Regnier-Coudert O, McCall J, Lothian R, Lam T, McClinton S, N'Dow J. Machine learning for improved pathological staging of prostate cancer: A performance comparison on a range of classifiers. *Artificial Intelligence in Medicine* 2012;**55**:25-35.

This is the final draft, after peer-review, of a manuscript published in *Artificial Intelligence in Medicine*. The definitive version, detailed above, is available online at www.elsevier.com

Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers.

Olivier Regnier-Coudert[a], John McCall[a], Robert Lothian[a], Thomas Lam[b], Sam McClinton[b], James N'Dow[b]

[a] *IDEAS Research Institute, Robert Gordon University, Aberdeen, United Kingdom*
[b] *Academic Urology Unit, University of Aberdeen, Aberdeen, United Kingdom*

Abstract

Objectives: Prediction of prostate cancer pathological stage is an essential step in a patient's pathway. It determines the treatment that will be applied further. In current practice, urologists use the pathological stage predictions provided in Partin tables to support their decisions. However, Partin tables are based on logistic regression (LR) and built from US data. Our objective is to investigate a range of both predictive methods and of predictive variables for pathological stage prediction and assess them with respect to their predictive power based on UK data.

Methods and Material: The latest version of Partin tables was applied to a large scale British dataset in order to measure their performances by mean of concordance index. The data was collected by the British Association of Urological Surgeons (BAUS) and gathered records from over 1700 patients treated with prostatectomy in 57 centers across UK. The original methodology was replicated using the BAUS dataset and evaluated using concordance index. In addition, a selection of classifiers, including, among others, LR, artificial neural networks and Bayesian networks (BN) was applied to the same data and compared with each other using an ROC analysis. Subsets of the data were created in order to observe how classifiers perform with the inclusion of extra variables. Finally a local dataset prepared by the Aberdeen Royal Infirmary was used to study the effect on predictive performance of using different variables.

Results: Partin tables have low predictive power when applied on UK data for comparison on patients with organ confined and extra prostatic extension conditions, patients at the two most frequently observed pathological stages. The use of replicate lookup tables built from British data shows an improvement in the classification, but the overall predictive power remains low.
Comparing a range of classifiers shows that BNs generally outperform other methods. Using the four variables from Partin tables, Naive Bayes is the best classifier for the prediction of each class label. When two additional variables are added, the results of LR, artificial neural networks and BN methods are overall improved. BNs show higher AUC than the other methods when the number of variables raises.

Conclusion: The predictive power of Partin tables can be described as low to moderate on UK data. This means that following the predictions generated by Partin tables, many patients would received an inappropriate treatment, generally associated with a deterioration of their quality of life. In addition to demographic differences between UK and the original US population, the methodology and in particular LR present limitations. BN represents a promising alternative to LR from which prostate cancer staging can benefit. Heuristic search for structure learning and the inclusion of more variables are elements that further improve BN models quality.

Keywords: Classification, Bayesian networks, Logistic regression, Prostate cancer staging, Partin tables

---

1. Introduction

Cancer is a widely spread disease responsible for many deaths all over the world. In 2008, the World Health Organization estimated the number of new cancer cases in the world to be over 7.5 million [1]. Among all types of cancer, prostate cancer is the most frequent in men. In 2008, around 900 000 new cases of prostate cancer were diagnosed, and approximately 260 000 men died from it over the same period [1]. In Britain, the same source shows that 37 000 men were affected with new occurrence of prostate cancer, accounting for nearly a quarter of all male cancer diagnosed annually. It is also the second commonest cause of cancer death in men in the UK after lung cancer [1].

This paper considers the use of different machine learning techniques in

order to improve the prediction of pathological stage in prostate cancer. A UK-wide dataset collected by the British Association of Urological Surgeons (BAUS) is used in order to build and assess predictive models. Results are compared with each other, but also against those of tools and methods currently in use clinically. Machine learning gathers a wide range of methods, particularly for classification purposes. Performance studies of such methods on different applications is an essential step towards the optimization of predictive tools. In [2], classifiers are compared with each other with respect to their performance on predicting different outcomes related to pancreatic cancer, including cancer staging. Similarly, in [3], classifiers were applied on breast cancer patient data to improve survivability prediction. In [2], models built using bayesian techniques and logistic regression presented the best prediction for the different outcomes while decision trees performed best in [3]. This highlights the importance of comparing methods on different domain as the most adapted technique can vary across them.

Partin tables [4] are the most commonly used tool for prostate cancer staging. They were originally created using Logistic Regression (LR) [5] on a database gathering records of patients that were treated with radical prostatectomy in a single US institution [4]. Since then, the tables have been updated using different up-to-date datasets [6–8]. The revision takes into account changes in population demographics, advances in health technology and improved health care systems, but the tables are still based on the same fundamental LR-based methodology.

Partin tables are a well-established and most widely used pathological staging tool in the urological community worldwide. However, concerns have been raised regarding their validity on non-US populations [9–19]. In some instances, Partin tables were considered to be unsuitable for the target population because of limitations with respect to their predictive power [9, 11, 16–19]. The appropriateness of the methodology behind Partin tables, especially in regard to the choice of predictive variables and classifier, was not addressed in those studies. In addition, it is widely recognized that prostate cancer staging is associated with a high level of uncertainty. All these considerations are compelling clinicians to explore alternative means of generating predictive tools, especially those which apply machine-learning techniques which have the potential of improving the quality and accuracy of predictive performance [20].

In this paper, we assess the Partin tables on a British population. Replicate lookup tables, based on British data are built and analyzed following the

original approach given in [8]. Additionally, we propose and compare alternative classification techniques, including Bayesian Networks (BN), which by simplifying the probability distribution, are recognized as a reference method to reason under uncertainty [21]. In addition, we consider more variables to include in the model. The paper is divided as follows. Section 2 presents important details related to the understanding of prostate cancer and its treatments. Section 3 describes the methodology of the study. It presents the different objectives and the data, provides technical background knowledge on classification and describes the experiments. Results are presented in section 4 and discussed in section 5. Finally, we describe our conclusions and introduce ideas for further work in section 6.

2. Medical Background

Cancer is a disease where malignant cells are developed and alter the function of their hosting organs or tissues. Typically, malignant cells reproduce and group together to form a tumor. Untreated tumors grow and affect surrounding healthy cells, leading to a spread of the cancer. Metastasis happens when the cancer reaches surrounding organs or tissues. The presence of cancer results in the deterioration of some body functions and can lead to death when vital organs are touched.

