

## ORIGINAL RESEARCH ARTICLE

# Prognostic factors and machine learning models for metastatic prostate cancer survival: Insights from the SEER database

Xian-Yong Yan<sup>1†</sup>, Jing-Zhi Huang<sup>1†</sup>, Qing-Fan Wei<sup>1</sup>, Rong-Yuan Wei<sup>1</sup>,  
Jiang Qin<sup>1</sup>, Hai Lu<sup>1\*</sup>, and Jie Lan<sup>2\*</sup>

<sup>1</sup>Department of Urology, The First Affiliated Hospital of Guangxi Medical University Hechi Hospital, Hechi, Guangxi Zhuang Autonomous Region, China

<sup>2</sup>Department of Surgery, Hechi Maternal and Child Health Hospital, Hechi, Guangxi Zhuang Autonomous Region, China

## Abstract

Metastatic prostate cancer remains a significant health concern, necessitating the use of accurate prognostic tools to predict patient outcomes. This study aims to develop a novel survival prediction model for patients with metastatic prostate cancer. Data for patients diagnosed with metastatic prostate cancer between 2010 and 2020 were sourced from the Surveillance, Epidemiology, and End Results database. Independent prognostic features were identified using the least absolute shrinkage and selection operator (LASSO)-Cox regression. Survival models, including XGBoost, RandomForestSRC, CoxBoost, SuperPC, and the DeepSurv model, were constructed, followed by an assessment of model predictive accuracy for 1-, 3-, and 5-year overall survival (OS) using the area under the receiver operating characteristic curve (AUC). The optimal model was selected based on the highest AUC, followed by feature importance evaluation and Kaplan–Meier survival analysis. A total of 10,408 eligible patients were enrolled, with 15 prognostic features identified by LASSO-Cox regression. The DeepSurv model showed the best overall performance, with a concordance index of 0.6488 and AUC values of 0.696, 0.685, and 0.720 for 1-, 3-, and 5-year OS, respectively. Gleason score and age were confirmed as core prognostic features across multiple algorithms. Survival analysis showed high-risk patients (risk score > median) had significantly shorter OS times than low-risk patients (risk score ≤ median;  $p < 0.001$ ). The DeepSurv model achieves high OS prediction accuracy for metastatic prostate cancer, and Gleason score and age are key core prognostic factors for this population. This model enables reliable individualized risk stratification, which can enhance clinical decision-making and improve patient outcomes.

**Keywords:** Metastatic prostate cancer; Machine learning; Deep learning; Survival model; SEER database

<sup>†</sup>These authors contributed equally to this work.

### \*Corresponding authors:

Hai Lu  
(lh89708@126.com);  
Jie Lan  
(JieLan9938@163.com)

**Citation:** Yan X-Y, Huang J-Z, Wei Q-F, *et al.* Prognostic factors and machine learning models for metastatic prostate cancer survival: Insights from the SEER database. *Cancer Plus*. 2026;8(3):025510090. doi: 10.36922/CP025510090

**Received:** December 18, 2025

**Revised:** January 24, 2026

**Accepted:** March 25, 2026

**Published online:** May 8, 2026

**Copyright:** © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

**Publisher's Note:** AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## 1. Introduction

Prostate cancer is the second most commonly diagnosed cancer and the fifth leading cause of cancer death among men worldwide.<sup>1</sup> Approximately 10 million men are currently living with prostate cancer, with approximately 700,000 cases classified as

metastatic. This condition results in more than 400,000 deaths each year, a number anticipated to more than double by 2040.<sup>2</sup> Despite progress in diagnostic and treatment methods, approximately 30% of prostate cancer cases develop into metastatic disease, primarily bone metastasis.<sup>3</sup> The 5-year survival rate of patients with advanced metastatic prostate cancer is approximately 28%.<sup>4</sup> Currently, the standard medical approaches for metastatic disease include androgen deprivation therapy, androgen receptor signaling inhibition, and chemotherapy. However, most patients develop castration resistance, which is associated with poor prognosis.<sup>5</sup> Timely detection and precise prognostic assessment are essential for enhancing patients' overall survival (OS) outcomes and customizing treatment strategies.<sup>6</sup> Therefore, it is crucial to identify the key features of metastatic prostate cancer.

Prior research has established that multiple factors, such as age, tumor-node-metastasis (TNM) stage, Gleason score, prostate-specific antigen (PSA) level, and surgical and distant metastases, are closely linked to the survival outcomes of patients with prostate cancer.<sup>7</sup> Early detection allows for more effective treatment options and better management of the disease, leading to improved patient outcomes.<sup>8</sup> Integrating PSA level, grade, and stage allows for the development of robust prognostic models for prostate cancer, and refining group stratification systems can enhance model discrimination.<sup>9</sup> It is also important to note that the prognosis of prostate cancer can be significantly influenced by the patient's overall health status, including comorbidities and lifestyle factors.<sup>10,11</sup> Moreover, recent studies have highlighted the role of genetic factors in the development and progression of prostate cancer, suggesting that personalized medicine approaches may be beneficial.<sup>12</sup> For example, Gravis *et al.*<sup>13</sup> used alkaline phosphatase to establish a simple prognostic prediction model for non-castrate metastatic prostate cancer; however, the model showed inadequate performance with a concordance index (C-index) of 0.64. Therefore, there is a need to develop robust prognostic models to improve the clinical outcomes of metastatic prostate cancer. Gleason score and age are well-recognized as the most robust prognostic indicators for prostate cancer, and their association with increased death risk has been repeatedly validated in previous population-based studies. However, the value of integrating these classic indicators with other clinical and metastasis-related features into a machine learning-based prognostic model—for calculating individual risk scores and predicting subsequent outcomes in *de novo* metastatic prostate cancer—has not been fully explored. Therefore, it is crucial to construct a practical prognostic model for individualized risk assessment. The core objective of this study is to develop such a comprehensive model, which

can provide actionable clinical references for treatment decision-making, and this constitutes the key innovation of our work.

With the advent of machine learning, alternative approaches for predictor selection have become feasible. Several predictive models exhibiting exceptional performance have been developed for clinical practice using machine learning.<sup>11,14</sup> The Surveillance, Epidemiology, and End Results (SEER) database (<https://seer.cancer.gov/>) encompasses a diverse group of patients from various geographical locations, offering comprehensive clinicopathological data and follow-up information, which provides a wealth of medical cases for analysis.<sup>15,16</sup> This real-world database provides a large cohort suitable for machine-learning model development.<sup>17</sup> However, few studies have focused on constructing prognostic models for metastatic prostate cancer based on data from the SEER database.

It is worth noting that the SEER database has been widely used in cancer research due to its large sample size and rich clinical information, which can provide reliable data support for the development of prognostic models.<sup>15</sup> For example, a study by Lee *et al.*<sup>11</sup> applied a novel machine learning framework to predict non-metastatic prostate cancer-specific mortality using the SEER database, demonstrating the potential of machine learning in cancer prognosis.

