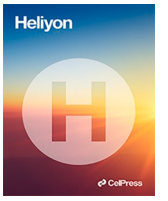
Heliyon 10 (2024) e38242

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Improve clinical feature-based bladder cancer survival prediction models through integration with gene expression profiles and machine learning techniques

Yali Tang [a], Shitian Li [b], Liang Zhu [b], Lei Yao [b], Jianlin Li [b], Xiaoqi Sun [b], Yuan Liu [b], Yi Zhang [b], Xinyang Fu [b,*]

[a] Department of Oncology, Kaiping Central Hospital, Kaiping, Jiangmen, China
[b] Department of Urology, Kaiping Central Hospital, Kaiping, Jiangmen, China

ARTICLE INFO

Keywords:
Bladder cancer
Gene sets
Machine learning
Prognosis

ABSTRACT

Background: Bladder cancer (BCa), one of the most common cancers worldwide, is characterized by high rates of recurrence, progression, and mortality. Machine learning algorithms offer promising advancements in enhancing predictive models. This study aims to develop robust machine learning models for predicting BCa survival using clinical and gene expression data.
Methods: Clinical data from BCa patients were obtained from the Surveillance, Epidemiology, and End Results database. Cox proportional hazards regression models assessed the association between clinical variables and overall survival. Machine learning algorithms, including logistic regression, random forest, XGBoost, decision tree, and LightGBM, were employed to predict survival at 1, 3, and 5 years. The TAGO database, combined with the data from The Cancer Genome Atlas and four databases from the Gene Expression Omnibus, which have available genomic data and clinical data, were selected. Gene expression data were transformed into gene sets data, and the performance of models based on clinical data and gene sets data and their combination were compared. Furthermore, the impact of model-derived scores on overall survival was evaluated.
Results: Among 138,741 BCa patients with available clinical data, key independent predictors of survival included age, race, marital status, surgery, chemotherapy, radiation, and TNM stages. Clinical data machine learning (CML) models used these clinical predictors to achieve AUC values of 0.860, 0.821, and 0.804 in the testing sets for predicting survival at 1, 3, and 5 years, respectively. In the TAGO database, which has 863 patients with clinical and genomic data, the integrated clinical and gene expression machine learning model (IML) outperformed the CML and gene expression machine learning (GML) models in survival prediction. Patients with higher IML and GML model scores exhibited poorer survival outcomes.
Conclusions: This study successfully identifies key clinical and genomic predictors, a significant step forward in BCa research. The development of predictive models for BCa survival underscores the potential of integrated data approaches in improving BCa management and treatment strategies.

* Corresponding author.
E-mail address: fxytyl@163.com (X. Fu).

https://doi.org/10.1016/j.heliyon.2024.e38242
Received 20 May 2024; Received in revised form 13 August 2024; Accepted 20 September 2024
Available online 21 September 2024

1. Introduction

Bladder cancer (BCa), one of the most common malignant tumors, is diagnosed in over 500,000 individuals worldwide annually, resulting in approximately 200,000 deaths each year [1]. BCa is classified into non-muscle invasive bladder cancer (NMIBC) and invasive (MIBC) types based on tumor invasion depth [2]. Patients with NMIBC generally exhibit a favorable 5-year survival rate (over 90 percent), although they have high recurrence and progression rates [3]. One of the primary treatments for NMIBC is transurethral resection of the bladder tumor (TURBT) [4]. In contrast, MIBC patients often have a poor prognosis and require more aggressive treatments, including radical cystectomy, chemotherapy, and immunotherapy [5]. Given MIBC's aggressive nature and tendency for metastasis, its 5-year survival rate falls below 50 percent [6]. Several prognostic factors are strongly linked to the outcomes in patients with BCa. These include anatomical extent like the TNM staging system [7]; histological details such as pathological subtype and tumor grade [7]; baseline characteristics including age and gender [8]; and various molecular biomarkers [9]. Accurate risk stratification is essential for clinicians to develop tailored surveillance and follow-up strategies because of its high progression and death rates. Implementing a systematic surveillance protocol will enhance survival benefits for BCa patients.

Machine learning is a powerful tool for integrating critical clinicopathological parameters and molecular biomarkers to predict clinical outcomes. The ability of machine learning models to predict 1, 3, and 5-year outcomes can significantly aid clinicians in making informed decisions. Despite several risk models being developed for BCa patient risk stratification using these clinical factors [10,11], they often face limitations such as small sample sizes or lack of external validation. In this study, we constructed and validated novel machine learning models using a large population dataset from the Surveillance, Epidemiology, and End Results (SEER) database, aiming to predict the overall survival (OS) of BCa patients. Moreover, the gene set-based genetic data was added to increase the accuracy of machine learning models. This approach underscores the necessity and advantages of machine learning in enhancing prognostic accuracy and providing robust, validated tools that support personalized treatment planning and improve patient outcomes.

Despite numerous advancements in treatment strategies, BCa continues to pose significant challenges for effective management. Immune dysregulation is central to the progression of BCa [12], which has catalyzed the development of immunotherapy as a promising treatment avenue. Immune checkpoint inhibitors (ICI) like atezolizumab, pembrolizumab, durvalumab, nivolumab, and avelumab have been approved for metastatic and advanced BCa [13]. However, the effectiveness of ICI remains limited, with response rates in BCa patients typically at most 10–30 % [14]. Given this variability in response, there is a critical need for predictive tools that can accurately identify which BCa patients are likely to benefit from ICI therapy.

In this study, we aimed to identify independent risk factors that affect overall survival in BCa patients and to develop predictive models. We constructed and validated machine learning models that estimate survival probabilities at 1, 3, and 5 years for patients with BCa. By integrating gene expression data with clinical information, the accuracy values of survival prediction models were enhanced. This integration of genomic and clinical data allowed us to create personalized prognosis prediction models for individual patients, ultimately aiding in more informed clinical decision-making.

2. Methods

2.1. Data sources

2.1.1. The database for clinical data machine learning (CML)

The workflow for this study is illustrated in Fig. 1. Since the SEER (Surveillance, Epidemiology, and End Results) is a publicly accessible database, ethical review or informed consent is not required in this study. The data used in this study was sourced from the November 2022 release of SEER Research Data. Patient records were obtained using the SEER*Stat software, containing cancer-related population data across 17 registries from 2000 to 2020, covering approximately 30 % of the U.S. population. The inclusion criteria were: (1) Subjects identified using the site codes C67.0–C67.9 for bladder cancer. (2) Patients diagnosed between 2004 and 2015, as the AJCC 6th edition was used during this timeframe. (3) Subjects flagged as the first malignant primary indicator. The exclusion criteria included: (1) Survival time of less than one month. (2) Patients under 18 years of age. (3) Missing values in survival or clinical information. For each patient, data were collected on age, gender, race, marital status, T stage, N stage, M stage, surgery, chemotherapy, radiation therapy, and follow-up information (overall survival status and time) from the SEER database.

2.1.2. The database for gene expression machine learning (GML) and Integrated Clinical Data and Gene Expression Machine Learning (IML)

The datasets utilized for constructing and validating Gene expression Machine Learning (GML) and Integrated Clinical and gene expression Machine Learning (IML) models were sourced from two primary platforms: The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO). For the TCGA dataset, we downloaded data from the TCGAbiolinks package [15]. Additionally, expression data from GEO datasets, specifically GSE13507 [16], GSE31684 [17], GSE32548 [18], and GSE48276 [19], were downloaded. The expression data and clinical information from these datasets were used to construct and validate models to predict the overall survival status in BCa patients. In the TCGA and GEO datasets, clinical variables (OS, OS Time, Gender, Age, T stage, and N stage) were selected due to their availability and relatively low rates of missing values (less than 30 %). Only samples containing both clinical and gene expression data were retained for further analysis. The data from TCGA and GEO were merged into a single cohort, named the TAGO set. Given the limited number of samples (863 in the TAGO set) and the missing value ratio for the N stage (approximately 22 %), directly deleting all samples with missing values could impair the machine learning prediction ability. Therefore, we imputed the missing values in the TAGO set using the most frequent values observed for each clinical variable.

