Classification of Pathological Stage of Prostate Cancer Patients Using Penalized Splines

Tathagata Banerjee
Tapabrata Maiti
Pashpal Mukhopadhyay

W.P. No. 2005-11-06
November 2005

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD-380 015
INDIA

Classification of Pathological Stage of Prostate Cancer Patients Using Penalized Splines

Tathagata Banerjee [a,*], Tapabrata Maiti[b] and Pushpal Mukhopadhyay[6,♀]

[a]*Indian Institute of Management Ahmedabad, Vastrapur, Ahmedabad, 380015, India.*

[b]*Department of Statistics, Iowa State University, Ames, IA, 50011, U.S.A.*

SUMMARY

We propose a penalized splines based method to predict the pathological stage of localized prostate cancer. A combination of prostate specific antigen, Gleason histological score, and clinical stage from a cohort study of 834 prostate cancer patients are used to build the penalized splines model. It turns out that the proposed methodology results in improved prediction of pathological stage compared to usual logistic regression after removing a few outliers. The improvement is shown to be statistically significant. Receiver operating characteristic curve is drawn and we show that the increase in area under the ROC curve over the commonly used logistic regression based classification method is also statistically significant.

KEY words: Logistic Regression; Receiver Operating Characteristic Curve, Nomogram, Organ Confined Disease.

---

*On leave from the Department of Statistics, University of Calcutta, Kolkata, W.B. 700019, India.
"♀"Corresponding author. Department of Statistics, Iowa State University, Ames, IA, U.S.A. Tel.: 5152947786; fax: 5152944433
E-mail address: pushpal@iastate.edu

1. Introduction

Incorrect classification of pathological stage (organ confined (OC) or non-organ confined (NOC)) of prostate cancer patients has an adverse implication on the decision of medical practitioners to go for surgery or not. Cumulative data from numerous contemporary studies involving clinically classified 12,984 patients with organ confined disease documented considerable under staging. Pathological review of the radical prostatectomy specimens from the above group showed that 6,810 patients (52.4%) had organ confined disease (Badalament et al. (1996)). Thus, using the traditional methods in practice, the ability of the urologist to predict non-organ confined disease appears to be no better than a flip of a coin for those receiving radical prostatectomy.

The customary procedure for classifying a patient into either an organ confined or a non-organ confined disease group is to use a nomogram. The nomogram is based on logistic regression of pathological stages (binary) on three pre-surgerical clinical variates viz., Prostate Specific Antigen (PSA) level, Biopsy Report (binary) and Gleason histological score. Examples of such nomograms include those by Partin et al. (1993), Narayan et al. (1995), Badalament et al. (1996), Partin et al. (1997), and Partin et al. (2001).

In section 2, we describe the data obtained from a cohort study (Ghosh et al. (2004)) of 834 cancer patients who were classified either into organ confined or non-organ confined group based on the above three clinical parameters. Afterwards, they were put to surgery and their true disease status viz., organ confined or non-organ confined were observed. We use this data set to develop a prediction model for true status based on the above three clinical parameters. First, we fit a standard logistic regression model to the data for this purpose. On closer examination it is found that logistic regression fails to predict reasonably the true status of the disease with PSA and Gleason score for the patients with unilateral prostate tumor (biopsy = 1). This leads us to consider a logit linear model with covariate PSA being replaced by a smooth function of it. We estimate the smooth function by penalized splines, often called P-splines. In particular, in this paper we use the recent connection between P-splines smoothing and likelihood based linear mixed models (Brumback et al. (1999), Ruppert et al. (2003)). An attractive consequence of this approach is that it can be fit using mixed model software, with maximum likelihood (ML) or residual maximum likelihood (REML) selecting the amount of smoothing. We then draw the receiver operating characteristic (ROC) curve (Pepe (2000)) for the above two prediction models. The proposed model does not show significant increase in area under the ROC curve over that of the standard model.

In section 3 we extend the semi-parametric model to include an interaction term between PSA and biopsy. This is a consequence of the fact that the scatter plot of PSA vs. organ confined rate at each level of biopsy shows the presence of a strong interaction. Examining the plot further and calculating standardized residuals we find that both the standard logistic regression and the penalized splines model detect the same seven patients as outliers. They were thus excluded from further analysis. The

area under the ROC curve of the classification rule based on the proposed penalized splines model shows statistically significant increase over the same corresponding to the standard logistic regression based classification. In section 4 the concluding remarks are given.

