Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Clinical Medicine Insights: Oncology

Development of Interpretable Predictive Models for BPH and Prostate Cancer

Pablo Bermejo[1], Alicia Vivo[2], Pedro J. Tárraga[2] and J. A. Rodríguez-Montes[3]

[1]Escuela Superior de Ingeniería Informática, Universidad de Castilla-La Mancha, Albacete, Spain. [2]Gerencia de Atención Primaria, Hospital Universitario de Albacete, Albacete, Spain. [3]Departamento de Cirugía, Hospital de la Paz, Madrid, Spain.

ABSTRACT

BACKGROUND: Traditional methods for deciding whether to recommend a patient for a prostate biopsy are based on cut-off levels of stand-alone markers such as prostate-specific antigen (PSA) or any of its derivatives. However, in the last decade we have seen the increasing use of predictive models that combine, in a non-linear manner, several predictives that are better able to predict prostate cancer (PC), but these fail to help the clinician to distinguish between PC and benign prostate hyperplasia (BPH) patients. We construct two new models that are capable of predicting both PC and BPH.

METHODS: An observational study was performed on 150 patients with PSA ≥3 ng/mL and age >50 years. We built a decision tree and a logistic regression model, validated with the leave-one-out methodology, in order to predict PC or BPH, or reject both.

RESULTS: Statistical dependence with PC and BPH was found for prostate volume (P-value < 0.001), PSA (P-value < 0.001), international prostate symptom score (IPSS; P-value < 0.001), digital rectal examination (DRE; P-value < 0.001), age (P-value < 0.002), antecedents (P-value < 0.006), and meat consumption (P-value < 0.08). The two predictive models that were constructed selected a subset of these, namely, volume, PSA, DRE, and IPSS, obtaining an area under the ROC curve (AUC) between 72% and 80% for both PC and BPH prediction.

CONCLUSION: PSA and volume together help to build predictive models that accurately distinguish among PC, BPH, and patients without any of these pathologies. Our decision tree and logistic regression models outperform the AUC obtained in the compared studies. Using these models as decision support, the number of unnecessary biopsies might be significantly reduced.

KEYWORDS: observational study, prostate pathology prediction, prostate cancer, benign prostate hyperplasia, primary care, prostate biopsy

CITATION: Bermejo et al. Development of Interpretable Predictive Models for BPH and Prostate Cancer. *Clinical Medicine Insights: Oncology* 2015:9 15–24 doi: 10.4137/CMO.S19739.

RECEIVED: August 27, 2014. **RESUBMITTED:** October 28, 2014. **ACCEPTED FOR PUBLICATION:** October 30, 2014.

ACADEMIC EDITOR: William C S Cho, Editor in Chief

TYPE: Original Research

FUNDING: This work has been partially funded by FEDER funds and the Spanish Government (MINECO) through project TIN2010–20900-C04–03, and the FISCAM grant AP-2008/07. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: pjtarraga@sescam.jccm.es

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Introduction

Prostate cancer (PC) is the most common type of tumor and the second cause of death in men from EU,[1] and the main prostate-related pathology associated with death in men from USA.[2] These high incidence rates led the scientific community to search for PC markers.

Marginal markers. PC is not the only common prostate-related pathology in men. Benign prostate hyperplasia (BPH) is also a prostate-related pathology, which consists mainly in a very high prostate volume. It can be difficult to differentiate BPH from PC, and there are a few studies that tried to predict it from prostate-specific antigen (PSA)[3] or derivatives[4,5] but without success, maybe because of the fact that severe inflammation may affect the level of PSA in blood.[6] Statistically, it is difficult to differentiate PC from BPH using intermediate PSA.[7] Taking this into account, as well as the interaction with other predictive variables, we suggest that PC and BPH can be disjoint values for a prostate-pathology prediction model. To the best of our knowledge, no successful model for BPH and PC prediction has yet been suggested in the literature.

Thus, over the years, PSA has become the main indicator for referring primary care patients to a urologist. Furthermore,

high levels of PSA or an abnormal digital rectal examination (DRE) became the main (and frequently the only) reason to recommend a surgical biopsy, resulting in the fact that around 75% of these biopsies were negative.[8] This low specificity was the main reason to define new variables derived from PSA in order to decrease the number of unnecessary biopsies while maintaining its sensitivity. Thus, variables such as PSA velocity (increment of PSA in a given time), PSA density (PSA divided by prostate volume as measured by transrectal ultrasound), age-specific PSA (PSA ranges accepted as normal given the patient's age), free/total PSA (PSA unbound to protein divided by total or standard PSA test), and complex PSA (PSA bound to alpha-1 anti-chymotrypsin; that is, PSA-ACT) have been a hot research topic in the scientific community. A comparison of several assays[3] shows that complex PSA may achieve better specificity with respect to the others at the same sensitivity levels, but of course the method to obtain this variable is more costly. On the other hand, this conclusion is negated in later studies,[9,10] which conclude that no real differentiation can be found among the PSA-derived variables. It becomes clear that the marginal consideration of just one variable for biopsy recommendation is not enough, whatever the variable and whatever the cut-off level chosen. In the same way that PSA is not valid as the sole indicator of biopsy, the use of the international prostate symptom score (IPSS) evaluation as the only marker may result in many BPH situations being unidentified or confused with PC. The joint value of other variables such as PSA, DRE, and prostate volume is more conclusive than IPSS. In fact, we found that IPSS was seldom selected in our experiments. So, given that PC and BPH may share the same probability for variables such as PSA, we propose that it is important to define models that provide a multivariate understanding of the variables available.

Multivariate markers and predictive models. Given the above, the next logical step found in the literature is the stratified use of a set of variables; that is, the surgical process is decided according to the concatenation of pre-defined values and to a given order for such variables. For example, Catalona et al.[11] claim that patients with a PSA level of 4–10 ng/mL have, in general, a 25% probability of having PC. It is possible to factor this PC probability according to six different levels of %free PSA, achieving three to five points more in specificity. This kind of data stratification regarding the values of ordered variables is commonly known as look-up tables, and these have been compared with other PC prediction methods,[12] leading to the conclusion that they do not work very well. We think that the potential reason is that this linear combination of variables, based just on frequencies, might not represent the real interaction between such variables.

