INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 77 (2008) 81–97

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Review

# Predictive data mining in clinical medicine: Current issues and guidelines

*Riccardo Bellazzi*[a,*], *Blaz Zupan*[b,c]

[a] *Dipartimento di Informatica e Sistemistica, Università di Pavia, via Ferrata 1, 27100 Pavia, Italy*
[b] *Faculty of Computer Science, University of Ljubljana, Slovenia*
[c] *Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX, United States*

ARTICLE INFO

*Article history:*
Received 27 October 2006
Accepted 17 November 2006

*Keywords:*
Data mining
Predictive models
Clinical medicine
Data mining process
Data analysis

ABSTRACT

*Background:* The widespread availability of new computational methods and tools for data analysis and predictive modeling requires medical informatics researchers and practitioners to systematically select the most appropriate strategy to cope with clinical prediction problems. In particular, the collection of methods known as 'data mining' offers methodological and technical solutions to deal with the analysis of medical data and construction of prediction models. A large variety of these methods requires general and simple guidelines that may help practitioners in the appropriate selection of data mining tools, construction and validation of predictive models, along with the dissemination of predictive models within clinical environments.

*Purpose:* The goal of this review is to discuss the extent and role of the research area of predictive data mining and to propose a framework to cope with the problems of constructing, assessing and exploiting data mining models in clinical medicine.

*Methods:* We review the recent relevant work published in the area of predictive data mining in clinical medicine, highlighting critical issues and summarizing the approaches in a set of learned lessons.

*Results:* The paper provides a comprehensive review of the state of the art of predictive data mining in clinical medicine and gives guidelines to carry out data mining studies in this field.

*Conclusions:* Predictive data mining is becoming an essential instrument for researchers and clinical practitioners in medicine. Understanding the main issues underlying these methods and the application of agreed and standardized procedures is mandatory for their deployment and the dissemination of results. Thanks to the integration of molecular and clinical data taking place within genomic medicine, the area has recently not only gained a fresh impulse but also a new set of complex problems it needs to address.



* *Corresponding author*. Tel.: +39 0382 505511; fax: +39 0382 505373.
E-mail address: Riccardo.Bellazzi@unipv.it (R. Bellazzi).
1386-5056/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2006.11.006

## Contents



## 1. Introduction

Over the last few years, the term 'data mining' has been increasingly used in the medical literature. In general, the term has not been anchored to any precise definition but to some sort of common understanding of its meaning: the use of (novel) methods and tools to analyze large amounts of data. Data mining has been applied with success to different fields of human endeavor, including marketing, banking, customer relationship management, engineering and various areas of science. However, its application to the analysis of medical data – despite high hopes – has until recently been relatively limited. This is particularly true of practical applications in clinical medicine which may benefit from specific data mining approaches that are able to perform predictive modeling, exploit the knowledge available in the clinical domain and explain proposed decisions once the models are used to support clinical decisions. The goal of predictive data mining in clinical medicine is to derive models that can use patient-specific information to predict the outcome of interest and to thereby support clinical decision-making. Predictive data mining methods may be applied to the construction of decision models for procedures such as prognosis, diagnosis and treatment planning, which – once evaluated and verified – may be embedded within clinical information systems.

In this paper, we give a methodological review of data mining, focusing on its data analysis process and highlighting some of the most relevant issues related to its application in clinical medicine. We limit the paper's scope to predictive data mining whose methods are methodologically ripe and often easily available and may be particularly suitable for the class of problems arising from clinical data analysis and decision support.

## 2. Background

Data mining is the process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst [1]. Coined in the mid-1990s, the term data mining has today become a synonym for 'Knowledge Discovery in Databases' which, as proposed by Fayyad et al. [2], emphasized the data analysis process rather than the use of specific analysis methods. Data mining problems are often solved by using a mosaic of different approaches drawn from computer science, including multi-dimensional databases, machine learning, soft computing and data visualization, and from statistics, including hypothesis testing, clustering, classification and regression techniques. The craft of data mining lies in the appropriate choice and combination of these techniques to efficiently and reliably solve a given problem.

Data mining tasks can, in general, be classified to tasks of description and prediction. While description aims at finding human-interpretable patterns and associations, after considering the data as a whole and constructing a model prediction seeks to foretell some response of interest. Although the goals of description and prediction may overlap (the models generated by some prediction methods may point out some interesting patterns), the main distinction is that prediction requires the data to include a special response variable. The response may be categorical or numerical, thus further classifying predictive data mining as, respectively, classification and regression. In this review we address classification problems in particular: while the difference between the two lies mainly in the set of methods used, the data mining process applied to regression and classification problems is quite similar.

Before we go on, we use a simple example to introduce the basic concepts in predictive data mining and to show the application of two popular but quite different data classification techniques.

### 2.1. *Introductory example*

For our example, consider a trauma surgeon specialized in hip arthroplasty who would like to know if and how she can predict a patient's long-term clinical status after the surgery. The fictitious and purposely simplified data set (Fig. 1A) we use in this example, whose structure was inspired by a real study [3], consists of 20 patient records each described by three attributes: 'health', giving the patient's overall health at the time of the operation, 'timing', which tells if the operation was on time or delayed, and 'complications', which reports on the degree of complications occurring during the operation. The data includes the response variable called 'outcome' that reports if the treatment was considered successful as evaluated at the follow-up examination at least 2 years after the operation. The snapshots of visualization of data mining models and results presented in this section were obtained using the Orange data mining suite [4].

Our example shows the use of two different modeling techniques to induce predictive models from our data set. The first one, called a naïve Bayesian classifier, is one of the simplest yet it is a useful and often a fairly accurate predictive data mining method [5]. We build a naïve Bayesian classifier by estimating various probabilities from the data. For instance, using a relative frequency estimate, an unconditional probability for a successful operation $P(\text{outcome} = \text{good})$ is 0.55, as there are 11 out of 20 patients in the data set labeled with this class. This is also the probability of a successful operation that we can predict for a patient in the absence of any other information. Prior probability gets updated when other attribute values are known. Following the naïve Bayesian rule the probability of the outcome is proportional to the prior probability times the conditional probability of the attribute value given the outcome. For instance, if we know that the timing for our patient has been good we update the prior by multiplying it with $P(\text{timing} = \text{good}|\text{outcome} = \text{good}) = 9/11$ thus obtaining 0.846. Similarly, for the outcome = bad, the prior equals to 0.45, the update related to timing is $P(\text{timing} = \text{good}|\text{outcome} = \text{bad}) = 5/9$, thus obtaining 0.250. After normalization, the probability of a good outcome for this patient equals 0.643. Knowing the value of other attributes requires a further update of these probabilities. For instance, knowing that the patient had many complications during the surgery the probability of a good outcome decreases to about 0.5.

The naïve Bayesian classifier is thus comprised of unconditional and conditional probabilities as estimated from the data set. The model can be nicely visualized with a nomogram [6,7] (Fig. 1B), a graphical representation that may serve both for analysis of the model (how and to what extent do particular values of predictive factors influence the outcome) and for making predictions. Our nomogram shows, for instance, that complications = no is an attribute-value combination that is the most significant indicator of a good outcome. On the other side, bad timing reduces the probability of a successful treatment the most. The nomogram in Fig. 1B also shows a classification of a patient with good timing and many complications. Nomograms have been frequently used to represent logistic regression models [8] and a number of them are in regular clinical use [9,10].

An explanation of classifications and model interpretability that allows the domain expert to inspect the inner-workings of the classifiers may both be very important in clinical medicine, where every decision should always be clearly motivated. Another popular data mining technique that addresses both aspects is the induction of decision trees. Decision trees are built through recursive data partitioning, where in each iteration the data is split according to the values of a selected attribute. The recursion stops at 'pure' data subsets which only include instances of the same class. Heuristics that include those for choosing the best split attributes and tree pruning aim at obtaining small but accurate trees by avoiding the overfitting of the data [11,12].

Fig. 1C shows a decision tree as induced from our data set. A simple pruning rule which does not allow data splitting if any resulting set contains fewer than two instances was used. The attribute timing is favored at a root node and splits 20 cases into a group of 14 (right branch, timing = good) and 6 (left branch, timing = bad). Notice that the latter leads to a leaf where the prevailing outcome is bad. Classification with a decision tree means following a path from the root node to the leaf, which also determines the outcome and its probability. For instance, for a patient with good health status, good timing and some complications we would reach the leftmost leaf at the bottom of the figure with a prediction outcome = good and associated probability of 0.75.

Decision trees are praised for their transparency, allowing the decision-maker to examine and understand the decision model and its workings. In addition, each path in the decision tree can be regarded as a decision rule. For instance, for the leftmost leaf at the bottom of the tree from Fig. 1C, the inferred classification rule is:

IF timing of the operation is good AND there were some complications during the operation AND patient's overall health at the time of the operation is good.
THEN a good outcome of the operation is expected, $P(\text{outcome} = \text{good}) = 0.75$.

Similar classification rules can also be inferred directly, that is, without going through the construction of classification trees. A popular approach to such inference is a rule-covering algorithm where the conditional part of the rule is iteratively refined so as to cover mostly the examples from one prevailing class. Once such a rule is discovered, the examples it covers are removed from the training set, and rule discovery continues by running the algorithm on the remaining examples. The procedure terminates once all the examples have been covered. Popular implementations of this approach are the CN2 [13] and AQ families of algorithms [14]; the result of running the former on our example set is presented in Fig. 2.