2. Semiparametric Penalized Spline Model

2.1 *Data Description*

The data were obtained from a cohort study of 834 cancer patients (Ghosh et al. (2004)). For each patient we have the Gleason score (in a scale of 2-10), Prostate Specific Antigen (in a scale of 0-30), and biopsy report (unilateral if the tumor occur on one side of the prostate or bilateral if both) based on current staging modalities. The Gleason score is based on observations of tissue samples from two different tumors. A pathologist will look for certain well-defined features of cancerous prostate tissue and classify each sample into one of five levels. The levels are the assigned scores of 1 (most benign) to 5 (severe). The Gleason score is the sum of the scores for the two tumors, it is an integer in the range of 2 to 10. Scores of two to four designate low aggressiveness, five to six mildly aggressive, seven moderately aggressive, and scores of eight to 10 highly aggressive. PSA occurs in a protein that is produced in the prostate and tends to seep into the bloodstream. Elevated levels of PSA may be caused by various disorders of the prostate; cancer is just one of several possible causes. Prostate biopsy is best performed under transrectal ultrasound guidance using a spring-loaded biopsy device coupled to the transrectal probe, which is placed in the rectum. The physician will first image the prostate using ultrasound noting the prostate gland's size and shape and whether or not any other abnormalities exist, the most common of which are shadows which might signify the presence of prostate cancer.

Score	Range	Classification	Score	Range	Classification
	0-4	low		2-4	low
	4-10	slightly elevated	Gleason	5-7	medium
PSA	10-20	moderately elevated		8-10	high
	>20	highly elevated			

Table 1
Categorical transformation for PSA and Gleason score.

Biopsy report is coded as 1 if the tumor occurs on one side of the prostate (unilateral) or 2 if the tumor occurs on both sides of the prostate (bilateral). The dependent variable *Y* denotes the disease status after the final pathological analysis (0-nonorgan confined, 1-organ confined), and can be thought of as a gold standard for our analysis. 2.2 *Standard Analysis*

We fit a logistic regression model

$-Y_i \overset{ind}{\sim} Ber(p_i)$

$$Logit(p_i)\beta_0 + \beta_1(PSA_i)+\beta_2(GS_i)+\beta_3(Biopsy_i) \quad (1)$$

where *Ber(.)* is the Bernoulli distribution, *pi* is the proportion of organ confined disease among patients having PSA = *PSAi,* Gleason score = *GSi,* and biopsy = *Biopsyi.* To find the proportion of organ-confined patient for a given level of covariates we first categorize our covariates into 24 groups. These groups are defined by four levels of PSA, three levels of Gleason score and two levels of biopsy. PSA and Gleason score are divided into following levels:

This practice of categorization of PSA, Gleason score and biopsy is quite standard in medical literature (Narayan et al. 1995, Badalament et l. 1996, and Partin et al. 1997). Number of patients with observed proportions of organ confined disease is given in table (2) for each of the 24 groups.

		GS:Low	GS:Medium	GS:High	Total
Biopsy 1	PSA:Low PSA:Slightly	0.69 (41)	0.56 (108)	1.00 (4)	0.52 (153)
	Elevated PSA:Moderately	0.85 (34)	0.83 (173)	0.68 (19)	0.82 (226)
	Elevated PSA:Highly	0.75 (20)	0.78 (72)	0.73 (11)	0.77 (103)
	Elevated	0.62 (8)	0.94 (17)	0.60 (5)	0.83 (30)
Total		0.72 (103)	0.74 (370)	0.72 (39)	0.74 (512)
Biopsy 2	PSA:Low PSA:Slightly	0.03 (32)	0.00 (50)	0.00 (4)	0.01 (86)
	Elevated PSA:Moderately	0.10 (30)	0.00 (74)	0.00 (18)	0.02 (122)
	Elevated PSA:Highly	0.00 (9)	0.04 (49)	0.00 (13)	0.03 (71)
	Elevated	0.00 (2)	0.09 (33)	0.00 (8)	0.07 (43)
Total		0.05 (73)	0.02 (206)	0.00 (43)	0.03 (322)

Table 2

Organ confined disease rate for each of 24 groups. Total number of patients are given in parenthesis. Group sizes are varies from 2 to 173. For biopsy = 2, the observed organ confined rate is always less than 10%.

Source	Estimate	Sum Square	*p* value
Intercept	5.66	0.54	<0.0001
GS	-0.05	0.07	0.4367
PSA	0.07	0.02	0.0003
Biopsy	-4.83	0.37	<0.0001

Table 3

Parameter estimates from model 1.

