Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
JNMU
Journal of Nanjing Medical University, 2007, 21(3): 190–195
www.elsevier.com/locate/jnmu
Research Paper

The diagnostic rules of peripheral lung cancer preliminary study based on data mining technique

Yongqian Qiang[a], Youmin Guo[b,*], Xue Li[c], Qiuping Wang[a], Hao Chen[c], Duwu Cui[c]
[a]*Imaging Center, the First Affiliated Hospital, Xi'an Jiaotong University, Xi'An 710061, China*
[b]*Imaging Center, the Second Affiliated Hospital, Xi'an Jiaotong University, Xi'An 710004, China*
[c]*Computer Faculty, Xi'an University of Technology, Xi`An 710048, China*
Received 22 November 2006

Abstract

Objective: To discuss the clinical and imaging diagnostic rules of peripheral lung cancer by data mining technique, and to explore new ideas in the diagnosis of peripheral lung cancer, and to obtain early-stage technology and knowledge support of computer-aided detecting (CAD). **Methods**: 58 cases of peripheral lung cancer confirmed by clinical pathology were collected. The data were imported into the database after the standardization of the clinical and CT findings attributes were identified. The data was studied comparatively based on Association Rules (AR) of the knowledge discovery process and the Rough Set (RS) reduction algorithm and Genetic Algorithm(GA) of the generic data analysis tool (ROSETTA), respectively. **Results**: The genetic classification algorithm of ROSETTA generates 5 000 or so diagnosis rules. The RS reduction algorithm of Johnson`s Algorithm generates 51 diagnosis rules and the AR algorithm generates 123 diagnosis rules. Three data mining methods basically consider gender, age, cough, location, lobulation sign, shape, ground-glass density attributes as the main basis for the diagnosis of peripheral lung cancer. **Conclusion**: These diagnosis rules for peripheral lung cancer with three data mining technology is same as clinical diagnostic rules, and these rules also can be used to build the knowledge base of expert system. This study demonstrated the potential values of data mining technology in clinical imaging diagnosis and differential diagnosis.

Keywords: peripheral lung cancer; tomography; X-ray computed; data mining; computer aided detecting(CAD)

INTRODUCTION

Medical data, such as peripheral lung cancer, often seem to contain a great number and uncertain or irrelevant features. How to extract enough necessary and useful diagnostic rules used to be highly depended on the clinical experience. Recently, intelligent techniques [1-3] have been proven to be an effective tool for data or "knowledge mining" in many real-world fields including medical diagnosis. Rough Set(RS) theory can deal with uncertainty and incompleteness in data analysis. This attribute, the reduction algorithm(which removes redundant information or features) selects a feature subset that has the same discernibility as to the original set of features. The Genetic Algorithm (GA) has been proven to be one of the search methods and optimization techniques for an optimal value of a complex objective function by simulation of the biological evolutionary process, as in genetics, on crossover and mutation. Association rule (AR) mining algorithm, as originally proposed in with its apriori algorithm, has developed into an active research area [4,5]. But few of these three algorithms have been used in knowledge mining in diagnosing the peripheral lung cancer [6,7]. The aim of the work was to determine the value of RS reduction algorithm, GA and AR mining algorithm in this question.

*Corresponding author.
E-mail address: Imagingqyq@163.com

MATERIALS AND METHODS

Study population

The study protocol was approved by the Research Ethics Committee of Xi'an Jiao Tong University. All patients received the informed consensus. According to inclusion and exclusion criteria, 58 patients (age range, 35-70 years old; median age, 50 years old), including 43 man and 15 women with peripheral lung cancer, were selected in from the year of 2004 to 2006. All patients received the CT scan three days before operation or biopsy.

Inclusion and exclusion criteria

The final diagnosis was determined by pathological result from operation or biopsy. The patients had not received any anti-cancer therapies before imaging. The clinical information was supplied direct from the patient. CT scans were obtained with an one-slice helical CT scanner (PHILIPS Co.) in our imaging center. A tube voltage of 120 KV and current of 200 mA were used. Slice thickness and reconstruction interval for routine scanning were 5mm, CT images were displayed at fixed setting (lung window center, -300~-500 Hu; lung width, 1300~1500 Hu; mediastinum window center, 30~50 Hu; mediastinum window width,400 Hu). Data were reconstructed to 512 × 512 matrices.

