J Med Syst (2007) 31:263–273
DOI 10.1007/s10916-007-9064-1

Predicting Metastasis in Breast Cancer: Comparing a Decision Tree with Domain Experts

Amir R. Razavi · Hans Gill · Hans Åhlfeldt · Nosrat Shahsavar

Received: 18 January 2007 / Accepted: 12 March 2007 / Published online: 19 April 2007

Abstract Breast malignancy is the second most common cause of cancer death among women in Western countries. Identifying high-risk patients is vital in order to provide them with specialized treatment. In some situations, such as when access to experienced oncologists is not possible, decision support methods can be helpful in predicting the recurrence of cancer. Three thousand six hundred ninety-nine breast cancer patients admitted in south-east Sweden from 1986 to 1995 were studied. A decision tree was trained with all patients except for 100 cases and tested with those 100 cases. Two domain experts were asked for their opinions about the probability of recurrence of a certain outcome for these 100 patients. ROC curves, area under the ROC curves, and calibration for predictions were computed and compared. After comparing the predictions from a model built by data mining with predictions made by two domain experts, no significant differences were noted. In situations where experienced oncologists are not available, predictive models created with data mining techniques can be used to support physicians in decision making with acceptable accuracy.

Keywords Data mining · Decision tree induction (DTI) · Breast cancer · Classification · Prediction · Domain expert · Decision support

A. R. Razavi (✉) · H. Gill · H. Åhlfeldt · N. Shahsavar
Department of Biomedical Engineering, Division of Medical Informatics, Linköping University, University Hospital, S-58185 Linköping, Sweden
e-mail: amirreza.razavi@imt.liu.se

N. Shahsavar
Regional Oncology Centre, University Hospital, Linköping, Sweden

Introduction

In recent times, information about cancer patients has increasingly been stored in large data sources. These databases are often built for studying changes in the incidence and behavior of cancers. Among cancers, breast malignancy is the most common cancer and the second highest cause of cancer death among women. It is a major health problem and represents a significant worry for many women and their physicians [1]. When this disease is diagnosed and determined to be localized without evidence of metastasis, it is still critical to identify patients who are at a substantial risk of experiencing cancer recurrence, especially distant metastasis.

Assessment of an individual woman's actual risk of recurrence of breast cancer is difficult. Among known risk factors are abnormal values for some morphological and pathological tumor specifications and biological tumor markers. Identification of risk factors that are associated with the recurrence of cancer makes it possible to tailor the most appropriate treatment for the individual. Patients assigned to high-risk groups get more intensive treatment and more frequent follow-ups. This assessment constitutes a very critical decision and the role of domain experts is important. However, the availability of these experienced oncologists is limited. The challenge is how to support less experienced oncologists when they need expert knowledge in order to care for their patients [2]. It would be of considerable benefit if knowledge about what to do and how to do it could be extracted from data sources. Electronic medical records and registers are data sources that can provide knowledge about how different patients have been diagnosed and treated. Knowledge discovery in databases (KDD) [3] can be used to create a representation for this knowledge. Data mining is a part of KDD that is

designed to look through data in search of patterns or relationships between variables, and then to validate the findings by applying the identified models to new data [4]. Decision tree induction (DTI) is a data mining method in the form of a tree structure, and it is used to classify cases in a dataset [5]. The resulting tree is a representation that can be verified by humans and can be used by either humans or computer programs [6]. DTI has been used in different areas of medicine including oncology [7, 8] and respiratory diseases [9]. Decision trees can be easily visualized and formulated into if–then rules. DTI has been compared in several studies with other techniques such as Artificial Neural Networks [10–12], and it has been shown that the accuracy of the techniques is similar. However, DTI produces an understandable model that explains the reasoning of the method, in contrast to the "black box" approach in ANN. In building predictive models, there is a risk of overfitting the training data, which leads to poor accuracy in future predictions. The solution here is pruning of the tree, and the most common method is post-pruning. In this method, the tree grows from a dataset until all possible leaf nodes have been reached, and then particular subtrees are removed. Post-pruning creates smaller and more accurate trees [13].

A prerequisite for successful knowledge discovery is the availability of quality data. A dataset that is representative of a population and contains all important variables affecting a specific event is needed.

By analyzing the data stored in a regional cancer register by a data mining method, we try to find rules for detecting high risk breast cancer patients. These patients may develop distant metastasis—invasion of other organs by malignant cells—and need special attention. A predictive model resulting from DTI can support less experienced oncologists. However, for any such use of a decision support model the model needs validity, transparency and an acceptable degree of accuracy.

In this study, we first analyzed a regional cancer register by DTI in order to develop a predictive model for predicting the occurrence of distant metastasis in breast cancer patients. Thereafter, the accuracy of the predictions for the 100 randomly selected cases were compared with predictions made by two domain experts to see if there was any significant difference between these different prediction sources.

Background

Recurrence of breast cancer and distant metastasis

Recurrence of breast cancer often occurs in the first 3–5 years after diagnosis. It can come back as a local/regional recurrence or as a distant metastasis. The most common sites of recurrence include the lymph nodes, bones, liver, and lungs [14].

In loco-regional recurrences malignant cells remain in the original site in a preserved breast, in the chest wall or in regional lymph nodes, and over time grow back. This may be because of failure of the primary treatment or return of the tumor cells. Distant metastasis is the fatal type of recurrence. When out of the breast, cancer usually spreads first to the axillary lymph nodes. In 25% of distant recurrences, breast cancer spreads from the lymph nodes to bone. Other sites to which breast cancer may spread include the bone marrow, lungs, liver, brain, or other organs. Unfortunately, the chance of recovery after this recurrence is low, and death due to breast cancer is very probable following the occurrence of distant metastasis.