We fit model 1 with PSA and Gleason score as continuous covariates and biopsy as nominal covariate. The results of the analysis are given in table (3). It shows that only Gleason score is not statistically significant with a *p* value = 0.44. High significance of biopsy variable is due to the fact that most of the patients with biopsy equal to 2 have NOC prostate cancer. Also our data show (Table (2)) that

marginal increase in PSA level tends to increase OC disease rate while increase in Gleason score does not seem to have any tangible effect. Though past studies confirmed that all these three variables are important predictors of disease status (Partin et al. (1997), (2001), Badalament et al. (1996) and Narayan et al. (1995)). Note in our paper we use a data set of 834 men compared to 4133 in (Partin et al. (1997)), 5079 in (Partin et al. (2001)). It may be that due to sampling errors our data fail to reflect the significance of Gleason score. Thus, we refrain from dropping Gleason score from our model.

2.3 *Penalized Splines Model for PSA*

Since PSA is found to be highly significant *(p* value = 0.0003) in model (1), we thought that it would be interesting to investigate the relationship between logit organ confined (pathological stage) rate and PSA. Observed organ confined rate is obtained by taking the ratio of patients with organ confined disease at each level of PSA. From the scatter plot (1) it is found that the relationship is highly nonlinear. We tried to fit a quadratic logistic model in PSA keeping other covariates linear. The fit does not improve significantly. We do not report the details of this investigation.

This leads us to consider a regression model like the following:

$$logit(p_i) = m(PSAi) + \beta_2(GSi) + \beta_3(Biopsy_i) + \epsilon_i \qquad (2)$$

where *pi, PSA*$_i$*, GS*$_i$*,* and *Biopsyi* are defined as in (1) , m(.) is an unknown smooth funϵ_in, and is a random error.

Following Brumback et al. (1999) we propose a linear mixed model formulation of P-splines. We write,

$$m(x) = \beta_0 + \beta_1 x + (x - \kappa_1)_+ b_1 + \ldots + (x - \kappa_k)_+ b_k \qquad (3)$$

where $b_1, \ldots, b_k$ are independently and identically distributed (i.i.d.) with

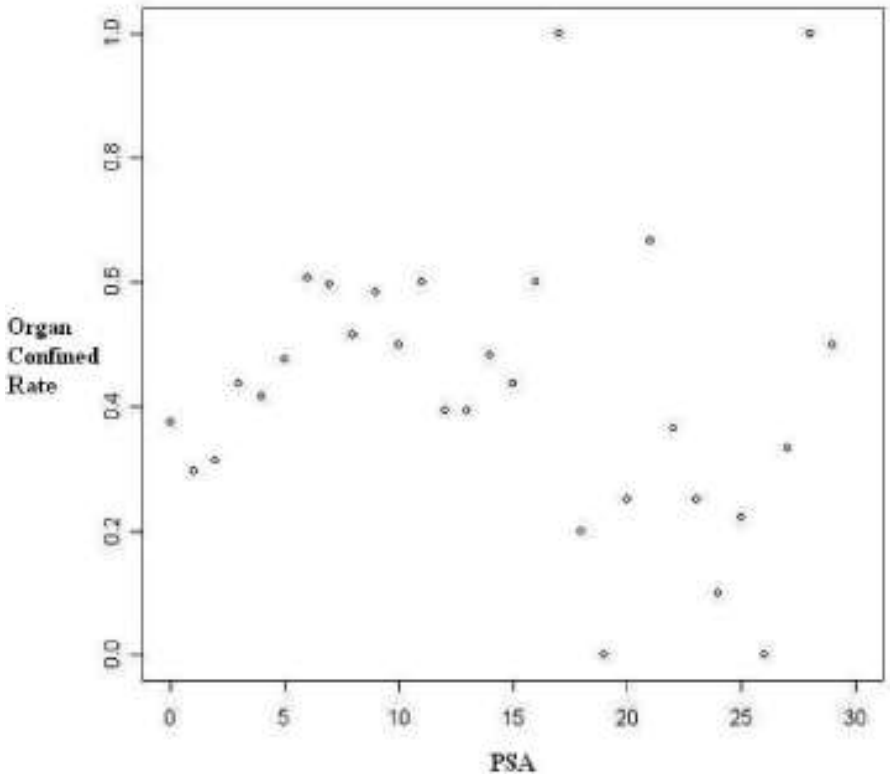

Figure 1. Distribution of organ confined rate at each level of PSA.

