J Med Syst (2013) 37:9899
DOI 10.1007/s10916-012-9899-y

ORIGINAL PAPER

Mining Medical Data: A Case Study of Endometriosis

Yi-Fan Wang · Ming-Yang Chang · Rui-Dong Chiang · Lain-Jinn Hwang · Cho-Ming Lee · Yi-Hsin Wang

Received: 5 July 2012 / Accepted: 6 September 2012 / Published online: 17 January 2013
© Springer Science+Business Media New York 2013

Abstract Ultrasound guided aspiration of ovarian endometrioma had been tried as an alternative therapeutic modality in patients whose desire to avoid surgery or surgical approach is contraindicated since 1991. Cyst puncture can reduce tumor volume and destruct the cyst wall, alleviate sticking circumstances and enhance the chance of recovery. But simple aspiration without other treatments results in high recurrence rate (28.5 % to 100 %). In order to reduce recurrence after aspiration, ultrasound-guided aspiration with instillation of tetracycline, methotrexate, and recombinant interleukin-2 has been combined and proven to be effective with the recurrence rates of 46.9 %, 18.1 %, and 40 % respectively. Noma et al. (2001) reported that conduct of ethanol instillation for more than 10 min particularly for a case with a single endometrial cyst is considered most effective from the standpoint of recurrence (14.9 %). Our goal is to analyze patients with recurrent pelvic cyst who underwent surgical intervention. The research data are based on clinical diagnosis, symptoms and medical intervention classification, and the cyst numbers are defined as forecast project target. The decision tree, methodology of data mining technology, is used to find the meaningful characteristic as well as each other mutually connection. The experimental result can help the clinical faculty doctors to better diagnose and provide treatment reference for future patients.

Keywords Endometriosis · Data mining · Decision tree · Medical data

Y.-F. Wang (✉)
Institute of Information and Decision Sciences, National Taipei College of Business, Taipei, Taiwan, Republic of China
e-mail: yfwang_tw@webmail.ntcb.edu.tw

M.-Y. Chang
Department of Obstetrics and Gynecology, Chang Gung Memorial Hospital, Taipei, Taiwan, Republic of China

R.-D. Chiang · L.-J. Hwang · C.-M. Lee
Department of Computer Science and Information Engineering, Tamkang University, Taipei, Taiwan, Republic of China

R.-D. Chiang
e-mail: Chiang@cs.tku.edu.tw

L.-J. Hwang
e-mail: micro@mail.tku.edu.tw

C.-M. Lee
e-mail: 896410049@s96.tku.edu.tw

Y.-H. Wang (✉)
Department of Information Management, Chang Gung University of Science and Technology, Taoyuan, Taiwan, Republic of China
e-mail: yishing@gw.cgust.edu.tw

Introduction

Endometriosis has been discovered in many women to be asymptomatic. Yet for some women endometriosis can produce extreme discomfort, pain and even lead to infertility [1]. Standard treatment measures include expectant management, medical suppression and surgical ablation, with surgical ablation being the most common method [9]. However, clinical research has shown that within 5 years of surgery there is still a 19 % recurrence rate [21, 22], which may be due to recurrent endometriomas or adhesion cysts. Enlarged pelvic cysts will change the relation between ovaries and fallopian tubes to result in infertility, and pressure onto nearby organs could lead to chronic pelvic pain and dyspareunia, while endometriotic lesions that have damaged the ovarian stroma will also decrease ovarian reserve.

Ultrasound-Guided Aspiration (UGA) of ovarian endometrioma with sclerotherapy has become an alternative treatment to avoid surgery [16, 18]. Aspiration of endometriomas can shrink the cyst size, cause collapse of cyst wall, decrease adhesions and promote healing. However, performing UGA alone without other additional treatments can lead to high recurrence rates (28.5–100 %) [12]. Therefore to lower the high recurrence rates of UGA, certain medical groups have begun utilizing UGA with tetracycline [13], methotrexate [11] or interleukin-2 [7] irrigation therapies, the recurrence rates of which is 46.9 %, 18.1 % and 40 % respectively. Although in 2003 Kafali had also performed 5-min irrigations on

endometriosis patients with 70 % ethanol and erythromycin, encouraging results were not achieved [4]. Thus in clinical use, Noma's research outcome of 14.9 % recurrence rate with 10-min ethanol irrigation of each endometrioma is still the preferred therapeutic reference guide [18].