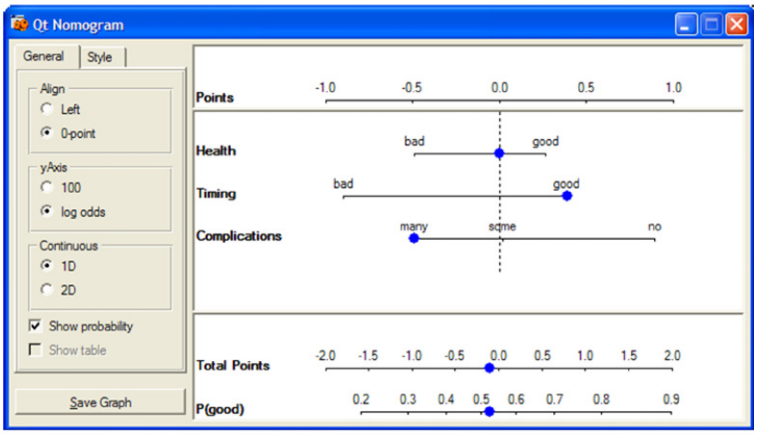
Fig. 3 shows how the naïve Bayesian classifier and the decision tree classified three new cases. While the classifications are qualitatively similar, there are some differences in the predicted probabilities. To estimate how well each of our

(A) Training Data Set

Qt DataTable

| | Health | Timing | Complications | Outcome |
|---|---|---|---|---|
| 1 | good | bad | some | good |
| 2 | bad | bad | many | bad |
| 3 | good | bad | many | bad |
| 4 | good | bad | many | good |
| 5 | good | bad | no | bad |
| 6 | good | bad | many | bad |
| 7 | bad | good | no | good |
| 8 | good | good | no | good |
| 9 | good | good | no | good |
| 10 | good | good | many | bad |
| 11 | bad | good | many | good |
| 12 | good | good | many | good |
| 13 | bad | good | some | good |
| 14 | bad | good | some | bad |
| 15 | bad | good | some | bad |
| 16 | bad | good | some | bad |
| 17 | good | good | some | good |
| 18 | good | good | some | good |
| 19 | good | good | some | good |
| 20 | good | good | some | bad |

(B) Naïve Bayesian Classifier




(C) Decision Tree

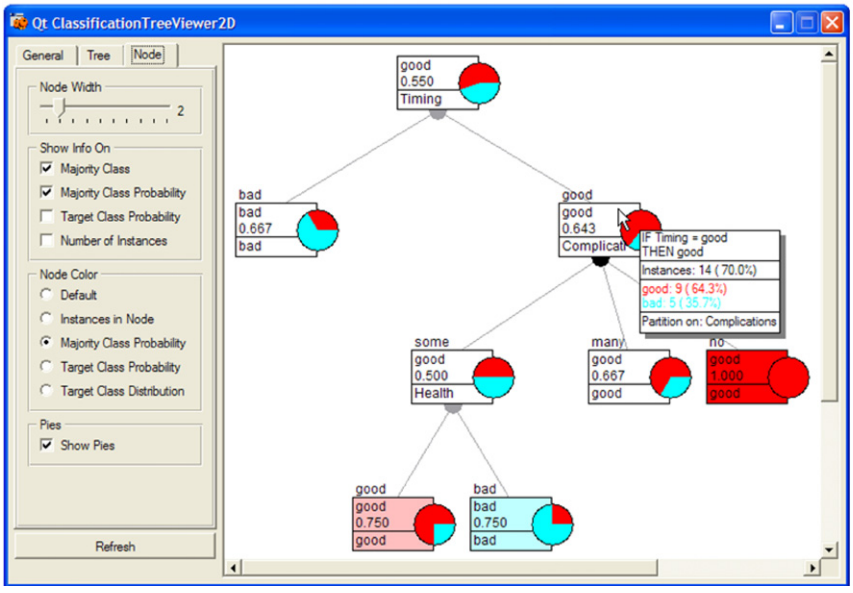



**Fig. 1 – Induction of prediction models. The figure shows an example of a training data set with three attributes, an outcome and 20 instances (A), a nomogram representing a naïve Bayesian classifier (B), and a decision tree developed from the same data set (C). To use a nomogram for prediction, each attribute value relates to the number of points (the topmost scale), which after summation give the total number of points and corresponding probability (the two scales on the bottom of B).**

algorithms would perform on unseen cases, however, data mining most often uses a hold-out procedure that repeatedly builds classifiers for one and tests them on another data set. One such procedure that is useful when data sets are small is called 'leave-one-out'. For our 20 training instances, it in turn selects one case, induces the classifier on the 19 remaining cases and tests it on the selected case. We can then report, for instance, in how many of the 20 runs the classifiers predicted the correct outcome (classification accuracy), or report some other performance measures such as sensitivity or specificity. Fig. 4 shows an example of such a report and also points out that, while both classifiers did not perform well, the decision tree performed somewhat better. The poor performance on this data set, manifested in low classification accuracy as well as the low values of other statistics, can be attributed to the many conflicting cases with the same attribute values but of a different class.

### 2.2. *Predictive data mining (classification) methods*

Predictive data mining methods originate from different research fields and often use very diverse modeling

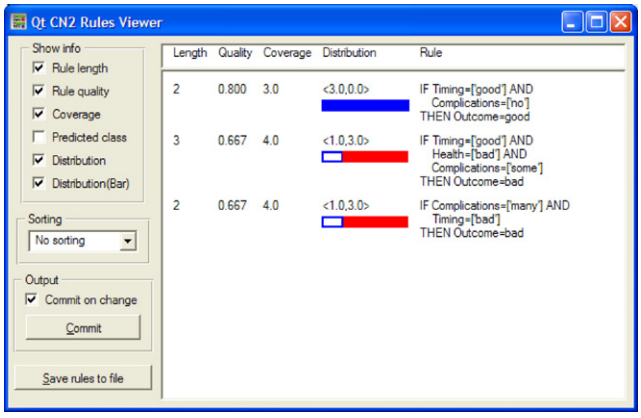



**Fig. 2 – Classification rules inferred by a CN2-like covering algorithm from the data set from Fig. 1A. While the first rule covers only those examples with a good outcome, the class distribution of the other two rules is mixed as the coverage includes one example from the minority (good outcome) class. Rule quality was assessed through a Laplace probability estimate.**

approaches. They come in various flavors and may be compared on the basis of:

- their handling of missing data and noise;
- their treatment of different types of attributes (categorical, ordinal, continuous);
- the presentation of classification models which may or may not allow the domain expert to examine it and understand the inner workings;
- the reduction of the number of tests [15], that is, the reduction of attributes needed to derive the conclusion;
- the computational cost for induction and the use of classification models;
- their ability to explain the decisions reached when models are used in decision-making;
- generalization, that is, the ability to perform well with unseen cases.

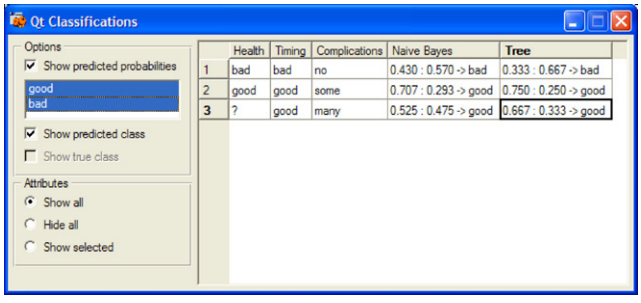



**Fig. 3 – Predictions of the naive Bayesian classifier (Fig. 1B) and decision tree (Fig. 1C) for three different cases. The question mark in the third case for the attribute Health signifies a missing (unknown) value. Probabilities by each classifier are given for both outcomes, 'good' and 'bad' (rightmost two columns, probabilities are separated by a column, the prevailing class label is also shown).**

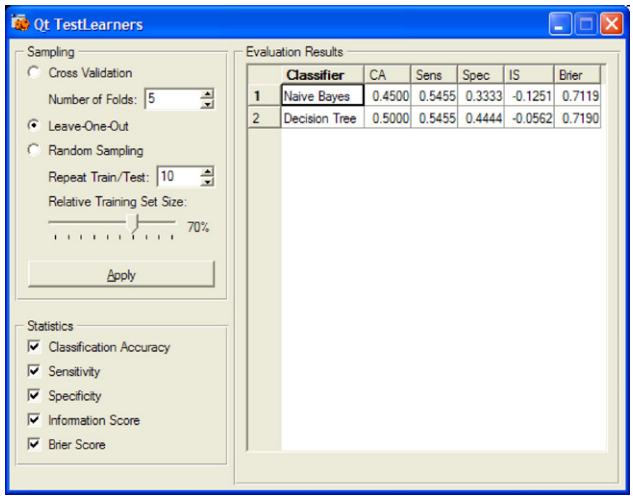



**Fig. 4 – Evaluation results for a naive Bayesian classifier and decision tree inference algorithm on an example data set from Fig. 1A using a 'leave-one-out' test.**

We list here some of the most commonly used predictive data mining methods and order them according to a recent ranking from the relevant pool at KDNuggets (http://www.kdnuggets.com/polls/2006/data_mining_methods.htm, April 2006) which asked data miners to name the techniques they most frequently use:

**Decision trees** use recursive data partitioning, induce transparent classifiers whose performance may suffer from data segmentation: the leaves in decision trees may include too few instances to obtain reliable predictions. The computational complexity of the induction algorithms is low due to powerful heuristics. Most current data mining suites include variants of C4.5 and its successor See5 and CART decision tree induction algorithms [11,12].

**Decision rules** in the form of 'IF condition-based-on-attribute-values THEN outcome-value' may be constructed from induced decision trees as in the C4.5rules [11], or can be derived directly from the data as is the case with AQ and CN2 algorithms [13,14]. While in their performance these algorithms share most of their characteristics with decision trees, they may be computationally more expensive.

**Logistic regression** is a powerful and well-established method from statistics [16]. It is an extension of ordinary regression and it can model a two-valued outcome which usually represents the occurrence or non-occurrence of some event. Like with the naïve Bayesian classifier, the underlying model for probability is multiplicative [17] but uses a more sophisticated method based on a maximum likelihood estimation to determine the coefficients in its probability formula. Handling of the missing attribute values is not straightforward. The model can be nicely represented through a nomogram [6,8].