$N(0, \sigma_b^2)$, $(x)_+ = x$ if $x > 0$ and 0, otherwise and $\kappa_1, \ldots, \kappa_k$ are k knots usually k quantiles of $(PSA)_i$. Choice of k depends on how non-linear $m(x)$ is. We have fitted splines with 2, 4, 8, 10, and 12 knots and in our case we find that 10 knots give adequately good fit. For further discussion on selection of k we refer to Carroll & Ruppert (2002).

Since ϵ_i's represent the sampling variability in the model we assume ϵ_i's are independently normally distributed with $E(\epsilon_i) = 0$, and $V(\epsilon_i) = V(ln(\frac{p_i}{1-p_i})) \simeq \frac{1}{n_i \pi_i (1-\pi_i)}$ where $\pi_i = E(p_i)$.

As π_i's are unknown we carry out the analysis based on P-splines model (2) assuming $V(\epsilon_i) \simeq \frac{1}{n_i p_i (1-p_i)} = \sigma_{e_i}^2$ (known). In matrix notation the model (2) can be written as,

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{b} + \boldsymbol{\epsilon} \qquad (4)$$

where, $\mathbf{y} = (y_1, ..., y_g)^T$, $y_i = ln(\frac{p_i}{1-p_i})$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$,

$$X = \begin{bmatrix} 1 & (PSA)_1 & (GS)_1 & (Biopsy)_1 \\ .. & .. & .. & .. \\ 1 & (PSA)_g & (GS)_g & (Biopsy)_g \end{bmatrix},$$

$$Z = \begin{bmatrix} PSA_1 - \kappa_1 & ... & PSA_1 - \kappa_k \\ ... & ... & ... \\ PSA_g - \kappa_1 & ... & PSA_g - \kappa_k \end{bmatrix},$$

$\mathbf{b} = (b_1, ..., b_k)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_g)^T$, and $g = 1, 2, ..., 24$.

The predicted value of $ln(\frac{\hat{p}_i}{1-\hat{p}_i})$ is given by,

$$\hat{ln}(\frac{\hat{p}_i}{1-\hat{p}_i}) = \hat{\beta}_0 + \hat{\beta}_1(PSA)_i + \hat{\beta}_2(GS)_i + \hat{\beta}_3(Biopsy)_i + (Z_{1i}...Z_{ki})(\hat{b}_1, ..., \hat{b}_k)^T \quad (5)$$

where $\hat{\boldsymbol{\beta}} = (X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}\mathbf{y}$, $\hat{\Sigma} = ZZ^T\hat{\sigma}_b^2 + diag(\sigma_{e_i}^2, i = 1, ..., m)$,

$(\hat{b}_1, ..., \hat{b}_k)^T = \hat{\sigma}_b^2 Z^T\hat{\Sigma}^{-1}(\mathbf{y} - X\hat{\beta})$. $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_b^2$, the estimates of $\boldsymbol{\beta}$ and σ_b^2, are obtained by maximizing the penalized spline likelihood. The kernel of the penalized splines likelihood is given by $-\frac{1}{2}ln|\Sigma| - \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})^T\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta})$.

To assess the predictive powers of the above two models we find the predicted values of *pi* using the above models for each of the 834 patients. The *pi* value corresponding to each of the 24 groups is then obtained by taking average of the predicted *pi* values corresponding to the patients belonging to the group. A scatter plot showing the predicted values from both the models along with the observed values are shown on figure (2). In ideal situation, the predicted values will be equal to the observed values. In other words, all the points in the scatter should lie on a line passing through the origin and making an angle 45° with the positive direction of the horizontal axis. From figure (2), it looks like that the penalized splines result in improved prediction.

The whole motivation of introducing penalized splines is to improve the prediction of disease status. This, in consequence, is expected to reduce the error in classification of patients into one of the two groups, viz., organ confined and non-organ confined. To see this we draw the receiver operating characteristic (ROC) curves (Pepe (2000)) of the classification rules based on the above models. The ROC curve is primarily a descriptive device displaying the range of the trade-offs between true-positive and false-positive rates. For a given threshold value c we classify each patient into organ confined or non-

organ confined group according as the predicted *pi* value exceeds or does not exceed c. We then have, for each c, a table as shown in (4). We calculate

