

Managing and Mining Clinical Outcomes

Hyoil Han¹, Il-Yeol Song¹, Xiaohua Hu¹,
Ann Prestrud², Murray F. Brennan³, and Ari D. Brooks²

¹ College of Information Science and Technology, Drexel University,
3141 Chestnut St., Philadelphia, PA 19104, USA

{hyoil.han, songiy}@drexel.edu, thu@cis.drexel.edu

² Department of Surgery, College of Medicine, Drexel University
Hahnemann Hospital, Broad & Vine St. Philadelphia, PA 19102, USA

{ap35, ari.d.brooks}@drexel.edu

³ Department of Surgery, Memorial Sloan-Kettering Cancer Center
New York, NY 10021, USA

m-brennan@mskcc.org

Abstract. In this paper, we describe clinical outcomes analysis for data in Memorial Sloan-Kettering Cancer Center Sarcoma Database using relational data mining and propose an infrastructure for managing cancer data for Drexel University Cancer Epidemiology Server (DUCES). It is a network-based multi-institutional database that entails a practical research tool that conducts On-Line Analytic Mining (OLAM). We conducted data analysis using relational learning (or relational data mining) with cancer patients' clinical records that have been collected prospectively for 20 years. We analyzed clinical data not only based on the static event, such as disease specific death for survival analysis, but also based on the temporal event with censored data for each death. Rules extracted using relational learning were compared to results from statistical analysis. The usefulness of rules is also assessed in the context of clinical medicine. The contribution of this paper is to show that rigorous data analysis using relational data mining provides valuable insights for clinical data assessment and complements traditional statistical analysis and to propose an infrastructure to manage and mine clinical outcomes used in multi-institutional organizations.

1 Introduction

The central hypothesis of our work is that data mining can identify previously unrecognized relationships and mining data compiled from multiple sources will produce more insightful outcomes analysis. Research on health care outcomes is a complex task. One of the reasons that outcomes research needs to be conducted in different manners is data quantity. The amount of data collected must be limited because of the inability to effectively process huge volume of data using current statistical methods. This problem is analogous to that faced by researchers trying to identify the relevant 30,000 genes out of the entire map of the human genome which spans 3 billion base pairs. In this paper, we propose

the use of relational learning (data mining) to identify unanticipated patterns in medical and epidemiologic data. These patterns could not be identified because of limitations in the ability to analyze such volumes of data for meaningful relationships in the current statistical methods. In Sect. 2.2 we show that data mining complements statistical analysis, pinpointing new factors associated with outcomes.

Certain data types lend themselves easily to data mining. However, health care outcomes data has an added complexity of censored data. Not every field in a health related database will be completed. This stems from the fact that biological systems are imperfect. Not all patients receive the same tests, undergo the same treatment, or have the same outcome. Data missing in a medical database cannot be extrapolated or assumed to be unimportant; the fact that it is missing may or may not be associated with outcome. The best example is survival length. Patients entered into a prospective database are entered on different dates. The database is studied on a given date, when some patients have been followed for years and the last patients were entered a few days before analysis. One might incorrectly extrapolate that survival length is longer for patients entered at the start of the study. A technique that enables utilization of data points that are complete and simultaneously drops patients from analysis when their status becomes unknown would clarify this misinterpretation of the data. In survival analysis, such as Kaplan-Meier and Cox [1], this situation is handled using censoring [2]. When the dependent variable (time of survival) is censored, it is included in the analysis until the time at which the data becomes unknown. When other variables are missing in a record, the whole record is left out of the analysis. Each time an event (death) occurs, all patients at risk at that time point are included in the analysis. Patients who have missing information are dropped from analyses following the time point from when the information became missing.

