

# Development and internal validation of a novel model and markers to identify the candidates for lymph node metastasis in patients with prostate cancer

Hai-ming Cao, MS<sup>a</sup>, Zi Wan, MD<sup>b</sup>, Yu Wu, MS<sup>a</sup>, Hong-yang Wang, PhD<sup>c</sup>, Chao Guan, BS<sup>a,\*</sup>

## Abstract

**Background** High-grade prostate cancer (PCa) has a poor prognosis, and up to 15% of patients worldwide experience lymph node invasion (LNI). To further improve the prediction lymph node invasion in prostate cancer, we adopted risk scores of the genes expression based on the nomogram in guidelines.

**Methods** We analyzed clinical data from 320 PCa patients from the Cancer Genome Atlas database. Weighted gene coexpression network analysis was used to identify the genes that were significantly associated with LNI in PCa (n=390). Analyses using the Gene Ontology and Kyoto Encyclopedia of Genes and Genomes databases were performed to identify the activated signaling pathways. Univariate and multivariate logistic regression analyses were performed to identify the independent risk factors for the presence of LNI.

**Results** We found that patients with actual LNI and predicted LNI had the worst survival outcomes. The 7 most significant genes (CTNNAL1, ENSA, MAP6D1, MBD4, PRCC, SF3B2, TREML1) were selected for further analysis. Pathways in the cell cycle, DNA replication, oocyte meiosis, and 9 other pathways were dramatically activated during LNI in PCa. Multivariate analyses identified that the risk score (odds ratio [OR]=1.05 for 1% increase, 95% confidence interval [CI]: 1.04–1.07,  $P < .001$ ), serum PSA level, clinical stage, primary biopsy Gleason grade (OR=2.52 for a grade increase, 95% CI: 1.27–5.22,  $P = .096$ ), and secondary biopsy Gleason grade were independent predictors of LNI. A nomogram built using these predictive variables showed good calibration and a net clinical benefit, with an area under the curve (AUC) value of 90.2%.

**Conclusions** In clinical practice, the application of our nomogram might contribute significantly to the selection of patients who are good candidates for surgery with extended pelvic lymph node dissection.

**Abbreviations:** AUC = area under the curve, DEG = differentially expressed genes, DFS = disease free survival, ePLND = extended pelvic lymph node dissection, GO = gene ontology, KEGG = Kyoto Encyclopedia of Genes and Genomes, LASSO = Least absolute shrinkage and selection operator, LNI = lymph node invasion, OS = overall survival, PRAD = Prostate adenocarcinoma, PCa = prostate cancer, PSA = prostate specific antigen, qRT-PCR = quantitative real time-PCR, RS = risk score, TCGA = The Cancer Genome Atlas, WGCNA = weighted gene coexpression network analysis.

**Keywords:** gene signature, lymph node invasion, nomogram, prostate cancer

Editor: Patricia Severino.

H-mC and ZW have contributed equally to this work.

This study was funded by the Natural Science Fund of Bengbu Medical College (BYKY17115).

The authors have no conflicts of interest to disclose.

<sup>a</sup> Department of Urology, The Second Affiliation Hospital, Bengbu Medical College, Bengbu, Anhui, <sup>b</sup> Department of Urology, The First Affiliation Hospital, Sun Yat-Sen University, Guangzhou, Guangdong, <sup>c</sup> Department of Urology, The First Affiliation Hospital, Qingdao University, Qingdao, Shandong, China.

\* Correspondence: Chao Guan, The Second Affiliated Hospital of Bengbu Medical College, Bengbu, Anhui, China (e-mail: guanchao139@foxmail.com)

Copyright © 2019 the Author(s). Published by Wolters Kluwer Health, Inc. This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Medicine (2019) 98:30(e16534)

Received: 4 March 2019 / Received in final form: 10 June 2019 / Accepted: 27 June 2019

<http://dx.doi.org/10.1097/MD.00000000000016534>

## 1. Introduction

High-grade prostate cancer (PCa) has a poor prognosis, and up to 15% of patients worldwide experience lymph node invasion (LNI).<sup>[1]</sup> Radical prostatectomy coupled with extended pelvic lymph node dissection (ePLND) remains the principal surgical procedure for these patients.<sup>[2]</sup> LNI is significant for the diagnosis of PCa, as it could not only predict the prognosis of patients but also play a decisive role in their surgical treatment. Therefore, predicting the occurrence of LNI is of vital clinical significance. At present, several nomograms can be used to predict the occurrence of LNI in PCa patients,<sup>[3–5]</sup> and they all show good predictive accuracy. However, these models are based on traditional biopsies or imaging-based diagnoses. Additionally, clinical imaging techniques have limited sensitivity for detecting LNI.<sup>[6,7]</sup> Hence, to overcome these drawbacks, a non-invasive and simple method to accurately predict LNI is urgently needed.

Published studies<sup>[8–10]</sup> have suggested that molecular biomarker analysis is a good method for predicting the prognostic outcome of PCa patients; it is also a promising and powerful method to predict the occurrence of LNI in PCa patients. Levels

of gene expression<sup>[11–13]</sup> have been shown to be associated with LNI in PCa patients. By utilizing microarray data, we were able to use the method of applied weighted gene coexpression network analysis (WGCNA) to identify the hub genes. Thus, using the identified hub genes, we investigated an optimal gene signature to predict the occurrence of LNI in PCa patients. Recently, some studies<sup>[14–17]</sup> have demonstrated that the combination of several genes had the ability to predict lymph node metastasis in malignant tumors. However, the predictive ability of gene signatures in PCa is still unknown.

The primary aim of this study was to evaluate the LNI-predictive value of the gene signature of PCa and to develop a gene-based risk score for predicting LNI. In the current study, we examined the predictive ability of the base module combined with the risk score for predicting LNI in PCa patients. We also found

that the risk score was significantly associated with poor prognosis in PCa patients.

## 2. Materials and methods

### 2.1. Patient population

The standardized level 3 RNA sequencing data of Prostate adenocarcinoma patients and the corresponding clinical records in The Cancer Genome Atlas (TCGA) were obtained from the FireBrowse (<http://firebrowse.org>). A total of 320 PCa patients' clinical and pathologic data were obtained from the TCGA database (without missing values of gene and clinical information). LNI in patients was confirmed by pathology. A total of 390 patients' genes data was complete. The RNAseq by Expectation-