Therefore, in this study, we integrated the least absolute shrinkage and selection operator (LASSO)-Cox regression with multiple machine learning and deep learning algorithms to establish an optimal survival prediction model for patients with metastatic prostate cancer. This model relies on clinical and histopathological features linked to the prognosis of metastatic prostate cancer in the SEER database. We further validated the predictive power of the survival model by stratifying the patients into risk categories and conducting a survival analysis. These findings may help inform clinical decision-making and improve patient care outcomes.

## 2. Methods

### 2.1. Data source

Data from patients diagnosed with prostate cancer between 2010 and 2020 were obtained from the SEER database using the SEER\*Stat software (version 8.4.3, National Institutes of Health, United States of America). The inclusion criteria were as follows: (i) site code of C61.9 according to the International Classification of Diseases for Oncology 3rd Edition; (ii) diagnosis date between 2010 and 2020; (iii) cases with positive histologic confirmation to ensure diagnostic accuracy; (iv) inclusion of samples

with complete survival time data; (v) exclusion of samples with only autopsy or death reports; and (vi) consideration of marital status, restricting samples to married and single patients to avoid potential misinterpretation from atypical marital statuses. The exclusion criteria were as follows: (i) patients without any metastases, and (ii) patients with incomplete data (such as age, stage, Gleason score, race, and radiotherapy information) and other tumor types. Strict exclusion of non-adenocarcinoma histotypes and secondary metastatic tumors to the prostate, combined with the above screening criteria, ensured that all enrolled patients were primary prostatic adenocarcinoma cases with complete clinicopathological and follow-up data. Based on these criteria, 10,408 patients were included in this study. The prognostic survival status and survival time of all enrolled patients were defined as OS, referring to the time from the initial diagnosis of metastatic prostate cancer to death from any cause. The detailed screening process is illustrated in Figure 1. As the data were obtained from an open database, ethical approval or informed consent was not required.

## 2.2. Feature selection

Statistical analyses were conducted using R (version 4.3.1; <https://www.r-project.org>) and Python (version 3.8.19; <https://www.python.org/>) software. To identify independent prognostic factors associated with metastatic prostate cancer, we applied LASSO-Cox regression to filter the features relevant to patients' prognosis. The selected features were subsequently incorporated into the predictive models. Patients were randomly split into training and testing sets in a 7:3 ratio using R. The optimal age cutoff value was determined using X-tile software.<sup>18</sup> To address potential biases from random sampling, we employed a chi-squared test to examine the variables between the two sets, with a significance threshold set at  $p < 0.05$ . The PSA levels were stratified into three categories (<10 ng/mL, 10–20 ng/mL, >20 ng/mL) in accordance with the clinical practice guidelines for prostate cancer, which is a widely accepted stratification criterion for prognostic assessment of prostate cancer patients in clinical settings. T stage was initially included as a candidate prognostic feature in the primary analysis, while LASSO-Cox regression revealed that T stage was not retained as a prognostic feature in the final model ( $p > 0.05$ ). Thus, T stage and other non-significant variables were excluded, and only statistically significant variables were retained for subsequent model construction and analysis. For lymph node (LN)-related features, N stage was used to represent regional LN involvement, whereas distant LN metastasis represented non-regional LN metastasis. The two features were included as independent variables in the subsequent

analysis due to their distinct clinical implications.

## 2.3. Construction of survival models

Based on the selected features, survival models were constructed using machine learning algorithms in the R package, including XGBoost (version 1.7.5.1)<sup>19</sup>, RandomForestSRC (version 3.2.2)<sup>20</sup>, CoxBoost (version 1.4)<sup>21</sup>, and SuperPC (version 1.12).<sup>22</sup> In addition, the DeepSurv<sup>23</sup> model, a novel deep learning approach combining Cox proportional hazards for survival analysis, was developed using Python. During the construction of these models, hyperparameter tuning was performed to select the optimal parameters for model building and training. Model training and hyperparameter tuning were performed on the training set. To evaluate the predictive accuracy of the survival models, we performed receiver operating characteristic (ROC) curve analysis using the R package timeROC (version 0.4)<sup>24</sup> based on the training and testing sets. The area under the receiver operating characteristic curve (AUC) was determined to evaluate the accuracy of the model for predicting OS at 1, 3, and 5 years. ROC curves were plotted for visual representation.

## 2.4. Analysis of key features from the optimal survival model

The final model was selected based on overall discrimination across the evaluated survival horizons together with the C-index. The importance of the features within this model was evaluated using the SHapley Additive exPlanations (SHAP) Python package (version 0.32.0).<sup>25</sup>