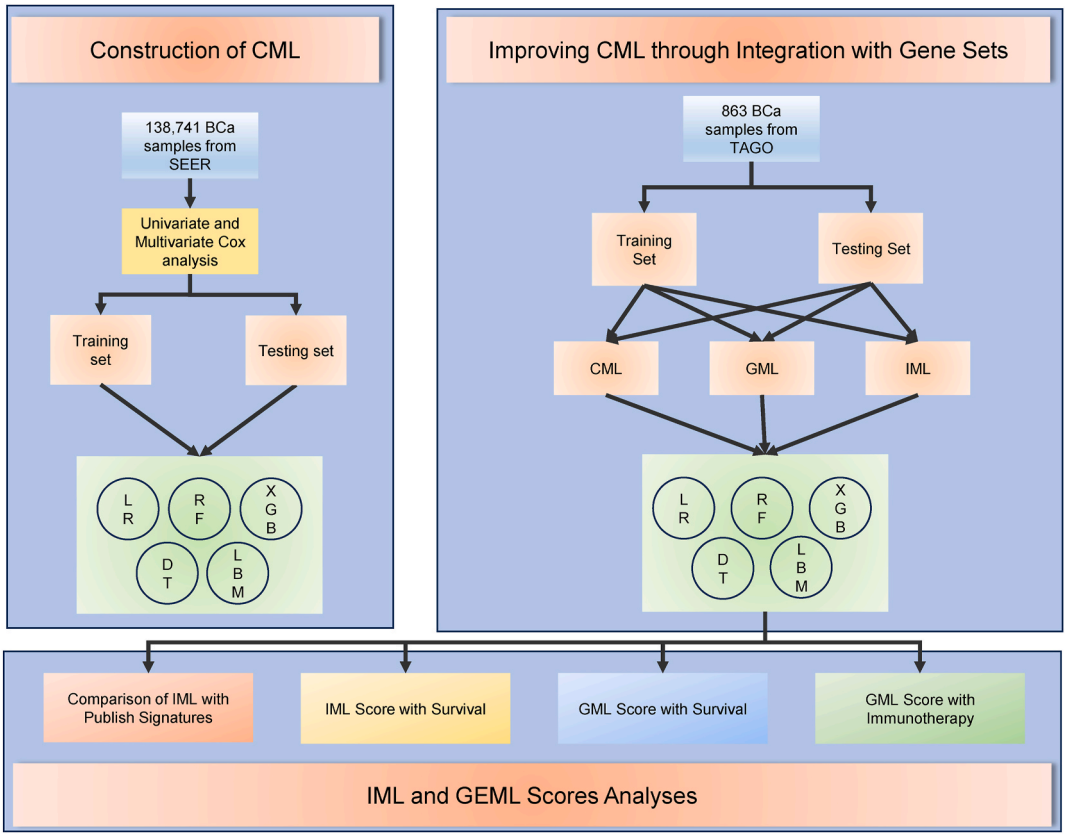


Fig. 1. The workflow of this study. Clinical Data Machine Learning (CML); Surveillance, Epidemiology, and End Results (SEER); Combined cohorts of TCGA and GEO (TAGO); Gene Expression Machine Learning (GML); Integrated Clinical Data and Gene Expression Machine Learning model (IML); logistic regression (LR); random forest (RF); XGBoost (XGB); decision tree (DT); LightGBM (LBM).

Additionally, to further validate our signature, we incorporated an independent cohort of bladder cancer patients receiving immunotherapy from the IMvigor210 trial [20]. This cohort's mRNA and clinical data were obtained from the "IMvigor210CoreBiologies" package.

2.2. Data preprocessing

2.2.1. Preprocessing of data from SEER

In the SEER database, bladder cancer (BCa) patients were divided into two subgroups: younger and older, based on the median age value. Gender contains male and female groups. Racial classification included Non-Hispanic White (NHW) and Others. Marital status was categorized as Married or Unmarried. Surgical options included None, Local Tumor Destruction (LTD), Partial Cystectomy (PC), Radical Cystectomy (RC), and Pelvic Exenteration (PE). To simplify the research, RC and PE patients were grouped into "RC/PE". Chemotherapy and radiation treatments were classified as 'No/Unknown' or 'Yes'. Staging was categorized as follows: T stage into T0 (including "T0", "Ta", "Tis"), T1, T2, T3, and T4; N stage into N0, N1-3 (including N1, N2, and N3); and M stage into M0 and M1. Survival differences across subgroups were analyzed using Kaplan-Meier (KM) survival curves, with p-values calculated by the log-rank test.

The SEER dataset was randomly divided into training (70 %) and testing (30 %) sets. Demographic data for these two sets were presented in supTable 1. Differences between the training and testing sets were assessed. In the comparison tests, the variable "OS Time" which was numerical, was analyzed using a *t*-test. A chi-squared test was used to measure other categorical variables. Univariate Cox regression and multivariate Cox analysis were conducted to identify potential predictive features for model construction. The results were displayed in Table 1. Only the variables with $p < 0.05$ in multivariate Cox analysis were used for model construction.

2.2.2. Preprocessing of data from TAGO

In the preprocessing phase, Gender was categorized into male and female groups. Age was divided into younger and older based on the median value. T stage was classified into T0 (including "T0", "Ta", "Tis"), T1, T2, T3, and T4, while N stage was grouped into N0 and

Table 1
Univariate and Multivariate Cox proportional hazards regression analysis in SEER database.

Characteristics		Univariate Cox		Multivariate Cox	
		HR (95 %)	p-value	HR (95 %)	p-value
Age	Older	Reference	<0.01	Reference	<0.01
	Younger	0.31 (0.31–0.32)		0.32 (0.32–0.33)	
Gender	Female	Reference	0.49	–	–
	Male	0.99 (0.97–1.01)			
Race	NHW	Reference	<0.01	Reference	<0.01
	Others	0.94 (0.92–0.96)		0.91 (0.89–0.93)	
Marital	Married	Reference	<0.01	Reference	<0.01
	Unmarried	1.37 (1.35–1.39)		1.28 (1.26–1.30)	
Surgery	None	Reference	<0.01	Reference	<0.01
	LTD	0.78 (0.76–0.81)	0.33	0.83 (0.80–0.86)	<0.01
	PC	0.96 (0.89–1.04)	<0.01	0.52 (0.48–0.56)	<0.01
	PE/RC	1.18 (1.13–1.23)		0.47 (0.45–0.50)	
Chemotherapy	No/Unknown	Reference	<0.01	Reference	<0.01
	Yes	1.19 (1.17–1.22)		0.83 (0.81–0.84)	
Radiation	No/Unknown	Reference	<0.01	Reference	<0.01
	Yes	3.57 (3.46–3.70)		1.23 (1.18–1.28)	
T stage	T0	Reference	<0.01	Reference	<0.01
	T1	1.56 (1.53–1.59)	<0.01	1.54 (1.51–1.58)	<0.01
	T2	3.06 (2.99–3.13)	<0.01	3.33 (3.24–3.42)	<0.01
	T3	3.22 (3.11–3.34)	<0.01	4.35 (4.15–4.56)	<0.01
	T4	5.82 (5.60–6.05)		5.74 (5.49–6.01)	
N stage	N0	Reference	<0.01	Reference	<0.01
	N1-3	3.73 (3.60–3.86)		1.75 (1.68–1.82)	
M stage	M0	Reference	<0.01	Reference	<0.01
	M1	8.38 (8.06–8.72)		3.45 (3.30–3.61)	

N1-3 (including N1, N2, and N3). The TAGO set was then randomly divided into training (70 %) and testing (30 %) sets. Demographic data for these two sets are presented in SupTable 2. Differences between the training and testing sets were assessed: a *t*-test was conducted on the variable "OS Time," while chi-squared tests were conducted on the other variables.

2.3. Gene set variation analysis (GSVA)

The comprehensive exploration of gene set value was undertaken through GSVA [21], including 50 hallmark gene sets sourced from the MSigDB website [22]. This analysis utilized the GSVA package in R, employing expression profiles from TAGO. Individual samples were scored based on gene sets using GSVA, thereby deriving GSVA scores indicative of the gene set's activity for each sample. The GSVA scores of 50 hallmark gene sets were used in the machine learning model to predict the survival of BCa.

2.4. Machine learning Prepare on clinical and gene set variables

We employed machine learning techniques to enhance the accuracy of models predicting the survival of bladder cancer (BCa) patients. These techniques aimed to predict survival status at 1, 3, and 5-year intervals. Our study focused on overall survival (OS), defined as the time from diagnosis until death from any cause. Patients were categorized into 'alive' or 'deceased' groups based on their survival status at these intervals. Patients with unavailable survival status at specific time points were excluded from the model construction and validation. For example, if a patient died of a heart attack 41 months after BCa diagnosis, the survival status would be recorded as "alive" at 1 year, "alive" at 3 years, and "deceased" at 5 years. Conversely, if another patient was last known to be alive 47 months after BCa diagnosis, the survival status would be recorded as "alive" at 1 year, "alive" at 3 years, and "not available" at 5 years. This patient would be excluded from models predicting 5-year survival.

For the clinical variables available in the SEER and TAGO datasets, we applied the following conversion rules to transform categorical variables into numerical values. Specifically, the Age variable was converted into a binary format: "Younger" was mapped to 0 and "Older" to 1. Gender was coded as 0 for "male" and 1 for "female". The Race variable was simplified, with "NHW" mapped to 0 and "Others" to 1. Marital status was coded as 0 for "Unmarried" and 1 for "Married". Surgery types were encoded as follows: "None" to 0, "LTD" to 1, "PC" to 2, and "RC/PE" to 3. Chemotherapy status was represented as 0 for "No/Unknown" and 1 for "Yes," while radiation treatment was similarly coded as 0 for "None/Unknown" and 1 for "Yes." The T stage variable, representing tumor stages, was encoded with "T0" as 0, "T1" as 1, "T2" as 2, "T3" as 3, and "T4" as 4. The N stage variable, indicating lymph node involvement, was simplified to "N0" as 0 and "N1-3″ as 1. Finally, the M stage variable, denoting metastasis status, was mapped from "M0" to 0 and "M1" to 1.

For the gene set variables, represented by GSVA scores of 50 hallmark gene sets in the TAGO dataset, we used min-max normalization. This technique scales the data to a range between 0 and 1. Min-max normalization works by subtracting the minimum value in the variable from each data point and then dividing by the range (the difference between the maximum and minimum values). This ensures that the minimum value of each gene set becomes 0 and the maximum value becomes 1, with all other values adjusted proportionally in between. This process ensures that no single gene set disproportionately influences the model.

2.5. Machine learning training and validation

To enhance the predictive value of clinical biomarkers, we developed three distinct categories of machine learning models: clinical data machine learning (CML), which utilizes only clinical variables; gene expression machine learning (GML), which employs gene set variables from gene expression profiles; and the integrated clinical data and gene expression machine Learning (IML), which combines both clinical and gene set variables. CML models were developed using datasets from the SEER and TAGO databases. Due to data availability constraints, the models based on GML and the IML were specifically developed using data only from the TAGO database. For the implementation of these models, we utilized several machine learning algorithms, including logistic regression (LR), decision tree (DT), random forest (RF), XGBoost (XGB), and LightGBM (LBM). LR, a generalized linear regression analysis model [23], uses a logistic function (Sigmoid function) to predict the probability of binary outcomes [24]. DT is a non-parametric, supervised learning method for classification and regression [25], forming decision rules from data features. RF builds on the decision tree approach [25], utilizing an ensemble of trees where each tree is constructed from a random subset of data and features, with final predictions based on averaging or majority voting to enhance generalizability and reduce overfitting. XGB [26] and LBM [27] further advance this concept using a gradient-boosting framework that includes regularization to control complexity and prevent overfitting.

