Boyce et al. BMC Medical Informatics and Decision Making 2013, 13:126
http://www.biomedcentral.com/1472-6947/13/126

BMC Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

Evaluation of prediction models for the staging of prostate cancer

Susie Boyce[1,4*], Yue Fan[3,4], Ronald William Watson[1,4] and Thomas Brendan Murphy[2]

Abstract

Background: There are dilemmas associated with the diagnosis and prognosis of prostate cancer which has lead to over diagnosis and over treatment. Prediction tools have been developed to assist the treatment of the disease.

Methods: A retrospective review was performed of the Irish Prostate Cancer Research Consortium database and 603 patients were used in the study. Statistical models based on routinely used clinical variables were built using logistic regression, random forests and k nearest neighbours to predict prostate cancer stage. The predictive ability of the models was examined using discrimination metrics, calibration curves and clinical relevance, explored using decision curve analysis. The N = 603 patients were then applied to the 2007 Partin table to compare the predictions from the current gold standard in staging prediction to the models developed in this study.

Results: 30% of the study cohort had non organ-confined disease. The model built using logistic regression illustrated the highest discrimination metrics (AUC = 0.622, Sens = 0.647, Spec = 0.601), best calibration and the most clinical relevance based on decision curve analysis. This model also achieved higher discrimination than the 2007 Partin table (ECE AUC = 0.572 & 0.509 for T1c and T2a respectively). However, even the best statistical model does not accurately predict prostate cancer stage.

Conclusions: This study has illustrated the inability of the current clinical variables and the 2007 Partin table to accurately predict prostate cancer stage. New biomarker features are urgently required to address the problem clinician's face in identifying the most appropriate treatment for their patients. This paper also demonstrated a concise methodological approach to evaluate novel features or prediction models.

Keywords: Prediction models, Model evaluation, Discrimination, Calibration, Prostate cancer

Background

Prostate cancer (PCa) is the most common cancer in European and North American men, and the second most common cause of male cancer deaths [1]. There are dilemmas associated with the diagnosis and prognosis of PCa which has lead to the over diagnosis and over treatment of the disease [2]. However, new treatments such as active surveillance are being introduced to overcome these issues [3-6].

Prediction tools for PCa have been developed to assist in the accurate diagnosis and treatment of the disease, and address a wide variety outcomes; e.g. the Partin tables [7-10], Partin nomogram [11], Kattan and Stephenson nomograms [12-14], D'Amico risk classification [15], CAPRA score [16] and many others [17-19]. For the prediction of stage at radical prostatectomy (RP), the Partin tables not only represent the most common prediction tool used by clinicians, but have also undergone extensive validation in a number of cohorts [20-26]. The Partin table uses clinical stage based on digital rectal exam (DRE), Gleason score (GS) of the prostate needle biopsy [27-31], and serum prostate specific antigen (PSA) to predict stage at RP. PCa stage indicates the extent or location of the cancer, and can be categorized as; organ confined (OC), extracapsular extension (ECE), seminal vesicle invasion (SVI) and/or lymph node involvement (LNI). Non-organ confined (NOC) disease represents any stage which extends beyond the prostate organ, i.e. ECE, SVI or LNI.

While the Partin table is well used by clinicians, excluding this, few other prediction tools are used in a clinical setting.

* Correspondence: susan.boyce@ucdconnect.ie
[1]UCD School of Medicine and Medical Science, University College Dublin, Dublin, Ireland
[4]Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland
Full list of author information is available at the end of the article

To overcome this issue, external validation of prediction models are ongoing. External validations which validate and compare two or more models are particularly useful. Chun et al. used this approach and compared five logistic regression (LR) based nomograms with other LR based models, namely look up table, classification and regression tree, artificial neural networks and risk group stratification [32]. However, each set of models being compared was developed in different patient cohorts and different outcomes were compared, i.e. nomogram for BCR and classification and regression tree for BCR, nomogram for stage and look up table for stage.

The Partin table was developed using multivariate logistic regression (MLR), however it isn't known whether other statistical modelling techniques would have been more accurate to use with this type data. By extending the work of Chun et al. and Partin et al., the aim of this study is to explore a number of classification techniques rather than just LR, each predicting the same outcome and developed and tested in one cohort of patients, using the same variables as those used in the Partin tables. We also aim to explore methods to evaluate prediction models, such as discrimination and calibration metrics, as well as decision curve analysis.

Methods

Study population

A retrospective review was performed of the Irish Prostate Cancer Research Consortium (PCRC) database. The PCRC was founded in 2003, and is a multi-disciplinary trans-institutional collaboration. Patient samples were sourced from four institutions; three tertiary referral centres and one private hospital. Eight consultant urologists and four distinct pathology departments are involved in the acquisition and grading of prostatic tissue. Ethical approval was awarded in each hospital (Mater Misericordiae University Hospital, St James's Hospital, Beaumont Hospital, Mater Private Hospital). Written informed consent was obtained from study participants. Inclusion criteria for this study were availability of pre-operative serum PSA, trans-rectal ultrasound guided needle biopsy Gleason Score [27-31], clinical T stage using TNM staging [33] identified by DRE and the corresponding RP pathology reports. All study participants had pathologically confirmed prostatic adenocarcinoma. Between February 2002 and October 2011, data relating to 705 patients who underwent RP was collected through the PCRC. A total of 102 patients were excluded due to benign prostatic hyperplasia (BPH) and missing data. This left a total of 603 patients.

Clinical and pathological assessment

The clinical stage was stratified as T1c (DRE negative) or T2 (DRE positive) [33]. Recording of the sub-stratification of T2 was not available for the analysis. The Gleason scoring system was used for needle biopsy grading [27-31]. RP specimens were assigned as organ confined (OC) if the tumour can be felt on examination, but has not spread outside the prostate, extra capsular extension (ECE) if the tumour has spread through the prostatic capsule on one or both sides, seminal vesicle invasion (SVI) if the tumour has invaded one or both seminal vesicles and lymph node involvement (LNI) if the pelvic lymph nodes exhibited prostate cancer [33]. Patients were then re-stratified as organ confined (OC) or non-organ confined (NOC), where NOC represents any pathological stage which is not OC.

Statistical analysis

Patient information included pre-operative PSA, clinical stage based on DRE, biopsy Gleason score (GS), age and family history. Descriptive statistics focused on frequencies and proportions for categorical variables. Means, medians, and ranges were reported for continuous data and error measures were reported as 95% confidence intervals (CI). The parametric independent samples *t*-test and non-parametric Mann Whitney U test were used to examine the statistical significance of differences in means for normal and non-normal data respectively. Chi-square test was used to compare frequencies for categorical data.