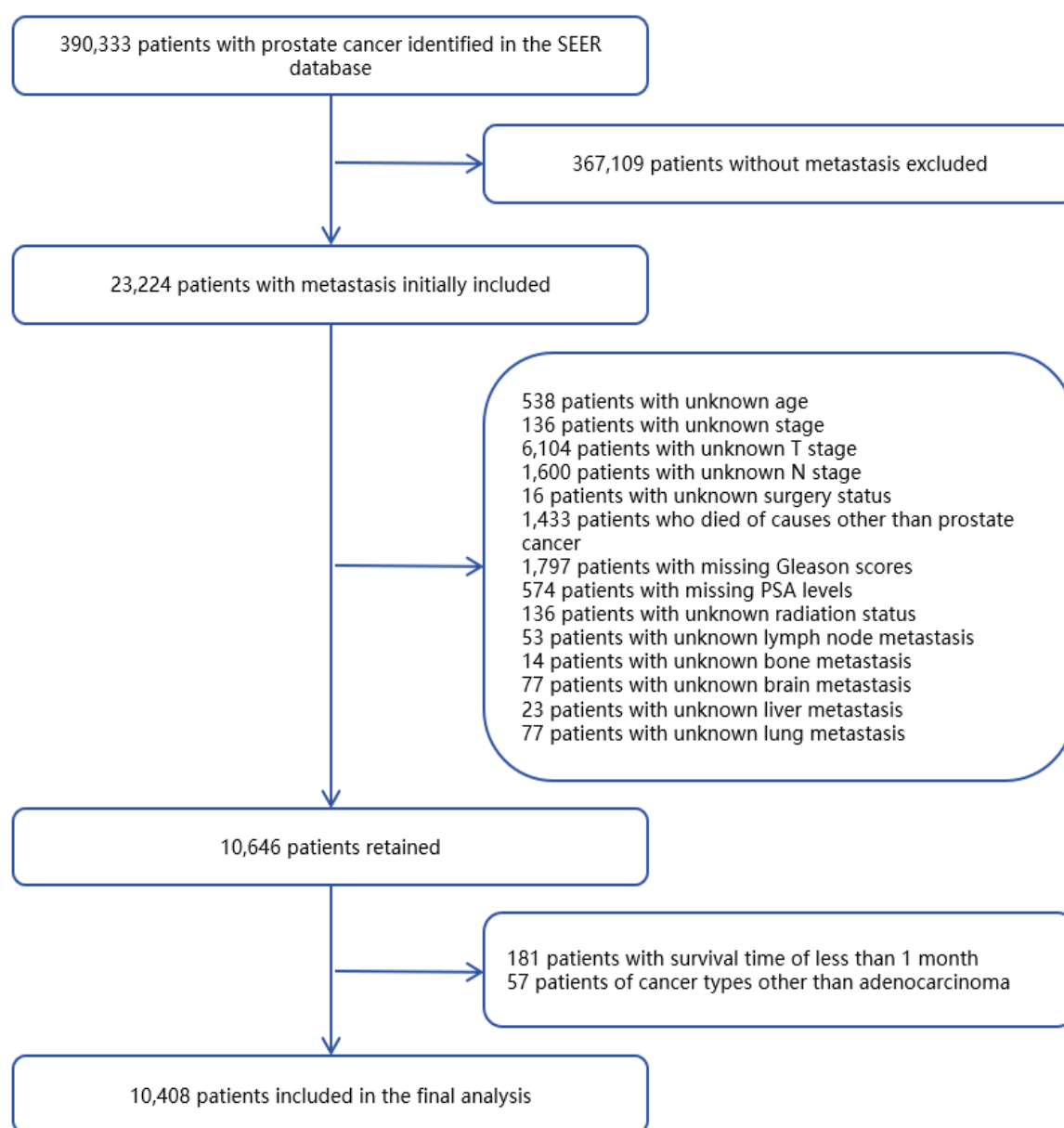
## 2.5. Kaplan–Meier curve analysis

Using the testing set, we calculated the risk scores of the samples using the optimal survival model in the R package survival (version 3.5-7).<sup>26</sup> The median risk score of all samples was set as the critical value for binary stratification: patients with a risk score higher than the median were classified into the high-risk group, while those with a risk score lower than or equal to the median were assigned to the low-risk group. The log-rank test was applied to assess the difference in survival outcomes between the two risk groups, and Kaplan–Meier (KM) survival curves were plotted to visualize the survival differences.

# 3. Results

## 3.1. Patient characteristics and feature selection

A total of 10,408 eligible patients were enrolled in this study. Subsequent LASSO-Cox regression analysis identified 15 features related to prognosis (Figure 2): marital status, age, race, N stage (regional LN metastasis), surgery, chemotherapy, radiation, LN metastasis (non-



**Figure 1.** Flowchart of patient inclusion and exclusion for the metastatic prostate cancer cohort. A total of 390,333 patients with prostate cancer were initially identified from the Surveillance, Epidemiology, and End Results (SEER) database. After exclusion of patients without metastasis ( $n = 367,109$ ), those with unknown age ( $n = 538$ ), unknown stage ( $n = 136$ ), unknown T stage ( $n = 6,104$ ), unknown N stage ( $n = 1,600$ ), unknown surgical status ( $n = 16$ ), deaths from causes other than prostate cancer ( $n = 1,433$ ), missing Gleason score ( $n = 1,797$ ), missing PSA levels ( $n = 574$ ), unknown radiation status ( $n = 136$ ), unknown lymph node metastasis status ( $n = 53$ ), unknown bone metastasis status ( $n = 14$ ), unknown brain metastasis status ( $n = 77$ ), unknown liver metastasis status ( $n = 23$ ), unknown lung metastasis status ( $n = 77$ ), survival time of less than 1 month ( $n = 181$ ), and non-adenocarcinoma histology ( $n = 57$ ), 10,408 patients with adenocarcinoma were included in the final analysis. Patients with missing key clinicopathologic or metastasis-related data were excluded to define the final analytic cohort.

regional LN metastasis), bone metastasis, brain metastasis, liver metastasis, lung metastasis, first malignant primary indicator (all cases were primary prostatic adenocarcinoma, with the 8% proportion attributed to sample size reduction after excluding missing data), Gleason score, and PSA level.

The patients were then randomly assigned into training ( $n = 7,285$ ) and testing ( $n = 3,123$ ) sets using R. Optimal age cutoff values were established via X-tile, resulting in three age groups:  $<73$ ,  $73-80$ , and  $>80$  years. Furthermore, chi-squared tests revealed no significant differences in features

between the training and testing sets (Table 1), indicating that the baseline characteristics were balanced between the training and testing sets and could be employed for further analyses.

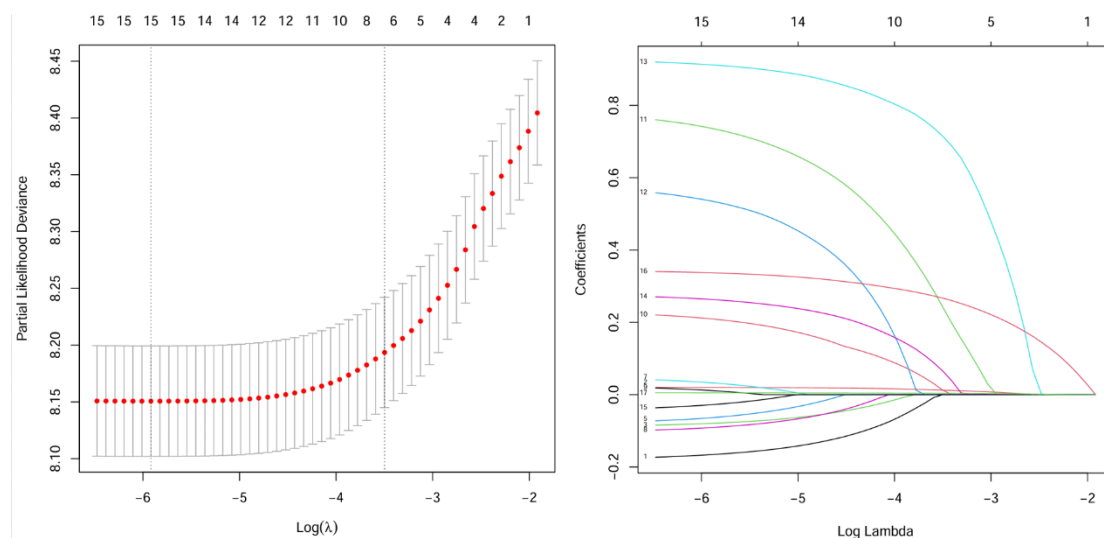
### 3.2. Construction and validation of survival models

Survival models were constructed using five algorithms, and their performances were validated based on the testing dataset. The C-indexes of the survival models constructed using the CoxBoost, RandomForestSRC, SuperPC, XGBoost, and DeepSurv algorithms were 0.6486, 0.6466, 0.6080, 0.6372, and 0.6488, respectively. ROC analysis showed that for 1-, 3-, and 5-year OS predictions, the AUC values of the CoxBoost model were 0.683, 0.686, and 0.697, respectively (Figure 3A); the AUC values of the RandomForestSRC model were 0.675, 0.683, and 0.700 (Figure 3B); the SuperPC model yielded AUC values of 0.633, 0.649, and 0.689 (Figure 3C); the AUC values of the XGBoost model were 0.688, 0.669, and 0.707 (Figure 3D); and the DeepSurv model produced AUC values of 0.696, 0.685, and 0.720 (Figure 3E). Owing to the highest AUC and C-index values of the DeepSurv model among these models, this model was chosen as the final survival model. The optimal DeepSurv model integrates 15 prognostic features (including Gleason score, age, and metastasis-related indicators) to generate individual risk scores for each patient (Figure 4). This score-based stratification enables precise differentiation of survival outcomes, as

evidenced by the significant survival difference between high-risk and low-risk groups ( $p < 0.001$ ; Figure 5), confirming the model's practical value for individualized prognostic assessment.

