Comparison of Bayesian Network and Binary Logistic Regression Methods for Prediction of Prostate Cancer

Selen Bozkurt
Akdeniz University, Faculty of Medicine Department of Biostatistics and Medical Informatics,
Antalya, Turkey

Asli Uyar
Akdeniz University, Faculty of Medicine Department of Biostatistics and Medical Informatics,
Antalya, Turkey

Kemal Hakan Gulkesen
Akdeniz University, Faculty of Medicine Department of Biostatistics and Medical Informatics,
Antalya, Turkey

***Abstract*—Prostate cancer is one of the most common cancers in men. Luckily, Serum PSA level, age, digital rectal examination (DRE), and clinical symptoms are helpful for early detection of this tumor. The aim of this study was to examine and compare the methods used for improving the diagnostic accuracy of serum PSA in Turkey, a country with low incidence of prostate cancer. The predictors used for early detection of prostatic carcinoma were identified by both Logistic Regression and Bayesian networks. The results of the methods were compared in terms of predicting performance and advantages**

Keywords-component; Prostate Cancer, Bayesian Networks, Logistic Regression

I. INTRODUCTION *(HEADING 1)*

Prostate cancer is one of the most common cancers in men [1-2]. In the United States, approximately 32,000 men have died of the prostate tumor as a part of the 217,730 men who were detected with that disease in 2010 [2]. Luckily, there are some predictors used for early detection of prostatic carcinoma. Serum PSA level, age, digital rectal examination (DRE), and clinical symptoms are helpful for early detection of this tumor [3-4].

When a patient is suspected to have a prostate tumor, a biopsy from the prostate is advised by the physician. Sometimes, because of the presence of strong indicators such as a very high serum PSA level, the decision for biopsy is easy. However, when the findings are in the grey zone, physicians and patients have to make a choice between the risk of missing an early detection of a tumor and the risk of an unnecessary biopsy [5]. There are several studies [3-4, 6] trying to establish methods to improve the sensitivity and specificity of different examinations in these grey zone cases. Generally, a PSA level between 4-10 ng/ml is accepted as having a 70% sensitivity and a 70% specificity [5, 7]. Since 1989, several concepts to further improve the diagnostic accuracy of PSA have been developed with the aim of avoiding unnecessary biopsies.

Likewise, the aim of this study was to examine and compare the methods used for improving the diagnostic accuracy of serum PSA in Turkey, a country with low incidence of prostate cancer. The predictors used for early detection of prostatic carcinoma were identified by both Logistic Regression and Bayesian networks. The results of the methods were compared in terms of predicting performance and advantages.

II. METHOD

A. Study Population

All the transrectal ultrasound (TRUS)-guided prostatic biopsy cases who were admitted to Akdeniz University Hospital, Department of Urology, between January 2000-April 2007 was retrospectively evaluated. TRUS-guided biopsy could be performed only in Akdeniz University in Antalya district which has a population around 1,800,000. Original dataset included medical records of 1453 patients, samples including missing variables were excluded from the study and the remaining 983 cases, whose serum PSA level ranged between 0.05 and 1000, have been analyzed. In patients with multiple biopsies only the first one was included in the study.

B. Dataset characteristics

Akdeniz University Hospital Information System (HIS) and the patients' medical records of the Department of Urology were used as data sources. HIS contains the demographic information about all patients, laboratory data of the last eight years, and pathology reports.

C. Statistical Analysis

Logistic Regression Analysis

978-1-4244-9352-4/11/$26.00 ©2011 IEEE

Forward and backward conditional stepwise methods were applied because of multi-collinearity of the variables. Entry and removal criteria were 0.5 and 0.10 respectively. Hosmer-Lemeshow goodness of fit test was also performed for each model. The forward method was selected because of better performance in Hosmer-Lemeshow goodness of fit test.

Bayesian Networks

A Bayesian network is a directed acyclic graphical model that represents the joint probability distribution over a collection of random variables [8]. Each node corresponds to a random variable, and edges are the direct correlations between the variables. A probability distribution for each node is computed depending on its parents. A Bayesian network defines a unique joint probability distribution over the set of random variables Xi in the network given by (1):

$$P(X_1, \cdots, X_n) = \prod P(X_i \mid \prod X_i) \qquad (1)$$

Defining the structure of the Bayesian network is an important research interest. The network structure can be learned from data or can be defined based on domain knowledge. There are various algorithms to learn network structure from training data. In order to evaluate the performance of the Bayesian network in the domain, initially we have constructed the network based on domain knowledge (Figure 1).