**Table 1**  
The clinicopathological characteristics of LNI patients versus non-LNI patients.

	Overall n = 320 (100%)	pN0 n = 258 (80.6%)	pN1 n = 62 (19.4%)	P value
Age at diagnosis, yr				.86
Mean (median)	60.9 (61)	60.9 (61)	60.7 (61)	
IQR	56–66	56–66	56–65.5	
History of prior malignancy				.79
Yes	305 (95.3)	245 (95)	60 (96.8)	
No	15 (4.7)	13 (5)	2 (3.2)	
PSA level, ng/mL				<.001
Mean (median)	11.19 (7.5)	10.05 (7)	15.89 (10.07)	
IQR	5.2–11.25	5–10.03	7.1–17.85	
Clinical stage (%)				<.001
T1	129 (40.3)	112 (43.4)	17 (27.4)	
T2	149 (46.6)	117 (45.3)	32 (51.6)	
T3	42 (13.1)	29 (11.3)	13 (21)	
Primary Gleason grade (%)				<.001
≤3	108 (33.8)	105 (40.7)	3 (4.9)	
≥4	212 (66.2)	153 (59.3)	59 (95.1)	
Second Gleason grade (%)				<.001
≤3	90 (28.1)	82 (31.8)	8 (13)	
≥4	230 (71.9)	176 (68.2)	54 (87)	
No. of removed and examined lymph nodes				.07
Mean (median)	13 (11)	12.3 (10)	15.6 (14)	
IQR	6–17	5–17	10.5–19	
No. of positive lymph nodes in patients with LNI				NA
Mean (median)	2.3 (1)	NA	2.3 (1)	
IQR	1–2.5	NA	1–2.5	
Positive surgical margins (%)				<.001
R0	213 (66.6)	191 (74)	22 (35.5)	
R1	90 (28.1)	55 (21.3)	35 (56.5)	
R2	3 (0.9)	2 (0.8)	1 (1.6)	
NA	14 (4.4)	10 (3.9)	4 (6.4)	
Pathologic stage				<.001
pT2	113 (35.3)	111 (43)	2 (3.2)	
pT3a	109 (34.1)	94 (36.4)	15 (24.2)	
pT3b	90 (28.1)	48 (18.6)	42 (67.7)	
pT4	6 (1.88)	3 (1.2)	3 (4.9)	
NA	2 (0.62)	2 (0.8)	0 (0)	
Pathologic Gleason score				<.001
<7	17 (5.3)	17 (6.6)	0 (0)	
=7	158 (49.4)	145 (56.2)	13 (21)	
>7	145 (45.3)	96 (37.2)	49 (79)	
Overall survival time, d				.96
Mean (median)	761.3 (545.5)	760.4 (529)	764.9 (636.5)	
IQR	212.5–1022.5	202.8–1000.2	227.5–1246.5	
NA (no. of patients)	6	3	3	

IQR = interquartile range; LNI = lymph node invasion; NA = not available; PSA = prostate-specific antigen.

Maximization (RSEM) values were utilized to quantify the mRNA expression levels.

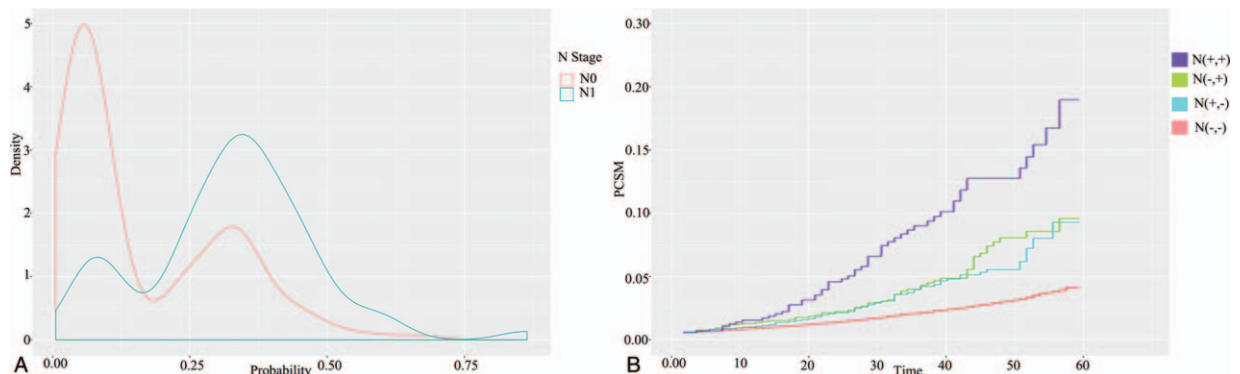
### 2.2. Weighted gene coexpression network analysis and functional enrichment analysis

A total of 390 patients' genes data was complete. We find significant genes by WGCNA. Subsequently, the KEGG signaling pathways were coordinated by R package components such as ClusterProfiler, Pathview (<http://www.bioconductor.org/>) and

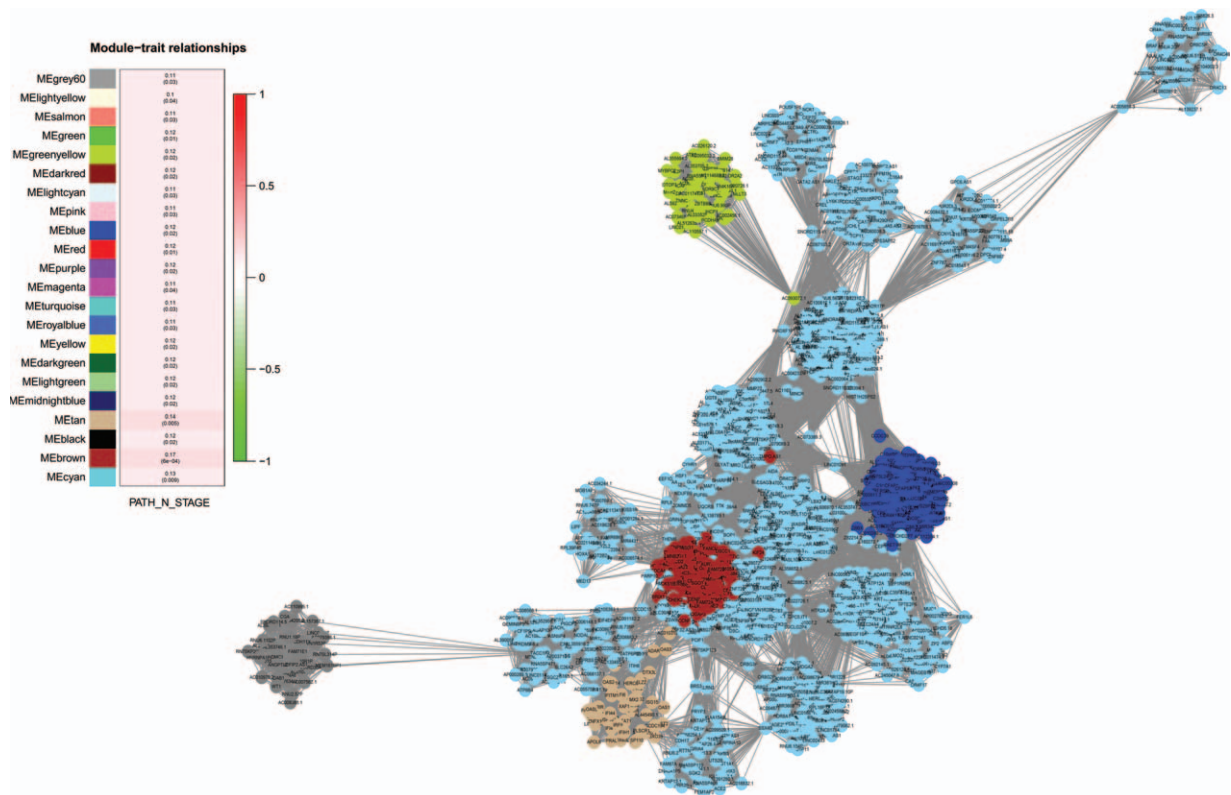
Stringi (<https://cran.r-project.org/>). Cytoscape software (<http://www.cytoscape.org/>) was then used to convert the enriched analysis data into visual images.