For any data mining techniques to be relevant in health care outcomes analysis, it must be capable of dealing with censored cases. We believe that data mining using relational learning techniques is well suited to censored data; we also propose an algorithm for utilizing time censored data. We used the Memorial Sloan-Kettering Cancer Center Sarcoma Database for experiments in this paper. The Memorial Sloan-Kettering Cancer Center Sarcoma Database is a prospective database that was designed and implemented in 1982 [3]. The purpose of the database is to provide comprehensive information on the diagnosis, treatment and survival of patients with soft-tissue sarcoma, a rare cancer, seen at this world-class institution. In 1995, the database was analyzed to determine factors associated with survival. Pisters and colleagues identified strong relationships between certain tumor and treatment factors and survival that have been validated by many subsequent analyses over time [3] [4] [5] [6]. The aim of the pilot study presented herein is to take this well characterized and validated dataset and utilize data mining techniques to evaluate the validated relationships identified and the new relationships identified. This project evaluates the

utility of relational data mining techniques for handling clinical outcomes data that require censoring.

At Drexel University, we have developed and maintained a multi-center cancer database tracking more than 7000 patients spanning over seven years. The current database, implemented using a FileMaker system, tracks 2166 properties in 34 tables. It will be updated and implemented in SQL Server to provide wide network bandwidth and ease the implementation of a web interface. For this, we propose an infrastructure to manage and mine clinical outcomes, a network-based multi-institutional database that entails a practical research tool that conducts On-Line Analytic Mining (OLAM) [7]. The architecture of the infrastructure is intended to overcome limitations of currently available national databases, such as the Surveillance, Epidemiology and End Results Program (SEER) (see Sect. 4), part of the National Cancer Institute, Division of Cancer Control and Population Sciences and the National Cancer Data Base (NCDB) (see Sect. 4), which is a combined effort of the Commission on Cancer, American College of Surgeons and the American Cancer Society (see Sect. 3 and Sect. 4).

The rest of this paper is organized as follows: section 2 describes relational learning modeling and experiments with results and section 3 presents infrastructure. Section 4 describes related work and section 5 concludes our work and shows our research directions.

2 Relational Learning

Inductive Logic Programming (ILP) is one of the techniques for learning sets of first-order Horn clauses (or rules) that contain variables. The clauses are represented as programs in the logic programming language Prolog. Inductive learning of first-order rules is referred to Inductive Logic Programming [8] [9]. ILP induces hypothesized predicate definitions from examples and relevant background knowledge. ILP combines inductive methods with the power of first-order representations, representing clauses/rules as logic programs. The relationship between examples can be easily expressed with first-order logic representations. In our work, Progol [10] is used for relational learning. Progol constructs concept definitions from examples and background knowledge by using a set cover algorithm. The set cover algorithm (also known as a sequential covering algorithm) generalizes an example in Progol input, adds the generalization to the background knowledge, and removes the examples that the new background knowledge covers. This is repeated until all examples from the input are covered [11]. Progol was successfully used in bioinformatics [12] and web ontology learning [13].

2.1 Relational Learning Modeling and Experimental Methodology

Our representation for propositional logic predicates is `has_gender(A,female)` instead of `has_gender(female)`. Both of them have the same meaning "the given example has gender." When ILP generalizes examples, `has(X,Y)` is more general

and the ILP system generalizes $\text{has}(X, Y)$ over Y . On the other hand, $\text{has}_Y(X)$ is more specific and limits the generalization of ILP. In general, more general predicates require larger numbers of training examples to generalize examples correctly than more specific predicates do. Using more general predicate representation is helpful to obtain richer patterns (or rules).

We conducted experiments in two different ways: static analysis and temporal analysis. We analyzed clinical data not only based on the static event, such as disease specific death for survival analysis, but also based on the temporal event with censored data for each death. To determine training and testing sets, we applied a Bootstrapping procedure [14] that allows sampling with replacement because our data set is not large: it randomly picks half of instances from the data set to form a training set, and the rest is used for a testing set, where each instance represents a patient’s record. We created 10 random splits of the datasets in half, training on one half of the data and testing on the other using the Bootstrapping method. To overcome low coverage, we merged the learned rules to create a rule base that consists of the learned patterns. This would presumably increase coverage slightly with a corresponding penalty in accuracy.