**Artificial neural networks** were up until recently the most popular artificial intelligence-based data modeling algorithm used in clinical medicine. This is probably due to their good predictive performance, albeit they may have a number of deficiencies [18] that include high sensitivity to the parameters of the method—including those that determine

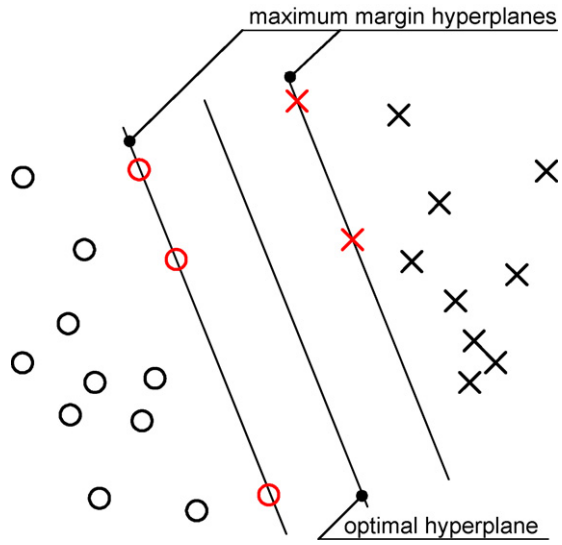



**Fig. 5 – Scatterplot of a two-class data set with maximum-margin hyperplanes found by a support vector machine induction algorithm with a linear kernel. The data instances along the hyperplanes that define the margin (plotted in red) are called support vectors.**

the architecture of the network, high computational cost in training, and induction of the model that may – at best – be hard to interpret by domain experts. Neural networks may be able to model complex non-linear relationships, comprising an advantage over simpler modeling methods like the naïve Bayesian classifier or logistic regression.

**Support vector machines (SVM)** are perhaps today's most powerful classification algorithm in terms of predictive accuracy [19]. They are based on strong mathematical foundations and statistical learning theory [20]. Central to the method is a procedure that finds a hyperplane that separates the examples of different outcomes (see Fig. 5). Being primarily designed for two-class problems, SVMs find a hyperplane with a maximum distance to the closest point of the two classes; such a hyperplane is called the optimal hyperplane. A set of instances that is closest to the optimal hyperplane is called a support vector. Finding the optimal hyperplane provides a linear classifier. Besides such linear kernels, support vector machines are also frequently used with other, non-linear kernels which in essence transform the original attribute space to a new, higher dimensional space in which the linear classifier is inferred. Popular kernel functions are, for instance, polynomial, radial basis and sigmoid functions. The choice of the appropriate kernel should in principle depend on the properties of the data set and problem domain.

For real data sets, the hyperplane that would clearly separate the examples of different classes most often does not exist. To solve this problem, a soft margin method was proposed [21] where the resulting hyperplane splits the data set into two sets that are as clean as possible, that is, where one class prevails to the highest possible degree.

Support vector machines are becoming increasingly popular in medicine and, in particular, in bioinformatics. With the exception of linear kernels, where the structure of the model can be easily revealed through the coefficients that define a linear hyperplane, support vector machines use a formalism that is often unsuitable for interpretation by human experts. As such, and if we are only interested in predictive accuracy, support vector machines are a serious contender to artificial neural networks, especially since their performance may be more robust and depend less on the specific selection of the method's parameters.

**The naïve Bayesian classifier** is an approach we have already introduced. Despite its simplicity, its performance is often at least comparable with other more sophisticated approaches [15,22]. Due to the fast induction of a classifier, it may be used as a baseline algorithm in comparative studies. When surpassed in predictive performance by other more sophisticated algorithms, this often indicates the presence of non-linear interactions between attributes.

**Bayesian networks** are probabilistic graphical models that are able to conveniently express a joint probability distribution over a number of variables through a set of conditional probability distributions. A Bayesian network is a directed acyclic graph where each node represents a stochastic variable and arcs represent a probabilistic dependency between a node and its parents. Each variable $x_i$ is assumed to be independent of its non-descendants given its set of parents, $pa(x_i)$. Under this assumption, known as a Markov assumption, the joint probability distribution of all variables ($x$) can be written following the so-called *chain rule*:

$$p(x) = \prod_{i=1}^{n} p(x_i|pa(x_i))$$

The network is fully specified by a set of conditional probability distributions which quantifies the qualitative relationships between the variables expressed by the graph. Such probability distributions depend on a set of parameters $\theta$, such as the entries of the conditional probability tables for discrete variables or the mean and variance of the Gaussian distribution for continuous variables. Although Bayesian networks have been traditionally applied in medicine as an instrument to perform probabilistic reasoning [23–26], several algorithms and tools are nowadays available to learn both the graph structure and the underlying probabilistic model from the data [27–30]. Bayesian networks can be easily applied in classification problems, where they can be seen as a generalization of the naïve Bayesian classifier by modeling the interactions between the problem variables. They are now increasingly used in both predictive data mining and in bioinformatics [31,32]. The main drawbacks of this method lie in learning the graph structure, which may require a large data set, and in interpretation of the inferred causalities [28].

The algorithms for learning Bayesian networks from data are based on the framework of Bayesian model selection. The goal is to learn the structure $S$ with the highest posterior probability distribution, given a data set $x$. Such a posterior probability distribution can be computed as:

$$p(S|x) = \frac{p(x|S)p(S)}{p(x)} \propto p(x|S)p(S)$$

The posterior is proportional to the product of two terms, namely the marginal likelihood $p(x|S)$, that measures how likely the model is with respect to the data $x$, and the prior

probability of the structure $p(S)$. The marginal likelihood is obtained as the average of the likelihood over the values of the parameter set $\theta$. The marginal likelihood can be computed in close form only when the variables are discrete [29] and when the model is conditionally Gaussian [33].

The comparison of the posterior distribution of the different structures requires the exploration of the space of all possible structures, a problem which turns out to be intractable with a brute-force approach. To cope with this problem, many heuristic algorithms have been proposed in the literature. The most widely applied is the greedy search algorithm K2, described in Ref. [29], but genetic algorithms [34] and Monte Carlo Markov Chain techniques [35] have also been successfully applied.

**The *k*-nearest neighbors** algorithm is inspired by the approach often taken by domain experts who make decisions based on previously seen similar cases [17]. For a given data instance, the *k*-nearest neighbors classifier searches for the *k* most similar training instances and classifies based on their prevailing class. The search for the most similar instances may be slow and requires the retrieval of a complete training set at the time of classification.

The methods listed above are often an integral part of most modern data mining suites and, alone or in combination with pre-processing, often perform well and sufficiently fast. The biggest differences when treating clinical data may arise in the predictive performance and interpretability of results. Throughout this review, we advocate that both of these are important and if methods perform similarly with respect to accuracy those which offer an explanation and interpretable models should be preferred.

### 2.3. *Standards*

Standards in predictive data mining are sparse but emerging. Those recently gaining attention and wide acceptance are CRISP-DM, SEMMA and PMML. The first two are standards that define the process of data mining. CRISP-DM was crafted by the Cross-Industry Standard Process for the Data Mining Interest Group (www.crisp-dm.org), which in the late 1990s defined a so-called CRISP-DM Process Model [36]. CRISP-DM breaks data mining into several phases: business and data understanding, data preparation, modeling, evaluation and deployment. It defines the inputs, outputs and general strategies to be applied in each phase. SEMMA (sample, explore, modify, model, assess) is a data mining methodology proposed by the SAS Institute and used within its powerful data mining suite. While CRISP-DM provides for a comprehensive project management template, SEMMA focuses mostly on the application to exploratory statistical and visualization-based data mining techniques.

Predictive Data Mining Markup Language (PMML, www.dmg.org) is very relevant to the communication, sharing and deployment of predictive models. PMML is an emerging vendor-independent open standard for defining data mining models. It defines an XML-based markup language for the encoding of many predictive data mining models that include decision trees and rules, Naïve Bayesian Classifiers and logistic regression models. The most recent version of popular data mining suites supports this standard by being able to export and import models encoded in complying XML files.

Related to predictive data mining are standards for data presentation and coding, like the Systemized Nomenclature of Human and Veterinary Medicine (SNOMED) and the Unified Medical Language System (UMLS). But as these standards are employed in clinical database management systems that data miners use to obtain their data, and as they help with the uniformity and consistency of the data sets [37], they have not (yet?) become an explicit part of any data mining system addressing the analysis of medical data. An exception to this but which lies beyond the scope of this review is medical text mining where, for instance, UMLS has been used to relate medical concepts and abstracts of papers cited in Medline [38,39].

### 2.4. *Predictive data mining and statistics*

In its brief history, data mining was at the start somewhat misleadingly associated solely with data analysis methods coming from fields other than statistics. The exposed characteristics of these methods were that they could address large quantities of data, make use of different data types (various types of attributes, text mining), were very flexible in modeling (e.g., inclusion of non-linearity) and could automate most of the analysis process. The initial success of the approaches in areas such as market basket analysis and text mining, along with the over-emphasis of machine-learning and pattern-recognition approaches in emerging data mining suites, provoked several statisticians to encourage their community to engage and contribute to the field [40]. Since then, the field has matured substantially and its coming of age is also reflected in the way today's data mining relates to statistics. Data mining suites are becoming part of large statistical packages, major books on data mining have been written by statisticians [17,41], while many recent developments in data mining have focused on bridging statistics, visualization and data analysis approaches from various fields of engineering.

A much-emphasized distinction between classical statistics and data mining involves the sheer size of data tackled by the latter [40]. Data mining deals with secondary analysis. There, the data is not purposely collected to test some research hypothesis but is obtained from legacy databases or data warehouses where the volume of data may be much greater. In this paper, though, we argue that for applications in clinical data analysis other aspects of data mining may be just or even more relevant. Most importantly, these include making the knowledge discovered from the data explicit and communicable to domain experts, the provision of an explanation when deploying and using this knowledge with new cases, and the ability to encode and use the domain knowledge in the data analysis process.

Data sets – including those drawn from clinical medicine – are often prone to different sources of noise, they may include various types of predictive features (e.g. nominal, real-valued, come from a time-series, etc.), may include a substantial number of missing feature values and may be governed by underlying non-linear processes. Modern data mining and statistical methods are often powerful enough to handle most of these cases, with the main difference being in the approach to

the discovery of predictive models. Explorative data analysis in statistics most often involves a manual search and the modeling of, for instance, non-linearities and attribute interactions, whereas when using data mining one would first rely on automatic techniques such as constructive induction [42], attribute interaction discovery [43] and approaches to non-linear modeling that systematically search through the data and attribute space. Existing knowledge in some problem domain would in statistics influence the composition of the data set to be collected, while – when appropriately encoded – it would help data mining to focus and report only on problem-relevant patterns found within secondary data analysis. For data mining algorithms, the data is only one source of information and others include any additional knowledge that can be obtained, encoded and made useful in the analysis.