Seven statistical and algorithmic classification techniques were used to develop models to predict stage at RP. These included logistic regression, linear discriminant analysis, classification and regression trees, k nearest neighbours, artificial neural networks, support vector machines and random forests. The objective of a classification model is to classify patients in two or more groups based on a predicted outcome associated with each patient. On examination of the individual model fit for each of the seven classification techniques, three models were chosen for further analysis and model evaluation: logistic regression [34], random forests [35] and k nearest neighbours [36].

The data was prepared for modelling using 5-fold cross validation (Figure 1). 5-fold cross validation involves randomly dividing the data into five evenly sized subgroups. Each group is called a fold. A model is then constructed using the data from the first four folds and applied to the fifth group. The model building and validation process is repeated five times with each fold of patients used once as the validation set. This results in no patient being used to both develop and test the model [37].

Model evaluation was carried out by examining calibration, discrimination and decision curve analysis [37-41]. The calibration of the models was measured using calibration curves [39]. Calibration measures how close the predicted probabilities are to actual probabilities. A calibration curve plots predictions on the x-axis and the true outcome on the y-axis. Due to the fact that the actual outcomes are 0 and 1, Loess smoothing [42,43] was used to estimate the observed probabilities of the outcome in

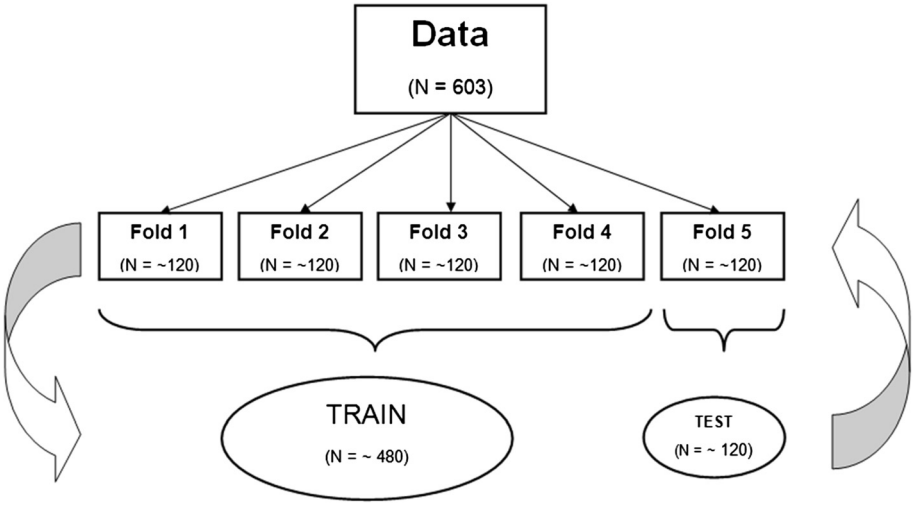

Figure 1 Illustration of 5-fold cross validation. The dataset is randomly split into five subgroups (called folds). Four folds are merged together to form a training dataset. The 5th fold is used as the testing dataset. This process is then repeated five times.

relation to the predicted probabilities. The discriminate ability of the models were compared by formulation of sensitivity [44], specificity [44], positive predictive value (PPV) [45], negative predictive value (NPV) [45], Youden index [46,47], Brier score [48] and area under the curve (AUC) values [49,50]. The discriminate ability of a model measures how well the model discriminates between patients with and without the outcome. The AUC value provides us with a probability that the model will correctly identify which of two individuals with different outcomes actually has the disease.

However, there has been much criticism of the AUC value in the last number of years [51,52]. This is due to the fact that patient's do not present to a clinician's office in pairs, one of whom has NOC disease and the other with OC disease. There is also concern regarding what the necessary AUC value should be for a model to be considered 'clinically useful'. To overcome these issues, decision curve analysis was used to measure the clinical relevance of the three models [37,40,41]. Decision curve analysis is a method for evaluating and comparing prediction models that incorporates clinical consequences. It is based on the principle that the probability at which a physician would advise treatment is informative on how the physician and patient weigh the harms of false-positive results in comparison with the harms of false-negative results. This probability is referred to as the threshold probability (P_t). This threshold probability (P_t) can then be used to derive the net benefit of the model across different threshold probabilities, where:

$$\text{Net Benefit} = \frac{\text{True Positive Count}}{n} - \frac{\text{False Positive Count}}{n} \cdot \left(\frac{P_t}{1-P_t}\right) \quad (1)$$

Plotting net benefit against threshold probability results the 'decision curve'. The decision curve gives the expected net benefit per patient relative to assuming all patients have OC disease, the expected benefit associated with assuming all patients have NOC and the expected benefit associated with using the classification model. The interpretation of net benefit is the model with the highest net benefit should be chosen.

The patient's clinical data was also applied to the 2007 Partin table for ECE [7] in order to evaluate how well this prediction tool can predict stage at RP compared to the three classification models developed in this study. The predictions from the Partin tables were measured for discrimination.

Statistical analysis was performed using R software, version 2.14.0 with the following packages: 'car', 'boot', 'rpart', 'randomForest', 'class', 'e1071', 'MASS', 'nnet', 'ROCR', 'pROC', 'Hmisc', 'rms', 'gmodels', 'gplots', 'epicalc'.

Results

The clinical and pathological characteristics of the N = 603 PCRC patient cohort are given in Table 1. Average patient age was 61 (C.I: 60.4, 61.6) years (median 62, range 42–74). Average PSA value was 7.96 (C.I: 7.57, 8.26) ng/ml (median 7.0, range 0.7-40). Most patients had clinical stage T1c (44.3%), biopsy Gleason score 6 (37.1%) and prostatectomy Gleason score 3 + 4 (39.3%). 54.9% had no family history of cancer, 20.1% had a history of cancer (excluding PCa) in the family and 25.0% had a family history of PCa. Most patients, 70%, had OC disease while the remaining 30% had NOC disease (Table 1). Of the NOC patients, 19% had ECE, 8% SVI and 3% LNI. Patients with NOC disease had a higher average PSA than OC patients (Mean: 8.6 ng/ml vs. 7.6 ng/ml), higher biopsy Gleason score (GS8: 9.9% vs.