2.2 Experimental Results

We performed analysis of disease specific survival (DSS) using two types of endpoints: static and temporal. First, static analysis was performed. Rules are learned for overall disease specific survival (DSS). In the static analysis, no patient record was censored. Patients who died from Sarcoma were compared to all other patients. Rules are evaluated based on accuracy, coverage and chi-square test. Chi-square test is done with $\alpha=0.05$ with degree of freedom 1. Accuracy and coverage are evaluated by averaging them respectively from 10 splits for training and testing data sets. FULLdata_accuracy means the rule accuracy when the rule is applied to all records including training and testing examples. TESTdata_accuracy means the rule accuracy when the rule is applied to only testing records. FULLdata_accuracy is always larger than TESTdata_accuracy because FULLdata set consists of training examples and testing examples, whereas TESTdata set contains only testing data. When we evaluate rules for unseen data, the accuracy and coverage of TESTdata should be used. FULLdata information is useful for medical doctors to interpret the learned rules based on what attributes affect disease specific death in the current database.

Fig. 1 shows the learned rules with coverage $>2\%$ and confidence level 95%. The accuracy of rules in Fig. 1 shows the best case 66% and the worst case 61%. The target function $\text{class}(A)$ means a patient A with disease specific death.

The first rule explains that the disease specific death of a patient with no gross margin, large newsizecategory and procedure type "a" is predicted with accuracy 61% and coverage 2%. Compared to statistical analysis, these rules show more specific attributes that coexist in the prediction.

Next, we performed temporal analysis by censoring patients who did not have follow-up information in each event. A temporal event is determined whenever a patient dies from disease under consideration (i.e. Sarcoma). The temporal

```

1  class(A) :- has_gross_margins(A,no), has_newsizecat(A,large),
           has_procedure_type(A,a).
FULLdata_accuracy=72%  FULLdata_coverage=4%
TESTdata_accuracy=61%  TESTdata_coverage=2%

2  class(A) :- has_histo1(A,leiomyo), has_newsizecat(A,large).
FULLdata_accuracy=76%  FULLdata_coverage=3%
TESTdata_accuracy=66%  TESTdata_coverage=2%

3  class(A) :- has_grade(A,high), has_procedure_type(A,a), has_subsite(A,thigh).
FULLdata_accuracy=82%  FULLdata_coverage=5%
TESTdata_accuracy=66%  TESTdata_coverage=2%

```

Fig. 1. Rules that have coverage higher than 2% in static analysis

event represents the death as duration from the first visit to hospital to the patient's death. In the temporal analysis, a patient is considered alive before his/her death event. After his/her death, s/he is considered dead. In other words, we performed censoring in our temporal analysis by following survival analysis, such as Kaplan-Meier and Cox [1] [2]. When the dependent variable (time of survival) is censored, a case (that patient) is included in the analysis until the time at which the data becomes unknown. When any other variables are missing in a record, those attributes are left out of the analysis and the other attributes except those in the record are used for relational data mining. Each time an event (death) occurs, all patients at risk for the time point are included in the analysis. Patients who have missing information before that time point are dropped from the analysis at that point. A patient p is classified into one of the following four cases in temporal analysis. Suppose that time of survival of a patient p is $T_s(p)$ and the time of the current event is t_e .

In case that p is dead with diseases that are not under consideration,
 if $T_s(p) < t_e$, p is censored (i.e., excluded from analysis).
 if $T_s(p) \geq t_e$, p becomes a neagative example.

In case that p is dead with disease under consideration,
 if $T_s(p) \leq t_e$, p becomes a positive example.
 if $T_s(p) > t_e$, p becomes a neagative example.

In the early events, there are too few positive examples to learn rules by relational data mining. During the early time period, very few patients have suffered a disease specific death. In this dataset there are 99 events of DSS. The first event occurred within the 1st month. The last event occurred at month 183. In the last event, the number of positive examples is maximized, because the formular $T_s(p) \leq t_e$ can be applied to all patients who are dead by disease under consideration.