### 3.3. Identification of key features from the optimal survival model

Using the SHAP library, we assessed the importance of the features within the DeepSurv model. The results showed that the five features with the highest importance were LN metastasis, PSA level, age, Gleason score, and lung metastasis. The importance ranking of the features and the error variations during the training process of DeepSurv are shown in Figure 4A and 4B. Moreover, we calculated the feature importance using the random forest algorithm, and the top five features were the Gleason score, liver metastasis, age, bone metastasis, and brain metastasis (Figure 4C). Notably, the Gleason score and age were ranked among the top five by comparison of the importance rankings of the two algorithms. Consistently, Gleason score and age were ranked among the top five in both algorithms, which further confirmed their stable and core prognostic role in patients with metastatic prostate cancer—even when multiple metastasis-related clinical features were included in the model. Notably, the integration of these core features with other clinical variables (e.g., LN metastasis, PSA level) in the DeepSurv model enhances the accuracy of individual risk prediction, which is the key advantage of



**Figure 2.** LASSO regression parameter selection. LASSO, least absolute shrinkage and selection operator. The left panel illustrates the partial likelihood deviances across different values of the regularization parameter  $\text{Log}(\lambda)$ , with the optimal value marked by the vertical dashed line. The right panel displays the trajectories of the coefficients for the features as  $\text{Log}(\lambda)$  decreases. Each line represents the coefficient path of a single feature, demonstrating how the feature importance changes with the regularization strength. The numbers above the lines in the right panel do not correspond to specific features but are identifiers generated by the algorithm to distinguish between different features.

**Table 1. Comparison of clinicopathological characteristics of patients with metastatic prostate cancer in the training and testing sets**

| Baseline characteristics           | Test ( <i>n</i> = 3,123) | Train ( <i>n</i> = 7,285) | <i>p</i> -value |
|------------------------------------|--------------------------|---------------------------|-----------------|
| Marital                            |                          |                           |                 |
| Single                             | 618 (19.8%)              | 1,524 (20.9%)             | 0.2             |
| Married                            | 2,505 (80.2%)            | 5,761 (79.1%)             |                 |
| Age                                |                          |                           |                 |
| <73                                | 2,007 (64.3%)            | 4,769 (65.5%)             | 0.483           |
| 73–80                              | 714 (22.9%)              | 1,598 (21.9%)             |                 |
| >80                                | 402 (12.9%)              | 918 (12.6%)               |                 |
| Race                               |                          |                           |                 |
| White                              | 2,422 (77.6%)            | 5,612 (77.0%)             | 0.946           |
| Black                              | 475 (15.2%)              | 1,132 (15.5%)             |                 |
| Asian or Pacific Islander          | 205 (6.6%)               | 493 (6.8%)                |                 |
| American Indian/Alaska Native      | 21 (0.7%)                | 48 (0.7%)                 |                 |
| T stage                            |                          |                           | 0.357           |
| T1–T2                              | 892 (28.6%)              | 2,071 (28.4%)             |                 |
| T3–T4                              | 2,231 (71.4%)            | 5,214 (71.6%)             |                 |
| N (regional lymph node metastasis) |                          |                           |                 |
| N0                                 | 1,904 (61.0%)            | 4,402 (60.4%)             | 0.62            |
| N1                                 | 1,219 (39.0%)            | 2,883 (39.6%)             |                 |
| Surgery                            |                          |                           |                 |
| No                                 | 2,696 (86.3%)            | 6,246 (85.7%)             | 0.446           |
| Yes                                | 427 (13.7%)              | 1,039 (14.3%)             |                 |
| Chemotherapy                       |                          |                           |                 |
| No                                 | 2,531 (81.0%)            | 5,868 (80.5%)             | 0.576           |
| Yes                                | 592 (19.0%)              | 1,417 (19.5%)             |                 |
| Radiation                          |                          |                           |                 |
| No                                 | 2,293 (73.4%)            | 5,318 (73.0%)             | 0.673           |
| Yes                                | 830 (26.6%)              | 1,967 (27.0%)             |                 |
| Lymph node metastasis              |                          |                           |                 |
| No                                 | 2,431 (77.8%)            | 5,720 (78.5%)             | 0.459           |
| Yes                                | 692 (22.2%)              | 1,565 (21.5%)             |                 |
| Bone metastasis                    |                          |                           |                 |
| No                                 | 233 (7.5%)               | 550 (7.5%)                | 0.907           |
| Yes                                | 2,890 (92.5%)            | 6,735 (92.5%)             |                 |

*(Cont'd...)*

Table 1. (Continued)

| Baseline characteristics          | Test (n = 3,123) | Train (n = 7,285) | p-value |
|-----------------------------------|------------------|-------------------|---------|
| Brain metastasis                  |                  |                   |         |
| No                                | 3,099 (99.2%)    | 7,252 (99.5%)     | 0.064   |
| Yes                               | 24 (0.8%)        | 33 (0.5%)         |         |
| Liver metastasis                  |                  |                   |         |
| No                                | 3,029 (97.0%)    | 7,047 (96.7%)     | 0.533   |
| Yes                               | 94 (3.0%)        | 238 (3.3%)        |         |
| Lung metastasis                   |                  |                   |         |
| No                                | 2,905 (93.0%)    | 6,757 (92.8%)     | 0.658   |
| Yes                               | 218 (7.0%)       | 528 (7.2%)        |         |
| First malignant primary indicator |                  |                   |         |
| No                                | 251 (8.0%)       | 578 (7.9%)        | 0.89    |
| Yes                               | 2,872 (92.0%)    | 6,707 (92.1%)     |         |
| Gleason score                     |                  |                   |         |
| 6                                 | 57 (1.8%)        | 131 (1.8%)        | 0.709   |
| 7                                 | 432 (13.8%)      | 962 (13.2%)       |         |
| 8                                 | 737 (23.6%)      | 1,776 (24.4%)     |         |
| 9                                 | 1,525 (48.8%)    | 3,596 (49.4%)     |         |
| 10                                | 372 (11.9%)      | 820 (11.3%)       |         |
| PSA                               |                  |                   |         |
| <10 ng/mL                         | 407 (13.0%)      | 991 (13.6%)       | 0.222   |
| 10–20 ng/mL                       | 410 (13.1%)      | 1,032 (14.2%)     |         |
| >20 ng/mL                         | 2,306 (73.8%)    | 5,262 (72.2%)     |         |

Note: N stage: Regional lymph node metastasis; Lymph node.metastasis: Non-regional lymph node metastasis.

our model compared to previous studies that only focused on validating single prognostic factors.

