- *Anonymous data* are data that were collected so that the patient-identification was removed at the time the information was collected. For example, a block of tissue may be taken from an autopsy on a patient with a certain disease, to serve as control tissue-block in the histology laboratory. The patient's identifiers are not recorded at the time of specimen collection, and thus can never be recovered.
- *Anonymized data* are data that are collected initially with the patient-identifiers, which are subsequently, irrevocably removed. That is, there can never be a possibility of returning to the patient's record and obtaining additional information. This research practice has been common in the past. However, anonymized data, as described above,

could be accidentally duplicated, and could not be verified for corrections or additional data.

- *De-identified data* are data that are collected initially with the patient-identifiers, which are subsequently encoded or encrypted. The patient can be re-identified under conditions stipulated by an appropriate agency, typically an Institutional Review Board (IRB).
- *Identified data* can only be collected under significant review by the institution, federal guidelines, etc. with the patient giving written informed consent.

Even for public Internet distribution, identifier-encrypted data which enter the database only once are fairly safe from attackers. For example, in the Johns Hopkins Autopsy Resource [31], a publicly posted Internet resource that lists over 50 000 deceased patients, each deceased patient enters the database only once, and is contributed by a single institution with an IRB-approved encryption procedure. On the other hand, data from multiple institutions are only as secure as the procedures from the least-secure contributing institution. Also, data from a single institution, in which there are multiple updates of the public database over time, are also less secure from a determined attacker.

There are a variety of encryption protocols suitable for such purposes [4,42]:

- double-brokered encryption;
- one-time-pad encryption (lookup table);
- public–private encryption.

The emerging US federal paradigm for using de-identified medical data for research purposes is minimal risk. That is, if one employs only data that are collected in the ordinary diagnosis and treatment of patients, and there is no change in patient management as a result of the research, including no pressure on the patient to accept or refuse certain management, and no call-back for additional data that might upset the patient or next-of-kin, then the only risk of using such data is the loss of confidentiality to the patient. This is called minimal risk data, and may be possible to use in research projects with a simple exemption from the IRB. There was a well-publicized case of a prominent researcher at a major institution a few years ago who called a family in order to verify certain data regarding a deceased patient under study; this is not allowed under the minimal risk paradigm.

3.4. Expected benefits

Any use of patient data, even de-identified, must be justified to the IRB as having some expected benefits. Legally and ethically one cannot perform data analysis for frivolous or nefarious purposes. However, the Internet is the cheapest and most convenient way to distribute data, and the most accessible to the public which may have legitimate reasons for access. For example, there may be rare-disease interest groups, medical watchdog groups, or even investigators with unconventional scientific perspectives, who have reasonable claims to mine the data, but who could not mount the financial and administrative resources to mine privately held databases. How is this conflict between public access and frivolous use of public human data to be resolved? There is as yet no answer to this question.

3.5. Administrative issues

The emerging US federal guidelines for patient privacy specify a number of administrative policies and procedures that would not ordinarily be required for non-medical data mining [41]. There must be policies to evaluate and certify that appropriate security measures are in place in the research institution. There must be legal contracts between the organization and any outside parties given access to individually identifiable health information, requiring the outside parties to protect the data. There must be contingency plans for response to emergencies, including a data backup plan and a disaster recovery plan. There must be a system of information access control that includes policies for the authorization, establishment, and modification of data access privileges. There must be an ongoing internal review of data-access records, in order to identify possible security violations. The organization must ensure supervision of personnel performing technical systems maintenance activities in order to maintain access authorization records, to ensure that operating and maintenance personnel have proper access, to employ personnel security procedures, and to ensure that system users are trained in system security. There must be termination procedures that are performed when an employee leaves or loses access to the data. There must be security training for all staff, including awareness training for all personnel, periodic security reminders, user education concerning virus protection, user education in the importance of monitoring login failures, password management, and how to report discrepancies.

These and many other rules impose constraints upon medical data miners that other academic researchers would regard as burdensome and stifling to scientific research creativity. Researchers must carefully weigh the perceived need for information such as zip codes (which might be necessary for epidemiological studies), that could also render the data re-identifiable in combination with other information [44].

4. Statistical philosophy

There is an emerging doctrine that data mining methods themselves, especially statistics, and the basic assumptions underlying these methods, may be fundamentally different for medical data. Human medicine is primarily a patient-care activity, and serves only secondarily as a research resource. Generally, the only justification for collecting data in medicine, or refusal to collect certain data, is to benefit the individual patient. Some patients might consent to be involved in research projects that do not benefit them directly, but such data collection is typically very small-scale, narrowly focused, and highly regulated by legal and ethical considerations. The major points of statistical philosophy in medicine may be organized under these general headings:

- Ambush in statistics
- Data mining as a superset of statistics
- Data mining and knowledge discovery process

4.1. Ambush in statistics

Classical statistical tests are designed from the idea of a repeatable experiment, with rules set up in advance. It is not fair to change rules in the middle of an experiment, because the formulas and distributions become meaningless. Thus, classical statistical tests employed in medicine may be subject to ambush. It is surprising, but true, that the intellectual paradigm of classical statistics depends not only upon the actual numbers collected, but also upon one's frame of mind (i.e. a priori assumptions) at the outset of the statistical investigation. If one changes one's mind during the investigation, then one pollutes the interpretation of the data, even if none of the observed values are changed.

