cancers

MDPI

Article

Gene Screening for Prognosis of Non-Muscle-Invasive Bladder Carcinoma under Competing Risks Endpoints

Chenlu Ke [1], Dipankar Bandyopadhyay [2,*] and Devanand Sarkar [3]

[1] Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VA 23284, USA
[2] Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23219, USA
[3] Department of Human Genetics, Virginia Commonwealth University, Richmond, VA 23219, USA
* Correspondence: dbandyop@vcu.edu; Tel.: +1-804-827-2058

Simple Summary: A vital task in contemporary cancer research is to discover clinically useful molecular markers for diagnosis and prognosis from microarray or sequencing data. However, reliable and efficient statistical tools are lacking in terms of marker screening and selection for high-throughput data with complicated survival endpoints, such as competing risks. Motivated by a study on progression of non-muscle invasive bladder carcinoma for 300 subjects with competing risk endpoints, this paper proposed a controlled screening procedure to fast eliminate most of irrelevant markers, before more precise selection can be further pursued. Combining screening with a boosting procedure, a significant six-gene signature for progression was identified subsequently, showing improved prediction performance over existing alternatives at a lower computational cost. The proposed method is readily applicable to other types of high-throughput cancer data with competing risk events, providing a desired addition to a biomedical researcher's toolbox.

Abstract: Background: Discovering clinically useful molecular markers for predicting the survival of patients diagnosed with non–muscle-invasive bladder cancer can provide insights into cancer dynamics and improve treatment outcomes. However, the presence of competing risks (CR) endpoints complicates the estimation and inferential framework. There is also a lack of statistical analysis tools and software for coping with the high-throughput nature of these data, in terms of marker screening and selection. Aims: To propose a gene screening procedure for proportional subdistribution hazards regression under a CR framework, and illustrate its application in using molecular profiling to predict survival for non-muscle invasive bladder carcinoma. Methods: Tumors from 300 patients diagnosed with bladder cancer were analyzed for genomic abnormalities while controlling for clinically important covariates. Genes with expression patterns that were associated with survival were identified through a screening procedure based on proportional subdistribution hazards regression. A molecular predictor of risk was constructed and examined for prediction accuracy. Results: A six-gene signature was found to be a significant predictor associated with survival of non–muscle-invasive bladder cancer, subject to competing risks after adjusting for age, gender, reevaluated WHO grade, stage and BCG/MMC treatment (p-value < 0.001). Conclusion: The proposed gene screening procedure can be used to discover molecular determinants of survival for non–muscle-invasive bladder cancer and in general facilitate high-throughput competing risks data analysis with easy implementation.

Keywords: bladder cancer; competing risk endpoints; gene screening; subdistribution hazards; survival analysis

Citation: Ke, C.; Bandyopadhyay, D.; Sarkar, D. Gene Screening for Prognosis of Non-Muscle-Invasive Bladder Carcinoma under Competing Risks Endpoints. *Cancers* **2023**, *15*, 379. https://doi.org/10.3390/cancers15020379

Academic Editor: Carlos S. Moreno

Received: 17 October 2022
Revised: 25 December 2022
Accepted: 27 December 2022
Published: 6 January 2023

1. Introduction

Bladder cancer is a common type of cancer associated with high morbidity and mortality [1] rates, if not treated optimally. The disease presents in two different forms:

non–muscle-invasive (NMI) tumors (stages Tis, Ta and T1), where the cancer is in its early stages, with cells only appearing in the inner lining of the bladder (and have not grown into the deeper bladder muscle layers), and muscle-invasive cancers (stages T2–T4), where cancer cells have spread into the detrusor muscle of the bladder. A variety of treatment protocols exist, stratified on the degree of risk of the disease, such as complete resection of the tumor followed by induction, and maintenance immunotherapy through intravesical BCG vaccine, chemotherapy, transurethral resection, and cystectomy. Most bladder cancers are diagnosed at an early stage, when the cancer is highly treatable. The non–muscle-invasive tumors account for roughly 75% of newly diagnosed cases and 50% of non-muscle-invasive bladder cancers (NMIBC) are low grade [2]. However, even early-stage bladder cancers can recur after successful treatment; more than 60% of patients experience recurrence, and some patients develop muscle-invasive tumors over time [3]. Therefore, early diagnosis with personalized treatment and follow-up is the key to a successful outcome.

Developments in microarray and sequencing technologies have allowed the collection of massive genomic information that substantially advances the understanding of molecular mechanisms, biomarker discovery, and personalized medicine. High-throughput data produced by those techniques are characterized by a large number of features that far exceeds the sample size ($p >> n$). In clinical studies, a vital research task is to find predictive features for survival outcomes and build prognostic models for cancer patients, which often requires techniques that were developed for specific time-to-event responses. In bladder cancer, one primary endpoint is time-to-progression, but competing events such as death from non-cancer causes can also be observed [4]. The proportional subdistribution hazard (PSH) model proposed by Fine and Gray [5] has become a popular semi-parametric model for competing risks (CR) data. On the basis of the PSH model, feature selection approaches have been developed for high dimensional data ($p > n$) including LASSO-type methods [6–10] and component-wise boosting [11], among others. These techniques have been used in identifying gene signatures for predicting bladder cancer progression [6,12]. However, when it comes to ultrahigh dimensional data ($p >> n$), exact feature selection is presumably very challenging, if not impossible, to achieve. The aforementioned methods may become statistical inaccurate and computationally expensive [13].

Recent years have seen an increasing attention to feature screening as a preliminary procedure to fast filter out unimportant features before using penalized or boosting methods for more precise selection. Feature screening was initially [13] introduced for a linear model through marginal independence learning based on the Pearson correlation. The screening mechanism asymptotically almost surely identifies all important predictors, and thus is called "sure independence screening" (SIS). Since in many applications researchers know from previous investigations that certain features are responsible for the outcomes or should be controlled for in the studies, conditional SIS [14] (CSIS) that allows multiple covariates to be adjusted for was also proposed for linear models. Although SIS and CSIS have been substantially extended to handle different types of data including right-censored survival outcomes [15], not much attention has been attracted to the competing risks data.