### 3.4. Kaplan–Meier curve analysis

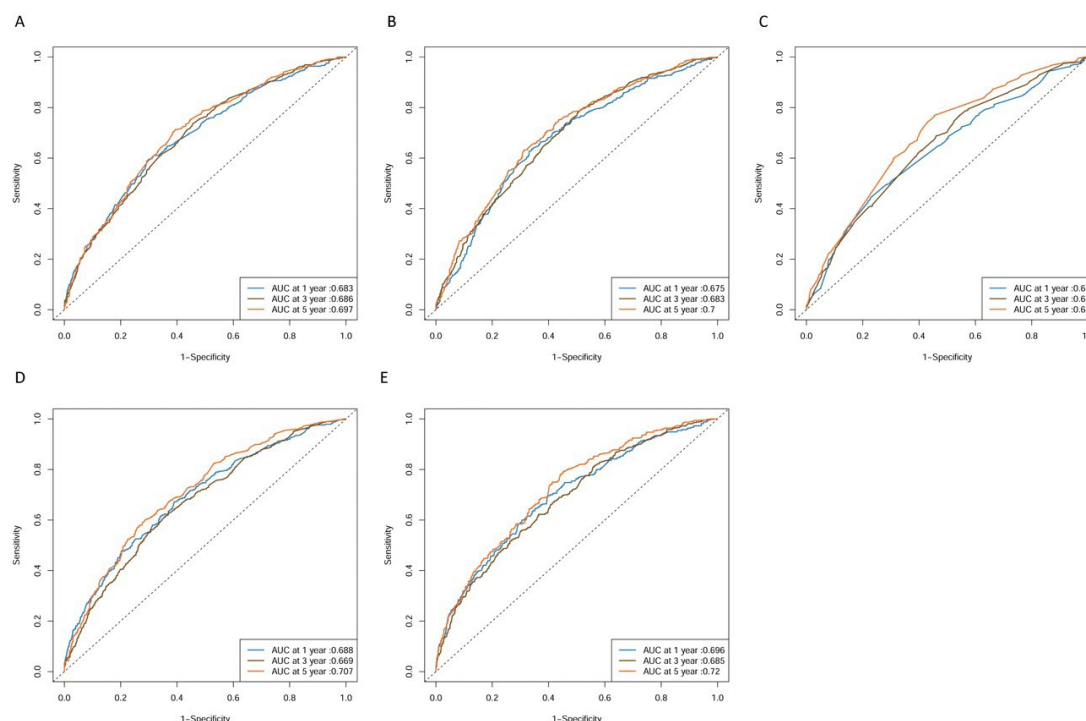
Using the testing set, we generated risk scores from the DeepSurv model for the testing cohort, and high-risk patients had significantly shorter OS than low-risk patients ( $p < 0.001$ ; Figure 5).

## 4. Discussion

Metastatic prostate cancer is linked to a substantial decrease in OS.<sup>27</sup> Given the large number of patients with metastatic disease and projected increase in mortality rates, the importance of identifying key prognostic indicators cannot be overemphasized. This study provides a comprehensive analysis of the prognostic factors influencing OS outcomes in patients with metastatic

prostate cancer, utilizing a robust dataset from the SEER database. Moreover, the DeepSurv model performed better than the other algorithms, showing high performance in predicting survival.

Our feature selection process, employing LASSO-Cox regression, enabled us to identify 15 critical features that were closely related to the prognosis of metastatic prostate cancer, including marital status, age, race, N stage, surgery, chemotherapy, radiation, LN metastasis, bone metastasis, brain metastasis, liver metastasis, lung metastasis, first malignant primary indicator, Gleason score, and PSA level. These features have been confirmed to be important prognostic factors in prostate cancer. For instance, Chen *et al.*<sup>28</sup> identified marital status, age, race, TNM stage, radiotherapy, chemotherapy, surgery, Gleason score, and PSA concentration as independent prognostic factors for prostate cancer. Li *et al.*<sup>29</sup> revealed



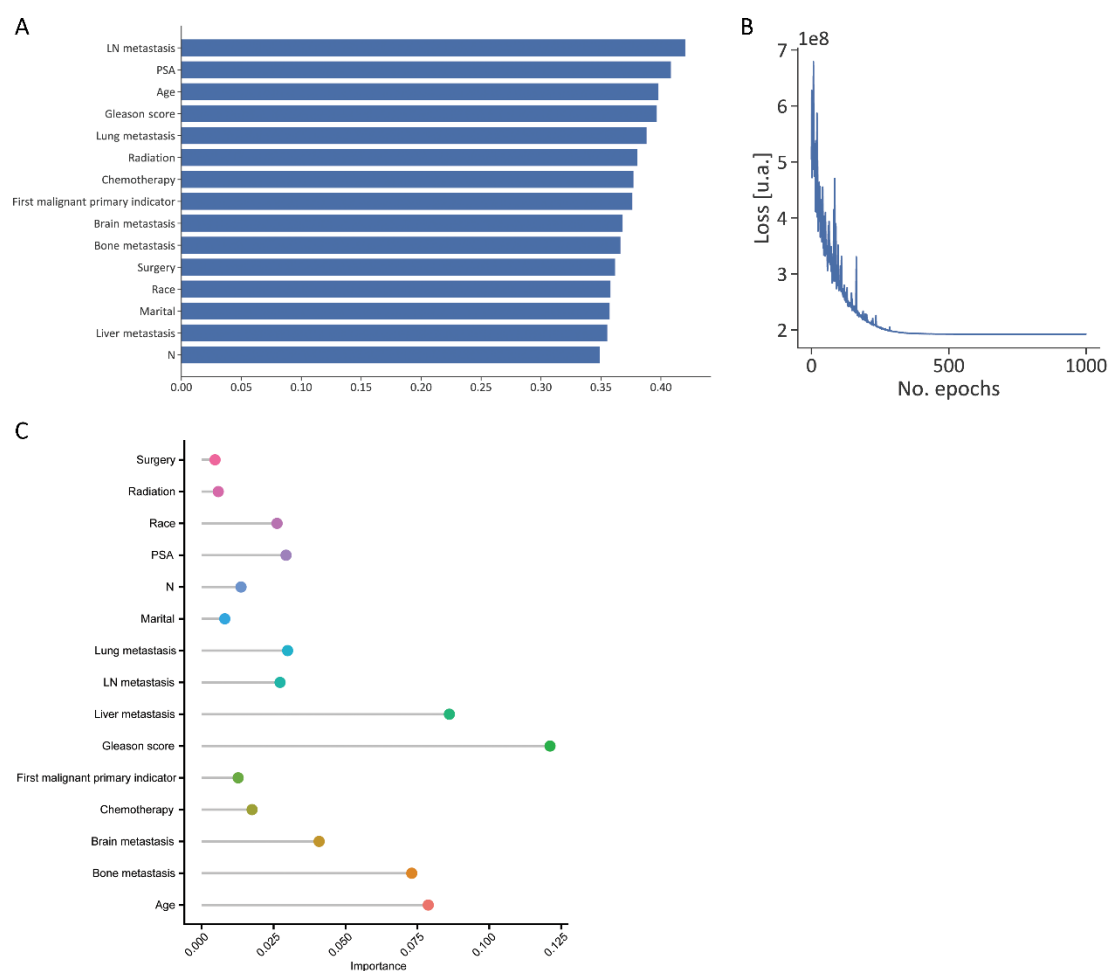
**Figure 3.** The 1-, 3-, and 5-year ROC curves of five algorithms. (A) CoxBoost model; (B) RandomForestSRC model; (C) SuperPC model; (D) XGBoost model; (E) DeepSurv model.