Predictor variables

Clinical information and CT findings were used as first-grade candidate predictors. Clinical information included fifteen clinical variables, and CT findings included twenty-four CT variables (***Fig 1***). All these predictors were collected from the corresponding literatures [8,9]. Thirty-eight candidate predictors were categorical variable except age(***Tab 1***).

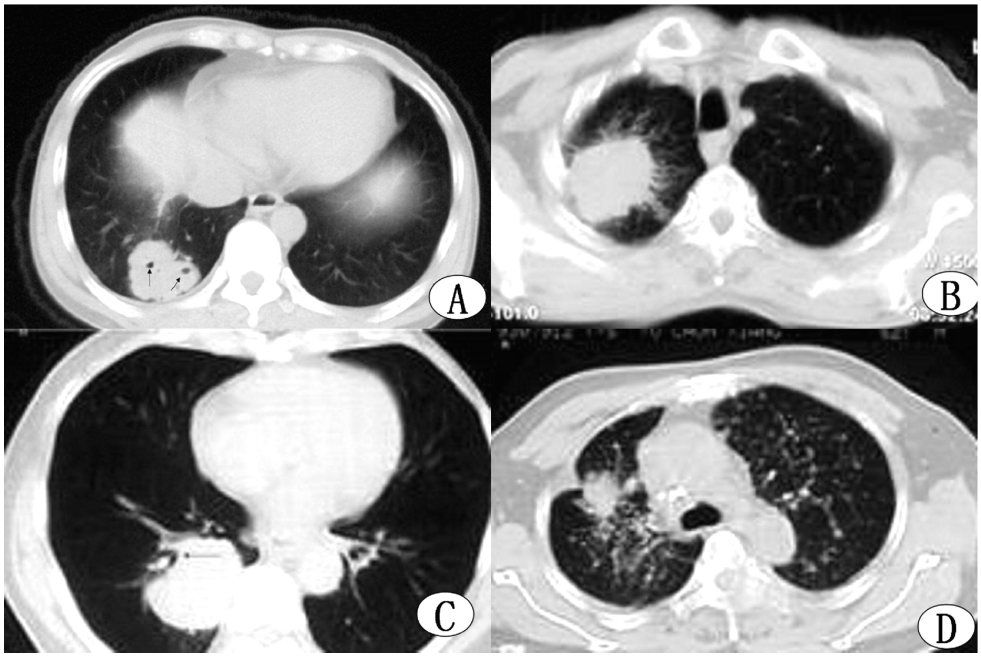

Fig 1 CT finding of peripheral lung cancer shows lobulation sign and vacuole sign (A), spiculation sign(B), notching sign (C) and pleural indentation sign(D).

Characteristic-recognizing of predictors

First, two senior attending radiologists with 15 to 20 years' experience blinded, independently recognized the category of each candidate predictor. Disagreements were resolved by discussion. The data was analyzed by AR algorithm and Johnson's Algorithm tool box in the software package ROSETTA.

RS

RS theory handles information or data with uncertainty based on equivalence relations and partitioning of finite sets [10]. An equivalence relation R partitions the set S into some non-empty disjoint subsets of S whose union equals S, which is the equivalence class. Both the lower and the upper approximation are defined to describe an arbitrary set X, which may not be accurately classified by all equivalence class. The lower approximation of X contains some equivalence classes guaranteed completely inside the set X. And the upper approximation of X contains some equivalence classes completely covered the set X. The difference between the lower and the upper approximation are those equivalence classes covered the boundaries.