Thus, in theory, one plans a clinical trial with a predetermined null hypothesis and a predetermined sample size, and performs the trial until the agreed-upon sample size is reached. One is not allowed to interrupt (ambush) the study if one has reached the numerical value for statistical significance, until the predetermined sample size has been reached. This is because the mathematical reasoning interprets the experimental results based upon the initial experimental design, and the expectations that this design implies, so-called a priori reasoning. One cannot recast the basic assumptions of a statistical test while the investigation is in progress. The dilemma created by this paradigm is that there may be compelling evidence that the a priori assumptions are wrong, long before the predetermined sample size is achieved, and that these wrong assumptions are injuring patients.

Several major, federally sponsored studies have been ambushed under these circumstances, including: chemotherapy of prostate cancer [53]; chemotherapy of breast cancer [27]; oral hypoglycemic therapy of adult-onset diabetes [22]; and steroid therapy of cystic fibrosis [25]. Many federally sponsored clinical trials now mandate the position of an ombudsman [52], who is empowered to interrupt the investigation when the well-being of study-patients is potentially threatened.

There is a similar ambush issue for data mining tools, say artificial neural networks, with the paradigm of the training/test set. When one has exhausted the observations in a training set to train the network, then one can no longer run a test on those observations: that is cheating. Rather, one must use elements from a different set, i.e. the test data set. In medical data the problem is, unlike in animal experiments, that one cannot recruit a few more subjects, perform the experiment again, and create another training set. One is, ethically, required to use the same observations over and over.

Another difficulty with the ambush paradigm in artificial neural networks is that one may wish to examine the entire data set for other reasons than network training, such as building an index of cases as a tissue-resource or rare-disease epidemiologic resource [32]. In this case, it is self-defeating to conceal a subset of cases from the training set. On the other hand, if one examines all the cases, then one pollutes the a priori reasoning required for training–test-set studies.

A related ambush issue is that different outcomes in a statistical distribution may not be perfectly random, but constrained by the fact that certain combinations of medical events are either common or rare. Typically, these events are recognized as common or rare by medical practitioners, but exact probabilities are not known. Can one lend credence to a

significant statistical result when the natural history of disease, and the likely events in this hypothesis, was nowhere used in formulating the null hypothesis?

Yet another ambush issue: *fairness* is required in statistical evaluations between competing medical hypotheses. That is, some statistical tests are designed, not as a search for truth, but as a search for the winning medical scientist. As a federal grant administrator recently observed, many medical researchers who seek grant funding typically formulate their ideas in terms of ''fairness for me'', rather than ''fairness for competing medical hypotheses.'' Working through some of these intellectual enigmas might provide suitable future employment for philosophers and theoretical statisticians, as well as for data miners.

4.2. Data mining as a superset of statistics

Although data mining shares a great deal in common with statistics, since both strive toward discovering some structure in data, data mining also draws heavily from many other disciplines, most notably machine learning and database technology. Data mining differs from statistics in that it must deal with heterogeneous data fields, not just heterogeneous numbers, as is the case in statistics. The best example of heterogeneous data is medical data that, say contains images like SPECT, signals like ECG, clinical information like temperature, cholesterol levels, urinalysis data, etc. as well as the physician's interpretation written in unstructured English. More success stories in data mining are due to advances in database technology rather than to advances in data mining tools. It is only after a subset of data is selected from a large database that most data mining tools are actually applied. Below we comment on some unique features of medical data.

- Because of the sheer volume and heterogeneity of medical databases, it is unlikely that any current data mining tool can succeed with raw data [36]. The tools may require extracting a sample from the database, in the hope that results obtained in this manner are representative for the entire database. Dimensionality reduction can be achieved in two ways. By sampling in the patient-record space, where some records are selected, often randomly, and used afterwards for data mining; or sampling in the feature space, where only some features of each data record are selected.
- Medical databases are constantly updated by, say, adding new SPECT images (for an existing or new patient), or by replacement of the existing images (say, a SPECT had to be repeated because of technical problems). This requires methods that are able to incrementally update the knowledge learned so far.
- The medical information collected in a database is often incomplete, e.g. some tests were not performed at a given visit, or imprecise, e.g. ''the patient is weak or diaphoretic.''
- It is very difficult for a medical data collection technique to entirely eliminate noise. Thus, data mining methods should be made less sensitive to noise, or care must be taken that the amount of noise in future data is approximately the same as that in the current data.
- In any large database, we encounter a problem of *missing values*. A missing value may have been accidentally not entered, or purposely not obtained for technical, economic, or ethical reasons. One approach to address this problem is to substitute missing values

with *most likely values*; another approach is to replace the missing value with *all possible values* for that attribute. Still another approach is intermediate: specify a *likely range of values*, instead of only one most likely. The difficulty is how to specify the range in an unbiased manner.

The missing value problem is widely encountered in medical databases, since most medical data are collected as a byproduct of patient-care activities, rather than for organized research protocols, where exhaustive data collection can be enforced. In the emerging federal paradigm of minimal risk investigations, there is preference for data mining solely from byproduct data. Thus, in a large medical database, almost every patient-record is lacking values for some feature, and almost every feature is lacking values for some patient-record.