Fig. 2 shows the learned rules in temporal analysis. Fig. 2 shows the learned rules with coverage > 10% and confidence level 95%. The accuracy of rules in

```

Event 183 months:
1  class(A) :- has_age(A,old), has_newsizecat(A,large).
FULLdata_accuracy=41%  FULLdata_coverage=23%
TESTdata_accuracy=83%  TESTdata_coverage=12%

2  class(A) :- has_grade(A,high), has_gross_margins(A,no), has_histo1(A,mfh).
FULLdata_accuracy=38%  FULLdata_coverage=24%
TESTdata_accuracy=85%  TESTdata_coverage=12%

3  class(A) :- has_age(A,old), has_grade(A,high), has_newsizecat(A,large).
FULLdata_accuracy=53%  FULLdata_coverage=21%
TESTdata_accuracy=91%  TESTdata_coverage=12%

4  class(A) :- has_grade(A,high), has_newsizecat(A,large), has_gender(A,male).
FULLdata_accuracy=51%  FULLdata_coverage=19%
TESTdata_accuracy=88%  TESTdata_coverage=11%

```

Fig. 2. The learned rules in temporal analysis

Fig. 2 shows the best case 91% and the worst case 83%. The second rule describes that the disease specific death of a patient with high grade, MFH (Malignant fibrous histiocytoma), and no margin is predicted with accuracy 85% and coverage 12%.

In 183-month event shown in Fig. 2, the rules include different attributes. Some of attributes in those rules also differ from attributes in the statistical output shown in Table 1. Therefore, we get clearer features (or attributes) associated with disease specific survival (DSS) by using data mining techniques compared to statistical results shown in Table 1.

Last, we performed statistical analysis as done in 1996 [3] for comparison with our data mining results.

Methods: Statistical analysis was performed with SPSS(tm). Univariate comparisons were conducted utilizing the Kaplan-Meier method, significance of comparisons is based on the log-rank test [1]. Cox regression modeling was conducted to evaluate the influence of all factors on disease-specific survival.

Results: Age, tumor size, site, depth, grade, and histology were each independently associated with disease specific-survival (Table 1). According to results from the univariate analysis, patients older than fifty had decreased survival. Other factors independently associated with decreased survival were large tumors, deep depth, and high grade. Individuals who had sarcomas in a lower extremity and certain pathologic sub-types, particularly MPNT (Malignant peripheral-nerve tumor) and other tumors, also had decreased survival. All factors available were placed into the Cox regression model. Presentation and tumor size, grade, and the histologic sub-type fibrosarcoma comprised the model. Age was marginally significant. Presentation at disease recurrence increases relative risk, RR: 1.1 - 2.1. Tumors greater than 5 cm and high grade tumors were also associated with decreased survival,

RR: 1.1 - 2.6 and 2.9 - 6.7, respectively. Tumors less than 5 cm and the pathologic sub-type fibrosarcoma both decreased relative risk, RR: 0.4 - 0.9 and 0.1 - 0.6.

3 Infrastructure for Cancer Epidemiology Server

In this section we propose a network-based multi-institutional database which entails a practical research tool that conducts On-Line Analytic Mining (OLAM). We have developed and maintained a multi-center cancer database tracking more than 7000 patients spanning seven years. Unlike SEER or NCDB (see Sect. 4), our database was designed with prospective research in mind. The current database, implemented using a FileMaker system, tracks 2166 properties in 34 tables. The properties collected include patients' demographics, cancer symptoms, treatment, and pathology. Our database provides a broader view of patients than that provided by NCDB or SEER. The database will be implemented in SQL Server to provide wide network bandwidth and ease the implementation of web interface. The architecture of our research server, Drexel University Cancer Epidemiology Server (DUCES), is shown in Fig. 3. In the initial phase of this project, three participating clinics will pilot the system. Each clinic runs its own database system, which is utilized daily to enter clinical data and query/browse data stored on its own machine. Each clinic is using the same exact database. The three current databases are stand-alone and isolated, in that data from one clinic cannot be browsed or queried by physicians at other clinics. This greatly limits both data availability and the research capability of the database. DUCES will develop a new architecture that integrates data from all sites into a new research server as shown in Fig. 3.

Data from existing clinics will be periodically replicated into the research server, such that the research server will integrate complete data from all participating clinics. This proposed architecture has at least three advantages.

1. It allows a physician in any clinic to browse or query any patient data from any participating clinic via the Web interface. We will allow each physician to view all data that s/he entered, but restrict access to the data entered by other physicians.
2. At a central location, we can perform sophisticated analysis on the integrated, de-identified data.
3. This model easily supports a future increase in participating clinics.