Predictors for high risk breast cancer

Variables that are predictors for the recurrence of breast cancer include some of the following. The S-phase fraction is a measure of the percentage of cells in cancer cells that are in the phase of the cell cycle during which DNA is synthesized. Some studies have shown that higher fractions are generally associated with poorer overall survival [15]. Examining lymph node involvement is essential when assessing the probability of breast cancer recurrence. The overall survival of patients has been shown to decrease as nodal involvement increases [16]. Periglandular growth of the malignant tumor [17], size of the tumor [18], and receptors for estrogen and progesterone [19] have also been found to be important predictors for recurrence of this disease. Some studies indicate that age plays a role [20], and very young patients have a poorer prognosis. Age is also important for loco-regional recurrence. Some other predictors might also be important, but they are not usually recorded in the breast cancer registers. The fact that the above mentioned variables are important predictors of breast cancer recurrence was confirmed in our previous study [21].

Cancer registers

Six regional cancer registers perform cancer registration in Sweden and one of these serves south-east Sweden, comprising the counties of Kalmar, Jönköping and Östrgötland with a population of about one million. The breast cancer register for the south-east region of Sweden has the following properties that make it a useful dataset. It covers more than 95% of breast cancer patients in the region [22]. Its quality is assessed regularly and probable mistakes are checked by directly contacting physicians or pathologists. In this region there are registers that are used

to provide data regarding additional risk factors and to give a better estimation of the recurrence of breast cancer, i.e. the tumor marker register and the death register. The tumor marker register includes values for some newer laboratory measurements for breast cancer such as receptors for estrogen and progesterone and S-phase fraction. The death register contains information about cause of death and can be linked to other registers by using the unique personal number.

Since there is information about tumor specification, treatment and follow-up in these registers, it is possible to find patterns describing the recurrence of breast cancer. Patients get their treatment based on the knowledge of clinicians and on protocols. The prognosis of the disease depends on the combination of each patient's disease specification and treatment. By analyzing these data, hidden knowledge may be discovered. By representing this knowledge and making a predictive model, it is possible to predict the outcome of new patients.

Knowledge discovery in databases (KDD)

KDD is the evolving field that provides automated analysis solutions for extraction of implicit, unknown knowledge and potentially useful information from data [23]. Data mining is the pattern extraction stage of the KDD process [24]. The extracted patterns may be used for diagnosis, screening, prognosis, monitoring, therapy support or overall patient management, and these methods have been successfully used in predicting survival in breast cancer [7].

To uncover and formulate the hidden knowledge, a number of steps should be considered [3]. After understanding the domain and finding suitable sources of data, the next step is preparing these data. Cleaning data from noise and outliers and handling missing values, and then finding the right subset of data, prepares them for successful data mining. Afterwards, in the data mining step, the processed data are used to create a model that can be employed for predicting recurrence in newly diagnosed patients. Data mining has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [25] and "the science of extracting useful information from large data sets or databases" [26]. One important goal of data mining is prediction, which is the most common type of data mining with the most direct practical applications.

Materials and methods

The first phase of this study consisted of preparing data sources, linking and matching datasets, pre-processing data, data mining and building a predictive model. In the next phase, prediction accuracies for the occurrence of distant metastasis or death because of breast cancer made by human experts were compared with prediction accuracies from decision tree induction. Afterwards, ROC curve analysis was used for validation. Figures 1 and 2 schematically describe methods that were applied in this study.

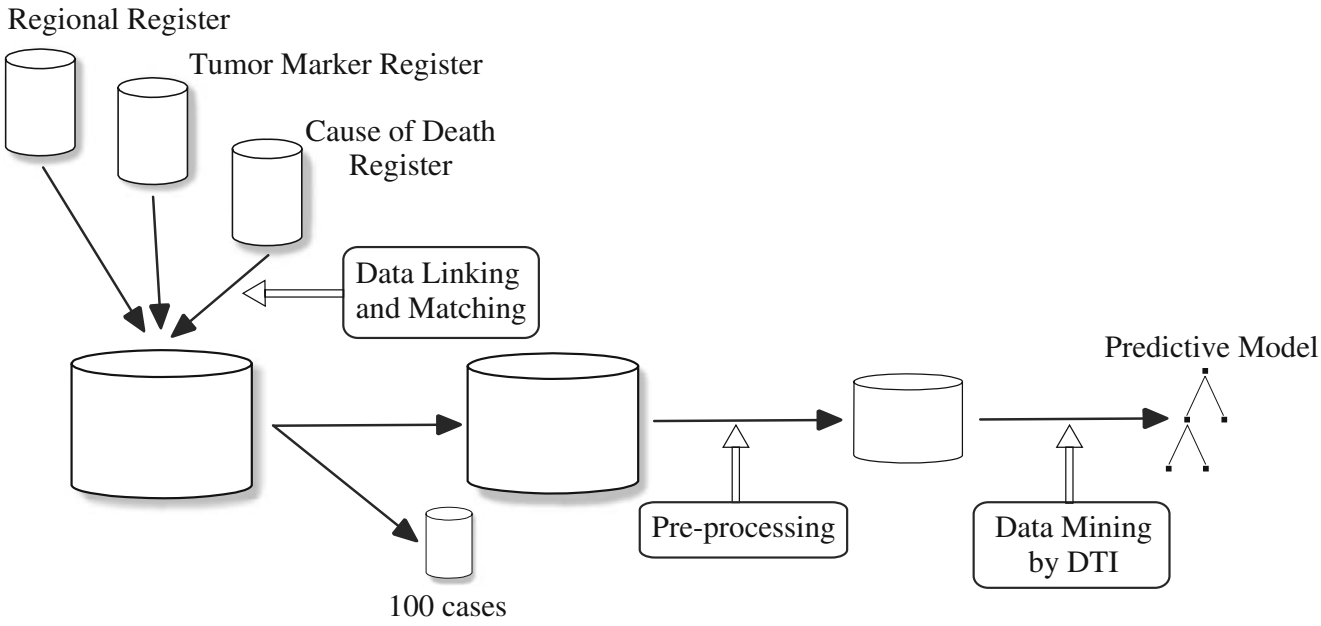
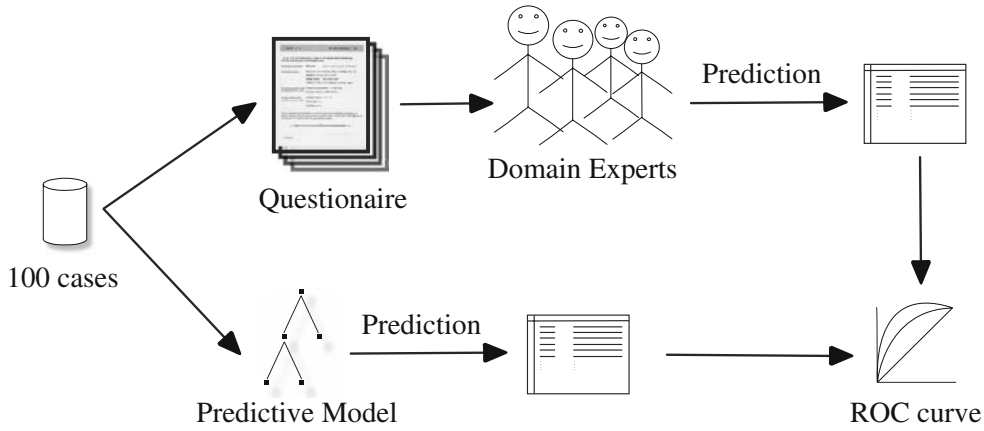