Table 1 Prostate cancer research consortium patient cohort characteristics

	PCRC data (n = 603)	OC (n = 427)	NOC (n = 176)	*p*
Age (y)				
Mean (Median)	61.1 (62)	60.7 (61)	62.2 (63)	*0.01*
Range	42-74	42-74	42-74	
Family history (%)				
PCa history	151 (25.0)	107 (25.1)	46 (28.9)	*0.70*
Ca history	121 (20.1)	85 (19.9)	26 (16.4)	
No history	331 (54.9)	235 (55.0)	87 (54.7)	
Clinical stage, DRE (%)				
T1c	267 (44.3)	188 (44.1)	79(44.9)	*0.69*
T2a	144 (23.9)	99 (23.1)	45 (25.6)	
Not reported	192(31.8)	140 (32.8)	52 (39.5)	
PSA (ng/ml)				
Mean (Median)	7.96 (7.0)	7.6 (6.7)	8.6 (7.5)	*0.02*
Range	0.7 – 40	0.7 – 40	2.1 – 36	
Biopsy GS (%)				
5	96 (17.4)	77 (19.7)	19 (11.8)	*P < 0.001*
6	205 (37.1)	163 (41.7)	42 (26.1)	
3 + 4 = 7	153 (27.7)	103 (26.3)	50 (31.1)	
4 + 3 = 7	46 (8.3)	22 (5.7)	24 (14.9)	
8	34 (6.2)	18 (4.6)	16 (9.9)	
9	18 (3.3)	8 (2.0)	10 (6.2)	
Prostatectomy GS (%)				
5	59 (9.8)	46 (10.8)	13 (7.4)	*P < 0.001*
6	151 (25.0)	125 (29.2)	26 (14.8)	
3 + 4 = 7	237 (39.3)	181 (42.4)	56 (31.8)	
4 + 3 = 7	90 (14.9)	44 (10.3)	46 (26.1)	
8	42 (7.0)	20 (4.7)	22 (12.5)	
9	24 (4.0)	11 (2.6)	13 (7.4)	
Pathological stage				
OC	427 (70)	427 (70)	-	
ECE	111 (19)	-	111 (63)	
SVI	47 (8)	-	47 (27)	
LNI	18 (3)	-	18 (10)	

4.6%), higher prostatectomy Gleason score (GS8: 12.5% vs. 4.7%) and were older (Mean: 62.2 years vs. 60.7 years). These findings were all statistically significant at the $P < 0.05$ level. No significant differences were recorded according to stage at RP for clinical stage or family history (both $P > 0.05$).

The Gleason score based on TRUS biopsy and the Gleason score recorded by pathology at RP were compared to measure the percentage of Gleason score upgrading or downgrading (Table 2). 52.9% of patients experience no Gleason upgrading or downgrading, i.e. the results of their TRUS biopsy were accurate. However, of the remaining 47.1% of patients, 34.1% experienced upgrading and 13.0% experienced down grading of their Gleason score. This indicates a 47% grading error based on TRUS biopsy.

Seven prediction models were developed using logistic regression, linear discriminant analysis, classification and regression trees, k nearest neighbours, artificial neural networks, support vector machines and random forests. On examination of the individual model fit for each classification technique, the linear discriminant analysis, classification and regression trees, artificial neural networks and support vector machines models were excluded as these classification techniques were deemed inferior in this study. This resulted in three prediction models; a model developed using logistic regression, a model developed using random forests and a model developed using k nearest neighbours. Each of these models contains the same predictor variables (PSA, clinical stage and biopsy GS) and have all been developed using the same 5-fold cross validation approach.

The discriminate ability of the three models was measured using discrimination metrics including sensitivity, specificity, Youden index, positive predictive value (PPV), negative predictive value (NPV), Brier score and AUC values (Table 3). The logistic regression (LR) model illustrates a sensitivity of 0.647 and a specificity of 0.601, indicating that this model correctly identified 64.7% of patients who had NOC disease and 60.1% of patients who had OC disease, i.e. the model discriminates between both NOC patients and OC patients to the same ability. However, these values for sensitivity and specificity, although high relative to the other results in Table 3, are low based on the fact that a perfect model would achieve a sensitivity and specificity of 1. The Youden index for the LR model is calculated as a summation of the sensitivity and specificity minus 1; therefore due to the fact that both the sensitivity and specificity are reasonably good, the Youden index for this model (0.248) is reasonably good relative to the others in Table 3. The Youden index is a useful metric when there is no preference between sensitivity and specificity. The LR model had a PPV of 0.495 and NPV of 0.800, indicating that 49.5% of patients in the sample who were predicted as being NOC by the model actually had NOC disease and 80% of patients who were predicted as being OC actually had OC disease. It should be noted that, unlike sensitivity and specificity, NPV and PPV are affected by the prevalence of disease in the sample. In this study, the prevalence of having NOC disease is 30% and of having OC disease is 70% (Table 1). When the prevalence is low the PPV will be low, regardless of the sensitivity and specificity. The Brier score for the LR model is 0.173. The maximum Brier score for a model with a prevalence of 30% is approximately 0.21. A model with a

Table 2 Percentage of Gleason score upgrading or downgrading

	Biopsy GS					
N (%)	**≤6**	**7 (3 + 4)**	**7 (4 + 3)**	**8**	**9-10**	**Total**
N	301	153	46	34	18	552
Decrease in GS	0	23 (15.1)	15 (32.6)	25 (73.5)	9 (50.0)	72 (13.0)
No change	170 (56.5)	92 (60.1)	16 (34.8)	5 (14.7)	9 (50.0)	292 (52.9)
Increase in GS	131 (43.5)	38 (24.8)	15 (32.6)	4 (11.8)	0	188 (34.1)

Brier score of 0.21 indicates that there are large differences between the predicted probabilities and the actual outcome. The AUC value for the LR model is 0.622, which is reasonably good, but an AUC of 0.70 and above would be the minimum required to consider a model useful for clinical application. When comparing the LR model AUC with those from the other classification models and clinical variables in isolation (Table 3), the AUC of 0.622 for the LR model is the highest. This is closely followed by biopsy Gleason score (AUC = 0.618, Sens = 0.623, Spec = 0.613, PPV = 0.396, NPV = 0.799, Brier = 0.179). These results would indicate that biopsy Gleason score is by far the individual predictor variable with the highest discriminate ability. The integration of biopsy Gleason score with the other clinical variables into a LR model improves the ability to predict PCa stage at RP, but this improvement is minimal, highlighting the strength of biopsy Gleason score alone. Neither the random forests (RF) model (AUC = 0.605, Sens = 0.673, Spec = 0.457, PPV = 0.339, NPV = 0.771, Brier = 0.206) nor the K nearest neighbours (kNN) model (AUC = 0.570, Sens = 0.673, Spec = 0.457, PPV = 0.339, NPV = 0.771, Brier = 0.215) achieve better overall discrimination than biopsy Gleason score alone or the LR model.

