Annals of Colorectal Research. 2013 September; 1(2): 68-71. DOI: 10.17795/acr-12649

Published Online 2013 September 14. **Research Article**

A Pilot study on Prediction of Pouchitis in Ulcerative Colitis Patients by Decision Tree Method Versus Logistic Regression Analysis

Saeedeh Pourahmad [1], Ali Reza Safarpour [2,*], Alimohammad Bananzadeh [3], Salar Rahimikazerooni [3], Zahra Zabangirfard [3]

[1] Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, IR Iran
[2] Gastroenterohepatology Research Center, Shiraz University of Medical Sciences, Shiraz, IR Iran
[3] Colorectal Research Center, Shiraz University of Medical Sciences, Shiraz, IR Iran

*Corresponding author: Ali Reza Safarpour, Gastroenterohepatology Research Center, Shiraz University of Medical Sciences, Shiraz, IR Iran, Tel.: +98-7112357282, Fax: +98-7112307594, E-mail: asafarpour@sums.ac.ir

Received: Jun 02, 2013; **Revised:** July 07, 2013; **Accepted:** July 20, 2013

Background: Pouchitis is a non-specific inflammation of the ileal reservoir and the most frequent complication that patients experience in long time periods. Diagnosis should be made on the basis of clinical, endoscopic, and histological aspects. Prediction of pouchitis is an important issue for the physician.

Objectives: Identifying the predictive factors of pouchitis and their importance is the study's objective.

Patients and Methods: In the present study, two classifier techniques namely decision trees method and logistic regression analysis are used to help the physician for prediction of pouchitis in ulcerative colitis (UC) patients. These patients are submitted to a specific surgery. The ability of these two methods in prediction is achieved by comparison of the accuracy of the correct predictions (the minimum error rate) and the interpretability and simplification of the results for clinical experts.

Results: The accuracy rate in prediction is 0.6 for logistic regression method and 0.45 for decision tree algorithm. In addition, the mean squared error is lower for logistic regression (0.41 versus 0.48). However, the area under the ROC is more for decision tree than logistic regression (0.52 and 0.45 respectively).

Conclusions: The results are not in favor of none of these two methods. However, the simplicity of decision tree for clinical experts and theoretical assumptions of logistic regression method make the choice clear. But more sample size may be needed to choose the best model with more confident.

Keywords: Pouchitis; Ulcerative Colitis; Decision Trees; Logistic Regression

1. Background

Up to 30% of patients suffering from ulcerative colitis (UC) will ultimately need to undergo a total colectomy (1). The most frequent indications for colectomy include intractable disease and the occurrence of dysplasia or cancer in case of longstanding colitis. A total proctocolectomy with ileal pouch-anal anastomosis (IPAA) has become the surgery of choice for the "definitive" management of UC, since it avoids a permanent stoma while removing all diseased colonic mucosa (1). The most frequent long-term complication is the occurrence of pouchitis, with cumulative incidence rates varying significantly between studies (7 to 59%).

The most frequent symptoms, which characterize pouchitis, include increased stool frequency and fluidity, rectal bleeding, abdominal cramping, urgency, malaise, tenesmus, and, in the most severe cases, incontinence and fever (2).

A clinical diagnosis should be confirmed by endoscopy and histology. Prediction of pouchitis in ulcerative colitis patients is a challenging issue for the physicians, a problem which calls the classifiers in data mining techniques for help.

There are different classification methods in data mining techniques. Some of them are parametric methods (depending on underlying theoretical assumptions) such as logistic regression model, and some others are nonparametric ones (assumption free), like artificial neural networks, decision trees, K-nearest neighborhood, etc. Logistic regression is a type of predictive model in which the output variable is a binary variable like healthy or unhealthy, dead or alive, win or loss, etc. This is used for prediction of the probability of the desired event (3). Logistic regression is widely applied in the medical sciences. The binary output variable can take one of two possible values, denoted by 1 and 0 (for example, Y = 1 if a disease is present; Y = 0 otherwise). The input variables are the at-

Implication for health policy/practice/research/medical education:
Two theoretical methods help the clinical experts to predict Pouchitis in Ulcerative Colitis Patients.

tributes involved in prediction of the probability of the desired event (Y = 1) denoted by X = (x_1 , x_2 ,, . . ., x_n). Logistic regression method models the relations between these variables through the following Formula :

$$\log\left\{\frac{P(Y=1)}{1-P(Y=1)}\right\}=b_0+b_1x_1+\cdots+b_nx_n$$

Formula

Where P stands for probability, b_0 is called the "intercept" and b_1, b_2, . . . are called the "regression coefficients" of x_1, x_2, . . . respectively. Each of the regression coefficients describes the importance of corresponding input attribute on output.

