Original Article

Turk J Med Sci
2010; 40 (5): 681-686
© TÜBİTAK
E-mail: medsci@tubitak.gov.tr
doi:10.3906/sag-0906-47

Prediction of prostate cancer using decision tree algorithm

Kemal Hakan GÜLKESEN[1,2], İsmail Türker KÖKSAL[3], Sebahat ÖZDEM[4], Osman SAKA[1]

Aim: Serum Prostate Specific Antigen (PSA) level is used for prediction of cancer, but this approach suffers from weak sensitivity and specificity. We applied binary-split decision tree (DT) algorithm to prostate cancer prediction problem.

Materials and methods: Quick, Unbiased and Efficient Statistical Tree (QUEST) algorithm was used in 750 patients who had a serum PSA levels between 0 and 10 ng/mL.

Results: The analysis indicated that following five nodes had different levels of cancer possibility: (1) PSA > 5.98 ng/mL; (2) PSA ≤ 5.98 ng/mL and Digital Rectal Examination (DRE) is suspicious or positive; (3) PSA ≤5 .98 ng/mL and DRE is negative and free PSA > 0.81; (4) PSA ≤ 5.98 ng/mL and DRE is negative and free PSA ≤ 0.81 and age ≤ 57 years; (5) PSA ≤ 5.98 ng/mL and DRE is negative and free PSA ≤ 0.81, and age more than 57 years. The incidences of cancer detection in these groups were 25%, 15%, 0%, 4%, and 16%, respectively. In cases where the nodes 3 and 4 were evaluated as negative, the system would detect 97 of 98 cancer cases with 0.99 sensitivity, saving 74 patients from biopsy (13% of the patients).

Conclusion: DT seems to be a valuable tool to increase specificity in prediction of prostate cancer.

Key words: Prostatic neoplasms, prostate-specific antigen, decision trees

Karar ağacı algoritması ile prostat kanserinin öngörülmesi

Amaç: Serum PSA (Prostate Specific Antigen) düzeyi prostat kanserinin öngörülmesi için kullanılır fakat duyarlılığı ve seçiciliği zayıftır. Bu çalışmada, prostat kanserinin öngörülmesi problemine Karar Ağacı (KA) analizi yöntemini uyguladık.

Yöntem ve gereç: Serum PSA düzeyi 0 to 10 ng/mL arasında olan 750 hastada QUEST (Quick, Unbiased and Efficient Statistical Tree) algoritması ile KA analizi yapıldı.

Bulgular: QUEST analizi sonucu beş farklı düzeyde kanser olasılığı gösteren düğüm elde edildi: (1) PSA > 5.98 ng/mL; (2) PSA ≤ 5.98 ng/mL ve DRM (Dijital Rektal Muayene) pozitif veya şüpheli; (3) PSA ≤ 5.98 ng/mL ve DRM negatif ve free PSA > 0.81; (4) PSA ≤ 5.98 ng/mL ve DRM negatif,,free PSA ≤ 0.81 ve yaş ≤ 57; (5) PSA ≤ 5.98 ng/mL, DRM negatif, free PSA ≤ 0.81 ve yaş > 57. Bu gruplarda kanser insidansı sırası ile % 25, % 15, % 0, % 4 ve % 16 idi. Düğüm 3 ve 4 negatif olarak değerlendirildiğinde, sistem 98 hastadan 97'sini 0.99 duyarlılık ile saptamakta, 74 hastayı biyopsiden kurtarabilmektedir (hastaların %13'ü).

Sonuç: KA analizi, prostat kanserinin öngörülmesinde değerli bir araç olarak görünmektedir.

Anahtar sözcükler: Prostat neoplazmları; prostate-spesifik antijen; karar ağaçları

Received: 13.09.2009 – Accepted: 03.03.2010

[1] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University , Antalya - TURKEY

[2] Department of Health Informatics, Informatics Institute, Middle East Technical University, Ankara - TURKEY

[3] Department of Urology, Faculty of Medicine, Akdeniz University, Antalya - TURKEY

[4] Central Laboratory Clinical Biochemistry Unit, Akdeniz University Hospital, Antalya - TURKEY

Correspondence: *Kemal Hakan GÜLKESEN, Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University, Antalya - TURKEY*
E-mail: hgulkesen@akdeniz.edu.tr

* The study was presented in the 5th Turkish Medical Informatics Congress, 13-16 November 2008, Antalya and Congreso Internacional de Informática en Salud, 9-13 February 2009, Havana, Cuba.

