Feature screening for survival trait with application to TCGA high-dimensional genomic data

Jie-Huei Wang, Cai-Rong Li and Po-Lin Hou

Department of Statistics, Feng Chia University, Taichung, Taiwan

ABSTRACT

Background: In high-dimensional survival genomic data, identifying cancer-related genes is a challenging and important subject in the field of bioinformatics. In recent years, many feature screening approaches for survival outcomes with high-dimensional survival genomic data have been developed; however, few studies have systematically compared these methods. The primary purpose of this article is to conduct a series of simulation studies for systematic comparison; the second purpose of this article is to use these feature screening methods to further establish a more accurate prediction model for patient survival based on the survival genomic datasets of The Cancer Genome Atlas (TCGA).

Results: Simulation studies prove that network-adjusted feature screening measurement performs well and outperforms existing popular univariate independent feature screening methods. In the application of real data, we show that the proposed network-adjusted feature screening approach leads to more accurate survival prediction than alternative methods that do not account for gene-gene dependency information. We also use TCGA clinical survival genetic data to identify biomarkers associated with clinical survival outcomes in patients with various cancers including esophageal, pancreatic, head and neck squamous cell, lung, and breast invasive carcinomas.

Conclusions: These applications reveal advantages of the new proposed network-adjusted feature selection method over alternative methods that do not consider gene-gene dependency information. We also identify cancer-related genes that are almost detected in the literature. As a result, the network-based screening method is reliable and credible.

Submitted 29 September 2021
Accepted 21 February 2022
Published 10 March 2022

Corresponding author
Jie-Huei Wang,
jhwang@mail.fcu.edu.tw

Academic editor
Alexander Schliep

Additional Information and Declarations can be found on page 16

DOI **10.7717/peerj.13098**

Subjects Bioinformatics, Statistics
Keywords Survival feature screening, High-dimensional genomic data, Network, Survival prediction, TCGA, Esophageal cancer, Pancreatic cancer, Head and neck squamous cell carcinoma, Lung adenocarcinoma, Breast invasive carcinoma

INTRODUCTION

In high-dimensional genomic data, identifying important genes related to clinical survival traits is a challenging and important problem in the field of bioinformatics. The discovery of important biomarkers that explain the phenotype of interest is essential for the development of phenotype prediction models. In particular, contaminated data and right-censored survival outcomes make relevant feature screening more challenging.

In the era of high-throughput biology, the number of potential features/biomarkers could be much larger than the research sample size. In this case, it is well-known that use of preliminary feature screening can substantially improve the model selection performed by the regularization approach (*Fan & Lv, 2008*). Univariate feature screening for right-censored survival outcomes has been a challenging topic receiving much attention in the literature. *Edelmann et al. (2020)* recently provided a comprehensive review and some useful suggestions for univariate feature independent screening methods, and pointed out these screening methods for survival traits can be roughly divided into the following categories: (semi-)parametric-based approaches, non-parametric ranking-based approaches, model-free approached based on conditional survival functions, and distance correlation-based approaches. All univariate independent feature screening methods are based on certain statistics with specific model assumptions, which are calculated for each variable without considering other variables. This statistic measure is the so-called marginal utility. The feature screening procedure can then be performed by selecting important features based on their corresponding marginal utility. Since the marginal model is low-dimensional, the main advantage of the marginal model is its computational stability and conceptual simplicity. Therefore, marginal programs are still popular in the fields of bioinformatics.

However, outlier-contaminated biomarker data pose a further challenge to the survival prediction problem based on high-dimensional genetic/genomic data. As is known, the developmental process of disease is complicated and may involve the interaction of multiple genes; that is, epistasis. In other words, to effectively identify disease-related genes, it is necessary to make full use of biological network information. *Wu, Zhu & Feng (2018)* pointed out that ignoring gene-gene dependency information might lead to bias in gene screening. To this end, *Wang & Chen (2021)* developed a network-adjusted Kendall's tau measure for feature screening by incorporating gene-gene dependency network information and compared a network-adjusted measure to a partial-likelihood screening method (*Fan, Feng & Wu, 2010*; *Zhao & Li, 2012*) and inverse probability-of-censoring weighted (IPCW) Kendall's tau statistics (*Song et al., 2014*; *Wang & Chen, 2020*). They proved that the network-adjusted Kendall's tau measurement method is superior to these two methods in variable screening for most of the network structures considered in terms of the average number of true predictors contained in the selected model and the minimum of model size.

In this article, we intend to perform a systematic comparison for these advanced feature screening methods and apply them to The Cancer Genome Atlas (TCGA, *The Cancer Genome Atlas Research Network, 2008*) survival genomic data to develop a more accurate prediction model for patient survival. The simulation studies under various scenarios are conducted to compare the performance of new network-adjusted IPCW Kendall's tau measure with several commonly used univariate independent feature screening methods. In the application of real data, we demonstrate that the new network-adjusted feature screening approach leads to more accurate survival prediction than alternative methods that do not account for feature network information or outlier-contamination. We also determine biomarkers that are associated with clinical survival outcomes of patients with

esophageal carcinoma (ESCA), pancreatic adenocarcinoma (PAAD), head and neck squamous cell carcinoma (HNSCC), lung adenocarcinoma (LUAD), and breast invasive carcinoma (BRCA) using TCGA genetic data.

MATERIALS AND METHODS

Data structure and methods partial review

We consider a study with n independent subjects. For a subject i, suppose that there are p genes expression $\left(x_{i1}, \ldots, x_{ip}\right)'$ related to clinical survival outcomes T_i. Note that the number of the genes is far greater than the sample size n. Usually, the survival outcome is subject to censoring, so we define C_i as censoring time, and use δ_i as the indicator of whether the survival time of subject i is censored, then define $V_i = \min(T_i, C_i)$ as observed survival time.

We list the common and effective survival feature screening methods, and provide readers with an overview summarized in Table 1. *Edelmann et al. (2020)* provided an R package "*MVS*" that can be downloaded from https://github.com/thomashielscher/MVS, which contains the first six screening methods discussed in Table 1, and R codes for the last two screening methods discussed in Table 1 are available at https://figshare.com/articles/software/CODE/16677070, laying the foundation for meaningful comparison.

TCGA cancer data source

TCGA RNA-Seq expression data and phenotypic data including survival time and censoring status data can be downloaded from the R package 'TCGAbiolinks' (*Colaprico et al., 2016*), or 'UCSCXenaTools' (*Wang & Liu, 2019*). The TCGA ESCA, PAAD, LUAD, and BRCA genomic data with survival traits analyzed during this study are all available at Figshare: https://figshare.com/articles/dataset/DATA/16677619. The TCGA HNSCC genomic data can be downloaded from the R package "GEInter" (*Wu, Qin & Ma, 2021*).

Evaluation performance in the simulation study

In performance measurement, we report the percentiles of the minimum model size (MMS) statistics among 200 replications through violin plot to view the distribution of MMS data and its summary statistics, where MMS is the minimum size of a selected model, including underlying effective predictors. MMS measures the complexity of the selected model and reflects the accuracy of the screening process; a smaller MMS value indicates the higher accuracy of feature screening. We note that a violin chart can be constructed through the "*vioplot*" R package (*Adler & Kelly, 2021*). We also performed additional simulation studies to investigate the survival time prediction errors by giving the average number of *c*-index (*Harrell, Lee & Mark, 1996*) among 200 replications as a function of the number of selected features for each method. The *c*-index metric compares the subjects' predicted survival time rankings with their real survival time rankings. A larger *c*-index indicates better prediction accuracy.

For our comparative study, we assess the similarity of different screening methods in terms of the list of detected features. To this end, for each simulation run, the Jaccard

Table 1 Reviews on univariate feature screening for survival outcomes (a partial list).