Tab 1 **Predictors Variables**

	Variables	Code	Attribute identification
Clinical manifestation	Sex	1 = male, 2 = female	Categorical variable
	Age	Years	Continuous Variable
	Cough	0 = no; 1 = yes	Categorical variable
	Expectoration	0 = no; 1 = yes	Categorical variable
	Hemoptysis	0 = no; 1 = yes	Categorical variable
	Chest pain	0 = no; 1 = yes	Categorical variable
	Fever	0 = no; 1 = yes	Categorical variable
	Dyspnea	0 = no; 1 = yes	Categorical variable
	Heart-throb	0 = no; 1 = yes	Categorical variable
	Cyanopathy	0 = no; 1 = yes	Categorical variable
	Cracked	0 = no; 1 = yes	Categorical variable
	Smoking	0 = no; 1 = yes	Categorical variable
	Losing-weight	0 = no; 1 = yes	Categorical variable
	Primary tumor	0 = no; 1 = yes	Categorical variable
	Family tumor history	0 = no; 1 = yes	Categorical variable
CT manifestation	Size	1 = D≤1 mm; 2 = 1 mm < D ≤ 2 mm; 3 = 2 mm <D ≤ 3mm	Categorical variable
	Location	1 = left upper lobe 2 = left lingular segment 3 = left lower lobe 4 = right upper lobe 5 = right middle lobe 6 = right lower lobe	Categorical variable
	Edge of lesion	0 = smooth; 1 = non-smooth	Categorical variable
	Notching sign	0 = no; 1 = yes	Categorical variable
	Lobulation	0 = no; 1 = shallow lobulation; 2 = deep lobulation	Categorical variable
	Spiculation sign	0 = no; 1 = long spiculation; 2 = short spiculation	Categorical variable
	Halo-sign	0 = yes; 1 = no	Categorical variable
	Convergence of vessels sign	0 = yes; 1 = no	Categorical variable
	Satellite lesion	0 = yes; 1 = no	Categorical variable
	Pleural indentation sign	0 = yes; 1 = no	Categorical variable
	Lesion density	0 = uniform;1 = non-uniform	Categorical variable
	Vacuole sign	0 = yes; 1 = no	Categorical variable
	Air-space sign	0 = yes; 1 = no	Categorical variable
	Cavity	0 = no; 1 = thin wall; 2 = thick wall	Categorical variable
	Calcification	0 = no; 1 = tiny sand-like; 2 = node-like; 3 = flake-like; 4 = mass-like	Categorical variable
	Lymph node enlargement	0 = no;1 = yes	Categorical variable
	Shape of lesion	0 = round; 1 = round-like 2 = irregular	Categorical variable
	Ground-glass attenuation	0 = no;1 = yes	Categorical variable
	Pleural indentation sign	0 = no;1 = yes	Categorical variable
	Pleural thickening	0 = no;1 =yes	Categorical variable
	Pleural involvement	0 = no;1 = yes	Categorical variable
	Pleural effusion	0 = no;1 = yes	Categorical variable
	Obstruction sign	0 = no;1 = yes	Categorical variable
	Enhancement	0 = no;1 = yes	Categorical variable

GA

The GA starts with the generation of the initial random population; then each member (design configuration) of the population is evaluated based on the fitness value. Each member of the next generation is formed with the process of cross-over that represents the exchange of genes of the parents to produce an offspring [11]. Since the higher probability to become the parent is assigned to the members with higher fitness value, the characteristics providing the best fit are carried to the offspring. Because the cross-over is the main engine of the evolution, the probability of the process is usually high (0.7~1.0). The processes of mutation and permutation might be applied to some of the members in the new generation to expand the search space by perturbing the genes. The corresponding probabilities are usually small (0.001~0.15) in order to ensure the evolution in the right direction and consequent convergence rather than exploring the entire search space. The best design is always transferred from generation to generation.

AR

In this section, the data was analyzed by AR mining algorithms by self-designed software based on classical Frequency Algorithm[12]. Step 1,reducing redundancy information: redundant information or features were removed and a new feature subset that has the same discernibility as the original set of features were selected by data-preparation; Step 2, producing frequent item set: assembly of n feature subset(n = 2,3,4,5…i,n represents the number of attribute) were randomly selected to determine the support. Such a procedure was repeated with adding one attribute to n, (n+1 attributes) until the support of assembly including all attributes is lower than the minimum support, which led to frequent item set. Step 3, diagnostic rule is $\{n\}$ to disease. As the rule-finding was performed in the same disease, peripheral lung cancer , the confidence of each rules is higher than the strong associated rule with the minimum support according to the formula, the number of (assembly of n attribute∪disease)/the number of assembly of n attributes=100%.

RESULTS

Fifty-one diagnostic rules were extracted by the RS reduction algorithm of Johnson`s Algorithm generates. Three typical rules were listed below: ① age≥60, no chest-pain, no smoking, lesion located in posterior segment of right upper lobe with short spiculation sign; ②age {56,60} no chest-pain, no smoking ,lesion located in posterior basal segment of right lower lobe with short spiculation sign and without satellite lesions; ③age {56,60}, no expectoration, lesion located in posterior basal segment of right lower lobe with short spiculation sign and without satellite lesions.