Whilst in the past prostate cancer was a disease which predominantly affected older men well into their seventies, the advent of Prostate Specific Antigen (PSA) testing over the past three decades has caused a shift in the age of presentation, such that men in their early fifties are increasingly being diagnosed. In addition, PSA testing has also resulted in a stage migration from late, symptomatic stages to early, asymptomatic stages of the disease. Men with raised PSA would then undergo prostate biopsies which will confirm the diagnosis and provide a grade of the disease expressed by means of the Gleason sum score (GS), with the grade of the cancer reflecting its aggressiveness. Once the diagnosis is confirmed, a digital rectal examination (DRE) is performed to assess the local clinical stage (CS). The stage is a means of indicating the spread of the disease, expressed by the TNM staging system [22], whereby the T stage refers to the local extent of spread. The treatment options available for localised prostate cancer include surgery, external beam radiotherapy, brachytherapy, active monitoring and minimally invasive localised therapy such as cryotherapy [23]. Surgery by way of radical prostatectomy, where the prostate is surgically removed completely, is one of the

leading options. Although most of the curative treatment options result in similar cure rates, surgery has the major advantage of removing the prostate completely as well as providing the actual pathological stage and grade of the disease, which in turn influence prognosis. The pathological stage is the most accurate determination of the actual stage of the disease (as opposed to clinical stage which is an estimate), being determined by pathological examination of the entire prostate specimen. However, the major drawback of surgery is its associated adverse effects, such as intra-operative complications (e.g. bleeding), prolonged hospital stay, urinary incontinence and erectile dysfunction. The pathological stage of prostate cancer significantly influences the prognosis; the presence of extra-prostatic extension reduces the chance of cure and increases the risk of adverse effects. Consequently, surgery may not be appropriate for every man with prostate cancer, and those with more advanced disease should be offered other options instead. Such decision-making crucially relies on the prediction of the likely pathological stage. It is for this purpose that predictive staging tools were created. Partin tables [4] are the most commonly used tool for prostate cancer staging. The tables are a means of predicting the likely pathological stage of the cancer using the pre-treatment variables of PSA, GS and CS, with the result being expressed as probabilities. Based on a patients PSA, GS and CS, probabilities are provided for each of four discrete pathological stage outcomes: organ confined (OC), extra-prostatic extension (EPE), seminal vesicle involvement (SVI) and lymph node involvement (LNI). The predicted probabilities of pathological outcomes are displayed by means of look-up tables organised according to the three pre-treatment variables, which are in turn divided into sub-groups.

3. Methodology

3.1. Objectives

This work introduces three main objectives that present interests for both medicine and machine learning communities.

First, we aim to critically assess the methodology which was used to construct Partin tables. This involves externally validating the version currently being used by practitioners, that is, studying how well it performs on a population that presents different characteristics. Here, the tool is evaluated on a large British cohort and results are compared to those of its internal

validation given in [8], where the original data was also used for testing. Using the British data and the approach described in [8], we build new lookup tables and assess the methodology itself. The results are compared against the previous validation studies [9–19] and provide additional understanding on Partin tables performances.

Second, we propose alternative classifying techniques to build lookup tables for prostate cancer staging. We run many classifiers, including LR, on our data and study the performances of the models produced by each. We compare the different methods with respects to their predictive power and propose alternatives to LR.

Finally, we investigate the impact of new variables being introduced into the model. Two different datasets are used for this purpose each using distinct set of predicting variables. Among them, variables that were originally excluded when Partin tables were built are considered. In [4], patient's age was tested against other combination of variables and did not show statistically significant improvement to the LR-based model. A range of classifiers is considered and applied to different subsets of data in order to observe the impact that inclusion of elements can have.

3.2. Data

BAUS gathered clinical and pathological data on over 7500 patients that were received with prostate cancer and underwent radical prostatectomy in one of the 57 different centers of the study between 1999 and 2008. This accounts for approximately 20% of the total number of prostatectomies that were performed in the whole of UK over this period [24]. The BAUS dataset can be considered as large and representative of the British population and consequently, well suited for the assessment of Partin tables for a use in the UK.

Variable Name	**Categories**
PSA	0-2.5, 2.6-4.0, 4.1-6.0, 6.1-10.0, >10.0
GS	5-6, 3+4, 4+3, $\geq$8
CS	T1c, T2a, T2b/c
PS	OC, EPE, SVI, LNI

From the original BAUS dataset, two subsets were created to meet the different objectives. To construct the first one, we only kept the records

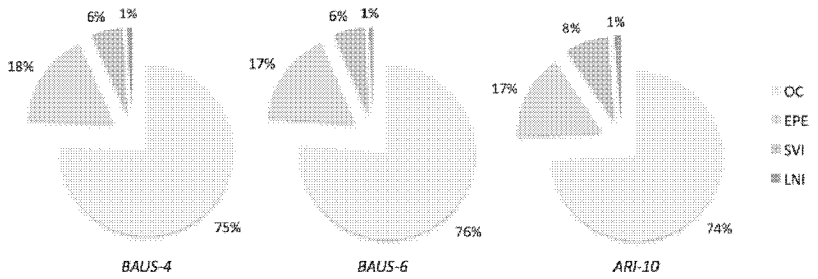


Figure 1: Distribution of the class variable pathological stage in the three datasets

where PSA, GS, CS and PS were set in order to match Partin tables variable settings. We call this dataset BAUS-4. Each variable in BAUS-4 was discretized following Partin method as described in Table 1. The final size of BAUS-4 was 1701 records, following the removal of cases where data was missing for any of the four variables and where input data was erroneous. The distribution of PS in the prepared data was compared with its distribution in the original BAUS dataset. No important differences were noticed and we assumed that the data remained unbiased after the preparation process.

The original BAUS dataset was used to create a second subset called BAUS-6. In BAUS-6, the two variables *age* and ASA are added to BAUS-4. ASA is a score which describes the severity of a patient's symptoms on a scale ranging from 1 to 5. No patient was received with an ASA of 4 or 5 and only three categories were kept for this variable. *Age* was discretized in five categories that were chosen as to ensure a balanced distribution between them, as described in Table 2. BAUS-6 contains 1535 records after preparation.

Variable Name	**Categories**
PSA	0-2.5, 2.6-4.0, 4.1-6.0, 6.1-10.0, >10.0
GS	5-6, 3+4, 4+3, ≥8
CS	T1c, T2a, T2b/c
Age	<55, 55-59, 60-64, 65-69, ≥70
ASA	1, 2, 3
PS	OC, EPE, SVI, LNI

A further dataset, ARI-10, was prepared from data collected at the Ab-

erdeen Royal Infirmary, UK. This data contains different variables and allows the exploration of variables that were not collected in the original BAUS dataset. Table 3 presents the variables that were selected for our study. In addition of the Partin variables, ARI-10 includes information on patients' age, pre-operative erection, prostate size following transrectal ultrasound (*TRUS size*) and stage prediction following Magnetic Resonance Imaging (*MRI stage*). Two variables are also included that relate to the patient's wellbeing. They both result from the International Prostate Symptom Score (IPSS) which is composed of seven questions related to the effect of the symptoms on the patients and an additional question which reflects his overall quality of life. We respectively call these variables *IPSS symptoms* and *IPSS QoL*. Being based on patients from a single institution, the size of ARI-10 is much smaller and contains 85 records. Such small size implies that the variance of classifiers built from this data is likely to be important [25] and results should be considered as preliminary.

Variable Name	**Categories**
PSA	0-2.5, 2.6-4.0, 4.1-6.0, 6.1-10.0, >10.0
GS	5-6, 3+4, 4+3, $\geq$8
CS	T1c, T2a, T2b/c
Age	<55, 55-59, 60-64, 65-69, $\geq$70
Pre-op erection	Full function, Partial, Absent, Unknown
IPSS Symptoms	Mild, Moderate, Severe, Unknown
IPSS QoL	0, 1, 2, 3, 4, 5, 6, Unknown
TRUS size	0-30, 31-60, >61, Unknown
MRI stage	T0/T1, T2, T3a, None
PS	OC, EPE, SVI, LNI

One important characteristic of both BAUS and ARI data relies in the distribution of the class variable being very skewed towards milder pathological stages. By consequence, the number of SVI and LNI cases is low in each dataset as illustrated in Figure 1.

In order to fairly compare LR to other methods, the selection of these techniques represents an important step of the study. A first set of runs was performed in order to select the best classifiers from the Weka platform [26]. In this section, we present the methods that showed the best initial performances and that were applied to the different datasets for the complete

study. In order to ensure the comparison covers a wide area of the machine learning landscape, we considered methods from the following machine learning families: decision tree learning, lazy learning, regression, Support Vector Machine, Artificial Neural Networks and Bayesian Networks. Most of these classifiers have already been applied to cancer applications in the past as summarized in [20].

