Artificial Intelligence in Medicine (2005) **34**, 113—127

ARTIFICIAL
INTELLIGENCE
IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Predicting breast cancer survivability: a comparison of three data mining methods

Dursun Delen*, Glenn Walker, Amit Kadam

*Department of Management Science and Information Systems, Oklahoma State University,
700 North Greenwood Venue, Tulsa, OK 74106, USA*

Received 13 January 2004; received in revised form 30 June 2004; accepted 15 July 2004

KEYWORDS
Breast cancer survivability;
Data mining;
k-Fold cross-validation;
SEER

Summary

Objective: The prediction of breast cancer survivability has been a challenging research problem for many researchers. Since the early dates of the related research, much advancement has been recorded in several related fields. For instance, thanks to innovative biomedical technologies, better explanatory prognostic factors are being measured and recorded; thanks to low cost computer hardware and software technologies, high volume better quality data is being collected and stored automatically; and finally thanks to better analytical methods, those voluminous data is being processed effectively and efficiently. Therefore, the main objective of this manuscript is to report on a research project where we took advantage of those available technological advancements to develop prediction models for breast cancer survivability.

Methods and material: We used two popular data mining algorithms (artificial neural networks and decision trees) along with a most commonly used statistical method (logistic regression) to develop the prediction models using a large dataset (more than 200,000 cases). We also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes.

Results: The results indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample (this prediction accuracy is better than any reported in the literature), artificial neural networks came out to be the second with 91.2% accuracy and the logistic regression models came out to be the worst of the three with 89.2% accuracy.

Conclusion: The comparative study of multiple prediction models for breast cancer survivability using a large dataset along with a 10-fold cross-validation provided us with an insight into the relative prediction ability of different data mining methods. Using sensitivity analysis on neural network models provided us with the prioritized importance of the prognostic factors used in the study.

* Corresponding author. Tel.: +1 918 594 8283; fax: +1 918 594 8281.
E-mail address: delen@okstate.edu (D. Delen).

0933-3657/$ — see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2004.07.002

1. Introduction

Breast cancer is a major cause of concern in the United States today. At a rate of nearly one in three cancers diagnosed, breast cancer is the most frequently diagnosed cancer in women in the United States. The American Cancer Society projected that 211,300 invasive and 55,700 in situ cases would be diagnosed in 2003 [1]. Furthermore, breast cancer is the second leading cause of death for women in the United States, and is the leading cause of cancer deaths among women ages 40—59 [1,2]. According to The American Cancer Society 39,800 breast cancer related deaths are expected in 2003 [2]. Though predominantly in women, breast cancer can also occur in men. In the United States, of the 40,600 deaths from breast cancer in 2001, 400 were men [3]. Even though in the last couple of decades, with increased emphasis towards cancer related research, new and innovative methods for early detection and treatment have been developed, which helped decrease the cancer related death rates [4—6], cancer in general and breast cancer in specific is still a major cause of concern in the United States.

Although cancer research is generally clinical and/or biological in nature, data driven statistical research is becoming a common complement. In medical domains where data and statistics driven research is successfully applied, new and novel research directions are identified for further clinical and biological research. For instance, Dr. John Kelsoe of the University of California, San Diego, demonstrated through his research study that a flawed gene appeared to promote manic-depression [7]. His data driven study found statistical evidence to tie the gene to the disease, and now researchers are looking for biological and clinical evidence to support his theory.

Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. Survival analyses is a field in medical prognosis that deals with application of various methods to historic data in order to predict the survival of a particular patient suffering from a disease over a particular time period. With the increased use of computers powered with automated tools, storage and retrieval of large volumes of medical data are being collected and are being made available to the medical research community who has been interested in developing prediction models for survivability. As a result, new research avenues such as knowledge discovery in databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers who seek to identify and exploit patterns and relationships among large number of variables, and be able to predict the outcome of a disease using the historical cases stored within datasets [8,9].

It is the combination of the serious effects of breast cancer, the promising results of prior related research, the potential benefits of the research outcomes and the desire to further understand the nature of breast cancer that provided the motivation for this research effort. In this paper, we report on our research project where we developed models that predict the survivability of diagnosed cases for breast cancer. One of the salient features of this research effort is the authenticity and the large volume of data processed in developing these survivability prediction models. We used the SEER cancer incidence database, which is the most comprehensive source of information on cancer incidence and survival in the United States [2]. We used three different types of classification models: artificial neural network (ANN), decision tree, and logistic regression along with a 10-fold cross-validation technique to compare the accuracy of these classification models.

1.1. Breast cancer

Breast cancer is a malignant tumor that develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division [2,3]. It is the most common cancer among women [1]. Although scientists do not know the exact causes of most breast cancer, they do know some of the risk factors that increase the likelihood of a woman developing breast cancer. These factors include such attributes as age, genetic risk and family history [3].

Treatments for breast cancer are separated into two main types, local and systematic. Surgery and radiation are examples of local treatments whereas chemotherapy and hormone therapy are examples of systematic therapies. Usually for the best results, the two types of treatment are used together [2]. Although breast cancer is the second leading cause of cancer death in women, the survival rate is high. With early diagnosis, 97% of women survive for 5 years or more [3,10].

1.2. Knowledge discovery in databases and data mining

The amount of data being collected and stored in databases (both in medical and in other fields) has increased dramatically due to the advancements in software capabilities and hardware tools that enabled the automated data collection (along with the decreasing trend of hardware and software cost) in the last decade. As a result, traditional data analysis techniques have become inadequate for processing such volumes of data, and new techniques have been developed. A major area of development is called KDD. KDD encompasses variety of statistical analysis, pattern recognition and machine learning techniques. In essence, KDD is a formal process whereby the steps of understanding the domain, understanding the data, data preparation, gathering and formulating knowledge from pattern extraction, and ''post-processing of the knowledge'' are employed to exploit the knowledge from large amount of recorded data [8]. The step of gathering and formulating knowledge from data using pattern extraction methods is commonly referred to as data mining [11]. Applications of data mining have already been proven to provide benefits to many areas of medicine, including diagnosis, prognosis and treatment [9].

The remainder of this paper is organized as follows. Section 2 provides the reader with the background information on breast cancer research, survivability analysis, commonly used prognosis factors and previously published relevant literature. In Section 3, we explained in detail the data and data processing, prediction methods and the 10-fold cross-validation algorithm. In Section 4, the prediction results of all three algorithms along with the sensitivity analysis of the ANN models are presented. In Section 5, the discussion of the research is given. The paper concludes with Section 6 where we summarize the research findings, and outline the limitations and further research directions.