Citation	Class/Method	Description
Fan, Feng & Wu (2010) and *Zhao & Li (2012)*	(semi-)parametric-based approach/partial likelihood (PL)	The PL approach takes the maximum value of the corresponding marginal Cox's partial likelihood as marginal utilities to rank the predictors.
Saldana & Feng (2018)	(semi-)parametric-based approach/sure independence screening (SIS)	The SIS approach is an *ad hoc* approach, which takes Pearson correlations between predictors and survival times as marginal utilities to rank the predictors.
Gorst-Rasmussen & Scheike (2012)	(semi-)parametric-based approach/feature aberration at survival times (FAST)	The FAST approach proposes a semi-parametric independent screening method for survival data which are described by single-index hazard rate models.
Chen, Chen & Wang (2018)	distance correlation-based approach/robust censored distance correlation screening (RCDCS)	The RCDCS approach takes the robust distance correlation (*Zhong et al., 2016*) by replacing the survival outcomes and predictors by their corresponding cumulative distribution functions' Kaplan–Meier estimator and empirical distribution function as marginal utilities to rank the predictors.
Chen, Chen & Wang (2018)	distance correlation-based approach/composite robust censored distance correlation screening (CRCDCS)	The CRCDCS approach modifies the robust distance correlation through the composite quantile distance correlation of *Chen, Chen & Liu (2019)* to the right-censored scenario by redistributing the masses of censored observations to the right with the indicator function being involved.
Harrell, Lee & Mark (1996)	non-parametric ranking-based approach/Harrell's concordance index (CINDEX)	The CINDEX approach takes the C-index as marginal utilities to rank the predictors.
Song et al. (2014) and *Wang & Chen (2020)*	non-parametric ranking-based approach/inverse probability-of-censoring weighted (IPCW) Kendall's tau	The IPCW Kendall's tau approach takes Kendall's tau rank correlation as marginal utilities to measure the association between survival trait and biomarkers, which uses the IPCW technique to accommodate right-censored survival outcomes.
Wang & Chen (2021)	network-based approach/: IPCW-tau (NPN-MB)	The NPN-MB approach modifies the IPCW Kendall's tau measure (*Wang & Chen, 2020*) to incorporate gene–gene dependency network information using the technique of Google's PageRank Markov matrix. The NPN-MB approach using the nonparanormal procedure (*Liu, Lafferty & Wasserman, 2009*) to transform the original predictors to follow multivariate normal distribution, and then using the MB method (*Meinshausen & Buhlmann, 2006*) to estimate the sparse precision matrix.

index ($J(A, B) = |A \cap B|/|A \cup B|$) of the true feature list is calculated by selection of by two methods under a specific model size of 500. The average Jaccard index of 200 simulation repetitions is used as a measure of similarity between the two methods; in this way, the similarity matrix of all the screening methods under consideration can be constructed and visualized, for example, using the "*corrplot*" R package (*Wei & Simko, 2017*).

In addition, we calculate the overlap coefficient ($O(A, B) = |A \cap B|/\min(|A|, |B|)$) set similarity analysis of features selected by each method with a ground truth set of predictors. The average number of overlap coefficient index among 200 replications as a function of the number of selected features for each method is used as a measure of similarity between the feature screening method and the ground truth set of predictors. A larger overlap coefficient index indicates highly similarity with a ground truth set of predictors.

Simulation scenarios

For the first simulated settings, we follow the simulation settings of *Song et al. (2014)*, and generate a cohort of 500 subjects. Each subject's survival time follows the linear transformation model

$$H(T_i) = -x_i'\beta_0 + \varepsilon,$$

where $H(t) = \log(0.5(e^{2t} - 1))$, the covariates x jointly follow a 2,000-dimensional multivariate standard normal distribution with the first-order autoregressive (AR(1)) structure that is $\mathrm{corr}(x_{.j}, x_{.k}) = 0.5^{|j-k|}$. The distribution of the error term ε follows a standard extreme value distribution, which corresponds to a proportional hazards (PH) model.

The true regression coefficient vector is sparse:

$$\begin{aligned}\boldsymbol{\beta}_0 = (&-1,\ -0.9,\ 0.5,\ 0.8,\ 0.6, 0_5, 0.3,\\ &0.7, -0.8, -0.5, -1,\ 0_5, -2,\ 1,\ 0,\ -0.5, -2, 0_{1,975}),\end{aligned}$$

so the underlying survival model has 14 true predictors. In the first simulated settings, only linear relationships were assumed with true parameter vector. The censoring time distribution follows a uniform distribution U(a, b), where (a, b) is chosen to control the censoring rate at about 30% (light censoring), 50% (middle censoring) and 70% (heavy censoring) respectively. Moreover, we consider the setting where, with a probability of 0.1, the covariates may be contaminated by outliers produced by a *t* distribution with two degrees of freedom.

For the second simulated settings, we follow the simulation settings of *Edelmann et al. (2020)*. The simulated settings are the same as the first simulated settings except for the relationships of true parameter vector, *i.e.*,

$$H(T_i) = -z_i'\beta_0 + \varepsilon,$$

where $z_i' = (g_1(x_{i1}), \ldots, g_p(x_{ip}))$, we assume $g_1(x) = \beta_1|x|$, $g_2(x) = \beta_2|x|, g_4(x) = \beta_4 x^2$, $g_5(x) = \beta_5 1(x > 0)$, and all other j, we assume $g_j(x) = \beta_j x$.

For the third simulated settings, we follow the simulation settings of *Wang & Chen (2021)*. The simulated settings are the same as the first simulated settings except for the correlation structure of the variables and the true regression coefficient. We generate the covariates jointly following a 2,000-dimensional multivariate standard normal distribution with different network structures, including "hub", "band", "cluster", and "scale-free". These network structures can be generated by the "*huge*" package with *huge.generator* function (*Zhao et al., 2012*) and the true regression coefficient vector is defined as

$$\begin{aligned}\boldsymbol{\beta}_0 = (&-1.5,\ -1.5,\ 1.5,\ 1.5,\ 1.5, 0_5, 1.5,\\ &1.5, -1.5, -1.5, -1.5,\ 0_5, -1.5,\ 1.5,\ 1.5,\ -1.5, -1.5, 0_{1,975})\end{aligned}$$

For each simulation scenario, we perform 200 replications to investigate the numerical performances of different methods.

Survival prediction measure in real data application

To evaluate the performance of survival prediction, we report three measures of prediction accuracy: the *c*-index, deviance, and the number of selected features (NOSF) metrics. Larger *c*-index/smaller deviance and number of selected features indicates better prediction accuracy. Since the *c*-index metric can be used to compare a subject's predicted survival time ranking with the true survival time ranking. And, the deviance metric is defined as

$$D = -2\left(logL\left(\hat{\boldsymbol{\beta}}\right) - logL(0)\right),$$

where $logL\left(\hat{\boldsymbol{\beta}}\right)$ is the log partial likelihood function of Cox's model from the testing set of the data, $\hat{\boldsymbol{\beta}}$ is an estimator of the penalized Cox's regression with the MCP penalty parameter in a prediction model obtained from the training set of the data, and $logL(0)$ is the log partial likelihood function of Cox's null model from the testing set of the data, where all predictors are assumed not related to the true survival time.

The deviance metric can therefore be considered as a comparison between the survival prediction model and null model (no predictors considered). The deviance metric is also a suitable survival prediction criterion.

Finally, we choose the number of selected features metric as a precision criterion; the reason is that the feature screening approach is used to find parsimonious precision models. Meaning, if we are modeling two sets of features with the same predictive accuracy, we want to choose the model that uses the features, as a smaller number of selected features are easier to interpret/evaluate in follow-up studies about biological function.

RESULTS

Simulation studies

In simulation studies, a series of simulation studies are conducted to investigate the performance of the existing feature screening methods for survival trait in identifying true associated predictors and survival prediction errors.

The simulation results are summarized in Figs. 1–6 and Figs. S1–S12. Note that Figs. 1–3 indicate the performances for the simulation study 1 with AR(1) structure; Figs. 4–6 indicate the performances for the simulation study 2 with nonlinear structure; Figs. S1–S3 indicate the performances for the simulation study 3 with band structure; Figs. S4–S6 indicate the performances for the simulation study 3 with hub structure; Figs. S7–S9 indicate the performances for the simulation study 3 with cluster structure; and Figs. S10–S12 indicate the performances for the simulation study 3 with scale-free structure. We note that the "IPCW-tau (MB)" method is also an IPCW Kendall tau method of network adjustment, but it uses the original predictors without using nonparanormal procedure to transform them.