Id3 [27] is a method to build a decision tree based on entropy. Entropy measures the uncertainty associated with a variable with respect to the data. In Id3, a decision variable is added to the tree if it presents the minimum entropy value among all remaining variables in the dataset. Variables with smallest entropy are the closest to the tree root as they have a bigger impact on classification.

In k-nearest neighbours (k-NN) [28], a new instance is classified according to the class value of its k most similar neighbours. A majority vote is used to infer a classification outcome from the k retrieved values. The distance between the test instance and the training instances can be computed in several ways. The most popular method for numerical attributes uses Euclidean distance. This has also been adapted to handle nominal variables, the distance between two instances corresponding to the number of attributes they have in common.

Logistic regression (LR) [5] associates a weight to each of the predictors (for binary variables) or to each of the predictor states (for multinomial variables). Weighted predictor observations are summed and fitted to a logistic curve to produce a probability for the response variable. As previously mentioned, LR is the technique which was used to generate the Partin tables and represents therefore an important element of comparison.

In Support Vector Machine (SVM), instances are represented as vectors and projected onto a n-dimension graph where n is the number of features in the dataset. Building such a classifier requires finding the optimal hyperplane that splits instances in clusters according to their class values. For the present study, the Sequential Minimal Optimization algorithm (SMO) [29] was used to train the SVM.

Artificial Neural Networks (ANN) consist of two or more layers of artificial neurons which receive information signals via their respective inputs. The value of the input information is weighted and processed by a neuron according to the value of its activation threshold. Using many layers of linked neurons, complex decision process can be modeled. Different approaches have been developed to learn the neuron weights or to set the activation

functions. From the wide range of available ANNs, and in order to cover more than one approach, Multilayer Perceptron (MLP) [30] and Radial Basis Function (RBF) [31] were selected for the study. MLP was applied for prostate cancer staging in previous studies and presented better performances than LR [32–34]. However, it has also been proven that MLP does not always outperform LR [35–37]. RBF was compared against LR in [35] but no significant difference was observed between the methods.

Bayesian Network (BN) [38] is a type of probabilistic graphical model (PGM). A BN is composed of a directed acyclic graph (DAG) and of a set of parameters that factorizes the joint probability distribution P of a set of variables X_i according to their respective parents $Pa(X_i)$ as shown in (1).

$$P = P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \tag{1}$$

In addition to BN's ability to handle data of any dimension, this property makes it a particularly suitable solution to deal with uncertainty in datasets where the number of features is large. However, the number of BN structures that can be built from a same dataset of size n grows as $O(n!2^{\frac{n}{2}})$ [39]. Evolutionary computation has been used to cope with the large size of the search space. Using a Genetic Algorithm (GA) and based on the CH metric described by Cooper and Herskovits [40], the K2GA search and score algorithm has proved successful for BN structure learning [41, 42].

Restrictions can also be applied on BNs and results on simplified models. In a Naive Bayes (NB) [43], a class variable is set prior to build the model. The BN which is created afterwards defines the class variable as a parent of all other variables. In a NB, no edges are allowed between the predictor variables. The concept of NB has been extended as to consider that relations may exist between the predictors. Such BN is called a Tree Augmented Naive Bayes (TAN) [44] and is built using a greedy search and the CH metric.

3.3. Experimental design

The main objective of the study is to evaluate the performance of the different classifiers. Area under the ROC curve (AUC) is a standard method to assess a model's predictive power [45]. AUC takes into consideration both sensitivity and specificity and represents an objective way to cope with data which is unbalanced between classes. An AUC value close to 1 describes a model with a good predictive power, while a value close to 0.5 shows that

the model is no better than a random decision. An AUC of zero describes a model that classifies all instances with a wrong label.

On the other hand, many medical studies have assessed Partin tables by computing their concordance index (c-index) [46], as an alternative to AUC. C-index reflects how good a model is at accurately distinguishing between two randomly selected patients with different outcomes. Algorithm 1 outlines how the concordance index CI_{c_i,c_j} between two distinct classes c_i and c_j is computed for a given model M. Let s_p and s_q respectively denote the class label of subject p and q. s_p and s_q belongs to the ordered set C of n class labels such as $c_1 < c_2 < ... < c_n$. Let S_i denotes the set of subjects p with class label c_i such as $S_i = \{p : s_p = c_i\}$ and $m_i = |S_i|$.

$P(s_p = c_i)$ represents the probability of subject p to be classified as c_i while $P(s_p > c_i)$ denotes the probability of subject p to be classified with a class label better than c_i. Similarly, $P(s_p < c_i)$ denotes the probability of subject p to be classified with a class label worse than c_i.

Algorithm 1 Concordance Index CI_{c_i,c_j} between classes c_i and c_j

Initialize $correct = 0$
for each pair $(s_p, s_q) \in S_i \times S_j$, $(i < j)$ **do**
  from model M, compute $P(s_p < s_q) = \sum_{k=1}^{n-1} P(s_p = c_k) * P(s_q > c_k)$
  and $P(s_q < s_p) = \sum_{k=1}^{n-1} P(s_q = c_k) * P(s_p > c_k)$
  if $P(s_p < s_q) > P(s_q < s_p)$ **then**
    $correct$ ++
  end if
end for
return $CI_{c_i,c_j} = correct \ / \ m_i * m_j$

In order to validate the Partin tables and their methodology, the approach that was used to build them [8] was carefully studied and replicated. As a result, multinomial LR was applied to the data using bootstrap resampling with 1000 replications. The variables were discretized in the same way as in Partin tables and as given in Table 1. C-indices for each pathological stage against OC were computed as to ensure the results can be compared with previous external validations and with original findings.

For the rest of the study, 10-fold stratified cross validation was performed 1000 times for each classifier on the three datasets and AUC was calculated. The number of folds and the choice for stratification was decided following

[47] to ensure the measure of accuracy reflects objectively the model's true abilities with respect to its variance and bias when data varies.

To assess Id3, k-NN, LR, MLP, RBF and the SVM classifiers, we used the Weka suite [26], while BN-based models were built and analyzed using implementations for BN learning developed at the Robert Gordon University [48]. To ensure the performance measures were consistent across the two tools, the random seeds for stratification and cross validation was set to the same value. We compared several BNs using both platforms and retrieved similar AUC values.

An initial 10-fold cross validation analysis was run for each of the Weka classifier in order to ensure that comparisons are fair. The parameter settings were hand optimized until no further improvement with respect to the AUC could be found. We regards these settings as optimum and kept them for the study. The final settings are presented in Table 4. K2GA, the algorithm used to search for an optimum BN was tuned with the settings presented in Table 5. K2GA and the greedy search for TAN were run on the complete datasets, providing an optimum structure for each. The parameters were calculated afterwards for each testing fold. We call the BN learned from K2GA, CHBN, as it is based on the CH metric. For BAUS-4, CHBN was found exhaustively because it was possible to test all possible structures, due to the small number of variables in the data.

