


Improved Prediction of the Pathologic Stage of Patient With Prostate Cancer Using the CART–PSO Optimization Analysis in the Korean Population

Technology in Cancer Research & Treatment
 2017, Vol. 16(6) 740–748
 © The Author(s) 2016
 Reprints and permission:
sagepub.com/journalsPermissions.nav
 DOI: 10.1177/1533034616681396
journals.sagepub.com/home/tct


Jae Kwon Kim, MS¹, Mi Jung Rho, PhD², Jong Sik Lee, PhD¹,
 Yong Hyun Park, PhD³, Ji Youl Lee, PhD³, and In Young Choi, PhD²

Abstract

Objective: In current practice, medical experts use the pathological stage predictions provided in the Partin tables to support their decisions. Hence, the Partin tables are based on logistic regression built from the US data. In the present study, we developed a data-mining model to predict the pathologic stage of prostate cancer. In this newly developed model, using the classification and regression tree-particle swarm optimization analysis of the Korean population data, we aim to improve the prediction accuracy of the pathologic state of prostate cancer. **Method:** A total of 467 patients from the smart prostate cancer database were evaluated. The results were intended to predict the pathologic stage of prostate cancer: organ-confined disease and non-organ-confined disease. The accuracy of 4 classification and regression tree-particle swarm optimization models was compared; furthermore, the models were validated with the Partin tables using the receiver operating characteristic curve. **Results:** Among the 467 evaluated patients, 235 patients had organ-confined disease and 232 patients had non-organ-confined disease. The area under the receiver operating characteristic curve of the proposed classification and regression tree-particle swarm optimization model (0.858 ± 0.034) was larger than the I in the Partin tables (0.666 ± 0.046). **Conclusion:** The proposed classification and regression tree-particle swarm optimization model was superior to the Partin tables in terms of predicting the risk of prostate cancer. Compared to the validation of the Partin tables for the Korean population, the classification and regression tree-particle swarm optimization model resulted in a larger receiver operating characteristic curve and a more accurate prediction of the pathologic stage of prostate cancer in the Korean population.

Keywords

classification and regression tree-particle swarm optimization algorithm, data mining, artificial intelligence, machine learning, pathology stage prediction, Smart Prostate Cancer Database

Abbreviations

ANN, artificial neural network; ANNA, artificial neural network analysis; BMI, body mass index; CART, classification and regression tree; IRB, institutional review board; LR, logistic regression; NOCD, non-organ-confined disease; NPV, negative prediction value; OCD, organ-confined disease; PPV, positive prediction value; PSA, prostate-specific antigen; PSO, particle swarm optimization; ROC, receiver operating characteristic

Received: August 1, 2016; Revised: August 16, 2016; Accepted: October 26, 2016.

Introduction

Globally, about 680 000 men are diagnosed with prostate cancer annually, which makes prostate cancer a most common disease in men.¹ In addition, 40% to 50% of men are estimated to have a potential extra prostatic disease.² In the United States, prostate cancer ranks the second in terms of death rate among the other cancer types. Since 2003, the incidence of prostate

¹ Department of Computer Science and Information Engineering, Inha University, Nam-gu, Incheon, Republic of Korea

² Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

³ Department of Urology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

Corresponding Author:

In Young Choi, PhD, Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul 137701, Republic of Korea.
 Email: iychoi@catholic.ac.kr



cancer in males ranks the fifth in terms of its death rate in South Korea.³ In addition, prostate cancer increases with the tumor incidence and aging of the population.

Among numerous methods of treatment of prostate cancer, prostatectomy and radiotherapy are the most effective treatments for patients with prostate cancer.⁴ The effective treatment of prostate cancer requires accurate determination of the clinical data. Here, pathology staging is an important factor that influences the choice of the most suitable method for prostate cancer treatment.⁵ For that reason, accurate prediction of the pathology stage before surgery is important to choose an effective prostate cancer treatment.⁶

Nomogram is a pathology stage prediction method, which can be applied to the patient population and includes the interval of the pathology staging-related data.⁷⁻⁹ However, nomogram is restricted to specific data interval, so it is not useful and valid. Also, applying nomogram to the Korean population can be different in case of clinical variability of prostate cancer. In particular, clinical experience of localized prostate cancer and seminal vesicle infiltration varies considerably between Asian and Western populations; in this context, applying nomogram to the US population is inadequate.¹⁰

The Partin tables are a method designed by nomogram's linear regression technique, a most popularized technique.^{11,12} The Partin tables use prostate-specific antigen (PSA), Gleason score, and clinical stage to predict the pathology stage. The Partin tables have been verified from 2001 to 2011, although the question remains about its applicability to the Korean and, more generally, Asian population.^{6,13} Therefore, for an accurate prediction of the pathology stage, it is necessary to have a decision-making method that would be applicable to the Korean population.

