Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks

Chung-Ho Hsieh, MD,[a,b] **Ruey-Hwa Lu, MD,**[b] **Nai-Hsin Lee, MD,**[b] **Wen-Ta Chiu, MD, PhD,**[e] **Min-Huei Hsu, MD,**[c] *and* **Yu-Chuan (Jack) Li, MD, PhD,**[a,c,d] *Taipei, Taiwan*

Background. *Diagnosing acute appendicitis clinically is still difficult. We developed random forests, support vector machines, and artificial neural network models to diagnose acute appendicitis.*
Methods. *Between January 2006 and December 2008, patients who had a consultation session with surgeons for suspected acute appendicitis were enrolled. Seventy-five percent of the data set was used to construct models including random forest, support vector machines, artificial neural networks, and logistic regression. Twenty-five percent of the data set was withheld to evaluate model performance. The area under the receiver operating characteristic curve (AUC) was used to evaluate performance, which was compared with that of the Alvarado score.*
Results. *Data from a total of 180 patients were collected, 135 used for training and 45 for testing. The mean age of patients was 39.4 years (range, 16–85). Final diagnosis revealed 115 patients with and 65 without appendicitis. The AUC of random forest, support vector machines, artificial neural networks, logistic regression, and Alvarado was 0.98, 0.96, 0.91, 0.87, and 0.77, respectively. The sensitivity, specificity, positive, and negative predictive values of random forest were 94%, 100%, 100%, and 87%, respectively. Random forest performed better than artificial neural networks, logistic regression, and Alvarado.*
Conclusion. *We demonstrated that random forest can predict acute appendicitis with good accuracy and, deployed appropriately, can be an effective tool in clinical decision making. (Surgery 2011;149: 87-93.)*

From the Institute of Biomedical Informatics,[a] *National Yang-Ming University, Taipei; Department of General Surgery,*[b] *Zhong-Xing Branch of Taipei City Hospital, Taipei; Graduate Institute of Biomedical Informatics,*[c] *Taipei Medical University, Taipei; Department of Dermatology,*[d] *Taipei Medical University - Wan Fang Hospital, Taipei; and Graduate Institute of Injury Prevention and Control,*[e] *Taipei Medical University, Taipei, Taiwan*

Acute appendicitis (AAP) is one of the most common conditions needing emergent operation. Deciding whether to operate is still a challenge for surgeons, with negative appendectomy rates of 10–25% in recent decades[1,2] leading to an unnecessary operation and suffering for patients. A false negative diagnosis may delay diagnosis and lead to increased morbidity, mortality and even legal problems.[3]

Accepted for publication March 25, 2010.

Reprint requests: Yu-Chuan (Jack) Li, MD, PhD, No. 250, Wu-Hsing Street, Taipei City, Taiwan 110. E-mail: jack@tmu.edu.tw.

0039-6060/$ - see front matter

doi:10.1016/j.surg.2010.03.023

Clinical suspicion of AAP usually needs further appropriate confirmatory test. Strategies to improve the accuracy of AAP diagnosis include active observation, clinical scoring systems, abdominal computed tomography (CT), ultrasonography (US), and diagnostic laparoscopy. Active observation is common.[4] Considering the risk of perforation, the threshold to decide to operate is usually low. The Alvarado score is a common clinical scoring system with sensitivity of 64% and specificity of 84% (Table I).[5,6] Abdominal CT may increase accuracy to 98%,[7] but it is a relatively high cost examination and is not always available. US is a useful tool, but its accuracy is highly operator dependent.[7,8] Diagnostic laparoscopy may increase the negative exploration rate but decrease the negative appendectomy rate.[9]

Table I. Alvarado scoring system for diagnosing acute appendicitis[5]

Clinical features	*Score*
Migration of pain to right lower quadrant	1
Anorexia	1
Nausea/vomiting	1
Tenderness in right lower quadrant	2
Rebound tenderness	1
Elevated temperature ≥37.3°C	1
Leukocytosis (WBC >10,000/μl)	2
Neutrophilic shift to the left >75%	1
Total	10
Recommendation	
Score <5	Appendicitis unlikely
Score 5 or 6	Appendicitis possible
Score 7 or 8	Appendicitis likely
Score 9 or 10	Appendicitis highly likely

Machine learning has contributed much in modern clinical decision making. Random forests, support vector machines, and artificial neural networks are among the most important models and all are supervised learning classifiers that can be used in complex clinical situations.