Classifier	Settings
k-NN	k = 3 (BAUS-4); k = 4 (BAUS-6) Search = linear based on Euclidean distance
ID3	None
LR	maxIts = -1 ridge = 10^{-8}
MLP	Hidden layers = 4 (number of classes) Learning rate = 0.3 Momentum = 0.2 Training time = 100 epochs
RBF	maxIts = -1 minStdDev = 0.1 numClusters = 2 ridge = 10^{-8}
SVM	Filter type = Normalize training data Complexity parameter = 1 Kernel type = Puk ($\Omega = 1$; $\Sigma = 1$)

GA parameter	Value
Number of runs	20
Population size	100
Selection type	Rank selection
Crossover type	Cycle crossover
Crossover rate	0.9
Mutation rate	0.1

4. Results

4.1. External validation of Partin tables

Similarly to the internal validation of the Partin tables, c-index was calculated for the three non-OC pathological stage vs. OC. Results are presented in Table 6 and illustrate how good the different LR models are at distinguishing between patients with each combination of stages. Such values can be understood relative to the scale given in [49]. The scale defines three levels of predictive power for a model according to its c-index. A model has low,

moderate or high prognostic accuracy if its c-index is respectively between 0.5 and 0.7; between 0.7 and 0.9; or greater than 0.9.

	Partin tables with US Data [8]	**Partin tables with BAUS Data**	**Multinomial LR with BAUS Data**
EPE vs. OC	0.696	0.602	0.610
SVI vs. OC	0.830	0.709	0.713
LNI vs. OC	0.894	0.819	0.873

Referring to Table 6, any model built following Partin approach is found to have a low predictive power when distinguishing between OC and EPE cases, regardless of the dataset used for validation. Internal validation numbers given in [8], show that the Partin tables predictive power is moderate for SVI vs. OC and LNI vs. OC cases. When applied to British data, we notice a drop in terms of c-index for every combination of outcome. Building a new model using the same methodology improves slightly the c-index but both Partin tables and new LR model can be described with a low predictive power for EPE vs. OC and a moderate predictive power for SVI vs. OC and LNI vs. OC cases.

4.2. Use of alternative classifiers

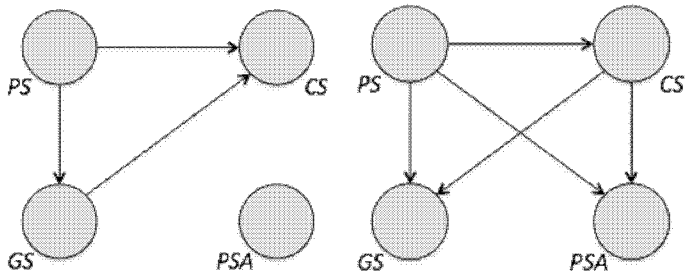
Using the BAUS-4 dataset, we evaluated a range of classifiers. Table 7 describes the AUC of the different classifiers for each pathological stage. Each of these values illustrates how good the model is at correctly classifying a new patient in the given category. Overall, the use of different methods gives rise to the variety of AUCs that are calculated.

Classifier	AUC (OC)	AUC (EPE)	AUC (SVI)	AUC (LNI)
NB	**0.662** (0.002)	**0.604** (0.003)	**0.702** (0.004)	**0.827** (0.012)
TAN	0.654 (0.003)	0.588 (0.005)	0.701 (0.007)	0.794 (0.015)
CHBN	0.630 (0.003)	0.578 (0.006)	0.693 (0.005)	0.809 (0.014)
LR	0.660 (0.002)	0.601 (0.004)	0.694 (0.004)	0.717 (0.036)
MLP	0.645 (0.006)	0.587 (0.008)	0.693 (0.012)	0.792 (0.031)
RBF	0.649 (0.006)	0.591 (0.009)	0.686 (0.012)	0.767 (0.046)
k-NN	0.632 (0.005)	0.569 (0.008)	0.666 (0.012)	0.700 (0.014)
Id3	0.632 (0.005)	0.574 (0.008)	0.661 (0.012)	0.468 (0.018)
SVM	0.525 (0.003)	0.492 (0.004)	0.585 (0.009)	0.491 (0.001)

The model built using NB offers the best AUC in any pathological stage. Bonferroni correction was applied to ensure a fair comparison between methods and the difference between NB and the other classifiers is found to be statistically significant (p-value<0.005).

Among models based on BN, CHBN and TAN do not offer any advantage over the simpler NB. It is interesting to note that, despite its performance, the naive structure is not found by the exhaustive search used in CHBN nor by the greedy search for TAN. Figure 2 illustrates the structures that were learned by the GA and by the TAN search algorithm. These are based on the CH metric and reflect relationships within the data. The structure showing the highest CH score does model relationships between PS, GS and CS but considers PSA as conditionally independent from the other variables. In the TAN, in addition to the naive structure, PSA is considered as conditionally dependent on CS. The latter model outperforms CHBN to a statistically significant extent for the prediction of OC, EPE and SVI cases. CH scores for the different BN structures on each dataset is summarized in Table 8. Learning algorithms considers a structure to be a good model of the data if its associated CH value is close to zero.

Dataset	NB	TAN	CHBN
BAUS-4	-6721.430	-6746.929	**-6720.597**
BAUS-6	-9507.130	-9544.971	**-9464.534**
ARI-10	-1066.562	-1066.360	**-1002.504**



(a) K2GA (b) TAN

Figure 2: BN Structures learned from BAUS-4 dataset

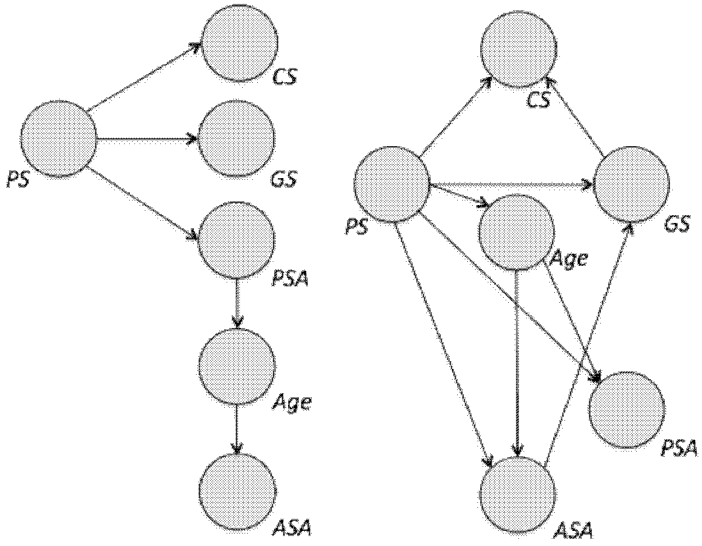
4.3. Addition of new variables

The BAUS-6 dataset introduces two extra variables that were not in BAUS-4. The same classifiers were run and AUC values are described in Table 9. With the probability distribution becoming harder to model, no classifier shows the best AUC on all four values of PS, as was the case on BAUS-4. Despite this difference, all best models are built using different BN techniques.

Classifier	AUC (OC)	AUC (EPE)	AUC (SVI)	AUC (LNI)
NB	**0.679** (0.002)	0.620 (0.004)	0.713 (0.005)	0.740 (0.007)
TAN	0.668 (0.004)	0.600 (0.006)	**0.735** (0.008)	0.627 (0.008)
CHBN	0.675 (0.002)	**0.622** (0.003)	0.724 (0.004)	**0.773** (0.006)
LR	0.675 (0.003)	0.615 (0.005)	0.699 (0.006)	0.731 (0.015)
MLP	0.650 (0.009)	0.597 (0.011)	0.694 (0.017)	0.746 (0.040)
RBF	0.656 (0.009)	0.599 (0.011)	0.692 (0.015)	0.648 (0.079)
k-NN	0.627 (0.007)	0.560 (0.009)	0.665 (0.012)	0.522 (0.049)
Id3	0.580 (0.009)	0.517 (0.009)	0.543 (0.014)	0.483 (0.001)
SVM	0.516 (0.004)	0.496 (0.006)	0.538 (0.010)	0.493 (0.001)

With respect to the assessment of LR for prostate cancer staging, both NB and CHBN significantly outperform LR for the prediction of all pathological stages, except for OC classification where LR yields an AUC not significantly different to that of CHBN.