2. Background

Due to improved screening techniques and improved treatments, death as a result of breast cancer has fallen in recent years [6,12]. These improvements are the direct result of scientific research [6]. Scientific research takes many forms and each form contributes to the body of knowledge from a different but complementary perspective [13].

2.1. Common breast cancer research methods

Among the most commonly used research methods for breast cancer are laboratory studies, observational studies and clinical trials. Laboratory studies test a hypothesis under controlled conditions. They can yield detailed results, but are limited by the controlled environment. Observational studies examine the characteristics of a population to establish the factors associated with a specific outcome. Observational studies can establish the association of the variables to the outcome, but they do not always establish the cause-and-effect relationship of the association. Clinical trials consist of medical studies involving humans. Clinical trials show the cause-and-effect relationship between the variables and the outcome. Clinical trials have been used extensively in the development of drugs and procedures for the treatment of breast cancer [13]. In this paper, we have used statistical analysis and data mining techniques as the research methods. The purpose of this research is to develop predictive models and discover/explain relationships between certain independent variables and the survivability in the context of breast cancer.

2.2. Medical prognosis and survival analysis

Medical prognosis is a field in medicine that encompasses the science of estimating the complication and recurrence of disease and to predict survival of patient or group of patients [14]. In other words, medical prognosis involves prediction modeling whereby different parameters related to patients' health could be estimated. These estimates can help to design treatment as per the expected outcomes. Survival analysis is a field in medical prognosis that deals with application of various methods to estimate the survival of a particular patient suffering from a disease over a particular time period. ''Survival'' is generally defined as a patient remaining alive for a specified period of time after the diagnosis of disease. Traditionally conventional statistical techniques such as Kaplan-Meier test and Cox-Propositional hazard models [15] were used for modeling survival. These techniques are conditional probability based models that provide us with a probability estimate of survival. With advances in the field of knowledge discovery and data mining a new stream of methods have came into existence. These methods are proved to be more powerful as compared to traditional statistical methods [14].

Many research projects define survival as a period of 10 years or longer. Survival estimates developed using such a definition of survival may not accurately reflect the current state of treatment and the probability of survival [15,16]. Recent improvements in early detection and treatment have increased the expectations of survival [16]. As a result, for the purposes of this research effort, we have defined ''survival'' as any incidence of breast cancer where the person is still living after sixty months (5 years) from the date of diagnosis.

2.3. Prognostic factors in breast cancer

The prognostic factors used in the prediction of survival of breast cancer can be separated into two categories: chronological (based on the amount of time present), or biological (based on the potential behavior of the tumor) [17]. Lymph node status, tumor size and histological grade are among the prognostic factors in use today [18]. Lymph node status is a time-dependent factor and is directly related to prognosis. As the number of nodes involved increases, the prognosis worsens [19]. Tumor size is also a time-dependent factor and is directly related to survival. Survival is inversely related to the size of the tumor. The probability of long-term survival is better with smaller tumors than with larger tumors [17,19]. Histological grade is a biological factor, and is based on a combination of three (3) factors: mitotic rate, nuclear grade and architectural morphological appearance [17,19]. Histological grade is highly correlated with long-term survival. Patients with a grade 1 tumor have a much greater chance of surviving than patients with grade 3 tumors [19].

In addition to the prognostic factors listed above, indices have been developed for the prediction of survival for breast cancer cases. Among the most commonly used ones, the Nottingham prognostic index (NPI) is calculated using a combination of the above listed factors:

$$\mathrm{NPI} = \mathrm{TS} + \mathrm{LS} + \mathrm{HS}$$

where TS is the tumor size (in cm) $\times$ 0.2; LP is the lymph stage that takes the values of 1, 2 or 3; and HS is the histological stage that takes the values of 1, 2 or 3. The NPI index separates the patients into three prognosis groups: good, intermediate and poor. The NPI has been proven to provide relatively good prognostic values, with results similar to those achieved with multivariate analysis [17—19].

Another index used for the prediction of survival of breast cancer is the breast cancer severity score (BCSS). The BCSS is based on tumor diameter, the number of positive lymph nodes, estrogen receptors and progesterone receptors. The BCSS has been shown to achieve greater degree of prediction accuracy than conventional staging systems [20].

2.4. Previous research

Burke et al. [21] compared the 5-year predictive accuracy of various statistical models with the predictive accuracy of ANNs. The statistical models included the pathological TNM staging model [21], principal component analysis, classification and regression trees, and logistic regression. The variation of ANN models included cascade correlation ANN, conjugate gradient descent ANN, probabilistic ANN, and backpropagation ANN. This study used the Patient Care Evaluation dataset collected by the Commission on Cancer of the American College of Surgeons during the period of 1983 with follow up information through October 1992. This data was divided into a training set of 3100 cases, a holdout set of 2069 cases and a test set of 3102 cases. The dataset contained 54 input variables. In this study, the backpropagation ANN was reported to be the most accurate achieving an area under the receiver operating characteristics (ROC) curve value of 0.784. The logistic regression model achieved an area under the ROC curve value of 0.776. The best classification and regression tree achieved an area under the ROC curve value of 0.762. These models were more accurate than the pathological TNM staging model, which achieved an area under the ROC curve value of 0.720. The only model that did worse than the pathological TNM staging model was the principal components analysis, which achieved an area under the ROC curve value of 0.714. In a follow-up study, Burke et al. [22] compared the predictive accuracy of the TNM staging system with the predictive accuracy of an ANN in predicting the 10-year survival of patients with breast carcinoma. In this study, he used only the TNM variables (tumor size, number of positive regional lymph nodes, and distant metastasis) from 1977 to 1982 SEER breast carcinoma dataset. According to the reported results, the TNM staging system achieved a prediction accuracy of 0.692, compared with the ANN, which achieved a prediction accuracy of 0.730.

In one of the recent comparative studies, Lundin et al. [23] used ANN and logistic regression models to predict 5-, 10- and 15-year breast cancer survival. In this study, the authors used a series of 951 breast cancer patients, which were randomly divided into a training set of 651 and a validation set of 300. The authors used tumor size, axillary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumor necrosis and age as input variables. For the 5-year prediction, this study was able to achieve an area under the ROC curve value of 0.909 for the ANN and 0.897 for the logistic regression model [23].

