RESEARCH ARTICLE

Improving the Accuracy of Early Diagnosis of Thyroid Nodule Type Based on the SCAD Method

Hadi Raeisi Shahraki[1], Saeedeh Pourahmad[1]*, Shahram Paydar[2], Mohsen Azad[3]

Abstract

Although early diagnosis of thyroid nodule type is very important, the diagnostic accuracy of standard tests is a challenging issue. We here aimed to find an optimal combination of factors to improve diagnostic accuracy for distinguishing malignant from benign thyroid nodules before surgery. In a prospective study from 2008 to 2012, 345 patients referred for thyroidectomy were enrolled. The sample size was split into a training set and testing set as a ratio of 7:3. The former was used for estimation and variable selection and obtaining a linear combination of factors. We utilized smoothly clipped absolute deviation (SCAD) logistic regression to achieve the sparse optimal combination of factors. To evaluate the performance of the estimated model in the testing set, a receiver operating characteristic (ROC) curve was utilized. The mean age of the examined patients (66 male and 279 female) was 40.9 ± 13.4 years (range 15- 90 years). Some 54.8% of the patients (24.3% male and 75.7% female) had benign and 45.2% (14% male and 86% female) malignant thyroid nodules. In addition to maximum diameters of nodules and lobes, their volumes were considered as related factors for malignancy prediction (a total of 16 factors). However, the SCAD method estimated the coefficients of 8 factors to be zero and eliminated them from the model. Hence a sparse model which combined the effects of 8 factors to distinguish malignant from benign thyroid nodules was generated. An optimal cut off point of the ROC curve for our estimated model was obtained (p=0.44) and the area under the curve (AUC) was equal to 77% (95% CI: 68%-85%). Sensitivity, specificity, positive predictive value and negative predictive values for this model were 70%, 72%, 71% and 76%, respectively. An increase of 10 percent and a greater accuracy rate in early diagnosis of thyroid nodule type by statistical methods (SCAD and ANN methods) compared with the results of FNA testing revealed that the statistical modeling methods are helpful in disease diagnosis. In addition, the factor ranking offered by these methods is valuable in the clinical context.

Keywords: Thyroid nodule type - benign - malignant - early diagnosis - ROC curve - SCAD

Asian Pac J Cancer Prev, **17 (4)**, 1861-1864

Introduction

Determining the type of nodule before the surgery has a great importance in diagnosis and treatment of the diseases. In some cases like thyroid disease the type of nodule (benign or malignant) determines the type of surgery (Pourahmad et al., 2015). Thyroid nodule is a typical problem in human society. Currently, fine needle aspiration (FNA) is the only effective minimally invasive method for the differential diagnosis of thyroid nodules. But it is subject to sampling and analysis uncertainties (Zhang and Berardi, 1998). In addition, it depends on the operator expertise. Therefore, the sensitivity and specificity values of this test may not be satisfactory. Thus, a powerful modeling method is urgently needed. However, some previous studies attempted to do this differently (Finley et al., 2004; Hong et al., 2009; Pourahmad et al., 2015).

To model the relationships among the factors involved in the diagnosis of diseases, there are various methods in feature selection and modeling technique (Ma and Huang, 2008). Conventional and penalized logistic regression, random forests, regression trees, artificial neural networks (ANN), decision trees, fuzzy models and discriminating methods by gene expression data are some recently applied methods in medical researches (Ghosh and Chinnaiyan, 2005; Mendonça et al., 2007; Yang et al., 2009; Lin et al., 2011; Yan et al., 2011; Talhaa and Al-Elaiwi, 2013; Mansiaux and Carrat, 2014).

Certainly, to achieve a model with minimum error and attenuate any possible biases, it is reasonable to use the whole potential factors that may be important in diagnosis. Hence, we encounter to high dimensional data. However, some of these factors are redundant. Utilizing them brings some complexity in the model without any significant improvement in its performance. Therefore, it is desirable to determine a sparse linear combination of factors with really effective detection. The smoothly clipped absolute

[1]*Department of Biostatistics,* [2]*Trauma Research Center, Department of Surgery, Shiraz University of Medical Sciences, Shiraz,* [3]*Mother and Child Welfare Research Center, Hormozgan University of Medical Sciences, Bandar Abbas, Iran *For correspondence: pourahmad@sums.ac.ir*

deviation (SCAD) method is a well known penalized regression which can be applied in a huge number of variables (Fan and Li, 2001).