Extensive hyperparameter tuning was conducted to optimize the performance of the algorithms using the grid search method, which systematically varies parameters to find the best-performing combination based on model accuracy. The parameter grids were defined as follows: for Logistic Regression, regularization strength (C) values of [0.1, 1, 10]; for Random Forest, number of trees (n_estimators) of 100 and 200 and maximum depth of the tree (max_depth) of 5 and 10; for XGBoost, number of gradient boosted trees (n_estimators) of 100 and 200, learning rates of 0.01 and 0.1, and maximum depth of a tree (max_depth) of 3, 5, and 7; for Decision Tree, maximum depth of a tree (max_depth) of 3, 5, 7, and 10 and minimum number of samples required to split an internal node (min_samples_split) of 2, 5, and 10; for LightGBM, number of leaves in one tree (num_leaves) of 31, 50, and 100, learning rates of 0.01 and 0.1, and number of boosting iterations (n_estimators) of 100 and 200. The tuning process was further enhanced through 5-fold cross-validation, solving model overfitting and ensuring that the hyperparameters generalize well across unseen data. The performance of each model was evaluated based on a suite of metrics in the testing sets, including the Area Under the Curve (AUC), accuracy score, precision score, F1 score, and recall score. These metrics provide a comprehensive view of each model's performance. All models were built and tested using the scikit-learn package [28], a widely used library in the machine learning community for building and deploying models.

2.6. Visualization of importance values of features in models

In the context of explainable machine learning, we utilized SHapley Additive exPlanations (SHAP) values to provide the accurate importance values for each feature in XGBoost model [29]. SHAP values, derived from Shapley values in cooperative game theory, assign an importance value to each feature based on its contribution to the prediction. These values quantify how much each feature shifts the output from the baseline prediction. SHAP is primarily used for tree-based machine learning models. SHAP values were visualized through SHAP summary plots, typically presented as bar graphs.

2.7. Comparison of prognostic signatures

We systematically reviewed 24 prognostic signatures associated with bladder cancer to identify the published bladder cancer signatures and obtained their gene names. These signatures include genes related to platinum resistance [30], tumor microenvironment [31], metabolism [32–35], B cells [36,37], and immune [38], FGFR3 alterations [39], cell cycle [40], autophagy [41,42], ferroptosis [43], focal adhesion [44], TGF-β pathway [45], IFN-γ signaling [46], anoikis [47], CpG methylation [48], aging [49], and pyroptosis [50]. Other signatures were obtained by the statistical analysis [51–53]. Before the model construction, clinical variables were integrated into these signatures. The same model training parameters and the same testing set used in the IML were adopted in these signatures. These steps are crucial to fairly compare the model performance based on different signatures. The ability of all signatures to predict survival at 1, 3, and 5 years was assessed by the AUC value in the testing set. These AUC values were then compared with the AUC value of our IML model.

2.8. The relationship of model score with prognosis

GML and IML scores were generated by the GML and IML models to predict 3-year survival. These scores indicated the likelihood of death for BCa patients 3 years after treatment. Samples were then divided into high and low groups based on the median score value, and survival curves were plotted for overall survival (OS) in these groups. In the TCGA-BLCA cohort, which included patients who underwent surgery and chemotherapy, both GML and IML scores were calculated due to the availability of genomic and clinical data.

In the IMvigor cohort, only GML scores could be predicted, as clinical features such as age and TNM staging were lacking. To evaluate the relationship between GML scores and immunotherapy, we selected the IMvigor210-bladder cohort, which contains complete overall survival information. IMvigor210 was a multicenter, single-arm, phase 2 trial that assessed the efficacy and safety of atezolizumab, a monoclonal antibody targeting the protein PD-L1. Survival curves for high and low GML score groups were plotted to analyze overall survival in the IMvigor210-bladder cohort.

3. Results

3.1. Demographic baseline characteristics of SEER

The process of this study is illustrated in Fig. 1. Utilizing inclusion and exclusion criteria, which are presented in Section 2.1.1., the study comprised 138,741 bladder cancer (BCa) patients. Age was divided into younger and older groups based on the median value (71 years old), with younger patients demonstrating better survival rates, whereas older patients exhibited poorer survival (Fig. 2A). No significant difference in survival rates was observed between genders (Fig. 2B). Statistically significant differences in survival were observed among racial subgroups (Fig. 2C). Marital status impacted survival outcomes, as married BCa patients exhibited a lower mortality risk compared to those who were unmarried (Fig. 2D). Patients undergoing local tumor destruction (LTD) presented the most favorable prognosis compared to other surgical interventions (Fig. 2E). The analysis of chemotherapy and radiation indicated poorer outcomes, as presented in Fig. 2F and **G**, respectively. Lastly, TNM staging, a critical prognostic factor, showed that advanced stages correlated with decreased survival Fig. 2H–J. The 138,741 patients were randomly categorized into the training set (97,121 patients) and validation set (41,620 patients) in a 7:3 ratio. There was no statistically significant difference in each variable between the two sets (supTable 1).

3.2. Prognostic factor analysis

Prognostic factor analysis of the entire cohort, presented in Table 1, revealed significant associations between poorer survival and several factors: increased age, NHW, unmarried status, absence of surgery, lack of chemotherapy, use of radiation treatment, and advanced TNM stages.

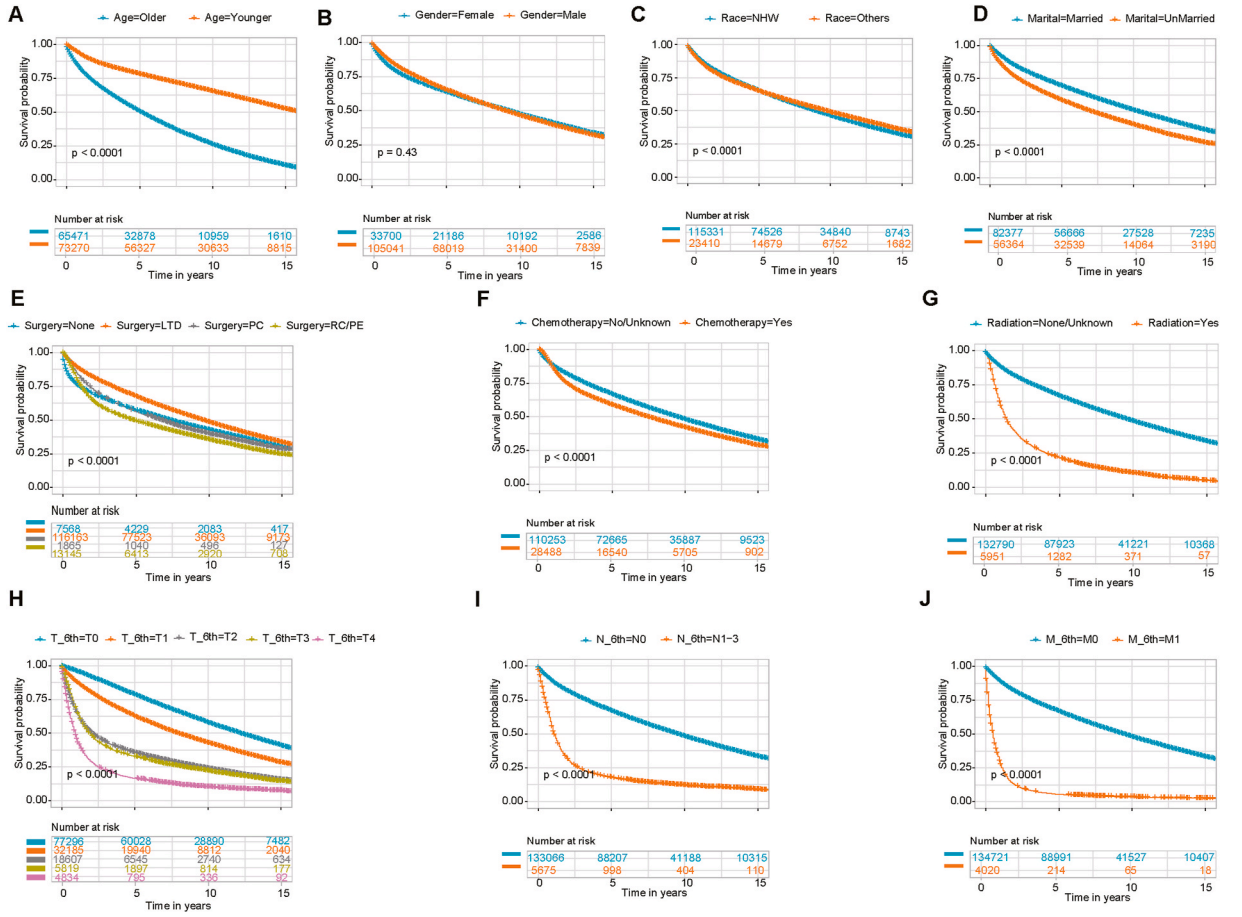


Fig. 2. Kaplan-Meier Survival Analysis Across Subgroups. Kaplan-Meier survival curves for various patient subgroups categorized by age (**A**), gender (**B**), race (**C**), marital status (**D**), surgical intervention (**E**), chemotherapy (**F**), radiation therapy (**G**), and tumor staging—T stage (**H**), N stage (**I**), M stage (**J**). Subgroup specifics include non-Hispanic White (NHW); types of surgical interventions: Local Tumor Destruction (LTD), Partial Cystectomy (PC), Radical Cystectomy (RC), and Pelvic Exenteration (PE); staging according to the AJCC 6th edition for T (T_6th), N (N_6th), and M (M_6th) stages. The p-values were calculated using the log-rank test, a nonparametric hypothesis test designed to compare the survival trends of two or more groups amidst censored observations. For comparisons involving three or more groups, an overall test result p-value is reported. An overall p-value less than 0.05 indicates statistically significant evidence that at least one of the groups differs from the others in terms of survival time.