The discrimination of the 2007 Partin table was also measured (Figure 2). It should be noted that the Partin tables predict four stages at RP (OC, ECE, SVI and LNI), whereas the models developed in this study predict NOC disease, where NOC is made up of ECE, SVI and LNI. The majority of the NOC patients are made up of ECE, therefore the most appropriate Partin table prediction to look at in comparison to this study is the ECE predictions. The 2007 Partin table can predict ECE with an AUC value of 0.572 for patients with clinical stage T1c (Figure 2a) and 0.509 for patients with clinical stage T2a (Figure 2b).

The calibration of each model was graphically measured by formulation of calibration curves (Figure 3a-f). The blue line represents the fit based on Loess smoothing. A model is well calibrated if the predicted probabilities or Loess smoothing fit (blue line) lie along the 45° line. Deviations away from this indicate mis–calibration. The error bars represent the 95% confidence interval for the predicted probabilities. The LR model (Figure 3a) is well-calibrated, although there appears to be very slight deviations from the 45° line at the very low and very high predicted probabilities, indicating that some of the lower predicted probabilities may slightly under estimate the true outcome and some of the higher predicted probabilities may slightly over-predict the true probability of the patient, but it should be noted that these deviations are minimal. The RF and kNN models illustrate some mis-calibration (Figure 3b-3c), indicating that the predicted probabilities for these models deviate from the true patient probability. Biopsy Gleason score (Figure 3d) is reasonably well-calibrated; although some of the error bars at predicted probabilities of approx 0.4 and 0.5 do not conform to Loess smoothing. PSA is reasonably calibrated (Figure 3e) although some clear over-prediction is occurring at higher probabilities based on Loess smoothing. The error bars indicate that the actua probabilities are well calibrated. The calibration curve for clinical stage (Figure 3f) illustrates how narrow the band of predicted probabilities is for the model built based on clinical stage (DRE) alone. The predicted probabilities vary between approx 0.25 and 0.35. Based on this it is difficult to examine the shape of the calibration of the error bars,

Table 3 Discrimination of prediction models and individual clinical variables

	Logistic regression	Random forests	K nearest neighbours	Biopsy Gleason score	PSA	Clinical stage
Sens	0.647	0.477	0.673	0.623	0.678	0
Spec	0.601	0.714	0.457	0.613	0.446	1
Youden	0.234	0.192	0.129	0.235	0.124	0
PPV	0.495	0.410	0.339	0.396	0.340	NA
NPV	0.800	0.767	0.771	0.799	0.767	NA
Brier	0.173	0.206	0.215	0.179	0.195	0.208
AUC	0.622	0.605	0.570	0.618	0.571	0.572

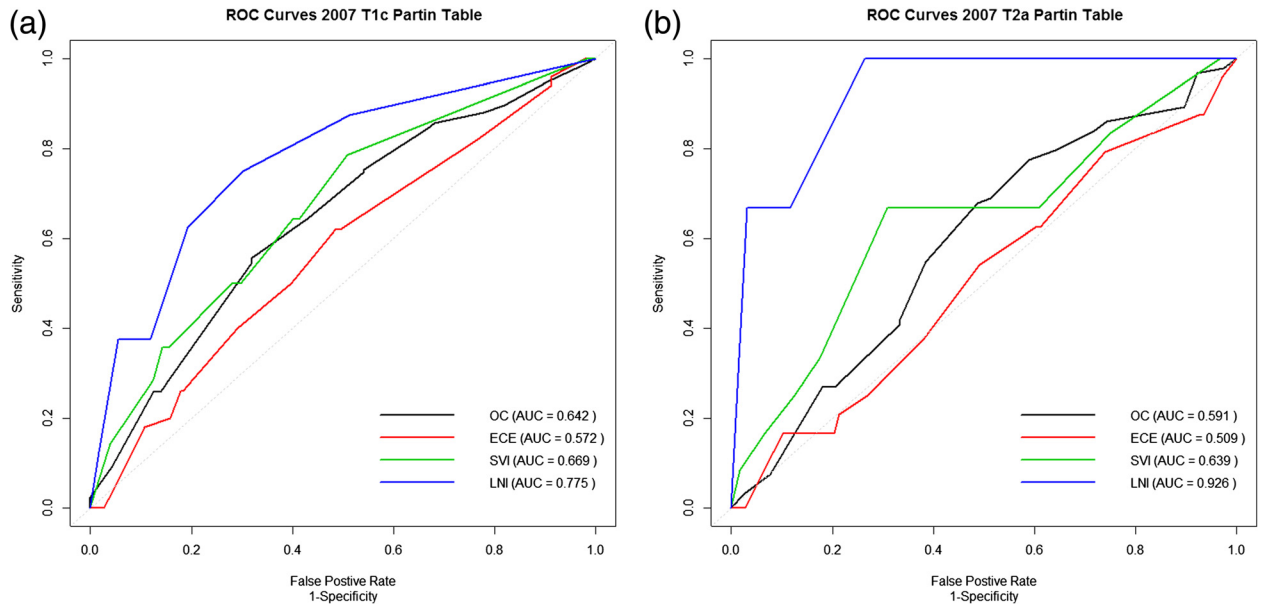

Figure 2 Predictions from the 2007 Partin tables. The patient's data was applied to the 2007 **(a)** T1c and **(b)** T2a Partin tables and the discriminate ability measured using ROC curves and AUC values. The AUC values of the ECE predictions are lower than those of the logistic regression model developed in this study.

