Asian Journal of Andrology (2017) 19, 586–590
www.asiaandro.com; www.ajandrology.com

Open Access

ORIGINAL ARTICLE

Prostate Disease

Prostate cancer prediction using the random forest algorithm that takes into account transrectal ultrasound findings, age, and serum levels of prostate-specific antigen

Li-Hong Xiao[1,2], Pei-Ran Chen[2], Zhong-Ping Gou[1], Yong-Zhong Li[3], Mei Li[1], Liang-Cheng Xiang[2], Ping Feng[1]

The aim of this study is to evaluate the ability of the random forest algorithm that combines data on transrectal ultrasound findings, age, and serum levels of prostate-specific antigen to predict prostate carcinoma. Clinico-demographic data were analyzed for 941 patients with prostate diseases treated at our hospital, including age, serum prostate-specific antigen levels, transrectal ultrasound findings, and pathology diagnosis based on ultrasound-guided needle biopsy of the prostate. These data were compared between patients with and without prostate cancer using the Chi-square test, and then entered into the random forest model to predict diagnosis. Patients with and without prostate cancer differed significantly in age and serum prostate-specific antigen levels ($P < 0.001$), as well as in all transrectal ultrasound characteristics ($P < 0.05$) except uneven echo ($P = 0.609$). The random forest model based on age, prostate-specific antigen and ultrasound predicted prostate cancer with an accuracy of 83.10%, sensitivity of 65.64%, and specificity of 93.83%. Positive predictive value was 86.72%, and negative predictive value was 81.64%. By integrating age, prostate-specific antigen levels and transrectal ultrasound findings, the random forest algorithm shows better diagnostic performance for prostate cancer than either diagnostic indicator on its own. This algorithm may help improve diagnosis of the disease by identifying patients at high risk for biopsy.

Asian Journal of Andrology (2017) **19**, 586–590; doi: 10.4103/1008-682X.186884; published online: 2 September 2016

Keywords: diagnosis; prostate cancer; prostate-specific antigen; random forest algorithm; transrectal ultrasound characteristics

INTRODUCTION

Prostate cancer accounts for 23%–26% of newly diagnosed cancers in men and 9%–10% of cancer-related deaths in the US and Europe, making it the most frequently diagnosed malignant tumor in males.[1,2] In 2008, prostate cancer incidence in Chinese men was 11 per 100 000, and it increased by 12% from 1998 to 2008.[3] Prostate cancer is closely related to age: more than 95% of all patients are older than 60 years.[1]

The standard method for diagnosing prostate cancer is pathology analysis of systematic, transrectal ultrasound-guided prostate biopsy.[4] However, this approach is invasive and can lead to bleeding and infection. The procedure also has relatively low sensitivity, with the recommended 12-core biopsy providing a sensitivity of 36%–58%[5] and 18-core biopsy providing only slightly higher sensitivity of 53%–58%.[6] In addition, biopsy is invasive and associated with some risks, such as bleeding and infection.

Transrectal ultrasound has been widely used to detect patients at risk of prostate cancer. This technique can offer relatively high sensitivity (44%–90%) and specificity (30%–74%),[7] but its performance is limited by variability in prostate cancer ultrasound signals.

To aid in early detection of prostate cancer, many clinicians also look for elevated serum levels of prostate-specific antigen (PSA).[8] PSA is produced by prostate epithelium and is a reliable marker of prostate disease. PSA levels above 4.0 ng ml^{-1} during screening are widely considered to indicate the need for biopsy,[9] and levels above 10 ng ml^{-1} are present in approximately 50% of patients with prostate cancer.[10] However, the specificity of PSA as a prostate cancer marker is only about 60% since it can also be elevated in patients with benign prostatic hypertrophy, prostatitis, and other nonmalignant conditions.[11] Its sensitivity is only around 56% based on a threshold of 4 ng ml^{-1}.[12] Indeed, a small proportion of patients with prostate cancer have PSA levels below 4.0 ng ml^{-1}.[13]