Advanced analysis for research purposes will be conducted via the Web interface, which will de-identify patient data according to the requirements of the Health Insurance Portability and Accountability Act (HIPAA) [15]. FileMaker data from each clinic will be replicated into the research server using the DTS (Data Transformation Services) of the SQL Server database system. Our web interface will include the following functions: browsing data and generate reports, searching data, inserting data by qualified participants, processing pre-defined reports in terms of various socio-demographic data including age, sex,

Table 1. Disease Specific Survival Analysis with SPSS

Attributes	5 yr DSS Rate %	Univariate p	Selection into Cox Model (score <i>p</i>)	SE	RR	95% CI
Age, years						
<= 50	82.5		0.0580	0.13	0.78	0.6 - 1.0
> 50	71.6	0.0003*				
Gender						
Female	77.7					
Male	77.0	0.7747				
Presentation						
Primary	77.3					
Recurrence	77.9	0.5880	0.0120	0.17	1.53	1.1 - 2.1
Size						
< 5	86.8		0.0240	0.24	0.58	0.4 - 0.9
5-10	76.8		0.0190	0.22	1.68	1.1 - 2.6
> 10	66.2	0.0000*				
Site						
Upper Extremity	83.5					
Lower Extremity	74.5	0.0048*				
Depth						
Superficial Depth	91.6					
Deep Depth	72.7	0.0000*				
Grade						
Low	94.8					
High	67.3	0.0000*	0.0000	0.21	4.4	2.9 - 6.7
Histology						
Liposarcoma	8.15					
MFH	75.0					
Synovial Sarcoma						
Fibrosarcoma	98.0		0.0040	0.49	0.24	0.1 - 0.6
Leiomyosarcoma	68.5					
MPNT	57.0					
Other tumors	69.7	0.0000*				
Micro Margins						
Negative	77.4					
Positive	76.2	0.4841				

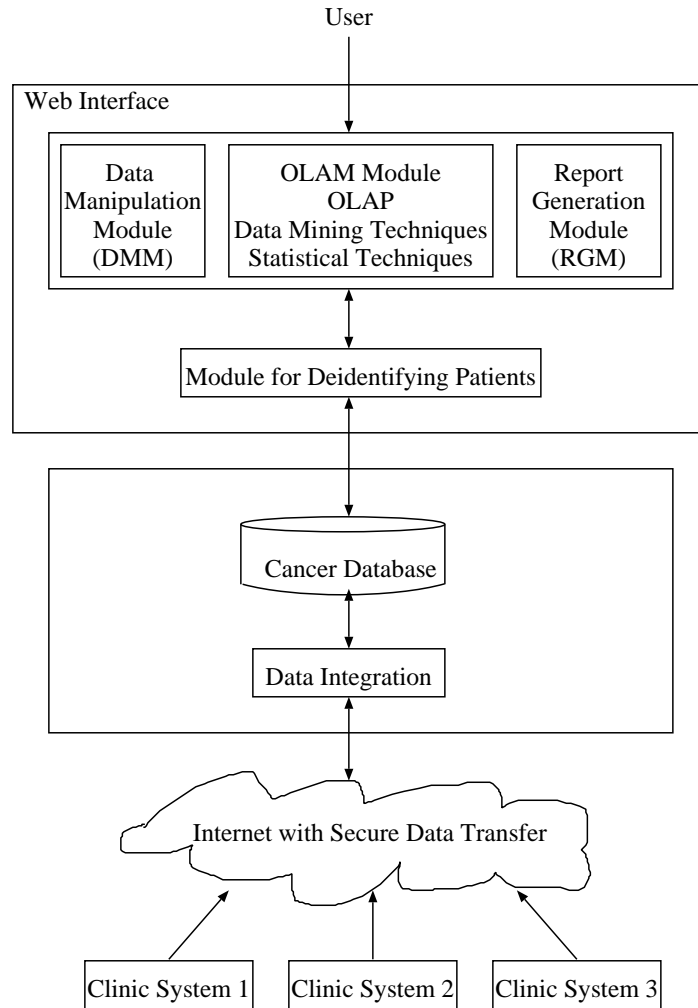


Fig. 3. The Architecture of Drexel University Cancer Epidemiology Server (DUCES)

race/ethnicity and medical conditions such as stage, etc., and applying Online Analytic Mining (OLAM) features for outcome analysis.