Abbreviation: AUC: Area under the receiver operating characteristic curve.

that age, marital status, bone metastasis, brain metastasis, liver metastasis, lung metastasis, distant LN metastasis, and N stage are significant factors affecting the prognosis of prostate cancer. To better understand the significance of these features, we calculated their importance using DeepSurv and random forest algorithms and found that factors such as Gleason score and age were key prognostic factors across different modeling approaches. An older age at diagnosis has been connected to a higher incidence of de novo metastatic prostate cancer and worse prostate cancer-specific survival.<sup>30–33</sup> The poorer prognosis of older adult patients with metastatic prostate cancer may be attributed to the more rapid advancement of resistant disease<sup>31</sup> and their limited capacity to undergo comprehensive treatments, such as chemotherapy, owing to frailty or existing comorbidities.<sup>30</sup> Additional research is needed to clarify the reasons for the poorer prostate cancer-specific survival in older adult men and create new treatment options suitable for frail patients with comorbidities. In addition, the risk and timing of metastatic progression in prostate cancer are closely associated with Gleason score in real-world scenarios, even after accounting for potentially unfavorable confounding factors.<sup>34</sup> A longer interval

between the initial diagnosis of prostate cancer and the onset of metastases is associated with higher survival rates. Therefore, the Gleason score should be incorporated into clinical follow-up and patient counseling after initial local treatment, particularly for patients with higher Gleason scores, to improve the prognosis of metastatic prostate cancer.

As well-recognized core prognostic indicators for prostate cancer, the Gleason score and age showed stable and high prognostic importance in our study of metastatic prostate cancer, which is consistent with the conclusions of previous population-based studies. Unlike previous studies that primarily validated the prognostic value of individual factors (e.g., Gleason score, age) in metastatic prostate cancer, the core innovation of our study lies in developing a comprehensive clinical prognostic model integrated with multiple clinical features. This model can calculate a personalized risk score for each patient to predict subsequent prognosis, which addresses the clinical need for individualized risk assessment in metastatic prostate cancer management. However, the present study further revealed three novel perspectives for these two indicators in the metastatic subgroup: First, in the context of multiple



**Figure 4.** Analysis of the importance of features within the DeepSurv model. (A) Importance ranking of features. (B) Error variation during the training process of DeepSurv. (C) Feature importance calculated by the random forest algorithm. Abbreviation: LN: Lymph node; N: Node.

metastasis-related features (e.g., LN metastasis, multi-organ metastasis) being included in the model, Gleason score and age still maintained top five prognostic importance, suggesting their irreplaceable predictive value even when metastatic status is confirmed. Second, the Gleason score outperformed most organ metastasis indicators (e.g., bone, brain, liver metastasis) in the random forest analysis, which indicated that the intrinsic tumor aggressiveness characterized by the Gleason score may have a more persistent impact on survival than the specific metastatic site in patients with metastatic prostate cancer. Third, age was a consistent top-five feature in both algorithms, and combined with the age stratification result (<73, 73–80, >80 years), it further confirmed that advanced age is an independent adverse prognostic factor in this subgroup—possibly due to the limited tolerance to systemic therapy and the higher incidence of comorbidities. These findings

provide new clinical insights for prognostic stratification of metastatic prostate cancer: in clinical practice, even for patients with confirmed metastasis, the routine assessment of Gleason score (at initial diagnosis) and age should still be prioritized, and their combination with metastasis-related indicators (e.g., LN metastasis, PSA level) can more accurately stratify the survival risk of patients, which is conducive to the formulation of individualized treatment strategies.

Taken together, these findings support the rationale for integrating such factors into clinical practice to refine risk stratification and approaches for managing patients with metastatic disease.

Survival analysis is a branch of statistics focused on the assessment of time-to-event outcomes such as death, metastasis, and recurrence.<sup>35</sup> This approach effectively