Introduction

Prostatic carcinoma is the most common form of cancer in men (1). Luckily, there are some predictors used for early detection of it. Serum Prostate Specific Antigen (PSA) level, age, and digital rectal examination are helpful in the early detection of the tumor.

When a patient has a suspected prostate tumor, a biopsy from the prostate is advised by the physician. Sometimes, because of the presence of strong indicators, the decision to perform a biopsy is easy. However, when the findings are in the grey area, the physician and patient have to make a joint decision between the risk of missing an early detection of the tumor and the risk of an unnecessary biopsy (2).

PSA is found in the serum of a patient and can easily be measured by a routine biochemical procedure. A serum level above 4 ng/mL was taken as an indicator of the possible presence of prostate cancer and used as the trigger for further clinical evaluation. Several studies using large number of men suggested that the quantization of serum PSA was a useful diagnostic tool for detecting the presence of prostate cancer, particularly when combined with Digital Rectal Examination (DRE) (3, 4). However, some other studies have questioned the sensitivity and specificity of the PSA test (5, 6).

Concepts for improving the specificity of PSA

Since 1989, several concepts to further improve the diagnostic accuracy of PSA have been developed with the aim of avoiding unnecessary biopsies (7). Of these concepts, PSA density (PSAD) and PSA density of transition zone were obtained by the help of transrectal ultrasonography (TRUS) (8, 9). Determination of prostate volume by TRUS may vary considerably. Furthermore, TRUS is not always available, and is time-consuming and relatively expensive.

Complex approaches for improving the specificity of PSA include regression analysis, neural networks, and nomograms (10-14). Although they are reported to produce useful results, these approaches are still in the evaluation phase and they are not used in daily clinical routine.

Free PSA, Percent free PSA

Percent-free PSA (serum-free PSA/PSAx100) is also suggested for improving specificity. Free PSA (fPSA) is the amount of PSA that is not bound to plasma proteins (15). According to a multicentre study on 773 men with a palpably benign prostate gland, compared with the PSA test alone, the "percent-free PSA" is thought to be more sensitive in identifying patients at risk for prostate cancer. A percent-free PSA of greater than 25% is 95% sensitive in excluding prostate cancers when PSA values are in the range of 4 to 10 ng/mL (16).

In another prospective trial from a defined geographic area, a 10.8% prostate cancer was detected in the total PSA range of 1–3 ng/mL when the free/total ratio was 0.2 or less, and the majority of these tumors were clinically relevant (17). According to another study, cancer detection rate is significantly higher in men with a percent-free PSA of < 15% compared to those with a percent free PSA of > 15% in screening population (18). In a Turkish study by Akduman et al, receiver operator characteristic (ROC) area under curve (AUC) for percent-free PSA was 0.553. On the other hand, in patients with prostate volume smaller than 50 cc, the value of percent-free PSA was encouraging. Area under the ROC curve was 0.615 (19).

Multivariate models which include percent-free PSA are generally accepted as having an additional value (20). However, the concept must be interpreted with caution. Several pre-analytical and clinical factors may influence the free/total PSA ratio, such as instability of free PSA both at 4 °C and at room temperature, assay characteristics (equimolar versus skewed response), and a "dilution effect" in large prostates (8).

Decision tree analysis

There are several algorithms for building decision trees (DT). One of them is QUEST (Quick, Unbiased and Efficient Statistical Tree) developed by Loh and Shih (21). The algorithm builds a decision tree structure and classifies subjects into several risk levels. It can be used simply to explore the data, identify possible high-risk subgroups, and uncover interactions or effect modifications among prognostic factors.

DT analysis was used in medicine for several problems including prediction of prostate cancer. To illustrate, Garzotto et al built a decision tree for patients suspected of having prostate cancer using classification and regression tree (CART) analysis (22).

Data were collected from 1433 referred men with serum PSA levels of ≤ 10 ng/mL who underwent a prostate biopsy. Analyzed factors included demographic, laboratory, and ultrasound data. Cancers that were missed by the CART were Gleason score 6 or less in 93.4% of cases. ROC curve analysis showed that CART and logistic regression models had similar accuracy (AUC = 0.74 vs. 0.72, respectively). Free PSA was not included in the data set.

Aim of the study

PSA is the most important marker for tumor screening, but suffers from weak sensitivity and specificity. Several approaches have been developed to obtain a better sensitivity and specificity to date. The aim of this study was to test the value of a decision tree which is obtained by the QUEST algorithm.