Fig. 1 Steps leading to building a predictive model

Fig. 2 Comparison study



Data preparation

In order to build the best possible predictive model, variables from different sources were collected. The main dataset was the regional breast cancer register for south-east Sweden.

Data were collected from female patients, mean age 61.9 years, with the diagnosis of malignant breast cancer. The earliest patient was diagnosed in 1986 and the last one in 1995. Because the outcome for this study was distant metastasis occurring up to 4 years after diagnosis, patients who were followed up for less than this period were omitted from the study. There were 664 (18%) patients with this type of recurrence in the dataset.

If patients developed symptoms following treatment they were referred to the hospital, but otherwise follow-up visits occurred at fixed time intervals for all patients.

The methodology for preparing the data for the main analysis was the same as in our previous studies [21, 27]. This step started with selecting appropriate variables, cleaning the raw data and removing outliers by running a set of logical rules. Some examples of these outliers were negative values for the time between cancer diagnosis and recurrence and very high values for patient age at the time of diagnosis. The register was searched for multiple entries (unknown to the authors in the previous study) and repeated cases were omitted. After eliminating repeated cases, the number of cases decreased to 3,699.

Subsequently, missing values for continuous variables were substituted using multiple imputation (MI) [28]. In this technique, missing values are replaced by final values resulting from repeated imputation, analysis and pooling steps. The variation among different imputations shows the uncertainty with which the missing values can be predicted from the observed ones. The result is several complete datasets. Thereafter, each of the simulated complete data sets is analyzed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing data uncertainty. Handling missing values by MI was done using the standalone version of NORM software written by Schafer [29]. The software starts by fitting models to incomplete data using the expectation maximization (EM) algorithm [30]. This algorithm is a parameter estimation method that falls within the general framework of maximum likelihood estimation and is an iterative optimization algorithm. Following

Table 1 Characteristics of study variables

Variable	Valid	Missing	Mean		SD	
			BHMVs	AHMVs	BHMVs	AHMVs
Age	3,695	4	61.85	–	13.10	–
Tumor size[a]	3,608	91	22.47	22.48	15.50	15.31
LN involvement[a]	3,606	93	1.89	1.89	5.48	5.41
LN involvement (N0)[b]	3,634	65	–	–	–	–
Perigland growth[a]	3,699	0	–	–	–	–
Estrogen receptor	3,641	58	2.66	2.66	4.37	4.34
Progesterone receptor	3,636	63	2.77	2.78	5.55	5.50
S-phase fraction	2,880	819	8.67	8.79	6.06	5.48

LN: Lymph Node; *BHMVs*: Before Handling Missing Values; *AHMVs*: After Handling Missing Values; *SD*: Standard Deviation
[a] from pathology report, [b] N0: Not palpable LN.

convergence of the EM algorithm, a data augmentation (DA) procedure was implemented. DA is an iterative process that utilizes the observed data to provide estimates of both the missing data and distributional parameters. The result of handling missing values is shown in Table 1. In this table, some statistics before and after handling missing values are presented.

After handling missing values, some variables were dichotomized and were transformed to binary variables. The rules for binarization of these variables are shown in Table 2. These rules are based on positive/negative or normal/abnormal values for those variables. An appropriate set of variables was then selected by using canonical correlation analysis (CCA) as a dimensionality reduction technique [21, 27]. In order to reduce the risk for bias in the study, 100 cases were randomly separated from the dataset for validation and the remaining 3,599 cases were analyzed with CCA and then used for data mining and model building. After analyzing the data with CCA, a clinically relevant outcome, i.e. distant metastasis or death because of breast cancer within 4 years, was associated with the predictors.

Data mining and predictive model building

Several data mining techniques have been examined in breast cancer studies. Predicting breast cancer survival using different data mining methods [7], and comparing the predictive accuracy of a staging system with artificial neural networks [31] are some examples. In comparison with different data mining methods, decision tree induction (DTI) performs well and the resulting predictive model is understandable. The algorithm uses information gain as a heuristic for selecting the variable that will best separate the cases into each outcome [32]. Good interpretability of acquired knowledge and fast execution make decision trees one of the most frequently used data mining techniques [33].

In this study, a predictive model was made by applying DTI to the prepared data. DTI was carried out using the J48 algorithm in WEKA [34]. WEKA is a set of machine learning algorithms for data mining tasks and the algorithms can either be applied directly to a dataset or called from other programs. As in our previous study, we used WEKA for mining (applying the J48 algorithm) to breast cancer register data [27]. The application contains tools for data preparation, classification, clustering and visualization. In WEKA, the J48 algorithm is the equivalent of the C4.5 algorithm written by Quinlan [5]. Post-pruning based on a 10-fold cross validation was also done to trim the resulting tree [13].