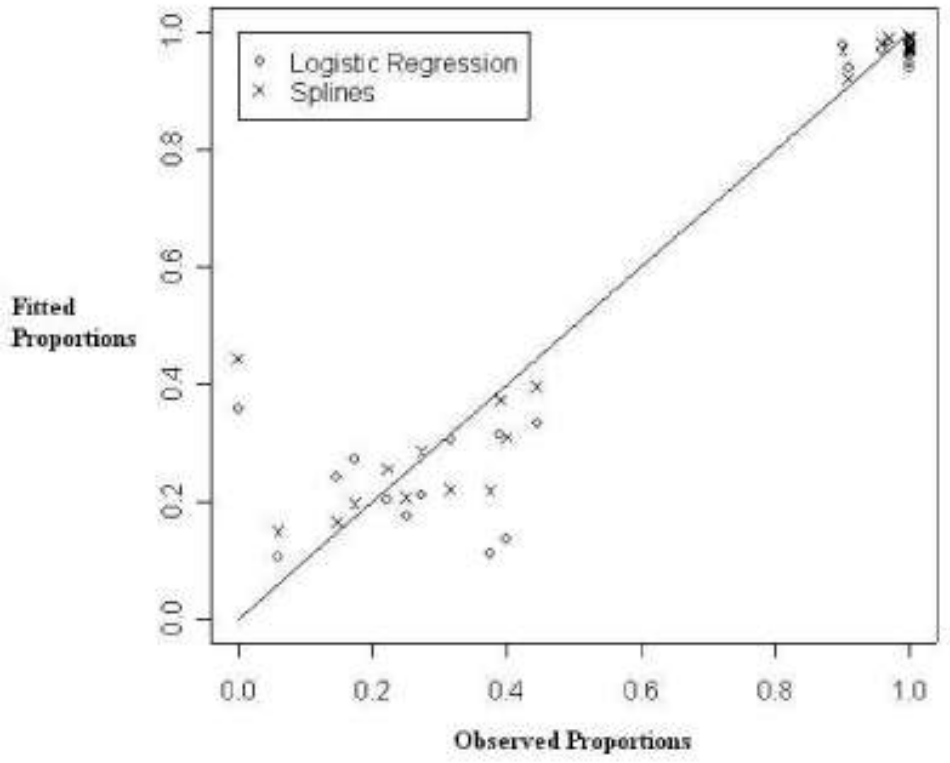

Figure 2. Scatter plot of observed NOC rate vs. fitted NOC rate from logistic regression and penalized splines model. Predicted values from the P-splines model are more close to the 45 degree line than predicted values from the logistic regression model.

		Predicted	
		OC	NOC
True	OC	a_{11}	a_{12}
	NOC	a_{21}	a_{22}

Table 4

true positive rate (TPR) = $\frac{a_{11}}{a_{11}+a_{12}}$ and false positive rate (FPR) = $\frac{a_{11}}{a_{21}+a_{22}}$. A plot of TPR vs. FPR for all possible choices of c is the ROC curve in this case. The better the classification rule the higher will be the curve. In other words, if the ROC curve of a classification rule say, C1 lies entirely above the ROC curve of another classification rule say, C2 then C1 is a better classifier than C2. From figure (3), it is evident that the classification rule based on penalized splines performs better than that of based on standard logistic regression. But the question is whether the increase in area under the

ROC curve (AUC) is statistically significant or not. In order to see it we carry out a Mann-Whitney type non-parametric test following Mason et al. (2002). The AUC for the logistic regression and the penalized splines are 0.874 and 0.890 respectively and the *p* value for the test is 0.35. Thus we find that the improvement is not statistically significant at levels less than 0.35.

3. Semiparametric Model With Interaction Between PSA And Biopsy

From figure (2), it is evident that for high values of observed *pi* the predicted values obtained from both the models are close to the observed values. However, for the smaller values of observed *pi* the predictive performances of the models differ. On closer scrutiny it is observed that most of the patients having biopsy equal to 2 have non-organ confined disease. This fact is clearly evident from the plot of organ confined rate against PSA for each level of biopsy as shown in figure (4). Thus it indicates that PSA and biopsy level have strong interaction. We now extend model (2) to incorporate the PSA-biopsy interaction.

$$logit(p_i) = \beta_0 + \beta_1(PSA)_i + \delta_i Z_i^T b_i + \beta_2(GS)_i + \beta_3(Biopsy)_i$$