Up to now, numerical indicators are being developed for the decision-making method of prostate cancer. There are representative analysis methods for prostate cancer prediction, such as univariate analysis, multivariable analysis, neural network, and classification and regression tree (CART) analysis, among others. Univariate analysis is simple, as it uses the basic statistics technique, and is easy to apply to the medical system. However, the limitation of univariate analysis is its lack of accuracy. Logistic regression (LR) is one of the multivariate analysis techniques that was applied to the nomograms and Partin tables. This method has a high prediction accuracy for prostate cancer but faces challenges in cases of complex variable relations. Finally, neural network is easy to apply to complex variable relations and is high in accuracy of prediction.^{14,15} However, the generated interpretation of the neural network model is very difficult. In addition, it is also difficult to apply to the medical system, as which the variables demonstrate effectiveness cannot be determined.¹⁶ Classification and regression tree analysis is a decision-tree method. It is high in accuracy of prediction and can determine which specific variable demonstrates effectiveness, which makes it easy to interpret the finally generated prediction model.^{17,18} Thus, CART is an appropriate method for the prediction of prostate cancer, as it is easy to apply to the medical system.^{19,20} For

this reason, this study uses the data-mining technique using the CART decision tree to predict the pathology stage of prostate cancer.

Several previous studies sought to predict the pathology stage of prostate cancer using the data-mining technique. For instance, Matsui used the artificial neural network analysis (ANNA) and demonstrated that its accuracy is higher than that of LR and the Partin tables, when it comes to prediction probability in the Japanese population.^{5,21} Input variables in ANNA are age, serum PSA, PSA density, Gleason score, positive scores, cancer length (mm), and clinical T stage. Furthermore, Olivier et al used machine-learning methods such as support vector machine, multilayer perceptron, radial basis function networks, and so on, to predict and classify the pathological stage of prostate cancer in the British population. Among many machine-learning methods, Bayesian network model particularly enhanced prediction accuracy of the prostate cancer.²² Tsao et al used artificial neural network (ANN) to predict the pathology stage of prostate cancer in the Taiwanese population. Tsao et al study has input variables such as age, body mass index (BMI), PSA, biopsy Gleason sum, primary Gleason grade, digital rectal examination, and transrectal ultrasound and produces a more accurate model than those available with LR and the Partin tables.⁶ Furthermore, Maria et al used the fuzzy expert system to reduce the uncertainty of the existing methods in predicting the pathologic stage of prostate cancer.²³ Also, using the genetic-fuzzy algorithm, Castanho et al demonstrated its enhanced performance in terms of predicting the pathology stage of prostate cancer.²⁴ A genetic-fuzzy system uses the extraction of fuzzy rule and optimized fuzzy membership function and represent a higher performance than the Partin tables. Considering the increase in the number of patients with prostate cancer in Korea, it is necessary to have an accurate method to accurately predict prostate cancer. Input variables used in the present study are PSA, Gleason score, and clinical T stage. Decision tree has a high accuracy and has the advantage of being easy to understand. However, it uses the greedy algorithm to create a tree. Therefore, as a way to focus on the statistical nature of certain data, it tends to fall into local optimization. In order to access the global optimization which takes all the data into account, an appropriate optimization technique should be applied.

The present study aims to predict the pathology stage of prostate cancer in the Korean patients using the data-mining method (CART model). The CART analysis uses Gini index and is based on the binary recursive partitioning method. In order to increase the prediction accuracy of the pathology stage of prostate cancer by optimizing the model generated from CART toward the global optimization, the particle swarm optimization (PSO) algorithm was applied. Particle swarm optimization is a method for optimizing the continuous nonlinear functions and applies the search algorithm to find the optimum solution in a multidimensional search space.

The CART builds a decision tree and classifies subsets into organ-confined disease (OCD) and non-organ-confined

disease (NOCD). In this study, we generate a prediction model that is suitable for the Korean population using CART. Our second aim is to increase the accuracy of the prediction model by accessing global optimization through the tree-structure optimization attained by applying PSO on the prediction model generated from CART (CART-PSO).

This study is the first to suggest a prediction model of the pathology stage of prostate cancer in the Korean population. For the verification of the CART-PSO model, we compare ANN, simple CART, and the Partin tables (2005-2011), with LR using the receiver operating characteristic (ROC) curve. This article is structured as follows. Materials and Methods section describes the data set and proposes the method. Results section outline the system implementation and compares its ability to discriminate prostate-confined cancer and probability tables. Discussion section provides a discussion of the proposed method. Finally, Conclusion section describes the conclusion and specifies further directions in future research.

Materials and Methods

The study data comprised a total of 467 male patients extracted from the Smart Prostate Cancer Data Base at Seoul St. Mary's Hospital between February and November 2013,²⁵ and the study protocol was approved and carried out in accordance with the approved guidelines by the institutional review board (IRB) at the Catholic University of Korea (IRB approval no. Kc14rimi0676). Six input variables—age, BMI, initial PSA value, percentage of the number of positive core (%), clinical Gleason score (sum), and clinical T stage—and 2 output variables—pathologic T stage and N stage—were used.

Preprocessing of the output variables in the analysis of 467 patient data uses the variables of pathologic T stage (pT2a, pT2b, pT2c, pT3a, pT3b, and pT3c) and N stage (pN1). Output variables are transformed by using the guidelines of the American Joint Committee on Cancer which was used to identify the pathologic stage between OCD (pT2+, 237 patients) and NOCD (pT3+ or N+, 232 patients) groups.^{26,27}

Classification and regression tree analysis uses the binary recursive partitioning method that produces a decision tree, which identifies a subset of patients at the pathologic stage. The CART model used in this application was the software IBM SPSS Modeler version 14.2. Classification and regression tree analysis was performed on the model building training data set.