Ninety-nine reduction sets were generated from Johnson`s Algorithm, extracting over 5000 diagnostic rules(***Fig 2***). Five typical rules were listed below: ① age≥60 years ,no expectoration or chest pain, lesions located in anterior basal segment of right lower lobe with short spiculation sign; ②age≥60 years, lesion size(cm) {5.01,10},located in apicoposterior segment of left upper lobe, lesion with smooth edge but without convergence of vessels sign. Lesion with pleural thickening and pleural indentation sign; ③ age {41,50}, cough, no acratia, the number of smoking per day {20,40}, lesion located in lateral basal segment of right lower lobe with irregular contour, like-umbilical notching sign and convergence of vessels signs; ④ male patient, age {56,60}, chest pain, lesion located in lateral basal segment of right lower lobe with round contour, smooth edge and ground-glass attenuation but without spiculation sign; ⑤ age {41,50}, cough, acratia, lesion located in lateral basal segment of right lower lobe with irregular contour, deep lobulation sign, long spiculation sign and pleural indentation sign.

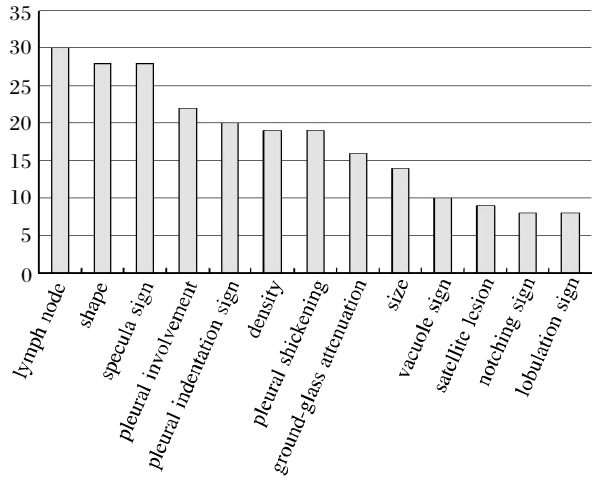

Fig 2 sensitivity ranking of radiological attributes from GA

One handred and twenty-three diagnostic rules were generated by the AR algorithm (***Fig 3***). Four typical rules were listed below: ① lesion in round shape with unsmooth edge and shallow lobulation sign, short spiculation sign and like-umbilical notching sign(24.1%);② patient with acratia feeling, le-

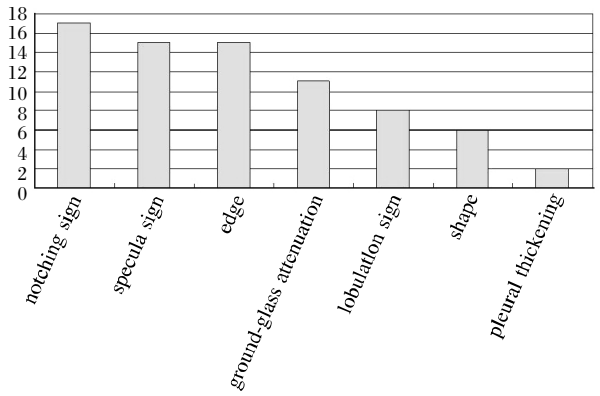

Fig 3 sensitivity ranking of radiological attributes from AR algorithm

sion in irregular shape with unsmooth edge, short spiculation sign and like-umbilical notching sign (25.8%); ③patient with cough and acratia feeling, lesions with shallow lobulation sign , short spiculation sign and like-umbilical notching sign (27.5%); ④patient with cough, lesions with unsmooth edge, shallow lobulation sign , short spiculation sign and like-umbilical notching sign(31%).