The BN structures learned using K2GA and TAN search are presented in Figure 3. K2GA retrieves the naive structure that exists between CS, GS, PSA and the class variable PS. In addition, relationships are discovered between PSA and age; and age and ASA. These two conditional dependencies are also found by the TAN.



(a) K2GA (b) TAN

Figure 3: BN Structures learned from BAUS-6 dataset

Along with Tables 9 and 7, Figure 4 illustrates the impact of adding features to the set of variables originally used in Partin tables. Performances of each classifier can be compared across the two datasets. LR-based models are significantly improved by addition of the age and ASA variables in all four categories. BN techniques and RBF present better AUC on BAUS-6 than on BAUS-4 for OC, EPE and SVI predictions, while MLP reaches a statistical significant level of difference for OC and EPE predictions. The k-NN and Id3 methods suffer from the addition of variables. As a consequence, their AUC values decrease for all of the PS categories but one, as LNI prediction is improved for Id3. LNI prediction is, over all classifiers, altered negatively by the inclusion of age and ASA in the study, except if combined with the use of LR or Id3.

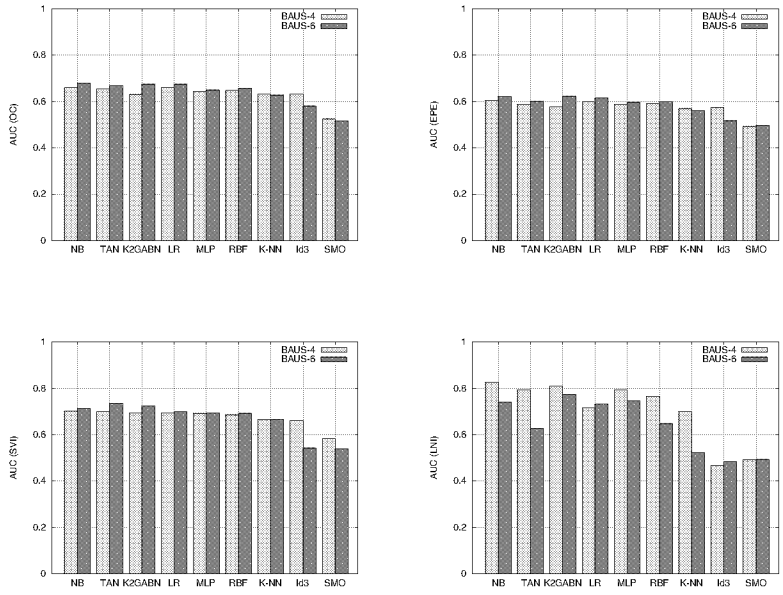


Figure 4: Difference in AUCs between BAUS-4 and BAUS-6

The ARI-10 dataset, prepared from a smaller number of records, but with more variables is used to explore the behavior of the classifiers and possible relationships among variables. AUC results are shown in Table 10 for our selection of classifiers. The range of AUC values is larger than previously observed with other datasets. The standard deviation is also larger for all methods. These two observations is likely to be linked with the small size of the dataset and the larger number of variables.

Classifier	AUC (OC)	AUC (EPE)	AUC (SVI)	AUC (LNI)
NB	0.523 (0.036)	0.410 (0.051)	0.528 (0.030)	0.008 (0.007)
TAN	0.592 (0.032)	**0.639 (0.046)**	0.488 (0.029)	0.011 (0.010)
CHBN	**0.668 (0.026)**	0.591 (0.045)	0.567 (0.029)	0.019 (0.008)
LR	0.534 (0.048)	0.342 (0.059)	0.582 (0.081)	0.380 (0.237)
MLP	0.500 (0.055)	0.379 (0.067)	**0.668 (0.075)**	**0.670 (0.232)**
RBF	0.490 (0.071)	0.487 (0.067)	0.475 (0.116)	0.546 (0.294)
k-NN	0.604 (0.041)	0.592 (0.047)	0.400 (0.067)	0.329 (0.172)
Id3	0.462 (0.048)	0.457 (0.034)	0.478 (0.026)	0.500 (0.000)
SVM	0.500 (0.000)	0.500 (0.000)	0.500 (0.000)	0.500 (0.000)

Similarly to their performances on the BAUS datasets, highest AUCs in three of the four pathological stages are obtained from BN techniques. The best model for prediction of OC and EPE are respectively built using CHBN and TAN. However, their AUC is much lower for EPE and close to zero for LNI prediction where MLP performs best.

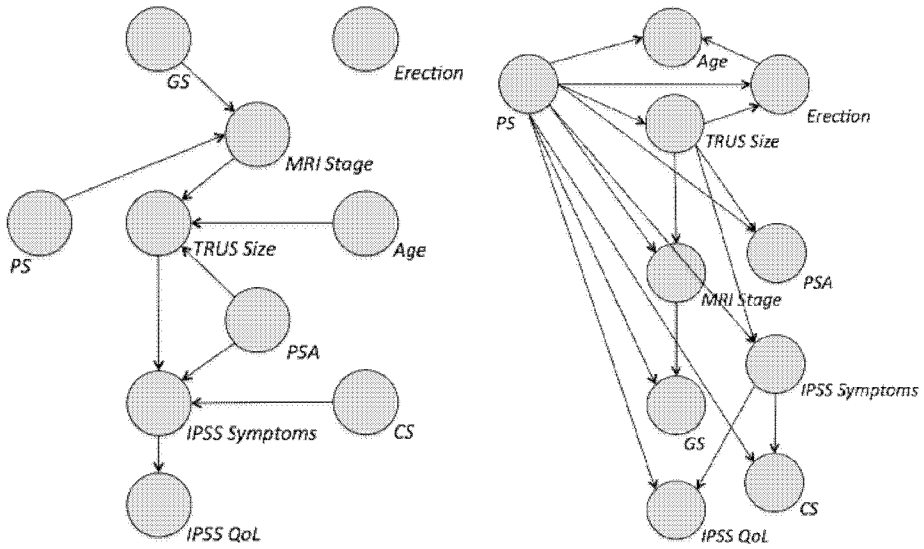
The CHBN and TAN structures presented in Figure 5 both show similar patterns. For example, edges are modeled between the variable *TRUS size* and *PSA*, *MRI stage* and *IPSS symptoms*. *IPSS symptoms* also appears as an important node, with relationships with *CS*, *TRUS size* and *IPSS QoL*. We note that K2GA found that *MRI stage* is the only variable conditionally dependent on *PS*. Finally, the *pre-operative erection* variable is isolated from other variables.

Over all datasets, and among a pool of various classifiers, techniques using BN offers the best AUC for *PS* prediction in 10 comparisons out of 12. Only MLP outperforms other methods in two domains. This is observed when MLP is applied on ARI-10 dataset and AUCs for SVI and LNI are measured. k-NN, Id3 and SVM are generally clearly behind the other techniques in terms of AUC.