Thus, the available evidence indicates that transrectal ultrasound, age, and PSA strongly correlate with prostate cancer but do not show sufficiently strong diagnostic performance on their own. This prompts the question of whether the three markers can be combined into a model that predicts prostate cancer reliably. Building such a model requires analyzing a large number of possibly predictive variables that are interrelated in complex ways, for which the random forest machine-learning algorithm appears to be superior to traditional statistical methods.[14] The random forest procedure relies on a large number of classifiers, helping to reduce bias, tolerate outliers, and avoid overfitting. The procedure has already been used to generate

[1]Institute of Clinical Trials, West China Hospital, Sichuan University, Chengdu, China; [2]Department of Epidemiology and Biostatistics, West China School of Public Health, Sichuan University, Chengdu, China; [3]Department of Ultrasound, West China Hospital, Sichuan University, Chengdu, China.
Correspondence: Dr. P Feng (pfyq@yahoo.com)
Received: 30 March 2016; Revised: 13 May 2016; Accepted: 01 July 2016

insights in many fields, but we are unaware of studies applying it to prostate cancer.

In the present study, we used the random forest algorithm to predict prostate cancer in patients with prostate diseases by combining age, PSA level, and such transrectal ultrasound findings as abnormal blood flow signals, prostate boundary, and the line dividing the prostate and rectum. Our goal was to examine whether we could achieve a technique sufficiently reliable to diagnose prostate cancer in the absence of prostate biopsy.

MATERIALS AND METHODS

Patients

The study was approved by the Independent Ethics Committee of West China Hospital, Sichuan University, Chengdu, Sichuan, China. Patients diagnosed and treated for prostate diseases at West China Hospital, Sichuan University, Chengdu, Sichuan, China between January 2008 and September 2011 were consecutively enrolled in this study. Data on patient age, serum PSA levels, transrectal ultrasound, and pathology analysis of prostate biopsies were collected. For comparing the distribution between patients with and without prostate cancer, age and serum PSA levels were treated as categorical variables: age, <50, 50–59, 60–69, and ≥70 years; PSA, <4, 4–10, 10–20, and ≥20 ng ml^{-1}.

Transrectal ultrasound

All patients were examined in the left lateral decubitus position using a 5-MHz or 7.5-MHz, convex or linear array biplane rectal probe connected to a PHILIPS HDI color Doppler ultrasound system. Two well-trained and experienced ultrasound practitioners performed the ultrasound evaluations. They followed standard operating procedures. The transducer probe was covered with a sterile condom and placed in the rectum. Gray-scale ultrasound images were examined in transverse and sagittal planes to classify the patient in terms of the following nine categorical variables: prostate shape (normal or abnormal), prostate boundary (clear or unclear), boundary between internal and external glands (clear or unclear), the line dividing prostate and rectum (clear or unclear), the line dividing prostate and seminal vesicle glands (clear or unclear), the presence of nodules (yes or no), enlargement of lymph nodes around the prostate (yes or no), uneven echo (yes or no), and presence of hypoechoic lesions (yes or no). In addition, color Doppler imaging was used to determine whether blood flow signal was normal or abnormal.

Prostate biopsy and pathology

Prior to biopsy, patients were informed of the risks and benefits of the procedure, and they gave written consent. Under the guidance of transrectal ultrasound, 12-core biopsy of suspicious lesions was conducted using a 16-gauge automatic biopsy needle (BARD, Covington, Georgia, USA). Prostate carcinoma tissue was identified based on cellular density, microvascularity, and loss of glandular architecture.[15]

Statistical analysis and random forest modeling

Statistical analysis was performed using the RStudio version 0.99 (RStudio Inc., Boston, Massachusetts, United States) statistical software package for 32-bit Windows, which runs R version 3.1.3 (R Core Team 2015, https://www.r-project.org). Differences in categorical ultrasound variables between patients with and without prostate cancer were assessed for significance using the Chi-square test. Differences in age and serum PSA levels were also assessed using the Chi-square test. The threshold of significance was defined as $P < 0.05$.