- The medical data set may contain redundant, insignificant, or inconsistent data objects and/or attributes. We speak about inconsistent data when the same data item is categorized as belonging to more than one mutually exclusive category. For example, a serum potassium value incompatible with life obtained from a patient who seemed reasonably healthy at the time the serum was drawn. A common explanation is that the specimen was excessively shaken during transport to the laboratory, but one cannot assume this explanation without additional investigation and data, which may be impractical in a data mining investigation.
- Often we want to find natural groupings (clusters) in large dimensional medical data. Objects are clustered together if they are similar to one another (according to some measure), and at the same time are dissimilar from objects in other clusters. A major concern is how to incorporate medical domain knowledge into the mechanisms of clustering. Without that focus and at least partial human supervision, one can easily end up with clustering problems that are computationally infeasible [42,44], or results that do not make sense.
- In medicine, we are interested in creating understandable to human descriptions of medical concepts, or models. Machine learning, conceptual clustering, genetic algorithms, and fuzzy sets are the principal methods used for achieving this goal, since they can create a model in terms of intuitively transparent if . . . then . . . rules. On the other hand, unintuitive black box methods, like artificial neural networks, may be of less interest.

4.3. Data mining and knowledge discovery process

It is important in medical data mining, as well as in other kinds of data mining, to follow an established procedure of knowledge discovery, from problem specification to application of the results. It is even more important for future successful medical applications to semi-automate the DMKD process. We elaborate on the two in the section that follows [16].

The goal of the DMKD process is to develop a set of processing steps that should be followed by practitioners when conducting data mining projects. The purpose of such design is to help to plan, work through, and reduce the cost of the project by outlining the DMKD process, and by describing procedures performed in each of the steps.

Knowledge discovery is a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from large collections of data [18]. One of

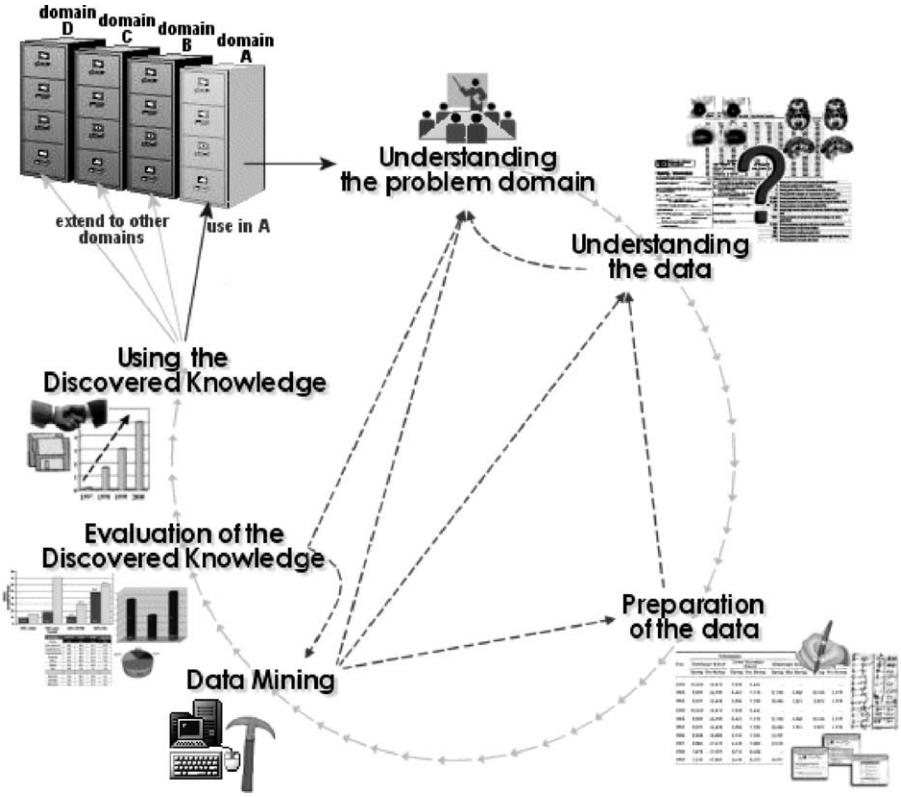

Fig. 1. The six-step DMKD process model.

the knowledge discovery steps is the data mining step concerned with actual extraction of knowledge from data. Design of a framework for a knowledge discovery process is important. Researchers have described a series of steps that constitute the KD process, which range from few steps that usually include data collection and understanding, data mining and implementation, to more sophisticated models like the nine-step model proposed by Fayyad et al. [19], or the six-step DMKD process model by Cios et al. [13] and Cios and Moore [14]. The latter model has been successfully applied to several medical problem domains [24,40]; it added several extensions to the CRISP-DM model [17].

The DMKD process is visualized in Fig. 1 [16]. The important issues are the iterative and interactive aspects of the process. Since any changes and decisions made in one of the steps can result in changes in later steps, the feedback loops are necessary.