For estimating the generalization error of the predictive model, the 10-fold cross validation technique was used [35]. The data (excluding the 100 cases) were divided into ten subsets of about the same size. Then the tree was trained ten times, each time leaving out one of the subsets from training. The omitted subset was used for testing and computing the error. These error estimates were used to adjust the extent of pruning the decision tree.

Validation

One hundred cases were selected by stratified random sampling after the data sources were linked and cases were matched. The ratio of outcome positive cases was the same between the whole population and the 100-case sample. This dataset was used for validating the model by comparing the predictions with those of domain experts.

The predictive model acquired from DTI was used to predict the occurrence of the outcome, and its probability for each case was recorded. The same 100 cases were given to two domain experts for the prediction of outcome (Fig. 2). The raw data for these 100 cases were presented to them, without any pre-processing, in a paper based

Table 2 Rules and characteristics for variables that underwent dichotomization

Variable	Categories	Coded as	*N*
LN involvement	No LN involvement	0	2,173
	Positive LN involvement	1	1,526
LN involvement (N0)	No palpable LN	0	638
	Palpable and/or fixed LNs	1	2,996
Periglandular growth	Absence of growth	0	2,977
	Presence of growth	1	722
Estrogen receptor	≥0.3 fmol/mg	0	1,117
	<0.3 fmol/mg	1	2,582
Progesterone receptor	≥0.3 fmol/mg	0	1,531
	<0.3 fmol/mg	1	2,168
S-phase fraction	<10%	0	2,264
	≥10%	1	1,435

questionnaire (Fig. 3). Information for each patient, representing age, physical examination, pathological investigation, and hormone receptor and tumor marker studies, was printed in the questionnaire. Then for each case, the oncologists were asked to place an "X" on a visual analog scale (VAS), from 0 to 100%, to indicate the probability of the occurrence of the outcome. A sample from the questionnaire is shown in Fig. 3.