### 2.5. *Predictive data mining and genomic medicine*

In recent years, predictive data mining has received a strong impulse from research in molecular biology. Data mining methods such as hierarchical clustering [44] or support vector machines [45] are routinely applied in the analysis of high-throughput data coming from DNA microarrays or mass-spectrometry. Quite interestingly, over the last few years several papers have highlighted the potential of predictive data mining to infer clinically relevant models from molecular data and to therefore provide decision support in the novel field of genomic medicine [46]. Nowadays, three different kinds of molecular data may be available to clinicians: (i) genotype data, often represented by a collection of single nucleotide polymorphisms (SNPs), DNA sequence variations that occur when a single nucleotide in the genome sequence is altered; since each individual has many SNPs, their sequence forms a unique DNA pattern for that person; (ii) gene expression data, which can be measured with DNA microarrays to obtain a snapshot of the activity of all genes in one tissue at a given time or with techniques that rely on a polymerase chain reaction (PCR) and real-time PCR when the expression of only a few genes needs to be measured with greater precision; (iii) protein expression data, which can include a complete set of protein profiles obtained with mass spectra technologies, or a few protein markers which can be measured with ad hoc essays.

The majority of papers published in the area of predictive data mining for genomic medicine deals with the goal of analyzing gene expression data coming from DNA microarrays, consisting of thousands of genes for each patient, with the aim to diagnose (sub)types of diseases and to obtain a prognosis which may lead to individualized therapeutic decisions. The published papers are mainly related to oncology, where there is a strong need for defining individualized therapeutic strategies [47]. A seminal paper from this area is that of Golub et al. [48] and focuses on the problem of the early differential diagnosis of acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). They were able to derive a classification model based on a weighted-voting approach relying on a list of about 50 genes. Today, there are many reports that show the potential usefulness of DNA microarray data for an outcome prediction in cancer treatment [49–51]. To improve classification accuracy and the clinical relevance of the prognostic models, some authors have proposed an integration of clinical and gene expression data. Nevins et al. [52] proposed a decision tree-based approach whereby the genes are first grouped into 'metagenes' and then used in a decision tree in conjunction with clinical data, such as lymphatic node status, to forecast a patient's survival. A different approach has been proposed by Futschik et al. [53], where clinical and microarray data are used to build two separate models for the outcome prediction of diffuse large B-cell lymphoma. The models employed are a Bayesian classifier and a Fuzzy Neural Network. The final prediction is obtained by means of an ensemble classifier whose parameters are also inferred from the training data. Clinical parameters and gene expressions have also been combined in Cox regression modeling in the risk stratification of Medulloblastomas [54].

In recent years some criticism has emerged against the gene-expression-based approaches to construct predictive models and derive lists of genes useful for outcome prediction [55]. Although several groups have published lists of predictive genes with very good predictive performance, it has been observed that these may vary widely from study to study. Such variability may be related to a lack of robustness due to the small number of clinical cases with respect to the number of attributes. A recent commentary by Berrar et al. [56] points out that many data mining papers in genomics and proteomics are affected by so-called selection bias since the feature selection is often (wrongly) performed on the entire data set prior to the cross-validation. This procedure adapts the classifier too much to the data set. The same problem, with reference to prominent early publications on the classification modeling of cancer gene expression data, was noted by Simon et al. [57]. Ein-Dor et al. [58] used a theoretical analysis to show that thousands of patients may be needed to obtain reliable gene lists. The integration of knowledge on the genes function and on the biomedical processes with clinical and gene expression data and the fusion of data coming from different studies [59] are promising directions for improving the robustness and practical impact of those studies.

Thanks to the possibility of measuring mass spectra from serum, proteomic profiles drawn from mass spectrometry techniques [45] have been analyzed to derive predictive models. In this case, the feature set is represented by a few hundreds or thousands of mass/charge ratios, in dependence on the resolution of the measurement technique. Predictive data mining approaches in this area have been applied to forecast patient outcomes in the case of prostate and ovarian cancer [45,60,61]. In those applications the pre-processing phase is crucial. For example, Yu et al. [45] developed a strategy based on a combination of feature filtering with the Kolmogorov–Smirnov test, wavelet analysis and support vector machines to define the predictive model. The high number of features combined with the need for pre-processing the raw spectra makes the problem of learning robust models very hard. As proteomic data is characterized by many features and much fewer cases, the risk of overfitting is even higher than with microarray data sets. Several proposals for systematic procedures to extract predictive models from mass spectrometry data have been recently proposed to avoid these problems [62,63].

Protein expression markers are also widely used for building prognostic models in cancer while recently there has been

great interest in applying statistical modeling and data mining to the analysis of tissue microarray data, which are a new high-throughput tool for the study of protein expression patterns in tissue [64,65]. We can expect that this area will give rise to several data mining applications in the next few years.

Another area where predictive data mining has been applied is the analysis of data on single nucleotide polymorphisms (SNPs). Genome-wide association studies most often include several hundred patients and controls and consider several hundred thousands of SNPs, with the goal of identifying those for which the risk of the disease is increased. The ambitious goal is to use SPN information to find the genetic basis of so-called complex traits, i.e. those traits that do not strictly follow Mendelian inheritance. The definition of a multi-variate prognostic model is a typical data mining task which has been studied in several papers. For example, Sebastiani et al. [32] use Bayesian networks to extract the relationships between SNPs and the risk of a stroke in patients suffering from Sickle cell anemia. They were able to extract a model with a small number of genes which were validated on a separate data set predicting the occurrence of a stroke in 114 individuals with 98.2% accuracy. Quite interestingly, in this paper the selection of genes and SNPs was performed by integrating prior knowledge in the data analysis process. The construction of models based on SNPs is not trivial since it has to face the same dimensionality problems in proteomics and genomics mentioned above. When it is important to model (or to discover) SNP interactions it is usually necessary to limit the analysis to several tens of SNPs due to the availability of data and complexity of the analysis [66]. To improve scalability, progress in the field will depend on the use of interaction analysis, constructive induction, and visualization [67]. Moreover, there is a need to integrate standard statistical analysis based on pedigrees and a linkage disequilibrium in order to reduce the number of SNPs.

With gene–gene interactions playing an important role in the susceptibility and progression of common diseases and response to treatment, and with the emerging case-control studies that collect genome-wide SNP data, the logical next step is a genome-wide, gene–gene interaction analysis. Yet, the data mining tools that could consider hundreds of thousands of SNPs and gene and protein expression profiles of thousands of patients do not exist yet. A major challenge to computer scientists is therefore to make these tools available and to design efficient heuristics to surpass the prohibitively complex exhaustive search for gene interactions. The challenge in designing such software is to provide an interactive, explorative analysis interface that provides users who are not computer scientists with seamless support in interaction discovery and the formation of new hypothesis to be then tested in a wet lab.

## 3. Contribution of data mining to predictive modeling in clinical medicine

Predictive models in clinical medicine are '... tools for helping decision making that combine two or more items of patient data to predict clinical outcomes' [68]. Such models may be used in several clinical contexts by clinicians and may allow a prompt reaction to unfavorable situations [69]. Data mining may effectively contribute to the development of clinically useful predictive models thanks to at least three inter-related aspects: (a) a comprehensive and purposive approach to data analysis that involves the application of methods and approaches drawn from different scientific areas; (b) the explanatory capability of such models; (c) the capability of using the domain (background) knowledge in the data analysis process.

### 3.1. *A systematic and integrated process*

As an engineering discipline, data mining relies on its associated process model which, being so important to the field, was recently regarded with much attention and for which several standards have been developed. The advantage but also the difficulty of data mining is that it is a framework that integrates various different approaches taken from diverse disciplines. Following the standard steps in studying a problem, data analysis and deployment can help researchers make systematic use of these various tools and appropriately choose from among the available techniques. Just like protocols in medicine, process standards in data mining help their users by guiding them through analysis process exposing those aspects that could otherwise be forgotten or neglected. Recently, a number of major data mining suites like that of the SPSS' Clementine (www.spss.com/spssbi/clementine) and the SAS' Enterprise Miner (www.sas.com/products/miner) have made the use of process standards explicit: there, the user chooses the phase he wants to address and is only shown a set of tools applicable to that phase.

### 3.2. *Explanation*

Data mining includes approaches that may play a double role: they may be used to derive a classification rule and to understand what information is contained in the available data. Inspired by early expert systems like Mycin [70] and Internist [71] that were quite rooted in medical applications, the explicit communication of knowledge discovered from the data and the subsequent explanation of decisions when this knowledge is used in the classification of new cases is what is emphasized by a number of data mining techniques. In the introductory example we have already demonstrated that classification trees can reveal interesting patterns in observational data. Examples where such analysis has led to the discovery of new medical knowledge include studies of brain injury [72], geriatrics [73] and trauma care [74].

Another formalism for representing classification models that allows for an easy explanation of the results are Bayesian networks. An interesting application of Bayesian network learning in the predictive data mining context has been recently published by Sierra and Larranaga [75]. In their work they compare the accuracy of a naïve Bayesian approach in forecasting the survival of malignant skin melanoma patients with that of the three different Bayesian networks induced from the data. Fig. 6A shows an example of an induced Bayesian network as reproduced from the original paper. We note that, for example, the variable sex is not considered to be useful to classify the cases, while the variables 'number of

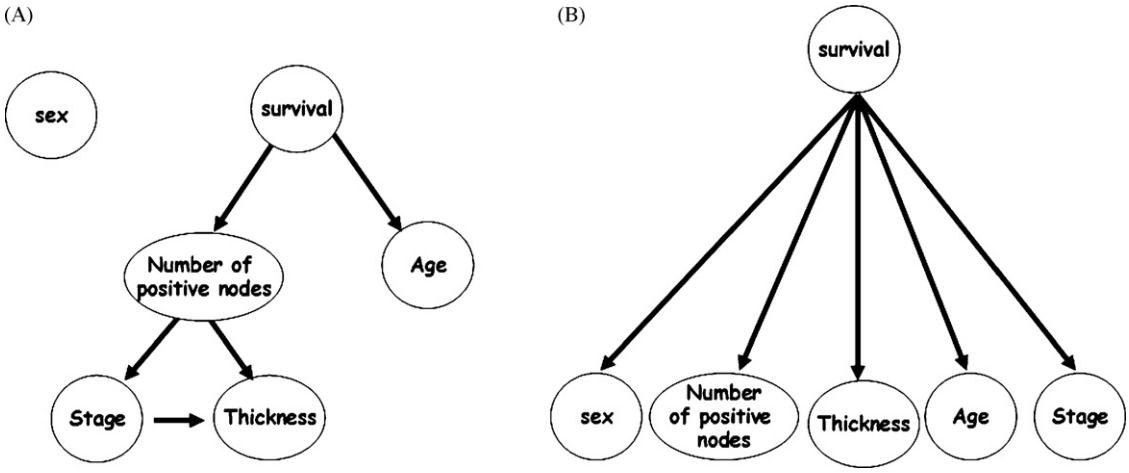



**Fig. 6 – The output of the survival prediction problem in a malignant skin tumor, presented by Sierra et al. [75]. Subfigure (A) shows the Bayesian network as induced from the data, while (B) shows the naive Bayesian model. Model (A) better describes the relationships between the variables and the outcomes.**

positive nodes', 'thickness' and 'stage' are found to be dependent on each other. Fig. 6B shows the naïve Bayes model of the same problem. Since the network structure remains fixed after the learning phase, the graphical output only reflects the a priori assumptions on the variable relationships, while the learned knowledge is hidden in the probability tables.