### 2.3. Statistical methods

Continuously coded variables were reported as the mean, median, and interquartile range (IQR) and analyzed by *t* test. Categorical variables were reported as frequencies and proportions and analyzed by chi-square test. We identified signatures



**Figure 1.** A. Pathological methods were used to predict the probability of LNI. N0: no lymph node invasion (LNI) confirmed by pathology; N1: lymph node invasion (LNI) confirmed by pathology; LNI: lymph node invasion. B. Comparison of 5-year patient survival between subgroups with different LNI statuses. N (+, +): actual LNI and predicted LNI; N (-, +): no actual LNI but predicted LNI; N (+, -): actual LNI but no predicted LNI; N (-, -): no actual LNI and no predicted LNI. LNI = lymph node invasion.



**Figure 2.** Weighted gene coexpression network analysis (WGCNA) of LNI prostate cancer. A. Correlation between 22 modules (2294 genes) and LNI. B. Visual representation of the WGCNA of the 22 modules (2294 genes). LNI = lymph node invasion.

using least absolute shrinkage and selection operator (LASSO) regression. Univariable and multivariable logistic regression models were used to predict the occurrence of LNI.

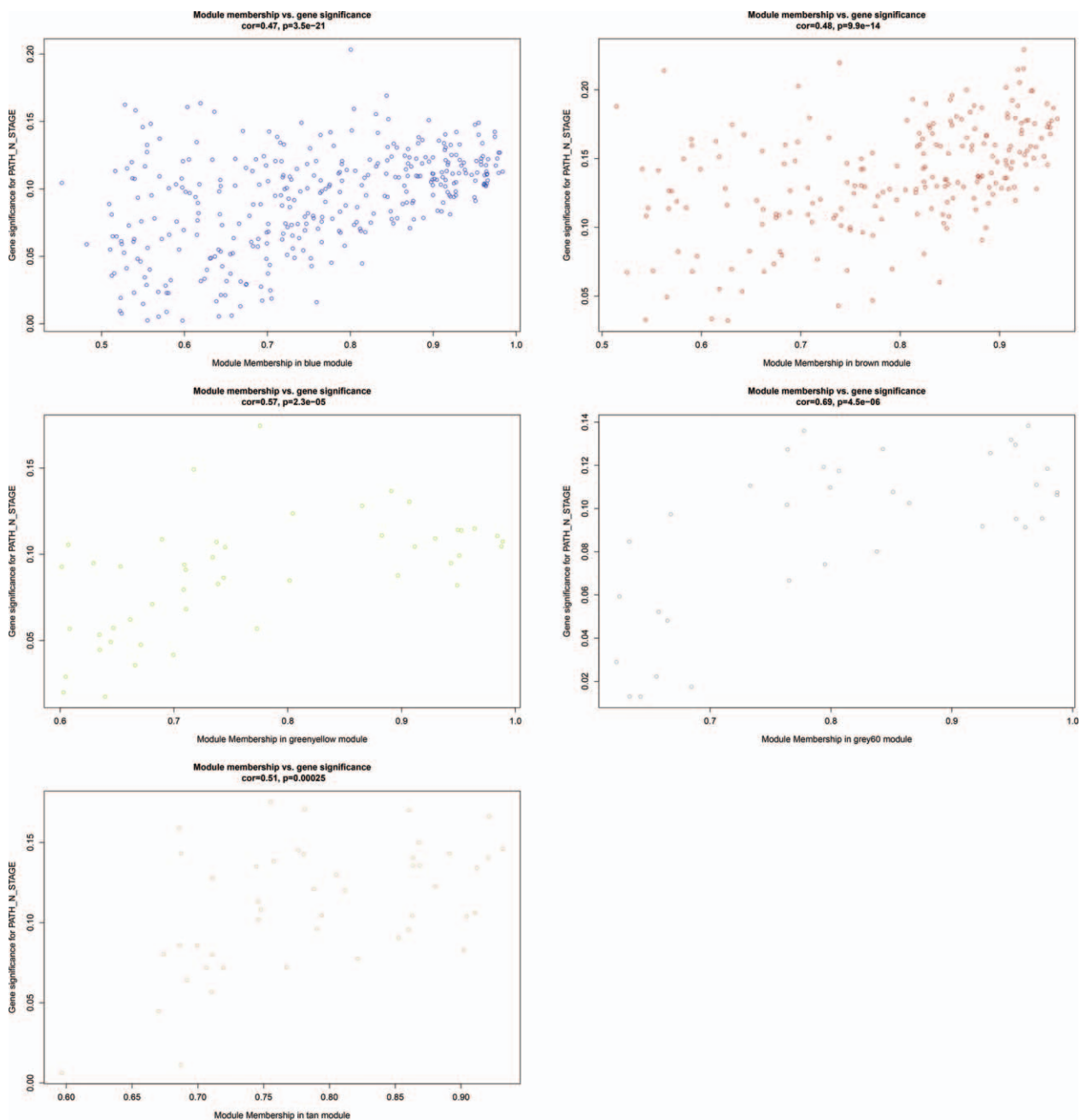
The discrimination accuracy of multivariable models based on these variables in our cohort was quantified by the value of the area under the curve (corrected-AUC was calculated using a 200-resample bootstrap). The extent of over- and underestimation of pathologically confirmed versus nomogram-predicted LNI was graphically explored using a calibration plot. To determine the clinical net benefit associated with the use of the nomogram, we conducted a decision-curve analysis (DCA). The calibration and DCA were corrected for overfit using 10-fold cross validation.

Statistical analyses in the study were performed using the R statistical package v.3.3.2 (R Project for Statistical Computing). All statistical tests were 2-sided, and  $P < .05$  was considered to be statistically significant.

### 3. Results

#### 3.1. Clinicopathological characteristics of PCa patients

First, we analyzed clinical data from 320 PCa patients obtained from TCGA database. The clinicopathological characteristics of LNI patients and non-LNI patients are compared in Table 1. We found that the serum level of prostate specific antigen (PSA) was



**Figure 3.** Gene significance for pathological lymph node metastasis in the 5 significant modules (blue, brown, greenyellow, grey60, tan) according to correlation and  $P$  value.

much higher in LNI patients than in non-LNI patients (15.89 vs 10.05 ng/mL,  $P < .001$ ). In addition, LNI patients had higher tumor grade (including clinical stage, primary and secondary biopsy Gleason grade, pathologic stage, and pathologic Gleason score) than non-LNI patients (all  $P < .001$ ).

### 3.2. Lymph node invasion is associated with the inferior outcome in PCa patients

Traditionally, LNI prediction models were commonly used to confirm the presence of LNI. However, LNI prediction models could not absolutely distinguish between LNI and non-LNI (Fig. 1A). Thus, it was very necessary to develop a better model to predict the presence of LNI.

Patients with actual LNI and predicted LNI had the worst survival outcomes, and patients with actual non-LNI or predicted non-LNI had the best survival outcomes. However, patients with actual non-LNI and predicted LNI or actual LNI and predicted non-LNI had significantly lower 5-year survival rates than patients with actual non-LNI and predicted non-LNI (Fig. 1B).

