Focus Article

Predictive data mining in clinical medicine: a focus on selected methods and applications

Riccardo Bellazzi,[1*] Fulvia Ferrazzi[2] and Lucia Sacchi[1]

Predictive data mining in clinical medicine deals with learning models to predict patients' health. The models can be devoted to support clinicians in diagnostic, therapeutic, or monitoring tasks. Data mining methods are usually applied in clinical contexts to analyze retrospective data, thus giving healthcare professionals the opportunity to exploit large amounts of data routinely collected during their day-by-day activity. Moreover, clinicians can nowadays take advantage of data mining techniques to deal with the huge amount of research results obtained by molecular medicine, such as genetic or genomic signatures, which may allow transition from population-based to personalized medicine. The current challenge is to exploit data mining to build models able to take into account the dynamic and temporal nature of clinical care and to exploit the variety of information available at the bedside. This review describes the main features of predictive clinical data mining and focus on two specific aspects of particular interest: the methods able to deal with temporal data and the efforts performed to translate molecular medicine results into clinically useful data mining models. © 2011 John Wiley & Sons, Inc. *WIREs Data Mining Knowl Discov* 2011 1 416–430 DOI: 10.1002/widm.23

INTRODUCTION

Biomedical informatics is ultimately aimed at organizing, storing, and processing information on molecular and cellular processes, tissues and organs, individuals, population, and society to support the definition of suitable decision-making strategies in health care.[1] Within such complex scenario, the availability of analytical methods and tools to automatically interpret patients' data is essential. As a matter of fact, data mining technologies can be applied to all areas of the biomedical informatics field: bioinformatics, imaging informatics, clinical informatics, and public health informatics (see Figure 1). In this review, we will focus on the exploitation of predictive data mining in clinical informatics; in this context, the goal of predictive data mining is to derive models that can use patient-specific information to predict the outcome of interest and thus improve clinical decision-making.

The capability of 'predicting health' is certainly a major challenge of biomedical research and clinical medicine. Predictions may range from the simple stratification of the patients' population on the basis of known risk factors, such as age or lifestyle, to the forecast of the effect that a treatment or drug may have on a single patient. Generally speaking, in a clinical context, predictions may support diagnostic, therapeutic, or monitoring tasks. Diagnosis is related to the classification of patients into disease classes or subclasses on the basis of patients' data. This activity covers a broad spectrum of clinical cases, including triage at hospital emergency departments, that is, prioritizing patients based on the severity of their condition, as well as assigning cancer patients to tumor subtypes, which may require different therapeutic strategies. Therapeutic prediction is related to the choice of the most suitable treatment for the patient; this kind of prediction is very common in clinical contexts too, either for drug treatment planning or for the prognosis of surgical interventions. Finally, predictions in clinical monitoring are crucial in several contexts, such as in intensive care units (ICUs), in which the outcome of stay is

*Correspondence to: riccardo.bellazzi@unipv.it

[1]Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy

[2]IRCCS Fondazione S. Maugeri, Pavia, Italy; and Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy

DOI: 10.1002/widm.23

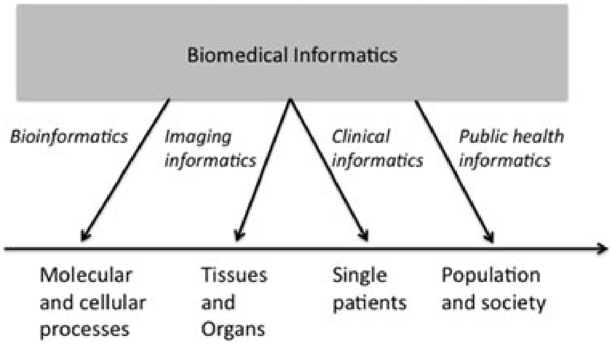

FIGURE 1 | Biomedical informatics is a discipline with several domains of applications. It deals with diverse data sources including molecular and cellular processes, tissues and organs, and individual patients and populations (adapted from Ref 1). Data mining can be successfully applied in all areas to support decision-making activities.

continuously updated on the basis of the monitoring data.

In several clinical contexts, time plays a crucial role as patient's care as well as data collection and decision-making activities are performed over time. It is therefore often mandatory to deal with the temporal aspects by deriving useful summaries of the patient's behavior, including physiological signals or measurement time series, and adapting the decisions to the accumulated data and information.

In recent years, the 'classical' prediction tasks involved by clinical activities have been empowered by new data coming from molecular medicine. Thanks to these data, it is now possible to build models based on a very large set of predictive 'biomarkers'. In general, biomarker is a set of physiological/biological variables used as an indicator of a biological or clinical state. Looking for biomarkers corresponds to searching for a set of variables within a set of measurable ones that well predict the state of interest.

Predictive medicine delivers the promise to change the paradigm of evidence-based medicine, which is aimed at applying the best available knowledge to medical decision-making.[2] Evidence-based clinical guidelines are often related to the definition of diagnostic or therapeutic strategies that are 'the average best choice' obtained for a large population. Evidence-based recommendations are typically formulated by exploiting the outcomes of clinical trials summarized by meta-analysis approaches. Predictive medicine may give the chance to adapt guidelines to the characteristics of the single patient, and/or to customize the decisions by taking into account the specific data set provided by a single hospital department.

Thanks to the growing availability of large amounts of experimental data, in all application contexts the role of computational methods has become fundamental to derive predictive models. Indeed, there are a growing number of papers that exploit data mining approaches for clinical prediction purposes. Because data mining is related to the data analysis process more than to the specific methods exploited,[3] it includes a mosaic of different computational techniques arising from both computer science and statistics, which may provide interesting solutions to the application needs.

This paper grounds on the review already presented by Bellazzi and Zupan[4] putting emphasis on some of the most promising areas in which predictive data mining has been used in clinical medicine over the last few years. After a brief general overview of the current trends of research, this review will focus on the analysis of temporal data and on the exploitation of '-omics' data coming from molecular medicine. These topics are somehow crucial to bridge the gap between the different kinds of information available in health care, from -omics to populations. On the one hand, time plays a fundamental role in all aspects of clinical settings, including clinical and home monitoring, patient surveillance, and public health informatics; its very nature requires the definition of specific data mining strategies. On the other hand, the integration of molecular information in clinical predictive models poses new challenges to data mining that have resulted in tools applied to clinical practice. This review will provide insights on the recent advances obtained in these interesting research areas.

DATA MINING APPLICATIONS IN CLINICAL MEDICINE: AN OVERVIEW

Clinical applications of data mining have a strong peculiarity. As reported in Ref 4, the most crucial aspect is that medical decision-making is safety critical: decisions always need to be strongly supported by arguments and models must be thoroughly evaluated. Second, data are precious: experiments are usually very expensive, often involve diseased people, and add further discomfort to the patients; moreover, the data sets may describe events that can be hardly reproduced in controlled conditions. Finally, the data may be also affected by several sources of uncertainty, including measurement errors, data entry and data coding mistakes, and missing data and textual reports. For these reasons, a distinctive aspect of medical data analysis is that it heavily relies on background clinical knowledge: data interpretation is grounded on

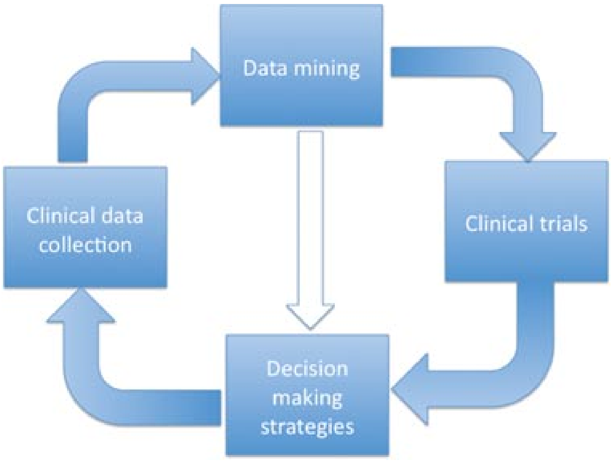

FIGURE 2 | Clinical decision-making is largely based on results coming from clinical trials. The analysis of clinical databases performed with data mining approaches provides a way to generate hypotheses for further trials and suggest changes in day-by-day practice based on the accumulated experience.

theories on the problem domain and on models of the population or of the single patient. Such theories and models allow overcoming the inherent uncertainty in the data and providing sound, although biased, interpretation of the events.