And so, in the last decade, we have seen an increase in the number of articles devoted to the development of predictive models coming from the machine learning[13] community. These models also use a pre-defined set of relevant features but combine their values according to different classification paradigms, which, moreover, can automatically re-define the predictive models built when adding new records to the database. Some of these models predict a binary response for PC,[8,12,14] and some others are able to compute a probability of PC.[6,15,16] However, the most important difference for clinicians is the fact that some of these predictive models make their prediction in a black-box manner; that is, they cannot explain in an interpretable manner the process of inference followed to decide the value of the response variable. Whatever the effectiveness of a predictive model is, these are always to be used just as decision support systems,[12] and so it is important that the clinician can understand what the reasons for a given prediction are. Thus, although some classifiers used for prediction such as artificial neural networks[14,17] usually obtain very good performance results after validation, they are unable to explain what interactions among the predictive variables have led them to make their prediction. Thus, focusing on the development of classifiers that can make an accurate prediction of PC and be self-explainable, we find that in the last few years there have been four predictive models that have caught the attention of the research community: nomograms,[18] decision trees,[8,12] logistic regression,[19,20] and Bayesian Networks (BNs).[21,22]

Nomograms are regarded as one of the most accurate and interpretable type of models. By means of graphical line-based connections, they show the interaction among variables from a mathematical formulation, which, commonly, is built using a regression model. Thus, this method is not self-dependent, which is probably the reason why we cannot find it in most of the state-of-the-art statistical and machine learning software packages. Furthermore, its representation and interpretation are very sensitive to the number of variables. With respect to BNs, the specialist may need a long training period and the aid of a visual application with a text explanation of the prediction and colored links in the graph.[21] Furthermore, the more variables are included, the more difficult the model is to interpret and the more disagreements may exist between structures built manually and automatically. Thus, it is not clear that BNs can be categorized or regarded as easy-to-interpret by the clinician. On the other hand, decision trees are self-descriptive models: using predictive variables as nodes of the tree, and one branch per nominal value of each node, any person can follow the reasoning of the prediction starting from the root node to a leaf node, which contains the predicted value of the goal or class variable. And as for logistic regression models, they provide a set of coefficients to apply to a given formula; these coefficients are very easy to interpret and help the clinician to understand the contribution of each variable to the outcome. Consequently, after reviewing the current state-of-the-art models, we find that decision trees and logistic regression are the only predictive models that successfully satisfy a three-fold criterion:

1. They clearly show the interaction among variables: we can interpret what the most important variables are in

reducing uncertainty about the prostate pathology, and it is straightforward to test the impact on prediction performance when removing or swapping variables, and furthermore, they are easily tunable.

2. They are easily interpretable: decision trees start with a root variable, and depending on its value, the clinician follows the corresponding branch, which leads to another variable, which can take another set of values, and so on until a leaf node is reached and its value is the prediction given. Regarding regression models, they provide odds ratios and a statistical significance that let us easily recognize the impact of variables on the model. Moreover, a clinical expert can even manually suggest interaction between variables.
3. Finally, these two models are implemented and are easily usable in well-known software packages such as R,[23] Weka[24] (which are free and open-source tools) and IBM SPSS.[25]

With regard to the predictive attributes, we can find a wide range of variables used in many PC-related articles. However, many of them are difficult and/or expensive to obtain, such as complex PSA, serum proteases such as hK2,[3] or expression of cytokines.[26] Moreover, they may even be impossible to obtain when the prediction needs to be performed in primary care consultations. Thus, in this article, we present a study with two main contributions: (1) we provide interpretable predictive models for PC and BPH, obtaining better results than those found in the literature and compared in this study, and thus making it possible to reduce the number of unnecessary biopsies, which is a non-definitive painful process that can cause infections; and (2) we build these models by using only variables that are easy to obtain from a primary care context; except in the case of prostate volume, which cannot be regarded as *easy* to obtain but which is, however, cheap.

In the next section, we explain the methodology used to obtain the target sample, and to create and evaluate the predictive models and perform the statistical comparisons. Then, we present the results of our statistical and predictive analyses, and finally, we discuss these results and draw our conclusions.

Materials and Methods

This is an observational study, with an original sample size of 150 men obtained from five primary care centers during a 3-year period, 2009–2012, in the city of Albacete, Spain. It was approved by the ethical committee of the CHUA (Hospital of Albacete). The research was conducted in accordance with the principles of the Declaration of Helsinki.

The inclusion criteria were men of age >50 years, PSA ≥3 ng/mL, and life expectancy >10 years. With respect to life expectancy, we used the statistics provided by the SESCAM service (health service in the region of Castilla-La Mancha, Spain), which claims that patients less than 78 years old may be included in the study. All patients participated upon explicit agreement, and briefings were conducted in the primary care centers to inform doctors of recruited patients about this study. A transabdominal vesico-prostatic ultrasound (an echography of the prostate) and DRE were conducted on all patients included in the study. The former provides the variable volume, that is, the volume of the prostate measured in cubic centiliters (cc), and the latter gives the variable DRE, this is a qualitative variable that describes the feeling on the specialist's finger when touching the prostate gland. Table 1 shows all variables obtained for this study. A patient without cancer was tagged as BPH when having values: PSA ≥8 ng/mL, volume ≥30 cc, and DRE augmented. Thus, BPH is always diagnosed by symptomatology in primary care centers. The hospitals in which prostate biopsies were performed include in their protocol not to verify BPH diagnosis through histology.

Patients with urologic diseases other than PC and BPH were excluded (orchitis, prostatitis, and urinary tract infections were found), so we finally ran our study over $N = 125$ patients: 25 tagged as PC, 75 as BPH, and 25 as none. Following the protocol established at the urology service in the hospital in which this research was carried out, a follow-up of 6 months was made. So, in our database, patients without a biopsy had been followed for at least 6 months.