In addition to the above listed research efforts, there are other studies related to using data mining for prediction in medical domains. For instance, Pendharker et al. [24] used several data mining techniques for exploring patterns in breast cancer. In this study, they showed that data mining could be a valuable tool in identifying similarities (patterns) in breast cancer cases, which can be used for diagnosis, prognosis and treatment purposes. In a related study, Abbass [25] used an evolutionary ANN (EANN) for breast cancer diagnosis. The EANN was able to achieve an average test accuracy of 0.981 with a standard deviation of 0.005. Abu-Hanna and Keizer [26] used a combination of logistic regression and classification trees to develop a model for the prediction of mortality for patients entering intensive care. This study showed that the hybrid model provides better prognostic performance than the global logistic regression model. Santos-García et al. [27] used an ANN ensemble to predict cardiorespiratory morbidity. The ANN ensemble in this study performed very well (compared to other related studies) achieving an area under the ROC curve value of 0.98. These studies are just a small representative of the existing large body of research in applying data mining to various medical domains for prediction and pattern recognition purposes.

3. Method

3.1. Data source

In order to perform the research reported in this manuscript, we used the data contained in the SEER Cancer Incidence Public-Use Database for the years 1973—2000. The SEER data files were requested through the Surveillance, Epidemiology, and End Results (SEER) web site (http://www.seer.cancer.gov). The SEER Program is a part of the Surveillance Research Program (SRP) at the National Cancer Institute (NCI) and is responsible for collecting incidence and survival data from the participating nine registries, and disseminating these datasets (along with descriptive information of the data itself) to institutions and laboratories for the purpose of conducting analytical research [28].

The SEER Public Use Data consists of nine text files, each containing data related to cancer for a specific anatomical sites (i.e., breast, colon and rectum, other digestive, female genital, lymphoma and leukemia, male genital, respiratory, urinary, and all other sites). There are 72 variables in each file, and each record in the file relates to a specific incidence of cancer. The data in the file is collected from nine different registries (i.e., geographic areas). These registries contain a population that is representative of the different racial/ethnic groups residing in the United States. The cancer incidence trends and mortality rates in SEER are assumed to be representative of the cancer incidence trends and mortality rates for the total United States [29]. The SEER database is considered to be the ''most comprehensive source of information on cancer incidence and survival in the USA'' [15]. The SEER program emphasizes quality and completeness of the data, and it has been estimated that databases provided by SEER program are 98% complete (or better) for each of these nine registries [15,29].

The SEER database is used often for analytical research purposes in a variety of projects. According to the testimonial [29], SEER sends out nearly 1500 copies of the public use data files a year [29]. ''A search of the National Library of Medicine's database (PUBMED) using only the term SEER identified in excess of 570 publications for the time period of 1978—1999, many of which either include an analysis of SEER data or refer to cancer statistics based on SEER data'' [29].

3.2. Data understanding and data preparation

The data understanding and the data preparation stages are among the most important steps in the data mining applications. Vast majority of time spend on developing data mining applications is accounted for this earlier stage [11]. Almost 80% of the time and effort in this research project was spent on cleaning and preparing the data for predictive modeling. The breast cancer dataset used herein was a single flat file in a fixed-width text format. SEER record description documentation provided the specification of each field needed to render the raw data into an appropriate format. Following this process, the raw data was uploaded into MS Access database, SPSS statistical analysis tool, Statistica data miner, and Clementine data mining toolkit. These software packages were used to explore and manipulate the data. The following section describes the surface complexities and structure of the data.

The SEER Breast cancer data consisted of 433,272 records/cases and 72 variables. These 72 variables provide socio-demographic and cancer specific information concerning an incidence of cancer. Each record represents a particular patient—tumor pair within a registry. Each record is assigned a case number for each patient, and a unique record number for each specific tumor. Therefore, the combination of the SEER registry, case number and record number used to uniquely identify a specific case in the database. Since the goal of this data-mining project is to develop models for predicting the survival of an incidence of breast cancer, a binary dependent variable representing the survival (as defined in the previous section) was created. In calculating the survival variable, along with other related fields, the Survival Time Recode field, which provides the number of years and months of survival after diagnosis, is used. Specifically, the survival variable used herein is encoded as a categorical dependent variable with values 0 and 1 (meaning *did not survive* and *survived*, respectively), based on the net survival in number of months for the properly recorded eligible cases.

The observed survival rate is the proportion of cancer patients who have ''survived'' the specified period of time after diagnosis. However, of those who did not survive, not all died as a result of the given cancer [28]. The observed survival rate is adjusted for expected mortality to yield the relative survival rate, which ''approximates the likelihood that a patient will not die from causes associated specifically with the given cancer'' [4,28]. Therefore, to adjust for expected mortality, and to predict survival based solely on the effects of breast cancer, we have removed the records where the patient did not survive for sixty months after the date of diagnosis, and the ''Cause of Death'' is coded as something other than breast cancer. Also, records for patients who were not followed for a full 60-month period were removed from the dataset [16].

The cancer data model (SCHEMA) used by SEER is the same for all of the datasets. Therefore, some of the variables in the breast cancer dataset do not relate to breast cancer. Additionally, there are variables in the breast cancer dataset that contain redundant information such as variable recodes and overrides. Based on the SEER documentation and our personal inquiries with the SEER personnel such variables were removed from the dataset. For instance, the Extent of Disease and Morphology variables contain aggregated information on different attributes of cancer. For example, the Morphology variable consists of Histology, Behavior and Grade code, each of which provides unique knowledge about the tumor. Similarly, Extent of Disease variable consists of six different characteristics of the tumor. Instead of using these aggregated variables, we chose to use their derivative variables with more detailed information.

Additionally, the dataset was explored for incorrect, inconsistent or missing data. For example, 40% of the records for the Extent of Disease and the AJCC Stage of Cancer contained missing data. An analysis determined that all of the missing data was for records prior to 1988 (these information variables were not being used by then). Since these variables are considered to be important in predicting survivability, rather than deleting the variables, the records containing the missing data were removed from the dataset. Further analysis was performed to check the affect on other variables of deleting these records. This analysis showed that there was no considerable change in the distribution of the other variables. For instance, the histograms presented in Fig. 1 shows that there is no significant change in the distribution of age variable before and after the deletion of these records.