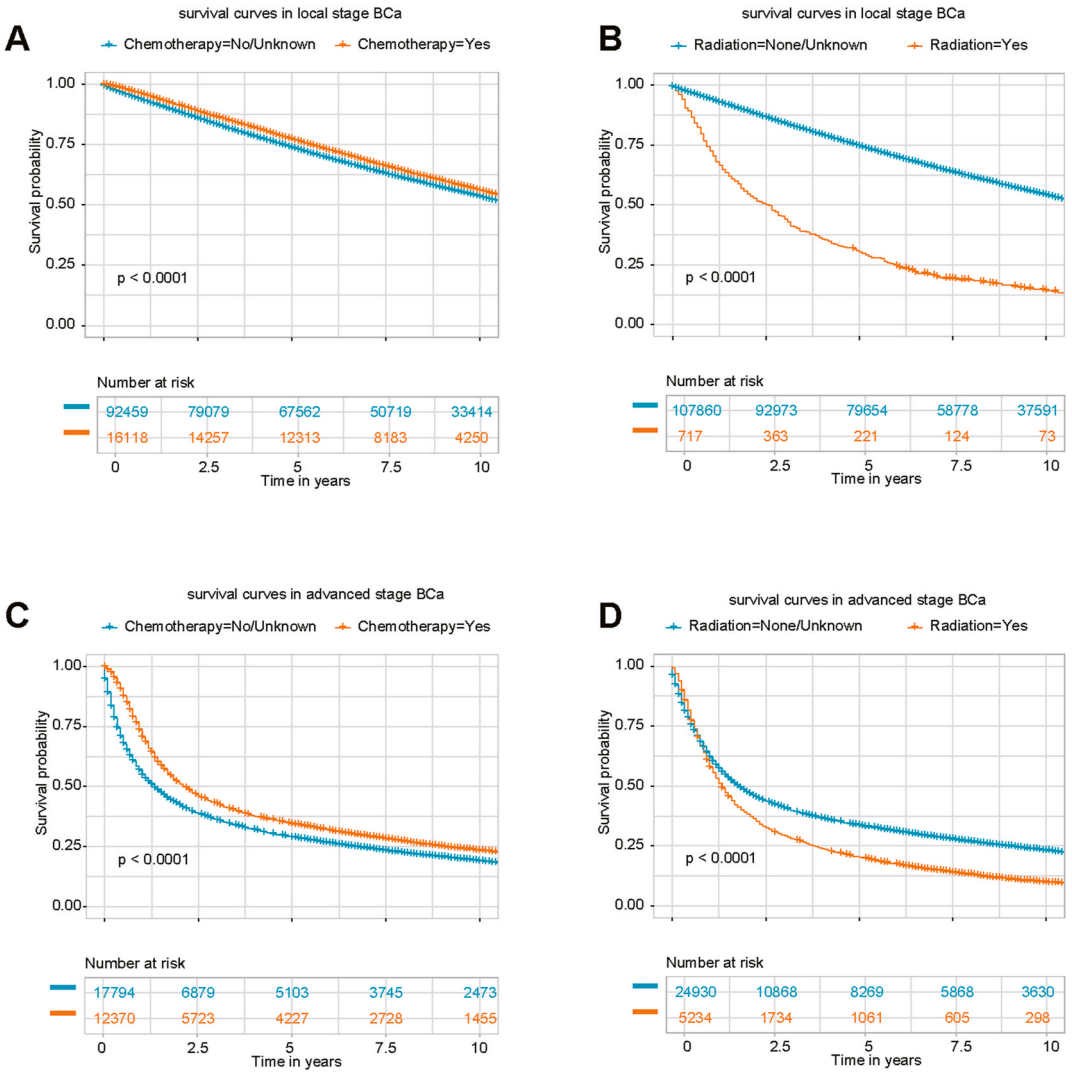


Fig. 3. Treatment Associations with Survival in Different Bladder Cancer (BCa) Stages. (**A**) The impact of chemotherapy on survival in the local stage of BCa. (**B**) The impact of radiation on survival in the local stage of BCa. (**C**) The impact of chemotherapy on survival in the advanced stage of BCa. (**D**) The impact of radiation on survival in the advanced stage of BCa. The p-values were calculated using the log-rank test, a nonparametric hypothesis test designed to compare the survival trends of two or more groups amidst censored observations.

Our survival analysis showed that BCa patients undergoing chemotherapy (Fig. 2F) and radiation (Fig. 2G) tend to exhibit lower survival rates. Additionally, chemotherapy was associated with worse survival in univariate Cox analysis but better survival in multivariate Cox analysis. Conversely, radiation was associated with worse survival outcomes in univariate and multivariate analyses. These findings contradict the common perception that treatments such as chemotherapy and radiation should increase the survival of cancer patients. This apparent contradiction may be because patients with advanced BCa, who inherently have lower survival rates, are more likely to receive chemotherapy and radiation. We stratified BCa patients into two subgroups based on TNM staging: local stage BCa (T0/T1, and N0, and M0) and advanced stage BCa (T2/T3/T4, or N1-3, or M1). In both the local stage and advanced BCa, chemotherapy enhanced survival (Fig. 3A and C). These results support our hypothesis that confounding factors, such as TNM stages, influenced the association of chemotherapy with worse survival in univariate Cox analysis. The use of radiation was linked to poorer survival outcomes in both local BCa (Fig. 3B) and advanced BCa (Fig. 3D). Given the potential effects of confounding factors, additional prospective research that addresses these confounding factors is essential to further investigate the role of radiation in BCa treatment.

3.3. CML Models for Survival Prediction in SEER

Based on the significant prognostic factors identified in the multivariate Cox analysis in Table 1, we selected age, race, marital status, surgery, chemotherapy, radiation, T stage, N stage, and M stage for constructing machine learning models. The prediction targets were the survival status of BCa patients at 1, 3, and 5 years after treatment. The number of alive/deceased samples in the

training sets for these time points were 84,000/11,926, 71,295/24,726, and 62,190/33,591, respectively. The numbers of alive/deceased samples in the testing sets for these time points were 35,891/5,180, 30,341/10,783, and 26,527/14,501, respectively.

After hyperparameter tuning on the training set, the CML models demonstrated excellent predictive performance across various algorithms on the testing set. XGBoost consistently emerged as the best-performing model for predicting 1-, 3-, and 5-year survival in the testing dataset of the SEER database, with AUC values of 0.860, 0.821, and 0.804, respectively (Fig. 4A). The highest accuracy values in the testing dataset for XGBoost were 0.896, 0.814, and 0.756 at these intervals (Fig. 4B). The F1 scores of XGBoost in the testing dataset were 0.449, 0.557, and 0.571 (Fig. 4C). The precision values of XGBoost in the testing dataset were 0.674, 0.742, and 0.751 (Fig. 4D). The recall values of XGBoost in the testing dataset were 0.337, 0.445, and 0.461 (Fig. 4E), respectively. These results underscore the efficacy of machine learning in utilizing clinical variables to accurately predict survival outcomes in bladder cancer patients.

Subsequently, we used SHAP summary plots from the XGBoost algorithm on the testing sets to identify the most influential features (**sup**Fig. 1A–C). The summary plot displays the relative impact of each feature on model predictions, ranking the top features by their mean absolute SHAP values. Our analysis highlighted the T stage as the most critical predictor of survival for 1 and 3 years. For 5-year predictions, the importance of age increased, with age proving to be the most vital factor.

3.4. Comparison of CML models, GML, and IML models in TAGO

In the combined cohorts of TCGA and GEO, referred to as the TAGO cohort, we developed three models for predicting survival: the clinical data machine learning (CML) model, the gene expression data machine learning (GML) model, and the integrated machine learning (IML) model, which combines clinical and gene set features. Using logistic regression, random forest, XGBoost, decision tree, and LightGBM algorithms, we evaluated the models' accuracy for predicting 1-, 3-, and 5-year survival. The number of alive/deceased samples in the training sets for these time points were 467/92, 171/201, and 140/234, respectively. In the testing sets, the numbers of alive/deceased samples for 1-, 3-, and 5-year survival were 198/45, 98/86, and 66/94, respectively. For comparison, we recorded the best performance values for each machine learning algorithm at each time point. The IML outperformed the other models, achieving the highest AUC values in the testing sets at 1, 3, and 5 years, with scores of 0.786, 0.819, and 0.813, respectively, as shown in Fig. 5A. The IML also demonstrated superior accuracy (Fig. 5B) and F1 scores (Fig. 5C) for predicting 3- and 5-year survival compared to CML and GML in the testing sets. These results were further supported by metrics such as precision (Fig. 5D). In terms of Recall value (Fig. 5E), CML and GML performed better than IML. Logistic regression significantly outperformed other machine learning algorithms among the IML models for 1-, 3-, and 5-year survival.

The most influential features from the IML models for predicting 1-, 3-, and 5-year survival were presented in **Sup**Fig. 2A–C. T stages consistently emerged as the most significant predictor of survival across all time frames. Additionally, gene sets such as Unfolded Protein Response (UPR), Epithelial-Mesenchymal Transitions (EMT), and PI3K/Akt/mTOR signaling (PI3KAM) were among the top influential features, highlighting their importance in survival prediction.