Variables. Prostate volume is the only variable included that cannot be regarded as an easy variable to obtain, since most primary care clinics do not have the necessary hardware to perform ultrasound scans. However, we included it because of the great importance it receives in the current literature for both PC and BPH, and moreover, if one has the necessary scanner, it is a cheap indicator to obtain.

The other variables acquired were: PSA, age, antecedents, DRE, meat consumption, physical activity, metabolic syndrome, origin, smoker, hematuria, body mass index (BMI), alcohol consumption, sexual activity, and the total sum of questions from IPSS. Questions and measurements were designed by the main investigator, and approved by the ethical board of the hospital.

Since the predictive models applied in our research use categorical variables, a new version of our numerical variables was added to our database by converting them to their multinomial version according to state-of-the-art values and the advice of a urologist from the same hospital where this study was conducted. All variables are shown in Table 1.

Under each urologist's criteria, referred patients underwent a fine-needle biopsy (six-cores or higher). If the biopsy was positive, they were tagged as PC. Patients who were not referred to the specialist and did not present PC for 2 years but suffered from BPH were tagged with the *BPH* diagnosis. Otherwise, they were tagged as *none*.

Statistical and predictive analysis. We ran Chi-square correlation tests for each predictive variable with the diagnosis variable (if a variable was numerical, we used its discretized version). For each numerical variable, the non-parametric Kruskal–Wallis analysis of variance test was used (after a prior Shapiro–Wilk test to reject the normal distribution assumption)

Table 1. Variables obtained from men included in the study. Mean and standard deviation are shown for numerical variables; frequency (%) is shown for each possible value of the categorical variables. The 95% CIs are computed with a bootstrap of 1000 samples.

VARIABLE	VALUES	MEAN/%	95% CI	STD DEV.
Antecedents	No	79.2	72.0–86.0	–
	1st grade (parents, brother, sister)	15.2	8.8–22.4	–
	2nd grade (grandparents, uncle, aunt)	5.6	1.6–9.6	–
Origin	City	67.2	59.2–75.2	–
	Town	32.8	24.8–40.8	–
	[50–59]	16.0	10.4–23.2	–
	[60–69]	48.0	40.0–56.0	–
Age (years)	[70–79]	30.4	23.2–38.4	–
	[80-]	5.6	2.4–10.4	–
Smoker	No	64.0	56.0–72.0	–
	Yes	36.0	28.0–44.0	–
Alcohol consumption	None	31.2	23.2–40.0	–
	Low	56.0	46.4–64.8	–
	High	12.8	7.2–18.4	–
BMI (kg/m^2)	Numerical/{low, normal, overweight, obesity}	28.56	27.8–29.4	4.27
Metabolic syndrome	No	84.8	77.6–90.4	–
	Yes	15.2	9.6 22.4	22.4
Meat consumption	None	0.0	0.0	0.0
	3–4 units/week	57.6	48.0–67.2	–
	Daily	42.4	32.8–52.0	–
PSA (ng/ml)	Numerical/{[3–6],[6–10],[11-]}	8.34	6.8–10.2	9.97
IPSS (points)	Numerical/{low, mid, high}	8.93	7.7–10.2	7.01
DRE	Normal	25.6	17.6–33.6	–
	Augmented	65.6	56.8–73.6	–
	Nodularity	8.8	4–14.4	–
Volume (cc)	Numerical/{[–30], [30–39], [40–50], [51-]}	57.0	51.8–62.9	32.78
Hematuria	Normal	75.2	68.0–83.2	–
	Evidence	8.0	4.0–12.8	–
	Microscopic	16.8	9.6–23.2	–
	Macroscopic	0.0	0.0	–
Physical activity (weekly sessions)	0–1	38.4	29.6–46.4	–
	2–4	43.2	34.4–52.0	–
	>= 4	18.4	12.0–25.6	–
Sexual activity	Low	31.2	23.2–38.4	–
	High	68.8	61.6–76.8	–
Diagnosis	PC	20.0	12.8–27.2	–
	BPH	60.0	52.0–68.8	–
	None	20.0	12.8–26.4	–

in order to compare its mean depending on diagnosis. All these methods were used as implemented by the R statistical package.

As explained in the Introduction, we chose two predictive models: decision tree and logistic regression. In particular, the decision tree built is C4.5,[27] which decides what variables to use in the tree based on an embedded feature-selection process regarding the reduction in uncertainty about the target variable (diagnosis). The well-known logistic regression model is constructed with a forward stepwise feature-selection process, which adds features to the final model based on an inner cross-validation driven by accuracy.

The metrics obtained to measure the performance of the two models are: accuracy (success rate) and area under the ROC curve (AUC).[28] Accuracy is a very common and intuitive

metric, which consists in dividing the number of correct predictions by the total number of predictions performed, and so it ranges from 0 to 100%. Of course, in contexts such as medical applications, the class or outcome variable presents a skewed distribution. Thus, if the classifier always predicts the skewed value of the class variable, its accuracy is high but less frequent values are never predicted correctly. However, AUC is insensitive to class skewness and changes of the class distribution from the training to the test sets; that is why it is very commonly used in classification problems related to medicine. It ranges from 0 to 100%, and it can be interpreted as the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.[28]

These metrics were computed with a special case of cross-validation: leave-one-out (LOO) validation. This kind of validation returns a very small biased, but with the potential of having a greater variance,[29] as can be seen in the standard deviation of the accuracies obtained in Section 3. However, it has been shown that this variance is not greater, theoretically, than the variance of the hold-out validation method,[30] and LOO is recommended for small sample sizes rather than common 10-cross-validation.[31] For comparison purposes with other prediction methods found in the literature, we use individual AUC for PC and BPH. We cannot use accuracy since this merges together the predictions for PC and BPH, and to the best of the authors' knowledge there does not exist any article that builds predictive models for these two pathologies.