In this paper, we propose a conditional gene screening procedure for the PSH model, controlling for important clinical covariates. The procedure can be combined with available stepwise selection [16], penalized or boosting methods to identify a short list of influential genes and build an interpretable prognostic model for subsequent inference. Although the goal of this paper is variable screening, predictive models were built to evaluate the selected variables, and to compare our proposal with existing selection methods. The screening step eases the computational burden of the penalized or boosting approaches, leading to an enhancement of their performance. We were particularly interested in applying the proposed procedure to discover a predictive gene signature for progression in early-stage bladder cancer. We performed screening followed by boosting and validated the prognostic value of the resulting gene signature after adjusting for the effect of some clinical covariates. As will be shown, this screening-and-selection paradigm has computational and statistical

advantages over classical selection tools for high-throughput CR data. The proposed procedure can be easily implemented with available `R` packages for CR regression, and readily applied to other cancer datasets.

2. Materials and Methods

2.1. The Proportional Subdistribution Hazards Model

Let T and C be the failure and censoring times, and $\epsilon \in \{1, \ldots, K\}$ be the cause of failure. Let $\mathbf{X} \in \mathbb{R}^p$ denote the vector of p covariates subject to selection and $\mathbf{Z} \in \mathbb{R}^{p_0}$ denote the vector of p_0 covariates to be controlled for in the analysis. For typical right-censored data, we observe $Y = \min(T, C)$ and $\delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Our goal is to model the cumulative incidence function (CIF) for failure from the cause of interest ($\epsilon = 1$) conditional on the covariates:

$$F_1(t; \mathbf{X}, \mathbf{Z}) = Pr(T \leq t, \epsilon = 1 | \mathbf{X}, \mathbf{Z}),$$

i.e., the probability of experiencing event 1 before time t and before the occurrence of any other types of event. The subdistribution hazard [17] associated with event 1 is defined as

$$\begin{aligned} \lambda_1(t; \mathbf{X}, \mathbf{Z}) &= \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t, \epsilon = 1 | T \geq t \cup (T \leq t \cap \epsilon \neq 1), \mathbf{X}, \mathbf{Z})}{\Delta t} \\ &= -\frac{d \log(1 - F_1(t; \mathbf{X}, \mathbf{Z}))}{dt} \end{aligned}$$

which measures the instantaneous risk of failure from event 1 for patients who have not yet experienced the event. Note that this risk set includes those who are currently event free as well as who have previously experienced a competing event. The subdistribution hazard for event 1 is assumed to follow a proportional hazard model [5]

$$\lambda_1(t; \mathbf{X}, \mathbf{Z}) = \lambda_{1,0}(t) \exp(\boldsymbol{\beta}^T \mathbf{X} + \boldsymbol{\gamma}^T \mathbf{Z}),$$

where $\lambda_{1,0}(t)$ is an unspecified baseline hazard, and $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\gamma} \in \mathbb{R}^{p_0}$ are regression coefficients. Given a finite sample $\{\mathbf{X}_i, \mathbf{Z}_i, Y_i, \delta_i, \epsilon_i\}_{i=1}^n$, the coefficients can be estimated by maximizing the log partial likelihood function

$$l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^n \int_0^\infty \{\boldsymbol{\beta}^T \mathbf{x}_i + \boldsymbol{\gamma}^T \mathbf{z}_i - \log \sum_{j=1}^n w_j(t) R_j(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_j + \boldsymbol{\gamma}^T \mathbf{z}_j)\} w_i(t) dN_i(t),$$

where $N_i(t) = I(Y_i \leq t, \epsilon_i = 1)$, $R_i(t) = 1 - N_i(t-)$, and $w_i(t) = I(C_i \geq Y_i \wedge t)\hat{G}(t)/\hat{G}(Y_i \wedge t)$ are weights to account for censoring with $\hat{G}(t)$ being the Kaplan–Meier estimate for the censoring time $G(t) = Pr(C \geq t)$. Denote the maximizer by $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \arg\max_{\beta,\gamma} l_n(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Having obtained the estimated regression coefficients, the estimated CIF is obtained by

$$\hat{F}_1(t) = 1 - \exp(-\hat{H}_{1,0}(t) \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_j + \hat{\boldsymbol{\gamma}}^T \mathbf{z}_j)),$$

where $\hat{H}_{1,0}(t)$ is the Breslow estimator of the cumulative baseline subdistribution hazard. The prediction error of the estimated CIF can then be calculated as

$$Err(t) = \frac{1}{n} \sum_{i=1}^n (N_i(t) - \hat{F}_1(t))^2 W_i(t)$$

after accounting for censoring, which is often used to evaluate the performance of the fitted PSH model. Here, $W_i(t) = \frac{I(Y_i \leq t) I(Y_i \leq C)}{\hat{G}(Y_i - | \mathbf{X}_i)} + \frac{I(Y_i > t)}{\hat{G}(t | \mathbf{X}_i)}$ is the inverse probability of censoring weights. The estimation, prediction and evaluation of the PSH model can be achieved via the R packages `riskRegression` and `pec`. However, the partial likelihood estimation is no

longer applicable when $p + p_0 > n$, demanding new techniques to be developed for high dimensional settings.

2.2. Conditional Sure Independence Screening for PSH

For high dimensional data, we realistically assume that the true parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is sparse. In other words, the subset

$$\mathbf{X}_{\mathcal{A}} = \{X_j : \beta_j \neq 0, j = 1, \ldots, p\}$$

is small. Our aim is therefore to identify the active subset $\mathbf{X}_{\mathcal{A}}$ and estimate $\boldsymbol{\beta}$. Approaches have been developed to obtain a sparse model by maximizing a penalized likelihood function [6–10]. Component-wise boosting [11] is an alternative way to obtain parsimonious model fits, and has been adapted to the PSH setting [12]. It uses a stepwise procedure that allows us to build up an overall model from many simple fits, and in each boosting step, the coefficient for one predictor is updated and the overall fit is refined. Boosting and LASSO-like penalized methods are known to have deep connections [11], and they produce similar sparse solutions. However, the aforementioned methods may become statistically inaccurate and computationally expensive when $p + p_0 >> n$ [13]. In the next section, we introduce a feature screening procedure for the PSH model to address the issue.