From simulation studies 1, 2, and 3 with band structure, we see that the IPCW-tau (NPN-MB) method outperforms all alternative methods in terms of the overlap coefficient (top three panels of Figs. 1, 4, S1), and MMS measure (Figs. 2, 5, S2). Overall, the variable

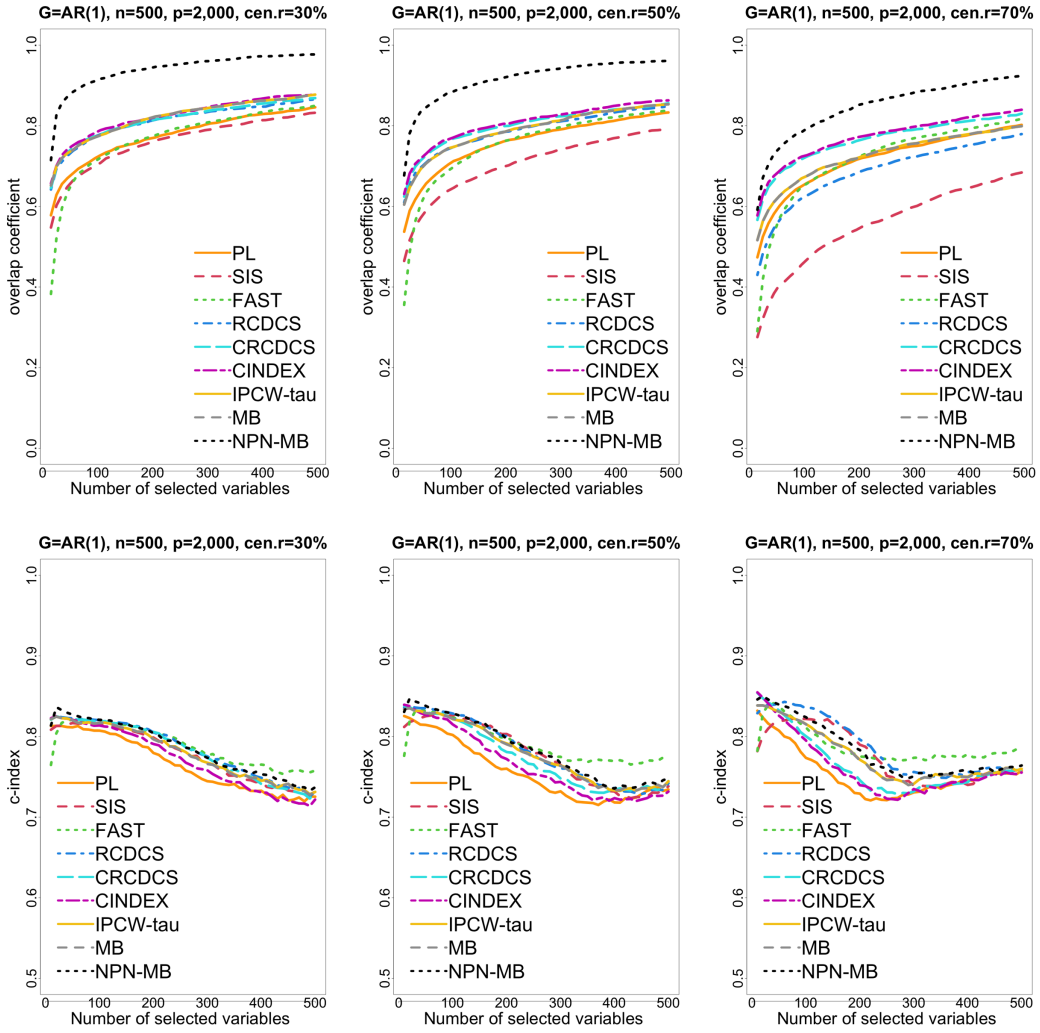

Figure 1 The multi-panel figure contains the mean number of overlap coefficient (top three panels) and *c*-index (bottom three panels) among 200 replications by the number of selected features for the simulation 1 with AR(1) structure based on PH model. The left, medium, and right plots are based on censoring rates of 30%, 50%, and 70%, respectively. A larger mean number of overlap coefficient indicates highly similarity with a ground truth set of predictors, and a larger c-index indicates better prediction accuracy. Note that the underlying survival model has 14 true predictors.

Full-size DOI: 10.7717/peerj.13098/fig-1

selection accuracy of the IPCW-tau (NPN-MB) method is substantially better than all other methods. From bottom three panels of Figures 1, 4, S1, the FAST method has a higher *c*-index when the number of selected variables is larger, and the IPCW-tau (NPN-MB) method outperforms most alternative methods when the number of selected variables is medium or small.

In simulation study 3 with hub structure, we see that the IPCW-tau (NPN-MB) method performs the best when the number of selected variables is larger or medium in terms of the overlap coefficient (top three panels of Fig. S4), then the CINDEX method performs the best when the number of selected variables is small. The *c*-index measure patterns (bottom three panels of Fig. S4) are similar to these in the previous simulation studies.

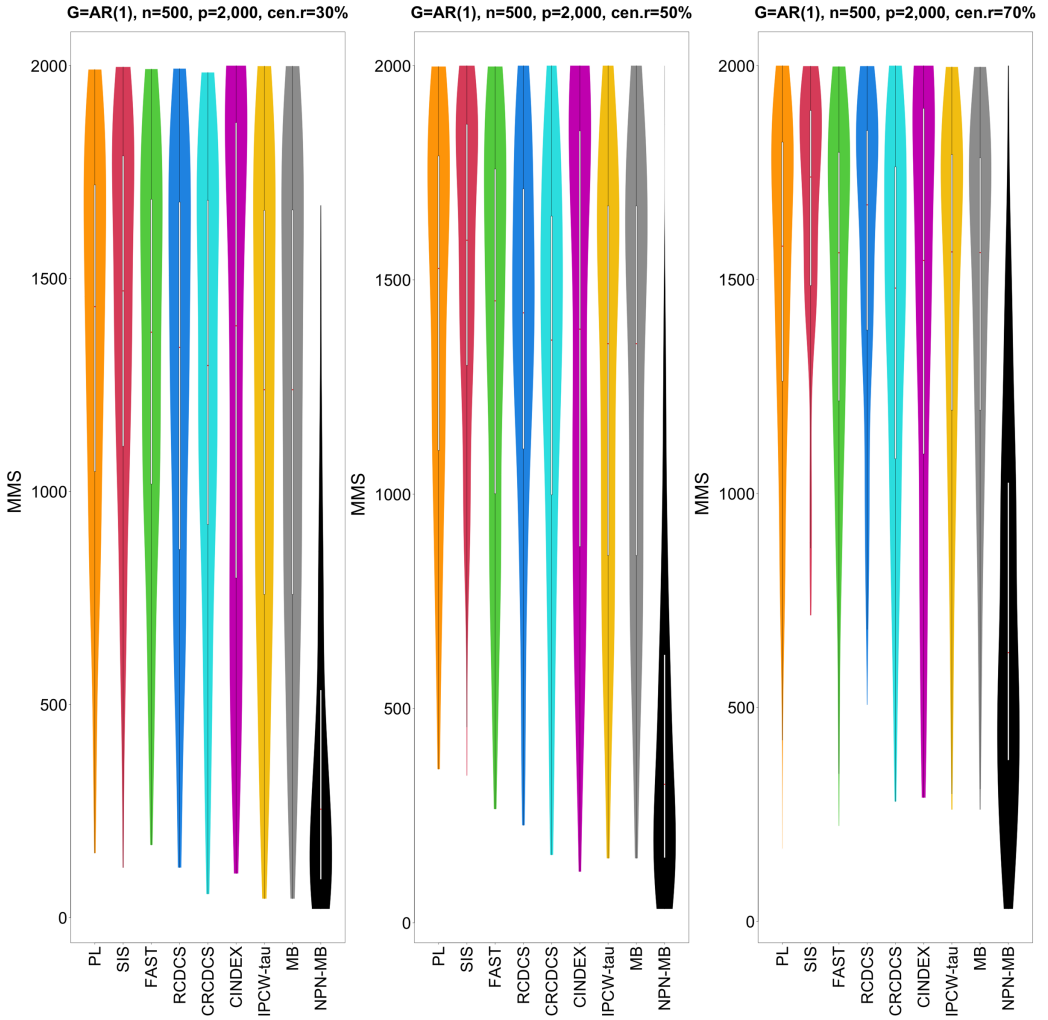

Figure 2 The violin chart of minimum of model size (MMS) measure among 200 replications for the simulation study 1 with AR(1) structure based on PH model. The left, medium, and right plots are based on censoring rates of 30%, 50%, and 70%, respectively. A smaller MMS value indicates the higher accuracy of feature screening.
Full-size DOI: 10.7717/peerj.13098/fig-2