Classification and regression tree analysis has important features. First, it can be used to find which output variables belong to the training data. Second, partitioning of variables is done using the Gini impurity measure.²⁸ Based on the 6 pre-operative pathological and clinical variables, we developed 5 CART models to find out the most powerful variables. The first model (CART model 1) includes 3 input parameters, namely initial PSA value, clinical T stage, and biopsy Gleason score, all of which were used in the Partin tables. The second model (CART model 2) includes BMI in addition to the previous 3 variables. The third model (CART model 3) includes the

Table 1. The CART Models.

Model (Number of Variables)	Input Variable
CART model 1 (3)	Initial PSA value, Gleason score, clinical T stage
CART model 2 (4)	Initial PSA value, Gleason score, clinical T stage, BMI
CART model 3 (5)	Initial PSA value, Gleason score, clinical T stage, BMI, positive core (%)
CART model 4 (5)	Initial PSA value, Gleason score, clinical T stage, age, positive core (%)
CART model 5 (6)	Initial PSA value, Gleason score, clinical T stage, age, BMI, positive core (%)

Abbreviations: BMI, body mass index; CART, classification and regression tree; PSA, prostate-specific antigen.

percentage of tumor cores in addition to the previous 4 parameters. In the fourth model (CART model 4), the age and percentage of the number of cores showing tumor traces were added to the previous 3 parameters. In the final model (CART model 5), age, BMI, and percentage of the number of cores showing tumor traces were added to the 6 parameters. The output variable was OCD or NOCD. The organization of all CART models is shown in Table 1.

Particle swarm optimization is a method belonging to the heuristics as one of the optimization techniques. Like an algorithm, heuristics does not require a specific end condition and is terminated when a specified number of iterations is met. Particle swarm optimization is used to perform the global optimization problems based on the community theory of objects. Therefore, it is used to solve the problem of local optimization in a decision tree. The procedures performed in PSO are as follows.

First, one configures the data that one wants to search for in the particles. Then, for the i th particle located in an arbitrary location $X_i^D = (x_i^1, x_i^2, \dots, x_i^D)$ inside the search space of dimension D , the position (X_i^{n+1}) of the next generation is determined by calculating the location for pbest, the best solution experienced by the researcher, and gbest, the best position for the solution experienced by the community. In order to experience the particle, the velocity function (V) as in Equation 1 is used.

$$V_i^{n+1} = wV_i^n + c_1r_1(P_i^n - X_i^n) + c_2r_2(G_i^n - X_i^n) \quad (1)$$

$$X_i^{n+1} = X_i^n + V_i^{n+1}$$

where, $i = 1, 2, \dots, N$, N denotes the size of the entire community and w , c_1 , and c_2 are the weights of each term as positive real numbers, r_1 and r_2 are random numbers between 0 and 1, and n is the current calculation step. By sharing the optimum positions P_i^n —which the particle has experienced—and G^n —which the whole communities have experienced—a local search around the particle and a global search for the entire space can be performed simultaneously. In this study, all of the attributes corresponding to the conditions of the tree model

Table 2. Confusion Matrix.

Outcome of the Launch Test		Prediction	
		Positive	Negative
Actual	Positive	TP	FP
	Negative	FN	TN

Abbreviations: FN, false negative; FP, false positive; TN, true negative; TP, true positive.

generated by the CART model were configured as particles and optimized using the PSO method.

The nonparametric Mann-Whitney *U* test, confusion matrix, and the ROC curve were used to compare the mean age, mean BMI, initial PSA, clinical T stage, and biopsy Gleason score, as well as the percentage of the number of total cores showing tumor-positive cores between the OCD and the NOCD groups. The software IBM SPSS Statistics version 22.0 was used for all statistical analyses.

Confusion matrix and ROC curves were used to compare the predictive ability. Confusion matrix evaluates the performance of the classifier (Table 2). As shown in Equation 2, the accuracy, sensitivity, specificity, positive predictive value, and negative prediction value (NPV) were measured. The ROC curve compares sensitivity versus specificity along a range. The limit of significance for all tests was set at $P < .05$.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \tag{2}$$

$$\text{PPV (positive prediction value)} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{NPV (negative prediction value)} = \text{TN} / (\text{TN} + \text{FN})$$