Methods

Patients

All transrectal ultrasonography (TRUS)-guided prostatic biopsy cases admitted to Urology Department of Akdeniz University Hospital between January 2001 and April 2007 inclusive were evaluated retrospectively. During the investigated period, 1830 patients underwent an initial prostate biopsy procedure. We included 750 of these patients in the study who had serum fPSA and PSA levels between 0-10 ng/mL. The patients were diagnosed histopathologically by at least sextant prostate biopsy.

Analysis

The data were evaluated by forward stepwise logistic regression analysis (LR) and decision tree analysis (DT). For each approach, ROC AUC was calculated. All the tests except decision tree analysis were performed by the help of SPSS 11.0 (SPSS Inc., Chicago). All statistical tests were 2-sided and alpha value was set to 0.05. DT analysis was performed by a freeware program, QUEST Classification Tree (version 1.9.2, http://www.stat.wisc.edu/~loh/quest.html).

Dependent variable of the current study was a binary coded diagnosis: benign and malign. Independent variables were age (numeric), PSA (numeric), free PSA (numeric), percent-free PSA (numeric), Digital rectal examination I (DRE I, class I: negative, class II: suspicious, class III: positive), DRE II (class I: negative, class II: suspicious or positive).

The data were divided randomly into a model building (75%) and a validation (25%) set. Unequal misclassification costs were specified so that there was a 6 times higher cost associated with misclassifying a cancer case as a noncancer case.

Results

The mean age and standard deviation of the study group (n=562) were 62.7 and 8.3 years respectively. The median PSA level in this group was 5.65 ng/mL. The median free PSA level was 0.95 ng/mL. DRE was normal in 47.5% of patients, suspicious in 15.3% of patients, and positive in 37.2% of patients. Cancer was detected 98 (17.4%) of patients and 21 (3.2%) of the cases were high grade malignant (Gleason grade ≥ 7).

QUEST algorithm identified the following five nodes (groups) having different levels of cancer possibility (Figure): (1) PSA ≥ 5.98 ng/mL; (2) PSA ≤ 5.98 ng/mL and DRE is suspicious or positive; (3) PSA ≤ 5.98 ng/mL, DRE is negative, and free-PSA ≥ 0.81;

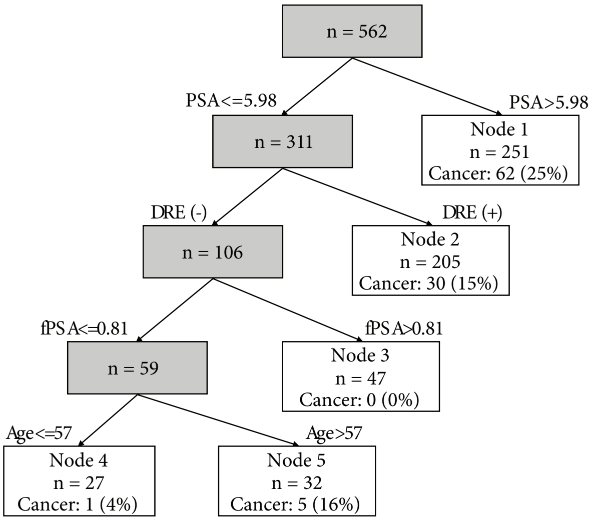

Figure. Decision tree obtained by QUEST analysis.

(4) PSA ≤ 5.98 ng/mL, DRE is negative, free-PSA ≤ 0.81, and age ≤ 57; (5) PSA ≤ 5.98 ng/mL, DRE is negative, free PSA ≤ 0.81, and age ≥ 57. The incidences of cancer detection in these groups were 25%, 15%, 0%, 4% and 16%, respectively. If the nodes 3 and 4 were considered as negative, the system would detect 97 of 98 cancer cases with 0.99 sensitivity; saving 74 patients from biopsy (13% of the patients).

The analysis was then carried out using the randomly selected validation set (n=188). The sensitivity was 0.97 (35 of 36 patients) in that set. Comparison of the study and validation groups is shown in Table. Node 3 contained cancer cases in none of the groups. Node 4 had two cancer patients, one in each group. These two patients were Gleason grade 5 and 6.

The results generated by the QUEST were then compared with a logistic regression model created using the same factors. The decision tree had a ROC curve AUC of 0.62 (95% confidence interval 0.58-0.68). When the same variables were entered into the logistic regression model, age (P = 0.036), DRE (P = 0.001) and PSA (P = 0.001) were detected as statistically significant variables. The AUC of LR was 0.68 (95% confidence interval (CI), 0.63 to 0.73). The AUC of free PSA/PSA ratio alone was calculated as 0.56 (95% CI, 0.50 to 0.62). The AUC's above were obtained using the complete data set.