Accordingly, our attempt in the present study was to select a subset of factors so that an appropriate linear combination of them can distinguish benign from malignant thyroid nodules, using the Smoothly Clipped Absolute Deviation (SCAD) method introduced by Fan and Li (Fan and Li, 2001).

Materials and Methods

In a prospective study from 2008 to 2012, all patients who had been admitted to Rajai and Namazi hospitals in Shiraz, Iran, for surgery of the thyroid nodule were recruited. FNA test was performed for all the patients and all of them underwent surgery. Effective characteristics such as, gender, age, type and growth of the thyroid gland, FNA test result, duration of the disease, family history of the disease and cancer, size of the right and left thyroid gland, size of the nodules in the left and right thyroid glands and their volumes were used as the predictor variables in the analysis. More details about the study protocol and data collection can be found in (Pourahmad et al., 2015).

SCAD method was used to identify the factors which affect the thyroid nodule types. By adding a penalty function in the maximum likelihood of the model, SCAD forces some of the coefficients shrink to zero. Variable selection and coefficients estimation performs simultaneously in this method, which leads to increase precision. In addition, SCAD has oracle property, i.e. it estimates both zero and non-zero factors truly, with a probability tending to one (Fan and Li, 2001).

Suppose that we have n observation and $\beta^T=(\beta_1, \beta_2, \ldots, \beta_P)$ is a vector of coefficients factors. For logistic regression, SCAD is defined as follows (Eq. 1& 2):

$$L(\beta;\lambda)=l_n(\beta)+\lambda\Sigma_p(\beta) \qquad \text{Eq. (1)}$$

$$p(\beta)=I(|\beta|\leq\lambda)+\frac{(3.7\lambda-|\beta|)_+}{2.7\lambda}I(|\beta|>\lambda) \qquad \text{Eq. (2)}$$

which $l_n(\beta)$ is the traditional maximum likelihood estimator and λ is a positive constant called tuning parameter (13). The amount of shrinkage depends on λ which is estimated by cross-validation method (Fan and Li, 2001; Shahraki et al., 2014).

In this article, we randomly split the dataset into a training set and testing set at a ratio of 7:3. Training set (242 observations) was used for estimation and variable selection using SCAD and obtaining a linear combination of factors. In this step, 10-fold cross-validation method was implemented to obtain the optimal lambda with minimal error. Finally, the estimated model evaluated in the testing set (103 observations) via receiver operating characteristic (ROC) curve. All the analyses were performed, using SPSS version 15 and ncvreg package in R.3.1.2 software.

Results

From 345 patients who were enrolled in the present study, 182 cases (52.8%) had two or more thyroid nodules. The mean age of the participated patients (66 male and 279 female) was 40.9 ± 13.4 years (range 15-90 years). 54.8% of them (24.3% male and 75.7% female) had benign thyroid nodule and 45.2% (14% male and 86% female) had malignant thyroid nodule while FNA test showed that 50.1% of the patients had benign thyroid nodules.

The performance indices of the estimated model for nodule type diagnosis in testing dataset are shown in Table 1. This Table also represents the selected factors and their coefficients by the SCAD logistic regression model (Table 1 & Eq. (3)). The amounts of 10-fold cross-validation errors are shown in Figure 1. Accordingly, the minimum

Table 1. Characteristics of the SCAD Logistic Regression Model for Thyroid Nodule Type

Variable in the model	Feature	SCAD Coefficient	Performance indices
X1	Cancer family history	1.94	
X2	FNA test	0.53	
X3	Multiple nodules	-0.3	AUC= 0.77
X4	Maximum size of left nodule	0.28	(0.68, 0 .85)
X5	Maximum size of right nodule	0.07	
X6	Recent enlargement	-0.05	Sensitivity = 0.72 (0.57, 0.84)
X7	Sex	-0.02	
X8	Maximum size of left lobe	-0.005	
X9	Age	0	Specificity=0.75 (0.62, 0.86)
X10	Disease family history	0	
X11	Duration of disease	0	
X12	Volume of left lobe	0	Accuracy=0.74
X13	Volume of left nodule	0	(0.69, 0.78)
X14	Maximum size of right lobe	0	
X15	Volume of right lobe	0	
X16	Volume of right nodule	0	