Clinical knowledge and clinical decision-making strategies are based on the results of clinical trials.[2] However, the process of deriving evidence from trials is slow and typically provides population-based results. The application of data mining techniques to clinical data gives the opportunity of generating hypotheses for further trials and experiments and, at the same time, timely proposing changes in decision-making strategies, which may be better tailored to the specific health care context (see Figure 2). Although clinical trials still represent the best way to perform counterfactual reasoning about clinical actions, a proper use of process data may lead to an optimization of clinical care in the daily routine.

This is a paradigmatic shift from clinical trials data analysis, as discovering predictive models requires a 'process-oriented' approach to avoid data dredging and limit, as much as possible, the impact of biases due to unspecific data collection.

The application of data mining in clinical medicine has been related to predictive and descriptive tasks, by applying both supervised and unsupervised techniques. As far as predictive tasks are concerned, both classification and regression problems have been largely studied, although the former is more often referred to as data mining in the biomedical literature (see Box *Predictive Data Mining Methods in Clinical Applications* for a brief overview of the most widely used classification approaches).

In the most recent papers in the field, supervised classifiers have been used for a variety of purposes including, for example, stratification of coronary heart disease patients' risk,[5] prediction of prostate cancer on the basis of tissue microarray data,[6] prediction of the outcome of bariatric surgery,[7] and population stratification for healthcare planning.[8]

An interesting trend of recent research is to exploit data mining strategies to perform feature selection before deriving predictive models based on statistical approaches, such as logistic and Cox regression (see Box *Feature Selection in Predictive Modeling*). For example, such an approach was applied to derive prognostic survival models after thoracic transplantations[9,10] and in the ICU.[11] Moreover, subgroup discovery methods were used to highlight critical variable subspaces when monitoring the blood glucose of ICU patients.[12] The issue of feature selection is of course one of the most interesting aspects of research on biomarker discovery, such as in the work of Baumgartner et al.,[13] who analyzed a large set of mass spectrometry data to extract metabolites relevant for infarction. Cox regression and a bootstrap elimination scheme were used to select variables useful for intermediate staging of renal cancer starting from tissue microarray data.[14]

Another crucial aspect is related to the evaluation and assessment of predictive models. As a matter of fact, several papers and commentaries have highlighted the lack of rigorous evaluation strategies in the biomedical data mining literature, because the feature selection step is often (wrongly) performed on the entire data set prior to the cross-validation, thus leading to overfitting and underestimation of the generalization error.[15]

The integration of background knowledge obout the clinical process under study and the fusion of information coming from related data sets are promising directions for improving the robustness of the obtained results.[16] For this reason, an interesting research direction concerns the development of methods able to automatically use the background (domain) knowledge in the data analysis process.

In recent years, a noteworthy effort has been performed to make the available biomedical knowledge accessible in electronic format, including formalized knowledge bases, ontologies, and the biomedical literature.[17] Such efforts have been leading to the implementation of integrated frameworks that allow querying and more efficiently analyzing large data sets.[18] This knowledge has been effectively used to improve feature selection, guiding model search, and post-processing the obtained results. Interesting results have been achieved in the area of rule learning, in which both rule search and rule pruning strategies have been successfully implemented, as reported by Raj et al.,[19] and Siadaty and Knaus.[20]

As mentioned in the previous section, in this review we will focus on two peculiar areas that are particularly novel and promising: the exploitation of temporal data mining to analyze time series and time sequences coming from healthcare clinical and administrative data repositories, and the extraction of prognostic models able to successfully combine molecular and clinical information. Following this, the basic methods and the most interesting applications will be reported and discussed.

PREDICTIVE CLINICAL DATA MINING WITH TEMPORAL DATA

Dealing with time in clinical practice represents a crucial but challenging issue.[21,22] Thanks to the new technologies that allow the collection and storage of a huge amount of temporal data, it is now possible to effectively explore the temporal dimension to improve clinical decision-making. Temporal reasoning and data mining have been working together to solve such a difficult task. The research field originated to fulfill these needs is commonly referred to as temporal data mining.[23–25]

Real clinical scenarios are often characterized by a variety of data collection settings that produce different kinds of temporal data outputs. On the one hand, in-hospital monitoring time series can be gathered. These include, for example, physiological signals monitored with electronic devices or specialized sensors (e.g., continuous blood glucose monitors) at the bedside of the patient (such as in the ICU). On the other hand, data from periodical visits or laboratory exams are collected. This is, for example, the case of cancer patients who undergo periodical visits during which some specific markers are evaluated. This second scenario typically originates data sets containing short time series (up to 10–25 time points), irregularly collected due to an uneven schedule of visits or measurements made. This results in a sampling grid that varies both along the patient-specific observation period and from patient to patient. In both data collection settings missing data and unreliable or noisy detections could sensibly affect the data analysis process.

All these considerations make clinical temporal data, in particular time series, not always suitable to be treated with traditional signal processing algorithms.

Along with time series of clinical data, administrative data are a very interesting source of information. Administrative records contain time-stamped data that depict the clinical history of patients: drug prescriptions, hospitalizations, outpatient visits, day-hospital activities, and so on. Such data are usually collected for administrative purposes, and in general do not contain clinical values (e.g., it is known that a patient has been admitted to the hospital with a specific diagnosis and that some laboratory tests have been performed but the specific results of such tests are not known). Databases containing administrative information as recorded, for example, by healthcare agencies, have been up to now only partially investigated. This is mainly due to the huge quantity of complex multivariate data stored therein. In particular, one of the main challenges in this case is to learn models that can be used for predictive purposes.[20,26,27]

The ultimate goal of temporal data mining applied to the clinical domain can be seen as the extraction of relevant patterns from data. The concept of temporal pattern is rather broad as it can assume different connotations when applied to different kinds of data. In general, when dealing with clinical data, a temporal pattern can be seen as the sequence of events that are clinically relevant for the onset of a particular condition. As already mentioned, the extraction of temporal patterns from clinical data is a multifaceted problem, which has been faced in the literature through different methodologies and which may find various solutions according to the type of data that have to be analyzed. Relying on this observation, the following section will give a review of the temporal data mining methods proposed to deal with two main categories of clinical data: time series and temporal sequences. This is a rather broad distinction and still very different clinical data can fall within the same group. Yet, each group is analyzed with methodologies that share common features, as it will be detailed below.

Mining Clinical Time Series

The main difference that exists between time series data and 'process' data (e.g., administrative health records) is that time series contain raw quantitative data, whereas administrative data are represented by sequences of events that have been already processed at a higher conceptual level.