Random forest, proposed by Breiman in 1999, is an ensemble classification algorithm consists of a collection of decision trees.[10] Each tree is built independently with techniques of so-called "bagging" and random selection of input variables. The result is based on the majority vote of the classification of all trees. Random forest has proven to be a highly accurate classifier, but it has rarely been applied in medical diagnosis.

Support vector machines, proposed by Vapnik, look for the optimal separating hyperplane between 2 classes based on least squares regression.[11,12] The points lying on the boundaries of 2 classes are called support vectors. The distance between the boundaries is called margin. The middle of the margin is the optimal separating hyperplane. When a linear separator cannot be found, data points are projected into a higher-dimensional space so that they can be separated linearly. Support vector machine is considered a good classification method but its performance in diagnosing AAP is unknown.

Artificial neural networks are computational models that emulate biologic neural networks; the technique is considered powerful in modeling nonlinear relationship and has been applied increasingly in medical diagnosis and prediction.[13-18] A typical neural network consists of a number of processing units that are called perceptrons. In medicine, the most common type of neural network is a multilayer perceptron, which consists an input layer, several hidden layers, and one output layer. Each layer consists of several perceptrons. Perceptrons between layers are heavily interconnected with different weights. The weights were adjusted by training algorithms. After repetitive cycle of training, similar to human's learning from experience, the weights were adjusted to achieve the best prediction ability.

Logistic regression is a widely used standard regression model for binary data, usually used as base for comparison in machine learning studies.[19]

The object of this study is to evaluate the performance of random forest, support vector machines, and artificial neural networks in diagnosing AAP. As a benchmark, the performance of those models was compared with that of logistic regression and the Alvarado scoring system.

MATERIALS AND METHODS

Data collection. We enrolled patients who had a consultation session with a surgeon for suspected appendicitis between January 2006 and December 2008 in a regional teaching hospital. Sixteen input variables that are commonly used in diagnosing AAP were initially selected for model construction. These variables, excluding urine occult blood and hemoglobin, were selected on the bases of previous studies.[20] The data were collected by reviewing medical records. Table II shows all the initial 16 variables and the variables that were entered in each model. Patients younger than 16 or those whose records were incomplete were excluded. The diagnosis of a patient who received an operation was based on the operation note and the pathologic report. Patients who did not receive an operation were followed by telephone interview to ascertain that they were not a false negative case. The study design was approved by the local institutional review board, and subject consent was not required.

Models construction. The data was assigned randomly into 2 mutually exclusive datasets. Of the data, 75% were used to construct the models, and the remaining 25% withheld for model testing.

We used Weka 3.6 (Waikato Environment for Knowledge Analysis, University of Waikato, New Zealand) to construct random forest, support vector machines, artificial neural networks, and logistic regression models.[21] Weka provides a uniform interface to many learning algorithms and can be easily run on a personal computer. Alvarado score was also calculated for comparison of performance.

Table II. Input variables used in models construction

Variable	Type	RF	SVM	ANN	LR	Alvarado
Age (yr)	Continuous					
Sex	Binary					
Migration of pain	Binary	*	*	*	*	*
Anorexia	Binary	*	*	*		*
Nausea/vomiting	Binary					*
RLQ tenderness	Binary					*
Rebounding pain	Binary	*	*	*	*	*
Diarrhea	Binary	*	*	*		
Progression of pain	Binary	*	*	*		
Right flank pain	Binary	*	*	*	*	
Body temperature(°C)	Continuous	*	*	*	*	*
WBC(/μl)	Continuous	*	*	*	*	*
Neutrophil(%)	Continuous	*	*	*	*	*
CRP(mg/L)	Continuous				*	
Urine occult blood(+)	Discrete					
Hemoglobin(g/dl)	Continuous					

*Variables enter each method after feature selection.
RF, Random forest; *SVM,* support vector machines; *ANN,* artificial neural networks; *LR,* logistic regression.

