Web-based Supplementary Information for "**Interaction Screening by Kendall's Partial Correlation for Ultrahigh-dimensional Data with Survival Trait**" by

Jie-Huei Wang and Yi-Hau Chen

Department of Statistics, Feng Chia University, Taichung, 40724, Taiwan

and

Institute of Statistical Science, Academia Sinica, Taipei, 11529, Taiwan

A. Proof of the unbiasedness of the inverse probability-of-censoring weighted estimators for the concordance and disconcordance probabilities:

$$\mathrm{E}\left\{\frac{\Delta_{(j)}}{S_C^2\left(V_{(j)}\right)} I\left(V_{(i)} > V_{(j)}, X_{(i)} > X_{(j)}\right)\right\} = \Pr\left(T_{(i)} > T_{(j)}, X_{(i)} > X_{(j)}\right),$$

$$\mathrm{E}\left\{\frac{\Delta_{(j)}}{S_C^2\left(V_{(j)}\right)} I\left(V_{(i)} > V_{(j)}, X_{(i)} < X_{(j)}\right)\right\} = \Pr\left(T_{(i)} > T_{(j)}, X_{(i)} < X_{(j)}\right).$$

Proof:

$$
\begin{aligned}
&\mathrm{E}\left\{\frac{\Delta_{(j)}}{S_C^2\left(V_{(j)}\right)} I\left(V_{(i)} > V_{(j)}, X_{(i)} > X_{(j)}\right)\right\} \\
&= \mathrm{E}\left\{\frac{\Delta_{(j)}}{S_C^2\left(V_{(j)}\right)} I\left(\min\left(T_{(i)}, C_{(i)}\right) > V_{(j)}, X_{(i)} > X_{(j)}\right)\right\} \\
&= \mathrm{E}\left\{\frac{\Delta_{(j)}}{S_C^2\left(V_{(j)}\right)} I\left(T_{(i)} > V_{(j)}, C_{(i)} > V_{(j)}, X_{(i)} > X_{(j)}\right)\right\} \\
&= \mathrm{E}\left\{\frac{\Delta_{(j)}}{S_C^2\left(V_{(j)}\right)} I\left(T_{(i)} > V_{(j)}\right) I\left(C_{(i)} > V_{(j)}\right) I\left(X_{(i)} > X_{(j)}\right)\right\} \\
&= E\left\{E\left[\frac{\Delta_{(j)}}{S_C^2\left(V_{(j)}\right)} I\left(T_{(i)} > V_{(j)}\right) I\left(C_{(i)} > V_{(j)}\right) I\left(X_{(i)} > X_{(j)}\right) \middle| X, T\right]\right\} \\
&= E\left\{I\left(T_{(i)} > T_{(j)}, X_{(i)} > X_{(j)}\right) E\left[\frac{\Delta_{(j)}}{S_C^2\left(T_{(j)}\right)} I\left(C_{(i)} > T_{(j)}\right) \middle| X, T\right]\right\} \\
&= Pr\left(T_{(i)} > T_{(j)}, X_{(i)} > X_{(j)}\right),
\end{aligned}
$$

where the conditional set $(X,T) = (X_{(i)}, X_{(j)}, T_{(i)}, T_{(j)})$, and the last equality follows from the fact that $C_{(i)}, C_{(j)}$ are independent of each other and are independent of (*X*,*T*), and $E(I(C_{(i)} > T_{(j)} \mid T_{(j)}) = S_C(T_{(j)})$, $E(I(C_{(j)} > T_{(j)} \mid T_{(j)}) = S_C(T_{(j)})$. The proof of the second equation is similar, we thus omit it.

B. Additional simulation results

B.1 Simulations in Section 3.1 of the main text: comparing the mean numbers of the true predictors with positive and negative coefficients selected by the IPCW-tau and IPCW(S) methods

We compare the mean numbers of the true predictors with positive and negative coefficients included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables among 200 replications between the IPCW-tau and IPCW(S) methods. We report only the results with the cohort size being 300; the results for the other cohort sizes are similar and hence omitted. From Table S1 (for the scenario with covariates contaminated) we see that when the censoring rate is 25%, both approaches perform similarly well in selecting the true predictors with positive and negative coefficients. However, when the censoring rate is 50% or 80%, the IPCW(S) approach tends not to select the true predictors with negative coefficients, while the IPCW-tau approach does not have this drawback and selects effects in either direction equally well.