Age and PSA were entered into the random forest procedure as continuous variables and transrectal ultrasound findings as 10 dichotomous variables. Data were sampled using random bootstrapping[14,16] to generate various training data sets, which were then classified according to the characteristics of the numerous variables using classification and regression trees. To minimize classification error in the training data, the model randomly selected a subset of feature variables (m_{try}) from the 12 input variables. Since the recommended number of feature variables was approximately 3.46, the square root of 12,[17] we tested values of 3, 4, and 5 to select the optimal results. Approximately one-third of the total data set was not randomly sampled; this out-of-bag (OOB) data[18] served as the testing set. Training data sets, and the corresponding trees, were repeatedly generated until the OOB error rate had stabilized. The random forest method then selected the model with the lowest OOB error rate and generated a confusion matrix, which included the predicted and actual classification data. Individual patients were then classified as having prostate cancer or not by a "polling procedure." The input variables in the optimized model were ranked by relative importance in predicting prostate cancer based on the mean decrease in accuracy and the mean decrease in Gini coefficient.[17] Every patient was displayed on a multidimensional scaling plot based on inter-patient distances estimated by the random forest model.[17]

The use of random sampling during modeling means that the random forest approach should generate a family of classification and regression trees that are superior to any single tree. The use of internal OOB error, which provides an unbiased estimate of error and predictive ability, means that the random forest approach does not require cross-validation for model optimization, in contrast to many other machine-learning algorithms. We treated transrectal ultrasound findings as ten variables rather than a single variable to avoid loss of information, and the random forest approach is well suited to capture potential complex interactions among the individual variables.

RESULTS

Patient characteristics

A consecutive series of 941 patients was analyzed in this study (**Table 1**), 358 of whom (38.04%) were diagnosed with prostate cancer. Median age of all patients was 71.00 years (range, 24.00–88.00 years), with just over half (530, 56.32%) older than 70. The age distribution differed significantly between those with prostate cancer and those with noncancerous prostate disease ($P < 0.001$). Across all patients, median level of PSA in serum was 15.63 ng ml^{-1} (interquartile range, 9.50–47.30 ng ml^{-1}), and patients fell into the following four subgroups: <4.0 ng ml^{-1}, 49 patients; 4–10 ng ml^{-1}, 212 patients; 10–20 ng ml^{-1}, 270 patients; and >20 ng ml^{-1}, 410 patients. There were 74.58% cancer patients who had >20 ng ml^{-1} PSA. The largest proportion of patients with noncancerous prostate disease (36.02%) had 10–20 ng ml^{-1} PSA.

Comparison of transrectal ultrasound findings between patients with cancer and noncancerous disease (**Table 1**) revealed significant differences in all variables measured ($P < 0.05$), with the exception of uneven echo ($P = 0.609$).

Random forest prediction of prostate cancer

Using the random forest model, the combination of age, serum PSA levels, and transrectal ultrasound findings was tested for its ability to predict the presence of prostate cancer. Cancer was definitively diagnosed based on pathology analysis of prostate biopsies. Of the three runs using m_{try} values of 3, 4, or 5, we obtained optimal results with 3, which gave a low OOB error rate of 16.90% (**Figure 1**). In addition, this model had a cross-validated error of 16.87%, similar to the OOB error rate. Accuracy was 83.10%; sensitivity, 65.64%; specificity, 93.83%; positive predictive value (PPV), 86.72%; and negative predictive

value (NPV), 81.64% (**Table 2**). The final random forest model was built from 500 trees (*ntree* = 500), and the OOB error rate decreased quickly with increasing tree number until approximately *ntree* = 50, after which it remained fairly constant (**Figure 1**).