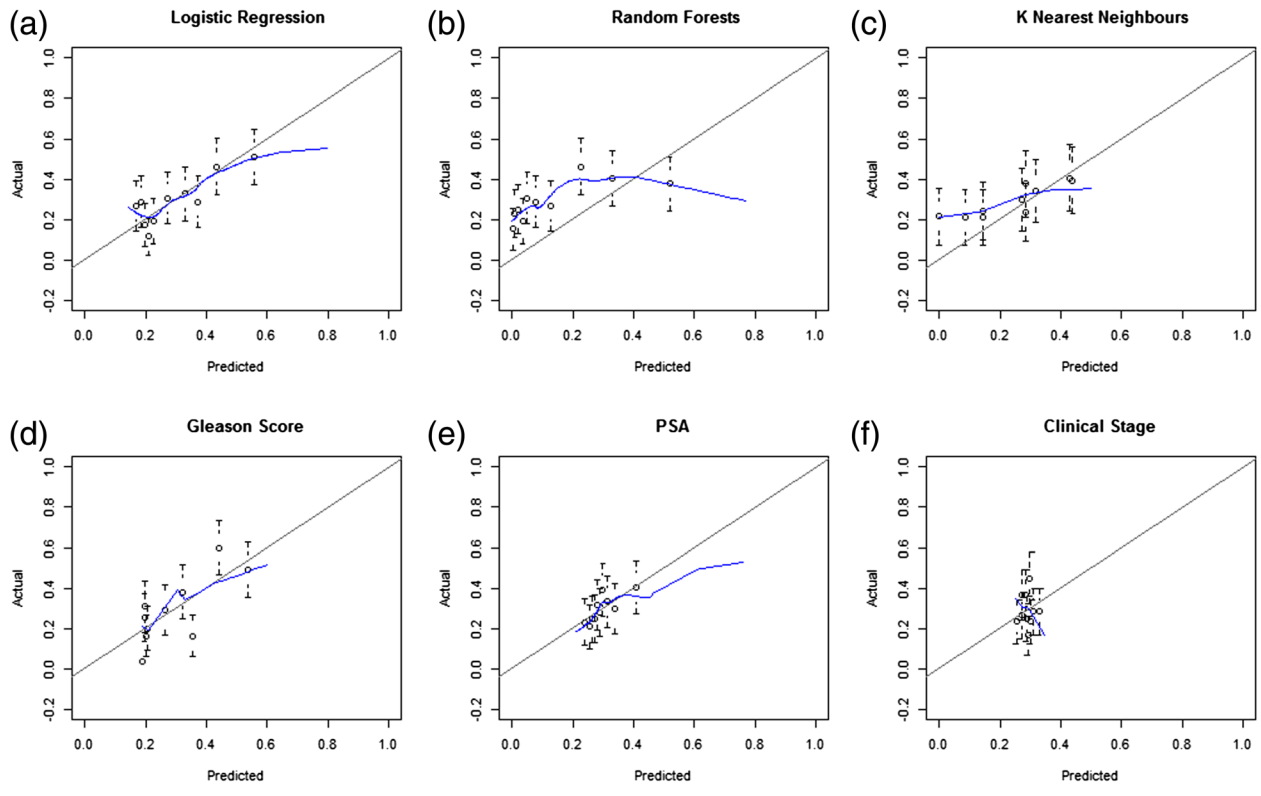

Figure 3 Calibration curves which illustrate predicted probabilities on the x-axis and the actual outcome on the y-axis. In this case the outcome is a binary variable hence Loess smoothing (blue line) was used to estimate the actual outcomes for **(a)** logistic regression model, **(b)** random forests model, **(c)** k nearest neighbours model, **(d)** biopsy Gleason score, **(e)** PSA and **(f)** clinical stage.

although the Loess smoothing blue line clearly appears to be mis-calibrated but this may in fact be due to the narrowness of the range of predicted probabilities. Regardless of the fact that the predicted probability calibration based on the error bars looks reasonably good, the narrow range of the predicted probabilities indicates the weakness of the clinical stage model and this has also been shown in previous results (Table 3).

The results of decision curve analysis are compared by means of decision curves (Figure 4), with separate decision curves for the classification models (Figure 4a) and the independent clinical variables in isolation (Figure 4b). For both figures, the straight black line at y = 0 represents the decision curve for the strategy of treating no patients for NOC disease and the grey line represents the decision curve for the strategy of treating all patients for NOC disease. The LR model is superior to the RF and kNN models as it has the highest net benefit at the majority of threshold probabilities along the x-axis (Figure 4a). From the same figure, it is also clear that the LR model is well calibrated: for the majority of threshold probabilities, the model never does worse than treating everyone (grey line) and treating no one (thin black line at net benefit = 0), unlike the other two models (RF and kNN), again illustrating that LR is the superior model in terms of discrimination, calibration and now also clinical relevance. An advantage of decision curves is the ability to identify the range of probabilities at which a model will be clinically relevant. For example, a clinician could input a new patient's clinical information into the model based on LR and calculate their predicted probability. The clinician would then refer to the decision curve and find the predicted probability along the x-axis and identify which prediction model has the highest net benefit at that point. If the LR model does not have the highest net benefit at that point, the LR model is not the most appropriate to use for this patient and an alternative (the model with the highest net benefit at that point) should be used instead. The RF and kNN models show mis-calibration at threshold probabilities between 0-25%. The range of threshold probabilities that these two models would be useful at is between 25-30%, however, at these threshold probabilities, the LR model would be the optimal prediction tool to use. The range of threshold probabilities that the LR model would be useful at is between 0-50%, and although there is a slight dip at approx 20%, at the majority of threshold probabilities this model has the highest net benefit. Of the individual clinical variables (Figure 4b), the Gleason score model appears to be the optimal model at threshold probabilities of 25% and above, below which the PSA model appears to be have a slightly higher net benefit. All of the models based on clinical variables in isolation appear to be reasonably well calibrated (except for clinical stage) as they are never worse than treating everyone and treating no one. Clinical stage (thick blue line) shows clear mis-calibration due to the fact that at threshold probabilities between approx 26-31%, the model is worse than treating everyone and this model also appears to be poorly discriminative due to the fact that between threshold probabilities of 30-35%, the model is worse than treating no one.