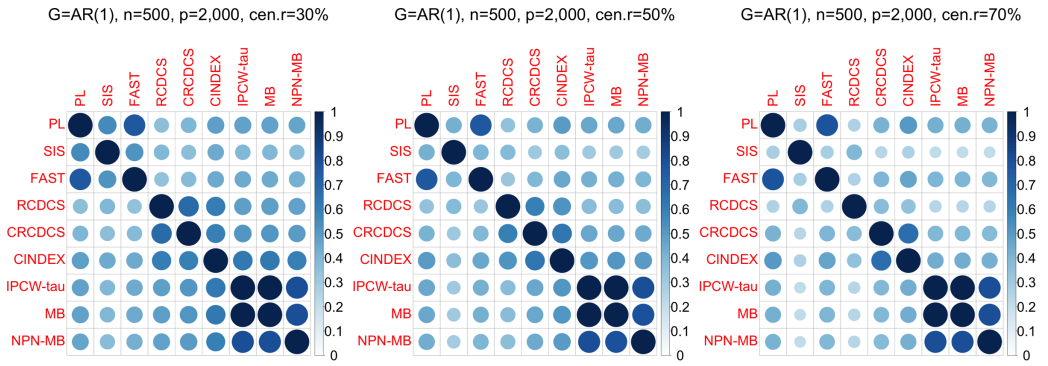

Figure 3 The average of Jaccard index among 200 replications for the simulation study 1 with AR(1) structure based on PH model. The left, medium, and right plots are based on censoring rates of 30%, 50%, and 70%, respectively.
Full-size DOI: 10.7717/peerj.13098/fig-3

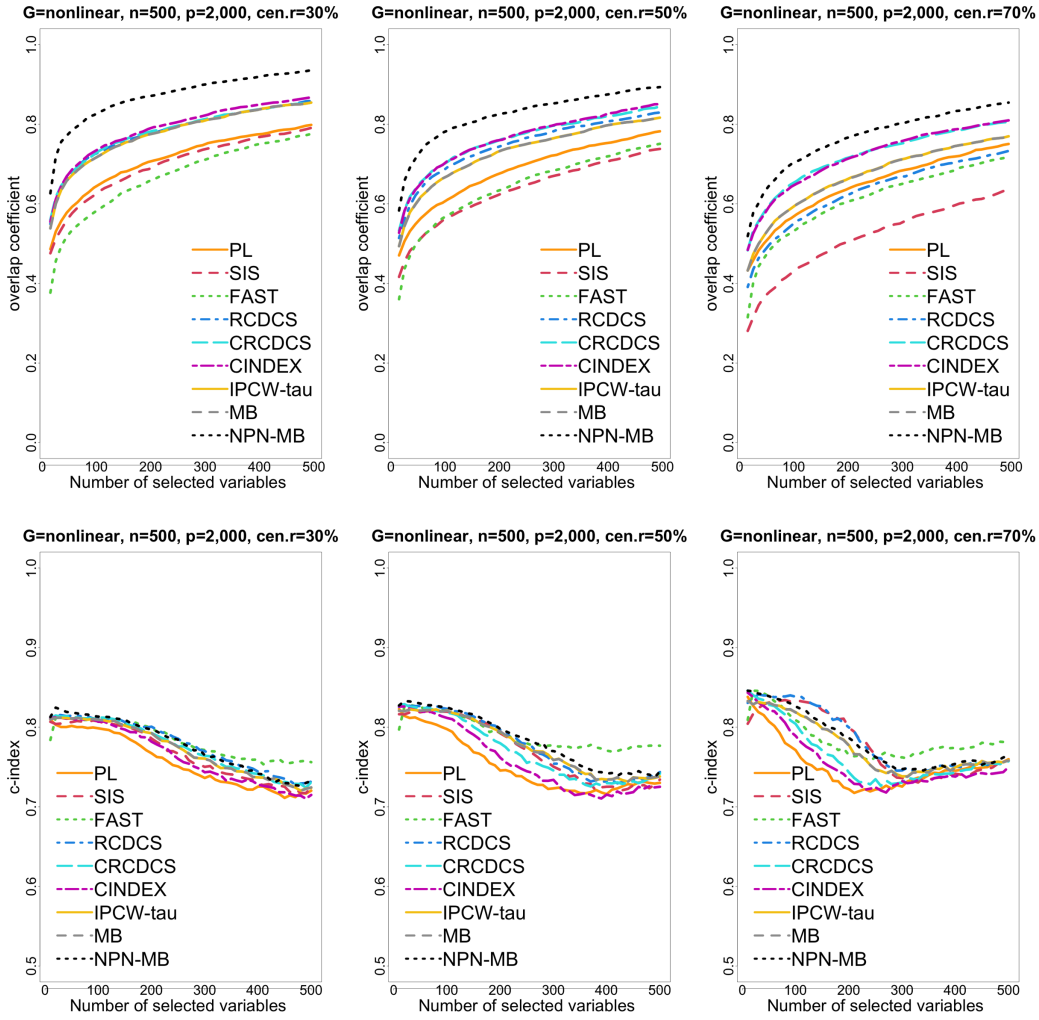

Figure 4 The multi-panel figure contains the mean number of overlap coefficient (top three panels) and *c*-index (bottom three panels) among 200 replications by the number of selected features for the simulation two with nonlinear structure based on PH model. The left, medium, and right plots are based on censoring rates of 30%, 50%, and 70%, respectively. A larger mean number of overlap coefficient indicates highly similarity with a ground truth set of predictors, and a larger c-index indicates better prediction accuracy. Note that the underlying survival model has 14 true predictors.

Full-size DOI: 10.7717/peerj.13098/fig-4

On the MMS measure metric (Fig. S5), the IPCW-tau (NPN-MB) method performs the best.

In simulation study 3 with cluster and scale-free structures, we see that the CINDEX method performs the best when the censoring rate is high in terms of the overlap coefficient (top three panels of Figs. S7, S10), then the IPCW-tau (NPN-MB) method performs the best when the censoring rate is medium or small. The *c*-index measure patterns (bottom three panels of Figs. S7, S10) are similar to these in the previous simulation studies. In the MMS measure metric (Figs. S8, S11), the performance of IPCW-tau (NPN-MB) method is better than other methods.