Furthermore, 16% of the records in the Site Specific Surgery variable contained missing data. Prior to 1998, the surgery variable was used to indicate whether surgery was performed on the patient, and the type of surgery performed. In 1998, due to I-C-D coding changes, this information was separated into four different variables based on the site of surgery. Therefore, the Site Specific Surgery variable contained missing values after 1998. A semantic mapping procedure was developed to map these four variables into the single site-

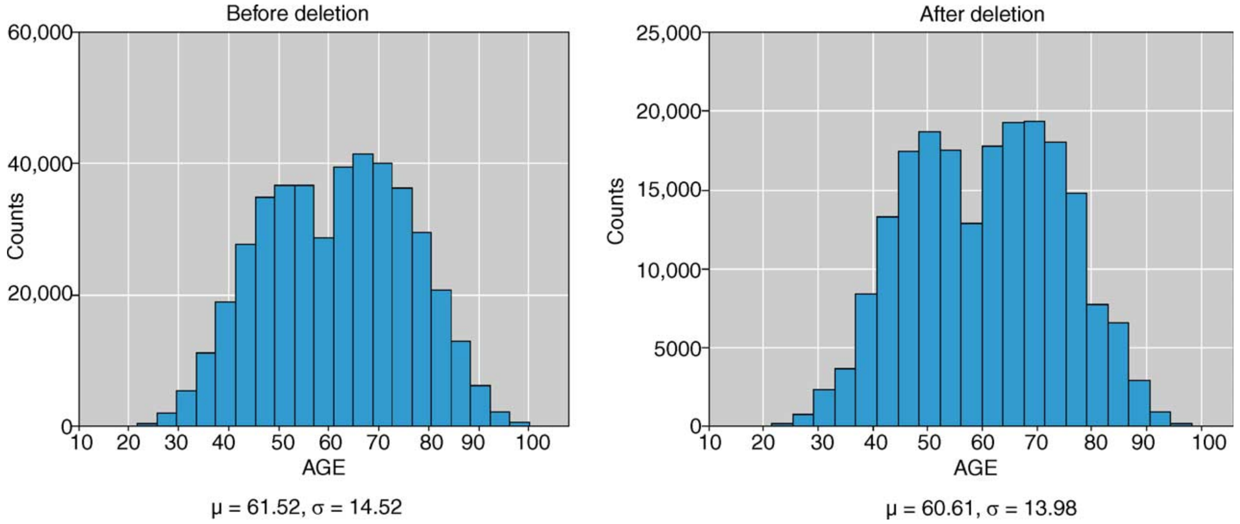

Figure 1 Distribution of the age variable.

specific surgery variable. Some of the granular information was lost during the mapping process. However, the mapping had no change in the distribution of the surgery variable. After the above mapping, there were still 1000 records with missing values, which were also removed from the dataset. Finally, the Tumor Size variable contained a small number of unusually large values (greater than 200 mm), which appeared to be incorrect due to data entry errors, and therefore removed from the dataset.

After using these data cleansing and data preparation strategies, the final dataset, which consisted of 17 variables (16 predictor variables and 1 dependent variable) and 202,932 records, was constructed. Table 1 shows the summaries of predictor variables.

The dependent variable is a binary categorical variable with two categories: 0 and 1, where 0 denoting *did not survive* and 1 denoting *survived*. The distribution of the dependent variable is shown in the Table 2.

3.3. Prediction models

We used three different types of classification models: artificial neural networks, decision trees, and logistic regression. These models were selected for inclusions in this study due to their popularity in the recently published literature as well as their better than average performance in our preliminary comparative studies. What follows is a short description of these classification model types and their specific implementations for this research.

Table 1 Predictor variables for survival modeling

Categorical variable name	Number of unique values		
Race	28		
Marital status	6		
Primary site code	9		
Histology	91		
Behaviour	2		
Grade	5		
Extension of disease	29		
Lymph node involvement	10		
Radiation	10		
Stage of cancer	5		
Site specific surgery code	11		
Continuous variable name	**Mean**	**S.D**	**Range**
Age	60.61	13.98	10—106
Tumor size	19.75	17.65	0—200
Number of positive nodes	1.423	3.659	0—75
Number of nodes	11.307	8.628	0—91
Number of primaries	1.23	0.491	1—8

Table 2 Distribution of dependent variable

Category	Frequency	Percentage
0 (did not survive)	109,659	54.0
1 (survived)	93,273	46.0
Total	202,932	100.0

3.3.1. Artificial neural networks

Artificial neural networks (ANNs) are commonly known as biologically inspired, highly sophisticated analytical techniques, capable of modeling extremely complex non-linear functions. Formally defined, ANNs are analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data [30]. We used a popular ANN architecture called multi-layer perceptron (MLP) with back-propagation (a supervised learning algorithm). The MLP is known to be a powerful function approximator for prediction and classification problems. It is arguably the most commonly used and well-studied ANN architecture. Our experimental runs also proved the notion that for this type of classification problems MLP performs better than other ANN architectures such as radial basis function (RBF), recurrent neural network (RNN), and self-organizing map (SOM). In fact, Hornik et al. [31] empirically showed that given the right size and the structure, MLP is capable of learning arbitrarily complex nonlinear functions to arbitrary accuracy levels. The MLP is essentially the collection of nonlinear neurons (a.k.a. perceptrons) organized and connected to each other in a feedforward multi-layer structure. Fig. 2 illustrates the graphical representation of the MLP architecture used in this study.

3.3.2. Decision trees

Decision trees are powerful classification algorithms that are becoming increasingly more popular with the growth of data mining in the field of information systems. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5 [32,33], and Breiman et al.'s CART [34]. As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms (e.g., information gain, Gini index, and Chi-squared test)

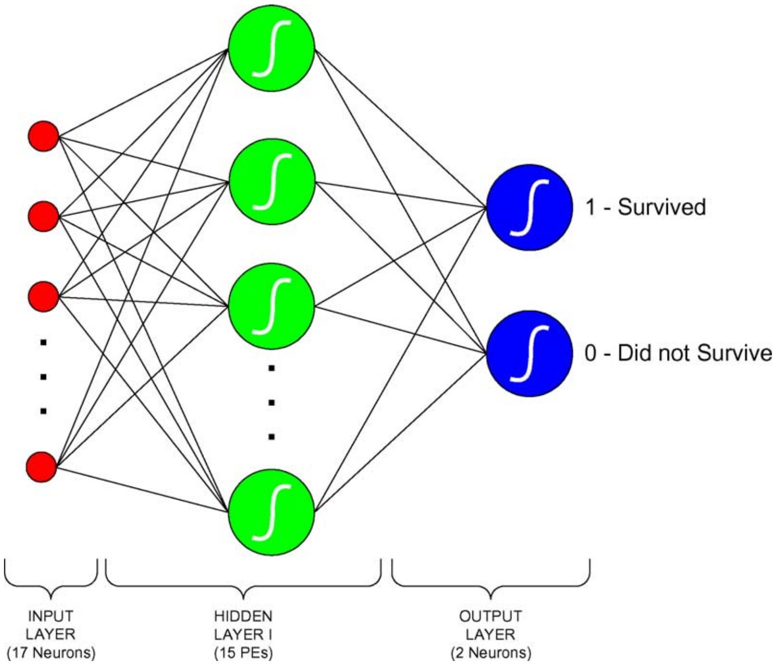

Figure 2 Graphical representation of our MLP ANN model.

to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is constructed. The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity (order) of the resulting two or more subgroups of samples. The most commonly used mathematical algorithm for splitting includes Entropy based information gain (used in ID3, C4.5, C5), Gini index (used in CART), and Chi-squared test (used in CHAID). Based on the favorable prediction results we have obtained from the preliminary runs, in this study we chose to use C5 algorithm as our decision tree method, which is an improved version of C4.5 and ID3 algorithms [33].