For this reason, clinical time series often need to be preprocessed in order to make them suitable for data mining applications. To this end, different techniques have been proposed. As far as monitoring time series or signals like electrocardiogram (ECG) or electroencephalogram (EEG) are concerned, approaches coming from traditional signal processing have been widely applied typically to solve classification problems. In Ref 28, nonlinear principal component analysis (PCA) and neural networks (NNs) have been applied to ECG signals to detect ischemia. Preprocessing using wavelet transforms has been applied to the classification of ECG and EEG signals and physiological time series.[29–32] Wavelet transforms were also applied in some studies on arrhythmias detection in which they were coupled with NNs or support vector machines (SVMs).[33,34]

The application of traditional signal processing techniques usually requires time series data to satisfy some assumptions. Especially when dealing with short and irregular time series collected in hospital by clinicians or nurses or from outpatient visits or examinations, some alternative techniques have to be applied.

In order to deal with the extraction of temporal patterns from time series data, several data mining approaches have been proposed to synthesize temporal information into temporal features. In this review, we will focus on temporal abstractions (TAs).[35–37] The formalization of a framework for knowledge-based TAs has represented an important breakthrough for the application of temporal reasoning to biomedical problems. It allowed bridging the gap between the qualitative knowledge of the domain experts and the need for a sound computational model of the data. The main goal of abstraction techniques is to detect specific qualitative patterns able to abstract high-level concepts from time-stamped data. Such patterns could be very simple, such as trends or states, or more complex and involving many clinical parameters. TAs are powerful instruments because they enable the translation of the clinical knowledge of domain experts into features that can then be efficiently searched in the data. Moreover, they are a very flexible instrument, as they can be conveniently applied to different kinds of data.

The TA framework was first formalized by Shahar,[35] and Shahar and Musen[38] and subsequently applied to several clinical problems. The main areas of interest have been ICU data analysis,[39,40] blood glucose control and diabetes mellitus management,[41–44] and hemodialysis service assessment.[45] Verduijn et al.[39] present a comparison of two methods for feature extraction applied to the problem of predicting prolonged mechanical ventilation in ICU patients. A set of features derived using state and trend TAs was compared with a set of data driven abstractions calculated as simple statistical summaries of the monitoring time series (mean, median, standard deviation, etc.). The extracted features were then used to learn classification trees to predict the outcome. In Ref 46, TAs were used to detect peculiar events in ECG time series and those events were then used as features in arrhythmias detection. In Ref 47, a set of basic and complex TAs was used to extract temporal patterns from time series coming from periodical (but irregularly sampled) laboratory tests related to hepatitis. These data were then used by a rule learning algorithm to evaluate the differences in the temporal patterns of patients affected by hepatitis B and C.

The availability of the TA framework coupled with the need to process large amounts of time-oriented multivariate data has recently led to the integration of TA tools into clinical temporal reasoning systems. This integration is mainly devoted to the management of multiple patients and to the automatic extraction and summarization of the most peculiar temporal patterns present in the data. The most important direction toward which this research has been oriented regarded the integration of TAs into temporal query systems and the development of intelligent analysis and visualization tools. As regards the first point, temporal patterns defined through TAs have been made available as instruments to perform temporal queries able to retrieve specific clinical data sets of interest with the possibility of directly defining TAs.[48–52] Among systems for intelligent temporal data analysis and visualization, Klimov et al.[53] presented the VISualizatIon of Time-Oriented RecordS (VISITORS) system. This system combines intelligent temporal analysis and information visualization techniques to represent both raw data and abstracted concepts. Moreover, it allows exploration of the data and a first insight into associations among the temporal patterns that characterize the patients. Another system for assisted visual analysis of clinical time series data was presented by Aigner et al.[54] Through this system, temporal data abstraction, PCA, and clustering techniques can be exploited together for the management of large volumes of time-oriented clinical data. In Ref 55, a human–computer collaborative approach is designed for the exploration of biomedical multivariate time series to model and extract typical scenarios and parameter evolution. The process-oriented temporal analysis (PROTEMPA) software library[56] is a complex architecture that contains modules for the definition of TAs and the processing of time series data to obtain interval-based abstract patterns. Other modules devoted to sequential pattern mining can then use such patterns.

Mining Sequential Patterns

When analyzing temporal data, researchers usually have to deal with temporal sequences of events. In temporal data mining, an event in general is defined as a temporal attribute with an associated time-stamp or interval of occurrence. An event sequence is defined as a list of events, in which each event is associated with the same individual.[57,58] In clinical applications, a sequence of events can be, for example, the clinical history of a patient in terms of hospital admissions, drug prescriptions, and laboratory tests. It can also be the sequence of the intervals in which blood glucose has been out of a normal range. Temporal sequences of events can thus originate both from administrative data and from properly preprocessed time series data. Extracting meaningful patterns from temporal sequences of events is a very important field of research in temporal data mining. In a clinical scenario, such patterns can be conveniently used as a decisional support tool to predict future healthcare events. Since the first introduction of such algorithms,[59] which represent an extension of more traditional association rule learning techniques, many efforts have been made to develop more efficient search strategies to maximize computational performance.[60–63] Both methodologies to mine frequent temporal patterns on single time point sequences and on time intervals have been proposed in the temporal data mining community.[58,64–68] As a particular case of these techniques, some approaches have been developed to derive temporal association rules (TARs), in which a set of contemporaneous events precedes another event of interest.[69,70]

Recently, techniques for temporal pattern mining have been applied also to clinical temporal data.[19,45,71–76] In Ref 76, the authors present KarmaLego, an algorithm for fast time intervals related patterns mining. The method is applied on a data set of diabetic patients, on which a set of clinical variables such as blood glucose, cholesterol, hemoglobin, and the medications purchased are collected. Raj et al.[19] have developed an ontology-driven temporal pattern mining system, ChronoMiner. Such system allows the dynamic extraction of temporal associations at different hierarchical levels. It was applied on a data set of HIV patients to investigate the temporal relationships between new mutations due to the administered therapy. In Refs 45 and 71, the authors present a methodology to mine TARs on a set of complex temporal patterns such as trends or up and down behaviors in the data set. The methodology was first applied to assess the quality of the service delivered by a hemodialysis center[45] and then to evaluate the relationships between complex patterns in heart rate and blood pressure on the hemodialyzed patients.[71]

In health care, the application of temporal data mining techniques to the analysis of administrative data is a field that started to be explored only recently. Temporal pattern mining methods have been applied in the area of detecting adverse drug reactions starting from drug prescriptions and clinical data stored in administrative healthcare databases.[72,77–79] In Ref 72, a method is proposed to mine unexpected TARs from administrative health data taking into account sequences of events of drug prescription. Norén et al.[79] proposed a method for pattern discovery in large data sets of patient records. The methodology is based on a statistical and graphical approach to represent the association between drug prescriptions and clinical events registered in the patients' temporal history. In Ref 74, the authors present a method for the analysis of the data stored in a data warehouse of a local healthcare agency. This method allows taking into account both administrative and clinical temporal data, which are suitably preprocessed through a TA step. The method is able to extract TARs on sequences of hybrid events, that is, events that can be characterized by duration or point-like events according to a specific temporal granularity.[80,81] It was applied to the evaluation of the relationships between drug prescriptions and variations in the clinical conditions of diabetic patients. In Ref 82, the same methodology is applied to the assessment of the costs related to pharmacological treatment of diabetes.

PREDICTIVE CLINICAL DATA MINING FOR MOLECULAR MEDICINE

Large-scale genomics and proteomics data hold the promise of characterizing the molecular mechanisms underlying disease development and progression. Not only has their availability revolutionized basic research, but it has also started to affect clinical practice, paving the way for personalized medicine.[83,84] Statistical and data mining approaches can offer a fundamental aid in the interpretation of this data.

Gene expression microarrays are a well-established high-throughput technique for the expression profiling of tissue/cell samples.[85] One of the earliest areas in which the potential of expression arrays has been recognized and exploited is oncology. Thanks to large-scale expression profiling, it has been possible to identify previously unknown tumor subclasses as well as discriminate patients in different risk classes. Thus, expression profiling has both diagnostic and prognostic relevance.