All variables were used to construct initial random forest, support vector machines, and artificial neural networks models. Then, variable selection was done with algorithm of consistency subset evaluation (CSE) method and exhausted search method, also using Weka. The result was used to construct simpler models. Nine variables were selected for model construction to predict AAP: body temperature, migration of pain, anorexia, rebounding pain, white blood cell count, neutrophil count, diarrhea, progression of pain, and right flank pain (Table II). In each type of model, we choose the best model for comparison.

For logistic regression, we used the backward stepwise method for variable selection. Seven variables were selected for constructing a logistic regression model: body temperature, C-reactive protein level, migration of pain, rebounding pain, white blood cell count, neutrophil count, and right flank pain.

Each model was built and validated with 10-fold cross-validations.[22] We used the default setting initially, then adjusted the parameters empirically to get the optimal result. In random forest, the number of trees was set to 200. LibSVM (www.csie.ntu.edu.tw/~cjlin/libsvm/, National Taiwan University) was used to construct support vector machines models. The SVM type parameter was set to nu-SVC, the kernel type set to polynomial, the normalize parameter set to true, and the "probabilityEstimates" parameter set to true. The artificial neural networks model was a multilayer perception network with a back propagation algorithm, which means that the weights of each perceptron can be adjusted and optimized based on the feedback from the difference between the network's output and the desire output. The learning rate was set to 0.1 and momentum was set to 0.1. The training time was set to 1,000 and the "nominalToBinaryFilter" parameter set to false. The Alvarado score was calculated with 8 variables.[5] Patients with Alvarado score ≥7 were considered positive for acute appendicitis.

Performance measure. We compared the performances of the models using the testing data set to calculate accuracy (AC), sensitivity (SN), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV). We drew a receiver operating characteristic (ROC) curve for each model and calculated the area under the ROC curve (AUC) to compare the performance of these models.[23] The AUCs were estimated and compared with MedCalc 9.3 (MedCalc Software, Mariakerke, Belgium).

RESULTS

From January 2006 to December 2008, we enrolled 180 patients with a mean age of 39.4 years (range, 16–85), 85 males (47%), and 95 females (53%). Median Alvarado score was 6.15 (1–10). Of these, 126 patients received an operation, 115 (91%) of whom were diagnosed with AAP after pathologic examination of tissue samples. A total of 28 (24%) patients were cases of perforated appendicitis. Normal appendices were removed in 8 (6%) patients and 3 (2%) received an operation

Table III. Demographic and clinical characteristics of 180 patients who had a consultation session with a surgeon for suspected acute appendicitis

N (%)	*Appendicitis* *115 (64)*	*Non-appendicitis* *65 (36)*	*Overall* *180*
Alvarado score*	6.9 ± 1.6	4.9 ± 2.0	6.2 ± 2.0
Operation†			
Yes	115 (100)	11 (17)	126 (70)
No	0 (0)	54 (63)	54 (30)
Age (yr)*	40.9 ± 17.1	36.9 ± 14.0	39.4 ± 16.0
Sex†			
Male	59 (51)	26 (40)	85 (47)
Female	56 (49)	39 (60)	95 (53)
Migration of pain†			
Yes	63 (55)	20 (31)	83 (46)
No	52 (45)	45 (69)	97 (54)
Anorexia†			
Yes	45 (39)	14 (2)	59 (33)
No	70 (61)	51 (68)	121 (67)
Nausea/vomiting†			
Yes	55 (48)	26 (40)	81 (45)
No	60 (52)	39 (60)	99 (55)
RLQ tenderness†			
Yes	112 (97)	60 (92)	172 (96)
No	3 (3)	5 (8)	8 (4)
Rebounding pain†			
Yes	84 (73)	17 (26)	101 (56)
No	31 (27)	48 (74)	79 (46)
Diarrhea†			
Yes	0 (0)	3 (5)	3 (2)
No	115 (100)	62 (95)	177 (98)
Progression of pain†			
Yes	109 (95)	51 (78)	160 (89)
No	6 (5)	14 (22)	20 (11)
Right flank pain†			
Yes	0 (0)	3 (5)	3 (2)
No	115 (100)	62 (95)	177 (98)
Body temperature (°C)*	36.9 ± 0.9	36.3 ± 0.3	36.7 ± 0.9
WBC(/μl)*	14,766 ± 4,090	11,371 ± 3,256	13,540 ± 4,138
Neutrophil(%)*	71.4 ± 13.5	81.3 ± 8.5	77.7 ± 11.6
CRP(mg/L)*	49.3 ± 86.3	26.5 ± 54.5	41.0 ± 76.9
Urine occult blood(+)*	0.23 ± 0.60	0.46 ± 0.89	0.31 ± 0.72
Hemoglobin(g/dl)*	13.45 ± 1.71	13.67 ± 1.77	13.5 ± 1.75