For each $j = 1, \ldots, p$, consider the PSH model containing an individual predictor X_j in addition to the control variables

$$\lambda_1(t; X_j, \mathbf{Z}) = \lambda_{1,0}(t) \exp(\beta_j X_j + \boldsymbol{\gamma}^T \mathbf{Z}),$$

and let

$$\hat{u}_j = \max_{\beta_j, \boldsymbol{\gamma}} l_n(\beta_j, \boldsymbol{\gamma})$$

denote its estimated maximum log partial likelihood. Then, $\hat{u}_j$ can be regarded as the change in likelihood associated with X_j given $\mathbf{Z}$, after ignoring the common constant $\max_{\boldsymbol{\gamma}} l_n(\boldsymbol{\gamma})$. That is, $\hat{u}_j$ measures the marginal contribution of X_j to the survival outcome after adjusting for the effect of $\mathbf{Z}$, and a large value suggests $\beta_j \neq 0$. We therefore propose to recruit the variables

$$\widehat{\mathbf{X}}_{\mathcal{A}} = \{X_j : \hat{u}_j \text{ is among the first } d \text{ largest of all}\}$$

for a pre-specified model size d. We henceforth refer to the above procedure as conditional sure independence screening for PSH model, or PSH-CSIS for short. According to the sure screening property [13], the choice of d can be relatively generous to ensure that all the important predictors are preserved with high probability. Conventional choices of d are $[n/\log(n)]$, $2[n/\log(n)]$, $3[n/\log(n)]$, and $n - p_0 - 1$ [13,18]. Once the dataset is sufficiently downsized by PSH-CSIS, existing lower dimensional methods can be used afterwards for more precise variable selection and statistical inference (Figure 1).

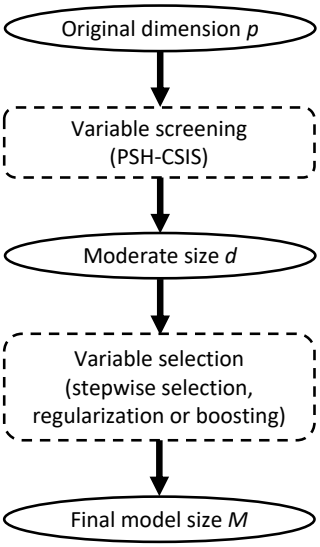

Figure 1. Overall diagram of variable screening and selection.

2.3. Non-Muscle-Invasive Bladder Carcinoma Data

Gene expression data and clinical data for patients with NMI bladder carcinoma were acquired from GEO database (accession number GSE5479 [19]). In total, 300 patients with complete information on 1381 microarray features and 5 important clinical covariates (age, sex, reevaluated WHO grade, reevaluated pathological disease stage and BCG/MMC treatment) were included for analysis. Table 1 summarizes the 5 clinical covariates. The primary endpoint, the time to progression or death from bladder cancer, was observed for 83 patients. Besides, 33 patients died from other/unknown causes and 184 patients were censored during follow-up. The progression-free survival time ranges from 0 to 185 months, with a median of 47 months.

Table 1. Summary of clinical and pathological characteristics.

Variables	Frequency (Percent)
Age	
Less than 60	42 (14.0%)
60–69	67 (24.0%)
70–79	105 (35.0%)
80 or greater	81 (27.0%)
Gender	
Female	59 (19.7%)
Male	241 (80.3%)
WHO Grade	
High	176 (58.7%)
Low	124 (41.3%)
Stage	
Ta	173 (57.7%)
T1	127 (42.3%)
Treatment	
BCG/MMC	82 (27.3%)
None	218 (72.7%)

The patients were divided into training and testing subgroups (4:1 ratio), such that the two cohorts share similar clinical characteristics. The sampling procedure was conducted via the R package `SPlit`. The training cohort was comprised of 240 samples and the

testing cohort was comprised of 60 samples. We first performed PSH-CSIS on the training subset and pre-selected $240 - 1 - 5 = 234$ genes after adjusting for the effect of the clinical covariates. The likelihood-based boosting approach [12] (CoxBoost) was then applied to the reduced training data for further gene selection and prognostic modeling simultaneously through the R package `CoxBoost`. The clinical covariates remained unpenalized in the boosting procedure and the optimal tuning parameters were determined through 10-fold cross validation. The training and testing prediction errors of the estimated CIF were calculated. As the apparent error evaluated on the training data will underestimate the true prediction error, bootstrap .632+ prediction error [12] was calculated instead. Besides, the time-dependent receiver operating characteristic (ROC) curve along with the area under the curve (AUC) were obtained on the testing data using the R package `riskRegression`. A patient's risk score was defined as the linear combination of the selected genes where the coefficients were extracted from the fitted model (i.e., the gene signature values). The risk score was used to classify the patient as having high or low risk, with the median score of the training group being the cutoff. The same cutoff value was also applied when assigning the test samples. The two risk groups were contrasted by cumulative incidence analysis [17] via the R package `cmprsk`. The performance of the proposed model is compared with a direct boosting procedure without screening as well as a PSH model containing only the clinical covariates. All three models were benchmarked against a null model of the Aalen–Johansen estimator that employs no genetic or clinical covariates. Finally, a PSH model was fitted to the entire dataset to conduct inference about independent prognostic factors associated with progression, where the gene signature identified by the PSH-CSIS+CoxBoost model and the five clinical covariates were used. As pointed out by a reviewer, age ≥ 70 is a potential risk factor for NMIBC. Hence, age was dichotomized at 70 in the final model. As suggested by another reviewer, two additional models were considered for sensitivity analysis: (a) PSH with Lasso penalty and (b) PSH-CSIS followed by PSH with Lasso penalty, using the same analysis scheme described above. The `R` package `fastcmprsk` was used for fitting the Lasso model. `R` code for implementing the PSH-CSIS+CoxBoost model is available at https://github.com/cke23/GeneScreeningBLCA (accessed on 10 October 2022).

3. Results

Table 2 lists the influential genes selected by the PSH-CSIS + CoxBoost model and the only CoxBoost model. The two models selected five genes in common and post-screening CoxBoost identified an additional influential gene. Computing times for running the PSH-CSIS + CoxBoost model and the CoxBoost model (including parameter tuning via cross validation and model fitting) were 2.95 min and 4.04 min, respectively, on a laptop with an i5 1.4 GHz processor and 16 G RAM, which suggests the computational benefit of performing screening before boosting.

Table 2. Genes selected by the two competing models. A risk gene with a positive coefficient from the fitted model is denoted by "+", while a protective gene with a negative coefficient is denoted by "-".