In present clinical medicine, there are many traditional treatment strategies targeted at infertility caused by endometriosis, but UGA combined with chemosclerotherapy has gradually become a new alternative treatment that can shrink cyst size, collapse cyst wall, decrease adhesions and promote healing, which not only effectively prevents surgical damage to female reproductive organs, but also decreases the discomfort caused by pain and the risk of recurrence [12].

Knowledge discovery and data mining have become very popular research topics in recent years. By applying data mining algorithms the characteristic and relationships between data can be discovered and further interpreted into meaningful information. For example, in DNA sequencing the information produced by data mining have been combined with visual computer diagrams to form 3D simulations for biologists and medical researchers to determine their functions and pathogenesis, which is a tremendous contribution to humans [2, 3, 17].

In addition to being widely applied in bioinformatics and molecular research, data mining is often used to test surgical efficacy and medical and pharmaceutical researches in clinical practice, and it is also used to discover the relationship between clinical and pathological data [19]. These data, collected from medical information systems of medical institutions, contain medical information that include the Picture Archiving and Communication System (PACS), related to X-rays, Clinical Information System (CIS), related to clinical paths, or testing systems related to sleep, ventilator, ECG, and gastro scope equipment, which if analyzed and researched with data mining techniques, should uncover key results that will benefit clinical or academic medicine.

This research had targeted patients whose endometrioses recurred after undergoing related surgeries. The data were analyzed for characteristics, similarities and differences to discover important key factors that are related to the data and clinical symptoms, which may be used by the clinical specialists as basis for diagnosis and treatment.

Technology used in this article

Data mining is a technique to perform data analysis, either automatically or semi-automatically, with various algorithms to extract valuable information from data [5]. The functions of data mining have been categorized below: Classification, Estimation, Prediction, Affinity Grouping, Clustering and Description and Visualization.

According to the definition of U.M. Fayyad, data mining is a step in the process of knowledge discovery in database [10]. In fact, knowledge discovery can also be regarded as the overall process of data mining. The methodology to effectively utilize data mining techniques and principles for locating valuable events hidden in massive databases has become an important topic for academic and practical fields.

The decision tree is one of data mining technology very suitable for performing classification and prediction. The decision tree can produce results according to different variable by repetition that can thus be used to analyze the characteristics, similarities and differences in data. Decision tree are often represented graphically. The process begins by classifying data into groups based on set threshold, and after repeated steps the classification method or predicted method for all data will be eventually determined.

Decision trees use the tree branch structure to produce classification rules suitable for all classification problems, such as bank credit grants (predicting bad debts), direct marketing response (predicting response) and customer loss expectation (predicting loss) etc., where the decision tree can be used to produce easily understandable rules.

In current practice the decision tree has become a very mature technique. The most famous algorithms of decision tree are C4.5 (Interactive Dichotometer 3, [20]), CART (Classification and Regression Tree, [6]), and CHAID (Chi-square Automatic Interaction Detector, [14]) which have gained wide prevalence in academic and practical applications.

Proposing a recommender system

For mining the medical data and proposing some recommends to cure endometriosis, we propose a recommender system. The system is designed on an IBM PC using JAVA and includes four major modules: preprocessing, field selection, classification, and Interpretation

In this proposed system, users can gain recommendations from the system. Since the process is application-oriented, different applications may need different classification approaches as appropriate. For present purposes, the decision tree is employed as the classification function. Figure 1 is a simplified architecture of the system.