In the case of breast cancers, the use of gene expression arrays has led to an improved molecular classification of tumors and to the development of commercially available expression-based prognostic assays.[86,87] The MammaPrint test, which aids in the recommendation of adjuvant therapy in patients with early stage breast cancer, has been the first *in vitro* diagnostic multivariate index assay to be approved by the U.S. Food and Drug Administration.[88] It is interesting to review the steps that led to the development of this test, starting from the gene expression profiling study by van't Veer et al.[89] They employed microarrays comprising about 25,000 human genes and identified around 5000 genes differentially expressed across 98 primary breast cancer samples (i.e., significantly regulated in more than five tumors). Hierarchical clustering found two distinct groups of 62 and 36 tumors, which can be labeled as 'good prognosis' and 'poor prognosis' on the basis of the associated clinical data. Next, the authors aimed at finding a prognostic gene expression signature, that is, a set of genes whose expression can be employed to predict the development of distant metastases. They concentrated the analysis on 78 samples from lymph node negative patients under 55 years of age at diagnosis; 34 of these developed distant metastases within 5 years, whereas the other 44 remained disease-free for over 5 years. The authors employed a three-step supervised classification approach: (1) about 5000 genes significantly regulated in more than three samples out of 78 were identified from the 25,000 genes; (2) the correlation coefficient of the expression of each gene with disease outcome was calculated and 231 genes whose absolute correlation coefficient was higher than 0.3 were selected and ordered on the basis of the magnitude of the coefficient; (3) the number of genes to be included in the final classifier was optimized by sequentially adding five genes from the ordered list and evaluating classification accuracy through leave-one-out cross-validation. Classification of a leave-one-out sample was obtained on the basis of the correlation of its expression profile with the mean expression profile of the remaining good prognosis and poor prognosis samples. The optimal number of genes in the final classifier was found to be 70; this is the so-called gene signature. This classifier achieved 83% prediction accuracy. An independent set of 19 young lymph node negative breast cancer patients was employed to validate the classifier. The set included seven patients who stayed metastasis free for at least 5 years and 12 patients who developed distant metastases within 5 years. The classifier correctly predicted disease outcome for 17 patients.

The gene signature by van de Vijver and colleagues[90] underwent further validation in a retrospective study on 295 breast cancer patients. This study was criticized because 130 patients had already received adjuvant chemotherapy or hormonal therapy and 61 patients were part of the group employed to develop the classifier itself, thus opening the way to overfitting issues. To address these concerns, a further retrospective validation study on 302 patients from several institutions was performed.[91] Thanks to the positive results obtained in this study, a large prospective validation trial has been planned. This trial, called MINDACT, started in 2007 and is expected to recruit about 6000 patients.[92] In the meanwhile, the MammaPrint diagnostic test, based on the 70-gene signature, was created.[93]

The 70-gene signature for metastasis prediction in breast cancer patients has been one of the most successful results obtained through the use of expression data for clinical outcome prediction. Different other works exist in which more advanced data mining and statistical techniques for prediction from gene expression data have been employed. An example is given by a study by Bovelstad et al.[94] who compared the performance of different methodologies that rely on Cox proportional hazard model, to predict survival from expression data. In particular, the authors evaluated seven-parameter estimation approaches: univariate selection, forward stepwise selection, principal component regression, supervised principal component regression, partial least squares regression, ridge regression, and the lasso (see Box *Feature Selection in Predictive Modeling* for a summary of modern feature selection methods in regression problems). They employed three microarray data sets to assess the prediction of the different methods: the breast cancer data set from van de Vijver et al.,[90] another breast cancer data set, and a data set on diffused large B-cell lymphoma. The study showed that more advanced methods outperformed simple selection rules, with ridge regression achieving the overall best performance. The authors also suggested that the use of lasso regression might be very useful for the development of diagnostic arrays as it is able to perform variable selection.

Different review papers have discussed the potential of expression data for diagnosis and prognosis in oncology and have also addressed issues and limitations of the studies so far performed, discussing possible future developments (see, e.g., Refs 87, 95, and 96). Guidelines to avoid statistical pitfalls that affected a significant number of earlier studies, above all overestimation of accuracy due to overfitting, have also been pointed out by some authors.[97]

Besides gene expression studies, genetic association studies have received increasing interest in recent years.[98] These studies aim at assessing whether a correlation exists between genotypic markers and a phenotype of interest (which can be occurrence of a disease or a quantitative trait such as height). Common types of association studies are case-control studies in which single nucleotide polymorphisms (SNPs) are measured for two groups of subjects, one of patients affected from a disease of interest and the other of healthy controls. Furthermore, a candidate gene approach or a genome-wide approach can be taken. In the former, SNPs are genotyped for a set of genes that are likely to be related to the disease of interest, whereas in the latter high-density arrays measuring several hundreds of thousands of SNPs are employed. It is worth noting that genome-wide studies require a number of samples in the order of at least some hundreds in order to achieve statistically significant results.

The majority of genetic association studies have limited themselves to the selection of the most informative single SNPs for phenotype prediction. Selection occurs through the use of univariate tests, or tests in which the association of the single SNP with the outcome is ‘adjusted’ taking into account possible confounding factors, such as sex, age, or smoking. In some cases, statistical tests to verify the presence of an effect due to the interaction of pairs of SNPs are performed.

Yet, the use of multiple SNPs for predictive modeling is expected to achieve better predictive power.[99] A successful study was performed by Sebastiani et al.,[100] who exploited Bayesian networks to build a predictive model for the risk of stroke in sickle cell anemia patients. The authors analyzed 108 SNPs in 39 candidate genes in 1398 sickle cell anemia subjects, 92 with reported stroke, and 1306 without stroke. Starting from the genotypic measurements for these SNPs and a number of available patient-specific clinical variables, a Bayesian network describing the probabilistic dependencies between the genotypic and phenotypic variables has been learned. The inferred network describes the association of stroke with 69 SNPs in 20 genes, fetal hemoglobin levels, total hemoglobin concentration, and thalassemia.

Probabilistic inference algorithms lead to the usage of the learned Bayesian network as a classifier, as they make it possible to predict the probability of the occurrence of stroke given all the other variables in the network. Moreover, a variable selection tool is implicitly provided by the Bayesian network learning algorithm as, by the global Markov property, a node is conditionally independent of all other variables in the network given its Markov blanket (i.e., its parent nodes, the children nodes, and the parents of the children). In the network inferred by Sebastiani et al.,[100] the risk of stroke can be predicted by relying on the knowledge of only 31 SNPs in 12 genes and fetal hemoglobin. The Bayesian network model had 98.5% predictive accuracy in fivefold cross-validation. The classifier was also evaluated on an independent test set of 114 subjects, seven with reported stroke and 107 without stroke, and its classification accuracy was 98.2%, with 100% true positive rate. As a comparison, the authors built a logistic regression model relying on a stepwise regression strategy. Validation of this model on the same independent test set had 88% accuracy.

The study showed that Bayesian networks are a powerful framework to build multivariate predictive models able to integrate genotypic and phenotypic data. Bayesian networks have been later successfully employed also in other clinical settings, for the prediction of cardioembolic stroke,[101] nicotine dependence,[102] and coronary artery calcification in atherosclerosis.[103]

As mentioned above, the use of multivariate genetic models is still limited in the biomedical literature, even if different methods are being investigated in more methodological studies. For example, Malovini et al.[104] proposed the use of gene-based Bayesian networks, in which a metavariable for SNPs mapping to the same gene is created by relying on classification trees; Wu et al.[105] proposed lasso logistic regression for use with a large number of SNPs, whereas Hoggart et al.[106] used a ‘Bayesian-inspired’ approach for simultaneous analysis of all SNPs from a genome-wide study. The use of multi-SNP models has also been advocated by Thomas[107] in his presentation of the methodologies for discrete traits proposed by the participants of Genetic Analysis Workshop 16 Group 1, who analyzed a case study data set on rheumatoid arthritis.