Table S1. Mean numbers of true predictors with the positive and negative coefficients included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables under the scenario with cohort size equal to 300 and contaminated covariates.

Cen.r=25%	15	30	45	60	75	90	120	150
IPCW-tau								
Beta>0	3.460	3.905	4.145	4.295	4.375	4.48	4.655	4.800
Beta<0	4.480	5.035	5.290	5.455	5.585	5.680	5.885	6.040
IPCW(S)								
Beta>0	3.615	4.030	4.300	4.415	4.565	4.700	4.855	5.010
Beta<0	4.090	4.635	4.925	5.085	5.235	5.355	5.530	5.705
Cen.r=50%	15	30	45	60	75	90	120	150
IPCW-tau								
Beta>0	2.995	3.350	3.600	3.775	3.910	4.030	4.185	4.335
Beta<0	3.715	4.350	4.650	4.815	4.990	5.100	5.355	5.515
IPCW(S)								
Beta>0	3.350	3.785	4.040	4.180	4.320	4.440	4.610	4.750
Beta<0	0.005	0.005	0.005	0.005	0.005	0.005	0.015	0.015
Cen.r=80%	15	30	45	60	75	90	120	150
IPCW-tau								
Beta>0	2.105	2.460	2.695	2.835	2.930	3.005	3.185	3.300
Beta<0	2.350	2.735	2.950	3.140	3.285	3.420	3.630	3.810
IPCW(S)								
Beta>0	2.465	2.825	3.020	3.200	3.315	3.44	3.62	3.84
Beta<0	0	0	0	0	0	0	0	0

B.2 Simulations in Section 3.1 of the main text with cohort sizes of 300 and 500

In Figures S1 and S2, we show the mean numbers of true predictors included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables for each method considered among 200 replications. The simulation setups are described in Section 3.1 of the main text; Figures S1 and S2 are results from the simulations with cohort sizes of 300 and 500. It can be seen that the proposed IPCW-tau method has higher mean numbers of true predictors included in the selected predictor set with a given set size.