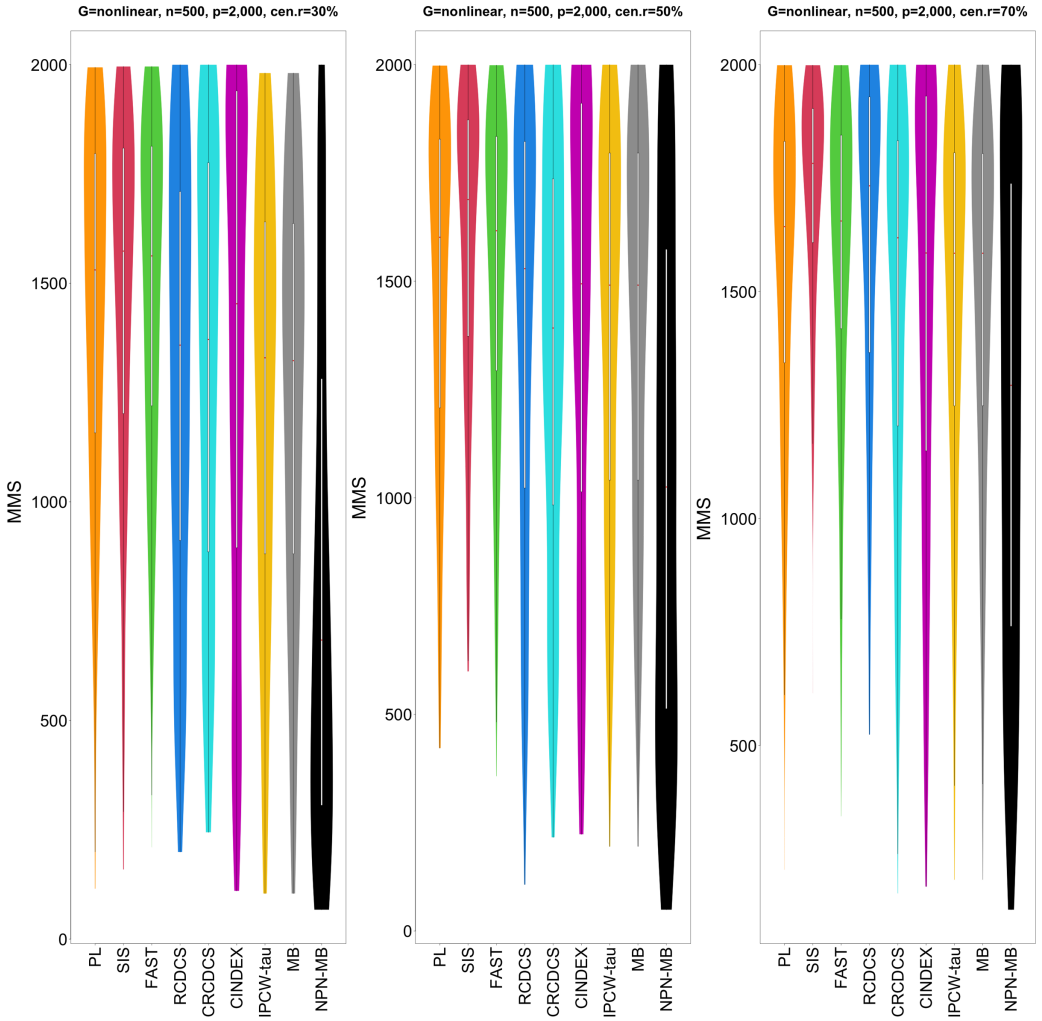

Figure 5 The violin chart of minimum of model size (MMS) measure among 200 replications for the simulation study 2 with nonlinear structure based on PH model. The left, medium, and right plots are based on censoring rates of 30%, 50%, and 70%, respectively. A smaller MMS value indicates the higher accuracy of feature screening. Full-size DOI: 10.7717/peerj.13098/fig-5

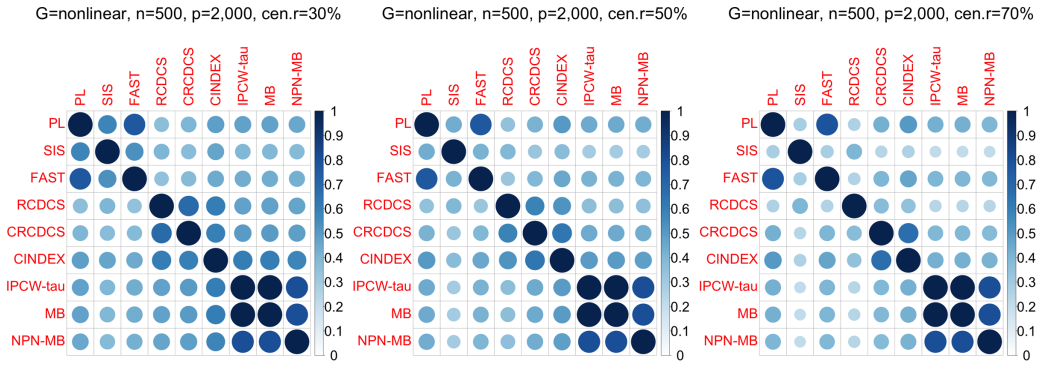

Figure 6 The average of Jaccard index among 200 replications for the simulation study 2 with nonlinear structure based on PH model. The left, medium, and right plots are based on censoring rates of 30%, 50%, and 70%, respectively. Full-size DOI: 10.7717/peerj.13098/fig-6

Table 2 Results (median of prediction accuracy of different methods in the TCGA ESCA data over five random splits of 294:74 training/test sets).

	PL	SIS	FAST	RCDCS	CRCDCS	CINDEX	IPCW-tau	NPN-MB	Ordinary-MCP	*Du et al. (2021)*
Deviance	17.4067	−5.6785	−6.2444	−2.79612	−3.7387	2.7191	1.0834	−19.7218	474.4513	0.3548
c-index	0.6542	0.6314	0.6690	0.6866	0.6943	0.7324	0.6236	0.7380	0.8466	0.5450
NOSF	14	4	13	8	8	12	12	9	36	2

Note:
All feature screening methods and a published biomarker genes model are applied together with the MCP penalized Cox regression.

According to the average of Jaccard similarity index (Figs. 3, 6, S3, S6, S9, S12), we find that screening methods belonging to the same category have higher similarity than screening methods not belonging to the same category.

A further simulation study is conducted under the scenario where the survival time follows a proportional odds (PO) model (*i.e.*, the error term follows a standard logistic distribution) and all other settings are the same as those in the previous simulation study. We still obtain similar numeric results and the proposed IPCW-tau (NPN-MB) method has a better performance than the alternative methods at most gene structures. These correspondence figures are all available at Figshare: https://figshare.com/articles/figure/Survival_Feature_Screening_based_on_proportional_odds_model/17089013.

Real data application with TCGA ESCA data

After excluding patients with missing survival time data, our analysis is focused on the subset of the TCGA ESCA data with 368 patients and 20,501 gene expression variables. The censoring rate in the data is about 58%. As the number of disease-associated biomarkers is not expected to be large, the top 2,000 genes with the smallest *p*-values based on marginal Cox's model are selected for downstream analysis. We take five random splits of the whole data into 294:74 training/test sets of the data to evaluate the performance of all methods for survival prediction in the TCGA ESCA data.

According to the procedure of *Wang & Chen (2021)*, we apply eight screening methods, "PL", "SIS", "FAST", "RCDCS", "CRCDCS", "CINDEX" "IPCW-tau", "IPCW-tau (NPN-MB)", to the TCGA ESCA data. After grid search from the top 10 to the top 300 ranked genes, the best overall prediction performance of all methods is attained by using the top 150 genes, so the top-ranked 150 predictors are selected as the candidate covariates for each method and the Cox's regression model with the candidate covariates and the MCP penalty (*Zhang, 2010*) is applied to the training data to establish the final prediction model. Besides, the MCP-penalized Cox model with the top 2,000 genes selected by the univariate Cox's test is applied to the training data to build the prediction model. We also take the published biomarker genes (*DNAJB1, BNIP1, VAMP7, TBK1*) related to ESCA (*Du et al., 2021*) as a survival prediction model to make comparisons.

The prediction accuracy performances for different methods are evaluated and the numerical results are provided in Table 2 that reports the median of the survival prediction results among five treatments. Overall, we can see that the proposed IPCW-tau (NPN-MB)

Table 3 Selected genes with their correspondence estimate by IPCW-tau (NPN-NB) screening procedure with MCP penalty for the whole TCGA ESCA data.

gene	Estimate	Citation
ATRX	0.7686	
C16orf80	0.9168	
C16orf87	−0.2953	
FAM189A2	−0.1655	
GFPT1	0.9318	*Zhang et al. (2020b)*
HNRNPC	1.1911	*Xu, Pan & Pan (2020)*
MAF	0.4708	
NEK1	−0.9591	
OSTM1	0.7367	
TSKU	−0.3521	

method outperforms the alternative methods for survival prediction in the TCGA ESCA test data.

In addition, we apply the proposed IPCW-tau (NPN-MB) method for whole data to identify several important biomarker genes and estimate the correspondence parameters by penalized Cox's regression model with the MCP penalty. Please see Table 3 for the list of selected associated predictors with their correspondence weights. We identify ten genes and find the two genes *(GFPT1, HNRNPC*) genes that are related to ESCA in the literature (*Zhang et al., 2020b*; and *Xu, Pan & Pan, 2020*).

Real data application with TCGA PAAD data

After excluding patients with missing survival time data, our analysis is focused on the subset of the TCGA PAAD data with 178 patients and 20,501 gene expression variables. The censoring rate in the data is about 48%. As the number of disease-associated biomarkers is not expected to be large, the top 2,000 genes with the smallest p-values based on marginal Cox's model are selected for downstream analysis. We take five random splits of the whole data into 142:36 training/test sets of the data to evaluate the performance of all methods for survival prediction in the TCGA PAAD data.