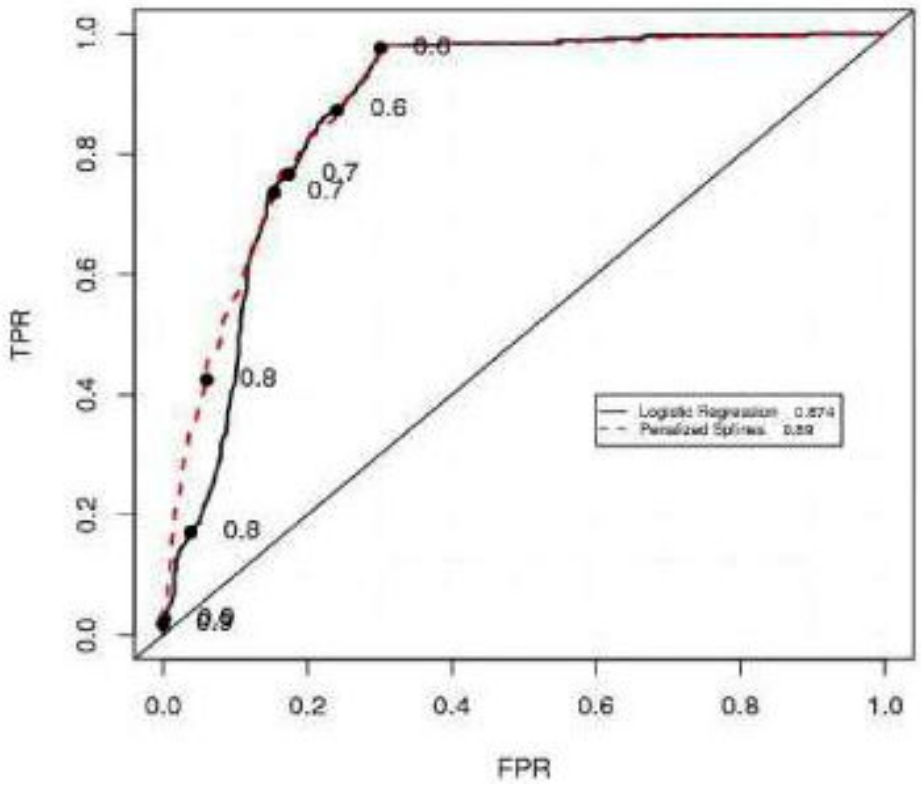

Figure 3. ROC curve for logistic regression prediction against P-splines prediction. The dashed line (predictions from the P-splines model) almost always dominates the solid line (predictions from the logistic regression model).

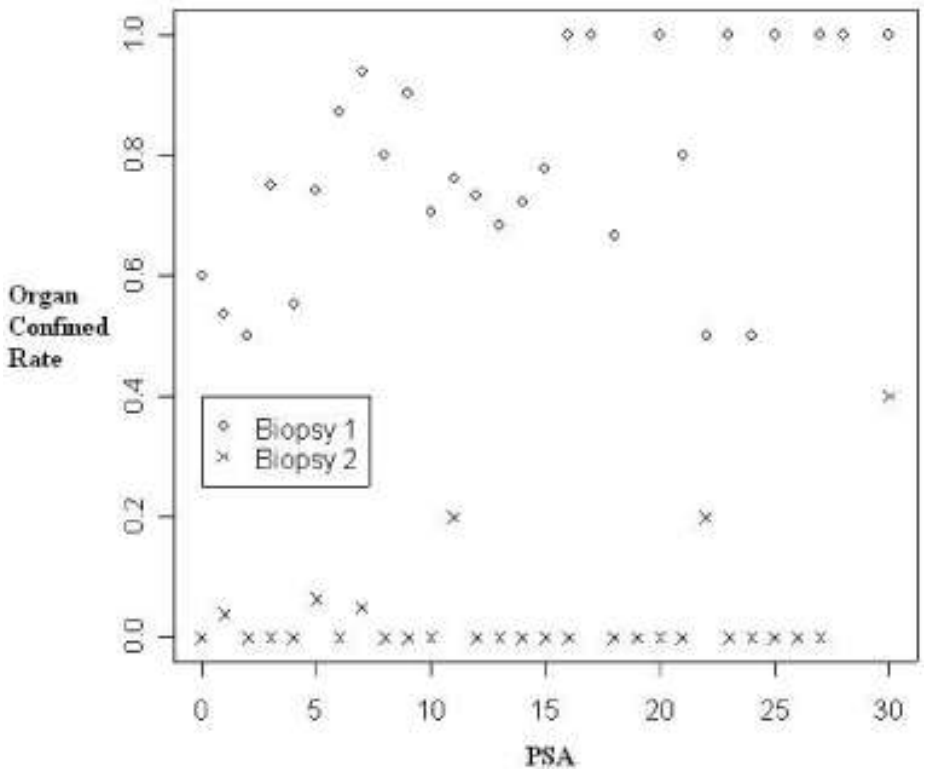

Figure 4. Distribution of organ confined rate at each PSA level for two different levels of biopsy. Clearly, there is an interaction between biopsy and PSA.