One of the key features of our system is OLAM [7], which combines data mining, data warehousing, and online analytic processing (OLAP) technologies. OLAP systems pre-calculate summary information to enable drilling, pivoting, slicing, dicing, and filtering to analyze data sets from multiple angles or dimensions. Data mining provides exploratory data analysis on data sets and additionally detects trend patterns in time dimension and correlation of different properties across other dimensions, as determined by the data mining process.

The multi-institutional comprehensive database will collect cancer data prospectively and the availability of OLAM tool will allow researchers' efforts to experiment with ad-hoc research questions through OLAP and perform follow-up analyses with data mining techniques. Based on the results of data mining techniques, cancer hypotheses can be formulated and tested. These steps can be iterative until research questions are answered.

4 Related Work

Even though data mining techniques have been used widely for prediction and prognosis in medicine and achieved great success [16], not much work has been done in applying data mining methods to outcomes analysis. The work in [17] reports on the analysis of electronic medical records from a Breast Care Center. Their data set consists of 887 patients with 6 attributes pre-chosen by a surgeon and 10% recurrence rate. In their study, various algorithms such as tree classifiers, rule inducers and naive Bayes were adopted to build predictive mode. In [18], Ong et al, designed a system CARES (Colorectal Cancer Recurrence Support), for colorectal cancer recurrence prediction analysis. Their method is based on case-based reasoning and their data set has 10000 curative resection patients.

The main reason that data mining has not been used widely in outcomes analysis is that patients may go through different paths during the study period. Patients may join and leave at different times during the time period. Some patients may die before the study ends, while other patients may still be alive after the study period ends and still other patients may discontinue the study after some time period. This creates a problem for data mining methods because data mining methods normally require the study object have the same starting and ending points. This is normally not true in outcomes analysis. Awareness of differences in outcome, particularly overall survival, is a function of population-based research conducted with large, multi-institutional databases. Currently, two national databases capture information on newly diagnosed cancers. The first is the Surveillance, Epidemiology and End Results Program (SEER), part of the National Cancer Institute, Division of Cancer Control and Population Sciences. Originally developed in 1973, SEER is a population-based database and captures information covering approximately 26% of the U.S. population. The primary focus of this database is reporting incidence and mortality. This project

enables evaluation of both prevention and treatment strategies implemented [19]. The second database is the National Cancer Data Base (NCDB), which is a combined effort of the Commission on Cancer, American College of Surgeons and the American Cancer Society. NCDB commenced data collection over ten years ago and receives data from over 1500 hospitals in all fifty states. NCDB enables surveillance at the local, state, regional, and national levels [20]. Like SEER, NCDB collects data primarily used to track cancer incidence and mortality [21].

However, SEER and NCDB have limitations in identifying the factors which determine outcomes and in their ability to answer new research questions [22]. One concern is that all data is entered retrospectively. This may create errors and makes entry of a complete record less likely than with a prospective approach. A second limitation of both SEER and NCDB is the limited scope of data collected. Diagnosis information is collected through review of pathology reports. Treatment information is limited to whether or not the patient received treatment. It does not include when treatment was received, what type of treatment, nor treatment duration. In 1999, the Committee on Cancer Research among Minorities and the Medically Underserved noted that SEER does not have the capability to explain differences in incidence and mortality rates, particularly among ethnic minorities and medically underserved communities. Furthermore, this prevents the development of effective prevention activities within these at-risk communities [23].

5 Conclusions and Future Research.

We showed that rigorous data analysis using relational data mining provides valuable insights for clinical data assessment and complements traditional statistical analysis and proposed an infrastructure to manage and mine clinical outcomes used in multi-institutional organizations. The rules from relational data mining have attributes that did not appear in statistical analysis. This suggests that rules learned by data mining techniques can be used to complement statistical analysis and even provide better explanations for outcomes analysis. The proposed architecture overcomes the limitations in SEER and NCDB such as in identifying the factors which determine these disparities and in their ability to answer new research questions as they arise. Clinical outcomes frequently have many missing values and rare events. We need a specific strategy to handle rare events. Clinical outcomes also have temporal properties, such as follow-up data and procedures. We are working towards a solution to handle analysis of this type of data.