3.3.3. Logistic regression

Logistic regression is a generalization of linear regression [35]. It is used primarily for predicting binary or multi-class dependent variables. Because the response variable is discrete, it cannot be modeled directly by linear regression. Therefore, rather than predicting point estimate of the event itself, it builds the model to predict the odds of its occurrence. In a two-class problem, odds greater than 50% would mean that the case is assigned to the class designated as ''1'' and ''0'' otherwise. While logistic regression is a very powerful modeling tool, it assumes that the response variable (the log odds, not the event itself) is linear in the coefficients of the predictor variables. Furthermore, the modeler, based on his or her experience with the data and data analysis, must choose the right inputs and specify their functional relationship to the response variable.

3.4. Measures for performance evaluation

3.4.1. Accuracy, sensitivity and specificity

In this study, we used three performance measures: accuracy (Eq. (1)), sensitivity (Eq. (2)) and specificity (Eq. (3)):

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives, respectively.

3.4.2. *k*-Fold cross-validation

In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, researchers tend to use *k*-fold cross-validation. In *k*-fold cross-validation, also called rotation estimation, the complete dataset

(D) is randomly split into k mutually exclusive subsets (the folds: D_1, D_2, ..., D_k) of approximately equal size. The classification model is trained and tested k times. Each time ($t \in \{1, 2, \ldots, k\}$), it is trained on all but one folds (D_t) and tested on the remaining single fold (D_t). The cross-validation estimate of the overall accuracy is calculates as simply the average of the k individual accuracy measures

$$\mathrm{CVA} = \sum_{i=1}^{k} A_i \tag{4}$$

where CVA stands for cross-validation accuracy, k is the number of folds used, and A is the accuracy measure of each folds.

Since the cross-validation accuracy would depend on the random assignment of the individual cases into k distinct folds, a common practice is to stratify the folds themselves. In stratified k-fold cross-validation, the folds are created in a way that they contain approximately the same proportion of predictor labels as the original dataset. Empirical studies showed that stratified cross-validation tend to generate comparison results with lower bias and lower variance when compared to regular k-fold cross-validation [36].

In this study, to estimate the performance of classifiers a stratified 10-fold cross-validation approach is used. Empirical studies showed that 10 seem to be an optimal number of folds (that optimizes the time it takes to complete the test while minimizing the bias and variance associated with the validation process) [34,36]. In 10-fold cross-validation the entire dataset is divided into 10 mutually exclusive subsets (or folds) with approximately the same class distribution as the original dataset (stratified). Each fold is used once to test the performance of the classifier that is generated from the combined data of the remaining nine folds, leading to 10 independent performance estimates (see Fig. 3). Specifically, we use the following three-step 10-fold cross-validation procedure to estimate the error rate:

1. We randomly divided the dataset (~200K records) into 10 disjoint subsets (folds), with each fold containing approximately the same number of records (~20K records). The sampling is stratified by the class labels to ensure that the subset class proportions are roughly the same as those in the whole dataset.
2. For each subset, a classifier is constructed using the nine of the 10 folds and tested on the tenth one to obtain a cross-validation estimate of its error rate.
3. The 10 cross-validation estimates are then averaged to provide an estimate for the classier accuracy constructed from all the data.

4. Results

4.1. Classification results

In this study, the models were evaluated based on the accuracy measures discussed above (classification accuracy, sensitivity and specificity). The results were achieved using 10 fold cross-validation for each model, and are based on the average results obtained from the test dataset (the 10th fold) for each fold. In comparison to the above studies, we found that the ANN model achieved a classification accuracy of 0.9121 with a sensitivity of 0.9437 and a specificity of 0.8748. The logistic regression model achieved a classification accuracy of 0.8920 with a sensitivity of 0.9017 and a specificity of 0.8786. However, the decision tree (C5) preformed the best of the three models evaluated. The decision tree (C5) achieved a classification accuracy of 0.9362 with a sensitivity of 0.9602 and a specificity of 0.9066. Table 3 shows the complete set of results in a tabular format. For each fold of each model type, the detailed prediction results of the validation datasets are presented in form of confusion matrixes. A confusion matrix is a matrix representation of the classification results. In a two-class prediction problem (such as

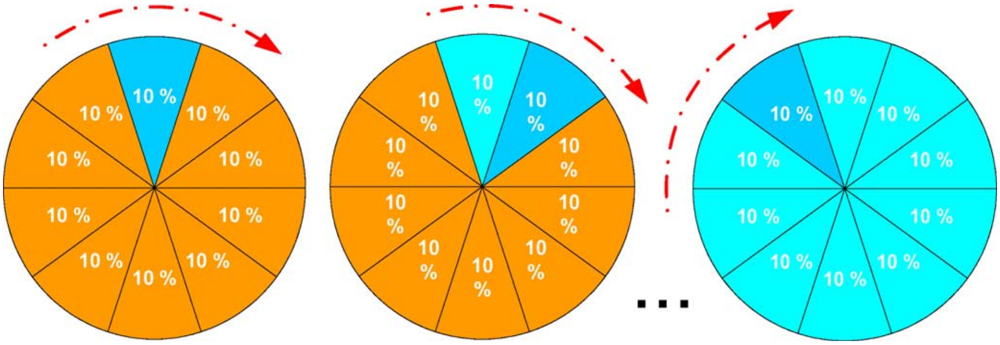

Figure 3 Graphical depiction of the ten-fold cross-validation procedure.

the one in this research) the upper left cell denotes the number of samples classifies as true while they we true (i.e., true positives), and lower right cell denotes the number of samples classified as false while they were actually false (i.e., true false). The other two cells (lower left cell and upper right cell) denote the number of samples misclassified. Specifically, the lower left cell denoting the number of samples classified as false while they actually were true (i.e., false negatives), and the upper right cell denoting the number of samples classified as true while they actually were false (i.e., false positives). Once the confusion matrixes were constructed, the accuracy, sensitivity and specificity of each fold were calculated using the respective formulas presented in the previous section.

4.2. Sensitivity analysis on ANN output

We have used sensitivity analysis to gain some insight into the decision variables used for the classification problem. Sensitivity analysis is a method for extracting the cause and effect relationship between the inputs and outputs of a neural network model. As has been noted by many investigators in the AI field, most of the time ANN may offer better predictive ability, but not much explanatory value. This criticism is generally true, however sensitivity analysis can be performed to generate insight into the problem. Recently, it has become a commonly used method in ANN studies for identifying the degree at which each input channel (independent variables or decision variables) contributes to the identification of each output channel (dependent variables).