Model	Gene Selected
PSH-CSIS + CoxBoost	AP1M2(-), CAT(-), CCL3(+), MCM7(+), NCF2(+), PKP4(-)
CoxBoost	AP1M2(-), CAT(-), CCL3(+), MCM7(+), NCF2(+)

Prediction error curves of the CIF for all models are displayed in Figure 2. The purely clinical model (dot curves) is seen to clearly improve over the null model (grey curves), indicating that valuable information is contained in the clinical covariates. Incorporating genetic information from the microarray data further assisted in predicting progression, as shown by the CoxBoost model (dash curves). The PSH-CSIS+CoxBoost model (solid curves) performed the best among all models. A similar observation can be found on the ROC curves of the predicted cumulative incidence at 1, 2, 3 and 5 years on the testing data, which are displayed in Figure 3. In addition, cumulative incidence analyses revealed that

the gene signature identified by the PSH-CSIS+CoxBoost model provides effective risk stratification (p-value < 0.001 for the training cohort and p-value = 0.033 for the testing cohort; Figure 4).

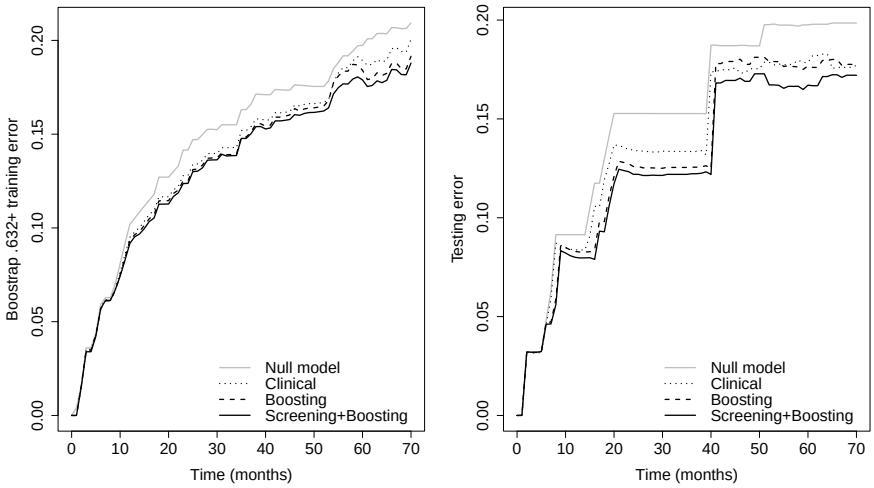

Figure 2. Bootstrap .632+ training error curve (**left panel**) and testing error curve (**right panel**) for prediction of the cumulative incidence function.

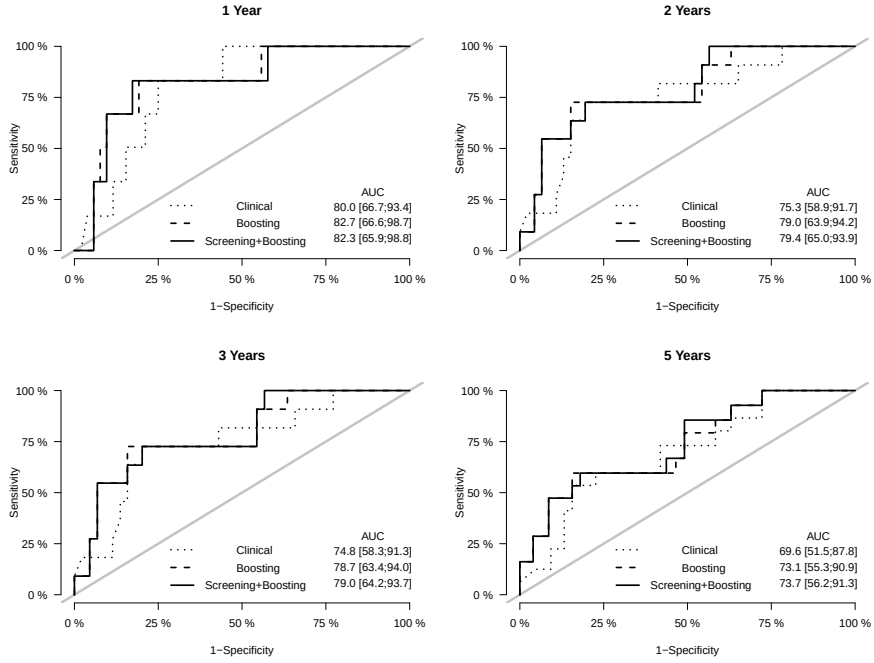

Figure 3. ROC curves and assocaited AUC values of the cumulative incidence predicted by the competing models on the testing data, at years 1, 2, 3 and 5.

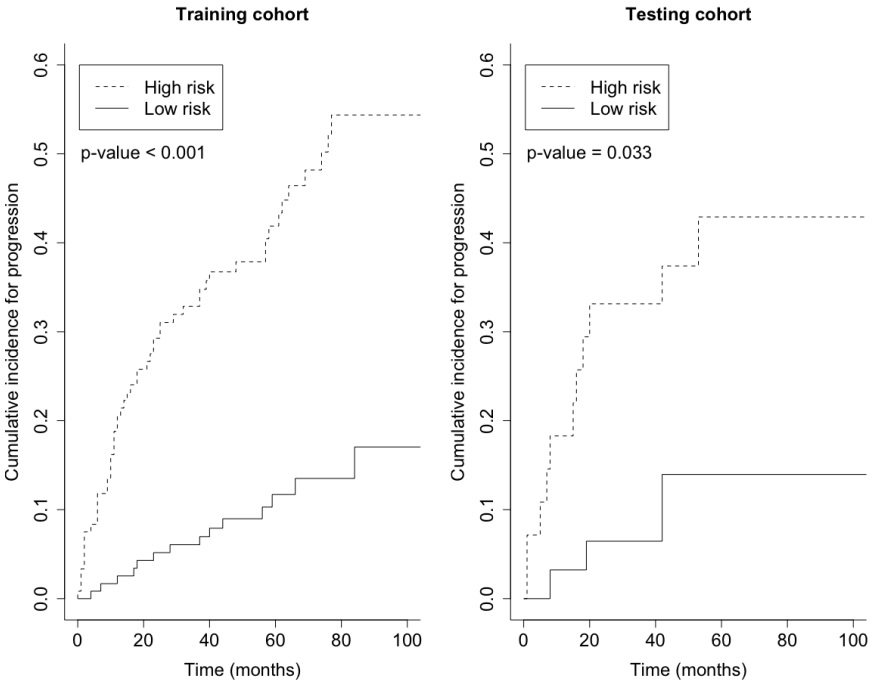

Figure 4. Cumulative incidence by risk group on the basis of the gene signature identified by the PSH-CSIS+CoxBoost model.