One of the technologies that can help in carrying out the DMKD process is XML (eXtensible Markup Language) [6]. All formatted text documents consist of text and *markup*. Markup is the set of commands, or *tags*, placed within the text, that control spacing, pagination, linkages to other documents, font style, size, color, and foreign alphabets. On the Internet, the most popular markup language is the *Hypertext Markup*

Language (HTML). In HTML, each *start-tag* begins with <and ends with>; each *end-tag* begins with </and ends with>. Thus, for example, the sequence <B> ... TEXT ... </B> causes the computer monitor or printer to display ... **TEXT** ... in boldface.

HTML owes its popularity on the Internet to its simplicity, its cost-free, non-proprietary status, and the widespread desire among web masters to display web pages with an attractive appearance and composition. Among scientists, the chief shortcoming of HTML is the web programmer's inability to designate *content* as easily as one designates *format* One might wish to designate LIVER-WEIGHT in the same way that one designates a 12-point Arial format or the Greek alphabet in HTML. This capability is offered by XML, a generalization of the HTML concept, that includes definable *content-tags* as well as format-tags. For example, a liver-weight of 1.6 Kg might be XML-tagged as:

<ORGAN><ORGAN-NAME>liver</ORGAN-NAME>
<ORGAN-WEIGHT>1.6 Kg</ORGAN-WEIGHT></ORGAN>

In medical DMKD, advances in XML technology will lead to greater sharing of medical research data across the Internet. The only serious remaining obstacle is agreement among professional medical societies on a consistent nomenclature for these XML-tags, and perhaps some paragraphing conventions. A cost-free standard medical nomenclature is already available to researchers as the Unified Medical Language System (UMLS) of the US National Library of Medicine (USNLM), which contains over 770,000 concepts and over 2 million synonyms in the 2002 edition [47]. XML allows one to store and describe structured or semi-structured data and their relationships.

One of the most exciting developments in XML standards are their emergence as a public, open-source standard for message traffic and biomedical data exchange. Open-source, non-proprietary standards are virtually a given for such purposes; for how can one exchange messages or data if one private company owns the language of discourse, and thus by its pricing policies can lock out certain customers or competitors?

An example of such a standard is the XML-open-source-standard for tissue micro-arrays (TMAs) [5]. In routine diagnostic pathology, a slide consists of a single piece of tissue from a single patient, laid over a 25 mm $\times$ 76 mm glass rectangle, which is stained for a particular chemical or gene, and which is evaluated by light microscopy. On a TMA, hundreds or even thousands of tissue fragments are precisely laid out in a checkerboard arrangement on a single slide. The position of each tissue fragment, and the corresponding patient history, are kept in a database. After staining and microscopic examination, a single slide can yield results and correlations on thousands of patients. Advantages listed by the author for open data standardization include: sharing TMA data among collaborators; submitting TMA data to journals or data repositories; merging TMA data with other TMA datasets; updating TMA datasets from related datasets, etc.

In the face of such complexity, it is important for individual research laboratories to share information [5,48]. Applications for US government funding in biomedicine that list open source software and open source solutions are actually much more likely to get funded than applications that use proprietary tools. This is emerging as a guiding principle of almost every federally funded genomics and proteomics effort.

One of the important features of XML is that it can be used to exchange data in a platform-independent manner. XML is easy to use, and one can employ a large number of

off-the-shelf tools for automatic processing of XML. From the DMKD point of view, XML is the key technology to:

- standardize communication between diverse data mining (DM) tools and databases;
- build standard data repositories sharing data between different DM tools that work on different software platforms;
- implement communication protocols between the DM tools;
- provide a framework for integration and communication between different DMKD steps. The information collected during domain and data understanding steps can be stored as XML documents. Then, the information can be used for data preparation and data mining steps, as a source of already-accessible information, cross-platforms, and cross-tools.

Since DMKD is a very complex process, the ability to automate (semi-automate) and consolidate the DMKD process cannot be overstated. Semi-automating the DMKD process is a very complex task, but necessary for many applications. User input is necessary to perform the entire DMKD task, since often knowledge about the domain and data must be provided by domain experts. In addition, some guidance is needed, such as evaluating results at each step by those performing medical data mining. To semi-automate the process, several technologies are necessary: a data repository that stores the data, background knowledge, and models; protocols for sending data and information between data repositories and DM tools, and between different DM tools; and finally standards for describing data and models.

XML has been one of the most exciting research and applications tools to emerge over last few years. This technology, along with other technologies that are built on top of XML can provide solutions to the problem of semi-automating of the DMKD process.

XML is a markup language for documents that contain structured information. Structured information consists of data-content (numbers, character strings, images, etc.) and information on what role that content plays (e.g. a rule is built out of selectors, i.e. pairs of attributes (name and value)). XML defines a markup standard, in other words, a standard to identify structures in documents.

XML is primarily used to create, share, and process information. XML enables user to define tags (element names) that are specific to a particular purpose. XML-tags are used to describe the meaning of information in a precisely defined manner. Because of these features, processing of XML documents can be performed automatically.

XML technology is widely used in industry to transfer and share data. One of the most important properties of XML is that current database management systems (DBMS) already support the XML standard. From the DMKD point of view, this means that XML can be used as a transport medium between, say, DM tools and XML-based knowledge repository, and the DBMS that is used to store the data.