Expression and SNP data are two significant categories of molecular data employed in predictive modeling but not the only ones. The availability of proteomics data has also been growing fast in recent years and their potential for clinical applications quickly recognized. Also in this area, machine learning methodologies can offer a great aid in the identification of biomarkers.[108]

Despite a number of significant examples in the literature, such as the ones discussed above, the number of studies in which the developed predictive models have shown a clear diagnostic or prognostic relevance is still low compared with the increasingly high amount of data available. Given the wealth of

BOX 1: PREDICTIVE DATA MINING METHODS IN CLINICAL APPLICATIONS

The application of data mining methods for predictive medicine is traditionally related to the exploitation of supervised classification approaches, in which the class is represented by a disease (in diagnostic problems) or by an event of clinical interest (e.g., mortality in therapy planning problems). Predictive data mining methods originate from different research fields and often employ very diverse modeling approaches.

Statistical methods are widely used in the medical literature. As classification problems are concerned, logistic regression is recognized to be a powerful and well-established method. Probabilistic classifiers, such as Naive Bayes classifiers and Bayesian networks, are also exploited in the medical literature, although less frequently. Survival prediction is often coupled with standard survival analysis; in particular, Cox regression is widely exploited to analyze the effect of covariates on time-to-event.

In the recent years, decision trees have also been increasingly applied, because they learn easy-to-interpret decision rules and can be used to highlight interesting stratification in the data. An extension of decision trees, called random forests, has also been applied successfully in the analysis of molecular medicine data.

Artificial NNs are black box artificial intelligence-based algorithms that have been until recently quite popular in clinical medicine, although they are prone to overfitting. A more robust approach, grounded in statistical learning theory, is provided by SVM, which may handle both linear and nonlinear decision boundaries. Although the learned classifiers may be difficult to interpret, the classification performance can be quite high.

BOX 2: TEMPORAL DATA MINING APPROACHES

Clinical medicine often deals with temporal data. Such data may be time series and/or sequences of events. Time series are collection of time-stamped quantitative measurements, whereas temporal sequences are usually labels associated to a clinical event, such as a drug prescription. In the literature, the two problems have been tackled with a variety of methods.

Time series have been studied with traditional signal processing and feature extraction methods, which are devoted to summarize the temporal information in attributes suitable for classification algorithms, such as NNs and SVMs. Another class of approaches is represented by the application of TA methods, aimed at extracting qualitative temporal patterns. These patterns are then associated to meaningful labels for the clinician. TAs are able to translate qualitative clinical knowledge into a computational mechanism that may be readily applied on the available data.

Time sequences have been analyzed exploiting algorithms able to extract frequent temporal patterns. Such algorithms are an extension of the well-known Apriori method for learning association rules. A subset of these algorithms is devoted to the extraction of TARs that can be used for predictive purposes.

applicable data mining techniques, this might seem surprising. Two main factors can account for this discrepancy. First of all, the high dimensionality of the data sets makes variable selection a key step in the analysis and also requires the availability of a truly independent test set to assess reproducibility of results. The second factor, easier to overlook at first for a methodologically focused researcher, is the specific nature of the data. Molecular biology data are noisy and their noise is due not only to the measurement techniques employed but also to the intrinsic and unavoidable heterogeneity between the studied subjects. A very careful study design, aimed at removing potential confounding factors, can partially mitigate the presence of noise and allow highlighting genetic variations and/or differential gene expression due only to the pathological process under examination. Also, the development of predictive methodologies should be guided by the biomedical knowledge already available for the studied problem. This intuition has led to the design of methodologies that exploit 'prior' knowledge, that is, knowledge already available in the literature and in online databases. In this area of research, Bayesian methods are particularly promising as they offer a powerful framework for integrating prior knowledge in data analysis.

Overall, a very close interplay between biologists, medical doctors, and computational researchers, from the very first phase of planning a study, is essential for increasing chances of establishing results that can be translated into clinical practice.

CONCLUSION

Clinical medicine is one of the most interesting areas in which data mining may have an important practical impact. The widespread availability of large clinical data collections enables thorough retrospective analysis, which may give healthcare institutions

BOX 3: FEATURE SELECTION IN PREDICTIVE MODELING

A predictive model is designed to forecast a response variable. Such variable may be categorical or numerical, so that predictive data mining may deal with classification and regression problems, respectively. Feature selection is an integral part of both classification and regression and needs to be included in the model validation process. Thus, care must be taken to ensure that examples included in the test data set have not been part of the data set used to select features. The selection of the features can occur prior to model estimation relying on unsupervised strategies, as in PCA. Principal components are calculated on the explanatory variables and the predictive model is then learned using only the first *k* principal components as covariates, which are able to explain a desired percentage of the input data variance.

A variety of other approaches rely on supervised strategies to perform feature selection. In the biomedical field, it is quite common to simply rank the features on the basis of their predictive capability, measured in terms of a suitable objective function, such as information gain or ReliefF, and then select a subset of those features. Another widely applied strategy involves forward or backward stepwise feature selection; this approach builds nested models of different complexity and is aimed at finding a proper compromise between model complexity and predictive accuracy. Several other feature selection strategies can be applied, such as bootstrap methods, to estimate features statistical significance.

Partial least squares regression performs regression on a smaller number of attributes, which are linear combinations of the original covariates. These components are chosen in order to maximize the covariance of the attributes with the outcome.

Ridge and lasso regression employ a penalized likelihood cost function in order to 'shrink' the regression coefficients. In ridge regression, the penalty term is proportional to the sum of the squared values of the coefficients, whereas in lasso regression, absolute values of the coefficients are employed. The proportionality constant is a 'tuning parameter' that determines how much shrinkage occurs. The effect of ridge regression is that some regression coefficients are shrunk toward zero; in lasso regression, instead, some of the estimated coefficients will become exactly equal to zero. Thus, lasso regression, recently exploited also in classification problems, embeds variable selection in the model estimation.

BOX 4: PREDICTIVE MEDICINE AND -OMICS SCIENCES

In the recent years, a variety of new molecular level data have been made available to researchers and clinicians by biotechnological advances.

First of all, it is now possible to obtain measurement of the genotype of each individual, that is, the specific DNA sequence of a subject that is common to all cells of that subject. As cellular DNA is made of millions of nucleotide bases, a genotype is usually evaluated by measuring specific bases which are known to vary in the population called SNPs. The individual genotype is correlated with the phenotype, which is any observable characteristic of an organism, including traits, such as the height of an individual or the presence of a disease. Other technologies allow obtaining several measurements that describe the activity of the cell. DNA plays the crucial role of storing the information that enables the cell to dynamically synthesize proteins by means of a rather complex biochemical process, which involves the transcription of DNA into an intermediate molecule called RNA. The DNA regions that contain such information are called genes. Although genomics deals with the study of DNA, genes, their transcription into RNA, and translation into proteins, proteomics is about the structure, role, and interaction of proteins. Data mining strategies may be applied to learn models able to predict a phenotype on the basis of the patient's genotype, relying on the transcriptional activity of a certain cell/tissue or finally looking at the proteomic profiles of cells, tissues, and serum.

an unprecedented opportunity to better understand the nature and peculiarity of the undergoing clinical processes. Moreover, the availability of large-scale molecular data may offer new insights on the single patient case and suggest changes in decision-making strategies. The area of predictive data mining may therefore support medicine in its transformation from population-based to personalized approaches. To this end, the use of methods able to deal with temporal information is crucial, as well as the development of novel data analysis tools able to integrate data and knowledge in a coherent framework.