All the predictive analysis methods referred to were used as implemented by the Weka toolkit, using the default parameters of classifiers C4.5 (decision tree) and Logistic. With respect to the validation process, the default seed (1) was used.

Results

Prior to building the predictive models, we ran statistical tests to identify categorical variables related to diagnosis. The aim of these tests is to search for variables that are statistically correlated with the variable diagnosis, which is categorical. Some of the predictive variables are categorical, and while others are numerical. Depending on the nature of these variables, a different test must be run.

Statistical analysis. Table 2 shows the well-known Chi-square test of independence for diagnosis and each one of the other categorical variables. Table 3 shows the results of a non-parametric group test for the difference in the mean of each numerical variable, given each of the three possible values of diagnosis. Although the purpose of Figure 1 is mainly descriptive, it can also be used as a support to the statistical findings in Table 2.

Categorical variables that proved to be correlated with diagnosis are shown in Table 2, volume, DRE, and PSA being the three variables with a more significant (and a stronger) correlation. Even using the most-conservative Bonferroni correction, the three variables identified as being the most relevant would still continue being so.

Table 2. Chi-squared correlation test for categorical (and discretized-numeric) variables with respect to diagnosis. Results shown are: *P*-value, the correlation test statistic, and the strength of the correlation as computed by Cramer's V (range from 0 to 1).

VARIABLE	*P*-VALUE	χ^2	CRAMER'S V
Volume (discretized)	4.21E-11	60.141	0.470
DRE	4.22E-09	44.878	0.408
PSA (discretized)	1.18E-07	42.983	0.398
Age	0.002	21.315	0.230
IPSS (discretized)	0.002	16.633	0.247
Antecedents	0.006	14.311	0.229
Meat consumption	0.028	10.878	0.200
Smoker	0.193	3.293	0.156
Origin	0.343	2.142	0.126
BMI (discretized)	0.373	4.256	0.125
Hematuria	0.564	4.844	0.133
Alcohol consumption	0.572	4.783	0.133
Metabolic syndrome	0.743	0.594	0.066
Sexual activity	0.788	0.477	0.062
Physical activity	0.805	1.623	0.077

With respect to numerical variables, a significant difference in their means with respect to diagnosis was found for volume, PSA, and IPSS. Since diagnosis has three possible values (PC, BPH, and none), we show in Table 3 the 95% confidence interval (CI) of the difference in the mean for each variable, comparing the mean for PC with respect to the other two factors.

Volume, PSA, and IPSS boxplots are shown in Figure 1 for each possible diagnosis. Outliers were checked and we did not remove them since they belong to actual and correct data. We can see that the PSA and volume distributions show very little overlapping for each type of diagnosis, so we can regard them as the stand-alone variables with the highest discriminating power. Although IPSS was found to be statistically different for each diagnosis, we see a huge overlap between PC and BPH, so it might not be very helpful when trying to differentiate those diagnoses.

Table 3. *P*-value and Chi-squared statistic of the Kruskal–Wallis non-parametric analysis of variance for numerical variables with respect to diagnosis. Results are shown only for the three variables for which the test was significant. The last two rows show the 95% CI of the difference of the mean of each variable, given PC and the other two possible diagnosis.

	VOLUME	PSA	IPSS
P-value	8.19E-12	1.25E-09	5.01E-05
Kruskal Wallis χ^2	51.057	41.004	19.804
BPH-PC 95% CI	[10.44, 40.71]	[–13.78, –6.21]	[0.44, 6.34]
None-PC 95% CI	[–34.93, 2.20]	[–17.11, –7.82]	[–6.80, 0.43]

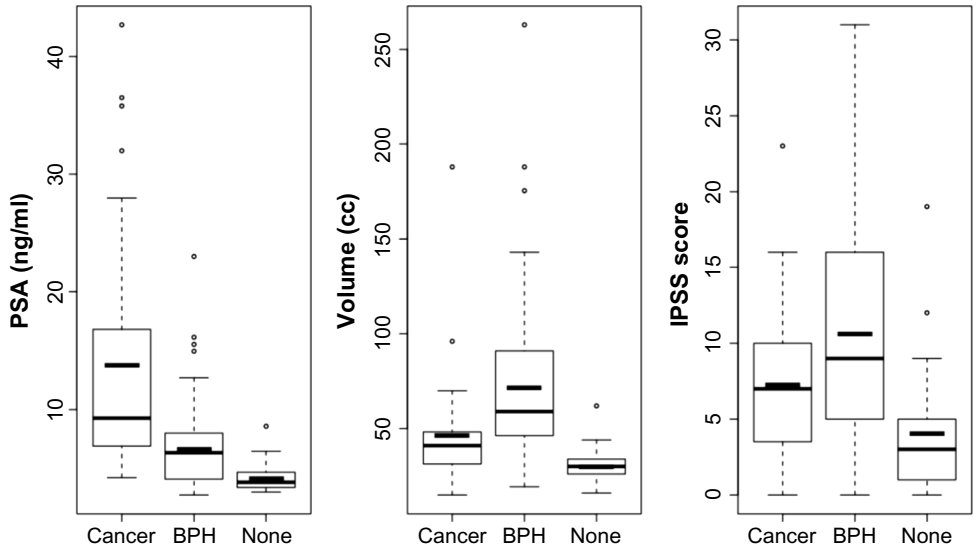

Figure 1. Boxplots of PSA, volume, and IPSS, for each possible diagnosis. The short segment inside each box represents the mean of the distribution, and the long segment represents the median.

Thus, after performing a statistical study, we found that the variables that are marginally related to diagnosis are: volume, PSA, IPSS (in both their numeric and categorical versions), DRE, age, antecedents, and meat consumption. Then, we conducted a predictive analysis using these variables as a start set, from which the models selected those which help them to improve their predictive performance the most.