According to the procedure of *Wang & Chen (2021)*, we apply eight screening methods, "PL", "SIS", "FAST", "RCDCS", "CRCDCS", "CINDEX" "IPCW-tau", "IPCW-tau (NPN-MB)", to the TCGA PAAD data. After grid search from the top 10 to the top 300 ranked genes, the best overall prediction performance of all methods is attained by using the top 20 genes, so the top-ranked 20 predictors are selected as the candidate covariates for each method, and the Cox's regression model with the candidate covariates and the MCP penalty (*Zhang, 2010*) is applied to the training data to establish the final prediction model. Besides, the MCP-penalized Cox model with the top 2,000 genes selected by the univariate Cox's test is applied to the training data to build the prediction model. We also take the published biomarker genes (*CDKN2A, TP53, TTN, KRAS*) related to PAAD (*Baek & Lee, 2020*) as a survival prediction model to make comparisons.

Table 4 Results (median of prediction accuracy of different methods in the TCGA PAAD data over 5 random splits of 142:36 training/test sets).

	PL	SIS	FAST	RCDCS	CRCDCS	CINDEX	IPCW-tau	NPN-MB	Ordinary-MCP	*Baek & Lee (2020)*
Deviance	−6.3363	−2.3062	−7.1140	2.5505	−6.31712	−5.1000	−3.3673	−9.7919	887.5797	−4.4676
c-index	0.6834	0.6457	0.6608	0.5955	0.6774	0.6387	0.6538	0.6834	0.5290	0.7048
NOSF	5	4	3	6	6	6	5	2	39	1

Note:
All feature screening methods and a published biomarker genes model are applied together with the MCP penalized Cox regression.

Table 5 Selected genes with their correspondence estimate by IPCW-tau (NPN-NB) screening procedure with MCP penalty for the whole TCGA PAAD data.

gene	Estimate	Citation
MET	0.5718	*Li et al. (2021)*, *Huang et al. (2021)*, *Wu et al. (2019a)*, *Vanderwerff et al. (2019)*, and *Li et al. (2019)*
ZMAT1	−0.1422	

The prediction accuracy performances for different methods are evaluated and the numerical results are provided in Table 4 that reports the median of the survival prediction results among five folds. Overall, we can see that the proposed IPCW-tau (NPN-MB) method outperforms the alternative methods for survival prediction in the TCGA PAAD test data.

In addition, we apply the proposed IPCW-tau (NPN-MB) method for whole data to identify several important biomarker genes and estimate the correspondence parameters by penalized Cox's regression model with the MCP penalty. Please see Table 5 for the list of selected associated predictors with their correspondence weights. We identify two genes (*MET*, *ZMAT1*) and find the *MET* gene that is related to PAAD in the literature (*Li et al., 2021*; *Huang et al., 2021*; *Wu et al., 2019a*; *Vanderwerff et al., 2019*; and *Li et al., 2019*).

Finally, the analysis results for TCGA HNSCC, TCGA LUAD, and TCGA BRCA data are provided in supplementary materials. Note that we take the published biomarker genes (*GIMAP6, SELL, TIFAB, KCNA3, CCR4*) related to HNSCC (*Ran et al., 2021*); (*ALK, BRAF, EGFR, ROS1*) related to LUAD (*Chen et al., 2021*); (*TMEM190, TUBA3D, LYVE1, LILBR5, CD209*) related to BRCA (*Liu et al., 2019*) as a survival prediction model to make comparisons. We identify nine genes and find the four genes (*PITPNM3*, *MXD4*, *ABCB1*, *BATF*) that are related to HNSCC in the literature (*Aravind et al., 2021*; *Wu et al., 2019b*; *da Silva et al., 2021*; *Duz & Karatas, 2021*; *Wang et al., 2020*; and *Wen et al., 2015*). We identify fifteen genes and find the seven genes (*EPB41L5*, *INPP5J*, *KRT16*, *MS4A1*, *MYLIP*, *PEBP1*, *SFTPB*) that are related to LUAD in the literature (*Li et al., 2020a*; *Zhang et al., 2020a*; *Yuanhua et al., 2019*; *Song et al., 2020*; *Liu et al., 2021b*; *Li et al., 2020b*; *Zhang et al., 2021*; *Cao et al., 2021*; and *Zhang et al., 2019*). We identify ten genes and find the four genes (*EDA2R*, *PCMT1*, *QPRT*, *SKP1*) that are related to BRCA in the literature (*Liu, Kain & Wang, 2012*; *Kyritsis et al., 2021*; *Liu et al., 2021a*; and *Tian*

et al., 2020). The proposed IPCW-tau (NPN-MB) method consistently performs well in these cancer datasets (refer to Tables S2, S4, and S6).

CONCLUSIONS AND DISSCUSSIONS

The identification of cancer-related genes in high-dimensional genetic/genomic data is a challenging and important issue. In particular, right-censored survival outcomes and contaminated biomarker data make relevant feature screening difficult. A two-step statistical algorithm is used to achieve this (*Fan & Lv, 2008*). The first step is preliminary feature screening to identify biomarkers that may be associated with cancer, then the regularization approach is used to conduct the final variable selection and parameter estimation simultaneously.

The first purpose of this article is to conduct a systematic simulation study to validate the performance of the advanced feature screening methods in variable selection and survival prediction error. We prove that for most types of gene structures, the performance of the new network-adjusted feature screening method is better than most effective univariate independent feature screening methods. The second purpose of this article is to establish a survival prediction model for TCGA survival genomic data. We prove that, compared with alternative methods that do not consider feature network information or outlier-contamination, and the published biomarker genes models, the new network adjustment feature screening method can lead to more accurate survival prediction, and determine biomarkers that are associated with clinical survival outcomes of patients with ESCA, PAAD, HNSCC, LUAD and BRCA using TCGA genetic data. These applications show that the new network-adjusted feature selection method performs well and outperforms the existing popular univariate independent feature selection methods and the published biomarker genes models. We have also identified cancer-related genes almost detected in the literature. Accordingly, the new network-based screening method is reliable and credible. R codes for the simulation studies and real data are available at Figshare: https://figshare.com/articles/software/CODE/16677070.

According to simulation studies for the *c*-index measure, we observe that the FAST method has a higher *c*-index when the number of selected variables is larger, and the IPCW-tau (NPN-MB) method outperforms most alternative methods when the number of selected variables is medium or small. Although in the real data application, the *c*-index measure of the IPCW-tau (NPN-MB) method is not always the best among all considered feature screening methods or the published biomarker gene models, the IPCW-tau (NPN-MB) method still outperforms most alternative methods by the different prediction metrics. In addition, we observe every method can also identify the biomarker genes that are related to TCGA cancer in the literature. However, according to the simulation studies and real data analysis, we can still infer that the IPCW-tau (NPN-MB) method has better performance in both variable selection and survival prediction. To this end, the IPCW-tau (NPN-MB) method is a good choice for developing a survival prediction model.

The main purpose of *Wang & Chen (2021)* is to develop the advanced network-based Kendall's tau feature screening, and in simulation studies, only compared a network-based

measure to partial likelihood screening method and IPCW Kendall's tau statistics, not provide a systematic comparison for some popular advanced feature screening methods with survival outcome. As a consequence, the first purpose of this article was to review multiple screening approaches systematically and make comparisons under the various simulated scenarios with more evaluation performance like *c*-index for prediction errors, overlap coefficient index, and the Jaccard index. Moreover, in real data applications, *Wang & Chen (2021)* apply a few feature screening methods to only two real data. We apply more feature screening methods and the published gene signature models to five TCGA cancer genomic data, and provide an optimal survival prediction model for patients. Furthermore, we also provide the selected biomarker genes with their correspondence weights, which has meaning for clinical significance.

In the real data application, we adopt the hard thresholding rule proposed by *Fan & Lv (2008)* to select the candidate set of predictors; that is, after ranking the predictors using a certain correlation measure, we select a prefixed number of top-ranked predictors as our candidate model. Several alternative strategies for thresholding rule can be considered, such as the soft thresholding rule proposed by *Zhu et al. (2011)*, a method based on the control of the false-positive rate or false discovery rate by *Zhao & Li (2012)*, and a method based on multiple testing procedure by *Song et al. (2014)*. In addition, we assume that the number of cancer-related biomarkers will not be large, so we select the top 2,000 genes with the smallest *p*-value for downstream analysis based on the marginal Cox's model. Different candidate models lead to different survival prediction models. How to define the number of cancer-related biomarkers for downstream analysis is a key and interesting issue.