As mentioned earlier, logistic regression method depends heavily on its theoretical assumptions, whereas the real dataset seldom follows the underlying theoretical assumptions of parametrical modeling methods, such as with clinical datasets. The variable nature of biological data and their vague relation does not consist with their ideal assumptions. In comparison, nonparametric methods in data mining techniques, by use of learning process from a set of existent prototypes, attract the attention of researchers in different fields. In these methods, without any specific underlying assumption, the relation among a large part of a dataset (training set) is discovered and the model's parameters are estimated in such a way that the error prediction is minimized. Then, the power of model's prediction is evaluated by the other part of dataset (testing set). Decision tree is one of these methods. Decision tree is a typical method for the classification of objects into decision classes (4). A decision tree classifier is a function stated as following:

$dt{:}dom(X_1)\times dom(X_2)\times\ldots\times dom(X_n)\rightarrow dom(Y)$

In which X_1, X_2 , . . ., X_n are input attributes and Y is the output, where X_i has domain $dom\ (X_i)$ and Y has domain $dom\ (Y)$. A decision tree is a directed, acyclic graph T in a form of a tree. Each node in a tree has either zero or more outgoing edges. If a node has no outgoing edges, then it is called a decision node (a leaf node); otherwise, a node is called a test node (or an attribute node). Each decision node N is labeled with one of the possible decision classes, $Y\in\{Y_1, Y_2\}$. Each test node is labeled with one input attribute, $X_i\in\{X_1, X_2, \ldots, X_n\}$ i.e. called the splitting attribute. Each splitting attribute X_i has a splitting function f_i associated with it. The splitting function f_i determines the outgoing edge from the test node, based on the attribute value X_i of an object O in question. It is in the form of $X_i\in Y_i$ where $Y_i\subset dom(X_i)$ if the value of the attribute X_i of object O is within Y_i, then the corresponding outgoing edge from the test node is chosen (5).

All this leads to some practical rules. In this way, it helps the person to make a decision. Decision tree plays a vital role in medical diagnosis (6). The derived rules from this method help the physician to decide about a patient based on his/her own clinical observations. Among all classifiers in data mining techniques, decision tree is preferred in medical researches as it provides human readable rules of classification, it is easy to interpret, and yields better accuracy. Furthermore, construction of decision tree is fast (6). These advantages encourage us to apply this method for prediction of pouchitis in ulcerative colitis patients. In the present study, the performance of this method is compared with logistic regression analysis in a real clinical data set. Since there is low number of patients and their clinical observations at hand, this study is done as a pilot. Obviously, to derive more reliable rules a larger sample size of patients is required.

2. Objectives

The aim of this study is to identifying the predictive factors of pouchitis and their importance.

3. Patients and Methods

All 43 patients who underwent a proctocolectomy with IPAA for ulcerative colitis (UC) at the Nemazi and Faghihi Hospitals in Shiraz University of medical sciences (tertiary referral centre) between 2001 and 2012 were identified through the pre and post operation data.

Clinical charts of all patients were reviewed to trace clinical, endoscopic, and histologic characteristics. The occurrence of pouchitis is considered as the binary output and 85 related attributes are used as the inputs or risk factors.

Logistic regression analysis and decision tree method are used for data analysis (7). Weka software is used for both types of data analysis. C4.5 algorithm is chosen for constructing decision tree. This algorithm is called J48 in Weka and it is chosen for its ability to handle binary output, both nominal and numeric input attributes, and missing values. The datasets are split into 70 percent for training set (constructing the tree) and the remainder percent for testing. To evaluate these two modeling methods, 10-fold cross validation procedure and area under the ROC (Receiver Operating Curve) are used (8).