Learning predictive models that can be applied in the clinics requires, however, considerable attention (see Figure 3). First of all, the model should be soundly statistically evaluated. After the first estimate of generalization error performed with cross-validation and bootstrap methods, the assessment of

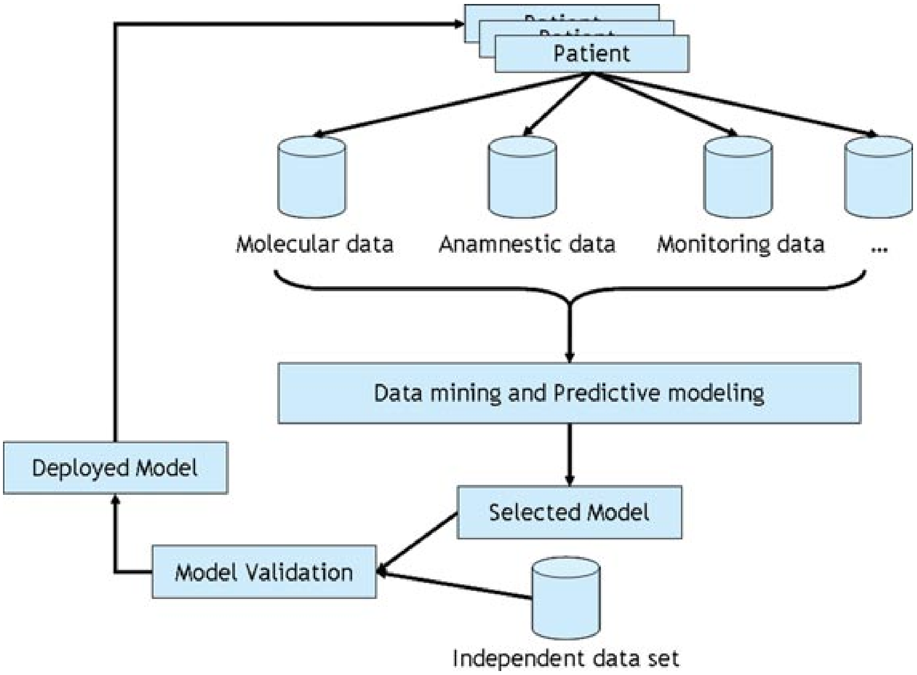

FIGURE 3 | Learning clinical predictive models requires a careful evaluation process. Different data sources need to be properly integrated and preprocessing and feature selection may turn out to be the most important parts of data analysis. Model evaluation requires an independent data set to assess the prediction performance. Finally, the model should be deployed carefully taking into account the clinical context.

the prediction performance on a sufficiently large independent test set is mandatory. This is also required given the very nature of retrospective data analysis, which can be biased by several confounding factors. Moreover, the learning process should be as transparent and reproducible as possible. Medicine is a safety critical context and all modeling steps, including the choice of design parameters, should be clear and well justified. Finally, the choice of the predictive model is also related to its capability of being deployed in a clinical context: models that are easy to be explained and are supported by statistical evidence are more likely to be adopted in clinical practice than black box ones. Given the complexity of the problem and grounding on successes and failures of data mining methods reported in the biomedical literature, Bellazzi and Zupan[4] have proposed a set of guidelines that should be applied to the construction of clinical predictive models that take into account the issues mentioned above.

It is likely that the incoming years will see an increase in the application of data mining and data analytics technology in medicine. It will be the responsibility of the researchers in the field to apply their methods and tools for properly moving from ‘bench’ to ‘bed’.

REFERENCES

1. Shortliffe EH. JBI status report. *J Biomed Inform* 2002, 35:279–280.
2. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn’t. *BMJ* 1996, 312:71–72.
3. Fayyad U, Uthurusamy R. Data mining and knowledge discovery in databases. *Commun ACM* 1996, 39:24–26.
4. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform* 2008, 77:81–97.

5. Gregori D, Bigi R, Cortigiani L, Bovenzi F, Fiorentini C, Picano E. Non-invasive risk stratification of coronary artery disease: an evaluation of some commonly used statistical classifiers in terms of predictive accuracy and clinical usefulness. *J Eval Clin Pract* 2009, 15:777–781.
6. Demichelis F, Magni P, Piergiorgi P, Rubin MA, Bellazzi R. A hierarchical Naive Bayes model for handling sample heterogeneity in classification problems: an application to tissue microarrays. *BMC Bioinformatics* 2006, 7:514.
7. Lee YC, Lee WJ, Lin YC, Liew PL, Lee CK, Lin SC, Lee TS. Obesity and the decision tree: predictors of sustained weight loss after bariatric surgery. *Hepatogastroenterology* 2009, 56:1745–1749.
8. Weinstein L, Radano TA, Jack T, Kalina P, Eberhardt JS 3rd. Application of multivariate probabilistic (Bayesian) networks to substance use disorder risk stratification and cost estimation. *Perspect Health Inf Manag* 2009, 6:1b.
9. Delen D, Oztekin A, Kong ZJ. A machine learning-based approach to prognostic analysis of thoracic transplantations. *Artif Intell Med* 2010, 49:33–42.
10. Oztekin A, Delen D, Kong ZJ. Predicting the graft survival for heart-lung transplantation patients: an integrated data mining methodology. *Int J Med Inform* 2009, 78:e84–96.
11. Minne L, Abu-Hanna A, de Jonge E. Evaluation of SOFA-based models for predicting mortality in the ICU: a systematic review. *Crit Care* 2008, 12: R161.
12. Nannings B, Bosman RJ, Abu-Hanna A. A subgroup discovery approach for scrutinizing blood glucose management guidelines by the identification of hyperglycemia determinants in ICU patients. *Methods Inf Med* 2008, 47:480–488.
13. Baumgartner C, Lewis GD, Netzer M, Pfeifer B, Gerszten RE. A new data mining approach for profiling and categorizing kinetic patterns of metabolic biomarkers after myocardial injury. *Bioinformatics* 2010, 26:1745–1751.
14. Dahinden C, Ingold B, Wild P, Boysen G, Luu VD, Montani M, Kristiansen G, Sulser T, Bühlmann P, Moch H, et al. Mining tissue microarray data to uncover combinations of biomarker expression patterns that improve intermediate staging and grading of clear cell renal cell cancer. *Clin Cancer Res* 2010, 16:88–98.
15. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003, 95:14–18.
16. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, et al. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 2006, 7:96.
17. Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. *Yearb Med Inform* 2008:91–101.
18. Tu SW, Tennakoon L, O'Connor M, Shankar R, Das A. Using an integrated ontology and information model for querying and reasoning about phenotypes: the case of autism. *AMIA Annu Symp Proc* 2008:727–731.
19. Raj R, O'Connor MJ, Das AK. An ontology-driven method for hierarchical mining of temporal patterns: application to HIV drug resistance research. In: Teich JM, Suermondt J, Hripcsak G, eds. *AMIA Annual Symposium*. Chicago: AMIA; 2007, 614–619.
20. Siadaty MS, Knaus WA. Locating previously unknown patterns in data-mining results: a dual data-and knowledge-mining method. *BMC Med Inform Decis Mak* 2006, 6:13.
21. Augusto JC. Temporal reasoning for decision support in medicine. *Artif Intell Med* 2005, 33:1–24.
22. Adlassnig KP, Combi C, Das AK, Keravnou ET, Pozzi G. Temporal representation and reasoning in medicine: research directions and challenges. *Artif Intell Med* 2006, 38:101–113.
23. Roddick JF, Spiliopoulou M. A survey of temporal knowledge discovery paradigms and methods. *IEEE Trans Knowledge Data Eng* 2002, 14:750–767.
24. Post AR, Harrison JH. Temporal data mining. *Clin Lab Med* 2008, 28:83–100.
25. Mitsa T. *Temporal Data Mining*. Boca Raton, FL: CRC Press; 2010.
26. Silberschatz A, Tuzhilin A. What makes patterns interesting in knowledge discovery systems. *IEEE Trans Knowledge Data Eng* 1996, 8:970–974.
27. Ohsaki M, Abe H, Tsumoto S, Yokoi H, Yamaguchi T. Evaluation of rule interestingness measures in medical knowledge discovery in databases. *Artif Intell Med* 2007, 41:177–196.
28. Stamkopoulos T, Diamantaras K, Maglaveras N, Strintzis M. ECG analysis using nonlinear PCA neural networks for ischemia detection. *IEEE Trans Signal Process* 1998, 46:3058–3067.
29. Sternickel K. Automatic pattern recognition in ECG time series. *Comput Methods Programs Biomed* 2002, 68:109–115.
30. Chaovalitwongse WA, Prokopyev OA, Pardalos PM. Electroencephalogram (EEG) time series classification: applications in epilepsy. *Ann Oper Res* 2006, 148:227–250.
31. Zhang H, Ho T, Lin M-S, Liang X. Feature extraction for time series classification using discriminating wavelet coefficients. In: *Advances in Neural Networks—ISNN 2006*. Heidelberg, Berlin: Springer; 2006, 1394–1399.
32. Chuah MC, Fu F. ECG anomaly detection via time series analysis. Frontiers of high performance computing and networking. ISPA 2007 Workshops.