D. Training and Testing Strategy

Two-thirds of the dataset was randomly selected for establishing a predictor model and the remaining one-third was utilized for testing. This random splitting has been performed using stratification principle in order to ensure that the proportions of positive and negative classes remain the same in both training and test sets as in the original dataset. The random two-thirds, one-third partitioning of dataset into training and test sets has been repeated 10 times in order to overcome sampling bias. The presented results are the mean and average of these 10 repetitions.

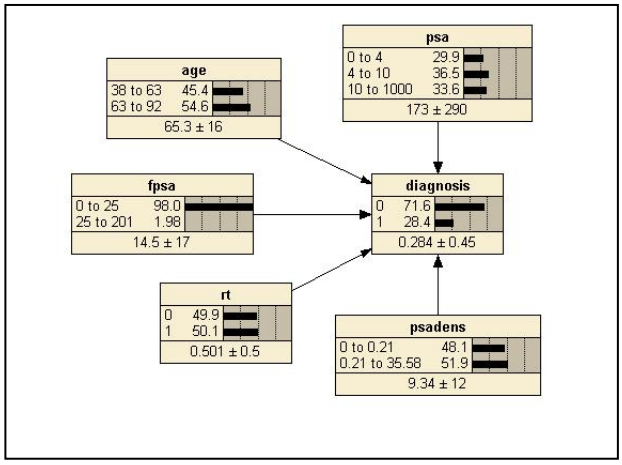

Figure 1. A simple network based on domain knowledge

Both of the analyses were performed using WEKA machine learning software [8]. In addition, Netica software was used for the visualization of Bayesian network based on domain knowledge. The predictive performances of both methods were evaluated on the test set in terms of sensitivity, specificity and Receiver Operating Characteristics (ROC) Area Under Curve (AUC). We computed the Mann-Whitney U test to check if the difference between the two models was significant.

III. RESULTS

The general patient characteristics are shown in Table 1. In addition, according to results of digital rectal examination, 54.1 % (532) of the cases were suspected and 45.9 % (451) of them were normal.

For the logistic regression model, while input variables were age, PSA, Free PSA, PSA density (PSAD) and digital rectal examination result (DRE), only age (p=0.001), PSA density (p=0.0001) and rectal examination result (p=0.0001) were found as significant determinants for prostate cancer with an ROC AUC of 0.775 (p=0.0001).

The results of prediction of Bayesian network over 10 fold cross validation is given in Table 1 in terms of accuracy, true positive (TP) rate (sensitivity) and false positive (FP) rate. In clinical side, TP rate indicates the correct prediction of cancer development while FP rate represents erroneous positive predictions. Minimization of FP rate should be favored in real world application.

TABLE I. GENERAL CHARACTERISTICS OF THE CASES

Characteristics	*n*	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
Age (years)	983	64	8.61	38	92
PSA (ng/ml)	983	13.7	43.51	0.05	1000
Free PSA	983	2.6	9.24	0.03	201
PSA density	983	0.52	2.01	0	35.58

TABLE II. PERFORMANCE COMPARISON OF TWO DIFFERENT METHODS

Bayesian Network	*AUC*[a]	*TPR*	*FPR*
Mean±SD	0.75±0.01	63.3±5.06	25.2±5.72

a. Area Under Curve (AUC)

Mann Whitney U test show that the Logistic Regression model produce significantly different results than Bayesian Network Model (p=0.0001). Logistic regression model (AUC=0.775) performs better in predicting adequate tumor development than Bayesian network model (AUC=0.75).

IV. DISCUSSION

Prediction studies with the help of PSA in prostate cancer is important because PSA is the first serum marker which can be used in cancer screening and the experience from these studies can be used in future possible cancer markers. Despite the

importance of predicting prostate cancer, so far this problem has not been sufficiently studied.

To compare the Bayesian network analysis results with conventional statistics, we applied binary logistic regression analysis on the same dataset. Logistic regression analysis had higher AUCs than Bayes network analysis. LR analysis of the cases showed that, in the LR set, PSA and free PSA could not take place in multivariate model, but age, PSAD, and DRE were significant. In a similar previous study, age, PSAD, DRE, and transrectal ultrasonography (TRUS) were significant variables [7]. Another study which did not contain f/tPSA in the analysis revealed only prostate volume as significant variable, a quite different result from our study [10]. However, the calculation of LR formula in the daily practice is not very easy. This method can be used for decision support systems but cannot be accepted as a simple rule that helps the physician.