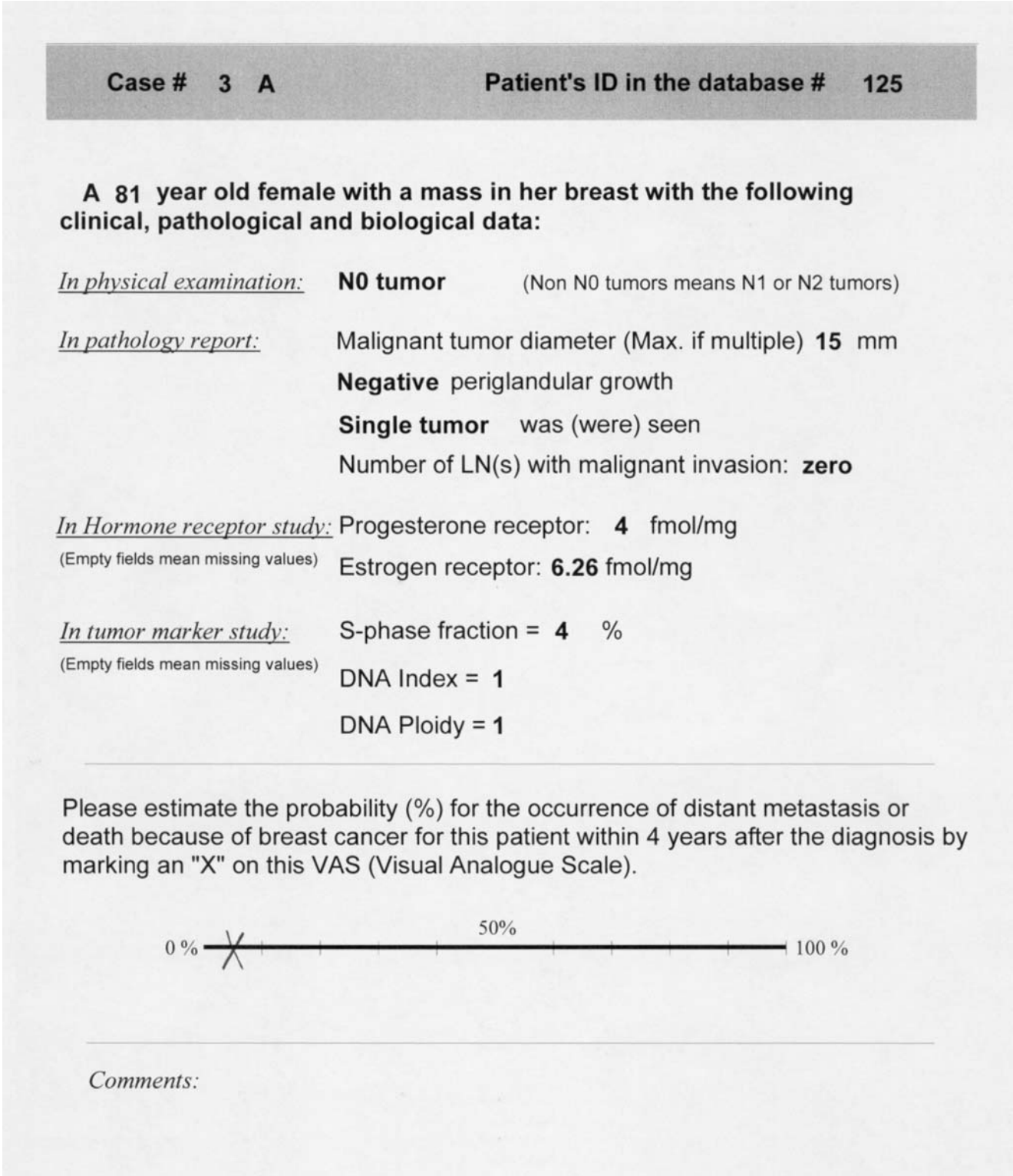


Fig. 3 Questionnaire form given to domain experts

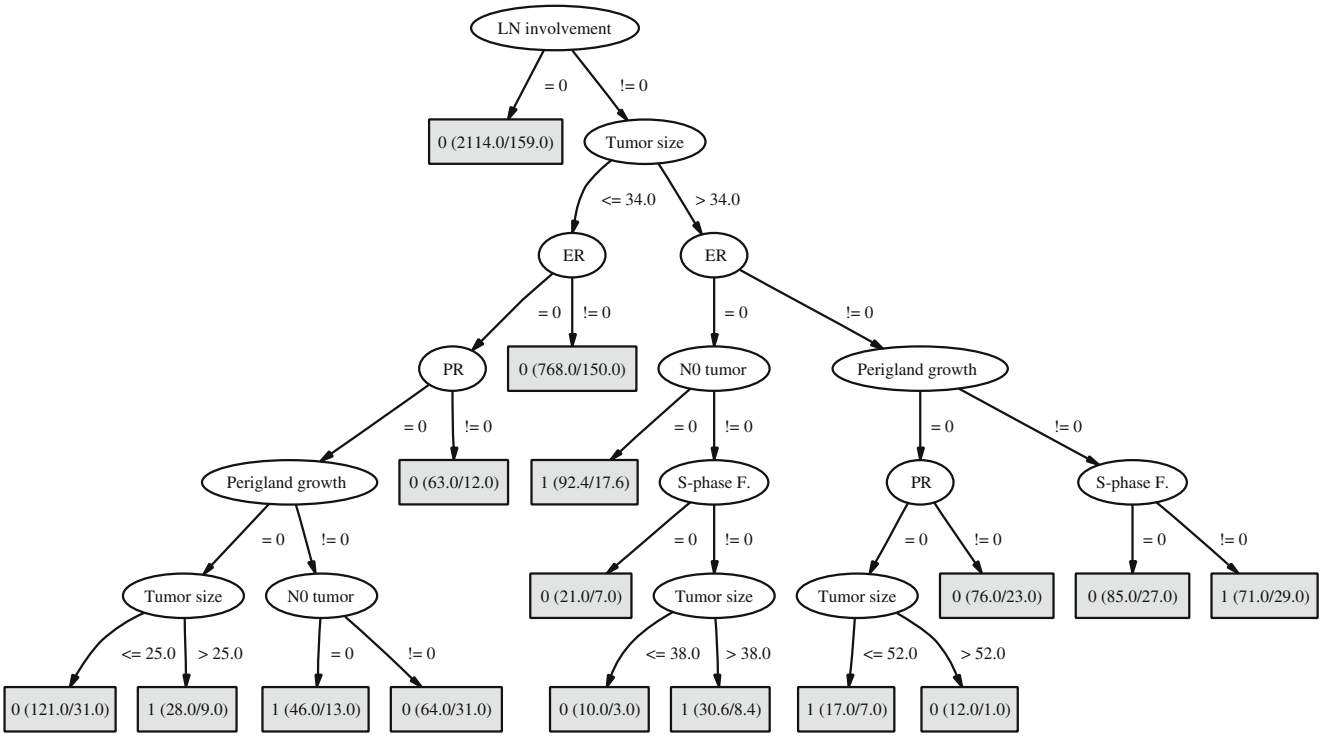


Fig. 4 The resulting decision tree. LN involvement 1/0 shows if the tumor has invaded/not invaded adjacent lymph nodes, tumor size is in millimeters and is obtained from the pathology report. If estrogen or progesterone receptor proteins are positive they are transformed to 1, and if not they are transformed to 0. If the tumor is not palpable in the physical examination then the variable N0 tumor is 1, and otherwise it is 0. Periglandular growth 1/0 indicates if the tumor has grown/not grown outside the tumor boundaries, and S-phase fractions less than 10% are transformed to 0 and larger amounts are transformed to 1. In the leaves (*gray boxes*), there are two numbers in parentheses. The *first number* shows the number of cases who reached this leaf and the *second* shows the number of cases for whom the leaf class was not predicted to happen. The *number outside the parentheses* indicates the class for cases that reach this leaf. 1 means cases with recurrence of the disease and 0 means absence of recurrence

The experts were asked to complete the questionnaires in one session. The number of cases, i.e. 100, was chosen after a discussion with the oncologists regarding the length of the session and how many cases they could read and predict in one session because of their busy schedules.

The discriminating power of the predictive model was tested and compared by calculating the areas under the ROC curves (AUCs) [36–38]. In the next step, these AUCs were compared using the pair-wise comparison method to show whether the differences were significant. Furthermore, differences between the DTI algorithm and each specialist, and the 95% confidence interval (CI), were calculated.

The Hosmer Lemeshow goodness-of-fit test [39, 40] was applied to evaluate how closely the predicted recurrence probabilities fit the observed recurrences.

Results

The decision tree was trained with 3,599 cases (after the exclusion of 100 cases). The complete decision tree and some statistics are shown in Fig. 4 and Table 3. This model was then used for predicting the probability of the occurrence of the outcome of the disease in 100 cases.

Probabilities resulting from the predictive model and from domain experts plus the real outcome for 100 cases were collected and ROC curves were drawn (Fig. 5). Areas under the ROC curve (AUC) for each method were 0.755, 0.810 and 0.847 for DTI, oncologist 1 and oncologist 2, respectively. The difference in AUCs between DTI and oncologist 1 was 0.055 (95% CI=−0.043–0.153) and the

Table 3 Performance of the predictive model created from all data minus 100 cases with 10-fold cross validation

Statistics		
Number of leaves		16
Size of the tree		31
Correctly classified cases (%)		82
Incorrectly classified cases (%)		18
Mean absolute error		0.25
TP rate	No recurrence	0.963
	Recurrence	0.211
Precision	No recurrence	0.839
	Recurrence	0.571

TP: True Positive.