5. Discussion

5.1. External validation of Partin tables

As shown in Table 6, original Partin tables achieve lower c-indices when applied to the BAUS data. This implies that when used on a UK population, Partin tables have a lower predictive power than on the native data from which they were derived. With a c-index below 0.70 for OC vs. EPE, Partin



(a) K2GA (b) TAN

Figure 5: BN Structures learned from ARI-10 dataset

tables can be considered as having poor predictive power for patients falling in these two categories. We recall that patients with OC and EPE pathological stages are the most frequent cases in the BAUS dataset, as they count for nearly 95% of the entire cohort. In addition, the decision making for radical prostatectomy is strongly guided by the probabilities of a patient to have OC or EPE conditions. Correct distinction between these two classes represents thus the most important feature of the Partin tables. Furthermore, the applicability and usefulness of predictive tools with a c-index lower than 0.70 has been questioned, although there is a lack of an accepted reference threshold beyond which the use of a predictive model becomes unacceptable [50].

The AUC analysis also supports this assertion and is in keeping with the trend seen in other validation studies [9–19]. In these studies, the AUC for OC varies between 0.604 [17] and 0.817 [12].

There are several possible explanations for the reduced performance of Partin tables when applied to a non-US population. Firstly, the original c-indices for Partin tables were derived from internal validation, indicating that

the same data was used to both generate and assess the tables. The resulting model should thus be very well suited for the population from which it was derived but not for different populations. Secondly, the updated Partin tables [8] were generated using a cohort of patients from a single institution. The tables reflect therefore the local population's demographics, genetic and ethnic mix, disease trends, environmental and social factors and health care system, and therefore may not perform as well on populations with different characteristics. The BAUS cohort presents some fundamental differences likely to affect the models predictive power. These differences may be due to significant discrepancies in health care system policies between the US and the UK. For instance, PSA screening is routinely practised in the US, while it is not the case in the UK, and clearly UK patients undergoing prostatectomy are presenting with a higher PSA, GS or clinical stage. Lymph node dissection is also a domain that differs between UK and US policies. While in the US, it is systematically performed on patients undergoing prostatectomy, in the UK, it is reserved for patients presenting high PSA and high GS. The lower rate of dissection in the UK may account for the differences in c-indices and AUC for LNI in the different models, as this pathological stage is only observed following dissection. Finally, the poor performance may also be due to the categorization of the pre-operative variables. As has already been shown, the disease characteristics and trends differ between the two populations and as such the original Partin sub-grouping of variables may not be entirely appropriate for the BAUS cohort.

Despite the overall low predictive power that can be associated with all models built following Partin approach, UK derived lookup tables show a better c-index. This observation supports our assertion that when applied locally, lookup tables generated from a UK population may have higher predictive power than those generated from a population with inherently different characteristics.

5.2. Importance of the methodology

To build and assess a predictive model, three main elements need to be taken into consideration. The choice of the classifier is important but it can only lead to good prediction if the input variables have been chosen carefully. The quality of the data is also a key factor in the process of building a model.

Nine methods were applied to the different datasets. The AUCs of the resulting models vary significantly. For example the AUCs for OC prediction on BAUS-4 range between 0.525, using SVM, and 0.662, using NB.

This results from the inherent characteristics and approaches of the different methods, along with the type of data being used. BN techniques have an overall higher predictive power than other methods. However, when measured on BAUS-4 and despite being statistically significant, the difference in mean between NB and LR, the two best performing techniques, is small (0.002). When applied on a dataset containing a larger number of variables, such as BAUS-6, the difference between the two same techniques becomes more marked (0.004). Similarly on ARI-10, LR is clearly outperformed in mean by the best BN model, CHBN (0.134). The dimension of the dataset impacts on the classifier's abilities to produce high quality models. When only four variables are involved, it is expected to see small differences in performances between the methods as the joint probability distribution is easier to model for low-dimensional data. When new variables are included in the study, some classifiers can lead to complex models. For example, the Euclidean distance, on which the k-NN algorithm is based, loses discriminating power when applied on high dimensional data. This characteristic of k-NN is illustrated when the variables *age* and ASA are included in the BAUS data by a drop in terms of AUC for all class labels. The tree algorithm Id3 also has difficulties in correctly classifying PS on BAUS-6. Explanation can be found by studying the final tree learned from data. With 6 variables, the tree is composed of 382 leaves, against 54 when BAUS-4 is used for training. This represents too many covariate patterns for Id3 to model using data of such size. On the other hand SVM techniques are known for their good performances on high dimensions. We suggest that the poor predictive power presented by SVM on the BAUS datasets can be explained by the number of variables being too few for this method.

As already mentioned, LR appears as a competitive solution for prostate cancer staging. However, it does not outperform any of the selected alternative methods on any domain. A potential drawback of LR resides in the hypothesis which states that all predictors are independent with each other. This property is also considered in a NB. Although this statement can be justified on BAUS-4 according to the performances of both LR and NB, it does not seem justified on BAUS-6 and ARI-10. On these two datasets, AUCs from TAN and CHBN are overall higher than those from NB and LR. Besides, the associated BN structures reveal relationships between some of the predictor variables. Independence between predictors can be assumed on datasets with a few variables. However, with addition of extra features, interactions are likely to appear that are beneficial for the model's quality.

The current version of Partin tables, based on LR may not suffer from the previously discussed assumption, but experiments have shown that LR is not the best method for PS classification. In addition, the performance gap with other approaches such as BN appears to increase with the dimension of the data to model.

When the first version of Partin tables was created [4], Chi-square tests were performed along with LR. The aim was to discover the combination of variables that best correlates with PS. Results led to the conclusion that PSA, GS and CS should be used together, and age of patients removed from the study. In the present study, experiments were run on two subsets of the BAUS dataset. The AUCs of the models built from BAUS-6, including the two new variables ASA and age, were higher than the ones obtained from BAUS-4 with the same methods. Among these methods, LR produced a better model when including age in the study, contradicting Partin's original assumption.

The quality of the data represents a key factor in the construction of a predictive model. BAUS dataset is the result of a large scale data collection, involving 57 different centers. Such data is extremely hard to gather as it involves collaboration between institutions, standardization of the data and ethical issues. As a consequence, some records had to be removed due to inconsistencies. Although the size of the datasets finally used was highly reduced, it remains large with respectively over 1700 and 1500 patient records for BAUS-4 and BAUS-6. The difficulties encountered in the data preparation process is to be taken into consideration as it reflects a current challenge of medical data mining [51]. Local data represents a good opportunity to explore ideas but can also suffer from too few records. Such difficulties were observed on ARI-10 and caution should be taken when analyzing resulting numbers. Another challenge resides in the skewed distribution of pathological stages in both BAUS and ARI datasets. Representing around 1% of all records, LNI condition is the most challenging stage to predict, and is illustrated by standard deviations higher than for other stages for most methods. These results should also be treated carefully as they are likely to vary with the data.

5.3. Performances and characteristics of Bayesian Networks

AUC measures show that the classifiers based on BN are better adapted to prostate cancer staging than other methods from our selection. On BAUS-4, NB outperforms TAN and CHBN. In other terms, setting the BN structure

to one of its simplest form was beneficial over the use of heuristic search for optimum networks. One could argue on the efficiency of the heuristic employed in CHBN and TAN, but the metric on which the search is based describes how well a structure reflects the data. The CH score assesses a BN in a general way, thus without focusing on a particular variable. Scores of the different BNs presented in Table 8 show that CH values are always higher for CHBN than for TAN and NB as the search strategy aims at maximizing it. However, CH scores and AUCs for PS are not affected in the same manner. A low CH score does not ensure that the corresponding AUC for PS will be high. The development of NB and TAN was originally motivated by this limitation on BN. These two restricted BNs are biased toward a specific purpose, such as classification of a predefined variable. Their performance on BAUS-4 comfort their efficiency on small dimensional datasets over unrestricted networks.