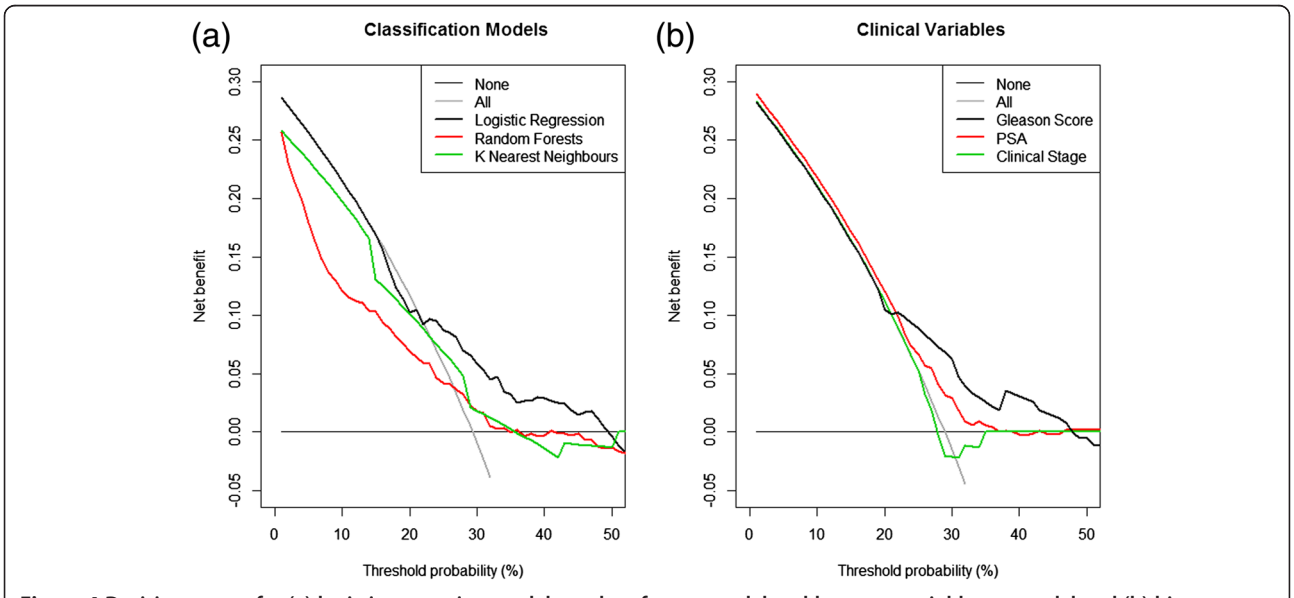

Figure 4 Decision curves for (a) logistic regression model, random forests model and k nearest neighbours model and (b) biopsy Gleason score, PSA and clinical stage. In each decision curve the solid, thin black line represents assuming no one has NOC PCa and the thin grey line represents assuming everyone has NOC PCa.

Discussion

70% of patients had OC disease while the remaining 30% had NOC disease (Table 1). This represents a 30% staging error, as the entire study cohort were assumed to have OC disease and hence underwent RP. 47.1% of patient's biopsy Gleason score was an incorrect estimate of their true Gleason score at RP, indicating that only 52.9% of patients did not experience an upgrading or downgrading of their Gleason score. Our group had previously shown a 42% Gleason score error between biopsy and RP in a smaller sample (N = 206) of the same patient cohort [24]. This level of upgrading or downgrading (42%) has been illustrated in other studies [53]. It was difficult to ascertain published figures for Gleason score upgrading or downgrading for the last number of years, particularly studies with a reasonably large sample size such as this one hence the result that 47.1% of patients experienced upgrading or downgrading of their biopsy Gleason score at RP is a significant finding of the paper.

The inclusion of the three clinical predictor variables into a statistical classification model provided a minimal improvement in predictive ability (discrimination, calibration and clinical relevance) compared to the model based on Gleason score in isolation, however, it was an improvement none-the-less. It is obvious that the statistical classification model is a welcome addition to PCa prediction, even more so due to the fact that the future of PCa staging is bound to contain complex new tests, biomarkers or features. There is no alternative to integrating multiple variables in a single prediction model [52]. This study has illustrated LR as a superior modelling technique.

Using the current clinical variables alone, excellent or even good discrimination, calibration and clinical utility will never be observed. Gleason score, PSA and clinical stage based on DRE do not contain enough information to accurately predict PCa stage at RP. New predictive features are urgently required for the prediction of PCa staging. The future of PCa prediction will likely involve the integration of novel biomarkers with existing clinical features. There are many ongoing biomarker discovery and validation studies, both published and in progress [54-63]. The modelling of such integrated data sets does not present a problem. This study has illustrated LR is as good and if not better than some of the newer more complex classification techniques. This is due in part to the fact that there are no complex relationships between PCa variables which need to be allowed for in a statistical model. The area which will require further, ongoing research is around methods to evaluate a new predictive marker/model. An initial framework to address this has been implemented in this study, an approach which examined discrimination, calibration and clinical relevance, based on previous work by Steyerberg et al. [39,64,65].

Conclusion

This study has illustrated the inability of the current clinical variables to accurately predict PCa stage. This is in part due to the fact that the most predictive clinical variable, Gleason score, over or underestimates the true Gleason score at RP in 47.1% of patients. New biomarkers or features are urgently required to address the problem clinician's face regarding accurately prognosticating the appropriate treatment for PCa patients. This paper has illustrated an approach which may be useful in the evaluation of such novel biomarkers or features, or prediction models in general.

Abbreviations

PCa: Prostate cancer; RP: Radical prostatectomy; PSA: Prostate specific antigen; DRE: Digital rectal exam; TRUS: Transrectal ultrasound; OC: Organ confined; ECE: Extracapsular extension; SVI: Seminal vesicle invasion; LNI: Lymph node involvement; NOC: Non-organ confined disease; BCR: Biochemical recurrence; ROC: Receiver operation characteristic; AUC: Area under the curve; PCRC: Prostate cancer research consortium; BPH: Benign prostatic hyperplasia; GS: Gleason score; CI: Confidence interval; LR: Logistic regression; RF: Random forests; kNN: K nearest neighbours; PPV: Positive predictive value; NPV: Negative predictive value.

Competing interests

The authors declare they have no completing interests.

Authors' contributions

RWW and TBM jointly contributed to conception and design. RWW developed the Prostate Cancer Research Consortium biobank and hence is responsible for the acquisition of data. SB carried out data preparation, data analysis and interpretation of data. YF assisted with the statistical analysis of the data. SB drafted and revised the manuscript, with TBM and RWW giving final approval. All authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge the following contributors to the Prostate Cancer Research Consortium (PCRC) bioresource: TH Lynch, TED McDermott, R Grainger, KJ O'Malley, GP Smyth, RP Power, DP Hickey, T Creagh, P Mohan, DM Little and the institution research nurses M Brenan, T Martin, C Morrow, C Schilling and R O'Connor.

Funding for this research was acquired from the Irish Research Council (IRC) previously known as Irish Research Council for Science, Engineering and Technology (IRCSET) through the University College Dublin Bioinformatics and Systems Biology PhD Programme. This funding source played no role in the study design; collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication. The Prostate Cancer Research Consortium bioresource is partially supported by the Wellcome Trust-Health Research Board (HRB) Dublin Centre for Clinical Research and by the Irish Cancer Society.

Author details

[1]UCD School of Medicine and Medical Science, University College Dublin, Dublin, Ireland. [2]UCD School of Mathematical Sciences, University College Dublin, Dublin, Ireland. [3]UCD School of Biomolecular and Biomedical Science, University College Dublin, Dublin, Ireland. [4]Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland.