Values in parentheses are percents.
*Data are presented as mean ± SD.
†Data are presented as mean (percent).
WBC, White blood count.

without appendectomy due to diagnosis of diverticulitis, pelvic inflammatory disease, or ventriculoperitoneal shunt infection. A total of 10 patients received abdominal CT, in which 8 patients were correctly diagnosed as AAP and 2 patients were correctly diagnosed as non-AAP. Table III summarizes the demographic and clinical characteristics of patients by final diagnosis (AAP/not AAP).

The AUC of random forest, support vector machines, artificial neural networks, logistic regression, and Alvarado were 0.98, 0.96, 0.91, 0.87, and 0.77, respectively. The AC of random forest, support vector machines, artificial neural networks, logistic regression, and Alvarado were 0.96, 0.93, 0.91, 0.82, and 0.80, respectively. The detailed performance of these models is shown in Table IV. An Alvarado score of 6 was the best cutoff value for prediction of AAP (AC = 0.80, SN = 0.84, SP = 0.69). Table V shows the result of pairwise comparison of ROC curves of the 5 models.

Table IV. Performance of random forests (RF), support vector machines (SVM), artificial neural networks (ANN), logistic regression (LR), and Alvarado score on diagnosis of acute appendicitis

	AUC	*AC*	*SN*	*SP*	*PPV*	*NPV*
RF	0.98 (0.017)	0.96	0.94	1.00	1.00	0.87
SVM	0.96 (0.027)	0.93	0.91	1.00	0.85	0.73
ANN	0.91 (0.047)	0.91	0.94	0.85	0.94	0.85
LR	0.87 (0.052)	0.82	0.91	0.62	0.85	0.73
Alvarado	0.77 (0.057)	0.80	0.84	0.69	0.87	0.64

AUC, Area under ROC curve; *AC,* accuracy; *SN,* sensitivity; *SP,* specificity; *PPV,* positive predictive value; *NPV,* negative predictive value.

Table V. Pairwise comparisons of area under receiver operating characteristic curves (AUCs) between random forest (RF), support vector machines (SVM), artificial neural networks (ANN), logistic regression (LR), and Alvarado score in predicting diagnosis of appendicitis

	RF	*SVM*	*ANN*	*LR*	*Alvarado*
RF	—	0.408	0.027*	0.011*	0.006*
SVM	0.408	—	0.094	0.043*	0.003*
ANN	0.027*	0.094	—	0.565	0.223
LR	0.011*	0.043*	0.565	—	0.51
Alvarado	0.006*	0.003*	0.223	0.51	—

*$P < .05$

Random forest performed better than artificial neural networks (P = .027), logistic regression (P = .011), and Alvarado (P = .006) in predicting AAP support vector machines performed better than logistic regression (P = .043) and Alvarado (P = .003). The figure summarizes the ROC curves of the 5 models.

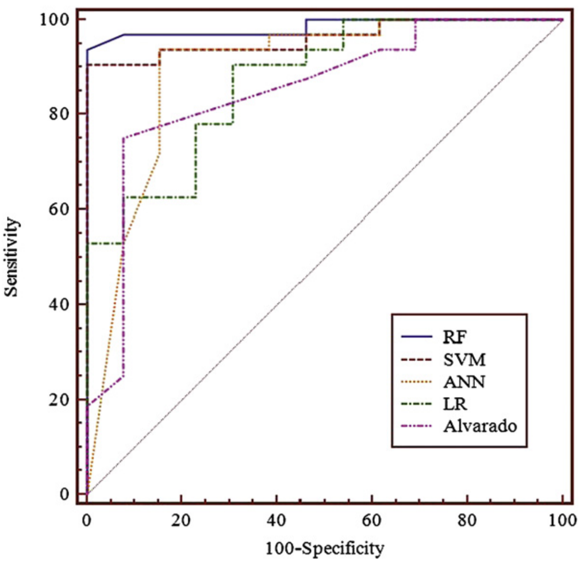

Figure. Receiver operating characteristic (ROC) curves of random forest (RF), support vector machines (SVM), artificial neural networks (ANN), logistic regression (LR), and Alvarado score for sensitivity and specificity in predicting diagnosis of appendicitis.