There are several public human databases, like METABRIC, TCGA, NCDB, and GEO. These useful databases can be utilized to determine the reproducibility of our findings. We consider that meta-analysis can be performed for discovery and validation of survival biomarker genes (*Xu et al., 2020*), which is worthy of further research and will be studied in our future work.

LIST OF ABBREVIATIONS

TCGA	The Cancer Genome Atlas
ESCA	esophageal carcinoma
PAAD	pancreatic adenocarcinoma
HNSCC	head and neck squamous cell carcinoma
LUAD	lung adenocarcinoma
BRCA	breast invasive carcinoma
PL	partial likelihood
FAST	feature aberration at survival times
SIS	sure independence screening
RCDCS	robust censored distance correlation screening
CRCDCS	composite robust censored distance correlation screening
CINDEX	Harrell's concordance index

IPCW	inverse probability-of-censoring weighted
NPN	nonparanormal
MMS	minimum model size
PH	proportional hazards
PO	proportional odds
NOSF	number of selected features
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
NCDB	National Cancer Database
GEO	Gene Expression Omnibus

ACKNOWLEDGEMENTS

We are very grateful to the associate editor and the referees for their very valuable comments and suggestions that helped to improve the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the grant MOST 110-2118-M-035-001-MY2 from the Ministry of Science and Technology of Republic of China (Taiwan). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Ministry of Science and Technology of Republic of China: MOST 110-2118-M-035-001-MY2.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Jie-Huei Wang conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Cai-Rong Li performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Po-Lin Hou performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

R codes for the simulation studies and real data are available at Figshare: Wang, Jie-Huei (2021): The R code for the paper entitled "Feature Screening for Survival Trait with

Application to TCGA High-dimensional Genomic Data". figshare. Software. https://doi.org/10.6084/m9.figshare.16677070.v3.

The TCGA ESCA, PAAD, LUAD, and BRCA genomic data with survival traits analyzed during this study are all available at Figshare: Wang, Jie-Huei (2021): The TCGA cancer data for the paper entitled "Feature Screening for Survival Trait with Application to TCGA High-dimensional Genomic Data". figshare. Dataset. https://doi.org/10.6084/m9.figshare.16677619.v2.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.13098#supplemental-information.

REFERENCES

Adler D, Kelly ST. 2021. vioplot: violin plot. R package version 0.3.6. *Available at https://github.com/TomKellyGenetics/vioplot*.

Aravind A, Palollathil A, Rex D, Kumar K, Vijayakumar M, Shetty R, Codi J, Prasad T, Raju R. 2021. A multi-cellular molecular signaling and functional network map of C-C motif chemokine ligand 18 (CCL18): a chemokine with immunosuppressive and pro-tumor functions. *Journal of Cell Communication and Signaling* **225(3)**:792 DOI 10.1007/s12079-021-00633-3.

Baek B, Lee H. 2020. Prediction of survival and recurrence in patients with pancreatic cancer by integrating multi-omics data. *Scientific Reports* **10(1)**:18951 DOI 10.1038/s41598-020-76025-1.

Cao B, Wang P, Gu L, Liu J. 2021. Use of four genes in exosomes as biomarkers for the identification of lung adenocarcinoma and lung squamous cell carcinoma. *Oncology Letters* **21(4)**:249 DOI 10.3892/ol.2021.12510.

Chen X, Chen X, Liu Y. 2019. A note on quantile feature screening via distance correlation. *Statistical Papers* **60**:1741–1762 DOI 10.1007/s00362-017-0894-8.

Chen X, Chen X, Wang H. 2018. Robust feature screening for ultra-high dimensional right censored data via distance correlation. *Computational Statistics & Data Analysis* **119(5)**:118–138 DOI 10.1016/j.csda.2017.10.004.

Chen L, Zeng H, Xiang Y, Huang Y, Luo Y, Ma X. 2021. Histopathological images and multi-omics integration predict molecular characteristics and survival in lung adenocarcinoma. *Frontiers in Cell and Developmental Biology* **9**:720110 DOI 10.3389/fcell.2021.720110.

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H. 2016. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44(8)**:71 DOI 10.1093/nar/gkv1507.

da Silva G, de Matos LL, Kowalski LP, Kulcsar M, Leopoldino AM. 2021. Profile of sphingolipid-related genes and its association with prognosis highlights sphingolipid metabolism in oral cancer. *Cancer Biomarkers: Section A of Disease Markers* **32(1)**:49–63 DOI 10.3233/CBM-203100.

Du H, Xie S, Guo W, Che J, Zhu L, Hang J, Li H. 2021. Development and validation of an autophagy-related prognostic signature in esophageal cancer. *Annals of Translational Medicine* **9(4)**:317 DOI 10.21037/atm-20-4541.

Duz MB, Karatas OF. 2021. Differential expression of ABCB1, ABCG2, and KLF4 as putative indicators for paclitaxel resistance in human epithelial type 2 cells. *Molecular Biology Reports* **48(2)**:1393–1400 DOI 10.1007/s11033-021-06167-6.

Edelmann D, Hummel M, Hielscher T, Saadati M, Benner A. 2020. Marginal variable screening for survival endpoints. *Biometrical Journal* **62(3)**:610–626 DOI 10.1002/bimj.201800269.

Fan J, Feng Y, Wu Y. 2010. High-dimensional variable selection for Cox proportional hazards model. *IMS Collect* **6**:70–86 DOI 10.1214/10-IMSCOLL606.

Fan J, Lv J. 2008. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B* **70(5)**:849–911 DOI 10.1111/j.1467-9868.2008.00674.x.

Gorst-Rasmussen A, Scheike TH. 2012. Coordinate descent methods for the penalized semiparametric additive hazards model. *Journal of Statistical Software* **47(9)**:1–17 DOI 10.18637/jss.v047.i09.

Harrell FE Jr, Lee KL, Mark DB. 1996. Multivariable prognostic models: issues in. developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15(4)**:361–387 DOI 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

Huang X, Tang T, Zhang G, Liang T. 2021. Identification of tumor antigens and immune subtypes of cholangiocarcinoma for mRNA vaccine development. *Molecular Cancer* **20(1)**:50 DOI 10.1186/s12943-021-01342-6.

Kyritsis KA, Akrivou MG, Giassafaki LN, Grigoriadis NG, Vizirianakis IS. 2021. Analysis of TCGA data of. differentially expressed EMT-related genes and miRNAs across various malignancies to identify potential biomarkers. *World Academy of Sciences Journal* **3(1)**:6 DOI 10.3892/wasj.2020.77.

Li A, Hou S, Chen J, Jiang Y. 2021. Development and validation of a novel glycolysis-related risk signature for predicting survival in pancreatic adenocarcinoma. *Clinica Chimica Acta; International Journal of Clinical Chemistry* **518(11)**:156–161 DOI 10.1016/j.cca.2021.03.020.

Li J, Li Z, Zhao S, Song Y, Si L, Wang X. 2020a. Identification key genes, key miRNAs and key transcription factors of lung adenocarcinoma. *Journal of Thoracic Disease* **12(5)**:1917–1933 DOI 10.21037/jtd-19-4168.

Li H, Tong L, Tao H, Liu Z. 2020b. Genome-wide analysis of the hypoxia-related DNA methylation-driven genes in lung adenocarcinoma progression. *Bioscience Reports* **40(2)**:BSR20194200 DOI 10.1042/BSR20194200.

Li Y, Zhu YY, Dai GP, Wu DJ, Gao ZZ, Zhang L, Fan YH. 2019. Screening and validating the core biomarkers in patients with pancreatic ductal adenocarcinoma. *Mathematical Biosciences and Engineering: MBE* **17(1)**:910–927 DOI 10.3934/mbe.2020048.

Liu L, Chen Z, Shi W, Liu H, Pang W. 2019. Breast cancer survival prediction using seven prognostic biomarker genes. *Oncology Letters* **18(3)**:2907–2916 DOI 10.3892/ol.2019.10635.