### 3.3. Weighted gene coexpression network analysis

WGCNA was built from 53,324 genes that were identified as being associated with LNI in 390 PCa, and 89 modules were

identified. Figure 2 shows the visual representation of WGCNA of 22 modules. Then, we selected the 5 most significant modules based on correlation with the LNI and  $P$  values from the 22 modules for further analysis (Fig. 3).

### 3.4. Gene ontology and pathway analysis

A total of 2294 genes and 22 modules associated with LNI were identified in PCa. The Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses are presented in Fig. 4. KEGG analysis revealed that pathways involved in the cell cycle, DNA replication, Fanconi anemia, oocyte meiosis, progesterone-mediated oocyte maturation, base excision repair, p53 signaling, cellular senescence, measles, alcoholism, homologous recombination, and mismatch repair were dramatically activated in PCa with LNI (Fig. 4D).

### 3.5. Identification of the 7-gene signature and its association with the survival of PCa patients

We used 7 genes (CTNNAL1, ENSA, MAP6D1, MBD4, PRCC, SF3B2, TREML1) to generate a signature using LASSO regression from 904 genes in the 5 most significant modules. The risk score was calculated for each of the 390 patients from TCGA, and patients in every grade were then successfully divided

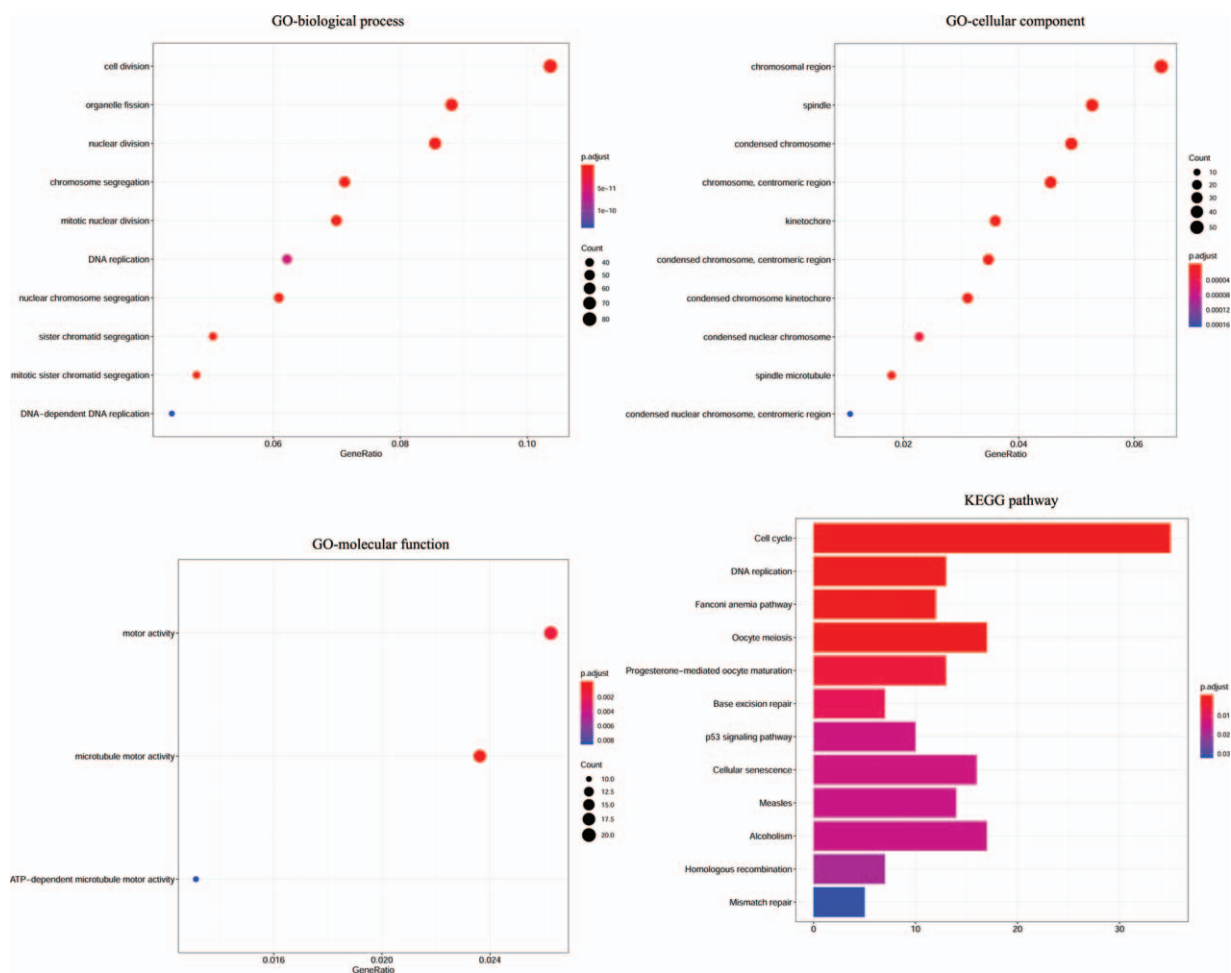
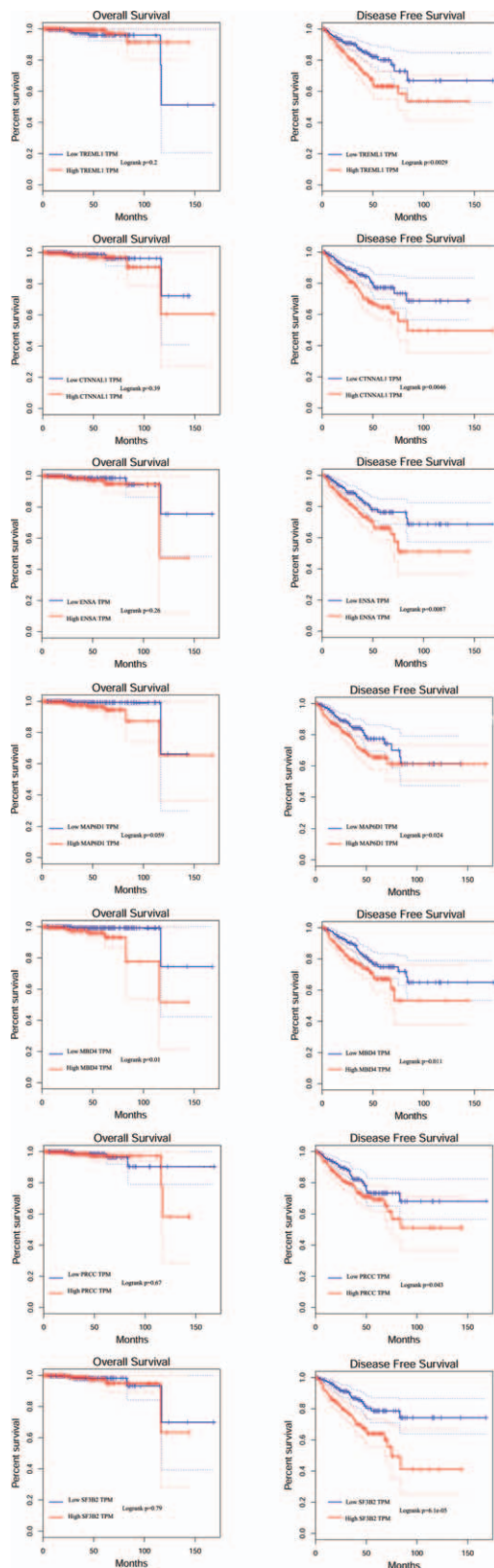


Figure 4. GO analysis and ingenuity pathway analysis of the 22 modules (2294 genes). GO= gene ontology.