DISCUSSION

In our study, random forest was significantly more accurate than artificial neural networks, logistic regression, and Alvarado. Support vector machines performed better than logistic regression and Alvarado. There was no significant difference between artificial neural networks, logistic regression, and Alvarado. Although results were not significantly different, ease of model construction was much greater for random forest than for support vector machines. In random forest, only one key parameter (number of trees) is adjusted; support vector machines models need to adjust at least 4–5 parameters. Additionally, the meaning of some parameters is unfamiliar to clinicians. Considering ease of model construction, random forest is a better model for clinical use in diagnosing AAP.

Many different kinds of machine learning algorithms have been developed, making the choice of proper algorithm in clinical practice difficult. Bayesian analysis is one of the most popular algorithms and has been found useful since the early days of computer-assisted diagnoses.[24,25] In choosing a proper algorithm, performance is the most important consideration, along with ease of use and interpretation. In our literature review, we found no use of random forest or support vector machines in diagnosing AAP, although our study indicates their potential value in diagnosing AAP in the clinical setting.

The negative exploration rate of our study was 9% and the perforation rate was 24%. Compared with Hale et al,[1] we had a lower negative exploration rate and the same perforation rate. The lower negative exploration rate may due to the fact that our practice combined strategies including active observation, ultrasonography, abdominal CT, and Alvarado score. The slightly higher perforation rate may due to pre-hospital delay since only 5

(18%) patients with perforation was stayed in hospital for more than 24 hours before operation.[26,27] While these results are clinically acceptable, negative exploration rates and perforation rate should be further reduced to minimize the expense and suffering of patients.

Several authors have used artificial neural networks to develop predictive models for diagnosing AAP with varying degrees of success.[28-30] Prabhudesai et al used artificial neural networks models to diagnose AAP with highly accurate results of 100% sensitivity, 97% specificity, 96% PPV, and 96% NPV.[30] Our results for artificial neural networks were less successful: 94% sensitivity, 85% specificity, 94% PPV, and 85% NPV. All previous studies enrolled patients with right lower quadrant (RLQ) abdominal pain, whereas we enrolled those who consulted with a surgeon for suspected AAP, a condition more compatible with the clinical scenario of AAP diagnosis. In this environment, the confusing clinical presentations of patients make diagnosis more difficult and warrants use of machine based models.

Variable selection is an important step in building models. In our study, we used CSE to select variables. CSE can evaluate the consistency of variables within a subset, which allows for the use of the smallest possible subset consistent with the full variable set.[31]

RLQ tenderness was not used as an input variable in building the final random forest models, although it is an important sign of AAP. All the cases we enrolled had this symptom, making it less important in classification. Therefore, the variables selected for model construction were those with the most power of differentiation, not those with the most cause–effect relationship. Clinical physicians should be careful to understand the prerequisites of the model before using it.

The result of our study showed random forest outperformed other models in diagnosing AAP. Compared with currently used strategies, our model provides an easy, fast, low cost, and noninvasive method to accurately diagnose AAP. Weka is open-source software. The algorithm can be used directly or called by the user's own code. With a web-based user interface including the input variables, the trained model can be used at any time where Internet is available. When combined with structured electronic medical record, it is also possible for the model to provide automatic alert for clinical physicians.

Although random forest performed well in our study, its performance in other hospital settings is unproven. Theoretically, random forest does not over-fit when the independent trees are increased, which means the models built with random forest have better generalizability. But the actual influence of local data would require further research in the future. In addition, the complex algorithm is not easily understood by clinicians, which may hinder its widespread use. Prospective external validation may extend the generalizability of our model.[32] Properly used, random forest can be an effective tool in clinical decision making.