In: *Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag; 2007, 4743:123–135.

33. Joshi A, Rajshekhar, Chandran S, Phadke S, Jayaraman V, Kulkarni B. Arrhythmia classification using local Hölder exponents and support vector machine. In: *Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag; 2005, 3776:242–247.
34. Asl BM, Setarehdan SK, Mohebbi M. Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. *Artif Intell Med* 2008, 44:51–64.
35. Shahar Y. A framework for knowledge-based temporal abstraction. *Artif Intell* 1997, 90:79–133.
36. Stacey M, McGregor C. Temporal abstraction in intelligent clinical data analysis: a survey. *Artif Intell Med* 2007, 39:1–24.
37. Catley C, Stratti H, McGregor C. Multi-dimensional temporal abstraction and data mining of medical time series data: trends and challenges. *Conf Proc IEEE Eng Med Biol Soc* 2008, 2008:4322–4325.
38. Shahar Y, Musen MA. Knowledge-based temporal abstraction in clinical domains. *Artif Intell Med* 1996, 8:267–298.
39. Verduijn M, Sacchi L, Peek N, Bellazzi R, de Jonge E, de Mol BA. Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data. *Artif Intell Med* 2007, 41:1–12.
40. Moskovitch R, Peek N, Shahar Y. Classification of ICU patients via temporal abstraction and temporal patterns mining. In: *Proceedings of the 14th Workshop on Intelligent Data Analysis In Biomedicine and Pharmacology (IDAMAP 2009)*. Verona, Italy; 2009, 35–40.
41. Bellazzi R, Larizza C, Magni P, Montani S, Stefanelli M. Intelligent analysis of clinical time series: an application in the diabetes mellitus domain. *Artif Intell Med* 2000, 20:37–57.
42. Silva A, Cortez P, Santos MF, Gomes L, Neves J. Rating organ failure via adverse events using data mining in the intensive care unit. *Artif Intell Med* 2008, 43:179–193.
43. Bellazzi R, Abu-Hanna A. Data mining technologies for blood glucose and diabetes management. *J Diabetes Sci Technol* 2009, 3:603–612.
44. Seyfang A, Paesold M, Votruba P, Miksch S. Improving the execution of clinical guidelines and temporal data abstraction high-frequency domains. *Stud Health Technol Inform* 2008, 139:263–272.
45. Bellazzi R, Larizza C, Magni P, Bellazzi R. Temporal data mining for the quality assessment of hemodialysis services. *Artif Intell Med* 2005, 34:25–39.
46. Carrault G, Cordier MO, Quiniou R, Wang F. Temporal abstraction and inductive logic programming for arrhythmia recognition from electrocardiograms. *Artif Intell Med* 2003, 28:231–263.
47. Takabayashi K, Ho TB, Yokoi H, Nguyen TD, Kawasaki S, Le SQ, Suzuki T, Yokosuka O. Temporal abstraction and data mining with visualization of laboratory data. *Stud Health Technol Inform* 2007, 129 1304–1308.
48. Spokoiny A, Shahar Y. A knowledge-based time-oriented active database approach for intelligent abstraction, querying and continuous monitoring of clinical data. *Stud Health Technol Inform* 2004, 107:84–88.
49. Shahar Y, Goren-Bar D, Boaz D, Tahan G. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. *Artif Intell Med* 2006, 38:115–135.
50. Post AR, Sovarel AN, Harrison JH Jr. Abstraction-based temporal data retrieval for a Clinical Data Repository. *AMIA Annu Symp Proc* 2007:603–607.
51. Mabotuwana T, Warren J. An ontology-based approach to enhance querying capabilities of general practice medicine for better management of hypertension. *Artif Intell Med* 2009, 47:87–103.
52. O'Connor M, Shankar R, Das A. An ontology-driven mediator for querying time-oriented biomedical data. In: *19th IEEE Symposium on Computer Based Medical Systems*. Salt Lake City, Utah; 2006, 264–269.
53. Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. *Artif Intell Med* 2010, 49:11–31.
54. Aigner W, Miksch S, Muller W, Schumann H, Tominski C. Visual methods for analyzing time-oriented data. *IEEE Trans Vis Comput Graph* 2008, 14:47–60.
55. Guyet T, Garbay C, Dojat M. Knowledge construction from time series data using a collaborative exploration system. *J Biomed Inform* 2007, 40:672–687.
56. Post AR, Harrison JH Jr. PROTEMPA: a method for specifying and identifying temporal sequences in retrospective data for patient selection. *J Am Med Inform Assoc* 2007, 14:674–683.
57. Mannila H, Toivonen H, Verkamo AI. Discovery of frequent episodes in event sequences. *Data Mining Knowledge Discov* 1997, 1:259–289.
58. Kam PS, Fu AWC. Discovering temporal patterns for interval-based events. In: Kambayashi Y, Mohania M, Tjoa AM, eds. *2nd International Conference on Data Warehousing and Knowledge Discovery*. London: Springer-Verlag; 2000, 317–326.
59. Agrawal R, Srikant R. Mining sequential patterns. In: Yu PS, Chen ALP, eds. *11th International Conference on Data Engineering*. Taipei: IEEE Computer Society; 1995 3–14.
60. Srikant R, Agrawal R. Mining sequential patterns: generalizations and performance improvements. In: Apers PMG, Bouzeghoub M, Gardarin G, eds. *5th International Conference on Extending Database*

Technology: Advances in Database Technology. Avignon: Springer-Verlag; 1996 3–17.