DISCUSSION

The diagnostic rules for peripheral lung cancer from RS reduction algorithm concentrated in male-population. The scope of age occurred above 55 years old, and especially above 60 years old. Few cases had fever, losing weight, chest-pain and like-umbilical notching sign. The lesion mainly located in the posterior segment of right upper lobe, lateral basal segment and posterior basal segment of right lower lobe. Lesions mainly had short spiculation sign. The clinical attributes including acratia, cough and expectoration associated with location of lesion (mainly lateral basal segment of right lower lobe). Lesion did not associate with smoking. The above diagnostic rules basically coincided with clinical diagnosis. But the association between smoking and lesion wasn't demonstrated by this algorithm. Rough set Johnson Reducer is a kind of minimum length reduction algorithms [13],producing the reduction assembly of rule's structure. Zhang *et al* [14] used the RS algorithm to remove fourteen redundant attributes from forty-three primary attributes of 110 cases with confirmed osteogenic sarcomas and to extract the diagnostic rules. The generated rules coincide with clinical diagnostic rules for osteogenic sarcomas. In this research, diagnostic rules for peripheral lung cancer from rough set associations with the five attributes including sex,age, expectoration, location and spiculation sign of lesions

The GA is one of the global search methods and optimization techniques for an optimal value of a complex objective function by simulation of the biological evolutionary process [15]. Diverse and a large amount of assembly of attributes often were generated by the reduction and rules generated by the GA. Attributes in high frequency of occurrence included location of lesions, age, cough, expectoration, chest-pain, lymphaden and the diameter of lymphaden, shape of lesion, spiculation, pleural involvement, acratia, pleural indentation sign, the number of smoking per day, lesion density, pleural thickening and ground-glass attenuation. The common characteristic between GA and RS was concentrated in male-population with the range of age locating in above 56 years old and lesion with short spiculation sign.

The differentia between two algorithms is location. Only posterior segment of right upper lobe and lateral basal segment of right lower lobe were regarded as associated with cancer by GA. Moreover, there is no association between expectoration and location of lesion. Lilla *et al* [16] used the Support Vector Machine on the base of GA to classify 189 solitary pulmonary nodules and effectively reduce the false positive value. More diagnostic rules extracting by GA for peripheral lung cancer need more concentration.

An AR algorithm is of the form $X \rightarrow Y$, where X and Y are both frequent item sets in the given database, and the intersection of X and Y is an empty set, i.e., $X \cap Y = \phi$. The support of the rule $X \rightarrow Y$ is the percentage of transactions in the given database that contain both X and Y, i.e,$P(X \cup Y)$. The confidence of the rule $X \rightarrow Y$ is the percentage of transactions in the given database containing X that also contains Y, i.e $P(Y \mid X)$. Therefore, AR algorithm is used to find all the associated rules among item sets in a given database, where the support and confidence of these associated rules must satisfy the user-specified minimum support and minimum confidence. By finding the associated rules between the basic manifestations with CT manifestation and the patients with head trauma, Susan *et al*[17] acquired the criteria by which patients with head trau ma need and therefore receive a CT scan. Ye *et al*[18] applied AR algorithm to extract the diagnostic rules from neurogliocytoma, whose average accuracy were over 80% , satisfying the clinical diagnostic requirements. Our self-designed AR algorithm program can mine the associated rule on the base of attributes' assembly. Each rule satisfying the manifestation of lesion represents the pathological change, moreover, support of rules was acquired from each

attribute. The associated rules between diagnostic rules and clinical attributes included: ① age ≥ 60 years old, cough, expectoration and chest-pain; ② rules was not sensitive to fever, losing-weight, primary tumor, night sweats, pulmonary function. Some rules contradict with present clinical diagnostic rules. Diagnostic rules recognized CT attributes including lesion's size (3.1-5.0 cm), shallow lobulation sign, short spiculation sign, like-umbilical notching sign, non-uniform density, ground-glass attenuation and pleural-thickening. Both Clinical attributes including age range from 55 to 60 years old and over 60 years old, male-population, insensitive to smoking number or fever or losing-weight, and CT imaging attributes including sensitivity to speculation were common points between ROSETTA and self-design AR algorithm. Difference included:①AR algorithm generated the larger attribute range like contour, lobulation, like-umbilical notching sign, lesion density, pleural-thickening and so on; ② AR algorithm suggested that cough, expectoration and acratia should be sensitive attributes but not lesion's location; ③ rules that can reduce the attributes to 'exclusion' has special ability of optimation; ④ rules from RS algorithm is smaller than AR algorithm. Moreover, rules from AR algorithm are smaller than GA. Diagnostic rules from AR algorithm had a higher coincidence. But because of the higher complexity and lower efficiency of AR algorithm, and a great number of redundancy information in rules[19], using AR algorithm in medical fields is still in the exploratory process.

Three algorithms were sensitive to sex, age, cough, location, spiculation, shape and ground-glass attenuation. After reduction, the range of attributes became different.