Results

Patients' Characteristics

The median age of the 467 patients was 71 years (range: 48-85; mean: 69.94). Median OCD and NOCD of age was 69 years (range: 48-84; mean: 68.46) and 73 years (range: 54-85; mean: 71.41). The median BMI was 23.66 (range: 16.53-38.06; mean: 23.80). Median OCD and NOCD of BMI was 23.18 (range: 16.53-33.31; mean: 23.17) and 24.38 (range: 17.31-38.06; mean: 24.43). The median initial PSA value was 7.3 ng/mL (range: 0.67-128.60; mean: 11.591). Median OCD and NOCD of initial PSA was 5.51 ng/mL (range: 0.669-27.3; mean: 6.638) and 10.28 ng/mL (range: 2.3-128.6; mean: 16.61). The proportion of the patients with PSA ≤ 4.0 , 4.01 to 10.0, 10.01 to 20.0, and ≥ 20.01 ng/mL was 12.42% (58 patients), 54.39% (254 patients), 20.03% (94 patients), and 13.06% (61 patients), respectively. The difference between the 2 groups (OCD and NOCD) in age, BMI, initial PSA, and positive core (%) was significant: $P = .009$ (age), $P = .068$ (BMI), $P = .0$ (initial PSA), and $P = .0$ (percentage of positive core). The median biopsy

Table 3. Distribution of Preoperative Variables Between Patients With OCD and NOCD.

	OCD (235 Patients)	NOCD (232 Patients)	<i>P</i> Value
Age			
Average	68.49	71.41	.009
50	3	0	
51-60	36	9	
61-70	91	85	
71-80	101	129	
81	4	9	
BMI			
Average	23.173	24.43	.068
Initial PSA (ng/mL)			
0-4	47	11	
4.01-10.0	156	98	
10.01-20.0	27	67	
20.01	5	56	
Positive core (%)			
Average	24.68	63.28	.000
Gleason score (sum)			
2-4	2	0	
5	2	2	
6	168	56	
7a (pri 3 + sec 4)	36	53	
7b (pri 4 + sec 3)	20	65	
8-10	7	56	
Clinical T stage			
T1c	62	17	
T2a	82	24	
T2b	41	46	
T2c	46	85	
T3a	4	37	
T3b	0	23	

Abbreviations: BMI, body mass index; NOCD, non-organ-confined disease; OCD, organ-confined disease; PSA, prostate-specific antigen.

showing the percentage of the number of tumor-positive cores was 37.5% (range: 1%-100%; mean: 43.85%). Median OCD and NOCD of positive core (%) was 21.05 (range: 1-100; mean, 24.68) and 69.23 (range: 5-100; mean: 63.28). Biopsy tumor grade was classified as Gleason score (sum) 2 to 4, 5, 6, 7a (pri 3+ sec 4), 7b (pri 4 + sec 3), and 8 to 10 in 0.43% (2 patients), 0.86% (4 patients), 47.97% (224 patients), 19.06% (89 patients), 18.20% (85 patients), and 13.49% (63 patients), respectively. The 467 patients were classified clinically as stage T1c (79 patients), T2a (106 patients), T2b (87 patients), T2c (131 patients), T3a (41 patients), and T3b (23 patients). The distribution of preoperative parameters between the patients with OCD and NOCD is shown in Table 3.

Experimental Results

The experiment was divided into training data set (70%, OCD: 328 patients, NOCD: 161 patients) and validation data set (30%, OCD: 141 patients, NOCD: 71 patients) to measure performance. Table 4 presents the comparison results of 5 CART models, ANN, LR, Partin tables, and the proposed

Table 4. The Area Under Confusion Matrix.

	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy (%)
Partin table	62.50	73.08	79.71	53.52	66.43
LR	78.38	83.33	84.06	77.46	80.71
ANN	75.90	89.47	91.30	71.83	81.43
CART model 1	75.32	82.54	84.06	73.24	78.57
CART model 2	67.82	81.13	85.51	60.56	72.86
CART model 3	78.87	81.16	81.16	78.87	80.00
CART model 4	81.94	85.29	85.51	81.69	83.57
CART model 5	82.09	80.82	79.71	83.10	81.43
CART-PSO	81.82	90.48	91.30	80.28	85.71

Abbreviations: ANN, artificial neural network; CART, classification and regression tree; CART-PSO, classification and regression tree-particle swarm optimization; LR, logistic regression; NPV, negative prediction value; PPV, positive prediction value.

Table 5. The Area Under ROC Curve.

	ROC Curve	P Value	95% Confidence Interval	
			Lower Bound	Upper Bound
Partin table	0.666 ± 0.046	.001	0.576	0.757
Logistic regression	0.808 ± 0.039	.000	0.732	0.883
ANN	0.816 ± 0.038	.000	0.741	0.890
CART model 1	0.786 ± 0.040	.000	0.708	0.865
CART model 2	0.730 ± 0.043	.000	0.645	0.815
CART model 3	0.800 ± 0.039	.000	0.723	0.877
CART model 4	0.836 ± 0.036	.000	0.765	0.907
CART model 5	0.814 ± 0.038	.000	0.739	0.889
CART-PSO (propose)	0.858 ± 0.034	.000	0.791	0.925

Abbreviations: ANN, artificial neural network; CART, classification and regression tree; CART-PSO, classification and regression tree-particle swarm optimization; ROC, receiver operating characteristic.

CART-PSO model (with regard to accuracy, sensitivity, specificity, positive prediction value [PPV], and NPV using the confusion matrix). The CART models, ANN, CART-PSO, and LR show a much better accuracy than the Partin tables (66.43%).