### 3.3. *The utility of domain knowledge*

Together with the capability to explain, some data mining algorithms can take so-called 'background knowledge' into account. Literally, background knowledge is the 'information that is essential to understanding a situation or a problem' [76]. In the process of building a predictive model, using background knowledge means being capable of taking into account information which is already known and should not be rediscovered from the data. This issue may be particularly important in the analysis of medical data [77].

Background knowledge can be expressed in different formats: examples may be found in the areas of decision rules [78], Bayesian Models [79], fuzzy sets [80] and concept hierarchies [3,81]. Among others, a method that may be particularly appropriate to deal and encode the background knowledge involves Bayesian networks. In Bayesian networks the background knowledge is exploited to define the network structure, i.e. number of variables, arcs and arc directions. Moreover, following the Bayesian paradigm prior probabilities on the conditional probability tables are given in order to take into account the background knowledge available on the relationships between the problem variables. Prior probabilities allow a model to be derived even when the information coming from the data is weak, and may help in avoiding overfitting where the derived model would reproduce the data too closely and fail to correctly classify the new and unseen cases [82]. An example of this approach is a study on a Bayesian network designed to assess the prophylaxis of graft versus host disease after bone marrow transplantation in children [79]. The network structure was assessed on the basis of the available background knowledge, while the probabilities were first defined by experts and then updated on the basis of a data set of fifty patients [79]. The use of background knowledge in building Bayesian networks and in eliciting their probabilities is an active area of research [82–84] where data mining and knowledge engineering frequently combine their efforts and results.

Background knowledge can also be easily exploited in the construction of classification rules: for instance, an incomplete set of classification rules provided by the expert can be refined and augmented on the basis of the available data, while the rule search can be driven by a certain number of monotonicity constraints [85,86].

## 4. Predictive data mining process: tasks and guidelines

Data mining is most often the application of a number of different techniques from various disciplines with the goal to discover interesting patterns from data. Given the large variety of techniques available and interdisciplinary fields, it is no surprise that data mining is often viewed as a craft that is hard to learn and even harder to master.

As we mentioned, several process models and standards have been proposed to introduce engineering principles, systemize the process and define typical data mining tasks. In the section on standards we introduced CRISP-DM, a data mining process standard that seems to be gaining the widest acceptance. While CRISP-DM enumerates a number of methods that may be used to accomplish data mining tasks, it is not meant to give precise guidelines on which techniques, evaluation schemes and statistics to use. Namely, these should all be specific to a problem domain, particular data mining tasks and the type of data under consideration. Predictive data mining in clinical medicine is an example of such a specific task and guidelines that in particular address different aspects of medical data analysis could be provided to accompany the CRISP-DM model and make it more useful in this domain. In

the following description of the predictive data mining process, we generally adhere to the CRISP-DM schema but we also try to be specific and list a number of problems, recommendations and guidelines that may apply to medical predictive mining and which have been proposed and evaluated by active researchers and developers in the field.

### 4.1. *Defining the problem, setting the goals*

Predictive data mining is concerned with analyzing data sets that are composed of data *instances* (e.g., cases or list of observations), where each instance is characterized by a number of *attributes* (also referred to as predictors, features, factors, or explanatory variables). There is a special additional attribute called an *outcome variable*, also referred to as a class, dependent or response variable. In general, the task of predictive data mining is to find the best fitting model that relates attributes to the outcome. Unlike standard data mining data sets, medical data sets may be smaller: typically, the number of instances are from several tens to several thousands. The number of attributes may widely range from several tens (classical problems from clinical medicine) to thousands (proteomics, genomics).

The goal of predictive data mining in clinical medicine is to construct a predictive model that is sound, makes reliable predictions and helps physicians improve their prognosis, diagnosis or treatment planning procedures. In terms of data analysis, there is a number of important questions that data mining may answer, including:

(a) Are the data and corresponding predictive features sufficient to construct a predictive model of acceptable performance?
(b) Which of the attributes are the most predictive? Which are those that need to be included in the predictive model?
(c) What is the relationship between the attribute and the outcome?
(d) Can we find any interesting combination (or relationship) between the attributes? Can any intermediate factors be derived from original attributes that may boost the performance of the predictive model and indicate an interesting phenomenon?

To find the answer to (a), it is very useful, if not required, that the measures of success are defined at this stage of data mining. This may include the decision upon which statistics to use for evaluating the predictive models and what the acceptable ranges are for these. Defining the criteria of acceptability of the resulting model prior to the actual data mining may help in producing less biased and a more objective evaluation of data mining results.

Data mining is rich in methods that may help find the answers to the other three questions. Techniques such as feature ranking [87], feature subset selection [88] and constructive induction [42] may help find the most relevant features and construct new ones from a combination of features from the original set (questions b and d). As we previously discussed, many data mining methods such as classification trees [11,12] and rules [13,14] focus on the construction of interpretable predictive models expressed in a textual form that can be communicated to and scrutinized by domain experts (questions c and d).

Data mining provides a large toolbox of techniques and so as to narrow the choice of which ones to use for a particular problem answering the following questions at the stage prior to actual data mining may help:

(1) Should the resulting model be 'transparent', i.e., defined through some language (like a set of rules) that the user may interpret?
(2) When used in decision-making, should the predictive model support the explanation?
(3) Should predictive models report the probabilities of outcomes? Should confidence intervals be reported?
Knowing how the model derives its prediction and being able to use the model's logic to explain how the conclusion was reached may significantly increase a physician's confidence in the model and help increase its acceptance. The current practice, though, may be different: one of the most often used artificial intelligence-based data mining techniques in building predictive models from clinical data involves artificial neural networks, from which it is far from trivial to understand the mechanisms that govern computation of the outcome and may in this respect be considered a 'black box'. Often, such models are reliable in terms of prediction but it has also been shown that some much simpler techniques such as, for instance, the naïve Bayes classifier, perform equally well [22] and may additionally accommodate for explanation [5] and model transparency [89]. Somehow similar in terms of its simplicity of the model, predictive power and ability to explain, statisticians often recommend that data mining techniques should be compared to logistic regression [18]. When performing clinical data mining, it may often be worthwhile to try the simple techniques first.
Question 3 above is relatively rhetorical: yes, to be useful in clinical practice, predictive models should model probabilities and, wherever possible, should report on confidence intervals. During the 1990s, this would have been quite an exception since most predictive data mining methods provided only crisp classifications, that is, they only reported which of the outcomes was the most probable one without quantifying this probability [90]. Only recently and through a relatively straightforward extension of existing algorithms most data mining methods do in fact allow the reporting of probabilities of outcomes. Rarely, however, do the data mining suites include implementations that are able to report the confidence intervals of predicted probabilities. Finally, another question arises that may highly influence the selection of data mining techniques:
(4) Is there additional knowledge that domain experts can explicitly make available for the modeling methods? If so, how can this knowledge be encoded?
Unfortunately, although several examples of the use of background knowledge in clinical data mining are available, as described in the previous section, no data mining standards are (yet) available on how to encode such knowledge. Techniques to allow the use of background knowledge are often crafted by specialized research groups and rarely, if at all, find their way into commer-

cially available data mining suites. However, while the inclusion of background knowledge is far from trivial it may have most significant impact on both performance and comprehensibility [85]. A substantial (but worthwhile!) effort is needed by the data mining community to standardize this area and make the existing academic tools available to the wider community of users.

### 4.2. *Data preparation*

For data mining, clinical data most often come from dedicated databases (or even paper-based forms) that were purposely collected to study a particular clinical problem. Although not yet widely available, another important source of clinical data is data warehouses. Currently, the most widely used data mining algorithms require data to be placed in a tabular form that includes predictive factors and outcomes and is constructed by querying single or several dedicated databases [91].

An important rule in the construction and evaluation of predictive models is that they should never be built and tested on the same data set. For this, techniques like *cross-validation* are used (see the next section) but it may also be a good idea at this stage to split the data into two sets: the first one, often referred to as the *learning set*, is used to compare different data mining algorithms, estimate their performance using some statistical metrics, find the best set of parameters for feature ranking, selection and learning methods and, finally, to select the modeling technique that performs best. Using this technique, a final model is to be developed from a complete *learning set* and tested on a second data set, commonly referred to as a *validation set*. The data split may be arbitrary or based on time or source label of the data instances.

Separate learning and validation sets are necessary to objectively assess the predictive performance. Data mining models may be complex and in extreme cases may 'remember' each data instance that they learned from. Such models perform perfectly on data that was used for learning, but poorly with any new case that does not match some data instance from the learning data. It is said that such models generalize poorly due to overfitting. Most contemporary data mining techniques include efficient mechanisms to avoid overfitting, like pruning for decision trees, limiting the complexity for the neural network, and the selection of only the most significant rules for decision rule modeling, but it is only the evaluation of an independent data set that can guarantee that the good performance did not result from overfitting.

### 4.3. *Modeling and evaluation*

Once the data is split into a learning and validation set, it is now time to employ our modeling techniques and trim their parameters to the learning data set. The goal of this phase is to determine which data mining algorithm performs best so we can use it to generate our target predictive model.