Decision tree predictive model. The decision tree construction process runs an internal (embedded) feature-selection process, which finally selects those variables that, in turn, reduce the uncertainty about the value of the target variable (diagnosis). When the inclusion of other variables does not reduce this uncertainty, then no more variables are selected. It is important to specify that the order of variables from the root node does not imply the order in which variables need to be acquired; all variables in the tree need to be obtained. The order among nodes is just the order of the inference of explanation of the output predicted.

Our decision tree (see Fig. 2) selected, from the start set defined above, the variables volume, PSA, and DRE, in that order. The variables volume and PSA are numeric, so the decision tree discretized them following the well-known discretization algorithm presented in ref. 32. After validation, the mean accuracy obtained is 79.2% with standard deviation SD = 40.7%. With respect to AUC, prediction for both PC and BPH was over a mean of 72% (SD = 1.9%), it being a little higher for normal.

Volume and PSA were selected in their numerical version, and the construction algorithm automatically discretized them in the optimal intervals with respect to diagnosis. Thus, volume became binomial with 43 cc as cut-off point; and PSA also became binomial with different cut-off points depending on the level of the tree: 5.63 ng/mL if volume <= 43 cc, and 23 ng/mL if volume >43 cc. For the case of volume <= 43 cc and PSA <= 5.63 ng/mL, the tree differentiates a second-level cut-off point depending on the DRE result: 3.72 ng/mL for DRE = augmented, and 4.69 for DRE = normal.

Logistic regression predictive model. We ran the construction method of a logistic regression model with a greedy forward stepwise process, using as candidates the same variables from the start set as for the decision tree model, and AUC as metric for the evaluation of feature subsets. The variables selected were volume, PSA, IPSS in their numerical version, and meat consumption. We found that if we dropped meat consumption from the model, AUC for PC increased by around 15%; thus we rejected it manually. Consequently, IPSS is the only different variable not selected in the decision tree (which used the DRE result instead). The logistic model and the corresponding odds ratios are shown in Table 4. This model, whose resulting equations are shown in Figure 3, obtained 74.4% accuracy with SD = 44.3%, and an AUC for each diagnosis value of over 80.0% (SD = 4.73), it being very similar for PC and BPH.

The odds ratio obtained in the model indicates that:

- For a patient with a given prostate volume, each 1-cc increase in such volume would increase the probability of PC by 9.6% (95% CI: 2.4–17.4%) and that of BPH by 13.3% (95% CI: 5.9–21.1%), with respect to not suffering from any of these pathologies.
- For a patient with a given PSA, each 1-ng/mL increase would double their previous probability of PC (95% CI: 43.8–280%) and BPH (95% CI: 15.9–200%), with respect to not suffering from any of these pathologies.
- Given two patients with the same volume and PSA values, the probability of one patient having PC increases for each extra point obtained in the IPSS questionnaire by 13.4% (95% CI: –2.6 to 32.1%); while the probability

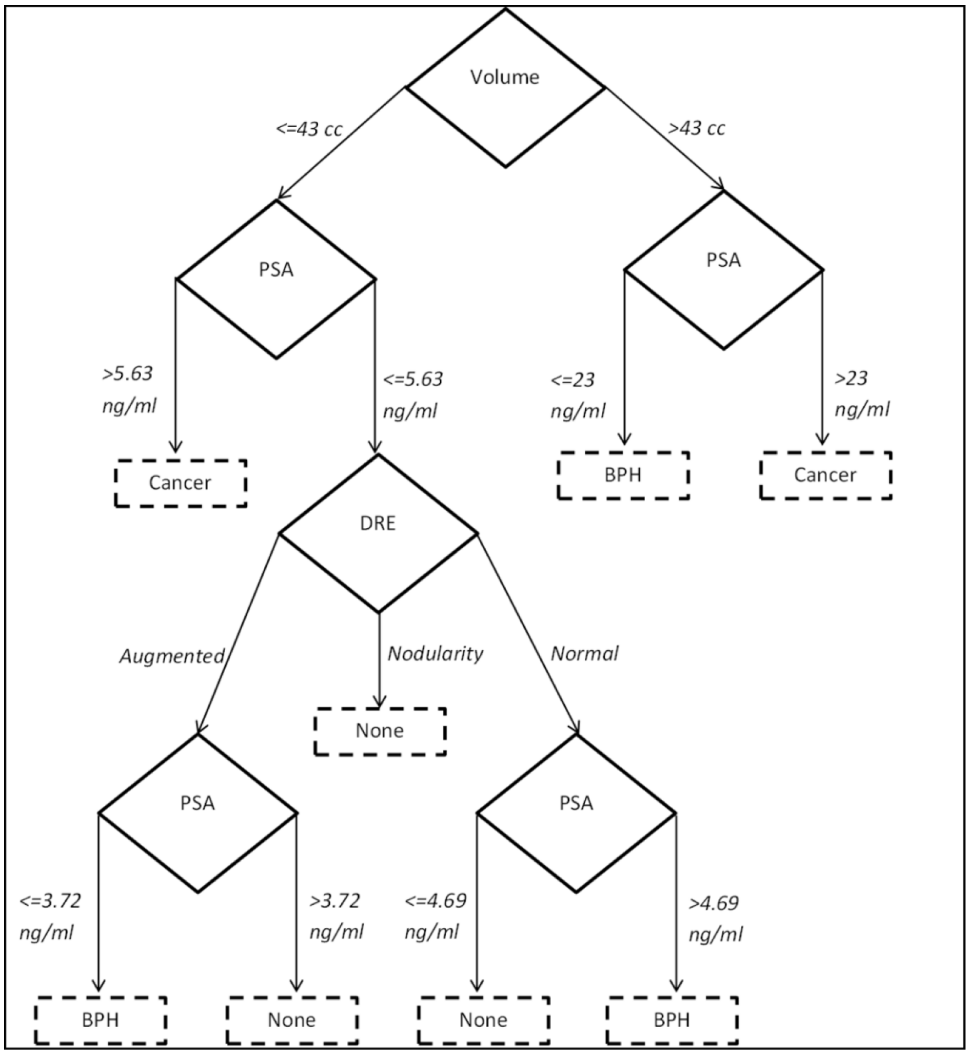

Figure 2. Decision tree: the diagnosis is predicted in the lean nodes, based on the patient's values for volume, DRE, and PSA. 79.2% accuracy, and AUC for cancer, BPH, and none is 71.4, 74.1, and 81.7%, respectively.

of BPH increases by 19.1% (95% CI: 3.7–36.8%), with respect to the probability of not suffering from any of these pathologies.