Based on mean decreases in accuracy and Gini coefficient, the three most important variables for predicting prostate cancer with the random forest model were age, serum PSA level, and the boundary between internal and external glands (**Figure 2**). A multidimensional scaling plot based on all input variables showed close clustering within and between groups of patients with prostate cancer or noncancerous disease (**Figure 3**).

Table 1: Age, serum levels of prostate-specific antigen (PSA), and transrectal ultrasound findings of 941 Chinese patients with prostate disease

Characteristic	*n*	*Prostate cancer, n (%)*	*No cancer, n (%)*	χ^2	*P value*
Age (year)					
<50	18	5 (1.40)	13 (2.23)	53.32	<0.001
50-59	86	14 (3.91)	72 (12.35)		
60-69	307	86 (24.02)	221 (37.91)		
≥70	530	253 (70.67)	277 (47.51)		
PSA (ng ml^{-1})					
<4	49	5 (1.40)	44 (7.55)	232.10	<0.001
4≤ PSA <10	212	26 (7.26)	186 (31.90)		
10≤ PSA <20	270	60 (16.76)	210 (36.02)		
≥20	410	267 (74.58)	143 (24.53)		
Ultrasound (shape)					
Normal	422	114 (31.84)	308 (52.83)	39.50	<0.001
Abnormal	519	244 (68.16)	275 (47.17)		
Ultrasound (prostate boundary)					
Clear	799	240 (67.04)	559 (95.88)	144.02	<0.001
Not clear	142	118 (32.96)	24 (4.12)		
Ultrasound (boundary between internal and external glands)					
Clear	717	178 (49.72)	539 (92.45)	223.30	<0.001
Not clear	224	180 (50.28)	44 (7.55)		
Ultrasound (dividing line between prostate and rectum)					
Clear	879	298 (83.24)	581 (99.66)	97.13	<0.001
Not clear	62	60 (16.76)	2 (0.34)		
Ultrasound (dividing line between prostate and seminal vesicle glands)					
Clear	890	308 (86.03)	582 (99.83)	82.34	<0.001
Not clear	51	50 (13.97)	1 (0.17)		
Ultrasound (nodule)					
No	455	155 (43.30)	300 (51.46)	5.92	0.015
Yes	486	203 (56.70)	283 (48.54)		
Ultrasound (lymph node enlargement					
No	930	348 (97.21)	582 (99.83)	11.03	0.001
Yes	11	10 (2.79)	1 (0.17)		
Ultrasound (uneven echo)					
No	71	25 (6.98)	46 (7.89)	0.26	0.609
Yes	870	333 (93.02)	537 (92.11)		
Ultrasound (hypoecho)					
No	538	151 (42.18)	387 (66.38)	53.06	<0.001
Yes	403	207 (57.82)	196 (33.62)		
Ultrasound (abnormal blood flow signals)					
No	519	115 (32.12)	404 (69.30)	123.92	<0.001
Yes	422	243 (67.88)	179 (30.70)		

PSA: prostate-specific antigen

DISCUSSION

Here, we show that the random forest algorithm trained with a relatively large data set of nearly 1000 patients can predict prostate

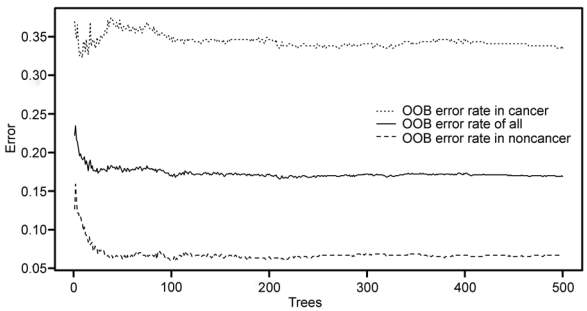

Figure 1: Out-of-bag (OOB) error rate to assess the quality of random forest prediction of prostate cancer, shown as a function of the number of decision trees generated during machine learning. The middle line depicts the OOB error rate of all data. The top line depicts the OOB error rate in the subgroup of patients with prostate cancer, and the bottom line shows the OOB error rate in the subgroup of patients with noncancerous prostate disease. For all data, $m_{try} = 3$.