**Figure 5.** Prognostic value of the expression of 7 genes in PCa patients. A. Analysis of overall survival (OS) and disease-free survival (DFS) in 390 PCa patients from TCGA database by TREML1 expression. B. Analysis of OS and DFS in 390 PCa patients from TCGA database by CTNNA1 expression. C. Analysis of OS and DFS in 390 PCa patients from TCGA database by ENSA expression. D. Analysis of OS and DFS in 390 PCa patients from TCGA database by MAP6D1 expression. E. Analysis of OS and DFS in 390 PCa

into a high-gene expression group and a low-gene expression group based on a cutoff value (the median risk score). PCa patients in the high-gene expression group had significantly shorter disease-free survival (DFS) and overall survival (OS) than those of the low-gene expression group (all  $P < .05$ , Fig. 5A–G).

Additionally, we found that the 7 genes were significantly differentially overexpressed in LNI PCa compared with non-LNI PCa (all  $P < .001$ , Fig. 6).

### 3.6. Risk factors for lymph node invasion

Univariate and multivariate logistic regression analyses were performed to identify independent risk factors for the presence of LNI (Table 2). In the univariate analysis, the variable of risk score was the most accurate predictor (corrected AUC = 86.7%), followed by clinical stage (74.3%), primary biopsy Gleason grade (73.2%), serum PSA level (68.1%), and secondary biopsy Gleason grade (64.1%). In the multivariate analysis, we found that the risk score (OR = 1.05 for 1% increase, 95% CI: 1.04–1.07,  $P < .001$ ), serum PSA level, clinical stage, primary biopsy Gleason grade (odds ratio [OR] = 2.52 for a grade increase, 95% confidence interval [CI]: 1.27–5.22,  $P = .096$ ) and secondary biopsy Gleason grade were independent predictors of LNI. Based on the 5 predictors, we developed a full model for predicting LNI, and the corrected AUC value was 90.2%. Interestingly, when the variable of risk score was removed from the full model, namely, the base model that identified the 4 predictors (serum PSA level, clinical stage, primary Gleason grade, and secondary Gleason grade), the corrected AUC value dropped to 83.7%.

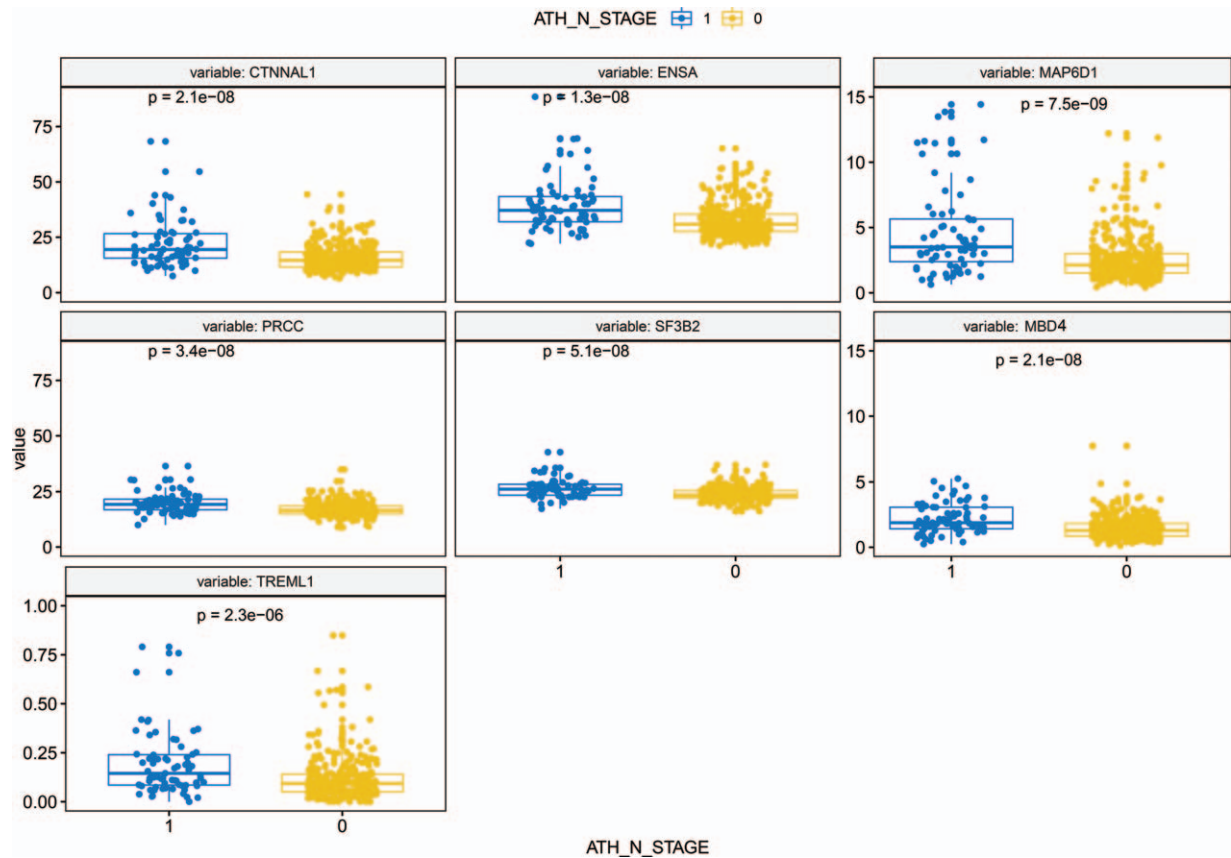
Then, a nomogram was developed (Fig. 7). The nomogram displayed the multivariable analysis effect of predictors on the risk of LNI. The calibration plot of the predicted probabilities against the observed LNI rates indicated good concordance (Fig. 8A). Additionally, the decision curve analysis demonstrated that the full model had the highest clinical net benefit across the entire range of threshold probabilities (Fig. 8B).

## 4. Discussion

Although ePLND represents the gold standard of treatment in LNI PCa,<sup>[2]</sup> given the increased and serious complications related to this procedure,<sup>[18]</sup> an ePLND should be considered only in men with very high risk of LNI. Therefore, accurate identification of these high-risk patients could greatly help to avoid unnecessary ePLND treatment. In the present study, on the basis of traditional predictive variables, we incorporated the risk scores of the gene signature and developed a novel nomogram to predict LNI in PCa patients. We showed that the nomogram has high accuracy in detecting LNI (AUC value = 90.2%); in addition, its calibration also has good concordance between predicted and observed LNI probabilities. The decision curve analysis demonstrated that the full model had the highest clinical net benefit across the entire range of threshold probabilities.