Preprocessing

The data source for performing the analytical processes of this research was collected from a reproductive medicine center affiliated with a certain national medical center. From July 2001 to July 2005 a total of 178 cases of recurrence following endometriosis surgeries were collected. Prior to conducting analysis, since the original data included many errors or irrelevant information, which may possibly interfere with

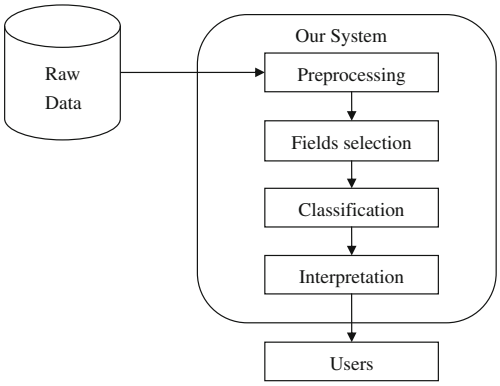


Fig. 1 A simplified system architecture

algorithms and thereby affect test accuracy, preprocessing of the original data was necessary. Generally, preprocessing of data mining include the following: standardization of missing or noise values, transformation of data formats and data, calculation of derived attributes and selection of data sample size.

Fields selection

Many field selection algorithms have been proposed, some of which have reported remarkable accuracy improvement [15]. Eventually the number of endometriomas as the reference for determining treatment efficacy was set, and the column containing values relevant to number of endometriomas were selected from the original data, as seen in Table 1. Parts of the original data are shown in Fig. 2.

Finally the clinical test values and information concerning treatments were selected the patients' basic information to conduct decision tree analysis. They are listed below in Table 2.

Classification

As mentioned in Section 2, scholars [8, 14, 20] had very well addressed the Decision Tree algorithms. In this part, we shall borrow from the CART algorithm which is using Gini impurity to measure the branch node values that help to understand the exact meanings of the rules and to discover more meaningful rules.

Table 1 Columns relevant to number of endometriomas

Name of column	Content
SN	Patient Case Number
Cyst No.	Number of endometriomas at initial visit
Cyst No. 3 m	Number of endometriomas at 3-month follow-up
Cyst No. 6 m	Number of endometriomas at 6-month follow-up
Cyst No. 12 m	Number of endometriomas at 12-month follow-up

	A	B	C	D	E
1	SN	Cyst no.	Cyst no.3m	Cyst no.6m	Cyst no.12m
2	1	2	2	2	
3	2	1			
4	3	6	2		
5	4	3	2		
6	5	1	1	1	1
7	6	3	1	1	1
8	7	1		0	0
9	8	4	1	0	0
10	9	1	1		
11	10	5	2	1	1
12	11	2	0	0	0
13	12	2	0	0	
14	13	2		1	2
15	14	5	1	0	0
16	15	4	1	1	1
17	:	:	:	:	:
18	:	:	:	:	:
19	:	:	:	:	:

Fig. 2 The original data of endometrioma counts

Interpretation

Due to the many generated rules, the discovered characteristics and information were made first into easily

Table 2 Overview of columns selected for decision tree analysis

Column name	Content
Age	Patient's age at initial visit
Cyst No	Number of endometriomas at initial visit
Cyst1 Size 01	Size of endometrioma at initial visit
CA125 01	Patient's CA125 value at initial visit
Med PreOP	Type of oral medication prior to surgery
	A: None
	B: Danazol, GnRHa
	C: MPA, HRT, Oral Pills, Mirena, Chinese, Others
AF Total	Number of follicles at initial visit
Assoc# Disease	Concomitant Disease
	0: None
	1: Adenomyosis
	2: PID
	3: Severe Adhesion
	4: CA-125 During Bleeding
	5: OCD Pregn.
Cyst Asp Min#	Length of drug retention time during UGA
Cyst asp retention	Type of medication during used UGA
	A: None
	B: Alcohol
	C: Glove Powder, Danazol, KM, Alcohol & Glove Powder, TC

CA125: This is an Ovarian Epithelial Tumor Antigen and its blood level can be used both as a guide to determine the severity of endometriosis.