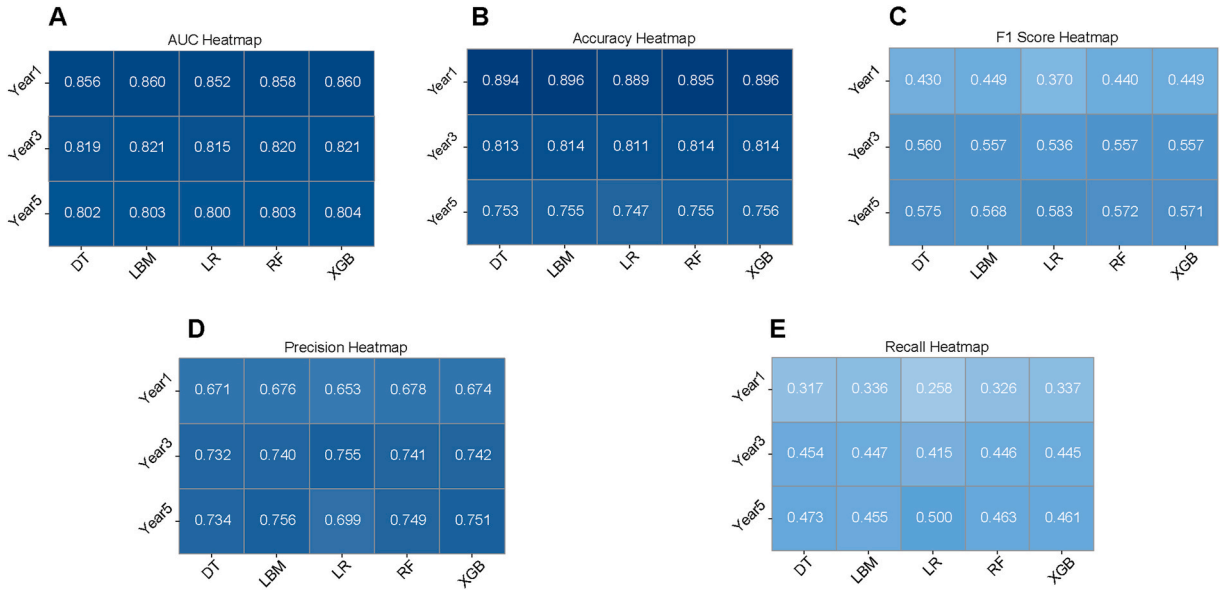


Fig. 4. Clinical Data Machine Learning (CML) Models for Survival Prediction at 1, 3, and 5 Years in the Testing Dataset of SEER Database. (**A**) AUC values for models predicting survival at 1, 3, and 5 years. (**B**) Accuracy values for models predicting survival at 1, 3, and 5 years. (**C**) F1 score values for models predicting survival at 1, 3, and 5 years. (**D**) Precision values for models predicting survival at 1, 3, and 5 years. (**E**) Recall values for models predicting survival at 1, 3, and 5 years. Decision Tree (DT); LightGBM (LBM); Logistic Regression (LR); Random Forest (RF); XGBoost (XGB).

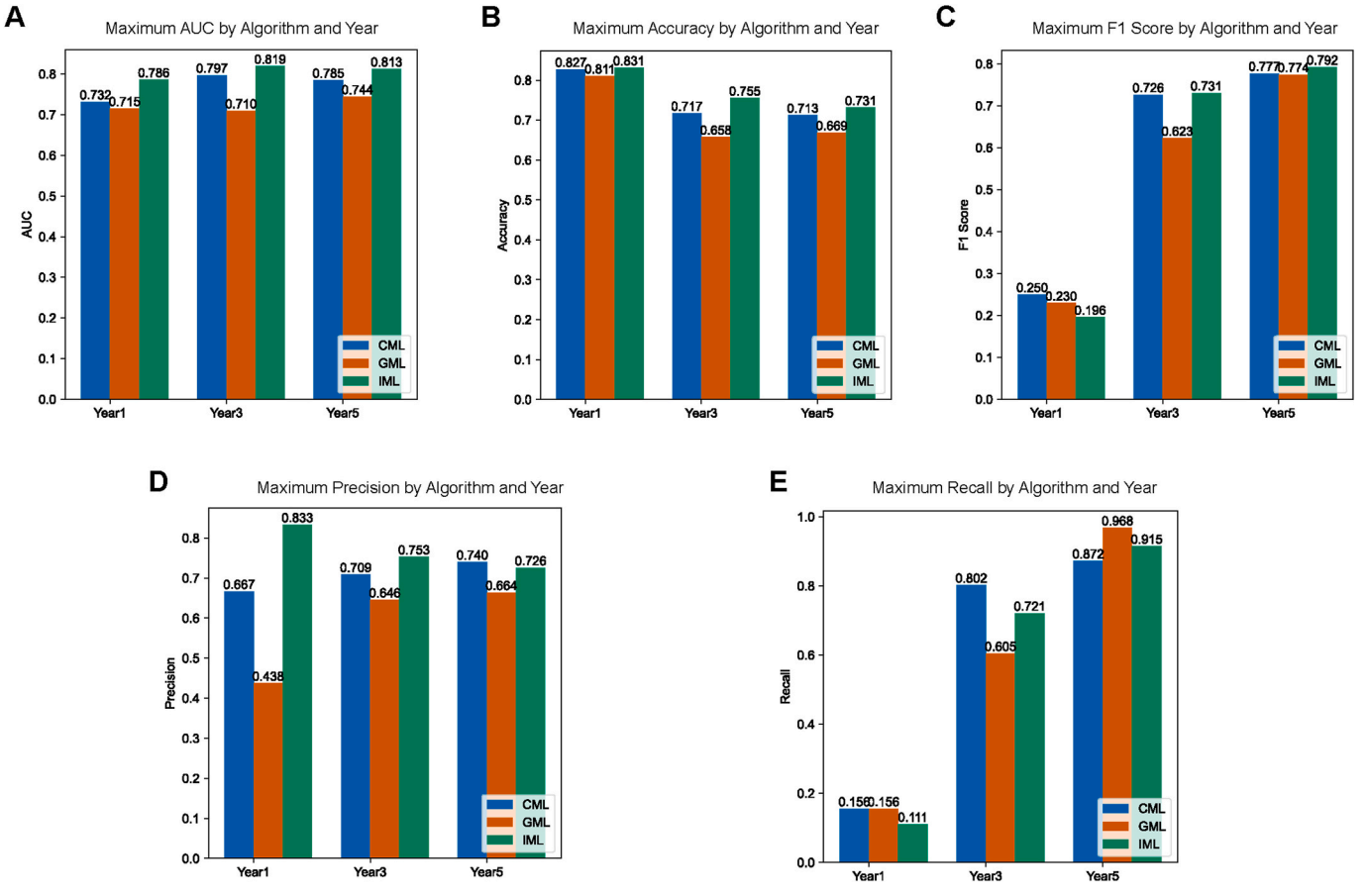


Fig. 5. Comparison of IML, CML, and GML based on AUC (A), accuracy (B), F1 score (C), precision (D), and recall (E) values for models predicting survival at 1, 3, and 5 Years in the testing sets.

3.5. Comparison of IML with published prognostic signatures

To facilitate a rigorous comparison between the IML and other published prognostic signatures, we systematically reviewed BCa literature over the past six years, incorporating 24 distinct signatures into our analysis (SupTable 3). These signatures cover various biological processes, including platinum resistance, tumor microenvironment interactions, metabolic pathways, B cell involvement, FGFR3 alterations, cell cycle dynamics, autophagy, focal adhesion, the TGF-β pathway, IFN-γ signaling, anoikis, CpG methylation, aging, and pyroptosis. We employed the same training and testing datasets used for IML (Fig. 5A), including the same clinical variables and modeling parameters, to ensure consistency across evaluations.

In the 1-year survival predictions, Signature 13 achieved the highest AUC at 0.791 (Fig. 6A). For 3-year predictions, Signature 15 led with an AUC of 0.832 (Fig. 6B) and topped the five-year predictions with an AUC of 0.806 (Fig. 6C). By comparison, the IML signature showed AUC values for 1, 3, and 5-year survival predictions of 0.786, 0.819, and 0.813, respectively (Fig. 5A), making it the best performer in the five-year category and second best for the one-year predictions. Signatures 15, 10, 24, and 13 outperformed our IML signature in the three-year predictions. Upon reviewing the original publications for these signatures, we discovered that they employed TCGA and GEO datasets to identify significant prognostic genes. Notably, the testing sets used in this comparison study also come from TCGA and GEO datasets, which might increase the performance of these signatures. Conversely, our IML signature derives from 50 hallmark gene sets sourced from the MSigDB website, and it does not utilize any selection steps for prognostic gene sets. In summary, our study highlights the importance of independent validation and original datasets in assessing the efficacy of predictive models.

3.6. Relationship of GML and IML scores with prognosis

This section investigated the relationship between IML and GML scores and prognosis. The overall survival (OS) data were sourced from two cohorts: the TCGA-BLCA cohort, comprising BCa patients who underwent surgery and chemotherapy, and the IMvigor cohort, which included BCa patients treated with the PD-L1 antibody. In the TCGA-BLCA cohort, both GML and IML scores were obtained due to the availability of expression data and clinical features. Conversely, in the IMvigor cohort, only GML could be predicted, as clinical features (age and TNM staging) were lacking. We utilized the GML and IML models to predict the scores of samples and estimate the probability of death. The BCa samples were then classified into high and low groups based on the median value of the IML or GML scores. We plotted their survival curves accordingly. The survival curves revealed that patients with high GML scores exhibited a worse prognosis (Fig. 7A). Similarly, patients with high IML scores also showed a poorer prognosis (Fig. 7B). To further evaluate the impact of the GML score on the effectiveness of immunotherapy, we analyzed the IMvigor cohort. This cohort contains the data of patients treated with atezolizumab monotherapy (an anti-PD-L1 antibody). The group with lower scores demonstrated a better prognosis (Fig. 7C), suggesting that the benefits of immunotherapy are more pronounced in this subgroup.