The sensitivity of CART model 5 (82.09%) is higher than that of the other models. The specificity of CART-PSO (90.48%) is higher than that of the other models. The PPV of ANN and CART-PSO (91.30%) is higher than those of the other models. The NPV of CART-PSO (80.28%) is higher than those of the other models.

Therefore, the accuracy, specificity, PPV, and NPV of CART-PSO (85.71%, 90.48%, 91.30%, and 80.28%, respectively) are higher than those of the others models. The CART model 4 has a higher accuracy than the other CART models. Therefore, CART-PSO was optimized using the predictive CART model 4.

The area under the ROC curves is summarized in Table 5. The CART model 1 (0.786 ± 0.040) was larger than the Partin tables (0.666 ± 0.046; Figure 1A). With an additional one, and more variables, the area under the ROC curve of CART model 2 (0.730 ± 0.043), CART model 3 (0.800 ± 0.039), CART

model 4 (0.836 ± 0.036), and CART model 5 (0.814 ± 0.038) was larger than the corresponding area under the ROC curve of the CART model 1 (Figure 1B). However, the area under the ROC curve of CART model 5 decreased to 0.814 (±0.038) and the Korean patients' BMI did not contribute to an increase in predictability in the present study. Therefore, CART model 4 is better than ANN, LR, and the Partin tables (Figure 1C).

Also, CART-PSO optimization by CART model 4 has the highest accuracy score (0.858 ± 0.034; Figure 1D). It can be seen that CART-PSO is the most effective as compared to other models (Figure 1).

Classification and Regression Tree-Particle Swarm Optimization Model

The best model of the CART-PSO procedure was carried out on the training set (329 patients). Variables such as age, initial PSA, Gleason score, clinical T stage, and percentage of the number of cores showing tumor traces were used to determine the prostate stage prediction. The detail decision tree of CART-PSO is shown in Figure 2. For example, the root node selected a percentage of the number of cores cutoff level of over 45.8% alone for the identification of child nodes (node 1, node 2). The node 1 selected a percentage of the number of cores cutoff level of over 24.04 mg/mL and node 3 selected a Gleason score cutoff level of over 7 alone for the identification of OCD. Using 2 cutoff values, 52.41% (sensitivity) of OCD (87 of 166 patients) was identified for further analysis, whereas 87 of the 161 (54.04%) participants with NOCD were correctly identified (specificity). Using this root node, node 1 and node 3 of cutoff alone, the percent overall reduction in pathologic stage was 27.83% in the training set (91 of 327 patients). The CART model 4 was found to have an accuracy of 91.44%, sensitivity of 90.02%, and specificity of 92.90% in the training set. The PPV was 93.37% and NPV was 89.44%.

Discussion

Decision tree derived from logic, management, and statistics is a very successful technique for predicting and explaining the relationship between the measured value and the target value.