However, as seen with NB, this is an efficient solution when the number of variables to model is low, but shows limitation in other contexts. On BAUS-6, the performances between NB and CHBN are close for OC prediction and NB is outperformed by CHBN for EPE and SVI classification. In a similar manner, experimental results on ARI-10 shows that NB has the worst performances of the three BN methods for AUCs on OC and EPE. In addition, CHBN presents with a AUC for OC of 0.668 against 0.604 for the second best performing technique, k-NN. This important difference reflects the ability of the GA to find a good solution from a large space of possible structures. Heuristic search appears like the right approach when more features are included. Search and score algorithms benefit from using datasets with a large number of variables. The larger search space associated offers more possible dependencies between variables that are likely to further improve the model's predictive power.

The structures found by K2GA represent the most relevant relationships associated with the data. On BAUS-4, the PSA variable is isolated from the rest of the features and it results in a model with a lower predictive power than the ones obtained from NB and TAN when setting PSA as dependent on PS. This reflects the importance of PSA and is consistent with medical understanding [4]. From the experimental results on BAUS-6 using K2GA and TAN, PSA is also considered as conditionally dependent of PS. PSA and ASA appear to be correlated with *age*, a medically meaningful finding. In addition, the NB structure on BAUS-4 is retrieved when K2GA is run on BAUS-6. CS, GS and PSA are indeed linked with PS, illustrating that

these variables are the most significant for PS. On ARI-10, only *MRI stage* and GS are contained in the Markov blanket of the class variable PS. In this latter model, only *MRI stage* and GS are needed to infer a patient's pathological stage. The presence of *MRI stage* in PS' Markov blanket while CS is not is consistent with the fact that MRI is a more accurate means of evaluating PS than DRE [52]. Other relationships were found to be relevant with medical expectations, such as the dependency between *TRUS size* and *age* and PSA [53], but others such as the relationship between *MRI stage* and *TRUS size* raised questions regarding their medical meaning [54]. This may be due to the small size of the ARI dataset implying that these results should only be considered as exploratory.

6. Conclusion and further work

In this paper, we have assessed one of the major tools used clinically for prostate cancer staging. The predictive power of Partin tables is much lower when it is applied on a British population than it was when originally measured on US data. In addition, a replicate Partin study shows that new lookup tables also exhibit a low to moderate predictive power. A range of alternative classifiers were selected and three datasets prepared in order to assess different aspects of the methodology, including the inclusion of new variables. Using the same variables as Partin, NB outperforms other techniques for the prediction of any pathological stage. With the addition of new variables, the difference in terms of predictive power between BN methods and the others becomes more marked. The choice for LR as the most adopted classifier to build Partin tables is not justified by our experiments. In addition, the inclusion of extra variables improves the quality of the prediction for most of the techniques. Overall, BNs exhibit the best predictive power. Their ability to deal efficiently with high dimensional data combined with the use of heuristic search make them ideal classifiers for prostate cancer staging. In addition, they provide comprehensible models where relationships between variables can be revealed that enhance disease understanding. Efforts should be made in improving data collection and a maximum number of features should be included in the study to benefit fully from BN abilities.

This paper has shown that BNs are suitable for the modelling of prostate cancer staging. In future work, BN should be applied to other medical problems as seen for ovarian [55] and breast [56] cancers in order to study their generalizability. Data properties also play an important part in the perfor-

mance of a classifier and the behavior of BN will be observed in greater details when the size and the dimensionality of the data vary. The high proportion of missing and inconsistent records in the data also indicates that techniques to handle these problems need to be incorporated to our methods. Despite the relatively good performances of CHBN and TAN, the metric upon which the heuristic is based, did not reflect very accurately the predictive power of the BN. Alternative metrics have been developed and it will be interesting to assess how each of them affects the quality of the models with respect to both structure discovered and classification. Finally, we believe BNs can be used efficiently to produce clinical tools for prostate cancer staging and help with current medical challenges. Efforts will be made to develop such instruments in the future.

Acknowledgements

This work was jointly funded by the Northern Research Partnership and NHS Grampian.

References

[1] International Agency for Research on Cancer - GLOBOCAN 2008. http://globocan.iarc.fr/; (Accessed: 21 September 2010).

[2] Hayward J, Alvarez S, Ruiz C, Sullivan M, Tseng J, Whalen G. Machine learning of clinical performance in a pancreatic cancer database. Artificial Intelligence In Medicine 2010;49(3):187–95.

[3] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. Artificial intelligence in medicine 2005;34(2):113–27.

[4] Partin A, Yoo J, Carter H, Pearson J, Chan D, Epstein J, et al. 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 1993;150(1):110–4.

[5] Hosmer D, Lemeshow S. Applied logistic regression. New York, NY, USA: Wiley-Interscience; 2000.

[6] Partin A, Kattan M, Subong E, Walsh P, Wojno K, Oesterling J. Combination of prostate-specific antigen, clinical stage, and Gleason score to predict pathological stage of localized prostate cancer: A multi-institutional update. Journal of the American Medical Association 1997;277(18):1445–51.

[7] Partin A, Mangold L, Lamm D, Walsh P, Epstein J, Pearson J. Contemporary update of prostate cancer staging nomograms (Partin Tables) for the new millennium. Urology 2001;58(6):843–8.

[8] Makarov D, Trock B, Humphreys E, Mangold L, Walsh P, Epstein J. 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):1095–101.

[9] Kattan M, Stapleton A, Wheeler T, Scardino P. Evaluation of a nomogram used to predict the pathologic stage of clinically localized prostate carcinoma. Cancer 1997;79(3):528–37.

[10] Blute M, Bergstralh E, Partin A, Walsh P, Kattan M, Scardino P, et al. Validation of Partin tables for predicting pathological stage of clinically localized prostate cancer. The Journal of Urology 2000;164(5):1591–5.

[11] Penson D, Grossfeld G, Li Y, Henning J, Lubeck D, Carroll P. How well does the Partin nomogram predict pathological stage after radical prostatectomy in a community based population? Results of the cancer of the prostate strategic urological research endeavor. The Journal of Urology 2002;167(4):1653–8.

[12] Graefen M, Augustin H, Karakiewicz P, Hammerer P, Haese A, Palisaar J, et al. Can predictive models for prostate cancer patients derived in the United States of America be utilized in European patients? A validation study of the Partin tables. European Urology 2003;43(1):6–11.

[13] Augustin H, Eggert T, Wenske S, Karakiewicz P, Palisaar J, Daghofer F, et al. Comparison of accuracy between the Partin tables of 1997 and 2001 to predict final pathological stage in clinically localized prostate cancer. The Journal of Urology 2004;171(1):177–81.

[14] Eskicorapci S, Karabulut E, Turkeri L, Baltaci S, Cal C, Toktas G, et al. Validation of 2001 Partin tables in Turkey: a multicenter study. European Urology 2005;47(2):185–9.

[15] Karakiewicz P, Lattouf J, Perrotte P, Valiquette L, Benard F, McCormack M, et al. Validation of 1997 Partin Tables lymph node invasion predictions in men treated with radical prostatectomy in Montreal Quebec. The Canadian Journal of Urology 2005;12(2):2588–92.

[16] Song C, Kang T, Ro J, Lee M, Kim C, Ahn H. Nomograms for the prediction of pathologic stage of clinically localized prostate cancer in Korean men. Journal of Korean Medical Science 2005;20(2):262–6.

[17] Gao X, Ren S, Lu X, Xu C, Sun Y. The Newer the Better? Comparison of the 1997 and 2001 Partin Tables for Pathologic Stage Prediction of Prostate Cancer in China. Urology 2008;72(5):1096–101.