From the multivariable PSH model fitted to the whole data (Table 3), the six-gene signature selected by PSH-CSIS+CoxBoost was a significant predictor with an hazard ratio of 12.55 (p-value < 0.001), adjusted for other clinical covariates. The hazard of progression on average increased by 55% after age 70 (p-value = 0.058). Having a low grade tumor led to a reduction of hazard by 57% (p-value = 0.004) compared to high grade tumors (p-value = 0.005). Diagnosis at an early stage decreased the hazard of progression by 40% (p-value = 0.056). Receiving BCG/MMC treatment also resulted in a reduction of hazard by 62% (p-value = 0.002). The high hazard ratio associated with the six-gene signature reflects the strong genetic effect on survival from progression given the clinical covariates already in the model. The effectiveness of the conditional screening-and-selection procedure is thus demonstrated.

Finally, in the sensitivity analysis, the Lasso model and the PSH-CSIS+Lasso model selected the same 5 genes (AP1M2, CAT, CCL3, MCM7, NCF2) as the CoxBoost model. Furthermore, their predictive performances were also similar to the CoxBoost model, but inferior to the PSH-CSIS+CoxBoost model. Prediction error curves of the cumulative incidence for the competing models are displayed in Figure 5. In addition, Figure 6 displays the ROC curves of the predicted cumulative incidences at 1, 2, 3 and 5 years on the testing data. We observe that the PSH-CSIS+Lasso model is better than the Lasso-only model in terms of AUC for all the years, while the PSH-CSIS+CoxBoost model is superior to the PSH-CSIS+Lasso model for the 3-year and 5-year predictions. Computing times for running the Lasso model and the PSH-CSIS+Lasso model (including parameter tuning via cross validation and model fitting) were 11.32 min and 2.82 min, respectively, on a laptop with i5 1.4 GHz processor and 16 G RAM. Although the two models selected identical genes and made similar predictions, the computational gain using the latter is substantial.

Table 3. Hazard ratio estimates, 95% confidence intervals (CI) and *p*-values for the 6-gene signature (identified by PSH-CSIS+CoxBoost) and other clinical covariates, corresponding to cumulative incidence function of Cause 1 (progression, or death from bladder cancer), obtained from fitting the univariate or multivariable competing risks Cox regression.

Variable	Univariate Analysis		Multivariable Analysis	
	Hazard Ratio (95% CI)	*p*-Value	Hazard Ratio (95% CI)	*p*-Value
6-gene signature	8.95 (4.75, 16.90)	<0.001	12.55 (6.11, 25.80)	<0.001
Age				
Age > 70	1.70 (1.11, 2.60)	0.015	1.55 (0.99, 2.45)	0.058
Age ≤ 70	-	-	-	-
Gender				
Male	0.80 (0.43, 1.33)	0.38	0.84 (0.46, 1.55)	0.580
Female	-	-	-	-
WHO Grade				
low	0.40 (0.28, 0.64)	<0.001	0.43 (0.24, 0.77)	0.005
high	-	-	-	-
Stage				
Ta	0.63 (0.41, 0.97)	0.034	1.66 (0.99, 2.80)	0.056
T2	-	-	-	-
Treatment				
None	2.04 (1.19, 3.48)	0.009	2.60 (1.44, 4.69)	0.002
BCG/MMC	-	-	-	-

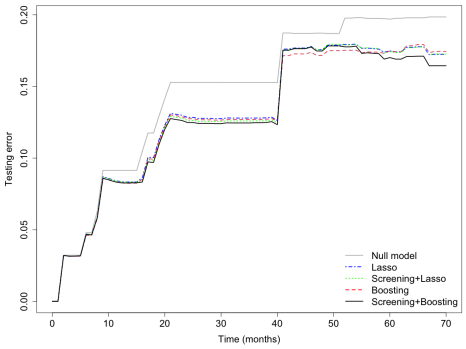

Figure 5. Sensitivity Analysis: testing error curves for the cumulative incidence predicted by the competing models.

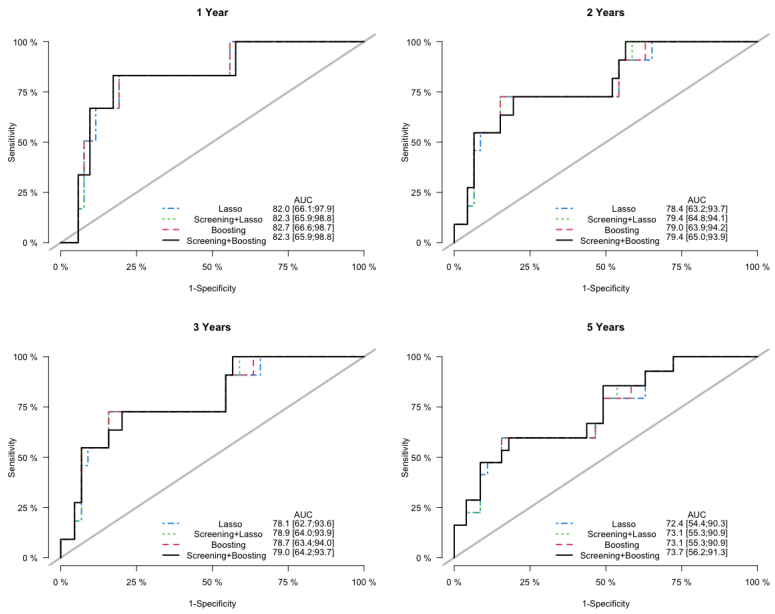

Figure 6. Sensitivity Analysis: ROC curves and assocaited AUC values of the cumulative incidence predicted by the competing models on the testing data, at years 1, 2, 3 and 5.