Finally, Figure 4 shows the ROC curves and their area values for each of the predictive models and the PC and BPH diagnosis (AUC is known to be equivalent to a Mann–Whitney *U* test). The logistic regression model obtains a higher AUC for both PC and BPH; however, the decision tree obtains a higher mean accuracy. This is because of its better behavior at predicting the *none* factor of diagnosis; that is, the decision tree gives less false negatives with respect to both pathologies. For the sake of clarity, we do not show the ROC curve for *none*, since it adds no extra information because its AUC is always better than the rest of the AUCs, in both predictive models.

With respect to accuracy, given the special case of LOO validation, the paired comparison between classifiers is not a comparison of means but a comparison of a binomial result: success or failure at classifying one instance

Table 4. Coefficients and odds ratio logistic regression model, the reference value is diagnosis = normal.

VARIABLE	COEFFICIENT		ODD RATIO		*P*-VALUE		ODD RATIO 95% CI	
DIAGNOSIS	P. CANCER	BPH	P. CANCER	BPH	P. CANCER	BPH	P. CANCER	BPH
Intercept (B_0)	−8.939	−8.159	–	–	–	–	–	–
Volume	0.092	0.125	1.096	1.133	0.009	0.000	[1.024, 1.174]	[1.059, 1.211]
PSA	0.850	0.623	2.339	1.865	0.001	0.010	[1.438, 3.805]	[1.159, 3.001]
IPSS	0.126	0.175	1.134	1.191	0.105	0.013	[0.974, 1.321]	[1.037, 1.368]

$$p(\textbf{Cancer}) = \frac{e^{-8.939 + 0.092\times\textbf{Volume} + 0.850\times\textbf{PSA} + 0.126\times\textbf{IPSS}}}{1 + e^{-8.939 + 0.092\times\textbf{Volume} + 0.850\times\textbf{PSA} + 0.126\times\textbf{IPSS}} + e^{-8.159 + 0.125\times\textbf{Volume} + 0.623\times\textbf{PSA} + 0.175\times\textbf{IPSS}}}$$

$$p(\textbf{BPH}) = \frac{e^{-8.159 + 0.125\times\textbf{Volume} + 0.623\times\textbf{PSA} + 0.175\times\textbf{IPSS}}}{1 + e^{-8.939 + 0.092\times\textbf{Volume} + 0.850\times\textbf{PSA} + 0.126\times\textbf{IPSS}} + e^{-8.159 + 0.125\times\textbf{Volume} + 0.623\times\textbf{PSA} + 0.175\times\textbf{IPSS}}}$$

$$p(\textbf{None}) = 1 - p(Cancer) - p(BPH)$$

Figure 3. Equations in the logistic regression model. By selecting as outcome the diagnosis value with a greater probability, this model obtained 74.4% accuracy, and for each possible diagnosis {PC, BPH, and none}, the AUC obtained was 83.3%, 85.3%, and 92.4%.

(the test set only has one instance in LOO). Thus, the statistical comparison of accuracy is performed by means of the Chi-square independence test using the contingency table of successes and failures. The result is that the accuracy obtained by the two classifiers is not statistically different (P-value = 0.008).

Discussion

This study presents some limitations, which should be taken into account when interpreting the results obtained.

1. The sample size after cleaning the database is 125; and it was obtained from a single center. The fact of having a not very large sample size has one main drawback, which is the existence of some inadequacies (paths in the tree with less clinical sense than others), which are created by the tree because it finds that the entropy (uncertainty) in data is reduced in that way. We present in this paper the decision tree that resulted in the best performance after validation. Another option would be to present the tree with most clinical adequacy.
2. We did not exclude patients with PSA >10 ng/mL. This range is commonly excluded from PC prediction studies since those values are related to having PC. However, as mentioned in the Introduction, using PSA as a sole marker for PC prediction has been shown not to have enough predictive power. Furthermore, patients with PSA >10 may present BPH but not PC, and our goal is to differentiate BPH from PC. In fact, our sample contains 13 patients with PSA in range,[10–20] nine of them being tagged as BPH but not PC (of those nine, four were not even recommended for biopsy by the urologist).
3. Biopsies were performed only on those patients who met the criteria for biopsy, in order to avoid unnecessary pain. Since the study was performed using primary care consultations, the criteria to perform a biopsy depended on the urologist who attended the patient at the hospital; commonly, these criteria are suspicious DRE, PSA >10, PSA >= 4, and free PSA <20%.

We studied 15 predictive variables for diagnosis that are not very expensive and are easy to acquire from primary care consultations (all except volume in many primary care buildings). Although the ultrasound scanner is not available in all primary care clinics, our results find that this variable is the most significant (more than PSA) for PC/BPH prediction, and so primary care doctors would find it a valuable aid when deciding whether to refer the patient to a urologist. Furthermore, the urologist could use any of the self-explainable predictive models developed in this work in order to make the final decision about the necessity of a surgical biopsy, which is a painful and infection-prone procedure.