Predictive models can be evaluated on the basis of their *predictive performance* and *comprehensibility*. Of the two, predictive performance is easier to quantify and typical statistics include metrics such as sensitivity, specificity, classification accuracy [92], area under the ROC curve [93] and Brier score [94]. Comprehensibility is a subjective measure that is assessed by participating domain experts. While it may be prohibitively hard to quantify the comprehensibility, preferable models may be found by answering questions like: 'given the two models, which one is easier to understand? which one explains the decisions better? which one do the experts have greater confidence using?' If comprehensibility and explanation are at stake, the data mining algorithms can be ranked first using the chosen predictive performance statistics and then, of the few top ranking models, domain experts may select the final model based on its comprehensibility and ability to explain.

As mentioned in the previous section, to estimate those statistics that evaluate the predictive performance a desirable approach is to apply the so-called hold-out strategy: a subset of the learning set, the training set, is used to construct the model while another subset, the test set, is used to estimate the accuracy of the model. However, the holdout procedure makes quite inefficient use of the data: the typical strategy is to learn from two-thirds of the data and then to test on the remaining one-third of the sample. Such a strategy may not be applicable with a small number of data since the algorithms for learning the prognostic model may have problems due to the reduced data set for learning, while the test may be still insufficient to reach the desired confidence interval limits. A popular contemporary method to be use in solving the abovementioned problems is $k$-fold cross-validation. With cross-validation, the data are split into a number ($k$) of data subsets which contain approximately an equal number of data instances and approximately match the outcome distribution of the learning set (stratified cross-validation). Typically, the learning data set is split into ten data subsets (10-fold cross validation). Then, data from the nine subsets are used for modeling while the remaining subset is used to test the resulting model and assess statistics. The process of training and testing is repeated 10 times, each time using a different testing subset. Averaged statistics are then reported and characterize the modeling method. Besides cross-validation, other data splitting approaches may be used such as 'leave-one-out' cross-validation, random sampling, bootstrap, etc. [95,96].

Special attention should be paid to parameter estimation. Most data mining methods depend on a set of parameters that define the behavior of the learning algorithm and directly or indirectly influence the complexity of the resulting models. For instance, the degree of pruning may be set for decision tree induction, the number of units in a hidden layer may be set for feed-forward neural network models and the required level of statistical significance may be defined for decision rules. While the finding of the best set of parameters can be characterized as a search in parameter space that employs some state-of-the-art optimization technique, practitioners often define a set of most likely values of parameters and, again through cross validation, evaluate each set separately to find the winner. The evaluation of data mining methods then yields not only the ranking of data mining techniques but also identifies the appropriate parameter set to be used with. Note also that feature ranking, subset selection and construction may have their own parameters, which also require optimization.

### 4.4. *Construction of the target predictive model*

The evaluation techniques described above provide the grounds for ranking data mining methods and identifies a suitable set of parameters. We can now use a complete evaluation data set and the best-ranked method to construct our predictive model. The resulting model is then evaluated on the validation data set and, if its predictive performance is acceptable, then this is now our target predictive model. Note that when reporting on the predictive qualities of the model, it is *only* the statistics obtained using the validation data set that have merit and should matter; reporting on results from learning data sets may be deceiving as they are prone to overfitting. If the task of data mining is to observe the relationships between the features and the features and the outcome, it is now time to scrutinize, analyze and visualize the resulting model.

### 4.5. *Deployment and dissemination*

Once the predictive model has been constructed and evaluated, this is where most clinical data mining projects stop. This is quite unfortunate because clinical data mining should also be concerned with the deployment of resulting models and discovered relationships and with studying their potential impact on clinical environments. For instance, would finding an interesting relationship change the current medical practice? Can the constructed predictive model be used for day-to-day decision support? What would the value be of such a system? Would, once the model is deployed, the quality of health care increase or would some of the related costs decrease? Reports on the utility of models constructed by predictive data mining are at best rare and so is the deployment of predictive models in clinical environments.

Technically, one of the issues that may the prevent smooth dissemination of constructed predictive models within the clinical environment is related to bridging predictive data mining and decision support [97,98]. Clinical decision support tools [99] often include predictive models and have dedicated interfaces that ease the utility of the particular application for physicians or medical personnel. Data mining tools are often quite complex, may be very expensive and are intended for specialists. Data mining tools are optimized for model development and usually do not provide specific interfaces for when deploying the model. While model development is, as we have discussed in this paper, nontrivial and often a complex task, computing an outcome using the predictive model is usually straightforward and does not need much computational power. We should therefore not expect the data mining suites to be appropriate environments for decision support. A technical difficulty of bridging the two technologies is that data mining tools usually do not offer encoding and the saving of the developed models in a form that is compatible with some decision support systems.

There have been some recent advances that may help us tackle this problem and ease the bridging of data mining and decision support. In our review, we have already mentioned the PMML standard for encoding the prediction models

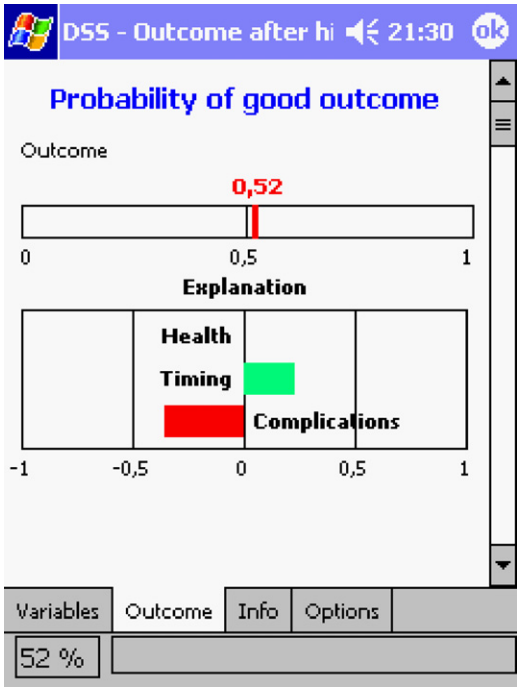



**Fig. 7 – Snapshot of decisions-at-Hand software on a PocketPC that shows the nomogram reporting on the outcome. The prediction was made on the same case as shown in Fig. 1B.**

in XML-based documents. If this or similar standards take ground, we may expect them to be supported by major data mining tools and, perhaps even more importantly, by decision support shells that would deploy a specific predictive model and provide an appropriate interface. As a demonstration of such technology, the decisions-at-hand schema by Zupan et al. [100] allows for the saving of data mining models in the XML format and provides either web- or handheld-based decision support shells. For example, for the naïve Bayesian classifier from Fig. 1B a PocketPC interface that supports the data entry and reports the outcome is shown in Fig. 7. The guiding idea of these and similar approaches is, on one side, to bridge data mining and decision support and, on the other side, to decouple the two technologies allowing users, in our case physicians and medical personnel, to use lightweight, inexpensive, computationally non-demanding and easy to use decision shells. Besides the utility per se, being able to save the predictive data mining model may also provide significant advantages in the communication of results in evaluation and testing phases of data mining projects [101].

Predictive data mining models can also be used as an instrument for assessing and comparing evidence-based medicine results with the outcomes collected in clinical practice. The availability of data collected in clinical institutions on their specific processes of disease management allows for the capability of integrating evidence-based strategies and hence of making prognostic reasoning, with local knowledge coming from the data collected in the clinical routine [102]. The merging of evidence-based medicine and local experience may be seen as a particular kind of data mining problem where the

background knowledge can be obtained from clinical studies. In this case, background knowledge is far more than a set of constraints; rather, it provides for a reference that can be complemented during the learning process [103]. The goal of data analysis is therefore related to the better comprehension of the information that is contained in the data, thus highlighting cases that do not confirm well-established knowledge or problems in the data collection procedures. Moreover, its aim is to uncover the peculiarities of the specific clinical center in order to better tailor their prognostic guidelines. In this respect, data mining can be seen as part of a medical institution's information infrastructure, where it is used to 'procure' knowledge that is maintained and operationalized through a knowledge management system [68].

## 5. Discussion

Compared to data mining in business, marketing and the economy, medical data mining applications have several distinguishing features [104]. The most important one is that medicine is a safety critical context [105] in which decision-making activities should always be supported by explanations. This means that the value of each datum may be higher than in other contexts: experiments can be costly due to the involvement of the personnel and use of expensive instrumentation and due to the potential discomfort of the patients involved. In clinical mining, the data sets can be small and report non-reproducible situations. The data may be further affected by several sources of uncertainty, like those from measurement errors or missing data or from errors in coding the information buried in textual reports. Physicians and researchers deal with such difficulties by exploiting their knowledge of the domain. Similarly, data mining can cope with these problems by carefully applying variable and model selection, by correctly evaluating the resulting models and by explicitly encoding this knowledge and using it in data analysis [106].

At present, data mining is a very diverse field with a number of techniques that may serve the same purpose and behave equally well. It may not be practical to explore all alternative methods when mining a particular data set, while the choice of which techniques to use is often guided by the instincts of expert data miners. While it is unlikely that with the current variety of approaches the community can come up with cookbooks and recipes, we have tried to provide some general task descriptions and a simple set of guidelines that may apply to the construction of clinical predictive models using data mining techniques. Overall, the ideas we have presented may be summarized in the following list:

- Define the success criteria in advance. Set acceptable ranges of evaluation statistics prior to modeling.
- If possible and for reference compare the performance results with those obtained from classical statistical modeling.
- Model probabilities, not crisp class membership. Prefer methods that report confidence intervals.
- Avoid overfitting. Never test models on data that was used in their construction. In the case of small data sets, use cross-validation or similar techniques to obtain evaluation statistics.
- If possible, test the resulting model on an independent separate data set.
- Report performance scores with confidence intervals.
- Prefer modeling techniques that expose relations and can present them in a readable form. If the discovery of relationships is a goal of data mining, avoid black-box models.
- If still of acceptable performance prefer simple modeling techniques, possibly those that derive models that can be reviewed and criticized by experts.
- Feature ranking, feature selection, constructive induction and so on, together with any parameter estimation, are all part of the modeling and should be tested within cross-validation. Using them in pre-processing that takes place prior to cross-validation leads to overfitting.
- The project is not finished when a good model is found. Think how to include your model within some clinical information or decision support system. If possible, perform a cost/benefit study.
- Explicitly assess the model's applicability and its potential for generalization. Here, in particular consider the type of data collection (retrospective, prospective, derived from a clinical trial or from clinical routine), the number of available data and performance of the model.