4. Discussion

Bladder cancer (BCa) represents a challenging medical condition with a diverse range of treatment modalities and varied prognostic outcomes. Recent advances in machine learning have led to the development of models that identify numerous risk factors related to tumor mortality, thereby enhancing the personalization of risk prediction. Accurate survival estimation is vital during patient consultations and treatment decisions. Our study used data from the SEER database to develop models that predict survival based on clinical features. However, these clinical variables could not capture the biological variance at the molecular level. To address this limitation, we integrated data from the TCGA and GEO databases to perform an integrated machine learning (IML) analysis combining clinical features and gene set data. The results of our integrated analysis are promising. We achieved AUC values of 0.786, 0.819, and 0.813 for predicting 1-year, 3-year, and 5-year survival, respectively. These performance metrics highlight the robustness of our model across various time intervals and demonstrate its utility in providing clinicians with critical insights into patient prognosis. Using clinical and gene expression data, our model not only accurately predicts outcomes but also enhances the understanding of the underlying biological mechanisms that drive tumor progression and response to treatment.

Our multivariate Cox regression analysis indicates that radiation therapy might negatively affect survival. Radiotherapy is a curative treatment for muscle-invasive bladder cancer, employing high-energy ionizing radiation to induce cancer cell death [54]. The primary objective of radiotherapy is to alleviate urinary symptoms associated with advanced cancer, such as hematuria. However, the prognosis remains challenging, as approximately a quarter of patients receiving palliative bladder radiotherapy either discontinued treatment or died within a month of initiation. According to a systematic review, radiotherapy has no clear benefit after radical surgery in BCa [55]. Another study indicated that radiation therapy patients' five-year OS rates are lower [56]. This adverse effect may be associated with the selective depletion of lymphocytes following radiotherapy. Radiation-induced lymphopenia (RIL), which occurs in approximately 70 % of patients undergoing external beam radiation therapy, has long been documented [57]. Studies have demonstrated a significant detrimental prognostic association between lymphopenia and survival in patients receiving radiation therapy for solid tumors [57,58]. Given these findings, the efficacy and safety of additional radiation therapy in treating BCa warrant further investigation.

The present study successfully established machine learning models to predict the individualized survival probabilities of bladder cancer (BCa) patients at critical time points of 1, 3, and 5 years post-treatment. These time points are crucial as they reflect the short-term, medium-term, and long-term survival outcomes of BCa patients, offering valuable insights into the efficacy of treatment protocols. Studies usually used the nomogram based on the multivariate Cox analysis to predict the survival rate. However, machine learning algorithms have significant advantages in accuracy compared to these statistical models. Our models outperformed the

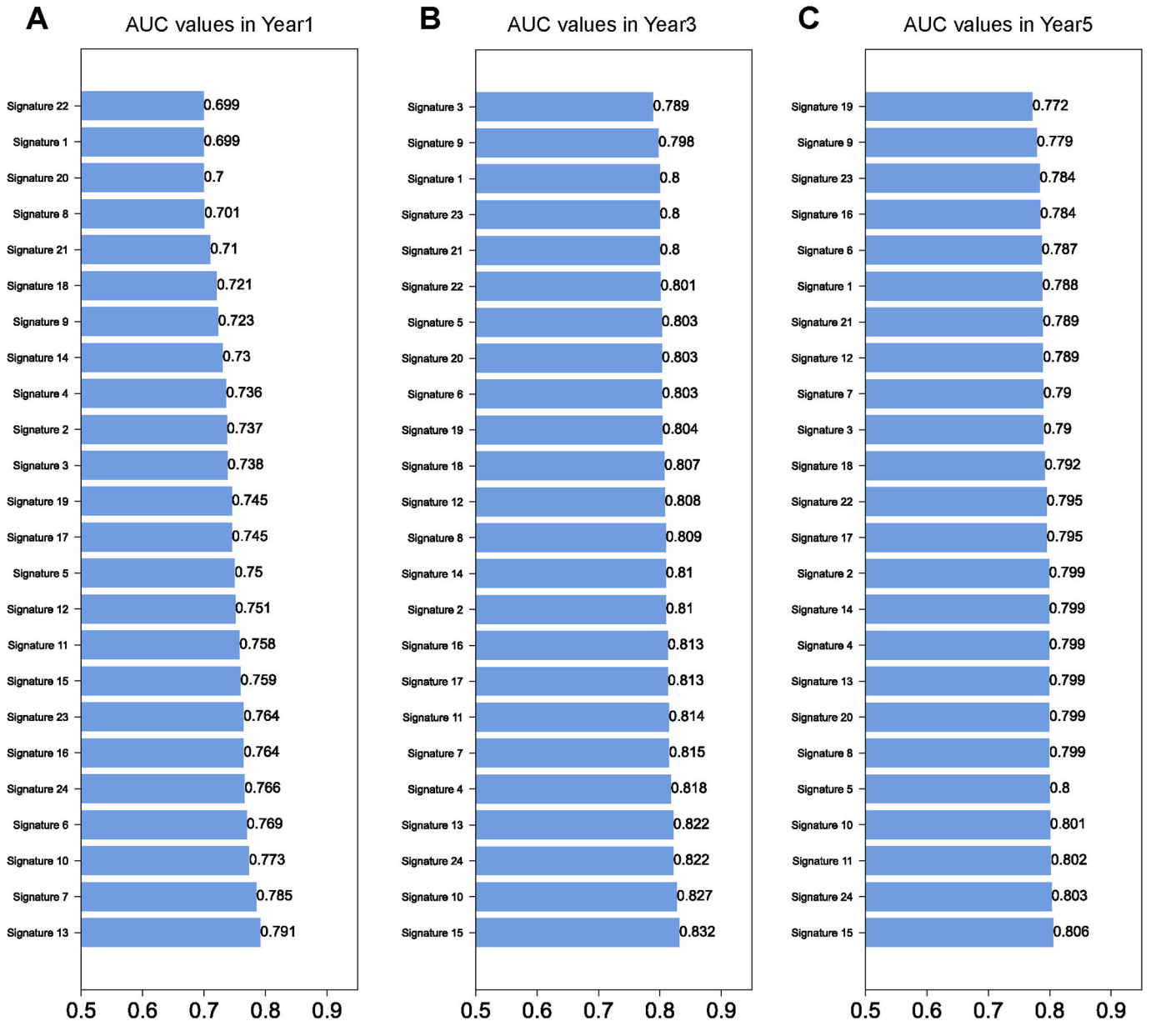


Fig. 6. Performance of published signatures for predicting survival status at 1, 3, and 5 Years in the testing sets.

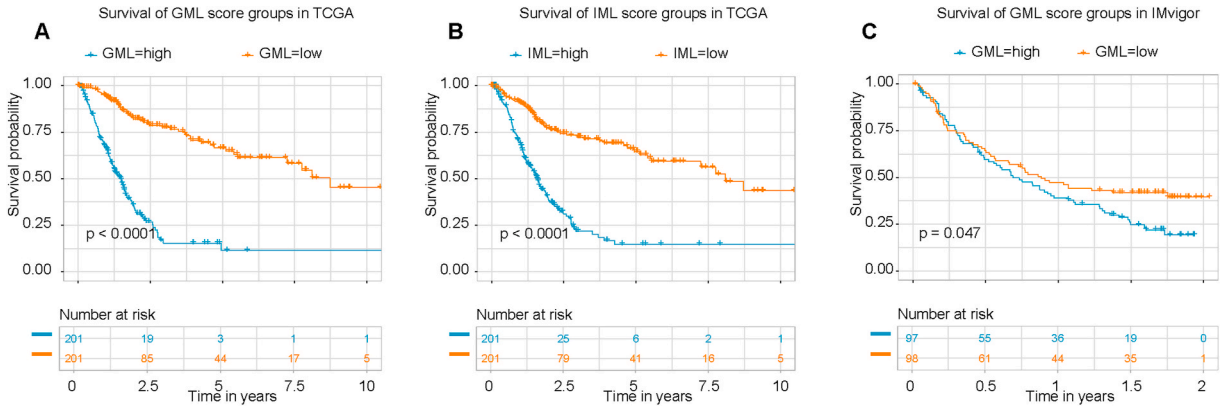


Fig. 7. Survival of GML and IML Scores in Different Cohorts. (**A**) The association of GML scores with overall survival (OS) in TCGA-BLCA cohort. (**B**) The association of IML scores with OS in TCGA-BLCA cohort. (**C**) The association of GML score with OS in the cohort treated with immunotherapy. The p-values were calculated using the log-rank test, a nonparametric hypothesis test designed to compare the survival trends of two or more groups amidst censored observations.

existing published signatures based on our comparative analysis results. Our research developed a model that integrates clinical and gene sets data. To our knowledge, this study is the first to report on machine learning models that utilize clinical data and gene set information to predict BCa survival at these specific time points. Integrating these diverse data types enhances the models' predictive accuracy and reliability, potentially setting a new standard in personalized cancer care.

Recent advances in Immune Checkpoint Blockade (ICB) therapy have demonstrated its potential to enhance survival rates in various metastatic cancers, including BCa. However, a significant subset of BCa patients remains unresponsive to these treatments [6]. Our research revealed that low GML score BCa patients with high survival rates when treated with immunotherapy. This observation underscores the potential of the GML score as a predictive biomarker for immunotherapy responsiveness. The GML score, a novel metric derived from machine learning models, can differentiate between patients likely to benefit from immunotherapy versus those who might not. This finding warrants further investigation to validate the GML score's effectiveness in clinical applications and to explore its implications for broader cancer treatment protocols.