Received: 8 January 2013 Accepted: 8 November 2013
Published: 15 November 2013

References

1. Siegel R, Naishadham D, Jemal A: **Cancer statistics.** *CA Cancer J Clin* 2012, **62:**10–29.
2. Oon SF, Pennington SR, Fitzpatrick JM, Watson RW: **Biomarker research in prostate cancer–towards utility, not futility.** *Nat Rev Urol* 2011, **8:**131–138.

3. Berglund RK, Masterson TA, Vora KC, Eggener SE, Eastham JA, Guillonneau BD: **Pathological upgrading and up staging with immediate repeat biopsy in patients eligible for active surveillance.** *J Urol* 2008, **180:**1964–1967. discussion 1967–1968.
4. Carter HB: **Management of low (favourable)-risk prostate cancer.** *BJU Int* 2011, **108:**1684–1695.
5. Etzioni R, Penson DF, Legler JM, di Tommaso D, Boer R, Gann PH, Feuer EJ: **Overdiagnosis due to prostate-specific antigen screening: lessons from U.S. prostate cancer incidence trends.** *J Natl Cancer Inst* 2002, **94:**981–990.
6. Tosoian JJ, Trock BJ, Landis P, Feng Z, Epstein JI, Partin AW, Walsh PC, Carter HB: **Active surveillance program for prostate cancer: an update of the Johns Hopkins experience.** *J Clin Oncol* 2011, **29:**2185–2190.
7. Makarov DV, Trock BJ, Humphreys EB, Mangold LA, Walsh PC, Epstein JI, Partin AW: **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:**1095–1101.
8. Partin AW, Kattan MW, Subong EN, Walsh PC, Wojno KJ, Oesterling JE, Scardino PT, Pearson JD: **Combination of prostate-specific antigen, clinical stage, and Gleason score to predict pathological stage of localized prostate cancer. A multi-institutional update.** *JAMA* 1997, **277:**1445–1451.
9. Partin AW, Mangold LA, Lamm DM, Walsh PC, Epstein JI, Pearson JD: **Contemporary update of prostate cancer staging nomograms (Partin Tables) for the new millennium.** *Urology* 2001, **58:**843–848.
10. Partin AW, Yoo J, Carter HB, Pearson JD, Chan DW, Epstein JI, Walsh PC: **The use of prostate specific antigen, clinical stage and Gleason score to predict pathological stage in men with localized prostate cancer.** *J Urol* 1993, **150:**110–114.
11. Huang Y, Isharwal S, Haese A, Chun FK, Makarov DV, Feng Z, Han M, Humphreys E, Epstein JI, Partin AW, Veltri RW: **Prediction of patient-specific risk and percentile cohort risk of pathological stage outcome using continuous prostate-specific antigen measurement, clinical stage and biopsy Gleason score.** *BJU Int* 2011, **107:**1562–1569.
12. Smaletz O, Scher HI, Small EJ, Verbel DA, McMillan A, Regan K, Kelly WK, Kattan MW: **Nomogram for overall survival of patients with progressive metastatic prostate cancer after castration.** *J Clin Oncol* 2002, **20:**3972–3982.
13. Stephenson AJ, Scardino PT, Eastham JA, Bianco FJ Jr, Dotan ZA, DiBlasio CJ, Reuther A, Klein EA, Kattan MW: **Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy.** *J Clin Oncol* 2005, **23:**7005–7012.
14. Stephenson AJ, Scardino PT, Eastham JA, Bianco FJ Jr, Dotan ZA, Fearn PA, Kattan MW: **Preoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy.** *J Natl Cancer Inst* 2006, **98:**715–717.
15. D'Amico AV, Whittington R, Malkowicz SB, Cote K, Loffredo M, Schultz D, Chen MH, Tomaszewski JE, Renshaw AA, Wein A, Richie JP: **Biochemical outcome after radical prostatectomy or external beam radiation therapy for patients with clinically localized prostate carcinoma in the prostate specific antigen era.** *Cancer* 2002, **95:**281–286.
16. Cooperberg MR, Pasta DJ, Elkin EP, Litwin MS, Latini DM, Du Chane J, Carroll PR: **The University of California, San Francisco cancer of the prostate risk assessment score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy.** *J Urol* 2005, **173:**1938–1942.
17. Haese A, Chaudhari M, Miller MC, Epstein JI, Huland H, Palisaar J, Graefen M, Hammerer P, Poole EC, O'Dowd GJ, *et al*: **Quantitative biopsy pathology for the prediction of pathologically organ-confined prostate carcinoma: a multiinstitutional validation study.** *Cancer* 2003, **97:**969–978.
18. Veltri RW, Chaudhari M, Miller MC, Poole EC, O'Dowd GJ, Partin AW: **Comparison of logistic regression and neural net modeling for prediction of prostate cancer pathologic stage.** *Clin Chem* 2002, **48:**1828–1834.
19. Veltri RW, Miller MC, Partin AW, Poole EC, O'Dowd GJ: **Prediction of prostate carcinoma stage by quantitative biopsy pathology.** *Cancer* 2001, **91:**2322–2328.
20. Karakiewicz PI, Bhojani N, Capitanio U, Reuther AM, Suardi N, Jeldres C, Pharand D, Peloquin F, Perrotte P, Shariat SF, Klein EA: **External validation of the updated Partin tables in a cohort of North American men.** *J Urol* 2008, **180:**898–902. discussion 902–893.
21. Augustin H, Isbarn H, Auprich M, Bonstingl D, Al-Ali BM, Mannweiler S, Pummer K: **Head to head comparison of three generations of Partin tables to predict final pathological stage in clinically localised prostate cancer.** *Eur J Cancer* 2010, **46:**2235–2241.
22. Bhojani N, Ahyai S, Graefen M, Capitanio U, Suardi N, Shariat SF, Jeldres C, Erbersdobler A, Schlomm T, Haese A, *et al*: **Partin tables cannot accurately predict the pathological stage at radical prostatectomy.** *Eur J Surg Oncol* 2009, **35:**123–128.
23. Bhojani N, Salomon L, Capitanio U, Suardi N, Shariat SF, Jeldres C, Zini L, Pharand D, Peloquin F, Arjane P, *et al*: **External validation of the updated Partin tables in a cohort of French and Italian men.** *Int J Radiat Oncol Biol Phys* 2009, **73:**347–352.
24. Fanning DM, Kay E, Fan Y, Fitzpatrick JM, Watson RW: **Prostate cancer grading: the effect of stratification of needle biopsy Gleason score 4 + 3 as high or intermediate grade.** *BJU Int* 2009, **105:**631–635.
25. Xiao WJ, Ye DW, Yao XD, Zhang SL, Dai B, Wang CF, Wang J, Zhang HL, Shen YJ, Zhu Y, *et al*: **Comparison of accuracy among three generations of Partin tables in a Chinese cohort.** *Can J Urol* 2011, **18:**5619–5624.
26. Yu JB, Makarov DV, Sharma R, Peschel RE, Partin AW, Gross CP: **Validation of the Partin nomogram for prostate cancer in a national sample.** *J Urol* 2010, **183:**105–111.
27. Gleason DF: **Classification of prostatic carcinomas.** *Cancer Chemother Rep* 1966, **50:**125–128.
28. Epstein JI, Allsbrook WC Jr, Amin MB, Egevad LL: **The 2005 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma.** *Am J Surg Pathol* 2005, **29:**1228–1242.
29. Epstein JI: **An update of the Gleason grading system.** *J Urol* 2010, **183:**433–440.
30. Albertsen PC, Hanley JA, Fine J: **20-year outcomes following conservative management of clinically localized prostate cancer.** *JAMA* 2005, **293:**2095–2101.
31. Albertsen PC, Hanley JA, Penson DF, Barrows G, Fine J: **13-year outcomes following treatment for clinically localized prostate cancer in a population based cohort.** *J Urol* 2007, **177:**932–936.
32. Chun FK, Karakiewicz PI, Briganti A, Walz J, Kattan MW, Huland H, Graefen M: **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.** *BJU Int* 2007, **99:**794–800.
33. Balch CM, Buzaid AC, Soong SJ, Atkins MB, Cascinelli N, Coit DG, Fleming ID, Gershenwald JE, Houghton A Jr, Kirkwood JM, *et al*: **Final version of the American joint committee on cancer staging system for cutaneous melanoma.** *J Clin Oncol* 2001, **19:**3635–3648.
34. Hosmer DW, Lemeshow S: *Applied Logistic Regression.* New York: Wiley; 1989.
35. Breiman L: **Random Forests.** *Mach Learning* 2001, **45:**5–32.
36. Duda RO, Hart PE, Stork DG: *Pattern classification.* New York: Wiley; 2001.
37. Vickers AJ, Cronin AM, Elkin EB, Gonen M: **Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers.** *BMC Med Inform Decis Mak* 2008, **8:**53.
38. Pepe MS: *The statistical evaluation of medical tests for classification and prediction.* Oxford: Oxford University Press; 2004.
39. Steyerberg EW: *Clinical prediction models.* New York: Springer; 2009.
40. Vickers AJ, Elkin EB: **Decision curve analysis: a novel method for evaluating prediction models.** *Med Decis Making* 2006, **26:**565–574.
41. Steyerberg EW, Vickers AJ: **Decision curve analysis: a discussion.** *Med Decis Making* 2008, **28:**146–149.
42. Fox J: *Robust regression: Appendix to an R and S-PLUS companion to applied regression.* California: SAGE Publications; 2002.
43. Harrell FE: *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.* New York: Springer; 2001.
44. Altman DC, Bland JM: **Diagnostic tests 1: sensitivity and specificity.** *BMJ* 1994, **308:**1552.
45. Altman DC, Bland JM: **Diagnostic tests 2: predictive values.** *BMJ* 1994, **309:**102.
46. Youden WJ: **Index for rating diagnostic tests.** *Cancer* 1950, **3:**32–35.
47. Fluss R, Faraggi D, Reiser B: **Estimation of the Youden index and its associated cutoff point.** *Biom J* 2005, **47:**458–472.
48. Brier GW: **Verification of forecasts expressed in terms of probability.** *Mon Weather Rev* 1950, **78:**1–3.
49. Bewick V, Cheek L, Ball J: **Statistics review 13: receiver operating characteristic curves.** *Crit Care* 2004, **8:**508–512.