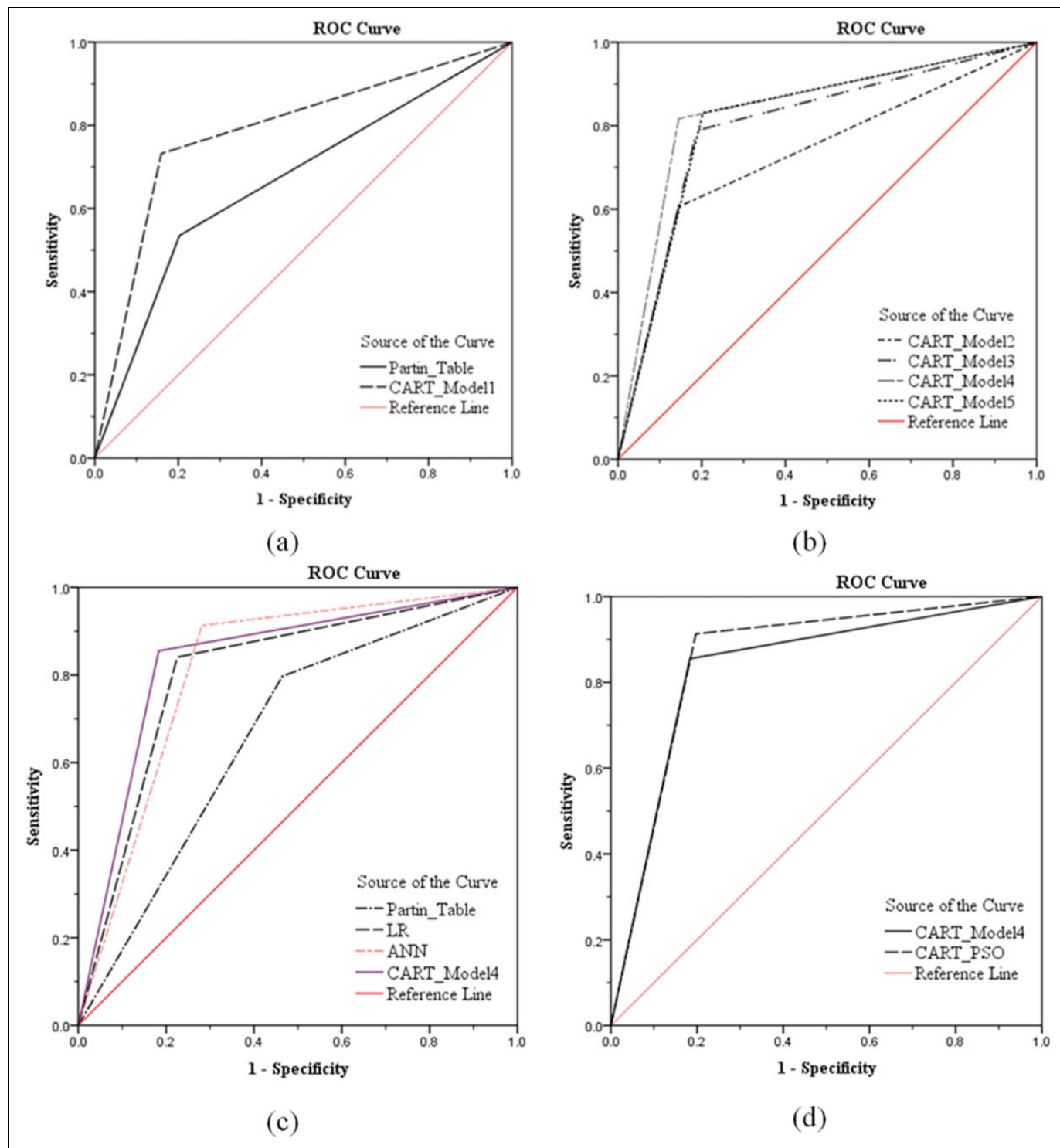


Figure 1. Area under the receiver operating characteristic (ROC) curve. A, ROC curves for each analysis with classification and regression tree (CART) model 1 and the Partin tables in the validation data set. B, ROC curves for CART models 2, 3, 4, and 5. C, ROC curves for CART model 4, logistic regression (LR), and the Partin tables. D, ROC curves for CART model 4 and classification and regression tree-particle swarm optimization (CART-PSO).

Decision tree is a predictive model for displaying the classification and regression models and a powerful technique that can help in decision-making involved in the problems of classification of deductions, prediction, and sequential reasoning.

Classification and regression tree is a nonparametric technique that can be applied to the classification analysis using regression analysis techniques. Impurity function is used in order to determine the data group of high homogeneity. Homogeneity