The sensitivity analysis provides information about the relative importance of the input variables in predicting the output field(s). In the process of performing sensitivity analysis, the ANN learning is disabled so that the network weights are not affected. The basic idea is that the inputs to the network are perturbed slightly, and the corresponding change in the output is reported as a percentage change in the output [37]. The first input is varied between its mean plus (or minus) a user-defined number of standard deviations, while all other inputs are fixed at their respective means. The network output is computed and recorded as the percent change above and below the mean of that output channel. This process is repeated for each and every input variable. As an outcome of this process, a report (usually a column plot) is generated, which summarizes the variation of each output with respect to the variation in each input. The sensitivity analysis performed for this research project and presented in a graphical format in Fig. 4, lists the input variables by their relative importance (from most important to least important). The value shown for each input variable is a measure of its relative importance, with 0 representing a variable that has no effect on the prediction and 1.0 representing a field that completely dominates the prediction.

Our sensitivity analysis results are based on the 10 different ANN models developed for the 10 data folds. After each of the 10 training, the network weights are frozen (testing stage) and the cause and effect relationship between the independent variables and the dependent variables are investigated as per the above-mentioned procedure. The aggregated results are summarized and presented as a

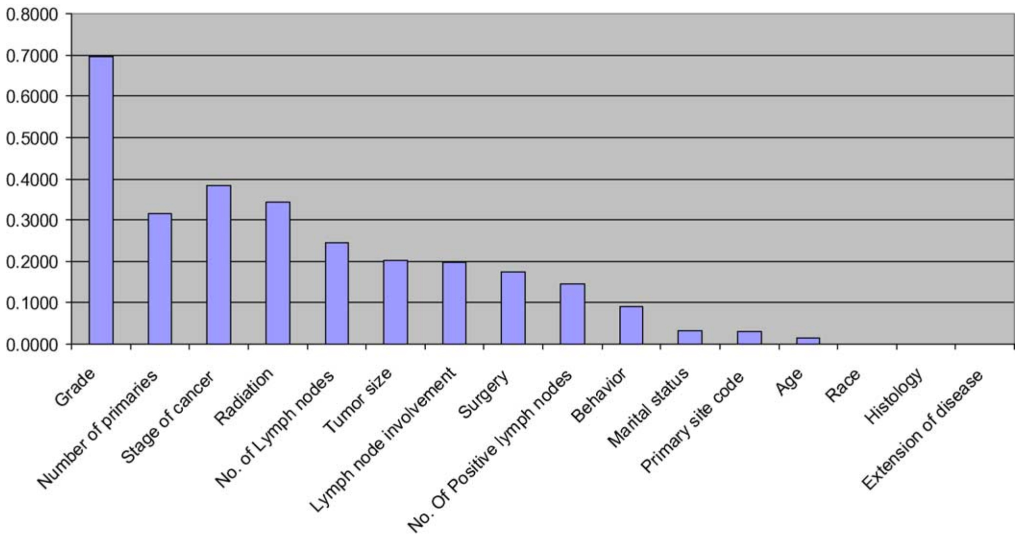

Figure 4 Sensitivity analysis graph.

Table 3 Tabular results for 10-fold cross-validation for all folds and all model types

Fold No	Neural Networks (MLP)					Decision Tree Induction (C5)					Logistic Regression				
	Confusion Matrix		Accuracy	Sensitivity	Specificity	Confusion Matrix		Accuracy	Sensitivity	Specificity	Confusion Matrix		Accuracy	Sensitivity	Specificity
1	7571	844	0.9165	0.9535	0.8719	7828	587	0.9363	0.9586	0.9078	7672	743	0.8912	0.9015	0.8766
	369	5747				338	5778				838	5277			
2	7589	729	0.9271	0.9578	0.8905	7737	581	0.9403	0.9639	0.9113	7543	773	0.8906	0.9018	0.8756
	334	5926				290	5970				821	5439			
3	7567	768	0.9214	0.9537	0.8818	7741	594	0.9356	0.9584	0.9065	7602	732	0.8915	0.9011	0.8779
	367	5730				336	5761				834	5261			
4	7508	796	0.9169	0.9480	0.8798	7703	601	0.9356	0.9582	0.9076	7607	696	0.8951	0.9017	0.8860
	412	5824				336	5900				829	5407			
5	7609	809	0.9239	0.9549	0.8903	7789	629	0.9348	0.9607	0.9021	7659	757	0.8908	0.9022	0.8747
	359	6565				319	5796				830	5284			
6	7390	908	0.8914	0.9179	0.8581	7694	604	0.9363	0.9604	0.9062	7554	743	0.8900	0.8992	0.8771
	661	5491				317	5835				847	5305			
7	7298	751	0.9081	0.9290	0.8823	7464	585	0.9374	0.9605	0.9095	7333	716	0.8949	0.9037	0.8831
	558	5631				307	5882				781	5408			
8	7069	977	0.9024	0.9442	0.8565	7436	610	0.9353	0.9594	0.9068	7269	773	0.8894	0.9001	0.8756
	418	5832				315	5935				807	5443			
9	7290	958	0.9040	0.9454	0.8559	7621	627	0.9360	0.9631	0.9027	7501	747	0.8923	0.9036	0.8767
	421	5691				292	5820				800	5310			
10	7475	764	0.9098	0.9330	0.8809	7625	614	0.9349	0.9591	0.9052	7518	716	0.8938	0.9022	0.8824
	537	5651				325	5863				815	5372			
Mean			**0.9121**	**0.9437**	**0.8748**			**0.9362**	**0.9602**	**0.9066**			**0.8920**	**0.9017**	**0.8786**
St. Dev.			**0.0111**	**0.0131**	**0.0135**			**0.0016**	**0.0019**	**0.0028**			**0.0020**	**0.0014**	**0.0038**

Confusion matrix shows the classification of the cases in the test dataset. In confusion matrix, the columns denote the actual cases and the rows denote the predicted. Accuracy = (TP + TN)/(TP + FP + TN + FN); sensitivity = TP/(TP + FN); specificity = TN/(TN + FP).

column plot in Fig. 4. The *x*-axis represents the input variables and the *y*-axis represents the percent change on the output variables, while the input variables (one at a time) are perturbed gradually around their mean with the magnitude of ±1 standard deviation.