References

1. Altman, D. G.: Practical Statistics for Medical Research. Chapman & Hall, (1999)
2. Daniel, W.: Biostatistics: A Foundation for Analysis in the Health Sciences. 7th Edn. John Wiley and Sons, Inc., New York (1999)

3. Pisters, P., Leung, D., Woodruff, J., Shi, W., Brennan, M.: Analysis of Prognostic Factors in 1,041 Patients with Localized Soft Tissue Sarcomas of the Extremities. *Clinical Oncology*. **14** (1996) 1679–1689
4. Stojadinovic, A., Leung, D.H.Y., Hoos, A., Jaques, D.P., Lewis, J.J., Brennan, M.F.: Analysis of the prognostic significance of microscopic margins in 2084 localized primary adult soft tissue sarcomas. *Ann. Surg.* **235** (2002) 424–434
5. Kattan, M.W., Leung, D.H., Brennan, M.F.: Postoperative nomogram for 12-year sarcoma-specific death. *J. Clin. Oncol.* **20** (2002) 791–796
6. Lewis, J.J., Leung, D., Woodruff, J.M., Brennan, M.F.: Retroperitoneal soft-tissue sarcoma: analysis of 500 patients treated and followed at a single institution. *Ann. Surg.* **222** (1998) 355–65
7. Han, J., Kamber, M.: *Data Mining: Concepts And Techniques*. Morgan Kaufmann Publishers, San Francisco (2001)
8. Russell, S.J., P. Norvig, P.: *Artificial Intelligence: A Modern Approach*. Prentice Hall (1995)
9. Mitchell, T.M.: *Machine Learning*. WCB/McGraw-Hill (1997)
10. Muggleton, S.: Inverse entailment and Progol. *New Generation Computing*. **13** (1995) 245–286
11. Muggleton, S.: *CProlgol4.4: A Tutorial Introduction*. Inductive Logic Programming and Knowledge Discovery in Databases (2001)
12. Cootes, A., Muggleton, S.H., Sternberg, M.J.E.: The automatic discovery of structural principles describing protein fold space. *Journal of Molecular Biology* (2003)
13. Han, H., Elamsri, R.: *Learning Rules for Conceptual Structure on the Web*. Journal of Intelligent Information Systems, Kluwer Academic Publishers (2004)
14. Liu H., Motoda, H.: *Feature Selection For Knowledge Discovery and Data Mining*. Kluwer Academic Publishers (1998)
15. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). (1996)
16. Zupan, B., Lavrac, N.: Data Mining Techniques and Applications in Medicine. *Artificial Intelligence in Medicine*, **16** (1999) 1–2
17. Mani, S., Pazzani, M.J., West, J.: Knowledge Discovery from a Breast Cancer Database: Proceedings on Artificial Intelligence in Medicine, Europe, (1997)
18. Ong, L.S., Shepherd, B., Tong, L.C., Seow-Choen, F., Ho, Y.H., Tang, C.L., Ho, Y.S., Tan, K.: The Colorectal Cancer Recurrence Support (CARES) System. *Artificial Intelligence in Medicine* **11** (1997) 175–188
19. Hankey, B., Ries, L., Edwards, B.: The surveillance, epidemiology, and end results program: a national resource. *Cancer Epidemiol Biomarkers Prev.* **8** (1999) 1117–1121
20. Menck, H., Bland, K., Eyre, H., Cunningham, M., Fremgen, A., Murphy, M.: Clinical highlights from the National Cancer Data Base. *Cancer J Clin.* **48** (1998) 134–145
21. McGinnis, L., Menck, H., Eyre, H., Bland, K., Scott-Conner, C., Morrow, M.: National Cancer Data Base survey of breast cancer management for patients from low income zip codes. *Cancer* **88** (2000) 933–945
22. Hindle, W.: Breast Cancer: Introduction. Diagnosis and treatment of nongynecologic cancer. *Clin Obstet Gynecol.* **45** (2002) 738–745
23. Shinagawa, S.: The excess burden of breast carcinoma in minority and medically underserved communities: application research, and redressing institutional racism. *Cancer* **88** (2000) 1217–1223