REFERENCES

1. Hale DA, Molloy M, Pearl RH, Schutt DC, Jaques DP. Appendectomy: a contemporary appraisal. Ann Surg 1997;225:252-61.
2. Colson M, Skinner KA, Dunnington G. High negative appendectomy rates are no longer acceptable. Am J Surg 1997;174:723-6.
3. Flum DR, Koepsell T. The clinical and economic correlates of misdiagnosed appendicitis: nationwide analysis. [see comment] Arch Surg 2002;137:799-804.
4. Jones PF. Suspected acute appendicitis: trends in management over 30 years. Br J Surg 2001;88:1570-7.
5. Alvarado A. A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med 1986;15:557-64.
6. Ohmann C, Yang Q, Franke C. Diagnostic scores for acute appendicitis. Abdominal Pain Study Group. Eur J Surg 1995;161:273-81.
7. Birnbaum BA, Jeffrey RB Jr. CT and sonographic evaluation of acute right lower quadrant abdominal pain. AJR Am J Roentgenol 1998;170:361-71.
8. Rioux M. Sonographic detection of the normal and abnormal appendix. AJR Am J Roentgenol 1992;158:773-8.
9. McGreevy JM, Finlayson SR, Alvarado R, Laycock WS, Birkmeyer CM, Birkmeyer JD. Laparoscopy may be lowering the threshold to operate on patients with suspected appendicitis. Surg Endosc 2002;16:1046-9.
10. Breiman L. Random forests. Machine Learning 2001;45:5-32.
11. Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag; 1995.
12. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20:273-97.
13. Drew PJ, Monson JR. Artificial neural networks. Surgery 2000;127:3-11.
14. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. Lancet 1995;346(8982):1075-9.
15. Baxt WG. Application of artificial neural networks to clinical medicine. Lancet 1995;346(8983):1135-8.
16. Lin CC, Wang YC, Chen JY, Liou YJ, Bai YM, Lai IC, et al. Artificial neural network prediction of clozapine response with combined pharmacogenetic and clinical data. Comput Methods Programs Biomed 2008;91:91-9.
17. Li Y, Liu L, Chiu W, Jian WS. Neural network modeling for surgical decisions on traumatic brain injury patients. Int J Med Inform 2000;57:1-9.
18. Lin CS, Chiu JS, Hsieh MH, Mok MS, Li YC, Chiu HW. Predicting hypotensive episodes during spinal anesthesia with the application of artificial neural networks. Comput Methods Programs Biomed 2008;92:193-7.
19. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35:352-9.

20. Andersson R. Meta-analysis of the clinical and laboratory diagnosis of appendicitis. Br J Surg 2004;91:28-37.
21. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.
22. Kohavi R, editor. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence; 1995; Montreal, Canada: Lawrence Erlbaum Associates Ltd.
23. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.
24. Gunn A. The acute abdomen: the role of computer-assisted diagnosis. Baillieres Clin Gastroenterol 1991;5(3 Pt 1):639-65.
25. de Dombal F, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. Br Med J 1972;2(5804):9.
26. Kearney D, Cahill RA, O'Brien E, Kirwan WO, Redmond HP. Influence of delays on perforation risk in adults with acute appendicitis. Dis Colon Rectum 2008;51:1823-7.
27. Bickell N, Aufses A, Rojas M, Bodian C. How time affects the risk of rupture in appendicitis. J Am Coll Surg 2006;202:401-6.
28. Eberhart R, Dobbins R, Hutton L, editors. Neural network paradigm comparisons for appendicitis diagnoses. Computer-based medical systems, proceedings of the Fourth Annual IEEE Symposium. May 12–14, 1991; Baltimore, MD.
29. Sakai S, Kobayashi K, Toyabe S, Mandai N, Kanda T, Akazawa K. Comparison of the levels of accuracy of an artificial neural network model and a logistic regression model for the diagnosis of acute appendicitis. J Med Syst 2007;31:357-64.
30. Prabhudesai SG, Gould S, Rekhraj S, Tekkis PP, Glazer G, Ziprin P. Artificial neural networks: useful aid in diagnosing acute appendicitis. World J Surg 2008;32:305-9.
31. Hall M, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering 2003;15:1437-47.
32. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol 2003;56:826-32.

Surgery is abstracted and/or indexed in *Index Medicus, Science Citation Index, Current Contents/ Clinical Medicine, Current Contents/Life Sciences,* and MEDLINE.