There are two mainstream developments of the DBMS capable of handling XML documents: native XML DBMS, and XML-enabled DBMS [16]:

- The majority of *native XML DBMS* are based on the standard DB physical storage model, like relational, object-relational, or object-oriented, but they use an XML document as the fundamental storage unit, just as a relational DBMS uses an n-tuple

as its fundamental storage unit. Their main advantage is the possibility of storing an XML document and then retrieving the same document without losing any information, both on the structural and data levels, which is not yet possible using the XML-enabled DBMS [28,43,34].

- The *XML-enabled DBMS* incorporates the XML document into the traditional database technology. There are numerous examples of successful approaches that use extensions to XML. Examples of commercial XML-enabled DBMS, all using the relational model are Oracle 8i [2], DB2 [9], Informix [23], Microsoft SQL Server 2000 [45], and Microsoft Access 2002 [54].

The DMKD community has developed several successful DM methods over the past several years; but just having a variety of DM methods does not solve any current problems of the DMKD, such as the necessity to integrate DM methods, to integrate them with the DBMS, and to provide support for novice users [21].

Let us define the difference between *data mining methods* and *data mining tools*. A DM method is simply an implementation of the DM algorithm, while a DM tool is the DM method that can communicate and operate in the DMKD environment. XML and XML-based technology provides tools for transforming DM methods into DM tools, combining them into DM toolboxes, and most importantly, for semi-automating the DMKD process. Several possible scenarios of how XML can be used during data preparation step, and as a medium to store, retrieve, and use the domain knowledge is described [8]. XML is a universal format for storing structured data. Since it is supported by current DBMS, it is already becoming a de facto standard, not only for data transportation but also for data storage [1].

The above described technologies can, and we think will, be used to support all stages of the DMKD process. The diagram showing design of the DMKD model based on these technologies, which enables semi-automation of the DMKD process, is shown in Fig. 2 [16].

The data database and knowledge database can be stored using a single DBMS that supports an XML format, because the PMML that is used to store the knowledge complies with the XML format. We separate the two to underscore the difference in format and functionality of the information they both store. The data database is used to store and query the data. All the DMKD steps, however, can store information and communicate using the knowledge database. The biggest advantages of implementing the knowledge database are automation of knowledge storage and retrieval; sharing the discovered knowledge between different domains, and support for semi-automation of these two DMKD steps: understanding the data, and preparation of the data. The architecture, shown in Fig. 2, has the additional advantage of supporting the iterative and interactive aspects of the DMKD process. This design supports the entire DMKD process rather than only a single DM step.

During the past few years businesses have shown growing interest in DMKD. The biggest DBMS vendors, like IBM, Microsoft, and Oracle, have integrated some of the DM tools into their commercial systems. IBM's DM tool, called Intelligent Miner, which integrates with DB2, consists of three components: Intelligent Miner for Data; Intelligent Miner for Text; and Intelligent Miner Scoring [3,39,37]. The Intelligent Miner for Data

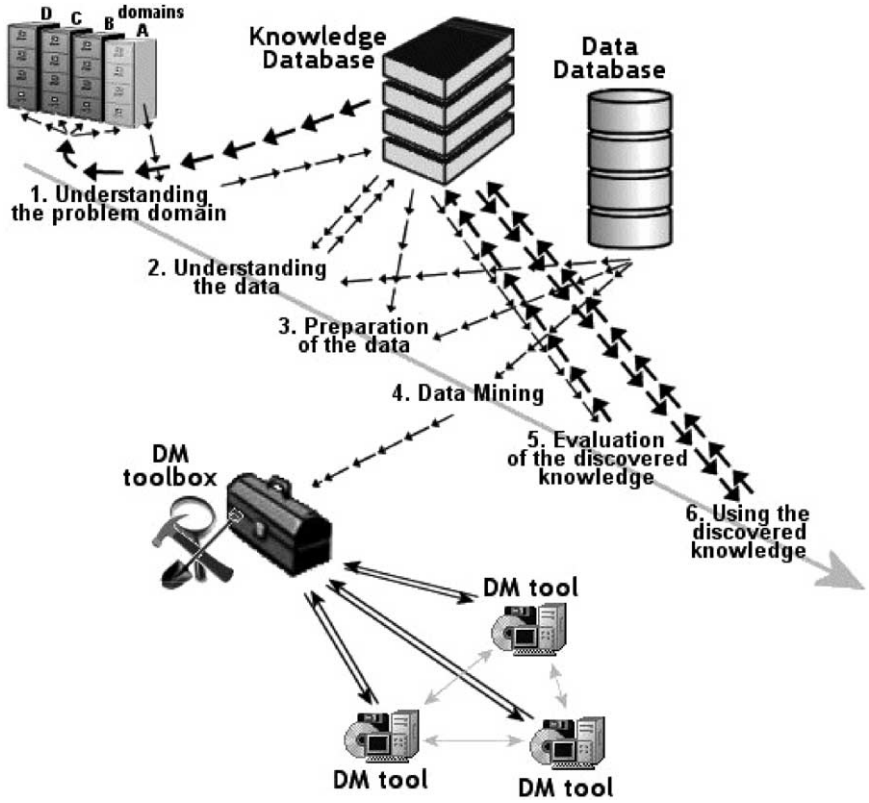

Fig. 2. The automation of the DMKD process using XML-based technologies.