4. Discussion

We first highlight some biological insights associated with the genes selected by PSH-CSIS+CoxBoost. Adaptor related protein complex 1 subunit mu 2 (AP1M2) is a component of clathrin adaptor complex, which is required for maintaining correct polarity of basolateral membrane proteins in epithelial cells [20]. As yet, there are no functional studies interrogating the role of AP1M2 in cancer. Bioinformatic analyses on AP1M2 are not conclusive, e.g., its level has no effect on patient survival in pancreatic cancer [21], it was identified as a hub gene in renal cancer in which its expression is downregulated [22], while its high expression predicted poor outcome in invasive breast carcinoma [23]. AP1M2 is required for maintaining cell polarity, a marker of differentiation. Dedifferentiation is a hallmark of cancer and it might be anticipated that loss of AP1M2 will prevent proper polarization of cells, thereby contributing to dedifferentiation. The observation that low levels of AP1M2 is associated with poor prognosis in bladder cancer patients is, therefore, meaningful. Reactive oxygen species (ROS) create oxidative stress, resulting in mutagenesis and tumor pathogenesis. The antioxidant enzyme catalase (CAT) provides protection against oxidative stress and downregulation or inactivity of catalase is observed in many cancers [24]. A decreased catalase expression in cancerous bladder tissues in comparison with normal tissues and its association with disease recurrence have been reported by several studies [25,26]. Plakophilin 4 (PKP4), also known as p0071, belongs to the family of armadillo-like proteins, and is involved in regulating adherens junction organization [27]. Overexpression of PKP4 in A431 cells inhibited the ability to close in vitro wounds, suggesting decreased motility of the cells [28]. As yet, the role of PKP4 in bladder cancer is not known and other members of this family play a variable role, e.g., PKP2 is upregulated and PKP3 is downregulated in invasive bladder cancer [29]. It is possible that low levels of PKP4 facilitate increased the motility of bladder cancer cells, conferring an aggressive, metastatic disease. Chronic inflammation plays a key role in bladder carcinogenesis [30]. The cytokine C-C motif chemokine ligand 3 (CCL3), also known as macrophage inflammatory protein 1 alpha (MIP1A), functions in the tumor microenvironment by recruiting tumor-associated macrophages (TAMs) and myeloid-derived suppressor cells (MDSC), thus creating an immunosuppressive environment and facilitating metastasis [31]. CCL3 has been shown to be produced by myeloid cells in bladder cancer patients, which contribute to the pathogenesis of inflammation and immunosuppression [32]. Mini-chromosome maintenance proteins (MCM) are key components of pre-replication complex and are essential for genome replication thereby functioning as oncogenes [33]. A member of this family, MCM7, has been shown to be up-regulated in numerous cancers including bladder cancer [34,35], especially in human papillomavirus (HPV)-positive tumors [36]. Neutrophil cytosolic factor 2 (NCF2) is a subunit of NADPH complex found in neutrophils. NCF2 plays a role in inflammation and cancer, e.g., it is highly expressed in gastric cancer promoting tumor metastasis and invasion by activating NF-κB signaling [37]. In bladder cancer, a high expression level of NCF2 was shown to be associated with an undesirable abundance of chemokine CXCL8, a marker of immunosuppressive MDSCs [38]. Overall, it might be inferred that the expression pattern of our six gene signature predicts a poor prognosis based on what is known about the function of each gene. Nonetheless, further studies, such as multi-center validation, are needed to quantify a stronger evidence and make these new discoveries feasible for use in clinical practice.

In addition to progression, recurrence (in $\leq$80% of patients) is another important issue in NMIBC, especially for patients with non-invasive papillary carcinoma (Ta). Several clinical and pathological factors of recurrence have been identified such as multiplicity, tumor size, and prior recurrence rate [39]. A variety of markers has also been linked to recurrence in bladder cancer, although the role for molecular markers to predict recurrence seems limited because multifocal disease and incomplete treatment are probably more important for recurrence than the molecular features of a removed tumor [40]. Recurrence information is not available in the dataset that we studied in this paper. However, we expect that similar analysis can be carried out to discover molecular drivers, if the survival

outcome for recurrence is provided. Metabolomic studies in bladder cancer for diagnosis and prognosis are also characterized by high dimensional data. Based on different available samples (urine, blood, tissue sample), advances in molecular pathology have driven efforts to identify prognostic and predictive markers and classify bladder cancer into subtypes [41]. Potential applications of the proposed screening procedure in metabolomic data analysis for diagnostic management and pathologic profiling are worthy of future research.

The proposed screening procedure is generalizable, can be combined with other existing feature selection techniques for the PSH model, and applied to other types of cancer data containing competing events. The method can also be adapted to some variations of the PSH model. In practice, the assumption of proportional hazards is unlikely to be satisfied for all covariates in a high-dimensional setting. Bellach et al. [42] thus developed a weighted likelihood function that generalizes the PSH model by allowing time-dependent covariate effects on the subdistribution hazard. On the other hand, control variables such as age can have dynamic impacts on the survival time, motivating us to consider a varying coefficient model to assess nonlinear interaction effects between the primary covariates and control variables [43]. Screening procedures to be developed for these more flexible models may lead to improved model fit and new marker discoveries.

5. Conclusions

The development of high-throughput biology provides a vast amount of information about various phenotypic data, including survival endpoints. Finding prognostic gene signatures for cancer survival became a vital task in biomedical research. The challenge is often threefold. First, the number of features in microarray or sequencing data is much larger than the sample size, creating a predicament to the use of classical statistical analysis tools. How to eliminate noise variables while preserving important features is the key to successful downstream analysis. Although penalized methods have been widely used for feature selection, they can be unstable and computational expensive for high-throughput data. Fast and effective feature screening tools for very high dimensional data have been emerging in the past decade in the statistics literature. However, the dissemination of these attractive methods to biomedical fields is limited. Second, survival endpoints are complicated by censoring and events other than that of primary interest. Data analysis including feature screening/selection must be tailored for the specific survival setting. Third, the desire of incorporating both clinical and genetic information also adds to the difficulty of prognostic modeling. One also has to tradeoff between prediction accuracy and model interpretability.

In this paper, we aimed to address the aforementioned issues for high-throughput competing risks data, in the context of bladder cancer prognosis. The PSH model proposed by Fine and Gray is commonly used for studying survival among competing events, and its estimation and inference has been well-studied. We therefore proposed a feature screening procedure for the PSH model and suggested an analysis pipeline of screening-and-selection for identifying influential genes while controlling for important clinical covariates. The implementation is straightforward with existing `R` packages for competing risks regression. It was demonstrated that our approach was able to reduce the dimension efficiently and improve the performance of the subsequent boosting procedure for feature selection. In particular, we identified a predictive six-gene signature, independent of numerous clinical covariates, for progression in early-stage bladder cancer. The gene signature was shown to be numerically effective and biologically sensible, but multi-center validation is needed to better evaluate its potential value in clinical practice.