Volume, DRE, PSA, age, IPSS, antecedents, and meat consumption were found to be the only variables with

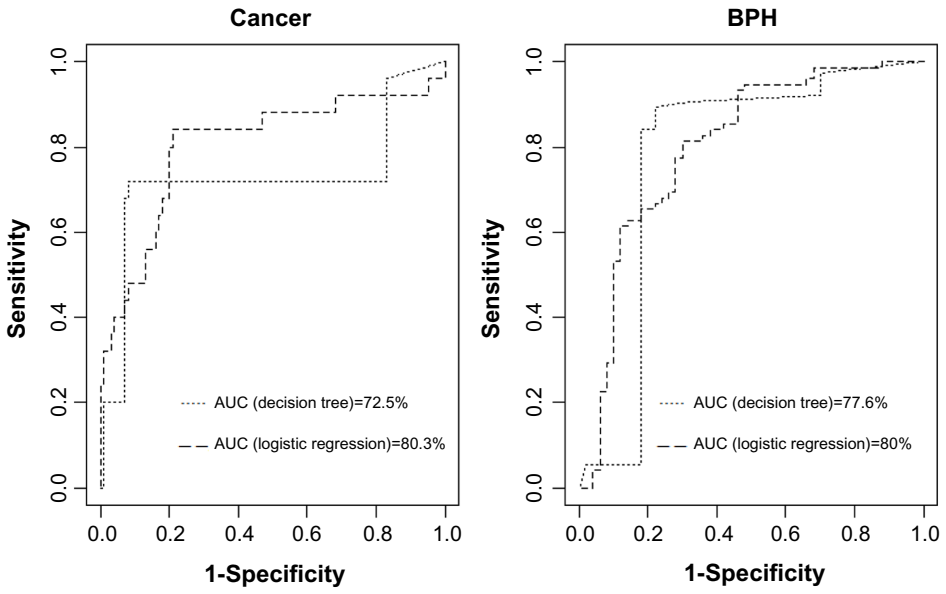

Figure 4. ROC curves for PC (left) and BPH (right) diagnosis. Dotted and segmented lines represent the curves for the decision tree and the logistic regression predictive models, respectively.

significant marginal correlation with diagnosis. However, the predictive models that were constructed finally selected only volume, PSA, and DRE (decision tree) or IPSS (logistic regression), three of which are numerical (DRE is categorical). The Kruskal–Wallis test and CIs for the difference of means (Table 3) show that the mean prostate volume is higher in BPH patients than in those with PC while, in contrast, the volume is higher in PC cases than in patients without any of these pathologies. With respect to PSA, this is greater in patients with PC than in patients with BPH or without any of these pathologies; furthermore, this difference of mean PSA is lower in the BPH-none cases. Moreover, in Figure 1 we find that boxplots for PSA and volume do not overlap for any diagnosis; thus, this corroborates with our assumption that PC and BPH can be predicted with the same predictive model (a distinction that is necessary[7]), and we share evidence with another work[8] that relates increase in volume with decrease in PC risk, but this requires further study since it may be the case that a six-core biopsy finds it easier to find PC tissues in a low-volume prostate.

Although our models are used to predict yes/no outcomes, soft models such as logistic regression are capable of predicting a numerical probability (0 to 1) for each possible diagnosis, which is useful if the medical user would rather not have just one rigid prediction. With respect to IPSS, it becomes clear from both the CIs and the boxplot that it might be of some help in distinguishing among the possible diagnoses, but there is a big overlap between BPH and PC, and PC and none. Furthermore, the distribution of IPSS in BPH patients is very widely spread, and so this variable should definitively be used together with other predictive variables. In fact, it is discarded by the decision tree embedded feature-selection process, although it is selected in the logistic regression model (but it is not significant with respect to PC, as shown in the *P*-value column in Table 4).

As we explained in the Introduction, the clinician needs the predictive model used as decision support to be interpretable, as is the case of decision trees and logistic regression. Maybe the decision tree is nearer to a natural language explanation of the inference process: starting from the root node, and following the corresponding branches according to the value of the variable in the node, a final diagnosis (leaf node) is reached, and the path followed is in fact the explanation of such a prediction. On the other hand, the logistic regression model is able to express the influence of variables on the diagnosis by means of its coefficients and the consequent odds ratios. Furthermore, the resulting equations can be easily inserted in any spreadsheet to easily compute the prediction outcome of a new patient.

For the sake of comparison with other state-of-the-art studies, we computed the AUC metric for PC and BPH in each model. In a study by Chun et al.[12], nomogram and black-box (neural network) prediction models were compared using the AUC metric for PC. Our decision tree and logistic regression obtained 75.1% and 80.3% AUC, respectively; that is, there is a 75.1% (or 80.3%) probability of correctly classifying a random patient with PC or BPH, instead of classifying another healthy random patient as having a prostate disease. However, ref. 12 reports an AUC for the nomogram and neural network of 70.6% (using variables age, DRE, PSA, and %free PSA) and 67.0% (using the variables age, DRE, PSA, %free PSA, and TRUS-measured volume), respectively. In another work by Garzotto et al,[8] a CART (tree) and a logistic regression were compared, and they obtained a PC AUC of 74% and 72%, respectively, using in both models the variables PSA, PSA density, presence of a TRUS hypoechoic lesion, age, and volume. Although our results are slightly better, this should only be interpreted as a heuristic comparison, since validations of cited works are performed on different patients, sample sizes, and set of predictive variables.

Consequently, we suggest the use of self-explainable predictive models as decision support in order to decide on the referral of patients for prostate biopsy. We tried to reject prostate volume (it is not easy to obtain from a primary care context, although it is cheap) as a predictive variable, but performance decreased significantly for BPH, a pathology that must be predicted and contrasted against PC. Some patients may present a high PSA and be recommended for biopsy, while by using this indicator together with volume, the clinician could tell whether the actual pathology of the patient is BPH, which is a probable scenario as we found in Figure 1 and depicted in the decision tree model. This may help to reduce the number of unnecessary surgical procedures, consequences of which comprise those related to diagnosis and treatment.

As future work, we intend to enlarge our sample with more patients from different health centers. Thus, we may obtain greater scientific evidence for the possibility of differentiating multivariate symptoms for PC and BPH prediction, thus avoiding unnecessary surgery. Once models are built with a greater sample, they could be validated in the context of primary care consultations with new patients.