employs clustering based on the Kohonen neural network, factor analysis, linear and polynomial regression, and decision trees to discover associations and patterns in data. The Intelligent Miner for Text includes search engine, Web access tools, and text analysis tools. Intelligent Miner Scoring is the DM component designed to work in real-time. The Intelligent Miner incorporates some data preprocessing methods, like feature selection, sampling, aggregation, filtering, cleansing, and data transformations like principal component analysis. The Microsoft SQL Server2000 incorporates two DM algorithms: decision trees and clustering [46]. The implementation is based on the OLE DB-DM specification. Oracle provides a DM tool called Oracle Darwin®, which is part of the Oracle Data Mining Suite [35]. It supports DM algorithms like neural networks, classification and regression trees, memory-based reasoning (based on the *k*-nearest neighbor approach), and clustering (based on *k*-means unsupervised learning algorithm). Their solution integrates with the Oracle 9i DBMS.

The above mentioned products provide tools to automate several steps of the DMKD process, like preparation of the data and DM. However, they only partially solve the issue of semi-automation of the entire DMKD process, because they do not provide the overall framework for carrying out the DMKD process.

Technologies like XML will play significant role in the design of the next-generation DMKD process framework and DM tools. These technologies will make it possible to build DM toolboxes which span multiple DM tools, to build knowledge repositories, to communicate and interact between DM tools, DBMS and knowledge repositories, and most importantly to semi-automate the entire DMKD process. These technologies also can be used to deploy DMKD processes that will include elements running on different platforms, since they are platform-independent. Because these technologies are standards influenced mostly by what is happening outside of the DMKD industry, they may move DMKD to the new level of usability. New users, who follow these standards, in spite of their possibly limited knowledge about the DMKD, will be exposed and attracted to DMKD research and its applications.

In addition to the design and implementation of a new DMKD framework, a more traditional course of action must be carried out, including design and implementation of a new generation of high performance DM systems that incorporate multiple DM methods [56], that are capable of mining heterogeneous sources of knowledge like multimedia data [57], that can visualize the results, and that handle huge amounts of complex data. One of the goals in designing such systems should be the design of much better user interfaces. This will result in a wider acceptance of the products, particularly by midsize and small companies where users may have only limited technical skills. Another very important issue is to learn about the user perception of the novelty, understandability, and easiness of the generated by the DMKD process knowledge. We must take into account the human cognitive processes, learn how people assimilate new knowledge to increase the usefulness of the new generation of better DMKD tools [38] if we are to make progress.

5. Special status of medicine

Finally, medicine has a special status in science, philosophy, and daily life. The outcomes of medical care are life-or-death, and they apply to everybody. Medicine is a necessity, not merely an optional luxury, pleasure, or convenience.

Among all the professions, medicine has the longest apprenticeship. Most medical specialists in the USA require at least 11 years of training after high school graduation, and some surgical subspecialties require up to 16. In the USA, medical care costs consume one-seventh of the gross domestic product. Licensed physicians represent about 0.2% of the US population; the incomes for full-time physicians are in the top several percent; and the average physician causes seven times his/her income to be spent on services ordered. The average citizen has high expectations of medicine and its practitioners. A sick person is expected to recover. Physicians are expected to be ethical, caring, and not too greedy. Medicine is a popular subject for the popular media. Medical care is sometimes risky, but when it fails, the desire for legal revenge is intense and punitive. Medical information about the individual patient is considered highly private, and the general public is extremely fearful about disclosure (US, 1999). We all enjoy the benefits of medical research conducted on other patients, but we are very often reluctant to contribute or release our own information for such purposes. When medical data are published it is expected that the researchers will maintain the dignity

of the individual patient, and that the results will be used for socially beneficial purposes [41].

It has been suggested that scientific truths are fundamentally amoral; they can be used for good or evil [11]. Yet although medicine is based upon science, there are certain tests that may not be performed, certain questions that may not be asked, and certain conclusions that may not be drawn, because of medicine's special status. There has been a vigorous public debate, for example, on whether data obtained from human experimentation, such as those obtained in Nazi Germany, should be published and used. Data from similar experiments, performed on laboratory animals, would be regarded as valid biological data without any further consideration. As we have seen in this article, this special status of medicine pervades our attitudes about medical data mining, as well as our attitudes about medical diagnosis and treatment.

6. Summary

In summary, data mining in medicine is distinct from that in other fields, because the data are heterogeneous; special ethical, legal, and social constraints apply to private medical information; statistical methods must address these heterogeneity and social issues; and because medicine itself has a special status in life.

Data from medical sources are voluminous, but they come from many different sources, not all commensurate structure or quality. The physician's interpretations are an essential component of these data. The accompanying mathematical models are poorly characterized compared to the physical sciences. Medicine is far, far from the intellectual gold-standard of a canonical form for its basic concepts.

The ethical, legal, and social limitations on medical data mining relate to privacy and security considerations, fear of lawsuits, and the need to balance the expected benefits of research against any inconvenience or possible injury to the patient.

Methods of medical data mining must address the heterogeneity of data sources, data structures, and the pervasiveness of missing values for both technical and social reasons. The natural history of disease affects statistical hypotheses in an unknown way. Statistical hypothesis tests often take the form of an ambush or a contest with a winner and a loser. The relevance of this model to the natural processes of medicine is questionable.