61. Ng RT, Lakshmanan LVS, Han J, Pang A: Exploratory mining and pruning optimizations of constrained associations rules. In: Clifford J, Lindsay B, Maier D, eds. *ACM-SIGMOD International Conference on Management of Data*. Seattle: ACM; 1998, 13–24.
62. Bayardo RJ, Agrawal R, Gunopulos D. Constraint-based rule mining in large, dense databases. In: Kitsurgawa M, Maciaszek L, Papazoglou M, Pu C, eds. *15th International Conference on Data Engineering*. Sydney: IEEE Computer Society; 1999, 188–197.
63. Pei J, Han J, Asl BM, Pinto H, Chen Q, Dayal U, Hsu M. PrefixSpan: mining sequential patterns by prefix-projected growth. In: *17th International Conference on Data Engineering*. Heidelberg: IEEE Computer Society; 2001, 215–224.
64. Zaki MJ: SPADE: an efficient algorithm for mining frequent sequences. *Machine Learn* 2001, 42:31–60.
65. Ayres J, Flannick J, Gehrke J, Yiu T. Sequential PAttern mining using a bitmap representation. In: Hand D, Keim D, Ng R, eds. *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton: ACM; 2002, 429–435.
66. Mörchen F, Ultsch A. Efficient mining of understandable patterns from multivariate interval time series. *Data Mining Knowledge Discov* 2007, 15:181–215.
67. Patel D, Hsu W, Lee ML. Mining relationships among interval-based events for classification. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. Vancouver: ACM; 2008.
68. Zhang L, Chen G, Brijs T, Zhang X. Discovering during-temporal patterns (DTPs) in large temporal databases. *Expert Systems Appl* 2008, 34:1178–1189.
69. Höppner F, Klawonn F. Finding informative rules in interval sequences. *Intell Data Anal* 2002, 3:237–256.
70. Winarko E, Roddick JF. ARMADA—An algorithm for discovering richer relative temporal association rules from interval-based data. *Data Knowledge Eng* 2007, 63:76–90.
71. Sacchi L, Larizza C, Combi C, Bellazzi R. Data mining with temporal abstractions: learning rules from time series. *Data Mining Knowledge Discov* 2007, 15:217–247.
72. Jin HW, Chen J, He H, Williams GJ, Kelman C, O'Keefe CM. Mining unexpected temporal associations: applications in detecting adverse drug reactions. *IEEE Trans Inf Technol Biomed* 2008, 12:488–500.
73. Batal I, Sacchi L, Bellazzi R, Hauskrecht M. Multivariate time series classification with temporal abstractions. *Int J Artif Intell Tools* 2009, 22:344–349.
74. Concaro S, Sacchi L, Cerra C, Bellazzi R. Mining administrative and clinical diabetes data with temporal association rules. *Stud Health Technol Inform* 2009, 150: 574–578.
75. Bellazzi R, Sacchi L, Concaro S. Methods and tools for mining multivariate temporal data in clinical and biomedical applications. *Conf Proc IEEE Eng Med Biol Soc* 2009, 2009:5629–5632.
76. Moskovitch R, Shahar Y. Medical temporal-knowledge discovery via temporal abstraction. *AMIA Annu Symp Proc* 2009, 2009:452–456.
77. Li J, Fu AW-c, He H, Chen J, Jin H, McAullay D, Williams G, Sparks R, Kelman C. Mining risk patterns in medical data. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. Chicago, IL: ACM; 2005.
78. Chazard E, Preda C, Merlin B, Ficheur G, Beuscart R. Data-mining-based detection of adverse drug events. *Stud Health Technol Inform* 2009, 150:552–556.
79. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining Knowledge Discov* 2010, 20:361–387.
80. Combi C, Franceschet M, Peron A. Representing and reasoning about temporal granularities. *J Logic Comput* 2004, 14:51–77.
81. Combi C, Pinciroli F, Pozzi G. Managing different time granularities of clinical information by an interval-based temporal data model. *Methods Inf Med* 1995, 34:458–474.
82. Concaro S, Sacchi L, Cerra C, Stefanelli M, Fratino P, Bellazzi R. Temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment. In: *AMIA Annual Symposium*. San Francisco: AMIA; 2009, 119–123.
83. Collins F. Has the revolution arrived? *Nature* 2010, 464:674–675.
84. Niederhuber JE. Translating discovery to patient care. *JAMA* 2010, 303:1088–1089.
85. Slonim DK, Yanai I. Getting started in gene expression microarray analysis. *PLoS Comput Biol* 2009, 5:e1000543.
86. Cianfrocca M, Gradishar W. New molecular classifications of breast cancer. *CA Cancer J Clin* 2009, 59:303–313.
87. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. *N Engl J Med* 2009, 360:790–800.
88. U.S. Department of Health and Human Services. Available at: http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2007/ucm108836.htm (Accessed October 8, 2010).
89. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002, 415:530–536.

90. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002, 347:1999–2009.
91. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, d'Assignies MS, Bergh J, Lidereau R, Ellis P, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006, 98:1183–1192.
92. Cardoso F, van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ. Clinical application of the 70-gene profile: the MINDACT trial. *J Clin Oncol* 2008, 26:729–735.
93. Glas AM, Floore A, Delahaye LJ, Witteveen AT, Pover RC, Bakx N, Lahti-Domenici JS, Bruinsma TJ, Warmoes MO, Bernards R, et al. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics* 2006, 7:278.
94. Bovelstad HM, Nygard S, Storvold HL, Aldrin M, Borgan O, Frigessi A, Lingjaerde OC. Predicting survival from microarray data—a comparative study. *Bioinformatics* 2007, 23:2080–2087.
95. Desmedt C, Ruiz-Garcia E, Andre F. Gene expression predictors in breast cancer: current status, limitations and perspectives. *Eur J Cancer* 2008, 44:2714–2720.
96. Wouters BJ, Lowenberg B, Delwel R. A decade of genome-wide gene expression profiling in acute myeloid leukemia: flashback and prospects. *Blood* 2009, 113:291–298.
97. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst* 2007, 99:147–157.
98. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008, 9:356–369.
99. Sebastiani P, Timofeev N, Dworkis DA, Perls TT, Steinberg MH. Genome-wide association studies and the genetic dissection of complex traits. *Am J Hematol* 2009, 84:504–515.
100. Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat Genet* 2005, 37:435–440.
101. Ramoni RB, Himes BE, Sale MM, Furie KL, Ramoni MF. Predictive genomics of cardioembolic stroke. *Stroke* 2009, 40:S67–S70.
102. Ramoni RB, Saccone NL, Hatsukami DK, Bierut LJ, Ramoni MF. A testable prognostic model of nicotine dependence. *J Neurogenet* 2009, 23:283–292.
103. McGeachie M, Ramoni RL, Mychaleckyj JC, Furie KL, Dreyfuss JM, Liu Y, Herrington D, Guo X, Lima JA, Post W, et al. Integrative predictive model of coronary artery calcification in atherosclerosis. *Circulation* 2009, 120:2448–2454.
104. Malovini A, Nuzzo A, Ferrazzi F, Puca AA, Bellazzi R. Phenotype forecasting with SNPs data through gene-based Bayesian networks. *BMC Bioinformatics* 2009, 10:S7.
105. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 2009, 25:714–721.
106. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ: Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet* 2008, 4:e1000130.
107. Thomas DC. Genome-wide association studies for discrete traits. *Genet Epidemiol* 2009, 33:S8–S12.
108. Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. Machine learning methods for predictive proteomics. *Brief Bioinform* 2008, 9:119–128.

FURTHER READING

Combi C, Keravnou-Papailiou E, Shahar Y. *Temporal Information Systems in Medicine*. Heidelberg, Germany: Springer; 2010.

Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Heidelberg, Germany: Springer; 2001.

Intelligent Data Analysis in bioMedicine and Phamacology (IDAMAP) Workshop Proceedings. Available at: http://www.idamap.org/. (Accessed November, 2010).

Lavrač N, Zupan B. Data mining in medicine. In: *Data Mining and Knowledge Discovery Handbook*. Heidelberg, Germany: Springer; 2005, 1107–1137.