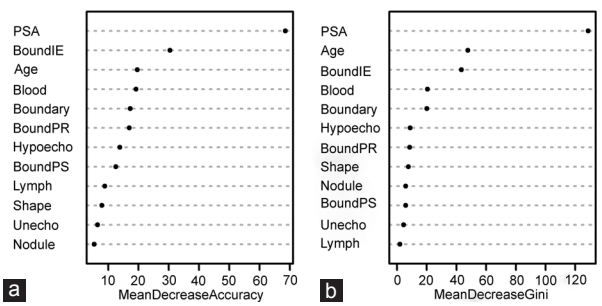

Figure 2: Ranking of input variables in the random forest model to predict prostate cancer. (**a**) Mean decrease accuracy. (**b**) Mean decrease gini. Variables are listed from most important to least important based on the mean decrease in accuracy and mean decrease in the Gini coefficient. Blood: abnormal blood flow signal; BoundIE: boundary between internal and external glands; BoundPR: dividing line between prostate and rectum; BoundPS: dividing line between prostate and seminal vesicle glands; Boundary: prostate boundary; Lymph: lymph node enlargement; PSA: serum levels of prostate-specific antigen; Shape: prostate shape.

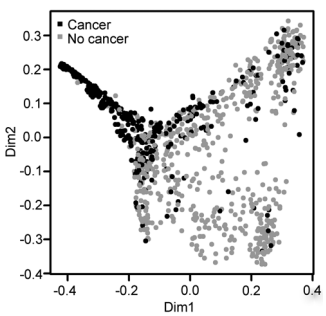

Figure 3: Multidimensional scaling plot of patients with prostate disease. Black dots indicate individual patients with prostate cancer; gray dots, patients with noncancerous disease. Dim 1 refers to dimension 1 and Dim 2 refers to dimension 2.

Table 2: Prediction of prostate cancer based on a random forest model incorporating age, serum PSA levels and transrectal ultrasound findings

Prediction	*Based on prostate biopsy*		*Total (n)*	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*PPV (%)*	*NPV (%)*
	Cancer (n)	*No cancer (n)*						
Cancer	235	36	271	83.10	65.64	93.83	86.72	81.64
No cancer	123	547	670					
Total	358	583	941					

m_{try}=3, *ntree*=500. PSA: prostate-specific antigen; Accuracy: the proportion of positive and negative results that are true positive and true negative results using the model; Sensitivity: the proportion of positives that are correctly identified as cancer using the model; Specificity: the proportion of negatives that are correctly identified as no cancer using the model; PPV: positive predictive value, the proportion of positive results that are true positive results using the model; NPV: negative predictive value, the proportion of negative results that are true negative results using the model

cancer in patients with prostate disease using a combination of transrectal ultrasound results, age, and serum PSA level. The diagnostic performance is superior to that obtained using serum PSA levels or ultrasound findings on their own. These results demonstrate the possibility of reliably predicting prostate cancer in patients who cannot be, or refuse to be, diagnosed based on prostate biopsy.

Our random forest model predicted prostate cancer with a diagnostic accuracy of 83.10%, sensitivity of 65.64%, specificity of 93.83%, and PPV of 86.72%, based on the gold standard of prostate biopsy pathology. This diagnostic performance is superior to that of PSA alone, which is associated with sensitivity of 46.44%[12] and specificity of 60.00%,[11] and of transrectal ultrasound on its own, which is associated with sensitivity of 14.98%–63.46%, specificity of 75.82%–92.28%, and PPV of 60.00%.[19,20] The sensitivity of our random forest model was much higher than the 20.50% and 32.20% specified by the American Cancer Society based on respective PSA cutoff values of 4.0 or 3.0 ng ml^{-1}, respectively.[8] The OOB error rate of our model was 16.90%, indicating good predictive ability. These findings suggest that random forest-based prediction can provide a satisfactory alternative to biopsy in patients with prostate disease. In particular, the high specificity may make our approach useful for screening.