As shown in the Fig. 4, the variable, Grade, was by far the most important variable in the prediction of survival. The information contained in this variable is concerned with the degree of differentiation of the cancer. The four codes represent the amount of differentiation, from Grade I, well differentiated, to Grade IV, undifferentiated. This result is consistent with some of the earlier studies where they have also found that histological grade is one of the most important prognostic factor in the prediction of breast cancer survival [17,19].

The second most important variable in the prediction of survival is Stage of Cancer. This variable is concerned with the degree to which the cancer has spread, from In situ (noninvasive) to Distant (spread other areas). This is a chronological factor based on the amount of time the disease is present. The longer the disease is present the more chance that it will spread to other areas, and the worse the prognosis.

5. Discussion

As shown in the results above, advanced data mining methods can be used to develop models that possess a high degree of predictive accuracy. However, there are several issues involved with the data collection, data mining and the predictive models that warrant for further discussion.

5.1. Data collection

One of the key components of predictive accuracy is the amount and quality of the data [22]. However, the data gathered in medicine is generally collected as a result of patient-care activity to benefit the individual patient, and research is only a secondary consideration. As a result, medical databases contain many features that create problems for the data mining tools and techniques. Medical databases may consist of a large volume of heterogeneous data, including heterogeneous data fields. The heterogeneity of the data complicates the use data mining tools and techniques. Additionally, as with any large database, medical databases contain missing values that must be dealt with prior to the use of the data mining tools. Further, as a result of the method of collection, medical databases may contain data that is redundant, incomplete, imprecise or inconsistent, which can affect the use and results of the data mining tools. Also, the collection method can introduce noise into the data, and can affect the results of the data mining tools. In addition to the collection problems, medical databases have the unique problem of incorporating medical concepts into an understandable form. All of the above may create problems for data mining, and as a result, may require more data reduction and data preparation than data derived from other sources [8,11].

Even with the inherent problems associated with medical databases, the use of medical data for research has many benefits. The research can provide useful information for diagnosis, treatment and prognosis [9]. As mentioned above, the results of data mining are directly affected by the quantity and quality of the data [22]. By improving the collection of the data, medical data mining can yield even greater results and benefits. By making the collection process a primary focus, the methods of collecting medical data can be formalized and standardized. Thus, reducing the problem of missing, redundant or inconsistent values data [11].

In addition to the above listed technical problems, the use of medical data in data mining involves the critical issues of privacy, security and confidentiality. The privacy of the individual should be respected concerning any medical data collected. Patient identification should be kept confidential and secure. Additionally, medical data governed under the rules of Common Rule (45 CFR 46) and HIPPA, and is subject to the penalties thereunder if the proper procedures are not followed [38]. The rules under these laws can generally be satisfied by the use of anonymous data, anonymized data and de-identified data. Anonymous data is data that is collected without any patient identification. Anonymized data is data that has the patient identification information permanently removed subsequent to collection. De-identified data is data in which the patient identification information is encoded or encrypted subsequent to collection. With de-identified data, the patient identification information can be retrieved with the appropriate approval. Identified data should never be used without the patient's prior consent [11].

5.2. Data mining

Data mining has been criticized by some for not following all of the requirements of classical statistics [11]. For example, most data mining tools use training and testing sets drawn from the same sample. Under classical statistics, it can be argued that the testing set used in this instance is not truly independent, and therefore, the results are biased

[9]. Despite these criticisms, data mining can be an important tool in the medical field. By identifying patterns within the large sums of data, data mining can, and should, be used to gain more insight into the diseases, generate knowledge that can potentially fuel lead to further research in many areas of medicine [9]. The high degree of predictive accuracy of the models included herein is just one example of the value of data mining in the medical field. However, for data mining to be accepted in the medical field, researchers must follow established procedures. Sophisticated models such as the six-step DMKD process model or the CRISP-DM model must be utilized from the problem definition to the end result [39,40].

5.3. Predictive models

As shown herein, based on certain predictive attributes, models can be developed that accurately predict the outcome of an incidence of cancer. These predictive models can be valuable tools in medicine. They can be used to assist in determining prognosis, developing successful treatment, or the avoidance of treatment [22]. However, there are areas of concern in the development of predictive models: (1) the model should include all clinically relevant data, (2) the model should be tested on an independent sample, and (3) the model must make sense to the medical personnel who is supposed to make use of it. It has been shown that not all predictive models constructed using data mining techniques satisfy all of these requirements [9].

5.4. Limitations of data mining

While data mining can provide useful information and support to the medical staff by identifying patterns that may not be readily apparent, there are limitations to what data mining can do. Not all patterns found via data mining are ''interesting''. For a pattern to be interesting, it should be logical and actionable. Therefore, data mining requires human intervention to exploit the extracted knowledge. For example, data mining can provide assistance in making the diagnosis or prescribing the treatment, but it still cannot replace the physician's intuition and interpretive skills [9].

6. Conclusion

In this paper, we report on a research effort where we developed several prediction models for breast cancer survivability. Specifically, we used three popular data mining methods: two from machine learning (ANN, decision trees) and one from statistics (logistic regression). We acquired a quite large dataset (433,272 cases with 72 prognosis factors) from the SEER program and after going though a long process of data cleansing and transformation used it to develop the prediction models. In this research, we defined survival as any incidence of breast cancer where person is still alive after 5 years (60 months) from the date of diagnosis. We used a binary categorical survival variable, which was calculated from the variables in the raw dataset, to represent the survivability where survival is represented with a value of ''1'' and non-survival is represented with ''0''. In order to measure the unbiased prediction accuracy of the three methods, we used a 10-fold cross-validation procedure. That is, we divided the dataset into 10 mutually exclusive partitions (a.k.a. folds) using a stratified sampling technique. Then, we used 9 of 10 folds for training and the 10th for the testing. We repeated this process for 10 times so that each and every data point would be used as part of the training and testing datasets. The accuracy measure for the model is calculated by averaging the 10 models performance numbers. We repeated this process for each of the three prediction models. This provided us with a less biased prediction performance measures to compare the tree models. The aggregated results indicated that the decision tree induction method (C5) performed the best with a classification accuracy of 93.6% which is better than any reported in the published literature, the ANN model (with multi layered perceptron architecture) came out to be second best with a classification accuracy of 91.2%, and the logistic regression model came out to be the worst with a classification accuracy of 89.2%. In addition to the prediction model, we also conducted sensitivity analysis on ANN model in order to gain insight into the relative contribution of the independent variables to predict survivability. The sensitivity results indicated that the prognosis factor ''Grade'' is by far the most important predictor, which is consistent with the previous research, followed by ''Stage of Cancer'', ''Radiation'', and ''Number of Primaries''. Why these prognostic factors are more important predictors than the other is a question that can only be answered by medical professional and further clinical studies.