Case ID	Logistic Regression	Penalized Splines
831	13.58	17.44
786	4.38	11.13
698	8.35	9.30
703	8.35	9.30
604	10.48	7.47
650	10.48	7.47
613	9.72	4.23

Table 5
Patients with high standardized residuals.

$$+(1-\delta_i)(\tilde{\beta}_1(PSA)_i + \tilde{Z}_i^T\tilde{b}_i) + \epsilon_i \qquad (6)$$

where δ_i is 1 if biopsy = 1 and 0 otherwise and Z_i is similar to Z_i except that it is based on the quantiles on $(PSA)_i$ when biopsy=2. Again we fit the model following procedure similar to that in Section 2.3.

Although the ROC curve of the classification rule based on the proposed model (6) shows some improvement over that based on standard logistic regression model but as before the improvement is not found to be statistically significant at levels 0.05 or less. Investigating further into the fittings of the models, from figure (4), it seems that there are a few outliers. We run the LR Model and the proposed PSP model (6) to the data and detected seven points giving high standardized residuals (table 5) by both the regression models. We then draw the ROC curves (figure 5) of the two models after removing the outliers from the data. The AUC are 0.882 for the LR model and 0.901 for the PSP model. The difference between the areas under ROC curve is found to be highly significant with a p value less than 0.0001. High statistical significance for an increase in area around 0.02 though seems surprising but

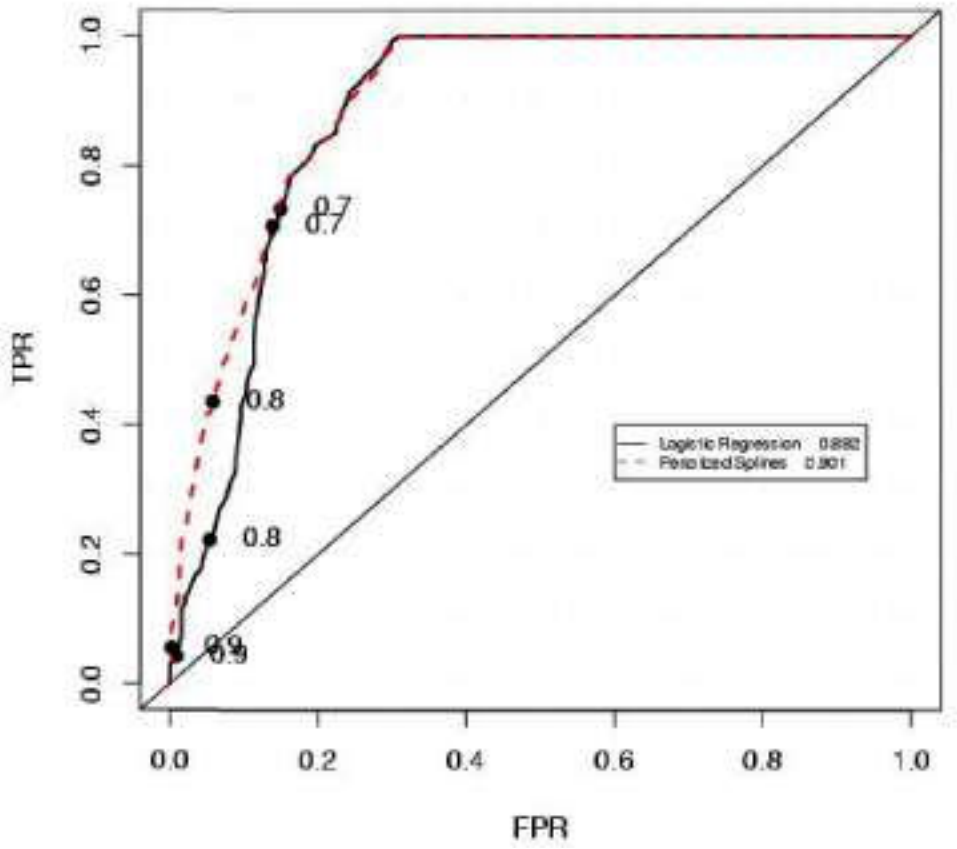

Figure 5. ROC curve for logistic regression prediction against P-splines prediction after removing seven outliers. The dashed line (predictions from the P-splines model) always dominates the solid line (predictions from the logistic regression model).