Random forest learning, unlike classical statistical models, can provide an unbiased ranking of the relative importance of input variables for predicting the outcome. The ranking in our model highlighted age, serum PSA levels, and the boundary between internal and external glands as the three most important predictors of prostate cancer. These findings are consistent with the widespread use of age and PSA levels as prostate cancer markers. In addition, our results argue that when using transrectal ultrasound to screen patients with prostate disease, clinicians should focus on the boundary between internal and external glands.

Transrectal ultrasound can detect several characteristic features of prostate cancer. Most carcinomas appear as diffuse, echo-poor changes in the peripheral, and periurethral transition zones;[21] an echo-poor nodule in the periurethral zone; a hypervascular echo-poor nodule in the periurethral zone; and/or as a nodule surrounded by altered echogenicity.[22] Another study suggests that the combination of a hypoechoic nodule in the peripheral zone, diffuse hypoechogenic prostate changes and loss of defined limits between the peripheral zone and internal gland is diagnostic of prostate carcinoma.[20] Therefore, we included transrectal ultrasound in our random forest model to predict prostate cancer. In fact, we incorporated 10 dichotomous ultrasound variables to capture as much information as possible that might help predict disease. We found that all but one variable (uneven echo, $P = 0.609$) differed significantly between patients with cancer and noncancerous disease. This highlights the value of transrectal ultrasound for detecting prostate cancer.

At the same time, this technique shows limited accuracy, sensitivity, and/or PPV. While prostate carcinoma tissue usually presents as a hypoechoic lesion in gray-scale images, it can sometimes present instead as hyperechoic or isoechoic.[23] It is true that hypoechoic lesions were observed in 57.82% of our patients with cancer (207 of 358), but in only 33.62% of patients with noncancerous disease (196 of 583) ($P < 0.001$). However, another study suggests that diagnosing prostate cancer based on hypoechoic lesions alone would miss as many as half of true cases (52.19%, 131 of 251).[24]

Patients aged 70 and older accounted for 70.67% of those with prostate cancer in our study population, consistent with the observation that prostate cancer incidence increases with age. The overall incidence of prostate cancer in our population, 38.04% (358 of 941), is similar to the 39.43% (69 of 175) reported in a study of German patients with suspected prostate cancer[19] and the 36.61% (41 of 112) reported in Chinese patients with elevated serum PSA or positive digital rectal exam.[25] Among patients with PSA ≥4 ng ml^{-1}, incidence of prostate cancer in our population (39.57%) was slightly lower than the 44.56% (487 of 1093 biopsies) in the large Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening follow-up trial involving patients with PSA ≥4 ng ml^{-1} from 10 North American centers.[26] The PLCO trial contained 4801 men with PSA >4 ng ml^{-1} or positive digital rectal exam and 31.00% were biopsied in the initial round of screening; in the end, 549 (36.90%) were diagnosed with prostate cancer, giving a PPV of 11.40% (95% CI: 10.50–12.30).[26]

Nearly all our prostate cancer patients (98.60%) had PSA values ≥4 ng ml^{-1}, in agreement with the screening cutoff value of 4.0 ng ml^{-1} recommended by the American Cancer Society.[8] However, most of our patients with noncancerous prostate disease (92.45%) also had PSA ≥4 ng ml^{-1}. In contrast, 74.58% of our patients with prostate cancer had PSA ≥20 ng ml^{-1} while only 24.53% of those with noncancerous disease had PSA ≥20 ng ml^{-1}. These results suggest that PSA >20 ng ml^{-1} is associated with high probability of prostate cancer, but that serum PSA levels, like transrectal ultrasound findings, on their own lack the sensitivity and specificity to reliably predict disease.[27]