Currently, there are several nomograms<sup>[19–21]</sup> that predict the occurrence of LNI in PCa patients. The accuracy of the first nomogram was not high (AUC = 76%), and the latest nomogram

patients from TCGA database by MBD4 expression. F. Analysis of OS and DFS in 390 PCa patients from TCGA database by PRCC expression. G. Analysis of OS and DFS in 390 PCa patients from TCGA database by SF3B2 expression. PCa = prostate cancer, TCGA = The Cancer Genome Atlas.



**Figure 6.** Upregulation of the expression of 7 genes in LNI PC (the blue in figure) compared with non-LNI PCa (the yellow in figure). The single spot represents the relative expression value of the gene from an individual patient. The thick line in the middle represents the median value. LNI=lymph node invasion.

had higher accuracy, but it was based on biopsy data. The previous prediction nomogram was based on detailed biopsy data obtained at the central pathologic review.<sup>[19]</sup> This highlights the need for simple and efficient methods that can detect LNI in PCa patients with increased accuracy. In our study, the nomogram was built on predictive variables, including a 7 gene signature-based risk score, serum PSA level, clinical stage, primary biopsy Gleason grade, and secondary biopsy Gleason grade. We showed that combining existing clinical variables with our newly developed gene signature-based risk scores could enhance the detection accuracy of LNI in PCa patients. We note

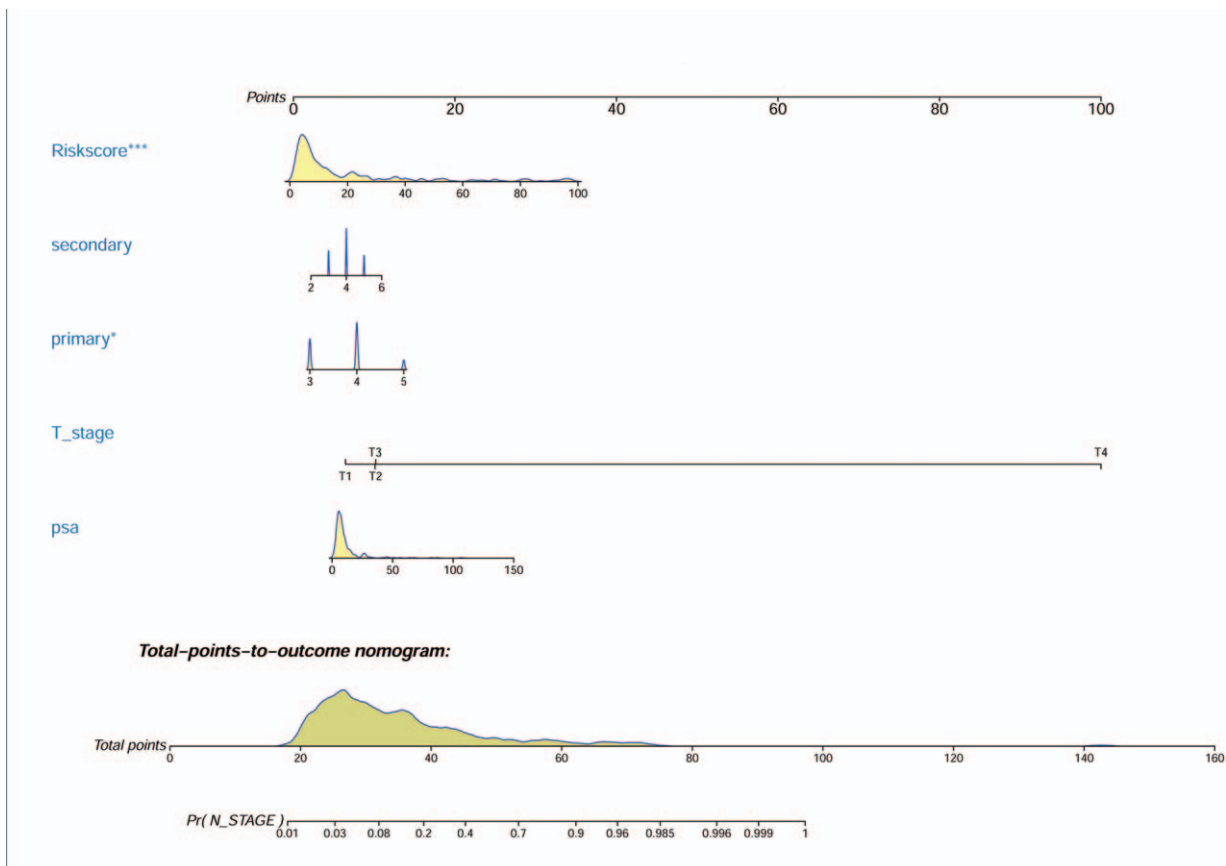
that these predictive variables were all convenient parameters; they do not depend on preoperative biopsies or image-based examination. Hence, our nomogram is a non-invasive and simple method for predicting LNI.

Prior studies have noted the importance of gene signatures in the prognosis of PCa. Jin et al<sup>[22]</sup> demonstrated that an NF-κB-activated recurrence predictor 21 gene signature could predict disease-specific survival and distant metastasis-free survival in patients with PCa, although the study used molecular biological methods. Recently, another study<sup>[23]</sup> identified a 24-gene signature that was significantly associated with the development

**Table 2**  
Univariable and multivariable logistic regression analysis predicting the presence of lymph node invasion.

Covariates	Univariable analysis			Multivariable analysis				
	Odds ratio (95% CI)	P value	AUC (%)	Base model Odds ratio (95% CI)	P value	Full model Odds ratio (95% CI)	P value	
PSA level, ng/mL	1.03 (1.01–1.06)	.002	68.1	1.02 (0.99–1.05)	.089	1.02 (0.99–1.05)	.15	
Clinical stage	–	<.001	74.3	–	.009	–	.006	
Primary Gleason grade	4.8 (2.84–8.5)	<.001	73.2	2.17 (2.9–9.77)	<.001	2.52 (1.27–5.22)	.096	
Second Gleason grade	2.23 (1.44–3.52)	<.001	64.1	2.2 (1.41–3.51)	<.001	1.41 (0.82–2.45)	.21	
Risk score, %	1.07 (1.05–1.09)	<.001	86.7	–	–	1.05 (1.04–1.07)	<.001	
AUC of multivariable models, %	–	–	–	83.7	90.2	–	–	

AUC=area under the curve; CI=confidence interval; PSA=prostate specific antigen.



**Figure 7.** A nomogram predicting the probability of LNI in PCa patients based on the risk score, secondary, primary, T stage, and serum PSA level. Instructions: Locate the risk score on the risk score axis. Draw a line straight upward to the point axis to determine how many points towards the probability of LNI that the patient receives for his risk score value. Repeat the process for each additional variable. Sum the points for each of the predictors. Locate the final sum on the total point axis. Draw a line straight down to find the patient's probability of LNI. LNI=lymph node invasion, PCa=prostate cancer; PSA=prostate specific antigen.

of metastasis and prostate cancer-specific mortality after radical prostatectomy. However, the importance of gene signatures in predicting LNI in PCa is not fully understood. In the present study, the predictive accuracy of the 7 genes to distinguish tissues with LNI and those without LNI was measured by ROC curve analysis. In the Cox analysis, gene signature-based risk scores were also the single most powerful predictor of LNI. The 7-gene signature could be of particular use in situations when predictions of the occurrence of LNI are ambiguous or borderline. Additionally, the 7-gene signature in our analysis could promote the initiation of additional therapies to treat LNI and allow for personalized treatment for patients.