	A	B	C	D	E	F
1	SN	Cyst no.	Cyst no.3m	Cyst no.6m	Cyst no.12m	MCystNo
2	1	2	2	2		1
3	2	1				0
4	3	6	2			1
5	4	3	2			1
6	5	1	1	1	1	1
7	6	3	1	1	1	1
8	7	1		0	0	1
9	8	4	1	0	0	1
10	9	1	1			0
11	10	5	2	1	1	1
12	11	2	0	0	0	1
13	12	2	0	0		1
14	13	2		1	2	0
15	14	5	1	0	0	1
16	15	4	1	1	1	1
17	:	:	:	:	:	:
18	:	:	:	:	:	:
19	:	:	:	:	:	:

Fig. 3 Part of data after data transformation of number of endometriomas

understandable reports and then modified in order to assist decision making.

Experimental result

With the premises of adhering to hospital guidelines, respect for medical ethics and protection of patient confidentiality, this research drew 178 case records of visits at the gynecology (Outpatient Department, OPD) between July 2001 and July 2005 concerning endometriosis recurrence following surgery to be used for analysis.

The research samples are accurate and sufficient because a total of 178 case records are collected within the population of 23 million in Taiwan. Samples 1 to 170 are used as training samples and the rest are used as testing samples. Afterward, Samples 9 to 178 are used as training samples and the rest (Sample 1 to Sample 8) are used as testing samples, and so forth. Just as expected, the result showed positive.

The case records included the patients' basic information, various data from the initial visits and three follow-up visits at the frequency of 3 and 6 months and 1 year. Based on the attributes the data source can be broadly classified into the following three groups:

1) Patients' Basic Information: Such as age, number of pregnancies, number of births, number of miscarriages, past histories, menstruations periods, regularity of menstruations, periods of menstrual flow, severity of dysmenorrhea, urges to defecate, dyspareunia, whether other pains exist and other concomitant histories.
2) Clinical Test Values: Such as endometrioma counts, sizes of endometriomas, follicle counts, CA125 blood values, sizes of uteruses (measurements), level of ovarian adhesions and contents of endometriomas.
3) Treatment-related Information: Such as medication prior to surgery, medication following surgery, route of drug administration, surgical method, surgical route, UGA method, UGA site, UGA with irrigation and medication used.

Goal analysis

Although clinically endometriosis is considered benign, accumulated endometriomas will only grow and cause adhesions to spread further and ultimately impair ovarian and fallopian tubal functions. Therefore after discussing with clinical doctors, the number of endometriomas before and after the treatment as the analysis target of this research was set.

Discussion with clinical doctors was conducted to better understand UGA and the types of medication for irrigation or retention, and whether the length of period of drug irrigation or retention would affect treatment outcome. The relevant columns were also included for consideration as targets of this research.

After determining the analysis targets transformation, the preprocessing of original data was underway. The research defined treatment as effective if the number of endometriomas at 12-month follow-up was less than that of the initial

Tree	Node ID	Score	Number of Records	Saturation
True	1	N	178 (100%)	52.8%
Cyst no# < 1.5	1.1	N	89 (50%)	79.8%
Cyst no# ≥ 1.5	1.2	N	89 (50%)	74.2%
CYST1 size 01 < 3.15	1.2.1	N	86 (48%)	76.7%
((Assoc# diseases = 4) 或 (Assoc# diseas	1.2.1.1	N	59 (33%)	69.5%
CA_125 01 < 15.55	1.2.1.1.1	Y	10 (6%)	100.0%
CA_125 01 ≥ 15.55	1.2.1.1.2	N	49 (28%)	63.3%
((Assoc# diseases ≠ 4) 和 (Assoc# diseas	1.2.1.2	Y	27 (15%)	92.6%
CYST1 size 01 ≥ 3.15	1.2.2	N	3 (2%)	100.0%