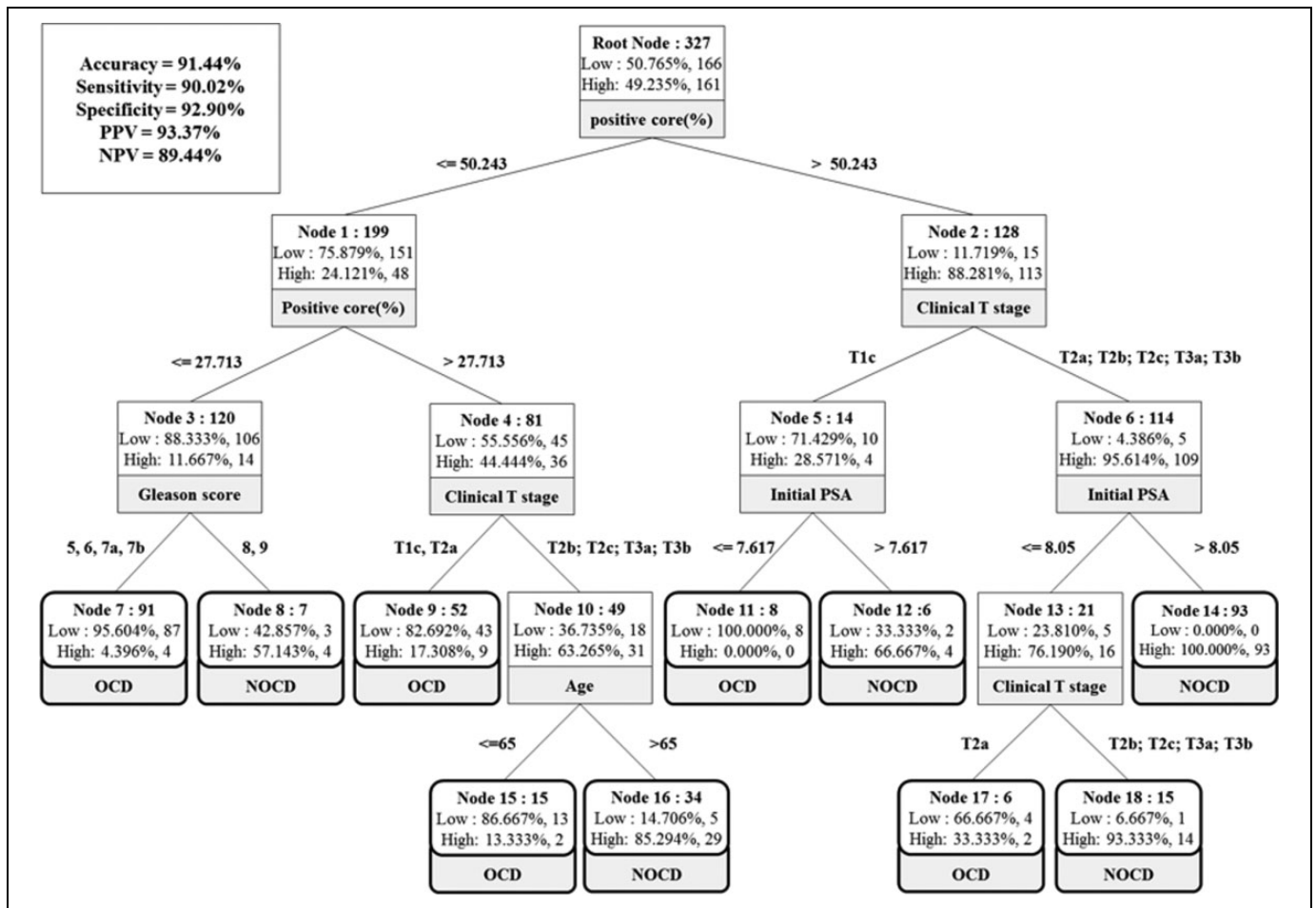


Figure 2. Decision tree (classification and regression tree-particle swarm optimization [CART-PSO]).

indicates that the data are similar in properties and shape, whereas the impurity function means the degree of dispersion, such as the dispersion that occurs among the data.

Substantial research has attempted to predict the pathology of prostate cancer; LR and the neural network were typically used to apply data mining. Both techniques were developed as the main means to increase the accuracy of predictive pathology; however, due to the difficulty in the interpretation of the model, they have the shortcomings of difficulty in the decision-making support. In this context, the CART analysis has advantages, such as the ease of interpretation of the model and a high accuracy of prediction.

Many studies have been conducted for the prediction of the pathology staging of the prostate cancer using data mining. The first relevant research in this area was the model proposed by Snow et al²⁹ that applied the neural network. Accordingly, studies on the Japan and British populations applying data mining have emerged and have demonstrated a higher performances compared to the previous studies which had relied on the Partin tables. Many studies have presented the pattern prediction method applying a statistical approach for pathological prediction and, in order to measure performance, compared the accuracy, sensitivity, specificity, and the ROC curve.

The decision-tree model is characterized by being easy to understand and has been used in many forecasting models showing a high degree of accuracy. However, due to the employment of statistical techniques, the interference for the specific data increases, leading it to easily fall into the local optimization. In this context, a technique that can help resolve the local optimization and ensure the access to the global optimization is required. Particle swarm optimization is 1 of the algorithms of evolutionary arithmetic operation and an optimized technique ensuring the access to a global optimization. Thus, it is possible to reconstruct the tree generated from CART to make the nearest model to the optimal point using the PSO.

The CART-PSO model proposed in the present study has demonstrated a higher predictive value than the Partin tables. In addition, among the shortcomings of the Partin tables is that, in their design, a number of samples depended on a specific population. Also, while 3 variables are used in the Partin tables, the CART model has a predictive value due to the use of 2 additional input variables. The results of the present study confirm that age and the percentage of positive core are significant factors for the pathology prediction of prostate cancer. The CART-PSO model, which applies a total of 5 variables, shows

a higher degree of accuracy (85.71%) than the previously used Partin tables and LR.

The results of the present study have confirmed that the percentage of positive core can be one of important variables in the pathology prediction of prostate cancer. Furthermore, BMI has been found to be not critical to the pathology prediction, so it can be seen that there is no seamless relevance between prostate cancer and the body weight.³⁰

Overall, this study shows that CART-PSO is effective in pathology prediction of prostate cancer. The CART-PSO model has a higher accuracy than the conventional techniques, and the results are presented in the form of a tree which allows for an easy interpretation of the model providing the decision support for the pathology prediction of prostate cancer.

Conclusion

There is the possibility that this CART-PSO analysis method will improve the pathology staging of prostate cancer and decision support in the suitable treatment. We found that BMI has low correlation while age and the percentage of positive core has high correlation. The CART-PSO analysis acknowledging such characteristics has been developed, which may aid in the treatment planning of these individuals.

Currently, among many variables on predicting the staging of prostate cancer, 5 variables have been used; however, one can make use of the additional information relating to the detailed medical history and survival time using the tracking data covering many years; this might deliver a great benefit to the patients in terms of predicting, beyond a simple prediction of the staging of cancer, the quantitative survival time. Also, if the CART-PSO analysis model can be applied to diseases other than prostate cancer, a self-diagnosis algorithm or a similar program could be developed and meaningfully applied.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIP, 2016R1A2B4015922).