50. Hanley JA, McNeil BJ: **A method of comparing the areas under receiver operating characteristic curves derived from the same cases.** *Radiology* 1983, **148:**839–843.
51. Pepe MS, Janes HE: **Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer.** *J Natl Cancer Inst* 2008, **100:**978–979.
52. Vickers AJ: **Prediction models in cancer care.** *CA Cancer J Clin* 2011, **61:**315–326.
53. King CR, Long JP: **Prostate biopsy grading errors: a sampling problem?** *Int J Cancer* 2000, **90:**326–330.
54. Steuber T, Helo P, Lilja H: **Circulating biomarkers for prostate cancer.** *World J Urol* 2007, **25:**111–119.
55. Shariat SF, Karam JA, Roehrborn CG: **Blood biomarkers for prostate cancer detection and prognosis.** *Future Oncol* 2007, **3:**449–461.
56. Sardana G, Dowell B, Diamandis EP: **Emerging biomarkers for the diagnosis and prognosis of prostate cancer.** *Clin Chem* 2008, **54:**1951–1960.
57. Reed AB, Parekh DJ: **Biomarkers for prostate cancer detection.** *Expert Rev Anticancer Ther* 2010, **10:**103–114.
58. Prensner JR, Rubin MA, Wei JT, Chinnaiyan AM: **Beyond PSA: the next generation of prostate cancer biomarkers.** *Sci Transl Med* 2012, **4:**127rv123.
59. Martin SK, Vaughan TB, Atkinson T, Zhu H, Kyprianou N: **Emerging biomarkers of prostate cancer (Review).** *Oncol Rep* 2012, **28:**409–417.
60. Margreiter M, Stangelberger A, Valimberti E, Herwig R, Djavan B: **Biomarkers for early prostate cancer detection.** *Minerva Urol Nefrol* 2008, **60:**51–60.
61. Bickers B, Aukim-Hastie C: **New molecular biomarkers for the prognosis and management of prostate cancer–the post PSA era.** *Anticancer Res* 2009, **29:**3289–3298.
62. Bensalah K, Lotan Y, Karam JA, Shariat SF: **New circulating biomarkers for prostate cancer.** *Prostate Cancer Prostatic Dis* 2008, **11:**112–120.
63. Artibani W: **Landmarks in prostate cancer diagnosis: the biomarkers.** *BJU Int* 2012, **110**(Suppl 1):8–13.
64. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B: **Assessing the incremental value of diagnostic and prognostic markers: a review and illustration.** *Eur J Clin Invest* 2012, **42:**216–228.
65. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW: **Assessing the performance of prediction models: a framework for traditional and novel measures.** *Epidemiology* 2010, **21:**128–138.

doi:10.1186/1472-6947-13-126
Cite this article as: Boyce *et al.*: Evaluation of prediction models for the staging of prostate cancer. *BMC Medical Informatics and Decision Making* 2013 **13**:126.