4. Results

Some descriptive attributes and the frequency of pouchitis in the patients sample are summarized in Table 1. In addition of those, the factors such as onset of disease symptoms, extra intestinal manifestations, complications, autoimmune disease, microscopic findings, WBC, HB, Na, K,. . . at time of surgery, instruments during surgery, some post operation, etc. are considered as the input risk factors. Logistic regression analysis and J48 decision tree algorithm are used to predict the occurrence of pouchitis in these patients. The results for both methods are shown in Table 2. However, the large number of categorical variables and their parameters may make some problems in the estimating process in logistic regression method such that the number of parameters is equal or even more than the observations in this pilot study. Therefore, there is no possibility to enter all the variables to the model. Consequently, the significant variables with the occurrence of pouchitis are chosen from univariate analysis (by chi-square test) and then enter to the model simultanously. Unfortunately, the variables which are important clinically are not entered to the model and therefore the formulation has not been shown here.

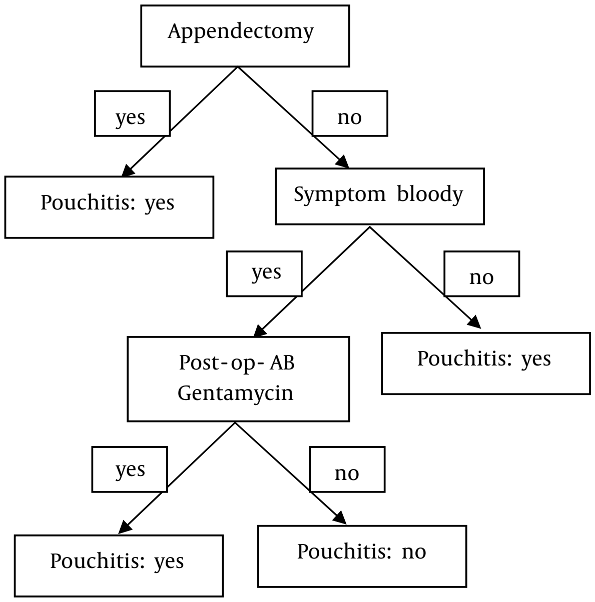

Figure 1. Decision Tree Derived from the Information of 43 Ulcerative Colitis Patients

The area under ROC near to 1 and far from 0.5 and also the mean squared error near to 0 is desired. The more the area under ROC is far from 0.5, the more is the model ability to distinguish between patient with and without pouchitis. Furthermore, the mean squared error near 0 detects the low error in prediction. Accordingly, findings show weak results for both methods. This fact is also detected in accuracy rate for both methods. The main reason for this result is the low sample size compared to the large number of attributes in the study. The derived rules from the constructed decision tree are summarized in Table 3. Furthermore, Figure 1 features these rules in flowchart view. These rules are evaluated by the testing set (30 percent of the observations which were not used in model construction). However, with larger sample size more reliable rules can be derived.

Table 1. Some Descriptive Attributes of our Patients' Samples

Attribute	Percent, %
Gender	
Male	44
Female	56
Marital Status	
Single	32.4
Married	67.6
Education level	
Illiterate	14.7
Primary	38.2
High school	23.5
University	14.8
Post graduate	8.8
Birth place	
Shiraz	39.3
Other cities in Fars	60.7
Family history of Pouchitis	
Yes	35.3
No	64.7
Under surgery	
Yes	29.4
No	70.6
Type of surgery	
Laparoscopy	61.8
Laparatomy	38.2
Malignancy	
No	81.3
Rectum	16.3
Liver and gall bladder	2.4 (1 case)
Occurrence of Pouchitis	
Yes	47.1
No	52.9

Table 2. Brief Results of two Applied Modeling Method on the Clinical Data Set

Method	TP rate*	FP rate*	Accuracy rate*	Area under ROC	Mean squared error
Logistic regression	0.6	0.43	0.6	0.45	0.41
Decision tree (J48 algorithm)	0.4	0.48	0.45	0.52	0.48

*Abbreviations: TP, True Positive; FP, False Positive; Accuracy: (TP+TN) / (TP+TN+FP+FN)