Johnson Reducer algorithm generates only one reduction set, but a reduction set in diversity was generated from GA, and reduction set had divergence sensitivity. Reduction set generated from GA had a middle level in set number and sensitivity to attributes. Johnson Reducer algorithm couldn't satisfy the requirement of diversity in clinical diagnostic rules. However, GA generated too many rules. AR balanced the diversity and number of rules. GA had the best ability of optimization and was seem to optimize the new generated and conventional rules. Totally, rules from the above three algorithms basically coincided with present clinical diagnostic rules and could supply important references in clinical practice.

In our future work, efficient data mining algorithms will be explored in a large number of samples to supply the optimized diagnostic rules.

References

[1] Katapka H, Sugiura T. The ideal form of laboratory information management. *Rinsho Byori* 2005;53:39-46.

[2] Skevofilakas M, Nikita K, Templakeksis P, Birbas K, Kaklamanos IG, Bonatsos GN.A decision support system for breast cancer treatment bases on data mining technologies and clinical practice guidelines. *Conf Proc IEEE Eng Med Biol Soc* 2005;3:2429-32.

[3] Lamma E, Mello P, Nanetti A. Artificial intelligence techniques for monitoring dangerous infection. *IEEE Trans Inf Technol Biomed* 2006;10:143-55.

[4] Agrawal R, Imielinski T, Swami A. Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering* 1993;5: 914 - 25 .

[5] Adepele O, Sylvanus E. A fast algorithm for mining association rules in medical image data. *Proceeding of the 2002 IEEE Canadian Conference on Electrical & Cpmputer Engineering* 2002:1181-7.

[6] Matsuki Y, Nakamura K, Watanabe H, Aoki T, Nakata H, Katsuragawa S, et al.Usefulness of artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT:evalution with receiver operating characteristic analysis. *AJR AM J Roentgenol* 2002;178 :657-63.

[7] Gurney JW. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis.2. *Application. Radiology* 1993;186:405-13.

[8] Erasmus JJ, Connolly JE, Page McAdams H, Roqqli VL. Solitary pulmonary nodules: part I. Morphologic evaluation for differentiation of benign and malignant lesions. *Radio Graphics* 2000;20:43-8.

[9] Ost D, Fein A. Evaluation and Management of the solitary pulmonary nodule. *Am J Respir Crit Care Med* 2002;162:782-7.

[10] Honghai F, Guoshun C, Yufeng W, Bingru Y, Yumei C. Rough Set Based Classification rules generation for SARS Patients. *Conf Proc IEEE Enq Med Biol Soc* 2005;7:6977-80.

[11] Boroczky L, Zhao L, Lee KP. Feature subset selection for improving the performance of false positive reduction in lung nodule CAD. *IEEE Trans Inf Technol Biomed* 2006;10:504-11.

[12] Xuanyang X, Yuchang G, Shouhong W, Xi L. Computer Aided Detection of SARS Based on Radiographs Data Mining. *Conf Proc IEEE Eng Med Biol Soc* 2005;7:7459-62.

[13] Zhai Peng. Evaluation and introduction of attribute reducing methods. *SCI/TECH Information Development and Econmy* 2004;14:98-9.

[14] Zhang H, Qian ZC, Qu JH. Application research for building bone tumor assisted diagnostic knowledge database based on rough set. *Medical Information* 2004;17:257-8.

[15] Cunrong li, Mingzhong Yong. Association Rules Data Mining in Manufacturing Information System based on Genetic Algorithms. *2004 3d International Conference on Computational Electromagnetice and Its Applications Proceedings* 2004:153-6.

[16] Lilla Boroczky, Luyin Z, K.P.Lee. Feature Subset Selection for Improving the Performance of False Positive Reduction in Lung Nodule CAD. *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)* 2005:85-90.

[17] Susan P, Bernard D, Hilary W. Using dependency/association rules to find indications for computed tomography in a head trauma dataset. *Artif Intell Med* 2002;26: 55-8.

[18] Ye CZ, Yang J, Geng DY. Data Mining in Diagnostic Knowledge Acquisition from Patients with Brain Glioma. *Chin J Biomed Enq* 2002;19:426-30.

[19] Georgii E, Richter L, Ruckert U, Kramer S. Analyzing microarray data using quantitative association rules. *Bioinformatics* 2005;21:123-9.