Author Contributions: Conceptualization, C.K. and D.B.; methodology, C.K. and D.B.; software, C.K.; validation, C.K.; formal analysis, C.K.; investigation, C.K., D.B. and D.S.; resources, D.B.; data curation, C.K.; writing—original draft preparation, C.K.; writing—review and editing, C.K., D.B. and D.S.; visualization, C.K. and D.B.; supervision, D.B.; project administration, D.B.; funding acquisition, D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grants P20CA252717, P20CA264067, R01DE031134, R21DE031879 from the United States National Institutes of Health (NIH) and the VCU Quest fund. Services and products in support of this research project were also generated by the VCU Massey Cancer Center Biostatistics Shared Resource, supported, in part, with funding from NIH-NCI Cancer Center Support Grant P30CA016059.

Institutional Review Board Statement: Not applicable; the study involved secondary statistical analysis of pre-existing, freely-downloadable and completely de-identified data.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5479 (accessed on 30 August 2022).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Teoh, J.Y.C.; Huang, J.; Ko, W.Y.K.; Lok, V.; Choi, P.; Ng, C.F.; Sengupta, S.; Mostafid, H.; Kamat, A.M.; Black, P.C.; et al. Global trends of bladder cancer incidence and mortality, and their associations with tobacco use and gross domestic product per capita. *Eur. Urol.* **2020**, *78*, 893–906. [CrossRef]
2. Kamat, A.M.; Hahn, N.M.; Efstathiou, J.A.; Lerner, S.P.; Malmström, P.U.; Choi, W.; Guo, C.C.; Lotan, Y.; Kassouf, W. Bladder cancer. *Lancet* **2016**, *388*, 2796–2810. [CrossRef]
3. Cookson, M.S.; Herr, H.W.; Zhang, Z.F.; Soloway, S.; Sogani, P.C.; Fair, W.R. The treated natural history of high risk superficial bladder cancer: 15-year outcome. *J. Urol.* **1997**, *158*, 62–67. [CrossRef]
4. Dignam, J.J.; Zhang, Q.; Kocherginsky, M. The Use and Interpretation of Competing Risks Regression ModelsModeling with Competing Risks. *Clin. Cancer Res.* **2012**, *18*, 2301–2308. [CrossRef]
5. Fine, J.P.; Gray, R.J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **1999**, *94*, 496–509. [CrossRef]
6. Fu, Z.; Parikh, C.R.; Zhou, B. Penalized variable selection in competing risks regression. *Lifetime Data Anal.* **2017**, *23*, 353–376. [CrossRef]
7. Hou, J.; Bradic, J.; Xu, R. Inference under fine-gray competing risks model with high-dimensional covariates. *Electron. J. Stat.* **2019**, *13*, 4449–4507. [CrossRef]
8. Kawaguchi, E.S.; Shen, J.I.; Suchard, M.A.; Li, G. Scalable algorithms for large competing risks data. *J. Comput. Graph. Stat.* **2021**, *30*, 685–693. [CrossRef]
9. Tapak, L.; Kosorok, M.R.; Sadeghifar, M.; Hamidi, O.; Afshar, S.; Doosti, H. Regularized Weighted Nonparametric Likelihood Approach for High-Dimension Sparse Subdistribution Hazards Model for Competing Risk Data. *Comput. Math. Methods Med.* **2021**, *2021*, 5169052. [CrossRef]
10. Sun, H.; Wang, X. High-dimensional feature selection in competing risks modeling: A stable approach using a split-and-merge ensemble algorithm. *Biom. J.* **2022**. [CrossRef]
11. Bühlmann, P.; Yu, B. Boosting with the L 2 loss: Regression and classification. *J. Am. Stat. Assoc.* **2003**, *98*, 324–339. [CrossRef]
12. Binder, H.; Allignol, A.; Schumacher, M.; Beyersmann, J. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* **2009**, *25*, 890–896. [CrossRef]
13. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2008**, *70*, 849–911. [CrossRef]
14. Barut, E.; Fan, J.; Verhasselt, A. Conditional sure independence screening. *J. Am. Stat. Assoc.* **2016**, *111*, 1266–1277. [CrossRef]
15. Hong, H.G.; Li, Y. Feature selection of ultrahigh-dimensional covariates with survival outcomes: A selective review. *Appl. Math.-A J. Chin. Univ.* **2017**, *32*, 379–396. [CrossRef]
16. Kuk, D.; Varadhan, R. Model selection in competing risks regression. *Stat. Med.* **2013**, *32*, 3077–3088. [CrossRef]
17. Gray, R.J. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Stat.* **1988**, *16*, 1141–1154. [CrossRef]
18. Li, R.; Zhong, W.; Zhu, L. Feature screening via distance correlation learning. *J. Am. Stat. Assoc.* **2012**, *107*, 1129–1139. [CrossRef] [PubMed]
19. Dyrskjøt, L.; Zieger, K.; Real, F.X.; Malats, N.; Carrato, A.; Hurst, C.; Kotwal, S.; Knowles, M.; Malmström, P.U.; de la Torre, M.; et al. Gene expression signatures predict outcome in non–muscle-invasive bladder carcinoma: A multicenter validation study. *Clin. Cancer Res.* **2007**, *13*, 3545–3551. [CrossRef]
20. Fölsch, H.; Ohno, H.; Bonifacino, J.S.; Mellman, I. A novel clathrin adaptor complex mediates basolateral targeting in polarized epithelial cells. *Cell* **1999**, *99*, 189–198. [CrossRef]
21. Zhang, X.; Liu, S.; Cai, Y.; Changyong, E.; Sheng, J. Screening and validation of independent predictors of poor survival in pancreatic cancer. *Pathol. Oncol. Res.* **2021**, *27*, 1609868.