[18] Bhojani N, Ahyai S, Graefen M, Capitanio U, Suardi N, Shariat S, et al. Partin Tables cannot accurately predict the pathological stage at radical prostatectomy. European Journal of Surgical Oncology 2009;35(2):123–8.

[19] Bhojani N, Salomon L, Capitanio U, Suardi N, Shariat S, Jeldres C, et al. External validation of the updated partin tables in a cohort of French and Italian men. International Journal of Radiation Oncology, Biology, Physics 2009;73(2):347–52.

[20] Cruz J, Wishart D. Applications of machine learning in cancer prediction and prognosis. Cancer Informatics 2006;2:59–77.

[21] Jensen F, Nielsen T. Bayesian networks and decision graphs. New York, NY, USA: Springer Verlag; 2007.

[22] Sobin L. TNM classification of malignant tumours. New York, NY, USA: Wiley-Blackwell; 2009.

[23] Kirby R. The prostate: small gland, big problem. Northwood, UK: Prostate Research Campaign UK; 2002.

[24] Shaida N, Malone P. Controversial topics in surgery. open versus laparoscopic radical prostatectomy: The case for open radical prostatectomy. Annals of The Royal College of Surgeons of England 2007;89:108–10.

[25] Brain D, Webb G, Richards D, Beydoun G, Hoffmann A, Compton P. On the effect of data set size on bias and variance in classification learning. In: Richards D, Beydoun G, Compton P, Hoffman A, editors. Proceedings of the Fourth Australian Knowledge Acquisition Workshop. Sydney, Australia: University of New South Wales; 1999, p. 117–28.

[26] Hall J, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The weka data mining software: An update. SIGKDD Explorations 2009;11(1):11–8.

[27] Quinlan J. Induction of decision trees. Machine learning 1986;1(1):81–106.

[28] Aha D, Kibler D, Albert M. Instance-based learning algorithms. Machine learning 1991;6(1):37–66.

[29] Platt J. Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods. Cambridge, MA, USA: MIT press; 1999, p. 185–208.

[30] Rumelhart D, Hinton G, Williams R. Learning internal representations by error propagation. Tech. Rep.; California University; 1985.

[31] Broomhead D, Lowe D, Signals R, Malvern R. Radial basis functions, multi-variable functional interpolation and adaptive networks. Tech. Rep.; Royal Signals and Radar Establishment Malvern; 1988.

[32] Han M, Snow P, Brandt J, Partin A. Evaluation of artificial neural networks for the prediction of pathological stage in prostate carcinoma. American Cancer Society 2001;91(8):1661–6.

[33] Matsui Y, Egawa S, Tsukayama C, Terai A, Kuwao S, Baba S, et al. Artificial neural network analysis for predicting pathological stage of clinically localized prostate cancer in the japanese population. Japanese Journal of Clinical Oncolology 2002;32(12):530–5.

[34] Veltri R, Chaudhari M, Miller M, Poole E, O´ Dowd G, Partin A. Comparison of logistic regression and neural net modeling for prediction of prostate cancer pathologic stage. Clinical Chemistry 2002;48(10):1828–34.

[35] Borque A, Sanz G, Allepuz C, Plaza L, Gil P, Rioja L. The use of neural networks and logistic regression analysis for predicting pathological stage in men undergoing radical prostatectomy: a population based study. The Journal of Urology 2001;166(5):1672–8.

[36] Anagnostou T, Remzi M, Lykourinas M, Djavan B. Artificial neural networks for decision-making in urologic oncology. European Urology 2003;43(6):596–603.

[37] Kawakami S, Numao N, Okubo Y, Koga F, Yamamoto S, Saito K, et al. Development, validation, and head-to-head comparison of logistic regression-based nomograms and artificial neural network models predicting prostate cancer on initial extended biopsy. European Urology 2008;54(3):601–11.

[38] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco, CA, USA: Morgan Kaufmann; 1988.

[39] Robinson R. Counting unlabeled acyclic digraphs. In: Little C, editor. Proceedings of the Fifth Australian Conference on Combinatorial mathematics. New York, NY, USA: Springer-Verlag; 1977, p. 28–43.

[40] Cooper G, Herskovits E. A bayesian method for the induction of probabilistic networks from data. Machine Learning 1992;9(4):309–47.

[41] Wu Y, McCall J, Corne D. Two novel ant colony optimization approaches for bayesian network structure learning. In: Sobrevilla P, Aranda J, Xambo S, editors. Proceedings of the 2010 World Congress on Computational Intelligence. Piscataway, NJ, USA: IEEE Press; 2010, p. 4473–9.

[42] Fournier F, McCall J, Petrovski A, Barclay P. Evolved bayesian network models of rig operations in the gulf of Mexico. In: Sobrevilla P, Aranda J, Xambo S, editors. Proceedings of the 2010 World Congress on Computational Intelligence. Piscataway, NJ, USA: IEEE Press; 2010, p. 1–7.

[43] Langley P, Iba W, Thompson K. An analysis of Bayesian classifiers. In: Rosenbloom P, Szolovits P, editors. Proceedings of the Tenth National Conference on Artificial Intelligence. Menlo Park, CA, USA: AAAI Press; 1992, p. 223–8.

[44] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning 1997;29(2):131–63.

[45] Metz C. Basic principles of ROC analysis. Seminars in nuclear medicine 1978;8:283–98.

[46] Harrell Jr F, Califf R, Pryor D, Lee K, Rosati R. Evaluating the yield of medical tests. Journal of the American Medical Association 1982;247(18):2543–6.

[47] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Walsh T, editor. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence; vol. 2. Menlo Park, CA, USA: AAAI Press; 1995, p. 1137–45.

[48] Kabli R, Herrmann F, McCall J. A chain-model genetic algorithm for bayesian network structure learning. In: Thierens D, Beyer H, Bongard J, Branke J, Clark J, Cliff D, et al., editors. Proceedings of the 9th Conference on Genetic and Evolutionary Computation. New York, NY, USA: ACM Press; 2007, p. 1271–8.

[49] Galfano A, Novara G, Iafrate M, Cavalleri S, Martignoni G, Gardiman M, et al. Mathematical models for prognostic prediction in patients with renal cell carcinoma. Urologia Internationalis 2008;80(2):113–23.

[50] Kattan M. Validating a prognostic model. American Cancer Society 2006;107(11):2523–4.

[51] Ramakrishnan N, Hanauer D, Keller B. Mining electronic health records. Computer 2010;43(10):77–81.

[52] Sanchez−Chapado M, Angulo J, Ibarburen C, Aguado F, Ruiz A, Viano J, et al. Comparison of digital rectal examination, transrectal ultrasonography, and multicoil magnetic resonance imaging for preoperative evaluation of prostate cancer. European urology 1997;32(2):140–9.

[53] Collins G, Lee R, McKelvie G, Rogers A, Hehir M. Relationship between prostate specific antigen, prostate volume and age in the benign prostate. British journal of urology 1993;71(4):445–50.

[54] Anastasiadis A, Lichy M, Nagele U, Kuczyk M, Merseburger A, Hennenlotter J, et al. MRI-guided biopsy of the prostate increases diagnostic performance in men with elevated or increasing PSA levels after previous negative TRUS biopsies. European urology 2006;50(4):738–49.

[55] Antal P, Fannes G, Timmerman D, Moreau Y, De Moor B. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. Artificial Intelligence in Medicine 2004;30(3):257–81.

[56] Gevaert O, Smet F, Timmerman D, Moreau Y, Moor B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 2006;22(14):184–90.