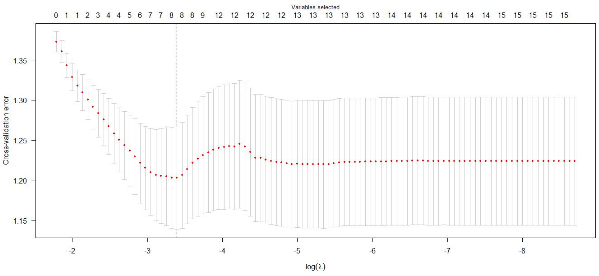

Figure 1. Error Values in Predicting the Malignant Thyroid Nodule Based on the Number of Factors Included in the SCAD Model

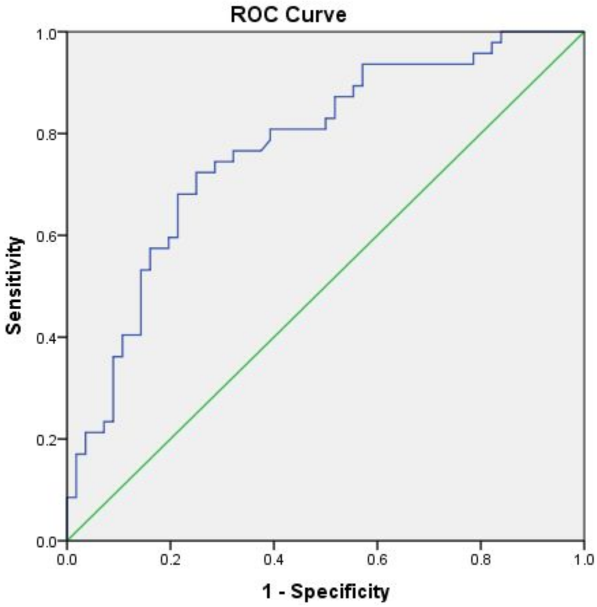

Figure 2. ROC Curve for the Estimated SCAD Model

amount of error is occurred when 10 factors were selected.

SCAD logistic regression introduced the history of cancer in family and result of FNA test as the most important factors in early diagnosis of the type of the thyroid nodule. Many less important features were removed from the model (zero coefficients in Table 1). Regarding the 8 eliminated variables, optimal combination of factors (SCAD model) was obtained as follows (Eq. (3)):

$$\log\left(\frac{p_i}{1-p_i}\right) = 1.94X_1 + 0.53X_2 - 0.30X_3 + 0.28X_4 + 0.07X_5 - 0.05X_6 - 0.02X_7 - 0.005X_8 \qquad \text{Eq. (3)}$$

Using Eq. (3), we calculated p_i (probability of malignant thyroid nodule) for the testing set data. Optimal cut off point of the ROC curve for this equation was obtained 0.44. The AUC of our estimated model was equal to 77% (95% CI: 68% - 85%), which was much greater than single factors (Figure 2). Sensitivity, specificity, positive predictive value and negative predictive value for this model were 72%, 75%, 71%, and 76%, respectively.

Discussion

Usually, the decision for thyroidectomy in patients with thyroid nodule problem is based on the result of FNA test. Indeed, if the test detects a benign nodule, then the right, left or subtotal lobectomy is applied. Otherwise, total lobectomy is performed (Pourahmad et al., 2015). Although FNA test is the only effective method for differential diagnosis of thyroid nodules, clinical texts reported some mistakes in decision based on this test (Pourahmad et al., 2015). The accuracy rate for FNA test in the present data set was 63% (The result is not shown here). Therefore, due to the importance of the early diagnosis of thyroid nodule type, the present study attempted to improve the accuracy of early diagnosis of the type for thyroid nodule based on SCAD method. Some previous studies have followed this purpose by different methods (Finley et al., 2004; Hong et al., 2009; Pourahmad et al., 2015). But, the simulated data showed a high accuracy rate in classification of this method (Yan et al., 2011). Our estimated model reached a sparse linear combination of factors which improves the accuracy of diagnosis. The results of the ROC curve in the present study also revealed both high sensitivity and specificity values for SCAD method (Figure 2). Furthermore, the type of penalty function used in the present model is more powerful than some other penalized functions like LASSO (Lin et al., 2011).