22. Wu, H.; Fan, L.; Liu, H.; Guan, B.; Hu, B.; Liu, F.; Hocher, B.; Yin, L. Identification of key genes and prognostic analysis between chromophobe renal cell carcinoma and renal oncocytoma by bioinformatic analysis. *BioMed Res. Int.* **2020**, *2020*, 4030915. [CrossRef]
23. Yi, Y.; Zhang, Q.; Shen, Y.; Gao, Y.; Fan, X.; Chen, S.; Ye, X.; Xu, J. System analysis of adaptor-related protein complex 1 subunit mu 2 (AP1M2) on malignant tumors: A pan-cancer analysis. *J. Oncol.* **2022**, *2022*, 7945077. [CrossRef]
24. Glorieux, C.; Calderon, P.B. Catalase, a remarkable enzyme: Targeting the oldest antioxidant enzyme to find a new cancer treatment approach. *Biol. Chem.* **2017**, *398*, 1095–1108. [CrossRef]
25. Islam, M.O.; Bacchetti, T.; Ferretti, G. Alterations of antioxidant enzymes and biomarkers of nitro-oxidative stress in tissues of bladder cancer. *Oxidative Med. Cell. Longev.* **2019**, *2019*, 2730896. [CrossRef]
26. Wieczorek, E.; Jablonowski, Z.; Tomasik, B.; Gromadzinska, J.; Jablonska, E.; Konecki, T.; Fendler, W.; Sosnowski, M.; Wasowicz, W.; Reszka, E. Different gene expression and activity pattern of antioxidant enzymes in bladder cancer. *Anticancer Res.* **2017**, *37*, 841–848. [CrossRef]
27. Keil, R.; Schulz, J.; Hatzfeld, M. p0071/PKP4, a multifunctional protein coordinating cell adhesion with cytoskeletal organization. *Biol. Chem.* **2013**, *394*, 1005–1017. [CrossRef] [PubMed]
28. Setzer, S.V.; Calkins, C.C.; Garner, J.; Summers, S.; Green, K.J.; Kowalczyk, A.P. Comparative analysis of armadillo family proteins in the regulation of a431 epithelial cell junction assembly, adhesion and migration. *J. Investig. Dermatol.* **2004**, *123*, 426–433. [CrossRef] [PubMed]
29. Takahashi, H.; Nakatsuji, H.; Takahashi, M.; Avirmed, S.; Fukawa, T.; Takemura, M.; Fukumori, T.; Kanayama, H. Up-regulation of plakophilin-2 and Down-regulation of plakophilin-3 are correlated with invasiveness in bladder cancer. *Urology* **2012**, *79*, 240.e1–240.e8. [CrossRef] [PubMed]
30. Michaud, D.S. Chronic inflammation and bladder cancer. In *Urologic Oncology: Seminars and Original Investigations*; Elsevier: Amsterdam, The Netherlands, 2007; Volume 25, pp. 260–268.
31. Ntanasis-Stathopoulos, I.; Fotiou, D.; Terpos, E. CCL3 signaling in the tumor microenvironment. *Tumor Microenviron.* **2020**, *1231*, 13–21.
32. Eruslanov, E.; Neuberger, M.; Daurkin, I.; Perrin, G.Q.; Algood, C.; Dahm, P.; Rosser, C.; Vieweg, J.; Gilbert, S.M.; Kusmartsev, S. Circulating and tumor-infiltrating myeloid cell subsets in patients with bladder cancer. *Int. J. Cancer* **2012**, *130*, 1109–1119. [CrossRef]
33. Yu, S.; Wang, G.; Shi, Y.; Xu, H.; Zheng, Y.; Chen, Y. MCMs in cancer: Prognostic potential and mechanisms. *Anal. Cell. Pathol.* **2020**, *2020*, 3750294. [CrossRef]
34. Fristrup, N.; Birkenkamp-Demtröder, K.; Reinert, T.; Sanchez-Carbayo, M.; Segersten, U.; Malmström, P.U.; Palou, J.; Alvarez-Múgica, M.; Pan, C.C.; Ulhøi, B.P.; et al. Multicenter validation of Cyclin D1, MCM7, TRIM29, and UBE2C as prognostic protein markers in non-muscle–invasive bladder cancer. *Am. J. Pathol.* **2013**, *182*, 339–349. [CrossRef] [PubMed]
35. Toyokawa, G.; Masuda, K.; Daigo, Y.; Cho, H.S.; Yoshimatsu, M.; Takawa, M.; Hayami, S.; Maejima, K.; Chino, M.; Field, H.I.; et al. Minichromosome Maintenance Protein 7 is a potential therapeutic target in human cancer and a novel prognostic marker of non-small cell lung cancer. *Mol. Cancer* **2011**, *10*, 1–11. [CrossRef] [PubMed]
36. Shigehara, K.; Sasagawa, T.; Kawaguchi, S.; Nakashima, T.; Shimamura, M.; Maeda, Y.; Konaka, H.; Mizokami, A.; Koh, E.; Namiki, M. Etiologic role of human papillomavirus infection in bladder carcinoma. *Cancer* **2011**, *117*, 2067–2076. [CrossRef]
37. Zhang, J.X.; Chen, Z.H.; Chen, D.L.; Tian, X.P.; Wang, C.Y.; Zhou, Z.W.; Gao, Y.; Xu, Y.; Chen, C.; Zheng, Z.S.; et al. LINC01410-miR-532-NCF2-NF-kB feedback loop promotes gastric cancer angiogenesis and metastasis. *Oncogene* **2018**, *37*, 2660–2675. [CrossRef]
38. Muthuswamy, R.; Wang, L.; Pitteroff, J.; Gingrich, J.R.; Kalinski, P. Combination of IFNα and poly-I: C reprograms bladder cancer microenvironment for enhanced CTL attraction. *J. Immunother. Cancer* **2015**, *3*, 1–10. [CrossRef] [PubMed]
39. Van Rhijn, B.W.; Burger, M.; Lotan, Y.; Solsona, E.; Stief, C.G.; Sylvester, R.J.; Witjes, J.A.; Zlotta, A.R. Recurrence and progression of disease in non–muscle-invasive bladder cancer: From epidemiology to treatment strategy. *Eur. Urol.* **2009**, *56*, 430–442. [CrossRef]
40. van Rhijn, B.W. Combining molecular and pathologic data to prognosticate non-muscle-invasive bladder cancer. In *Urologic Oncology: Seminars and Original Investigations*; Elsevier: Amsterdam, The Netherlands, 2012; Volume 30, pp. 518–523.
41. di Meo, N.A.; Loizzo, D.; Pandolfo, S.D.; Autorino, R.; Ferro, M.; Porta, C.; Stella, A.; Bizzoca, C.; Vincenti, L.; Crocetto, F.; et al. Metabolomic Approaches for Detection and Identification of Biomarkers and Altered Pathways in Bladder Cancer. *Int. J. Mol. Sci.* **2022**, *23*, 4173. [CrossRef]
42. Bellach, A.; Kosorok, M.R.; Rüschendorf, L.; Fine, J.P. Weighted NPMLE for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **2019**, *114*, 259–270. [CrossRef]
43. Tian, B.; Liu, Z.; Wang, H. Non-marginal feature screening for varying coefficient competing risks model. *Stat. Probab. Lett.* **2022**, *190*, 109648. [CrossRef]