For all its perils, medical data mining can also be the most rewarding. For an appropriately formulated scientific question, thousands of data-elements can be brought to bear on finding a solution. For an appropriately formulated medical question, finding an answer could mean extending a life, or giving comfort to an ill person. These potential rewards more than compensate for the many extraordinary difficulties along the pathway to success.

References

[1] Apps E. New mining industry standards: moving from monks to the mainstream. PC AI 2000;14(6):46–50.
[2] Banerjee S, Krishnamurthy V, Krishnaprasad M, Murthy R. Oracle8I—the XML-enabled data management system. In: Proceedings of the 16th International Conference on Data Engineering, San Diego (CA), 2000. p. 561–8.

[3] Bauer, CJ. Data mining digs. Special advertising recruitment supplement to the Washington Post. Washington Post, Sunday, 15 March 1998.
[4] Berman JJ, Moore GW, Hutchins GM. Maintaining patient confidentiality in the public domain Internet autopsy database (IAD). Proc AMIA Annu Fall Symp 1996:328–32.
[5] Berman JJ. Tissue microarray data exchange standards: frequently asked questions, 2002 (http://www.pathinfo.com/jjb/tmfaqv1.htm).
[6] Bray T, Paoli J, Maler E. eXtensible Markup Language (XML) 1.0. 2nd ed. W3C recommendation, October 2000 (http://www.w3.org/TR/2000/REC-xml-20001006).
[7] Brewka G, Dix J, Konolige K. Nonmonotonic reasoning: an overview. CSLI Lecture Notes No. 73, ISBN 1-881526-83-6, 1997. p. 179.
[8] Büchner AG, Baumgarten M, Mulvenna MD, Böhm R, Anand SS. Data mining and XML: current and future issues. In: Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00), Hong Kong, 2000. p. 127–31.
[9] Cheng J, Xu J. IBM DB2 extender. In: Proceedings of the 16th International Conference on Data Engineering, San Diego (CA), 2000. p. 569–73.
[10] Ceusters W. Medical natural language understanding as a supporting technology for data mining in healthcare. In: Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg: Springer, 2000. p. 32–60 [chapter 3].
[11] Changeux J-P, Connes A. Conversations on mind, matter, and mathematics [DeBevoise MB, Trans.]. Princeton (NJ): Princeton University Press, 1995.
[12] Cios KJ, Pedrycz W, Swiniarski R. Data mining methods for knowledge discovery. Boston: Kluwer Academic Publishers, 1998.
[13] Cios KJ, Teresinska A, Konieczna S, Potocka J, Sharma S. Diagnosing myocardial perfusion SPECT bull's-eye maps—a knowledge discovery approach. IEEE Eng Med Biol 2000;19(4):17–25[special issue on medical data mining and knowledge discovery].
[14] Cios KJ, Moore GW. Medical data mining and knowledge discovery: an overview. In: Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg: Springer, 2000. p. 1–16 [chapter 1].
[15] Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg: Springer, 2001 (http://www.springer.de/cgi-bin/search_book.pl?isbn=3-7908-1340-0).
[16] Cios KJ, Kurgan LA. Trends in data mining and knowledge discovery. In: Pal NR, Jain LC, Teodoresku N, editors. Knowledge discovery in advanced information systems. Berlin: Springer, 2002, in press.
[17] CRISP-DM, 1998 (www.crisp-dm.org).
[18] Fayyad UM, Piatesky-Shapiro G, Smyth P, Uthurusamy R. Advances in knowledge discovery and data mining. Boston: AAAI Press/MIT Press, 1996.
[19] Fayyad UM, Piatetsky-Shapiro G, Smyth P. Knowledge discovery and data mining: towards a unifying framework. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD96), Portland (OR): AAAI Press, 1996.
[20] Friedman C, Hripcsak GW. Evaluating natural language processors in the clinical domain. Meth Inform Med 1998;37:334–44.
[21] Goebel M, Gruenwald L. A survey of data mining software tools. SIGKDD Explor 1999;1(1):20–33.
[22] Goldner MG, Knatterud GL, Prout TE. Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. 3. Clinical implications of UGDP results. JAMA 1971;218(9):1400–10.
[23] Informix object translator, 2001 (http://www.informix.com/idn-secure/webtools/ot/).
[24] Kurgan LA, Cios KJ, Tadeusiewicz R, Ogiela M, Goodenday LS. Knowledge discovery approach to automated cardiac SPECT diagnosis. Artif Intell Med 2001;23(2):149–69.
[25] Lai HC, FitzSimmons SC, Allen DB, Kosorok MR, Rosenstein BJ, Campbell PW, Farrell PM. Risk of persistent growth impairment after alternate-day prednisone treatment in children with cystic fibrosis. N Engl J Med 2000;342(12):851–9.
[26] Manning CD, Schuetze H. Foundations of statistical natural language processing. Cambridge (MA): MIT Press, 2000.
[27] Mansour EG, Gray R, Shatila AH, Osborne CK, Tormey DC, Gilchrist KW, Cooper MR, Falkson G. Efficacy of adjuvant chemotherapy in high-risk node-negative breast cancer: an intergroup study. N Engl J Med 1989;320(8):485–90.