Author Contributions

Conceived and designed the experiments: PB, AV, PJT, JAR Analyzed the data: PB, AV. Wrote the first draft of the manuscript: PB, AV. Contributed to the writing of the manuscript: PB, AV, PJT, JAR. Agree with manuscript results and conclusions: PB, AV, PJT, JAR. Jointly developed the structure and arguments for the paper: PB, AV, PJT, JAR. Made critical revisions and approved final version: PB, AV, PJT, JAR. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Ferlay J, Autier P, Boniol M, Heanue M, Colombet M, Boyle P. Estimates of the cancer incidence and mortality in Europe in 2006. *Ann Oncol*. 2007;18:581–92.
2. Landis S, Murray T, Bolden S, Wingo P. Cancer statistics. *A Cancer J Clin*. 1999;49:8–31.

3. Brawer MK. Prostate-specific antigen: current status. *A Cancer J Clin*. 1999;49: 264–81.
4. Mochtar C, Kiemeney L, Laguna M, Debruyne F, de la Rosette J. PSA velocity in conservatively managed BPH: can it predict the need for BPH-related invasive therapy? *Prostate*. 2006;66:1407–12.
5. Merrick G, Butler W, Wallner K, Lief J, Hinerman-Mulroy A, Galbreath R. Prostate-specific antigen (PSA) velocity and benign prostate hypertrophy predict for PSA spikes following prostate brachytherapy. *Brachtherapy*. 2003;2:181–8.
6. Yuksel S, Dizman T, Yildizdan G, Sert U. Application of soft sets to diagnose the prostate cancer risk. *J Inequal Appl*. 2013;229.
7. Meigs J, Barry M, Oesterling J, Jacobsen S. Interpreting results of prostate-specific antigen testing for early detection of prostate cancer. *J Gen Intern Med*. 1996;11:505–12.
8. Garzotto M, Beer T, Hudson R, et al. Improved detection of prostate cancer using classification and regression tree analysis. *J Clin Oncol*. 2005;23:4322.
9. Jung K, Elgeti U, Lein M, et al. Ratio of Free or Complexed Prostate-specific Antigen (PSA) to Total PSA: which ratio improves differentiation between benign prostatic hyperplasia and prostate cancer? *Clin Chem*. 2000;46:55–62.
10. Filella X, Alcover J, Molina R, Corral JM, Carretero P, Ballesta AM. Measurement of complexed PSA in the differential diagnosis between prostate cancer and benign prostate hyperplasia. *Prostate*. 2000;42:181–5.
11. Catalona W, Partin A, Slawin K, 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. *J Am Med Assoc*. 1998;279:1542–7.
12. Chun F, Karakiewicz P, Briganti A, et al. A critical appraisal of logistic regression-based nomograms, artificial neural networks, classification and regression-tree models, look-up tables and risk-group stratification models for prostate cancer. *Br J Urol Int*. 2007;99:794–800.
13. Mitchel T. *Machine Learning*. McGraw-Hill; 1997.
14. Carsten S, Cammann H, Semjonow A, 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.
15. Seker H, Odetayo M, Petrovic D, Naguib R. A fuzzy logic based-method for prognostic decision making in breast and prostate cancers. *IEEE Trans Inf Technol Biomed*. 2003;7:114–22.
16. Saritas I, Allahverdi N, Sert I. A fuzzy expert system design for diagnosis of prostate cancer. In: International Conference on Computer Systems and Technologies, New York; 2003:341–51.
17. Benecchi L. Neuro-fuzzy system for prostate cancer diagnosis. *Urology*. 2006;68:357–61.
18. Shariat S, Karakiewicz G. Comparison of nomograms with other methods for predicting outcomes in prostate cancer: a critical analysis of the literature. *Clin Cancer Res*. 2008;14:4400–7.
19. Shipley W, Thames H, Sandler H, et al. Radiation therapy for clinically localized prostate cancer: a multi-institutional pooled analysis. *JAMA*. 1999;281:1598–604.
20. Ghafar M, Golliday E, Bingham J, Mansukhani M, Anastasiadis A, Katz A. Regression of prostate cancer following administration of Genistein Combined Polysaccharide (GCP), a nutritional supplement: a case report. *J Altern Complement Med*. 2002;8:493–7.
21. Lacave C, Luque M, Díez FJ. Explanation of Bayesian Networks and Influence Diagrams in Elvira. *IEEE Trans Syst Man Cybern B Cybern*. 2007;37:952–65.
22. Lacave C, Díez FJ. Knowledge Acquisition in PROSTANET – A Bayesian Network for Diagnosing Prostate Cancer. In: Knowledge-Based Intelligent Information and Engineering Systems (KES), Oxford; 2003:1345–50.
23. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2012. Report No.: ISBN 3–900051–07–0.
24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: an Update. *SIGKDD Explor*. 2009.
25. IBM Corp. *IBM SPSS Statistics for Windows, Version 21.0*. Armonk, NY: IBM Corp; 2012.
26. Sakellariou CAS, Galustian C, Elhage O, et al. Expression of IL-15 and IL-2 cytokines and receptors on prostate cancer cells–significance in prostate immunotherapeutics. *Br J Surg*. 2013;100:67–8.
27. Quinlan R. Induction of decision trees. *Mach Learn*. 1986;1:81–106.
28. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;8:861–74.
29. Kohav R. 14th International Joint Conference on Artificial Intelligence. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, Stanford, CA; San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1995:1137–43.
30. Elisseeff A, Pontil M. Leave-one-out error and stability of learning algorithms with applications. In: Suykens J. et al. ed. *Advances in Learning Theory: Methods, Models and Applications, NATO Science Series III: Computer and Systems Sciences*, Vol. 190. IOS Press.
31. Molinaro A, Simon R, Pfeiffer R. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 2005;21(15):3301–7.
32. Fayyad U, Irani K. Multi-interval discretization of continuous valued attributes for classification learning. In: Thirteenth International Joint Conference on Artificial Intelligence, San Mateo, CA; 1993: 1022–7.