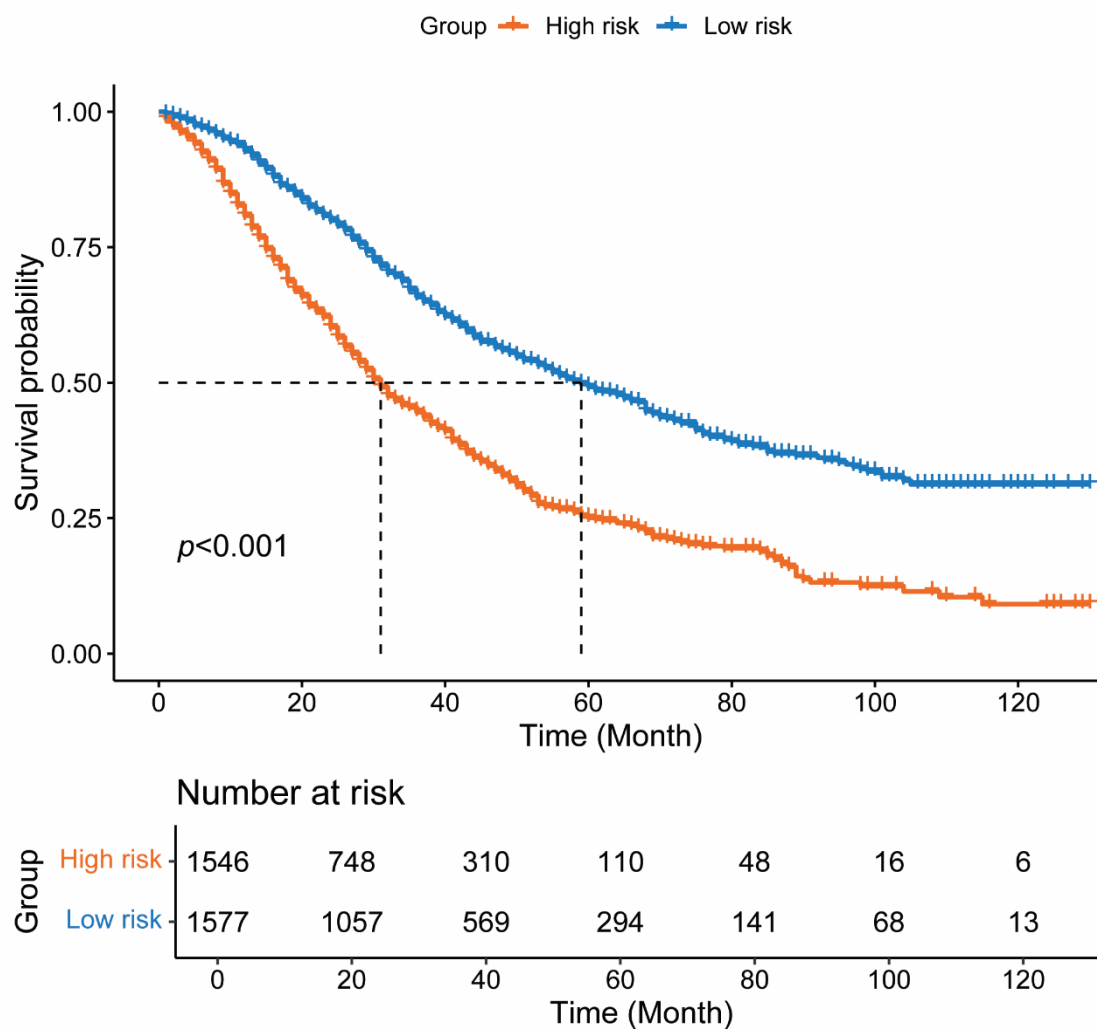


Figure 5. Kaplan–Meier curve analysis based on the testing set

addresses censored data and offers a more dynamic perspective of the probability of an event occurring over time.<sup>36</sup> The application of machine learning and deep learning algorithms has made significant progress in the domain of survival analyses. Several machine learning and deep learning algorithms have been utilized to develop survival models to predict cancer survival.<sup>37–39</sup> In this study, the DeepSurv model demonstrated superior predictive capabilities compared to other algorithms, as evidenced by the high AUC and C-index values across the survival endpoints. Patient stratification, followed by KM survival analysis based on the testing dataset, confirmed the validity of our model, highlighting its potential for clinical applications. These risk categories allow clinicians to tailor treatment plans for individual patients more effectively,

emphasizing the need for personalized approaches to metastatic prostate cancer management. It is worth mentioning that the development of prognostic models using machine learning and deep learning algorithms has shown great potential in improving the accuracy of cancer survival prediction. For example, Saito *et al.*<sup>37</sup> used machine learning to predict time-series prognosis factors in metastatic prostate cancer patients treated with androgen deprivation therapy, achieving good predictive performance. This indicates that machine learning and deep learning algorithms can effectively capture the complex relationships between various factors and survival outcomes, thereby providing more accurate prognostic information for clinicians and patients.

This study has some limitations. First, this study was

retrospective, and the potential for selection bias could not be eliminated. Second, this study was based exclusively on data from the SEER database, and the lack of some variables in SEER inevitably constrained our findings. Third, although the study used data from the SEER database, which includes patients from various geographic locations, it did not explicitly address the applicability of the model across different populations or subgroups. It is recommended to analyze the performance of the model in various subgroups, such as race, age, and region, and explore potential differences and limitations in its clinical application across different global regions. Finally, the predictive model was not externally validated, and its applicability should be further validated in clinical cohorts. In the future, more prospective multicenter studies across diverse ethnic populations should be conducted to validate the effectiveness of the model and explore guidelines and strategies for its specific application in clinical practice. The practical significance of our study is reflected in the translation of research findings into a usable tool: the developed DeepSurv model can be applied in clinical practice to calculate individual risk scores, helping clinicians identify high-risk patients who may benefit from more aggressive treatment and low-risk patients for whom overtreatment can be avoided. This personalized approach is conducive to improving the overall management efficiency and patient outcomes of metastatic prostate cancer.

## 5. Conclusion

This study successfully developed a comprehensive prognostic model for metastatic prostate cancer based on the DeepSurv algorithm, which integrates 15 clinical features, including Gleason score and age (core prognostic factors). The model can calculate individualized risk scores to predict subsequent patient prognosis, demonstrating high survival prediction accuracy and reliable risk stratification capability by integrating these classic indicators with metastasis-related features. This study not only reinforces the role of established prognostic factors in metastatic prostate cancer but also advocates for the integration of advanced statistical methodologies into routine clinical practice. Future investigations should prioritize the validation of these models in larger independent cohorts and examine the incorporation of additional genomic and molecular data to provide further insights into patient outcomes and potential treatment targets. The successful application of machine learning and deep learning algorithms in this study highlights the potential of these advanced methodologies to transform cancer prognosis and treatment. As the field of oncology continues to evolve, the integration of artificial intelligence

and big data analytics will undoubtedly play a crucial role in improving patient outcomes and advancing personalized medicine. Therefore, it is essential to continue exploring and refining these technologies to better serve the needs of patients with metastatic prostate cancer and other complex diseases. Additionally, the collaboration between clinicians, data scientists, and researchers is vital for the development of more accurate and reliable prognostic models. By leveraging the power of machine learning and deep learning, we can potentially uncover new biomarkers and therapeutic targets, leading to more effective treatment strategies for metastatic prostate cancer.

## Acknowledgments

We would like to express our gratitude to the staff at The First Affiliated Hospital of Guangxi Medical University Hechi Hospital for their support and assistance.

## Funding

None.

## Conflict of interest

The authors declare that they have no competing interests.

## Author contributions

*Conceptualization:* Xian-Yong Yan, Hai Lu

*Formal analysis:* Jiang Qin

*Investigation:* Xian-Yong Yan, Jing-Zhi Huang, Qing-Fan Wei

*Methodology:* Rong-Yuan Wei

*Writing - original draft:* Xian-Yong Yan

*Writing - review & editing:* Hai Lu, Jie Lan

## Ethics approval and consent to participate

As the data were obtained from an open database, ethical approval and informed consent were not required.

## Consent for publication

There are no identifiable patient data included in the manuscript that would require individual consent for publication.

## Availability of data

The data used in this study were obtained from the Surveillance, Epidemiology, and End Results (SEER) database, which is publicly accessible. The datasets are available for researchers under specific access conditions. For further information and to access the data, please visit the SEER website at <https://seer.cancer.gov>. The materials and datasets generated or analyzed during this study are

included in this published article and can be provided by the corresponding authors upon reasonable request.