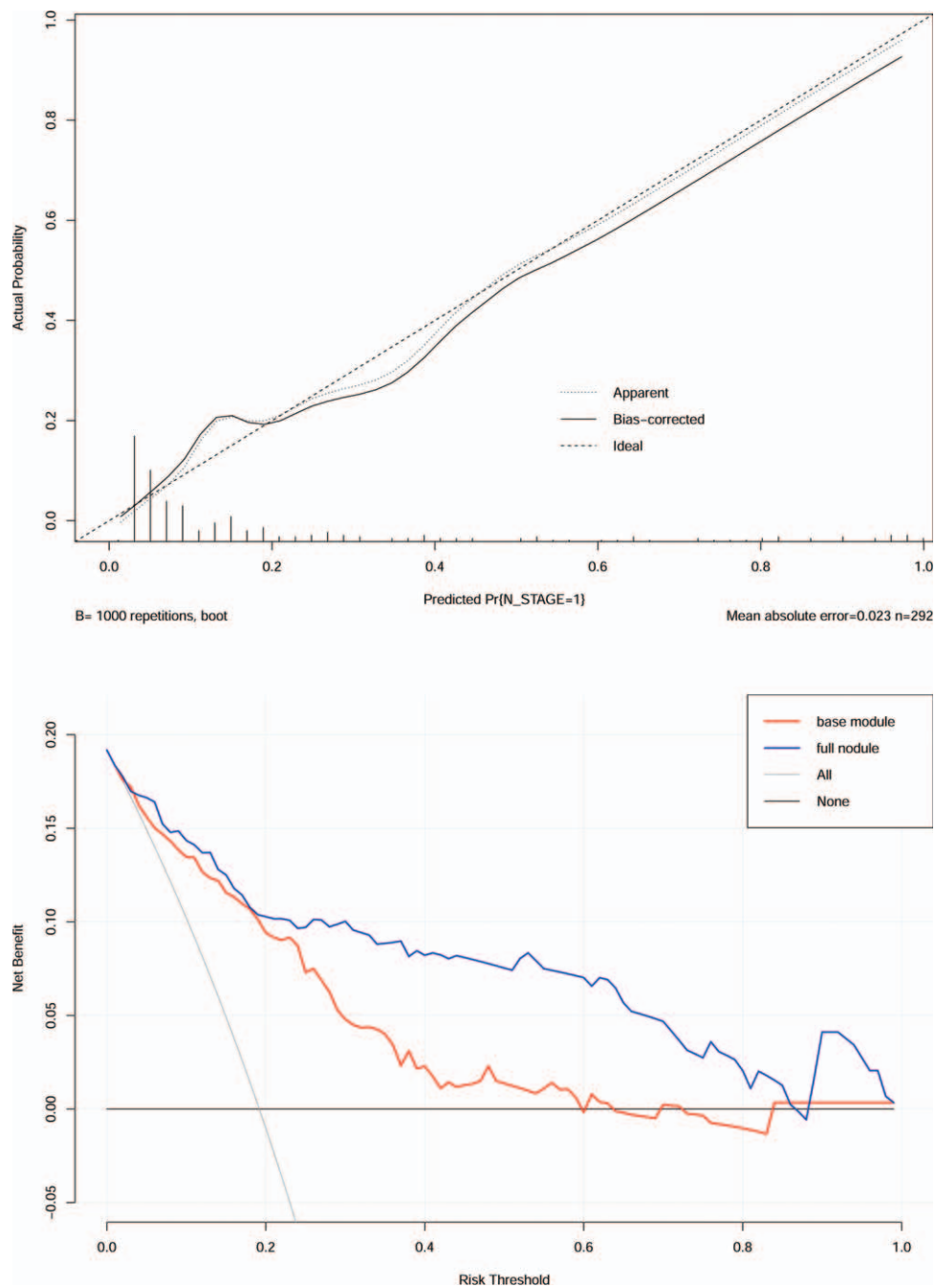
In our study, by using bioinformatic analyses, we found that the 7 gene signature-based risk score was an important independent predictor of LNI. Clearly, we need to further<sup>[24]</sup> understand how the 7 key genes affect the development of LNI. A previous study reported that CTNNAL1 was associated with pelvic lymph node metastasis in early-stage cervical cancer; in addition, CTNNAL1 can downregulate E-cadherin and promote melanoma progression and invasion.<sup>[25]</sup> It has been demonstrated the MBD4 gene was associated with PCa progression, and MBD4 was upregulated in metastatic PCa samples when compared with the expression in primary tumors.<sup>[26]</sup> SF3B2, as one of the genes of the spliceosome pathway, was overexpressed in hepatocellular carcinoma,<sup>[27]</sup> but its specific role in

PCa remains unknown. It has been reported that MAP6D1 was overexpressed in late stage clear cell renal cell carcinoma.<sup>[28]</sup> Based on the evidence described above, these genes appear to play important roles in the progression or metastasis of malignant tumors.

By applying novel analysis methods, we developed a new nomogram for predicting the occurrence of LNI in PCa patients. However, this study has several limitations that should be noted. First, our study had a small sample size, which was obtained from TCGA. Our study included 320 PCa patients, while previous nomograms included >500 patients each.<sup>[3,4]</sup> Additionally, although our nomogram has good concordance between predicted and observed LNI probabilities, it was not validated by an external validation cohort from another hospital. Further research should be undertaken to verify the predictive accuracy of our nomogram. Moreover, we identified the top 7 hub genes (TREM1, CTNNAL1, ENSA, MAP6D1, MBD4, PRCC, and SF3B2), which were closely related to LNI in PCa. However, their expression was not validated in PCa tissue samples by quantitative real-time PCR (qRT-PCR).

In summary, we have established that the risk score of the 7-gene signature was associated with a high risk of LNI in PCa patients. The present model improves the ability to identify patients at a high risk of LNI, and it could provide a practical guide for clinicians to more accurately identify patients who





**Figure 8.** Calibration plot for the nomogram and decision curve analysis of the logistic model. A. A calibration plot for the nomogram. The dotted line indicates the location of the ideal nomogram, in which the predicted and actual probabilities are identical. The broken line indicates the actual nomogram performance. The expected performance on future data is represented by the solid line. B. A decision curve analysis of the logistic model. It demonstrated that the net benefit associated with the use of the nomogram with (full model) or without (base model) the inclusion of the risk score as covariate. The net benefit can be interpreted as the proportion of all patients who have LNI and are recommended for extended pelvic lymph node dissection if no patients with negative lymph nodes were treated. LNI=lymph node invasion.

require surgery with ePLNDs. In clinical practice, the application of our nomogram might contribute significantly to the selection of patient candidates for surgery with ePLND.