Our study shows the potential of random forest learning for predicting prostate cancer, extending the handful of biomedical contexts to which it has been applied.[16] Random forest learning does not require strict assumptions about data, and it can be less sensitive to outliers than classical methods. Random forest learning can be robust to missing data, for which it calculates proximities based on the data present.[28,29] Random forest models can handle the large numbers of input variables common in clinical situations, and they are less prone to overfitting bias, which can result when the number of input variables is large relative to the sample size. During random forest learning, training data sets can be generated not only by random sampling but also by randomly selecting different input variables. This helps reduce cross-correlation among the resulting decision trees, making the final voting classification less biased.

CONCLUSION

This study lays the foundation for applying random forest learning to prediction of prostate cancer. Random forest machine-learning

algorithm can generate a model that combines transrectal ultrasound findings, age, and serum PSA levels to predict prostate cancer with good diagnostic performance compared to the gold standard of prostate biopsy pathology. This model may be a more accurate and reliable alternative to ultrasound or PSA on their own for deciding whether invasive biopsy is necessary. The random forest approach may prove useful for other clinical problems as well.

There are a few limitations in our study. First, only lymph nodes in regions around the prostate were examined, rather than all major lymph nodes. Second, interpreting ultrasound images can be subjective, which can limit the generalizability and reliability of our results. Third, we did not take into account all factors that may be useful for prostate cancer diagnosis, such as family history of prostate cancer, digital rectal exam results, and Gleason score. Future studies should examine whether adding these factors can improve our diagnostic model. It may also be possible to improve our model by classifying transrectal ultrasound images based on ordered category predictors and by incorporating more advanced ultrasound techniques.

AUTHOR CONTRIBUTIONS

LHX contributed to study conception and design, data analysis and interpretation, and manuscript drafting. PRC performed statistical analysis. ZPG participated in study design and data interpretation. YZL participated in study design and ultrasound evaluation. ML and LCX acquired and managed data. PF conceived the study, participated in its design and coordination, and critically revised important intellectual content in the manuscript. All authors have read and approved the final version of the manuscript.

COMPETING INTERESTS

All authors declared no competing interests.

ACKNOWLEDGMENTS

We would like to thank the research staff for their helpful advice on data collection and analysis and peer reviewers for their valuable comments.