In comparison with the previous study on this data using artificial neural networks (ANN) method, among thirteen batch learning algorithms, the maximum accuracy rate and AUC were found for GD (Gradient Descent) learning algorithm (Accuracy= 0.71 and AUC=0.837) (Pourahmad et al., 2015). Accordingly, SCAD method in the present study represented higher accuracy and lower AUC than ANN model. In addition, the order of ten first variables in determining the type of nodule assigned by ANN method was as follows: Multiple nodule, Sex, Rapid enlargement of the Thyroid gland, Age, Size of Thyroid nodule, Family history of Thyroid disease, Family history of cancer and Maximum size of Thyroid nodule. Comparing with results summarized on Table 1, the order of variables with non-zero coefficients by SCAD model was further confirmed by the physicians (Table 1). Indeed, the family history of cancer and FNA test result are two most important factors in malignancy of thyroid nodules in clinical texts. Furthermore, the important role of maximum size of nodules compared with their volumes and the sizes of lobes in malignancy prediction was another interesting result derived from our model. The statistical basis of SCAD method may be the cause of this superiority over ANN method (Fan and Li, 2001). Furthermore, the medium sample size (equal or more than 100) provides acceptable accuracy rate for SCAD method, while a large sample size is required for ANN method. However, there are some limitations concerning the SCAD method in practice. For example, adding or removing a variable leads to sensible changes in the variables' coefficients. In addition, like other penalized methods, SCAD suffers from missing data and outliers.

In conclusion, an increase of 10 percent and a greateraccuracy rate in early diagnosis of thyroid nodule type by statistical methods (SCAD and ANN methods) compared with the results of FNA testing revealed that the presetnly adopted statistical modeling methods are helpful in disease diagnosis. In addition, the factor ranking offered by these methods would be expected to be valuable in the clinical context.

Acknowledgements

This work was supported by the grant number 91-6167 from Shiraz University of Medical Sciences Research Council. The authors are thankful to the Trauma Research Center staff for their assistance in data gathering and Dr. HR. Abbasi for his collaboration in cancer diagnosis. The authors would like to thank Shiraz University of Medical Sciences, Shiraz, Iran and also Center for Development of Clinical Research of Nemazee Hospital and Dr. Nasrin Shokrpour for editorial assistance.

References

Fan J, Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Statistical Associat*, **96**, 1348-60.

Finley DJ, Zhu B, Barden CB, et al (2004). Discrimination of benign and malignant thyroid nodules by molecular profiling. *Ann Surg*, **240**, 425.

Ghosh D, Chinnaiyan AM (2005). Classification and selection of biomarkers in genomic data using LASSO. *Bio Med Res Int*, **2005**, 147-54.

Hong Y, Liu X, Li Z, et al (2009). Real-time ultrasound elastography in the differential diagnosis of benign and malignant thyroid nodules. *J Ultrasound Med*, **28**, 861-7.

Lin H, Zhou L, Peng H, et al (2011). Selection and combination of biomarkers using ROC method for disease classification and prediction. *Canadian J Statistics*, 39, 324-43.

Ma S, Huang J (2008). Penalized feature selection and classification in bioinformatics. *Briefings Bioinformatics*, **9**, 392-403.

Mansiaux Y, Carrat F (2014). Detection of independent associations in a large epidemiologic dataset: a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with H1N1pdm influenza infections. *BMC Med Res Methodol*, **14**, 99.

Mendonça LF, Vieira SM, Sousa J (2007). Decision tree search methods in fuzzy modeling and classification. *Int J Approximate Reason*, **44**, 106-23.

Pourahmad S, Azad M, Paydar S (2015). Diagnosis of malignancy in thyroid tumors by multi-layer perceptron neural networks with different batch learning algorithms. *Global J Health Sci*, **7**, 46.

Shahraki H, Salehi A, Zare N (2014). Survival prognostic factors of male breast cancer in Southern Iran: a LASSO-Cox regression approach. *Asian Pac J Cancer Prev*, **16**, 6773-7.

Talhaa M, Al-Elaiwi A (2013). Enhancement and classification of mammographic images for breast cancer diagnosis using statistical algorithms. *Life Sci J*, **10**, 764-772.

Yan F-R, Lin J-G, Liu Y (2011). Sparse logistic regression for diagnosis of liver fibrosis in rat by using SCAD-penalized likelihood. *Bio Med Res Int*, **8**, 875309.

Yang F, Wang H-z, Mi H, et al (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinformatics*, **10**, 22.

Zhang GP, Berardi VL (1998). An investigation of neural networks in thyroid function diagnosis. *Health Care Management Sci*, **1**, 29-37.