Discussion

The prsent study is the first one applying QUEST decision tree analysis in prediction problem of prostate cancer. In the previous literature, the same problem was investigated by Garzotto et al., who performed CART analysis which has a similar algorithm to QUEST. Their study was performed on 1433 patients, using demographic data, DRE and TRUS results and serum PSA levels (22). Their results were not comparable to our study due to differences in included parameters because their analysis contains TRUS data and does not contain serum-free PSA level.

The decision tree, which is created in the current study is valuable because it defines two subgroups (nodes) of patients who have a very low possibility of being cancer; (a) men who have serum PSA level below 6 ng/mL, DRE negative, and serum-free PSA level over 0.81 ng/mL, (b) men who have serum PSA level below 6 ng/mL, DRE negative, serum free-PSA level below 0.81 ng/mL, and age below 57 years. The model shows 0.99 sensitivity in study group and 0.97 sensitivity in validation group. According to the model, 13% of the patients may be saved from biopsy with a minimal loss in sensitivity. The AUC of the decision tree is slightly higher than free PSA/PSA ratio alone. The model has slightly lower AUC than LR, but it may be accepted as a significant tool because of its advantage of simplicity for both understanding the model and using it in clinical application.

This model has also revealed an interesting result: the patients with serum PSA levels below 6 ng/mL, negative DRE and low serum-free PSA level (≤0.81) were divided into two nodes. Men below 58 years old had a low possibility of malignancy (4%) compared to older man (16%) with the same features. It should also be noted that, the decision tree algorithm has

Table. Comparison of study group and validation group

Node	Study group		Validation group		Total	
	cancer/total	%	cancer/total	%	cancer/total	%
1	62/251	25	22/98	22	84/349	24
2	30/205	15	11/58	19	41/263	16
3	0/47	0	0/14	0	0/61	0
4	1/27	4	1/7	14	2/34	6
5	5/32	16	2/11	18	7/41	17
Total	98/562	17	36/188	19	134/750	18

preferred fPSA to percent-free PSA. The latter is a derivative of the former, and a low free PSA shows the tendency to a low percent-free PSA. As generally accepted, a low free PSA/PSA ratio is suggestive of prostate cancer; however, there has been no well-known effect of age on this relation. The decision tree revealed that low free PSA does not increase possibility of malignancy in young patients with serum PSA level below 6 ng/mL and negative DRE. A recent study also reported that percent free PSA was not valuable in prediction of prostate carcinoma in men between 44-50 ages (23). We would like to note that clinical value of fPSA in different ethnical groups may be variable; in a previous study on Turkish population, percent free PSA was reported as a poor predictor of biopsy outcome (19).

In conclusion, the findings of the present study suggested that a man with a serum PSA below 6 ng/mL, serum-free PSA over 0.81 ng/mL, and negative DRE has a very low possibility of prostate cancer. Additionally, if serum PSA level is below 6 mg/dL and DRE is negative in a patient younger than 58 years old, there may be no need for ordering a free PSA level, since it may not be helpful in prediction of prostate cancer. It also seemed possible that free PSA and percent-free PSA did not have the same predictive value in different ethnical groups. Further studies are needed to understand the relations of free PSA and free PSA/PSA ratio with age and different ethnical groups. Given that the application of DT analysis revealed unexpected results in the present study; it might be considered as a useful data mining technique with the capacity to give valuable clues in complex medical problems.

Acknowledgement

This study was supported by Akdeniz University Research Projects Unit.