These guidelines relate to newly emerging issues in personalized and genomic medicine. Today, the construction of reliable predictive models may require the integration of data drawn from heterogeneous sources that include clinical, laboratory, genetic, genomic and proteomic data. The full availability of data repositories and warehouses able to concurrently provide such information about a single patient, and the methods to integrate it within a decision support system are issues which remain to be resolved.

## 6. Conclusion

At present, many ripe predictive data mining methods have been successfully applied to a variety of practical problems in clinical medicine. As suggested by Hand [40], data mining is particularly successful where data are in abundance. For clinical medicine, this includes the analysis of clinical data warehouses, epidemiological studies and emerging studies in genomics and proteomics. Crucial to such data are those data mining approaches which allow the use of the background knowledge, discover interesting interpretable and non-trivial relationships, construct rule-based and other symbolic-type models that can be reviewed and scrutinized by experts, discover models that offer an explanation when used for prediction and, finally, bridge model discovery and decision support to deploy predictive models in daily clinical practice. With the promises offered by genomic medicine and upcoming needs to integrate molecular and clinical data, data mining and other knowledge-intensive computational approaches are becoming required for advancing the state-of-the-art of both research and real-life applications [107,108]. Last but not least, clinical data mining deals with 'bed-side'

problems, that is, with models that forecast the patient's outcome. Decision-making that uses a particular prediction model should therefore also take into account the issues of ethics and the cost of prediction while being concerned with the analysis of outcomes.

## Acknowledgements

The authors would like to acknowledge the help given by the International Medical Informatics Association and its Working Group on Intelligent Data Analysis and Data Mining, which they are chairing. The work was supported by a Slovenian-Italian Bilateral Collaboration Project. RB is also supported by the Italian Ministry of University and Scientific Research through the PRIN Project 'Dynamic modeling of gene and protein expression profiles: clustering techniques and regulatory networks', and BZ by the Slovenian Research Agency's Program Grant.

## REFERENCES

[1] P. Giudici, Applied Data Mining Statistical Methods for Business and Industry, Wiley & Sons, 2003.
[2] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Data mining and knowledge discovery in databases, Commun. ACM 39 (1996) 24–26.
[3] B. Zupan, J. Demsar, D. Smrke, K. Bozikov, V. Stankovski, I. Bratko, J.R. Beck, Predicting patient's long-term clinical status after hip arthroplasty using hierarchical decision modelling and data mining, Meth. Inf. Med. 40 (2001) 25–31.
[4] J. Demsar, B. Zupan, G. Leban, T. Curk, Orange: from experimental machine learning to interactive data mining, in: European Conference of Machine Learning, Springer Verlag, Pisa, Italy, 2004, 537-539.
[5] I. Kononenko, Inductive and Bayesian learning in medical diagnosis, Appl. Artif. Intelligen. 7 (1993) 317–337.
[6] J. Lubsen, J. Pool, E. van der Does, A practical device for the application of a diagnostic or prognostic function, Meth. Inf. Med. 17 (1978) 127–129.
[7] M. Mozina, J. Demsar, M.W. Kattan, B. Zupan, Nomograms for visualization of naive bayesian classifier, in: Proceedings of the Principles Practice of Knowledge Discovery in Databases (PKDD-04), Pisa, Italy, 2004, pp. 337–348.
[8] F.E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer, New York, 2001.
[9] M.W. Kattan, J.A. Eastham, A.M. Stapleton, T.M. Wheeler, P.T. Scardino, A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer, J. Natl. Cancer Inst. 90 (1998) 766–771.
[10] M. Graefen, P.I. Karakiewicz, I. Cagiannos, D.I. Quinn, S.M. Henshall, J.J. Grygiel, R.L. Sutherland, P.D. Stricker, et al., International validation of a preoperative nomogram for prostate cancer recurrence after radical prostatectomy, J. Clin. Oncol. 20 (2002) 3206–3212.
[11] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, Calif, 1993.
[12] L. Breiman, Classification and Regression Trees, Chapman & Hall, New York, London, 1993.
[13] P. Clark, T. Niblett, The CN2 Induction Algorithm, Mach. Learn. 3 (1989) 261–283.
[14] R.S. Michalski, K. Kaufman, Learning patterns in noisy data: the AQ approach, in: G. Paliouras, V. Karkaletsis, C. Spyropoulos (Eds.), Machine Learning and its Applications, Springer-Verlag, Berlin, 2001, pp. 22–38.
[15] N. Lavrac, I. Kononenko, E. Keravnou, M. Kukar, B. Zupan, Intelligent data analysis for medical diagnosis: using machine learning and temporal abstraction, AI Commun. 11 (1998) 191–218.
[16] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd ed., Wiley, New York, 2000.
[17] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2001.
[18] G. Schwarzer, W. Vach, M. Schumacher, On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology, Stat. Med. 19 (2000) 541–561.
[19] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, UK, New York, 2000.
[20] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[21] C. Cortes, V. Vapnik, Support-vectors networks, Mach. Learn. 20 (1995) 273–297.
[22] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artif. Intell. Med. 23 (2001) 89–109.
[23] S. Andreassen, F.V. Jensen, K.G. Olesen, Medical expert systems based on causal probabilistic networks, Int. J. Biomed. Comput. 28 (1991) 1–30.
[24] P.W. Hamilton, R. Montironi, W. Abmayr, M. Bibbo, N. Anderson, D. Thompson, P.H. Bartels, Clinical applications of Bayesian belief networks in pathology, Pathologica 87 (1995) 237–245.
[25] S.F. Galan, F. Aguado, F.J. Diez, J. Mira, NasoNet, modeling the spread of nasopharyngeal cancer with networks of probabilistic events in discrete time, Artif. Intell. Med. 25 (2002) 247–264.
[26] D. Luciani, M. Marchesi, G. Bertolini, The role of Bayesian Networks in the diagnosis of pulmonary embolism, J. Thromb. Haemost. 1 (2003) 698–707.
[27] D.J. Spiegelhalter, S.L. Lauritzen, Sequential updating of conditional probabilities on directed graphical structures, Networks 20 (1990) 579–605.
[28] W.L. Buntine, A guide to the literature on learning probabilistic networks from data, IEEE Trans. Know. Data Eng. 8 (1996) 195–210.
[29] G.F. Cooper, E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, Mach. Learn. 9 (1992) 309–347.
[30] M. Ramoni, P. Sebastiani, Robust learning with missing data, Mach. Learn. 45 (2001) 147–170.
[31] E.H. Herskovits, J.P. Gerring, Application of a data-mining method based on Bayesian networks to lesion-deficit analysis, Neuroimage 19 (2003) 1664–1673.
[32] P. Sebastiani, M.F. Ramoni, V. Nolan, C.T. Baldwin, M.H. Steinberg, Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia, Nat. Genet. 37 (2005) 435–440.
[33] D. Geiger, D. Hackerman, Learning Gaussian networks, in: R.L. de Mantaras, D. Poole (Eds.), Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA/Seattle, WA, 1994, pp. 235–243.
[34] P. Larrañaga, B. Sierra, M.Y. Gallego, M.J. Michelena, P J M. Learning Bayesian networks by genetic algorithms: a case study in the prediction of survival in malignant skin melanoma, in: E. Keravnou, C. Garbay, R. Baud, C.J. Wyatt (Eds.), Artificial Intelligence in Medicine Europe, Grenoble, France, 1997, pp. 261–272.