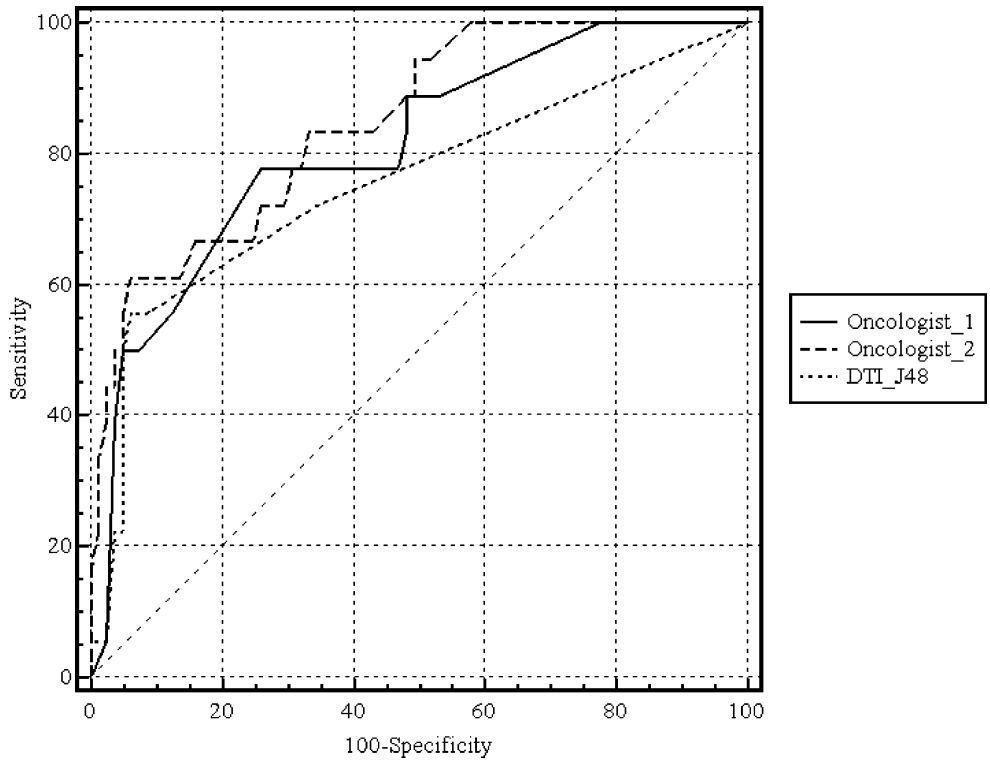


Fig. 5 A comparison between ROC curves

significance level for the difference was 0.27. The difference in AUCs between DTI and oncologist 2 was 0.092 (95% CI=−0.001–0.186) and the significance level for the difference was 0.053. In Table 4, a confusion matrix shows predictions done by domain experts and the decision tree and their comparison with the real values.

After performing the Hosmer Lemeshow goodness-of-fit test, chi-squared values were 3.29, 10.47 and 27.74 and *p*-values were 0.19, 0.16 and 0.0005 for DTI, oncologist 1 and oncologist 2, respectively.

Discussion

Predicting the probability of occurrence of distant metastasis is a very critical task. Both false positive and false negative predictions have unwanted effects on the patient and on the health care system. Accordingly, it is important to discuss the feasibility of the methodology proposed in this study.

The scope of data mining methods

The main aim of constructing cancer registers is not data mining. The data are not gathered for this purpose and registers may not contain all the necessary information. For successful data mining, a maximum number of relevant variables should be available in addition to high quality data. An ordinary breast cancer register may not contain all of the important predictors of recurrence. With the addition of high-tech laboratory tests, the estimation of recurrence may be improved; however, the main research question addressed in this study was whether useful knowledge could be extracted from an ordinary clinical database.

Table 4 Confusion matrix showing predictions done by oncologists and the decision tree in comparison with the real outcomes (no reply from oncologist 1 for one of the cases 1)

	Real outcomes	Oncologist 1		Oncologist 2		DTI	
		No rec	Rec	No rec	Rec	No rec	Rec
No recurrence	81	79	2	70	11	78	3
Recurrence	19	17	1	8	11	15	4

No rec: No recurrence; *Rec*: Recurrence; *DTI*: Decision Tree Induction.

Different arguments have been used concerning how to prepare the data, what method of data mining to use and how to evaluate the results. The good thing about data mining methods is that when they are trained with high quality, relevant data, they perform well. This is why data preparation is an important step in knowledge discovery [41].

Decision support via the predictive model

Convincing clinicians about the usefulness of a clinical decision support model is an important task. In order to do so, we should be able to show the goodness and usefulness of the model. To provide patients with quality health care, clinicians with good knowledge of the specific domain as well as extensive clinical experience are necessary. Senior oncologists use their experience and knowledge to study the risk factors of individual patients. This experience is gained after years of practice and cannot be learned through theoretical education alone. This experience-rich knowledge can be visualized, preserved and reallocated by a predictive model that is to be integrated in a decision support application for use by less experienced oncologists. This is a challenging task, and if it can be done successfully, it will help to increase the quality of health care. However, the most important issue is the attitude of clinicians toward using such a clinical decision support application. In arguing that the extracted knowledge expressed as a predictive model works as well as experienced clinicians, AUC is used to compare the predictions .

Clinicians tend to overestimate the severity of diseases. False positive predictions are more acceptable than false negative predictions. This means that the cut-off for predictions is not the traditional 0.5, and sensitivity and specificity could be different for different diseases based on their severity. To handle this problem we used AUC, which analyzes the whole range of cut-off levels and constitutes a more general validation.