[28] McHugh J, Abiteboul S, Goldman R, Quass D, Widom J. Lore: a database management system for semistructured data. SIGMOD Rec 1997;26(3):54–66.
[29] Moore GW, Hutchins GM. Effort and demand logic in medical decision making. Metamedicine 1980;1:277–304.
[30] Moore GW, Hutchins GM, Miller RE. Token swap test of significance for serial medical databases. Am J Med 1986;80:182–90.
[31] Moore GW, Berman JJ, Hanzlick RL, Buchino JJ, Hutchins GM. A prototype Internet autopsy database: 1625 consecutive fetal and neonatal autopsy facesheets spanning 20 years. Arch Pathol Lab Med 1996;120:782–5.
[32] Moore GW, Berman JJ. Anatomic pathology data mining. In: Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg: Springer, 2000. p. 61–108 [chapter 4].
[33] Nagao M. Machine translation. In: Shapiro SC, editor. Encyclopedia of artificial intelligence, vol. 2. M-Z (New York): Wiley/Interscience, 1992. p. 898–902.
[34] Native XML DBMS, 2001 (http://www.rpbourret.com/xml/XMLDatabaseProds.htm).
[35] Oracle data mining suite, Oracle Darwin®, 2001 (http://technet.oracle.com/products/datamining/htdocs/datasheet.htm).
[36] Pawlak Z. Rough classification. Int J Man-Mach Stud 1984;20:469–83.
[37] Rennhackkamp M. IBM's intelligent family. DBMS, August 1998 (http://www.dbmsmag.com/980d17.html).
[38] Pazzani MJ. Knowledge discovery from data? IEEE Intell Syst 2000;March/April:10–3.
[39] Reinschmidt J, Gottschalk H, Kim H, Zwietering D. Intelligent miner for data: enhance your business intelligence. IBM international technical support organization (IBM Redbooks). IBM Corporation, 1999.
[40] Sacha JP, Cios KJ, Goodenday LS. Issues in automating cardiac SPECT diagnosis. IEEE Eng Med Biol 2000;19(4):78–88.
[41] Saul JM. Legal policy and security issues in the handling of medical data. In: Cios KJ, editor. Medical data mining and knowledge discovery. Heidelberg: Springer, 2000. p. 17–31 [chapter 2].
[42] Schneier, B. Applied cryptography. Protocols, algorithms, and source code in C. 2nd ed. New York: Wiley, 1996.
[43] Schoening H. Tamino—a DBMS designed for XML. In: Proceedings of the 17th IEEE International Conference on Data Engineering, Los Alamos (CA, USA), 2001. p. 149–54.
[44] Sweeney L. Computational disclosure control: a primer on data privacy protection. PhD Thesis. Spring: Massachusetts Institute of Technology, Draft, 2001 (http://www.swiss.ai.mit.edu/classes/6.805/articles/privacy/sweeney-thesis-draft.pdf).
[45] SQL Server magazine. SQL Server magazine: the XML files, 2000 (http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnsqlmag2k/html/TheXMLFiles.asp).
[46] Tang Z, Kim P. Building data mining solutions with SQL Server 2000. DM Review. White Paper Library, 2001 (http://www.dmreview.com/whitepaper/wid292.pdf).
[47] US National Library of Medicine. Unified medical language system. 13th ed. Knowledge sources, 2002 (http://www.nlm.nih.gov/research/umls).
[48] US Department of Health and Human Services. National Institutes of Health Statement on Data Sharing, 2002 (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-035.html).
[49] US Department of Health and Human Services. 45 CFR (Code of Federal Regulations). Parts 160–164. Standards for Privacy of Individually Identifiable Health Information. Final Rule. Fed Regist 28 December 2000;65(250):82461–610 (http://aspe.hhs.gov/admnsimp/).
[50] US Code of Federal Regulations, 45 CFR Subtitle A, 10-1-95 ed. Part 46. 101 (b) (4). US Department of Health and Human Services (Common Rule). 56 Fed Regist 18 June 1991;56:28003 (http://ohrp.osophs.dhhs.gov/humansubjects/guidance/45cfr46.htm).
[51] US National Cancer Institute's Confidentiality Brochure, 2000 (http://www-cdp.ims.nci.nih.gov/policy.html).
[52] US Food and Drug Administration. Delegations of authority and organization; Office of the Commissioner—FDA. Final rule. Fed Regist 21 November 1991;56(225):58758.
[53] US Veterans Administration Co-operative Urological Research Group. Treatment and survival of patients with cancer of the prostate. The Veterans Administration Co-operative Urological Research Group. Surg Gynecol Obstet 1967;124(5):1011–17.

[54] XML and Access 2002. Exploring XML and Access 2002, 2001 (http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnacc2k2/html/odc_acxmllnk.asp).
[55] Zadeh LA. Fuzzy sets and information granularity. In: Gupta MM, et al., editors. Advances in fuzzy set theory and applications. Dordrecht: North-Holland, 1979. p. 3–18.
[56] Yaginuma Y. High-performance data mining system. Fujitsu Sci Tech J 2000;36(2):201–10[special issue: information technologies in the Internet era].
[57] Zaïane OR, Han J, Li ZN, Hou J, Mining multimedia data. In: Proceedings of the CASCON'98: Meeting of minds, Toronto, Canada, 1998. p. 83–96.