Liu CL, Cheng SP, Chen MJ, Lin CH, Chen SN, Kuo YH, Chang YC. 2021a. Quinolinate phosphoribosyltransferase promotes invasiveness of breast cancer through myosin light chain phosphorylation. *Frontiers in Endocrinology* **11**:621944 DOI 10.3389/fendo.2020.621944.

Liu R, Kain M, Wang L. 2012. Inactivation of X-linked tumor suppressor genes in human cancer. *Future Oncology* **8(4)**:463–481 DOI 10.2217/fon.12.26.

Liu H, Lafferty J, Wasserman L. 2009. The nonparanormal semiparametric estimation of. high dimensional undirected graphs. *Journal of Machine Learning Research* **10(80)**:2295–2328 DOI 10.1145/1577069.1755863.

Liu T, Yang C, Wang W, Liu C. 2021b. LncRNA SGMS1-AS1 regulates lung adenocarcinoma cell proliferation, migration, invasion, and EMT progression via miR-106a-5p/MYLI9 axis. *Thoracic Cancer* **12(4)**:2104–2112 DOI 10.1111/1759-7714.14043.

Meinshausen N, Buhlmann P. 2006. High dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34(3)**:1436–1462 DOI 10.1214/009053606000000281.

Ran QC, Long SR, Ye Y, Xie C, XuXiao ZL, Liu YS, Pang HX, Sunchuri D, Teng NC, Guo ZL. 2021. Mining TCGA database for prognostic genes in head and neck squamous cell carcinoma microenvironment. *Journal of Dental Sciences* **16(2)**:661–667 DOI 10.1016/j.jds.2020.09.017.

Saldana DF, Feng Y. 2018. SIS: an R package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software* **83(2)**:1–25 DOI 10.18637/jss.v083.i02.

Song C, Guo Z, Yu D, Wang Y, Wang Q, Dong Z, Hu W. 2020. A prognostic nomogram combining immune-related gene signature and clinical factors predicts survival in patients with lung adenocarcinoma. *Frontiers in Oncology* **10**:1300 DOI 10.3389/fonc.2020.01300.

Song R, Lu W, Ma S, Jeng XJ. 2014. Censored rank independence screening for high-dimensional survival data. *Biometrika* **101(4)**:799–814 DOI 10.1093/biomet/asu047.

The Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455(7216)**:1061–1068 DOI 10.1038/nature07385.

Tian Z, He W, Tang J, Liao X, Yang Q, Wu Y, Wu G. 2020. Identification of important modules and biomarkers in breast cancer based on WGCNA. *OncoTargets and Therapy* **13**:6805–6817 DOI 10.2147/OTT.S258439.

Vanderwerff BR, Church KJ, Kawas LH, Harding JW. 2019. Comparative characterization of the HGF/Met and MSP/Ron systems in primary pancreatic adenocarcinoma. *Cytokine* **123(185–203)**:154762 DOI 10.1016/j.cyto.2019.154762.

Wang JH, Chen YH. 2020. Interaction screening by Kendall's partial correlation for. ultrahigh-dimensional data with survival trait. *Bioinformatics* **36(9)**:2763–2769 DOI 10.1093/bioinformatics/btaa017.

Wang JH, Chen YH. 2021. Network-adjusted Kendall's Tau measure for feature screening with application to high-dimensional survival genomic data. *Bioinformatics* **37(15)**:2150–2156 DOI 10.1093/bioinformatics/btab064.

Wang S, Liu X. 2019. The UCSCXenaTools R package: a toolkit for accessing genomics. data from UCSC Xena platform, from cancer multi-omics to single-cell RNA-seq. *The Journal of Open Source Software* **4(40)**:1627 DOI 10.21105/joss.01627.

Wang Y, Xu Y, Hua Q, Jiang Y, Liu P, Zhang W, Xiang R. 2020. Novel prognostic model based on immune signature for head and neck squamous cell carcinoma. *BioMed Research International* **2020(11)**:4725314 DOI 10.1155/2020/4725314.

Wei T, Simko V. 2017. Visualization of a correlation matrix. Version 0.84. *Available at https://CRAN.R-project.org/package=corrplot*.

Wen H, Tang J, Liu B, Sun C. 2015. The expression and clinical significance of BATF2 in oral tongue squamous cell carcinoma. *Chinese Journal of Stomatology* **50(1)**:13–17 DOI 10.3760/CMA.J.ISSN.1002-0098.2015.01.004.

Wu M, Li X, Zhang T, Liu Z, Zhao Y. 2019a. Identification of a nine-gene signature and establishment of a prognostic nomogram predicting overall survival of pancreatic cancer. *Frontiers in Oncology* **9**:996 DOI 10.3389/fonc.2019.00996.

Wu M, Qin X, Ma S. 2021. GEInter: an R package for robust gene-environment interaction analysis. *Bioinformatics* **37(20)**:3691–3692 DOI 10.1093/bioinformatics/btab318.

Wu Y, Wang Y, Diao P, Zhang W, Li J, Ge H, Song Y, Li Z, Wang D, Liu L, Jiang H, Cheng J. 2019b. Therapeutic targeting of BRD4 in head neck squamous cell carcinoma. *Theranostics* **9(6)**:1777–1793 DOI 10.7150/thno.31581.

Wu M, Zhu L, Feng X. 2018. Network-based feature screening with applications to genome data. *Annals of Applied Statistics* **12(2)**:1250–1270 DOI 10.1214/17-AOAS1097.

Xu M, Li Y, Li W, Zhao Q, Zhang Q, Le K, Huang Z, Yi P. 2020. Immune and stroma related genes in breast cancer: a comprehensive analysis of tumor microenvironment based on The Cancer Genome Atlas (TCGA) database. *Frontiers in Medicine* **7**:64 DOI 10.3389/fmed.2020.00064.

Xu LC, Pan JX, Pan HD. 2020. Construction and validation of an m6A RNA methylation regulators-based prognostic signature for esophageal cancer. *Cancer Management and Research* **12**:5385–5394 DOI 10.2147/CMAR.S254870.

Yuanhua L, Pudong Q, Wei Z, Yuan W, Delin L, Yan Z, Geyu L, Bo S. 2019. TFAP2A induced KRT16 as an oncogene in lung adenocarcinoma via EMT. *International Journal of Biological Sciences* **15(7)**:1419–1428 DOI 10.7150/ijbs.34076.

Zhang CH. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38(2)**:894–942 DOI 10.1214/09-AOS729.

Zhang Y, Fan Q, Guo Y, Zhu K. 2020a. Eight-gene signature predicts recurrence in lung adenocarcinoma. *Cancer Biomarkers : Section A of Disease Markers* **28(4)**:447–457 DOI 10.3233/CBM-190329.

Zhang L, Li M, Deng B, Dai N, Feng Y, Shan J, Yang Y, Mao C, Huang P, Xu C, Wang D. 2019. HLA-DQB1 expression on tumor cells is a novel favorable prognostic factor for relapse in early-stage lung adenocarcinoma. *Cancer Management and Research* **11**:2605–2616 DOI 10.2147/CMAR.

Zhang C, Lian H, Xie L, Yin N, Cui Y. 2020b. LncRNA ELFN1-AS1 promotes esophageal cancer progression by up-regulating GFPT1 via sponging miR-183-3p. *Biological Chemistry* **401(9)**:1053–1061 DOI 10.1515/hsz-2019-0430.

Zhang A, Yang J, Ma C, Li F, Luo H. 2021. Development and validation of a robust ferroptosis-related prognostic signature in lung adenocarcinoma. *Frontiers in Cell and Developmental Biology* **9**:616271 DOI 10.3389/fcell.2021.616271.

Zhao T, Han L, Roeder K, Lafferty JD, Wasserman LA. 2012. The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research* **13**:1059–1062 DOI 10.3744/JNAOE.2012.4.4.403.

Zhao SD, Li Y. 2012. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis* **105(1)**:397–411 DOI 10.1016/j.jmva.2011.08.002.

Zhong W, Zhu L, Li R, Cui H. 2016. Regularized quantile regression and robust feature screening for single index models. *Statistica Sinica* **26(1)**:69–95 DOI 10.5705/ss.2014.049.

Zhu L, Li L, Li R, Zhu L. 2011. Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association* **106(496)**:1464–1475 DOI 10.1198/jasa.2011.tm10563.