Fig. 4 Result of decision tree analysis where number of endometriomas ≥1.5

Tree	Node ID	Score	Number of Records	Saturation
True	1	Y	178 (100%)	52.8%
Cyst no# < 1.5	1.1	Y	89 (50%)	79.8%
CA_125 01 < 9.7	1.1.1	Y	18 (10%)	50.0%
CA_125 01 ≥ 9.7	1.1.2	N	71 (40%)	87.3%
((Assoc# diseases = 4) 或 (Assoc# diseases = 2))	1.1.2.1	Y	12 (7%)	66.7%
((Assoc# diseases ≠ 4) 和 (Assoc# diseases ≠ 2))	1.1.2.2	N	59 (33%)	91.5%
Cyst no# ≥ 1.5	1.2	Y	89 (50%)	74.2%

Fig. 5 Result of decision tree analysis where number of endometriomas <1.5

visit—this was represented by "1"; an ineffective treatment was represented by "0" otherwise. After transformation the column name was set to "MCystNo" and was marked red, and parts of data is shown in Fig. 3.

Outcome analysis

After analyzing the decision tree, the result generated by the algorithm according to number of endometriomas and UGA, is shown below.

Figure 4 shows that the decision tree outcome with the analytical target field—the number of endometriomas (Cyst no#). The tree was divided into two branches determined by the number of endometriomas which were greater than or equal to (≥) 1.5. Two rules were obtained:

1) If the patient was discovered to have ≥1.5 endometriomas, with endometriomas smaller than 3.15 cm and concomitant (pelvic inflammatory disease, PID) or CA125 during bleeding at less than 15.55U/ml, then the number of endometriomas would improve (decrease) after treatment, was 63.3 % satisfied.
2) If the patient had ≥1.5 endometriomas, with endometriomas smaller than 3.15 cm and without concomitant PID or CA125 during bleeding, then the number of endometriomas would improve (decrease) after treatment, was 92.6 % satisfied.

In view of the above results, we discovered the treatment is effective when the numbers of endometriomas were ≥1.5 and endometriomas were less than 3.15 cm, regardless of concomitant PID or CA125 during bleeding.

However, there were 89 cases with the number of endometriomas was less than (<) 1.5. For research completeness, we analyzed the 89 cases further. The generated decision tree is shown in Fig. 5.

The outcome showed that, if the patient's number of endometriomas was <1.5, the CA125 value was ≥9.7U/ml and without concomitant PID or CA125 during bleeding, then the number of endometriomas decreased after treatment with satisfaction of 91.5 %.

As shown in Fig. 6, 59 cases had numbers of endometriomas of <1.5 (Only one among 55 cases) and CA125 values of ≥9.7U/ml. But without concomitant PID or CA125 during bleeding, it should be considered as the main cause of infertility.

To determine the relationship between the types of chemotherapy used in UGAs, the lengths of drug retention time and whether the number of endometriomas had improved after treatment, we targeted irrigation or drugs or drugs retention time to generate a decision tree; the result is shown below:

We discovered that when UGA was performed with Alcohol for irrigation or with alcohol for retention, in 27 cases the numbers of endometriomas were ≥1.5 and the endometriomas were <6.55 cm. After treatment the number