## References

- Bray F, Laversanne M, Sung H, *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229-263.  
doi: 10.3322/caac.21834
- Sandhu S, Moore CM, Chiong E, Beltran H, Bristow RG, Williams SG. Prostate cancer. *Lancet.* 2021;398(10305):1075-1090.  
doi: 10.1016/S0140-6736(21)00950-8
- Arriaga JM, Ronaldson-Bouchard K, Picech F, *et al.* In vivo genome-wide CRISPR screening identifies CITED2 as a driver of prostate cancer bone metastasis. *Oncogene.* 2024;43(17):1303-1315.  
doi: 10.1038/s41388-024-02995-5
- Pan J, Tong F, Ren N, *et al.* Role of N(6)-methyladenosine in the pathogenesis, diagnosis and treatment of prostate cancer (Review). *Oncol Rep.* 2024;51(6).  
doi: 10.3892/or.2024.8747
- Posdzich P, Darr C, Hilser T, *et al.* Metastatic prostate cancer—A review of current treatment options and promising new approaches. *Cancers.* 2023;15(2):461.  
doi: 10.3390/cancers15020461
- Shah N, Ioffe V. Early detection of prostate cancer: AUA/SUO guideline part I: Prostate cancer screening. Letter. *J Urol.* 2023;210(5): 731.  
doi: 10.1097/JU.0000000000003682
- Cao G, Li Y, Wang J, *et al.* Gleason score, surgical and distant metastasis are associated with cancer-specific survival and overall survival in middle aged high-risk prostate cancer: A population-based study. *Front Public Health.* 2022;10:1028905.  
doi: 10.3389/fpubh.2022.1028905
- Liu D, Kuai Y, Zhu R, *et al.* Prognosis of prostate cancer and bone metastasis pattern of patients: a SEER-based study and a local hospital based study from China. *Sci Rep.* 2020;10(1):9104.  
doi: 10.1038/s41598-020-64073-6
- Zelic R, Garmo H, Zugna D, *et al.* Predicting prostate cancer death with different pretreatment risk stratification tools: A head-to-head comparison in a nationwide cohort study. *Eur Urol.* 2020;77(2):180-188.  
doi: 10.1016/j.eururo.2019.09.027
- Gnanapragasam VJ, Lophatananon A, Wright KA, Muir KR, Gavin A, Greenberg DC. Improving clinical risk stratification at diagnosis in primary prostate cancer: A prognostic modelling study. *PLoS Med.* 2016;13(8):e1002063.  
doi: 10.1371/journal.pmed.1002063
- Lee C, Light A, Alaa A, Thurtle D, Van Der Schaar M, Gnanapragasam VJ. Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database. *Lancet Digit Health.* 2021;3(3):e158-e165.  
doi: 10.1016/s2589-7500(20)30314-9
- Sanda MG, Cadeddu JA, Kirkby E, *et al.* Clinically Localized Prostate Cancer: AUA/ASTRO/SUO Guideline. Part I: Risk stratification, shared decision making, and care options. *J Urol.* 2017;199(3):683-690.  
doi: 10.1016/j.juro.2017.11.095
- Gravis G, Boher JM, Fizazi K, *et al.* Prognostic factors for survival in noncastrate metastatic prostate cancer: validation of the glass model and development of a novel simplified prognostic model. *Eur Urol.* 2014;68(2):196-204.  
doi: 10.1016/j.eururo.2014.09.022
- Li T, Huang H, Zhang S, *et al.* Predictive models based on machine learning for bone metastasis in patients with diagnosed colorectal cancer. *Front Public Health.* 2022;10:984750.  
doi: 10.3389/fpubh.2022.984750
- Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data. *Med Care.* 2002;40(Supplement):IV-3.  
doi: 10.1097/00005650-200208001-00002
- Adamo M, Boten JA, Coyle LM, *et al.* Validation of prostate-specific antigen laboratory values recorded in Surveillance, Epidemiology, and End Results registries. *Cancer.* 2016;123(4):697-703.  
doi: 10.1002/cncr.30401
- Yi X, Xu W, Tang G, *et al.* Individual risk and prognostic value prediction by machine learning for distant metastasis in pulmonary sarcomatoid carcinoma: a large cohort study based on the SEER database and the Chinese population. *Front Oncol.* 2023;13:1105224.  
doi: 10.3389/fonc.2023.1105224
- Camp RL, Dolled-Filhart M, Rimm DL. X-Tile. *Clin Cancer Res.* 2004;10(21):7252-7259.  
doi: 10.1158/1078-0432.ccr-04-0713
- Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.  
doi: 10.1145/2939672.2939785
- Ishwaran H, Kogalur UB, Kogalur MUB. Package

- 'randomForestSRC'. *Breast*. 2022;6(1).
21. Binder H. CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks [Computer software]. R package version 1.10. 2013;1(4). doi: 10.1198/016214505000000628
  22. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. *J Am Stat Assoc*. 2006;101(473):119-137. doi: 10.1198/016214505000000628
  23. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. *J Am Stat Assoc*. 2006;101(473):119-137. doi: 10.1198/016214505000000628
  24. Blanche P, Blanche MP. Package 'timeROC'. 2019, updated 2019-12-18. Available from: <https://cran.r-project.org/web/packages/timeROC> [Last accessed on 15 September 2024].
  25. Lundberg SM, Erion G, Chen H, *et al*. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. doi: 10.1038/s42256-019-0138-9
  26. Lin H, Zelterman D. *Modeling survival data: extending the Cox model*. 2002, Taylor & Francis.
  27. Achard V, Putora PM, Omlin A, Zilli T, Fischer S. Metastatic prostate cancer: treatment options. *Oncology*. 2021;100(1):48-59. doi: 10.1159/000519861
  28. Chen Q, Zhu X, Hu Y, *et al*. A study on the impact of marital status on the survival status of prostate cancer patients based on propensity score matching. *Sci Rep*. 2024;14(1):6162. doi: 10.1038/s41598-024-56145-8
  29. Li Y, Ma X, Guan C, Yang X. Analysis of the influence of marital status on prognosis of prostate cancer patients based on big data. *Am J Clin Exp Urol*. 2022;10(5):320-326.
  30. Bernard B, Burnett C, Sweeney CJ, Rider JR, Sridhar SS. Impact of age at diagnosis of de novo metastatic prostate cancer on survival. *Cancer*. 2019;126(5):986-993. doi: 10.1002/cncr.32630
  31. Humphreys MR, Fernandes KA, Sridhar SS. Impact of Age at Diagnosis on Outcomes in Men with Castrate-Resistant Prostate Cancer (CRPC). *J Cancer*. 2013;4(4):304-314. doi: 10.7150/jca.4192
  32. MacKintosh FR, Sprenkle PC, Walter LC, *et al*. Age and Prostate-Specific Antigen Level Prior to Diagnosis Predict Risk of Death from Prostate Cancer. *Front Oncol*. 2016;6:157. doi: 10.3389/fonc.2016.00157
  33. Scosyrev E, Messing EM, Mohile S, Golijanin D, Wu G. Prostate cancer in the elderly. *Cancer*. 2011;118(12):3062-3070. doi: 10.1002/cncr.26392
  34. Wenzel M, Lutz M, Hoeh B, *et al*. Influence of tumor characteristics and time to metastatic disease on oncological outcomes in metachronous metastatic prostate cancer patients. *Clin Genitourin Cancer*. 2024;22(5):102158. doi: 10.1016/j.clgc.2024.102158
  35. Schober P, Vetter TR. Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. *Anesth Analg*. 2018;127(3):792-798. doi: 10.1213/ane.0000000000003653
  36. George B, Seals S, Aban I. Survival analysis and regression models. *J Nucl Cardiol*. 2014;21(4):686-694. doi: 10.1007/s12350-014-9908-2
  37. Saito S, Sakamoto S, Higuchi K, *et al*. Machine-learning predicts time-series prognosis factors in metastatic prostate cancer patients treated with androgen deprivation therapy. *Sci Rep*. 2023;13(1):6325. doi: 10.1038/s41598-023-32987-6
  38. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. *Sci Rep*. 2021;11(1):13505. doi: 10.1038/s41598-021-92799-4
  39. Moradmamand H, Aghamiri SMR, Ghaderi R, Emami H. The role of deep learning-based survival model in improving survival prediction of patients with glioblastoma. *Cancer Med*. 2021;10(20):7048-7059. doi: 10.1002/cam4.4230