On examination we find that the standard error of the difference becomes significantly lower. Also figure 5 clearly shows an appreciable improvement in the true positive rate by the PSP model over the same by the LR model for an interval of false positive rate near zero.

4. Concluding Remarks

In this paper we propose a methodology for classifying prostate cancer patients into organ confined & non-organ confined cases. The P-splines based classification shows significant improvement over the standard logistic regression based classification.Though we have illustrated the methodology with a specific data set, it is generally applicable to similar classification problems in other areas of applications too. Since the proposed classification method is based on a model like $logit(p_i) = m_1(x_1) + ... + m_p(x_p)$ where $m_1(.), .., m_p(.)$ are unknown smooth functions, it can be considered as a generalization of the same based on the loglinear model. The advantages of using the mixed linear model formulation of P-splines for classification, envisaged in this paper, are: the methodology

can be implemented using any standard software with a module for mixed linear models procedure and the method provides a data adaptive choice of the smoothing parameter either by maximum likelihood or residual maximum likelihood.

Similar to Section 3, the method can incorporate the interactions between any number of binary and continuous covariates. If in a given situation, the PSP based classifier performs better then we would expect that a nomogram created by using this model will be better compared to that based on the standard LR model.

ACKNOWLEDGEMENTS

This research is supported in part by NIH grants R01 CA-85414 and by NSF grants ACS0318184. We are indebted to the reviewers and an associate editor for their helpful comments.

References

Badalament, R. A., Craig, M. M., Peller, P. A., Young, D. C., Bahn, D. K., Kochie, P., O'Dowd, G. J., and Veltri, R. W., 1996. An algorithm for predicting nonorgan confined prostate cancer using the results obtained from sextant core biopsies with prostate specific antigen level. The Journal of Urology, 156 1375-1380.

Brumback, B. L., Ruppert, D., and Wand, M. P., 1999. Comments on Shively, Kohn, and Wood. Journal of the American Statistical Association, 94 794-797.

Dodd, E. L., and Pepe, M. S., 2003. Semiparametric regression for the area under the receiver operating characteristic curve. Journal of the American Statistical Association, 98(462) 409-417.

Ghosh, M., Maiti, T., Kim, D. H., Chakraborty, S., and Tewari, A., 2004. Hierarchical Bayesian neural networks: an application to a prostate cancer study. Journal of the American Statistical Association, 99(467) 601-608.

Mason, S. J., and Graham, N. E., 2002. Areas beneath the relative operating characteristics (ROC) and levels (ROL) curves: statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society, 128 2145-2166.

Narayan, P., Gajendran, V., Taylor, P. S., Tewari, A., Presti, J. C. Jr., Leidich, R., Lo, R., Palmer, K., Shinohara, K., and Spaulding, J. T., 1995. The role of transrectal ultrasound-guided biopsy-based staging, pre-operative serum prostate-specific antigen, and biopsy Gleason score in prediction of final pathologic diagnosis in prostate cancer. Urology, 46(2) 205-211.

Partin, A.W., Mangold, L.A., Lamm, D.M., Patrick, C.W., Epstein, J.I., and Pearson, J.D., 2001. Contemporary update of prostate cancer staging nomograms (Partin Tables) for the new millennium. Eurology, 58(6) 843-848.

Partin, A. W., Subong, E. N. P., Walsh, P. C., Wojno, K. J., Oesterling, J. E., Kattan, M. W., Scardino, P. T., and Pearson, J. D., 1997. Combination of prostate-specific antigen, clinical stage, and Gleason score to predict pathological stage of localized prostate cancer. JAMA, 227(18) 1445-1451.

Partin, A. W., Yoo, J., Carter, H. B., Pearson, J. D., Chan, D. W., Epstein, J. I., and Walsh, P. C., 1993. The use of prostate-specific antigen, clinical stage, and Gleason score to predict pathological stage in men with localized prostate cancer. The Journal of Urology, 150 110-114.

Pepe, M. S., 2000. Receiver operating characteristic methodology. Journal of the American Statistical Association, 95(449) 308-311.

Ruppert, D., Wand, M. P., and Carrol, R. J., 2003. Semiparametric Regression. Cambridge University Press.