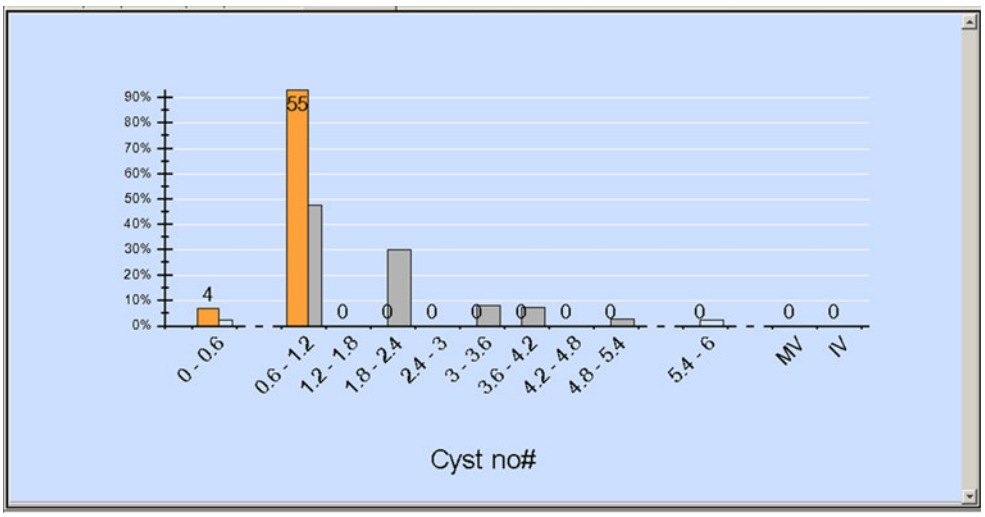
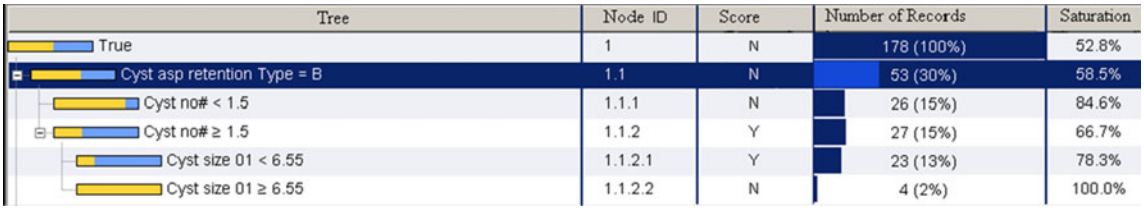


Fig. 6 Distribution of patients with number of endometriomas <1.5 and CA125 values of ≥9.7U/ml without concomitant PID or CA125 during bleeding



Tree	Node ID	Score	Number of Records	Saturation
True	1	N	178 (100%)	52.8%
Cyst asp retention Type = B	1.1	N	53 (30%)	58.5%
Cyst no# < 1.5	1.1.1	N	26 (15%)	84.6%
Cyst no# ≥ 1.5	1.1.2	Y	27 (15%)	66.7%
Cyst size 01 < 6.55	1.1.2.1	Y	23 (13%)	78.3%
Cyst size 01 ≥ 6.55	1.1.2.2	N	4 (2%)	100.0%

Fig. 7 Decision tree result of patients after UGA treatment

of cases with improvement (decrease) in endometriomas was 23, which accounted for 85 % and why the decision tree's rule satisfaction reached 78.3 %. In contrast, in the 4 cases where the endometriomas were ≥6.55 cm, improvement (decrease) in number of endometriomas did not occur after UGA with Alcohol irrigation, and the decision tree's rule satisfaction was 100 %.

According to the outcome of the decision tree in Fig. 7 on patients who received UGA, we discovered that if the number of endometriomas were ≥1.5, and endometriomas were <6.55 cm, then though the proportion of improvement (decrease) in number of endometriomas after treatment had reached 85 %, the decision tree's rule satisfaction had also reached 78.3 %.

Conclusion

Since the advent of UGA with drug retention or irrigation therapy in 1991, few academic researches and theses have been published. Some alternative measures are needed to avoid open surgery for removal of endometriomas, clinically used in infertility patients.

After our reviewing the results and discussion with the clinical doctors, the medical care decision rules that were discovered by this research through data mining techniques, such as in diagram 4, 5 and 7 are exciting. The medical decision rules targeting number of endometriomas derived by this research have already been confirmed by clinical doctors and listed as reference for clinical diagnosis and treatment.

For future study, this research will collaborate with more hospitals and construct a "Data Mining Orientated" database to further develop software for data collection and analysis. By continuous efforts to increase sample size and repeated cross inference, an effective and more realistic predictive model might be constructed for better treatment for Endometriosis patients.