A comparison of AUCs shows that the three approaches for predicting recurrence have no significant differences in discriminating power. The test result provides a p-value where higher values ($p>0.05$) indicate non-significant differences between observed and predicted probabilities. In this case, it implies that the model's estimates fit the data at an acceptable level. However, calibration as assessed by Hosmer Lemeshow goodness-of-fit statistics shows that the DTI model has a higher p-value and works more reliably than the oncologists in predicting the probabilities for recurrence of breast cancer. A predictive model cannot be both perfectly reliable (i.e. calibrated) and perfectly discriminatory [40]. One is increased at the expense of the other, and this may be the reason that the oncologist who had the highest AUC got the lowest calibration (lower p-value).

Our proposed model tends to predict cases with no recurrence better (Table 3, Table 4), because our training database is dominated by "no recurrence" cases. However, since the performance of the model is comparable to the predictions made by domain experts for the 100 test cases, this constitutes an argument in favor of the model and its usability when domain experts are not available.

One way of improving the prediction accuracy for cases with positive recurrence is to use a balanced dataset. With this approach, the number of cases in both classes should be the same. However, because of the rather low number of recurrences, this might result in a small dataset. Training DTI with a small dataset may not result in a meaningful predictive model. Another way is to use sampling techniques to make balanced datasets. This approach combines over-sampling the minority (abnormal) class and under-sampling the majority (normal) class for achieving a better classifier performance [42]. However the dataset is artificially manipulated and may not be representative of the dataset [43].

Using a visual analog scale (VAS) to capture judgments makes it easier for clinicians to express their predictions because they are looking at the whole risk from 0–100%, and this is easier than just writing a percentage.

Validation of the decision tree model

Physicians use all available information about their patients when deciding about the severity of their diseases. This also includes the appearance and mood of the patient and a complete physical examination at the first visit. The decision tree model, on the other hand, is dependent on the availability and quality of the data stored in the register. However, for a realistic comparison between the predictions made by domain experts and the decision tree model, the datasets should be similar. For this reason, the same 100 cases and the same variables for each case were provided to domain experts and the DTI predictive model.

The method used for random sampling is important. In this study, the method is stratified according to the outcome. The ratio of the 1/0 values for the outcome in the sample is equal to that for the whole population. This is important, because due to the random nature of the sampling it is possible to get very high or very low ratios, which can distort the results.

The greater the number of domain experts in a study, the more reliable the results will be. The participation of more experts makes it possible to examine the variability between experts. However, a study of inter-rater variability was not part of the objectives of this study, and the two experts were selected as representatives of clinical practice. The busy schedules of the oncologists and the need to

complete the forms in one session made it difficult to use more than 100 cases.

Future work

It is also possible to review different data mining studies concerning a specific cancer, i.e. breast cancer, and improve the register by adding more relevant predictors. This will improve the quality of the register as a source for a better training set for data mining later on.

In continuing this study, the aim will be to combine the DTI predictive model with guidelines for the treatment and management of cancer in a clinical decision support application integrated with routine daily work.

In following up this study, another step would be to use more domain experts. In order to be able to generalize the performance of our methodology, we should have more domain experts for further validation of the results.

Conclusion

Comparison of the results of human experts and those of the predictive model show that it is possible to formulate the knowledge that is hidden in registers in the form of a decision tree. Since a DTI model is easy to understand and implement, while at the same time producing predictions with the same accuracy as domain experts, the proposed methodology can be used as a semi-automatic knowledge discovery for building predictive models in oncology.

Acknowledgment This study was supported by grant no. F2003-513 from FORSS, the Health Research Council in the South-East of Sweden. Special thanks are due to the South-East Swedish Breast Cancer Study Group for fruitful collaboration and support in this study.