On the other hand, there are two main advantages of Bayesian network in modeling prostate cancer prediction: first, a Bayesian network can be used to learn cause effect relationships, and second, it is an ideal representation for combining prior knowledge and data. It is also reported that Bayesian networks are better suited to capture the complexity of the underlying decision-making process, taking into account the many (inter)dependencies among the variables [11]. There have been a few previous efforts at developing prognostic Bayesian networks in cancer [12-13]. Although the uncertainty present in predicting the future, Bayesian network formalism is well suited to this task, and the usefulness of Bayesian networks for medical prognostication is clearly recognized [11].

However, Bayesian networks may link more variables in complex, direct and indirect ways, making interpretation more problematic. In contrast, decision rules can easily be derived from decision trees and provide a simpler and more direct interpretation tool for physicians. For this reason decision trees are popular within the medical field [7].

In a recent study it is declared that a Bayesian network is well suited to assist with prognosis in intensity modulated radiation therapy plan selection because physicians draw upon many sources of information to predict an outcome. Clinical trials report on the predictive power of a set of variables chosen before the study begins. Retrospective studies mine past results for predictive variables, many of which are different than those covered in clinical trials. Physicians must combine these sources and they supplement that information with their subjective degree of belief in outcomes [14].

The present study included 983 biopsied patients. A small fraction of the patients in the included data set were deficient of TRUS data, although all of them had TRUS-guided biopsy. The strong aspect of the study is that it reflects one geographic region, Antalya district, because nearly all the patients in this region were referred to Akdeniz University hospital which had the only TRUS centre in the region during the study period.

This study is the first study that examines the usefulness of Bayes Networks in the prediction of prostate cancer in Turkish population. These results are promising, and further studies with larger and/or different patient groups should be considered.

REFERENCES

[1] A. Jemal et al. “Cancer statistics, 2008," CA Cancer J Clin, 58(2) 2008, pp. 71-96.

[2] A. Jemal, “Cancer statistics, 2010” CA Cancer J Clin, 60(5), 2010, pp. 277-300.

[3] J.M. Marroquin, “To screen or not to screen: ongoing debate in the early detection of prostate cancer,” Clin J Oncol Nurs, 15(1), 2011, pp. 97-98.

[4] P.R. Carroll, J.M. Whitson, M.R. Cooperberg, Serum prostate-specific antigen for the early detection of prostate cancer: always, never, or only sometimes? J Clin Oncol, 29(4), 2011, pp. 345-347.

[5] I.M. Thompson, D.P. Ankerst, “Prostate-specific antigen in the early detection of prostate cancer,” CMAJ, 176(13), 2007, pp. 1853-1858.

[6] P. Carroll, et al., “Prostate-specific antigen best practice policy--part I: early detection and diagnosis of prostate cancer,”. Urology, 57(2), 2001, pp. 217-224.

[7] K.H. Gulkesen, et al. “Comparison of methods for prediction of prostate cancer in Turkish men with PSA levels of 0-10 ng/mL.” J BUON, 15(3), 2010, pp. 537-542.

[8] F.N. Jensen, “Bayesian Networks and Decision Graphs,” 2007, Verlag; New York: Springer.

[9] E. Frank et al., “Data mining in bioinformatics using Weka. Bioinformatics,” 20(15), 2004, pp. 2479-2481.

[10] K. Shigemura et al., “Potential predictive factors of positive prostate biopsy in the Japanese population,” Int Urol Nephrol, 40(1), pp. 91-96.

[11] P.J. Lucas, L.C. van der Gaag, A. Abu-Hanna, “Bayesian networks in biomedicine and health-care,” Artif Intell Med, 30(3), 2004, pp. 201-214.

[12] P.J. Lucas, H. Boot, B.G. Taal, “Computer-based decision support in the management of primary gastric non-Hodgkin lymphoma,” Methods Inf Med, 37(3), 1998, pp. 206-219.

[13] S.F. Galan et al., “NasoNet, modeling the spread of nasopharyngeal cancer with networks of probabilistic events in discrete time,” Artif Intell Med, 25(3), 2002, pp. 247-264.

[14] W.P. Smith, et al. “A decision aid for intensity-modulated radiation-therapy plan selection in prostate cancer based on a prognostic Bayesian network and a Markov model,” Artif Intell Med, 46(2), 2009, pp. 119-130.