In the TAGO database, the CML model for predicting 1-, 3-, and 5-year survival rates in the testing sets are AUC values of 0.732, 0.797, and 0.785, respectively. Notably, AUC values increased significantly from the 1-year to the 3-year before slightly decreasing at the 5-year. A similar trend was observed in the IML model (AUC values of 0.786, 0.819, and 0.813 at 1, 3, and 5 years) from the same database, presenting an unusual phenomenon where models are more accurate for predicting longer-term rather than shorter-term survival. We propose several reasons for this observation: (1) The ratio of alive/deceased significantly impacts model performance. For example, the training set is more 'balanced' at the 3-year interval (171 alive vs. 201 deceased) compared to the 1-year (467 alive vs. 92 deceased) and 5-year intervals (140 alive vs. 234 deceased). This suggests that the more imbalanced the dataset, the lower the AUC value tends to be. Machine learning models often struggle to learn effectively from imbalanced data [59]. (2) In the TAGO database, only 863 samples have survival follow-up data, and the training and testing sets are randomly divided at a 7:3 ratio. The limited sample size and random division might influence model performance and contribute to this unusual phenomenon.

Our research has several limitations. First, essential clinical data, such as preoperative laboratory results, were unavailable, which may have affected the predictive accuracy of the CML model. Developing a more comprehensive predictive model that includes these elements is advisable. Second, the sample size based on gene set data is limited. Increasing the number of samples could significantly improve these models' robustness and predictive capabilities. Additionally, deep learning techniques have generally outperformed traditional machine learning methods in various predictive applications. Therefore, incorporating deep learning techniques should be considered in future studies. Finally, the role of medical imaging techniques, such as radiomics (e.g., CT scans) and pathomics (e.g., whole slide images), in predicting bladder cancer survival rates warrants further investigation. These imaging assessments could substantially enhance the precision of future prognostic models, offering a more detailed approach to predicting patient outcomes in bladder cancer.

5. Conclusions

To enhance the accuracy of our models by clinical data, gene expression data features were also incorporated alongside these clinical variables. We constructed and validated machine learning models that estimate survival probabilities at 1, 3, and 5 years for BCa patients. This integration allowed us to develop personalized prognosis prediction models for individual patients, facilitating more informed clinical decision-making. It was observed that models incorporating clinical and gene expression data offered more accurate survival estimates than models based solely on clinical or gene expression data. However, prospective studies are needed to perform external validation of these models.

Data availability

All the data in this study, including gene expression profiling data and clinical data supporting this analysis, come from previously published studies and datasets SEER (https://seer.cancer.gov/), TCGA (https://portal.gdc.cancer.gov/) and GEO (https://www.ncbi.nlm.nih.gov/geo/). The patients participating in these studies and datasets have acquired ethical permission. We have uploaded the relevant data and code for the machine learning models training to the GitHub repository, which is accessible at https://github.com/UrologyFxy/machineLeanring.

Funding statement

The research did not receive specific funding.

CRediT authorship contribution statement

Yali Tang: Writing – original draft, Investigation, Conceptualization. **Shitian Li:** Writing – original draft, Conceptualization. **Liang Zhu:** Investigation. **Lei Yao:** Methodology. **Jianlin Li:** Visualization. **Xiaoqi Sun:** Visualization. **Yuan Liu:** Visualization. **Yi Zhang:** Writing – review & editing. **Xinyang Fu:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Not applicable.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e38242.

References

[1] S. Wu, K. Nitschke, T.S. Worst, A. Fierek, C.-A. Weis, M. Eckstein, et al., Long noncoding RNA MIR31HG and its splice variants regulate proliferation and migration: prognostic implications for muscle invasive bladder cancer, J. Exp. Clin. Cancer Res. 39 (2020) 288, https://doi.org/10.1186/s13046-020-01795-5.
[2] H. Jin, X. Ying, B. Que, X. Wang, Y. Chao, H. Zhang, et al., N6-methyladenosine modification of ITGA6 mRNA promotes the development and progression of bladder cancer, EBioMedicine 47 (2019) 195–207, https://doi.org/10.1016/j.ebiom.2019.07.068.
[3] A.R. Zlotta, T. Roumeguere, C. Kuk, S. Alkhateeb, S. Rorive, A. Lemy, et al., Select screening in a specific high-risk population of patients suggests a stage migration toward detection of non–muscle-invasive bladder cancer, Eur. Urol. 59 (2011) 1026–1031, https://doi.org/10.1016/j.eururo.2011.03.027.
[4] K. Hendricksen, W.P.J. Witjes, J.G. Idema, J.J.M. Kums, Trip OB. van Vierssen, M.J.F.M. de Bruin, et al., Comparison of three schedules of intravesical epirubicin in patients with non–muscle-invasive bladder cancer, Eur. Urol. 53 (2008) 984–991, https://doi.org/10.1016/j.eururo.2007.12.033.
[5] T.L. Rose, M.R. Harrison, A.M. Deal, S. Ramalingam, Y.E. Whang, B. Brower, et al., Phase II study of gemcitabine and split-dose cisplatin plus pembrolizumab as neoadjuvant therapy before radical cystectomy in patients with muscle-invasive bladder cancer, J. Clin. Oncol. 39 (2021) 3140–3148, https://doi.org/10.1200/JCO.21.01003.
[6] Y. Xu, H. Zeng, K. Jin, Z. Liu, Y. Zhu, L. Xu, et al., Immunosuppressive tumor-associated macrophages expressing interlukin-10 conferred poor prognosis and therapeutic vulnerability in patients with muscle-invasive bladder cancer, J Immunother Cancer 10 (2022) e003416, https://doi.org/10.1136/jitc-2021-003416.
[7] L.P. Huang, D. Savoly, A.A. Sidi, M.E. Adelson, E. Mordechai, J.P. Trama, CIP2A protein expression in high-grade, high-stage bladder cancer, Cancer Med. 1 (2012) 76–81, https://doi.org/10.1002/cam4.15.
[8] D.-H. Mun, S. Kimura, S.F. Shariat, M. Abufaraj, The impact of gender on oncologic outcomes of bladder cancer, Curr. Opin. Urol. 29 (2019) 279–285, https://doi.org/10.1097/MOU.0000000000000606.
[9] E. Liow, B. Tran, Precision oncology in urothelial cancer, ESMO Open 5 (2020) e000616, https://doi.org/10.1136/esmoopen-2019-000616.
[10] J. Zhu, X. Ye, L. Zhou, Z. He, J. Jin, W. Yu, Treatment decisions of bladder cancer in patients older than 85 years: a SEER-based analysis 2011–2015, Transl Cancer Res TCR 11 (2022) 3584–3592, https://doi.org/10.21037/tcr-22-944.
[11] W. Wang, J. Liu, L. Liu, Development and validation of a prognostic model for predicting overall survival in patients with bladder cancer: a SEER-based study, Front. Oncol. 11 (2021) 692728, https://doi.org/10.3389/fonc.2021.692728.
[12] Z. Zheng, S. Mao, W. Zhang, J. Liu, C. Li, R. Wang, et al., Dysregulation of the immune microenvironment contributes to malignant progression and has prognostic value in bladder cancer, Front. Oncol. 10 (2020) 542492, https://doi.org/10.3389/fonc.2020.542492.
[13] R. Yadollahvandmiandoab, M. Jalalizadeh, K. Buosi, H.A. Garcia-Perdomo, L.O. Reis, Immunogenic cell death role in urothelial cancer therapy, Curr. Oncol. 29 (2022) 6700–6713, https://doi.org/10.3390/curroncol29090526.
[14] B.A. Maiorano, U. De Giorgi, D. Ciardiello, G. Schinzari, A. Cisternino, G. Tortora, et al., Immune-checkpoint inhibitors in advanced bladder cancer: seize the day, Biomedicines 10 (2022) 411, https://doi.org/10.3390/biomedicines10020411.
[15] A. Colaprico, T.C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, et al., TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data, Nucleic Acids Res. 44 (2016) e71, https://doi.org/10.1093/nar/gkv1507.
[16] W.-J. Kim, E.-J. Kim, S.-K. Kim, Y.-J. Kim, Y.-S. Ha, P. Jeong, et al., Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer, Mol. Cancer 9 (2010) 3, https://doi.org/10.1186/1476-4598-9-3.