References

1. Sakorafas, G. H., Krespis, E., and Pavlakis, G., Risk estimation for breast cancer development; a clinical perspective. *Surg. Oncol.* 10(4):183–192, 2002 May.
2. Fieschi, M., Dufour, J. C., Staccini, P., Gouvernet, J., and Bouhaddou, O., Medical decision support systems: Old dilemmas and new paradigms? *Methods Inf. Med.* 42(3):190–198, 2003.
3. Fayyad, U., PiatetskyShapiro, G., and Smyth, P., From data mining to knowledge discovery in databases. *AI Mag.* 17(3): 37–54, 1996 Fal.
4. Han, J., and Kamber, M., *Data mining concepts and techniques*. San Francisco: Morgan Kaufmann, 2001.
5. Quinlan, J. R., *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann, 1993.
6. Podgorelec, V., Kokol, P., Stiglic, B., and Rozman, I., Decision trees: An overview and their use in medicine. *J. Med. Syst.* 26(5): 445–463, 2002 Oct.
7. Delen, D., Walker, G., and Kadam, A., Predicting breast cancer survivability: A comparison of three data mining methods. *Artif. Intell. Med.* 34(2):113–127, 2005 Jun.
8. Vlahou, A., Schorge, J. O., Gregory, B. W., and Coleman, R. L., Diagnosis of ovarian cancer using decision tree classification of mass spectral data. *J. Biomed. Biotechnol.* 4(5):308–314, 2003 Dec.
9. Gerald, L. B., Tang, S., Bruce, F., Redden, D., Kimerling, M. E., Brook, N., et al., A decision tree for tuberculosis contact investigation. *Am. J. Respir. Crit. Care Med.* 166(8):1122–1127, 2002 Oct.
10. Atlas, L., Cole, R., Muthusamy, Y., Lippman, A., Connor, J., Park, D., et al., *A performance comparison of trained multilayer perceptrons and trained classification trees. IEEE International Conference on Systems, Man and Cybernetics; 1989 Oct.* Cambridge, MA, USA: Institute of Electrical and Electronic Engineers, pp. 1614–1619, 1989.
11. Brown, D. E., Corruble, V., and Pittard, C. L., A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems. *Pattern Recogn.* 26(6):953–961, 1993 Jun.
12. Talmon, J., Dassen, R., and Karthaus, V., Neural nets and classification trees: A comparison in the domain of ECG analysis. In: Gelsema, E. S., and Kanal, L. N., (Eds.), *Pattern Recognition in Practice IV: Multiple Paradigms, Comparative Studies and Hybrid Systems; 1994.* The Netherlands: Vlieland, pp. 415–423, 1994.
13. Esposito, F., Malerba, D., and Semeraro, G., A comparative analysis of methods for pruning decision trees. *IEEE Trans. Pattern Anal. Machine Intel.* 19(5):476–491, 1997 May.
14. Mehrotra, J., Vali, M., McVeigh, M., Kominsky, S. L., Fackler, M. J., Lahti-Domenici, J., et al., Very high frequency of hypermethylated genes in breast cancer metastasis to the bone, brain, and lung. *Clin. Cancer Res.* 10(9):3104–3109, 2004 May.
15. Wenger, C. R., and Clark, G. M., S-phase fraction and breast cancer—a decade of experience. *Breast Cancer Res. Treatment* 51(3):255–265, 1998.
16. Sundquist, M., Thorstenson, S., Brudin, L., Wingren, S., and Nordenskjold, B., Incidence and prognosis in early onset breast cancer. *Breast* 11(1):30–35, 2002 Feb.
17. Adami, H. O., Graffman, S., Johansson, H., and Rimsten, A., Survival and recurrences five years after selective treatment for breast carcinoma. *Br. J. Cancer* 38(5):624–630, 1978 Nov.
18. Sundquist, M., Thorstenson, S., Brudin, L., and Nordenskjold, B., Applying the Nottingham Prognostic Index to a Swedish breast cancer population. South East Swedish Breast Cancer Study Group. *Breast Cancer Res. Treat.* 53(1):1–8, 1999 Jan.
19. Ciocca, D. R., and Elledge, R., Molecular markers for predicting response to tamoxifen in breast cancer patients. *Endocrine* 13(1): 1–10, 2000 Aug.
20. Lyman, G. H., Lyman, S., Balducci, L., Kuderer, N., Reintgen, D., Cox, C., et al., Age and the risk of breast cancer recurrence. *Cancer Control* 3(5):421–427, 1996 Oct.
21. Razavi, A. R., Gill, H., Stal, O., Sundquist, M., Thorstenson, S., Ahlfeldt, H., et al., Exploring cancer register data to find risk factors for recurrence of breast cancer—Application of Canonical Correlation Analysis. *BMC Med. Inf. Decis. Mak.* 5:29, 2005 Aug.
22. Tejler, G., Norberg, B., Dufmats, M., and Nordenskjold, B., Survival after treatment for breast cancer in a geographically defined population. *Br. J. Surg.* 91(10):1307–1312, 2004 Oct.
23. Piatetskyshapiro, G., Knowledge discovery in databases. *IEEE Intell. Syst. Appl.* 6(5):74–76, 1991 Oct.
24. Lavrac, N., Selected techniques for data mining in medicine. *Artif. Intell. Med.* 16(1):3–23, 1999 May.
25. Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J., Knowledge discovery in databases—An overview. *AI Mag.* 13:57–70, 1992.

26. Hand, D. J., Smyth, P., and Mannila, H., *Principles of data mining.* Cambridge: MIT Press, 2001.
27. Razavi, A. R., Gill, H., Åhlfeldt, H., and Shahsavar, N., A data pre-processing method to increase efficiency and accuracy in data mining. In: Miksch, S., Hunter, J., and Keravnou, E., (Eds.), *10th Conference on Artificial Intelligence in Medicine; 2005 July 23–27*. Aberdeen, UK: Springer-Verlag GmbH, pp. 434–443, 2005.
28. Rubin, D. B., and Schenker, N., Multiple imputation in health-care databases—An overview and some applications. *Stat. Med.* 10(4): 585–598, 1991 Apr.
29. Schafer, J. L., *Analysis of incomplete multivariate data*. London: Chapman & Hall, 1997.
30. McLachlan, G. J., and Krishnan, T., *The EM algorithm and extensions*. New York: Wiley, 1997.
31. Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell, F. E. Jr., et al., Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 79(4): 857–862, 1997 Feb.
32. Luo, Y., and Lin, S., Information gain for genetic parameter estimation with incorporation of marker data. *Biometrics* 59(2): 393–401, 2003 Jun.
33. Zorman, M., Eich, H. P., Stiglic, B., Ohmann, C., and Lenic, M., Does size really matter-using a decision tree approach for comparison of three different databases from the medical field of acute appendicitis. *J. Med. Syst.* 26(5):465–477, 2002 Oct.
34. Witten, I. H., and Frank, E., *Data mining: Practical machine learning tools with Java implementations.* San Francisco: Morgan Kaufmann, 2000.
35. Stone, M., Cross-validation choice and assessment of statistical predictions. *J. Royal Stat. Soc. Ser. B* 36:111–147, 1974.
36. Bradley, A. P., The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recogn.* 30(7): 1145–1159, 1997 Jul.
37. Holmes, J. H., *Quantitative methods for evaluating learning classifier system performance in forced two-choice decision tasks. 2nd International Workshop on Learning Classifier Systems.* pp. 250–257, 1999.
38. Ling, C. X., Huang, J., and Zhang, H., AUC: A better measure than accuracy in comparing learning algorithms. *Adv. Artif. Intell. Proc.* 2671:329–341, 2003.
39. Hosmer, D. W., and Lemeshow, S., *Applied logistic regression*. New York: Wiley, 1989.
40. Jaimes, F., Farbiarz, J., Alvarez, D., and Martinez, C., Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Crit. Care* 9(2):R150–R156, 2005 Apr.
41. Duhamel, A., Nuttens, M. C., Devos, P., Picavet, M., and Beuscart, R., A preprocessing method for improving data mining techniques. Application to a large medical diabetes database. *Stud. Health Technol. Inf.* 95:269–274, 2003.
42. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16:321–357, 2002.
43. Crockett, K., Bandar, Z., and O'Shea, J., *On producing balanced fuzzy decision tree classifiers.* pp. 1756, 2006.