References

1. Aboulghar, M. A., Mansour, R. T., Serour, G. I., and Rizk, B., Ultrasonic transvaginal aspiration of endometriotic cysts: An optional line of treatment in selected cases of endometirosis. *Human Reproduction* 6(10):1408–1410, 1991.
2. Alexandrov, N. N., and Mironov, A. A., Application of a new method of pattern recognition in DNA sequence analysis: A study of E.coli promoters. *Nucleic Acids Research* 18(7):1847–1852, 1990.
3. Alnahi, H., and Alshawi, S., Knowledge discovery in biomedical databases—a machine induction approach. *Computer Methods and Programs in Biomedicine* 39(3–4):343–349, 1993.
4. Aloulghar, M. A., Mansour, R. T., Serour, G. I., Sattar, M., Ramzy, A. M., and Amin, Y. M., Treatment of recurrent chocolate cysts by transvaginal aspiration and tetracycline sclerotherapy. *Journal of Assisted Reproduction and Genetics* 10(8):531–533, 1993.
5. Berry, M. J., and Linoff, G., *Data mining techniques: for marketing, sales, and customer support*. Wiley, New York, NY, 1997.
6. Breiman, L., Friedman, K. H., Olshen, R. A., Stone, C. J. (1984). *Classification and regression trees*. Chapman & Hall.
7. Brosens, I. A., Puttemans, P. J., and Deprest, J., The endoscopic localization of endometrial implants in the ovarian chocolate cyst. *Fertil Steril* 61(6):1034–1038, 1994.
8. Chiang, D. A., Chen, W., Wang, Y.-F., and Hwang, L.-J., Rules generation from the decision tree. *Journal of Information Science and Engineering* 17:325–339, 2001.
9. Dicker, D., Goldman, J. A., Feldber, D., Ashkenazi, J., and Levy, T., Transvaginal ultrasonic needle-guided aspiration of endometriotic cysts before ovulation induction for in vitro fertilization. *Journal of Assisted Reproduction and Genetics* 8(5):286–289, 1991.
10. Fayyad, U., Piatetsky-Shapiro, G.., Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 37–54.
11. Hughesdon, P. E., The structure of endometrial cysts of the ovary. *The Journal Obstetrics and Gynaecology of the British Empire* 64 (4):481–487, 1957.
12. Ikuta, A., Tanaka, Y., Mizokami, T., Tsutsumi, A., Sato, M., Tanaka, M., Kajihara, H., and Kanzaki, H., Management of transvaginal ultrasound-guided absolute ethanol sclerotherapy for ovarian endometriotic cysts. *Journal of Medical Ultrasonics* 33(2):99–103, 2006.
13. Kafali, H., Yurtseven, S., Atmaca, F., and Ozardali, I., Management of non-neoplastic ovarian cysts with sclerotherapy. *International Journal of Gynecology and Obstetrics* 81(1):41–45, 2003.
14. Kass, G. V., An exploratory technique for investigating large quantities of categorical data. *Journal of Applied Statistics* 29 (2):119–127, 1980.
15. Liu, H., Motoda, H. (1998). *Feature selection for knowledge discovery and data mining*. Kluwer Academic Publishers.
16. Mittal, S., Kumar, S., Kumar, A., and Verma, A., Ultrasound guided aspiration of endometrioma—a new therapeutic modality to improved reproductive outcome. *International Journal of Gynecology & Obstetrics* 65(1):17–23, 1999.
17. Miyano, S. (1995). Learning theory toward genome informatics, *IEICE Transactions on Information and Systems*, 560–567.
18. Noma, J., and Yoshida, N., Efficacy of ethanol sclerotherapy for ovarian endometriomas. *International Journal of Gynecology and Obstetrics* 72(1):35–39, 2001.

19. Prather, J.C., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L., Hammond, W.E. (1997). Medical data mining: knowledge discovery in a clinical data warehouse, *Proceedings of the AMIA Annual Fall Symposium*, 101–105.
20. Quinlan, J. R., Learning efficient classification procedures and their application to chess end games. In: Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (Eds.), *Machine learning: an artificial intelligence approach*. Morgan Kaufmann, CA, USA, pp. 463–482, 1983.
21. Redwine, D. B., Conservative laparoscopic excision of endometriosis by sharp dissection: Life table analysis of reoperation and persistent or recurrent disease. *Fertil Steril* 56(4):628–634, 1991.
22. Wheeler, J. M., and Malinak, L. R., Recurrent endometriosis. *Contrib Gynecol Obstet* 16:13–21, 1987.