[17] M. Riester, J.M. Taylor, A. Feifer, T. Koppie, J.E. Rosenberg, R.J. Downey, et al., Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer, Clin. Cancer Res. 18 (2012) 1323–1333, https://doi.org/10.1158/1078-0432.CCR-11-2271.
[18] D. Lindgren, G. Sjödahl, M. Lauss, J. Staaf, G. Chebil, K. Lövgren, et al., Integrated genomic and gene expression profiling identifies two major genomic circuits in urothelial carcinoma, PLoS One 7 (2012) e38863, https://doi.org/10.1371/journal.pone.0038863.
[19] W. Choi, S. Porten, S. Kim, D. Willis, E.R. Plimack, J. Hoffman-Censits, et al., Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy, Cancer Cell 25 (2014) 152–165, https://doi.org/10.1016/j.ccr.2014.01.009.
[20] S. Mariathasan, S.J. Turley, D. Nickles, A. Castiglioni, K. Yuen, Y. Wang, et al., TGF-β attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells, Nature 554 (2018) 544–548, https://doi.org/10.1038/nature25501.
[21] S. Hänzelmann, R. Castelo, J. Guinney, GSVA: gene set variation analysis for microarray and RNA-Seq data, BMC Bioinf. 14 (2013) 7, https://doi.org/10.1186/1471-2105-14-7.
[22] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J.P. Mesirov, Molecular signatures database (MSigDB) 3.0, Bioinformatics 27 (2011) 1739–1740, https://doi.org/10.1093/bioinformatics/btr260.
[23] J.C. Stoltzfus, Logistic regression: a brief primer, Acad. Emerg. Med. 18 (2011) 1099–1104, https://doi.org/10.1111/j.1553-2712.2011.01185.x.
[24] T. Yan, W. Xu, J. Lin, L. Duan, P. Gao, C. Zhang, et al., Combining multi-dimensional convolutional neural network (CNN) with visualization method for detection of Aphis gossypii glover infection in cotton leaves using hyperspectral imaging, Front. Plant Sci. 12 (2021) 604510, https://doi.org/10.3389/fpls.2021.604510.
[25] Y.-C. Wang, D.-J. Tsai, L.-C. Yen, Y.-H. Yao, T.-T. Chiang, C.-H. Chiu, et al., Clinical characteristics of COVID-19 patients and application to an artificial intelligence system for disease surveillance, J. Clin. Med. 11 (2022) 1437, https://doi.org/10.3390/jcm11051437.
[26] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016) 785–794, https://doi.org/10.1145/2939672.2939785.
[27] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., LightGBM: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[29] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
[30] Y. Hao, C. Wang, D. Xu, Identification and validation of a novel prognostic model based on platinum resistance-related genes in bladder cancer, Int. Braz J. Urol. 49 (2023) 61–88, https://doi.org/10.1590/s1677-5538.ibju.2022.0373.
[31] Z. Wang, L. Tu, M. Chen, S. Tong, Identification of a tumor microenvironment-related seven-gene signature for predicting prognosis in bladder cancer, BMC Cancer 21 (2021) 692, https://doi.org/10.1186/s12885-021-08447-7.
[32] X. Li, S. Fu, Y. Huang, T. Luan, H. Wang, J. Wang, Identification of a novel metabolism-related gene signature associated with the survival of bladder cancer, BMC Cancer 21 (2021) 1267, https://doi.org/10.1186/s12885-021-09006-w.
[33] X. Liu, C. Chen, P. Xu, B. Chen, A. Xu, C. Liu, Development and experimental validation of a folate metabolism-related gene signature to predict the prognosis and immunotherapeutic sensitivity in bladder cancer, Funct. Integr. Genomics 23 (2023) 291, https://doi.org/10.1007/s10142-023-01205-x.
[34] C. Huang, Y. Li, Q. Ling, C. Wei, B. Fang, X. Mao, et al., Establishment of a risk score model for bladder urothelial carcinoma based on energy metabolism-related genes and their relationships with immune infiltration, FEBS Open Bio 13 (2023) 736–750, https://doi.org/10.1002/2211-5463.13580.
[35] L. Wei, L. Ji, S. Han, M. Xu, X. Yang, Construction and validation of a prognostic model of metabolism-related genes driven by somatic mutation in bladder cancer, Front Biosci (Landmark Ed) 28 (2023) 242, https://doi.org/10.31083/j.fbl2810242.
[36] J. Zhou, R. Zhou, Y. Zhu, S. Deng, B. Muhuitijiang, C. Li, et al., Investigating the impact of regulatory B cells and regulatory B cell-related genes on bladder cancer progression and immunotherapeutic sensitivity, J. Exp. Clin. Cancer Res. 43 (2024) 101, https://doi.org/10.1186/s13046-024-03017-8.
[37] R. Zhou, J. Zhou, B. Muhuitijiang, X. Zeng, W. Tan, Construction and experimental validation of a B cell-related gene signature to predict the prognosis and immunotherapeutic sensitivity in bladder cancer, Aging (2023), https://doi.org/10.18632/aging.204753.
[38] G. Qu, Z. Liu, G. Yang, Y. Xu, M. Xiang, C. Tang, Development of a prognostic index and screening of prognosis related genes based on an immunogenomic landscape analysis of bladder cancer, Aging (Albany, NY) 13 (2021) 12099–12112. https://doi.org/10.18632/aging.202917.
[39] T. Xu, W. Xu, Y. Zheng, X. Li, H. Cai, Z. Xu, et al., Comprehensive FGFR3 alteration-related transcriptomic characterization is involved in immune infiltration and correlated with prognosis and immunotherapy response of bladder cancer, Front. Immunol. 13 (2022) 931906, https://doi.org/10.3389/fimmu.2022.931906.
[40] W.-W. Shi, J.-Z. Guan, Y.-P. Long, Q. Song, Q. Xiong, B.-Y. Qin, et al., Integrative transcriptional characterization of cell cycle checkpoint genes promotes clinical management and precision medicine in bladder carcinoma, Front. Oncol. 12 (2022) 915662, https://doi.org/10.3389/fonc.2022.915662.
[41] R. Cao, B. Ma, G. Wang, Y. Xiong, Y. Tian, L. Yuan, Identification of autophagy-related genes signature predicts chemotherapeutic and immunotherapeutic efficiency in bladder cancer (BLCA), J Cellular Molecular Medi 25 (2021) 5417–5433, https://doi.org/10.1111/jcmm.16552.
[42] C. Shen, Y. Yan, S. Yang, Z. Wang, Z. Wu, Z. Li, et al., Construction and validation of a bladder cancer risk model based on autophagy-related genes, Funct. Integr. Genomics 23 (2023) 46, https://doi.org/10.1007/s10142-022-00957-2.
[43] S. Liu, J. Zhai, D. Li, Y. Peng, Y. Wang, B. Dai, Identification and validation of molecular subtypes' characteristics in bladder urothelial carcinoma based on autophagy-dependent ferroptosis, Heliyon 9 (2023) e21092, https://doi.org/10.1016/j.heliyon.2023.e21092.
[44] J. Hu, L. Wang, L. Li, Y. Wang, J. Bi, A novel focal adhesion-related risk model predicts prognosis of bladder cancer —— a bioinformatic study based on TCGA and GEO database, BMC Cancer 22 (2022) 1158, https://doi.org/10.1186/s12885-022-10264-5.
[45] Z. Liu, T. Qi, X. Li, Y. Yao, B. Othmane, J. Chen, et al., A novel TGF-β risk score predicts the clinical outcomes and tumour microenvironment phenotypes in bladder cancer, Front. Immunol. 12 (2021) 791924, https://doi.org/10.3389/fimmu.2021.791924.
[46] H. Deng, D. Deng, T. Qi, Z. Liu, L. Wu, J. Yuan, An IFN-γ-related signature predicts prognosis and immunotherapy response in bladder cancer: results from real-world cohorts, Front. Genet. 13 (2023) 1100317, https://doi.org/10.3389/fgene.2022.1100317.
[47] S. Zhu, Q. Zhao, Y. Fan, C. Tang, Development of a prognostic model to predict BLCA based on anoikis-related gene signature: preliminary findings, BMC Urol. 23 (2023) 199, https://doi.org/10.1186/s12894-023-01382-8.
[48] Y. Guo, J. Yin, Y. Dai, Y. Guan, P. Chen, Y. Chen, et al., A novel CpG methylation risk indicator for predicting prognosis in bladder cancer, Front. Cell Dev. Biol. 9 (2021) 642650, https://doi.org/10.3389/fcell.2021.642650.
[49] D. Wang, H. Ning, H. Wu, Y. Song, Y. Chu, F. Liu, et al., Construction and evaluation of a novel prognostic risk model of aging-related genes in bladder cancer, Curr. Urol. 17 (2023) 236–245, https://doi.org/10.1097/CU9.0000000000000218.
[50] J. Zhao, C. Wu, Y. Wang, M. Li, Y. Jiang, Y. Luo, Identification of a pyroptosis related gene signature for predicting prognosis and estimating tumor immune microenvironment in bladder cancer, Transl Cancer Res TCR 11 (2022) 1865–1879, https://doi.org/10.21037/tcr-22-177.
[51] J. Li, J. Cao, P. Li, Z. Yao, R. Deng, L. Ying, et al., Construction of a novel mRNA-signature prediction model for prognosis of bladder cancer based on a statistical analysis, BMC Cancer 21 (2021) 858, https://doi.org/10.1186/s12885-021-08611-z.
[52] F. Tang, Z. Li, Y. Lai, Z. Lu, H. Lei, C. He, et al., A 7-gene signature predicts the prognosis of patients with bladder cancer, BMC Urol. 22 (2022) 8, https://doi.org/10.1186/s12894-022-00955-3.
[53] J. Chu, N. Li, F. Li, A risk score staging system based on the expression of seven genes predicts the outcome of bladder cancer, Oncol. Lett. (2018), https://doi.org/10.3892/ol.2018.8904.
[54] S. Ashley, A. Choudhury, P. Hoskin, Y. Song, P. Maitre, Radiotherapy in metastatic bladder cancer, World J. Urol. 42 (2024) 47, https://doi.org/10.1007/s00345-023-04744-x.

[55] T. Iwata, S. Kimura, M. Abufaraj, F. Janisch, P.I. Karakiewicz, V. Seebacher, et al., The role of adjuvant radiotherapy after surgery for upper and lower urinary tract urothelial carcinoma: a systematic review, Urol. Oncol.: Seminars and Original Investigations 37 (2019) 659–671, https://doi.org/10.1016/j.urolonc.2019.05.021.
[56] Y. Yamamoto, A. Kawashima, T. Morishima, T. Uemura, A. Yamamoto, G. Yamamichi, et al., Comparative effectiveness of radiation versus radical cystectomy for localized muscle-invasive bladder cancer, Advances in Radiation Oncology 8 (2023), https://doi.org/10.1016/j.adro.2022.101157.
[57] H. Paganetti, A review on lymphocyte radiosensitivity and its impact on radiotherapy, Front. Oncol. 13 (2023), https://doi.org/10.3389/fonc.2023.1201500.
[58] P.J.J. Damen, T.E. Kroese, R van Hillegersberg, E. Schuit, M. Peters, J.J.C. Verhoeff, et al., The influence of severe radiation-induced lymphopenia on overall survival in solid tumors: a systematic review and meta-analysis, Int. J. Radiat. Oncol. Biol. Phys. 111 (2021) 936–948, https://doi.org/10.1016/j.ijrobp.2021.07.1695.
[59] M. Zheng, F. Wang, X. Hu, Y. Miao, H. Cao, M. Tang, A method for analyzing the performance impact of imbalanced binary data on machine learning models, Axioms 11 (2022) 607, https://doi.org/10.3390/axioms11110607.