References

1. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin.* 2005;55(2):74-108.
2. Pound C, Partin AW, Eisenberger MA, Chan DW, Pearson JD, Walsh PC. Natural history of progression after PSA elevation following radical prostatectomy. *JAMA.* 1999;281(17):1591-1597.
3. Ahmedin J, Rebecca S, Elizabeth W, Taylor M, Jiaquan XM, Michael JT. Cancer statistics 2007. *CA Cancer J Clin.* 2007; 57(1):43-66.
4. Oesterling JE, Brendler CB, Epstein JI, Kimball AW Jr, Walsh PC. Correlation of clinical stage, serum prostatic acid phosphatase and preoperative Gleason grade with final pathological stage in 275 patients with clinically localized adenocarcinoma of the prostate. *J Urol.* 138(1):92-98.
5. Matsui Y, Egawa S, Tsukayama C, et al. Artificial neural network analysis for predicting pathological stage of clinically localized prostate cancer in the Japanese population. *Jpn J Clin Oncol.* 2002;32(12):530-535.
6. Tsao CW, Liu CY, Cha TL, et al. Artificial neural network for predicting pathological stage of clinically localized prostate cancer in a Taiwanese population. *J Chin Med Assoc.* 2014;77(10): 513-518.
7. Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *J Natl Cancer Inst.* 1998;90(10):766-771.
8. Michael WK, Alan MFS, Thomas MW, Peter TS. Evaluation of a nomogram used to predict the pathological stage of clinically localized prostate carcinoma. *Cancer.* 1997;79(3):528-537.
9. Andrew JS, Peter TS, James AE, et al. Preoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *J Natl Cancer Inst.* 2006;98(10): 715-717.
10. Song CR, Kang TJ, Lee MS, et al. Clinico-pathological characteristics of prostate cancer in Korean men and nomograms for the prediction of the pathological stage of the clinically localized prostate cancer: a multi-institutional update. *Korean J Urol.* 2007;48(2):125-130.
11. Partin AW, Kattan MW, Subong ENP, et al. Combination of prostate-specific antigen, clinical stage and Gleason score to predict pathological stage of localized prostate cancer: a multi-institutional update. *JAMA.* 277(18):1445-1451.
12. Makarov DV, Trock BJ, Humphreys EB, et al. Updated nomogram to predict pathologic stage of prostate cancer given prostate-specific antigen level, clinical stage, and biopsy Gleason score (Partin tables) based on cases from 2000 to 2005. *Urology.* 2007; 69(6):1095-1101.
13. Jeong CW, Jeong SJ, Hong SK, et al. Nomograms to predict the pathological stage of clinically localized prostate cancer in Korean men: comparison with western predictive tools using decision curve analysis. *Int J Urol.* 2012;19(9):846-852.
14. Saritas I, Ozkan IA, Sert IU. Prognosis of prostate cancer by artificial neural networks. *Expert Syst Appl.* 2010;37(9): 6646-6650.
15. Sargent DJ. Comparison of artificial neural networks with other statistical approaches. *Cancer.* 2001;91(8 suppl):1636-1642.
16. Michael WK. Comparison of Cox regression with other methods for determining prediction models and nomograms. *J Urol.* 2003; 170(6):S6-S10.
17. Kattan MW, Randolph BC. A simulation of factors affecting machine learning techniques: an examination of partitioning and class proportions. *Omega.* 2000;28(5):501-512.
18. Andrew JS, Shahrokh FS, Michael JZ, et al. Salvage radiotherapy for recurrent prostate cancer after radical prostatectomy. *J Am Med Assoc (JAMA).* 2004;291(11):1325-1332.

19. Shariat SF, Karakiewicz PI, Suardi N, Kattan MW. Comparison of nomograms with other methods for predicting outcomes in prostate cancer: a critical analysis of the literature. *Clin Cancer Res.* 2008;14(14):4400-4407.
20. Norma T, Christopher HS, John LG, Ralph BD Sr, Harry PS. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. *J Clin Epidemiol.* 2003;56(8):721-729.
21. Yoshiyuki M, Shin E, Chotatsu T, et al. Artificial neural network analysis for predicting pathological stage of clinically localized prostate cancer in the Japanese population. *Jpn J Clin Oncol.* 2002;32(12):530-535.
22. Olivier RC, John M, Robert L, Thomas L, Sam M, James N. Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers. *Artif Intell Med.* 55(1):25-35.
23. Maria J, de PC, Laecio C, de B, Akebo Y, Laercio LV. Fuzzy expert system: an example in prostate cancer. *Appl Math Comput.* 2008;202(1):78-85.
24. Castanho MJP, Hernandez F, De Ré AM, et al. Fuzzy expert system for predicting pathological stage of prostate cancer. *Expert Syst Appl.* 2013;40(2):466-470.
25. Choi IY, Park BJ, Chung BH, et al. Development of prostate cancer research database with the clinical data warehouse technology for direct linkage with electronic medical record system. *Prostate Int.* 2013;1(2):59-64.
26. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol.* 2010;17(6):1471-1474.
27. Favaretto RL, Shariat SF, Savage C, et al. Combining imaging and ureteroscopy variables in a preoperative multivariable model for prediction of muscle-invasive and non-organ confined disease in patients with upper tract urothelial carcinoma. *BJU Int.* 2012;109(1):77-82.
28. Ngai EW, Xiu L, Chau DC. Application of data mining techniques in customer relationship management: a literature review and classification. *Expert Syst Appl.* 2009;36(2), 2592-2602.
29. Snow PB, Smith DS, Catalona WJ. Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. *J Urol.* 1994;152(5 pt 2):1923-1926.
30. Çınar M, Engin M, Engin EZ, Ateşçi YZ. Early prostate cancer diagnosis by using artificial neural networks and support vector machines. *Expert Syst Appl.* 2009;36(3):6357-6361.