Table 3. The Derived Rules from the Trained Decision Tree based on Clinical Findings of 43 Ulcerative Colitis Patients

The rules	The Occurrence of Pouchitis
Appendectomy, yes	Yes
Appendectomy, no and symptom (onset of disease) bloody, no	Yes
Appendectomy, no and symptom (onset of disease) bloody, yes and Post-op-AB [*] Gentamycin, yes	No
Appendectomy, no and symptom	Yes
(onset of disease) bloody, yes and Post-op-AB Gentamycin, no	No

[*] Abbreviation: Post-op-AB; Post Operation Antibiotics

5. Discussion

The attempt in the present study was to predict pouchitis in ulcerative colitis patients, the issue which its diagnosis is accompanied with some doubt for the physicians (1). In other words, there are no clear laboratory tests or wholly accepted diagnosis criteria for occurrence of this phenomenon in ulcerative colitis patients. Therefore, we tried to model the relations among clinical findings of samples of these patients by two classifier methods. One method was more theoretic with ideal underlying assumptions to be submitted before use, namely logistic regression analysis, and the other one was more practical and flexible to the real data circumstances but required a larger data set to be trained, namely decision tree technique (5). Unfortunately, the number of available patients for this research was low and consequently this study was done as a pilot. Therefore, the derived results could not reveal the real performance of these two methods. Nevertheless, the simplicity and interpretability of decision tree was obvious from the results of this study. Since the underlying theoretical assumptions of logistic regression analysis had not been checked before modeling procedure, the use of this method was not applicable and should be expressed with more care. Furthermore, logistic regression method cannot handle the large amount of categorical variables especially with low sample size (5). Therefore, although the significant variables in primary univariate analysis were entered to logistic regression model, the clinical important variables did not remain in the final model and consequently, the estimated model was not valuable for clinical experts. Whereas decision tree algorithm lead to practical rules which led the user to the decision. Obviously, with larger sample size, the more accurate results can be achieved for both methods.

Acknowledgements

With great thanks to the colorectal research center staff and for their attempt in data gathering process.

Authors’ Contribution

Saeedeh Pourahmad: Data analyzer and The main writer of the manuscript. Ali Reza Safarpour, Alimohammad Bananzadeh and Salar Rahimikazerooni: Clinical research group. Zahra Zabangirfard: Data gathering secratory.

Financial Disclosure

The authors declare no conflict of interest.

Funding/Support

This work is financially supported by Colorectal Research Center, Shiraz University of Medical Sciences, Shiraz, Iran.

References

1. Pardi DS, D'Haens G, Shen B, Campbell S, Gionchetti P. Clinical guidelines for the management of pouchitis. *Inflamm Bowel Dis.* 2009;**15**(9):1424-31.
2. Yu ED, Shao Z, Shen B. Pouchitis. *World J Gastroenterol.* 2007;**13**(42):5598-604.
3. Srivatsa SK. Evaluation of Logistic Regression and Neural Network Model With Sensitivity Analysis on Medical Datasets. *Int J Computer Sci Secur (IJCSS).* 2011;**5**(5).
4. Kokol P, Pohorec S, Štiglic G, Podgorelec V. Evolutionary design of decision trees for medical application. *Wiley Interdisciplin Rev: Data Mining and Knowledge Discovery.* 2012;**2**(3):237-54.
5. Pourahmad S, Azad M, Paydar S. Prediction of Malignancy in Suspected Thyroid Tumour Patients by Three Different Methods of Classification in Data Mining. 2012; Available from: Available from: http://airccj.org/CSCP/vol2/csit2501.pdf.
6. Lavanya D. Performance Evaluation of Decision Tree Classifiers on Medical Datasets. *Int J Comput App.* 2011;**26**(4):1.
7. Endo A, Shibata T, Tanaka H. Comparison of Seven Algorithms to Predict Breast Cancer Survival. *Biomed Soft Comput Human Sci.* 2008;**13**(2):11-16.
8. Long WJ, Griffith JL, Selker HP, D'Agostino RB. A comparison of logistic regression to decision-tree induction in a medical domain. *Comput Biomed Res.* 1993;**26**(1):74-97.