Although data mining methods are capable of extracting patterns and relationships hidden deep into large medical datasets, without the cooperation and feedback from the medical professional, their results are useless. The patterns found via data mining methods should be evaluated by medical professionals who have years of experience in the problem domain to decide whether they are logical, actionable and novel to fuel new biological and clinical research directions. In short, data mining is not aiming to replace medical professionals and researchers, but to complement their invaluable efforts to save more human lives.

Some of the research extensions to compensate the limitations of the presented effort can be listed as follows. First, in studying breast cancer survivability, we have not considered the potential correlation to other cancer types. It would be an interesting study to investigate if a person having, say, a skin cancer has a worse survivability rating. This can be done by including all possible cancer types and their prognostic factors to investigate the correlations, commonalities and differences among them. Second, new and more promising methods such as support vector machines and rough sets can be used to see if the prediction accuracy can be further improved. Another viable option to improve the prediction accuracy would be a hybrid intelligent system where the prediction results of data mining methods are augmented with expert opinions, which are captured and embedded into an expert system.

Our ongoing research efforts are geared toward incorporating new capabilities into our prediction system along the lines of extended research directions listed above. We also want to make the system available to the general public via a website to be used as a web-based decision support system (DSS). In this web-based DSS, we would like to incorporate most accurate prediction models and their hybrids for all possible cancer types.

References

[1] Calle J. Breast cancer facts and figures 2003—2004. American Cancer Society 2004. p. 1—27 (http://www.cancer.org/).

[2] Breast cancer Q&A/facts and statistics (http://www.komen.org/bci/bhealth/QA/q_and_a.asp).

[3] Jerez-Aragonés JM, Gomez-Ruiz JA, Ramos-Jimenez G, Munoz-Perez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med 2003;27:45—63.

[4] Edwards BK, Howe HL, Ries Lynn AG, Thun MJ, Rosenberg HM, Yancik R, et al. Annual report to the nation on the status of cancer, 1973—1999, featuring implications of age and aging on US cancer burden. Cancer 2002;94:2766—92.

[5] Ross E. Breast cancer drug may prolong life. Associated Press; 2003 (http://aolsvc.news.aol.com/news/article.ade?Fid=20030924090209990002&mpc=news%2e10%2e7).

[6] Warren J. Cancer death rates falling, but slowly. WebMD medical news; 2003 (http://aolsvc.health.webmd.aol.-com/content/Artcile/73/82013.htm).

[7] Ritter M. Gene tied to manic-depression. Newspaper article in Tulsa World June 16, 2003: D8.

[8] Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16:3—23.

[9] Richards G, Rayward-Smith VJ, Sonksen PH, Carey S, Weng C. Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med 2001;22:215—31.

[10] O'Malley CD, Le GM, Glaser SL, Shema SJ, West DW. Socioeconomic status and breast carcinoma survival in four racial/ethnic groups: a population-based study. Am Cancer Soc 2003;1303—11.

[11] Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26:1—24.

[12] Progress shown in death rates from four leading cancers (http://cancer.gov/newscenter/pressreleases/2003 Report Release).

[13] The ABCs of breast cancer—types of research studies (http://www.komen.org/bci/abs/chap_01.asp).

[14] Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. J Biomed Inform 2001;34:428—39.

[15] Cox DR. Analysis of survival data. London: Chapman & Hall; 1984.

[16] Brenner H, Gefeller O, Hakulinen T. A computer program for period analysis of cancer patient survival. Eur J Cancer 2002;38(5):690—5.

[17] Bundred NJ. Prognostic and predictive factors in breast cancer. Cancer Treatment Rev 2001;27:137—42.

[18] D'Eredita' G, Giardina C, Martellotta M, Natale T, Ferrarese F. Prognostic factors in breast cancer: the predictive value of the Nottingham Prognostic Index in patients with a long-term follow-up that were treated in a single institution. Eur J Cancer 2001;37:591—6.

[19] Rampaul RS, Pinder SE, Elston CW, Ellis IO. Prognostic and predictive factors in primary breast cancer and their role in patient management: the Nottingham breast team. Eur J Surg Oncol 2001;27:229—38.

[20] Jimenez-Lee R, Ham B, Vetto J, Pommier R. Breast cancer severity score is an innovative system for prognosis. Am J Surg 2003;186:404—8.

[21] Burke HB, Rosen D, Goodman P. Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. In: Tesauro G, Touretzky D, Leen T, editors. Advances in neural information processing systems, vol. 7. Cambridge, MA: MIT Press; 1995 p. 1063—7.

[22] Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997;79:857—62.

[23] Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology 1999;57:281—6.

[24] Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M. Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Syst Applic 1999;17:223—32.

[25] Abbass HA. An evolutionary artificial neural networks approach for breast cancer diagnosis. Artif Intell Med 2002;25:265—81.

[26] Abu-Hanna A, de Keizer N. Integrating classification trees with local logistic regression in intensive care prognosis. Artif Intell Med 2003;29:5—23.

[27] Santos-Garcia G, Varela G, Novoa N, Jimenez MF. Prediction of postoperative morbidity after lung resection using an artificial neural network ensemble. Artif Intell Med 2004; 30:61—9.

[28] SEER Cancer Statistics Review. Surveillance, Epidemiology, and End Results (SEER) program (www.seer.cancer.gov) public-use data (1973—2000). National Cancer Institute, Surveillance Research Program, Cancer Statistics Branch, released April 2003. Based on the November 2002 submission. Diagnosis period 1973—2000, Registries 1—9.

[29] Hankey BF. The surveillance, epidemiology, and end results program: a national resource.. Cancer Epidemiol Biomarkers Prev 1999;8:1117—21.

[30] Haykin S. Neural networks: a comprehensive foundation. New Jersey: Prentice Hall; 1998.

[31] Hornik K, Stinchcombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward network. Neural Netw 1990;3:359—66.

[32] Quinlan J. Induction of decision trees. Mach Learn 1986;1:81—106.

[33] Quinlan J. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.

[34] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.

[35] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer-Verlag; 2001.

[36] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Wermter S, Riloff E, Scheler G, editors. The Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)San Francisco, CA: Morgan Kaufman; 1995. p. 1137—45.

[37] Principe JC, Euliano NR, Lefebvre WC. Neural and adaptive systems. New York, NY: Wiley; 2000.

[38] Berman JJ. Confidentiality issues for medical data miners. Artif Intell Med 2002;26:25—36.

[39] Bostwick DG, Burke HB. Prediction of individual patient outcome in cancer: Cancer supplement 2001;91:1643—6.

[40] Shearer C. The CRISP-DM model: the new blueprint for data mining. J Data Warehousing 2000;5:13—22.