### Acknowledgments

The authors would like to acknowledge the help provided by Zebin Zhu (The First Affiliated Hospital, University of Science

and Technology of China), Jianming Zeng (CEO of biotraine.com), and Qiannan Yang (his wife of HMC).

### Author contributions

**Conceptualization:** Zi Wan, Chao Guan.

**Data curation:** Haiming Cao, Zi Wan, Hongyang Wang.

**Formal analysis:** Haiming Cao, Zi Wan.

**Funding acquisition:** Haiming Cao.

**Investigation:** Yu Wu.

**Methodology:** Haiming Cao, Yu Wu.

**Project administration:** Chao Guan.

**Software:** Haiming Cao.

**Supervision:** Hongyang Wang.

**Validation:** Haiming Cao.

**Visualization:** Haiming Cao.

**Writing – original draft:** Haiming Cao.

**Writing – review & editing:** Zi Wan, Chao Guan.

## References

- [1] Wilczak W, Wittmer C, Clauditz T, et al. Marked prognostic impact of minimal lymphatic tumor spread in prostate cancer. *Eur Urol* 2018; 74:376–86.
- [2] Mottet N, Bellmunt J, Bolla M, et al. EAU-ESTRO-SIOG guidelines on prostate cancer. Part 1: Screening, diagnosis, and local treatment with curative intent. *Eur Urol* 2017;71:618–29.
- [3] Briganti A, Larcher A, Abdollah F, et al. Updated nomogram predicting lymph node invasion in patients with prostate cancer undergoing extended pelvic lymph node dissection: the essential importance of percentage of positive cores. *Eur Urol* 2012;61:480–7.
- [4] Gandaglia G, Fossati N, Zaffuto E, et al. Development and internal validation of a novel model to identify the candidates for extended pelvic lymph node dissection in prostate cancer. *Eur Urol* 2017;72:632–40.
- [5] Gandaglia G, Ploussard G, Valerio M, et al. A novel nomogram to identify candidates for extended pelvic lymph node dissection among patients with clinically localized prostate cancer diagnosed with magnetic resonance imaging-targeted and systematic biopsies. *Eur Urol* 2018; 75:506–14.
- [6] Hovels AM, Heesakkers RA, Adang EM, et al. The diagnostic accuracy of CT and MRI in the staging of pelvic lymph nodes in patients with prostate cancer: a meta-analysis. *Clin Radiol* 2008;63:387–95.
- [7] von Eyben FE, Kairemo K. Meta-analysis of (11)C-choline and (18)F-choline PET/CT for management of patients with prostate cancer. *Nucl Med Commun* 2014;35:221–30.
- [8] Benzon B, Zhao SG, Haffner MC, et al. Correlation of B7-H3 with androgen receptor, immune pathways and poor outcome in prostate cancer: an expression-based analysis. *Prostate Cancer Prostatic Dis* 2017;20:28–35.
- [9] Cooperberg MR, Erho N, Chan JM, et al. The diverse genomic landscape of clinically low-risk prostate cancer. *Eur Urol* 2018;74:444–52.
- [10] Walker SM, Knight LA, McCavigan AM, et al. Molecular subgroup of primary prostate cancer presenting with metastatic biology. *Eur Urol* 2017;72:509–18.
- [11] Chen C, Cai Q, He W, et al. AP4 modulated by the PI3K/AKT pathway promotes prostate cancer proliferation and metastasis of prostate cancer via upregulating L-plastin. *Cell Death Dis* 2017;8:e3060.
- [12] Lu X, Pan X, Wu CJ, et al. An in vivo screen identifies PYGO2 as a driver for metastatic prostate cancer. *Cancer Res* 2018;78:3823–33.
- [13] Oh JJ, Park S, Lee SE, et al. A clinicogenetic model to predict lymph node invasion by use of genome-based biomarkers from exome arrays in prostate cancer patients. *Korean J Urol* 2015; 56:109–16.
- [14] Cai W, Li Y, Huang B, et al. Esophageal cancer lymph node metastasis-associated gene signature optimizes overall survival prediction of esophageal cancer. *J Cell Biochem* 2019;120:592–600.
- [15] Chen X, Wang YW, Zhu WJ, et al. A 4-microRNA signature predicts lymph node metastasis and prognosis in breast cancer. *Hum Pathol* 2018;76:122–32.
- [16] Kang S, Thompson Z, McClung EC, et al. Gene expression signature-based prediction of lymph node metastasis in patients with endometrioid endometrial cancer. *Int J Gynecol Cancer* 2018;28:260–6.
- [17] Sonohara F, Gao F, Iwata N, et al. Genome-wide discovery of a novel gene-expression signature for the identification of lymph node metastasis in esophageal squamous cell carcinoma. *Ann Surg* 2017; 269:879–86.
- [18] Ploussard G, Briganti A, de la Taille A, et al. Pelvic lymph node dissection during robot-assisted radical prostatectomy: efficacy, limitations, and complications—a systematic review of the literature. *Eur Urol* 2014;65: 7–16.
- [19] Briganti A, Chun FK, Salonia A, et al. Validation of a nomogram predicting the probability of lymph node invasion among patients undergoing radical prostatectomy and an extended pelvic lymphadenectomy. *Eur Urol* 2006;49:1019–26. discussion 1026-1027.
- [20] Briganti A, Karakiewicz PI, Chun FK, et al. Percentage of positive biopsy cores can improve the ability to predict lymph node invasion in patients undergoing radical prostatectomy and extended pelvic lymph node dissection. *Eur Urol* 2007;51:1573–81.
- [21] Godoy G, Chong KT, Cronin A, et al. Extent of pelvic lymph node dissection and the impact of standard template dissection on nomogram prediction of lymph node involvement. *Eur Urol* 2011; 60:195–201.
- [22] Jin R, Yi Y, Yull FE, et al. NF-kappaB gene signature predicts prostate cancer progression. *Cancer Res* 2014;74:2763–72.
- [23] Pellegrini KL, Sanda MG, Patil D, et al. Evaluation of a 24-gene signature for prognosis of metastatic events and prostate cancer-specific mortality. *BJU Int* 2017;119:961–7.
- [24] Noordhuis MG, Fehrmann RS, Wisman GB, et al. Involvement of the TGF-beta and beta-catenin pathways in pelvic lymph node metastasis in early-stage cervical cancer. *Clin Cancer Res* 2011;17:1317–30.
- [25] Kreiseder B, Orel L, Bujnow C, et al. alpha-Catulin downregulates E-cadherin and promotes melanoma progression and invasion. *Int J Cancer* 2013;132:521–30.
- [26] Bianco-Miotto T, Chiam K, Buchanan G, et al. Global levels of specific histone modifications and an epigenetic gene signature predict prostate cancer progression and development. *Cancer Epidemiol Biomarkers Prev* 2010;19:2611–22.
- [27] Xu W, Huang H, Yu L, et al. Meta-analysis of gene expression profiles indicates genes in spliceosome pathway are up-regulated in hepatocellular carcinoma (HCC). *Med Oncol* 2015;32:96.
- [28] Bhalla S, Chaudhary K, Kumar R, et al. Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Sci Rep* 2017;7:44997.