REFERENCES

1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA *Cancer J Clin* 2015; 65: 5–29.
2 Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, *et al*. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer* 2013; 49: 1374–403.
3 Sujun H, Siwei Z, Wanqing C, Changling L. Analysis of the status and trends of prostate cancer incidence in China. *Chin Clin Oncol* 2013; 18: 330–4.
4 Simmons MN, Berglund RK, Jones JS. A practical guide to prostate cancer diagnosis and management. *Cleve Clin J Med* 2011; 78: 321–31.
5 Abd TT, Goodman M, Hall J, Ritenour CW, Petros JA, *et al*. Comparison of 12-core versus 8-core prostate biopsy: multivariate analysis of large series of US veterans. *Urology* 2011; 77: 541–7.
6 Delongchamps NB, de la Roza G, Jones R, Jumbelic M, Haas GP. Saturation biopsies on autopsied prostates for detecting and characterizing prostate cancer. *BJU Int* 2009; 103: 49–54.
7 Aigner F, Mitterberger M, Rehder P, Pallwein L, Junker D, *et al*. Status of transrectal ultrasound imaging of the prostate. *J Endourol* 2010; 24: 685–91.
8 Wolf AM, Wender RC, Etzioni RB, Thompson IM, D'Amico AV, *et al*. American Cancer Society guideline for the early detection of prostate cancer: update 2010. *CA Cancer J Clin* 2010; 60: 70–98.
9 Hayes JH, Barry MJ. Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence. *JAMA* 2014; 311: 1143–9.
10 Pallwein L, Mitterberger M, Pelzer A, Bartsch G, Strasser H, *et al*. Ultrasound of prostate cancer: recent advances. *Eur Radiol* 2008; 18: 707–15.
11 Sciarra A, Cattarino S, Gentilucci A, Salciccia S, Alfarone A, *et al*. Update on screening in prostate cancer based on recent clinical trials. *Rev Recent Clin Trials* 2011; 6: 7–15.
12 Gann PH, Hennekens CH, Stampfer MJ. A prospective evaluation of plasma prostate-specific antigen for detection of prostatic cancer. *JAMA* 1995; 273: 289–94.
13 Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, *et al*. Prevalence of prostate cancer among men with a prostate-specific antigen level ≤4.0 ng per milliliter. *N Engl J Med* 2004; 350: 2239–46.
14 Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
15 Bigler SA, Deering RE, Brawer MK. Comparison of microscopic vascularity in benign and malignant prostate tissue. *Hum Pathol* 1993; 24: 220–6.
16 Biau G. Analysis of a random forests model. *J Mach Learn Res* 2012; 13: 1063–95.
17 Malley JD, Malley KG, Pajevic S. Statistical Learning for Biomedical Data. Cambridge: Cambridge University Press; 2011. p. 137–54.
18 Breiman L. Out-of-Bag Estimation. Available from: http://www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.3712. [Last accessed on 2015 Sep 19].
19 Brock M, von Bodman C, Palisaar RJ, Löppenberg B, Sommerer F, *et al*. The impact of real-time elastography guiding a systematic prostate biopsy to improve cancer detection rate: a prospective study of 353 patients. *J Urol* 2012; 187: 2039–43.
20 Kökeny GP, Cerri GG, de Oliveira Cerri LM, de Barros N. Correlations among prostatic biopsy results, transrectal ultrasound findings and PSA levels in diagnosing prostate adenocarcinoma. *Eur J Ultrasound* 2000; 12: 103–13.
21 Trabulsi EJ, Sackett D, Gomella LG, Halpern EJ. Enhanced transrectal ultrasound modalities in the diagnosis of prostate cancer. *Urology* 2010; 76: 1025–33.
22 Loch T. Urologic imaging for localized prostate cancer in 2007. *World J Urol* 2007; 25: 121–9.
23 Hwang SI, Lee HJ. The future perspectives in transrectal prostate ultrasound guided biopsy. *Prostate Int* 2014; 2: 153–60.
24 Flanigan RC, Catalona WJ, Richie JP, Ahmann FR, Hudson MA, *et al*. Accuracy of digital rectal examination and transrectal ultrasonography in localizing prostate cancer. *J Urol* 1994; 152 (5 Pt 1): 1506–9.
25 Zhao HX, Zhu Q, Wang ZC. Detection of prostate cancer with three-dimensional transrectal ultrasound: correlation with biopsy results. *Br J Radiol* 2012; 85: 714–9.
26 Grubb RL, Pinsky PF, Greenlee RT, Izmirlian G, Miller AB, *et al*. Prostate cancer screening in the prostate, lung, colorectal and ovarian cancer screening trial: update on findings from the initial four rounds of screening in a randomized trial. *BJU Int* 2008; 102: 1524–30.
27 Nam RK, Oliver TK, Vickers AJ, Thompson I, Kantoff PW, *et al*. Prostate-specific antigen test for prostate cancer screening: American Society of Clinical Oncology provisional clinical opinion. *J Oncol Pract* 2012; 8: 315–7.
28 Askland KD, Garnaat S, Sibrava NJ, Boisseau CL, Strong D, *et al*. Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy. *Int J Methods Psychiatr Res* 2015; 24: 156–69.
29 Stekhoven DJ, Bühlmann P. MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012; 28: 112–8.