[35] P. Le Phillip, A. Bahl, L.H. Ungar, Using prior knowledge to improve genetic network reconstruction from microarray data, In. Silico. Biol. 4 (2004) 335–353.
[36] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, CRISP-DM 1. 0: Step-by-Step Data Mining Guide: The CRISP-DM Consortium, 2000.
[37] G.W. Moore, J.J. Berman, Anatomic pathology data mining, in: K.J. Cios (Ed.), Medical Data Mining and Knowledge Discovery, Springer-Verlag, Berlin/Heidelberg, 2001, pp. 61–108.
[38] D. Hristovski, J. Stare, B. Peterlin, S. Dzeroski, Supporting discovery in medicine by association rule mining in Medline and UMLS, Medinfo 10 (2001) 1344–1348.
[39] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, Proc. AMIA Symp. (2001) 17–21.
[40] D.J. Hand, Data mining: statistics and more? Am. Statist. 52 (1998) 112–118.
[41] D.J. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, Cambridge, Mass, 2001.
[42] E. Bloedorn, R.S. Michalski, Data-driven constructive induction, IEEE Intell. Syst. 13 (1998) 30–37.
[43] A. Jakulin, I. Bratko, D. Smrke, J. Demsar, B. Zupan, Attribute interactions in medical data analysis, in: M. Dojad, E. Keravnou, P. Barahona (Eds.), Proceedings of the Ninth Conference on Artificial Intelligence in Medicine in Europe (AIME 2003), Protaras, Cyprus: Springer, 2003, pp. 229–238.
[44] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 14863–14868.
[45] J.S. Yu, S. Ongarello, R. Fiedler, X.W. Chen, G. Toffolo, C. Cobelli, Z. Trajanoski, Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data, Bioinformatics 21 (2005) 2200–2209.
[46] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, P. Tarczy-Hornoch, Data integration and genomic medicine, J. Biomed. Inform. 40 (2007) 5–16.
[47] P.S. Mischel, T. Cloughesy, Using molecular information to guide brain tumor therapy, Nat. Clin. Pract. Neurol. 2 (2006) 232–233.
[48] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[49] L.J. van't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002) 530–536.
[50] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y. Kim, L.C. Goumnerova, et al., Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (2002) 436–442.
[51] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C. Aguiar, M. Gaasenbeek, M. Angelo, et al., Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nat. Med. 8 (2002) 68–74.
[52] J.R. Nevins, E.S. Huang, H. Dressman, J. Pittman, A.T. Huang, M. West, Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction, Hum. Mol. Genet. (2003) R153–R157, 12 Spec No 2:.
[53] M.E. Futschik, M. Sullivan, A. Reeve, N. Kasabov, Prediction of clinical behaviour and treatment for cancers, Appl. Bioinform. 2 (2003) S53–S58.
[54] A. Fernandez-Teijeiro, R.A. Betensky, L.M. Sturla, J.Y. Kim, P. Tamayo, S.L. Pomeroy, Combining gene expression profiles and clinical parameters for risk stratification in medulloblastomas, J. Clin. Oncol. 22 (2004) 994–998.
[55] J.D. Brenton, L.A. Carey, A.A. Ahmed, C. Caldas, Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J. Clin. Oncol. 23 (2005) 7350–7360.
[56] D. Berrar, I. Bradbury, W. Dubitzky, Avoiding model selection bias in small-sample genomic datasets, Bioinformatics 22 (2006) 1245–1250.
[57] R. Simon, M.D. Radmacher, K. Dobbin, L.M. McShane, Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification, J. Natl. Cancer Inst. 95 (2003) 14–18.
[58] L. Ein-Dor, O. Zuk, E. Domany, Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 5923–5928.
[59] Z. Hu, C. Fan, D.S. Oh, J.S. Marron, X. He, B.F. Qaqish, C. Livasy, L.A. Carey, et al., The molecular portraits of breast tumors are conserved across microarray platforms, BMC Genom. 7 (2006) 96.
[60] B.L. Adam, Y. Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares, O.J. Semmes, P.F. Schellhammer, et al., Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, Cancer Res. 62 (2002) 3609–3614.
[61] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, et al., Use of proteomic patterns in serum to identify ovarian cancer, Lancet 359 (2002) 572–577.
[62] N. Barbarini, P. Magni, R. Bellazzi, A new approach for the analysis of mass spectrometry data for biomarker discovery, AMIA Annu Symp. Proc. (2006) 26–30.
[63] R.L. Somorjai, B. Dolenko, R. Baumgartner, Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions, Bioinformatics 19 (2003) 1484–1491.
[64] X. Liu, V. Minin, Y. Huang, D.B. Seligson, S. Horvath, Statistical methods for analyzing tissue microarray data, J. Biopharm. Stat. 14 (2004) 671–685.
[65] T.A. Bismar, F. Demichelis, A. Riva, R. Kim, S. Varambally, L. He, J. Kutok, J.C. Aster, et al., Defining aggressive prostate cancer using a 12-gene model, Neoplasia 8 (2006) 59–68.
[66] B.A. McKinney, D.M. Reif, M.D. Ritchie, J.H. Moore, Machine learning for detecting gene-gene interactions: a review, Appl. Bioinform. 5 (2006) 77–88.
[67] J.H. Moore, J.C. Gilbert, C.T. Tsai, F.T. Chiang, T. Holden, N. Barney, B.C. White, A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility, J. Theor. Biol. 241 (2006) 252–261.
[68] C.J. Wyatt, D.G. Altman, Prognostic models: clinically useful or quickly forgotten? BMJ (1995) 311.
[69] M.W. Kattan, M.J. Zelefsky, P.A. Kupelian, P.T. Scardino, Z. Fuks, S.A. Leibel, Pretreatment nomogram for predicting the outcome of three-dimensional conformal radiotherapy in prostate cancer, J. Clin. Oncol. 18 (2000) 3352–3359.
[70] E.H. Shortliffe, R. Davis, S.G. Axline, B.G. Buchanan, C.C. Green, S.N. Cohen, Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system, Comput. Biomed. Res. 8 (1975) 303–320.
[71] R.A. Miller, H.E. Pople Jr., J.D. Myers, Internist-1, an experimental computer-based diagnostic consultant for general internal medicine, N. Engl. J. Med. 307 (1982) 468–476.

[72] P.J. Andrews, D.H. Sleeman, P.F. Statham, A. McQuatt, V. Corruble, P.A. Jones, T.P. Howells, C.S. Macmillan, Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression, J. Neurosurg. 97 (2002) 326–336.
[73] V.S. Stel, S.M. Pluijm, D.J. Deeg, J.H. Smit, L.M. Bouter, P. Lips, A classification tree for predicting recurrent falling in community-dwelling older persons, J. Am. Geriatr. Soc. 51 (2003) 1356–1364.
[74] E.A. Eastwood, J. Magaziner, J. Wang, S.B. Silberzweig, E.L. Hannan, E. Strauss, A.L. Siu, Patients with hip fracture: subgroups and their outcomes, J. Am. Geriatr. Soc. 50 (2002) 1240–1249.
[75] B. Sierra, P. Larranaga, Predicting survival in malignant skin melanoma using Bayesian networks automatically induced by genetic algorithms An empirical comparison between different approaches, Artif. Intell. Med. 14 (1998) 215–230.
[76] C. Fellbaum, C. Fellbaums, C. Fellbaum, WordNet An Electronic Lexical Database, MIT Press, 1998.
[77] B. Zupan, J.H. Holmes, R. Bellazzi, Knowledge-based data analysis and interpretation, Artif. Intell. Med. 37 (2006) 163–165.
[78] N. Lavrac, S. Dzeroski, V. Pirnat, V. Krizman, The utility of background knowledge in learning medical diagnostic rules, Appl. Artif. Intelligen. 7 (1993) 273–293.
[79] S. Quaglini, R. Bellazzi, F. Locatelli, M. Stefanelli, C. Salvaneschi, An influence diagram for assessing GVHD prophylaxis after bone marrow transplantation in children, Med. Decis. Mak. 14 (1994) 223–235.
[80] R. Silipo, R. Vergassola, W. Zong, M.R. Berthold, Knowledge-based and data-driven models in arrhythmia fuzzy classification, Meth. Inf. Med. 40 (2001) 397–402.
[81] S. Mani, W.R. Shankle, M.B. Dick, M.J. Pazzani, Two-stage machine learning model for guideline development, Artif. Intell. Med. 16 (1999) 51–71.
[82] P. Lucas, Expert knowledge and its role in learning Bayesian Networks in medicine: an appraisal, in: S. Quaglini, P. Barahona, S. Andreassen (Eds.), Artificial Intelligence in Medicine, Springer, Berlin, 2001, pp. 156–166.
[83] M.J. Druzdzel, L.C. van der Gaag, Building probabilistic networks: "Where do the numbers come from?", IEEE Transn. Knowl. Data Eng. 12 (2000) 481–486.
[84] V.M.H. Coupe, L.C. Van der Gaag, J.D.F. Habbema, Sensitivity analysis: an aid for belief-network quantification, Knowl. Eng. Rev. 15 (2000) 215–232.
[85] M. Pazzani, D. Kibler, The utility of background knowledge in inductive learning, Mach. Learn. 9 (1992) 57–94.
[86] M.J. Pazzani, S. Mani, W.R. Shankle, Acceptance of rules generated by machine learning among medical experts, Meth. Inf. Med. 40 (2001) 380–385.
[87] I. Kononenko, Estimating attributes: analysis and extensions of RELIEF, in: European Conference on Machine Learning (ECML), 1994, pp. 171–182.
[88] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
[89] B. Zupan, J. Demsar, M.W. Kattan, J.R. Beck, I. Bratko, Machine learning for survival analysis: a case study on recurrence of prostate cancer, Artif. Intell. Med. 20 (2000) 59–75.
[90] R. Bellazzi, B. Zupan, Intelligent data analysis in medicine and pharmacology: a position statement, in: Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP), Brighton, UK, 1998, pp. 2–5.
[91] D. Pyle, Data preparation for data mining, Morgan Kaufmann Publishers, San Francisco, CA, 1999.
[92] N. Lavrac, P. Flach, B. Zupan, Rule evaluation measures: a unifying view, in: Workshop on Inductive Logic Programming, 1999, pp. 174–185.
[93] J.R. Beck, E.K. Shultz, The use of relative operating characteristic (ROC) curves in test performance evaluation, Arch. Pathol. Lab. Med. 110 (1986) 13–20.
[94] G.W. Brier, Verification of forecasts expressed in terms of probability, Month. Weather Rev. 78 (1950) 1–3.
[95] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations, Morgan Kaufmann, San Francisco, CA, 1999.
[96] D.J. Hand, Construction and assessment of classification rules, Wiley, Chichester; New York, 1997.
[97] M. Bohanec, B. Zupan, Integrating Decision Support and Data Mining by Hierarchical Multi-Attribute Decision Models, in: Intl. Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning, Helsinki, Finland, 2001, pp. 25–36.
[98] S.E. de Rooij, A. Abu-Hanna, M. Levi, E. de Jonge, Factors that predict outcome of intensive care treatment in very elderly patients: a review, Crit. Care 9 (2005) R307–R314.
[99] J.H.v. Bemmel, M.A. Musen, J.C. Helder, Handbook of Medical Informatics, Springer Verlag, Heidelberg, Germany, 1997.
[100] B. Zupan, A. Porenta, G. Vidmar, N. Aoki, I. Bratko, J.R. Beck, Decisions at hand: a decision support system on handhelds, Medinfo 10 (2001) 566–570.
[101] B. Zupan, J. Demsar, M.W. Kattan, M. Ohori, M. Graefen, M. Bohanec, J.R. Beck, Orange and Decisions-at-Hand: bridging predictive data mining and decision support., in: Intlerationa Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning, Helsinki, Finland, 2001, pp. 151–162.
[102] S.S. Abidi, Knowledge management in healthcare: towards 'knowledge-driven' decision-support services, Int. J. Med. Inf. 63 (2001) 5–18.
[103] M. Pazzani, Knowledge discovery from data? IEEE Intell. Syst. (March–April 2000) 10–13.
[104] K.J. Cios, G.W. Moore, Uniqueness of medical data mining, Artif. Intell. Med. 26 (2002) 1–24.
[105] J. Fox, S.K. Das, Safe and Sound: Artificial Intelligence In Hazardous Applications, MIT Press, Cambridge, Mass, 2000.
[106] R. Bellazzi, B. Zupan, Intelligent data analysis, Meth. Inf. Med. 40 (2001) 362–364.
[107] R. Haux, E. Ammenwerth, W. Herzog, P. Knaup, Health care in the information society A prognosis for the year 2013, Int. J. Med. Inf. 66 (2002) 3–21.
[108] Towards 2020 Science, Available at http://research.microsoft.com/towards2020science.