References

1. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. Cancer statistics, 2007. CA Cancer J Clin 2007; 57: 43-66.
2. Tricoli JV, Schoenfeldt M, Conley BA. Detection of prostate cancer and predicting progression: current and future diagnostic markers. Clin Cancer Res 2004; 10(12 Pt 1): 3943-53.
3. Labrie F, Dupont A, Suburu R, Cusan L, Tremblay M, Gomez JL, et al. Serum prostate specific antigen as pre-screening test for prostate cancer. J Urol 1992; 147(3 Pt 2): 846-51.
4. Catalona WJ, Smith DS, Ratliff TL, Dodds KM, Coplen DE, Yuan JJ, et al. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. N Engl J Med 324: 1156-61, 1991. Erratum in: N Engl J Med 1991; 325: 1324, 1991.
5. Guinan P, Bhatti R, Ray P. An evaluation of prostate specific antigen in prostatic cancer. J Urol 1987; 137: 686-9.
6. Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E. Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. N Engl J Med 1987; 317: 909-16
7. Schmid HP, Riesen W, Prikler L. Update on screening for prostate cancer with prostate-specific antigen. Crit Rev Oncol Hematol 2004; 50: 71-8.
8. Stephan C, Lein M, Jung K, Schnorr D, Loening SA. The influence of prostate volume on the ratio of free to total prostate specific antigen in serum of patients with prostate carcinoma and benign prostate hyperplasia. Cancer 1997; 79: 104-9.
9. Djavan B, Remzi M, Zlotta AR, Ravery V, Hammerer P, Reissigl A, et al. Complexed prostate-specific antigen, complexed prostate-specific antigen density of total and transition zone, complexed/total prostate-specific antigen ratio, free-to-total prostate-specific antigen ratio, density of total and transition zone prostate-specific antigen: results of the prospective multicenter European trial. Urology 2002; 60(4 Suppl 1): 4-9.
10. Borden LS Jr, Wright JL, Kim J, Latchamsetty K, Porter CR. An abnormal digital rectal examination is an independent predictor of Gleason > or =7 prostate cancer in men undergoing initial prostate biopsy: a prospective study of 790 men. BJU Int 2007; 99: 559-63.
11. Chun FK, Graefen M, Briganti A, Gallina A, Hopp J, Kattan MW, et al. Initial biopsy outcome prediction--head-to-head comparison of a logistic regression-based nomogram versus artificial neural network. Eur Urol 2007; 51: 1236-40; discussion 1241-3.
12. Stephan C, Cammann H, Semjonow A, Diamandis EP, Wymenga LF, Lein M, et al. Multicenter evaluation of an artificial neural network to increase the prostate cancer detection rate and reduce unnecessary biopsies. Clin Chem 2002; 48: 1279-87
13. Porter CR, Gamito EJ, Crawford ED, Bartsch G, Presti JC Jr, Tewari A, et al. Model to predict prostate biopsy outcome in large screening population with independent validation in referral setting. Urology 2005; 65: 937-41.

14. Karakiewicz PI, Benayoun S, Kattan MW, Perrotte P, Valiquette L, Scardino PT, et al. Development and validation of a nomogram predicting the outcome of prostate biopsy based on patient age, digital rectal examination and serum prostate specific antigen. J Urol 2005; 173: 1930-4.

15. Lilja H, Christensson A, Dahlen U, Matikainen MT, Nilsson O, Pettersson K, et al. Prostate-specific antigen in serum occurs predominantly in complex with alpha 1-antichymotrypsin. Clin Chem 1991; 37: 1618-25.

16. Catalona WJ, Partin AW, Slawin KM, Brawer MK, Flanigan RC, Patel A, et al. Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: a prospective multicenter clinical trial. JAMA 1998; 279: 1542-7.

17. Recker F, Kwiatkowski MK, Huber A, Stamm B, Lehmann K, Tscholl R. Prospective detection of clinically relevant prostate cancer in the prostate specific antigen range 1 to 3 ng./ml. combined with free-to-total ratio 20% or less: the Aarau experience. J Urol 2001; 166: 851-5.

18. Pelzer AE, Volgger H, Bektic J, Berger AP, Rehder P, Bartsch G, et al. The effect of percentage free prostate-specific antigen (PSA) level on the prostate cancer detection rate in a screening population with low PSA levels. BJU Int 2005; 96: 995-8.

19. Akduman B, Alkibay T, Tuncel A, Bozkirli I. The value of percent free prostate specific antigen, prostate specific antigen density of the whole prostate and of the transition zone in Turkish men. Can J Urol 2000; 7: 1104-9.

20. Stephan C, Cammann H, Meyer HA, Lein M, Jung K. PSA and new biomarkers within multivariate models to improve early detection of prostate cancer. Cancer Lett 2007; 249: 18-29.

21. Loh WY, Shih YS: Split selection methods for classification trees. Statistica Sinica 1997; 7: 815-40.

22. Garzotto M, Beer TM, Hudson RG, Peters L, Hsieh YC, Barrera E, et al. Improved Detection of Prostate Cancer Using Classification and Regression Tree Analysis. J Clin Oncol 2005; 23: 4322-9.

23. Vickers AJ, Ulmert D, Serio AM, Björk T, Scardino PT, Eastham JA, et al. The predictive value of prostate cancer biomarkers depends on age and time to diagnosis: Towards a biologically-based screening strategy. Int J Cancer 2007; 121: 2212-7.