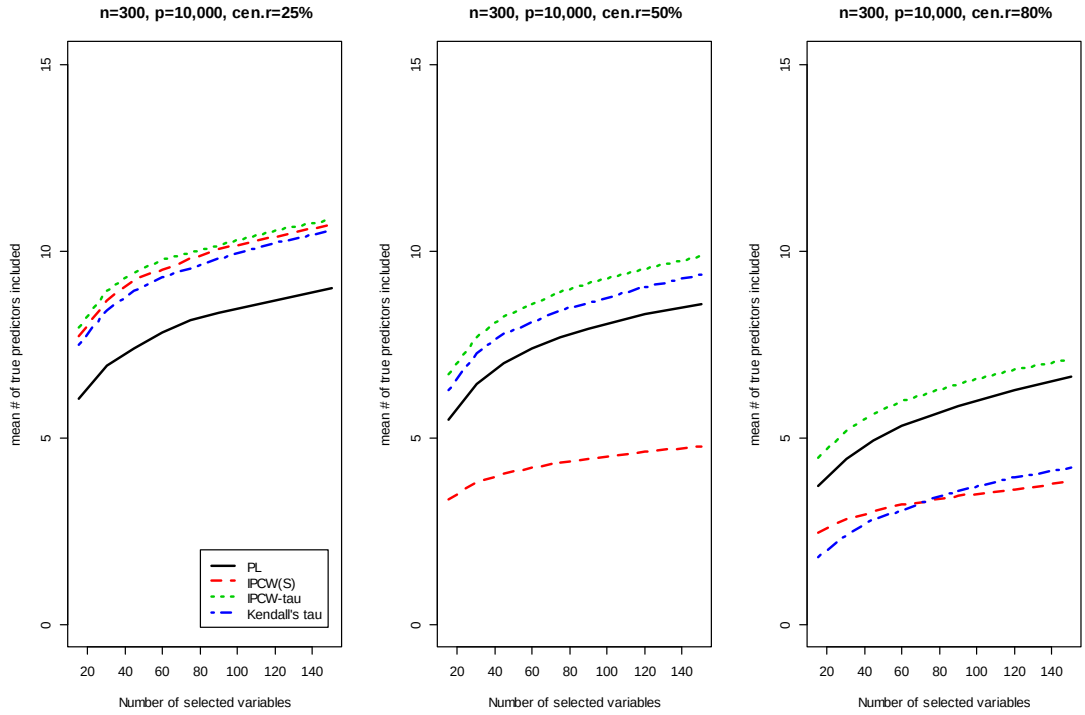


Fig. S1. Mean numbers of true predictors included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables under the scenario with cohort size equal to 300 and contaminated covariates.

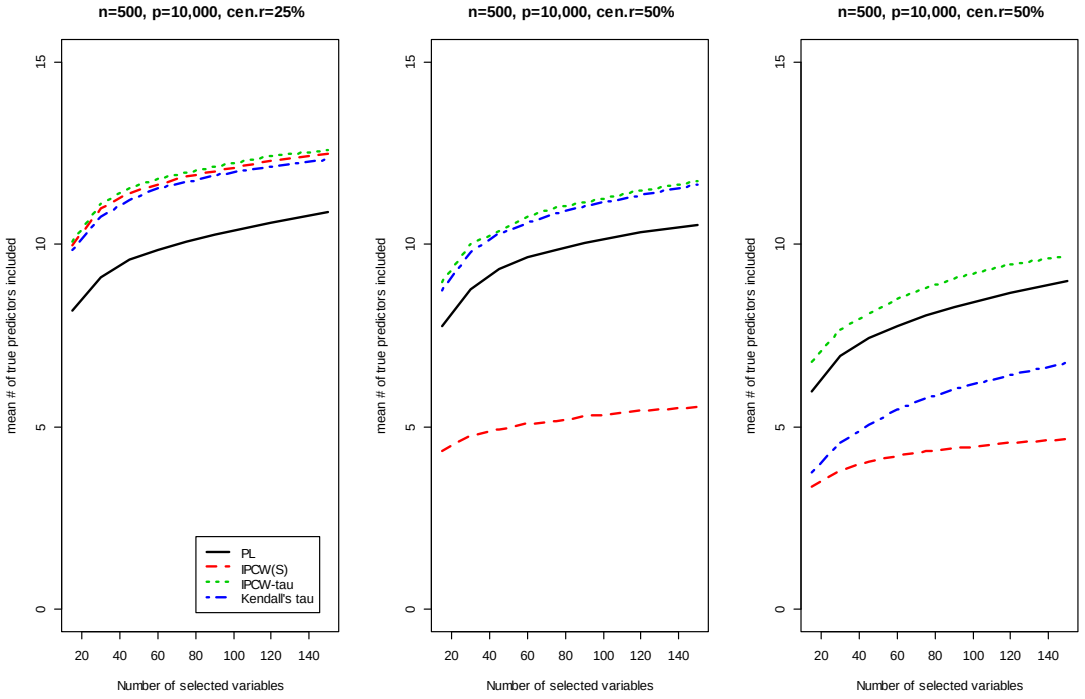


Fig. S2. Mean numbers of true predictors included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables under the scenario with cohort size equal to 500 and contaminated covariates.

B.3 Simulations in Section 3.1 of the main text with uncontaminated covariates

We report in Figures S3, S4, and S5 of the Supplementary Materials the results of simulations with covariates uncontaminated and all other settings same as those in Section 3.1 of the main text. From these figures, we see that the PL approach performs best when the censoring rate is 50% or 80%, while the IPCW-tau approach performs best when the censoring rate is 25%. The simulation results are expectable given that the PL approach is with approximately correct model specification when the covariates are multivariate normal and uncontaminated. Results in Table S2 of the Supplementary Materials, under the scenario with uncontaminated covariates (all the other setups same as those with contaminated covariates [Table S1]), lead to the same conclusion for comparison between the IPCW-tau and IPCW(S) methods as those in Table S1 under the scenario with contaminated covariates.

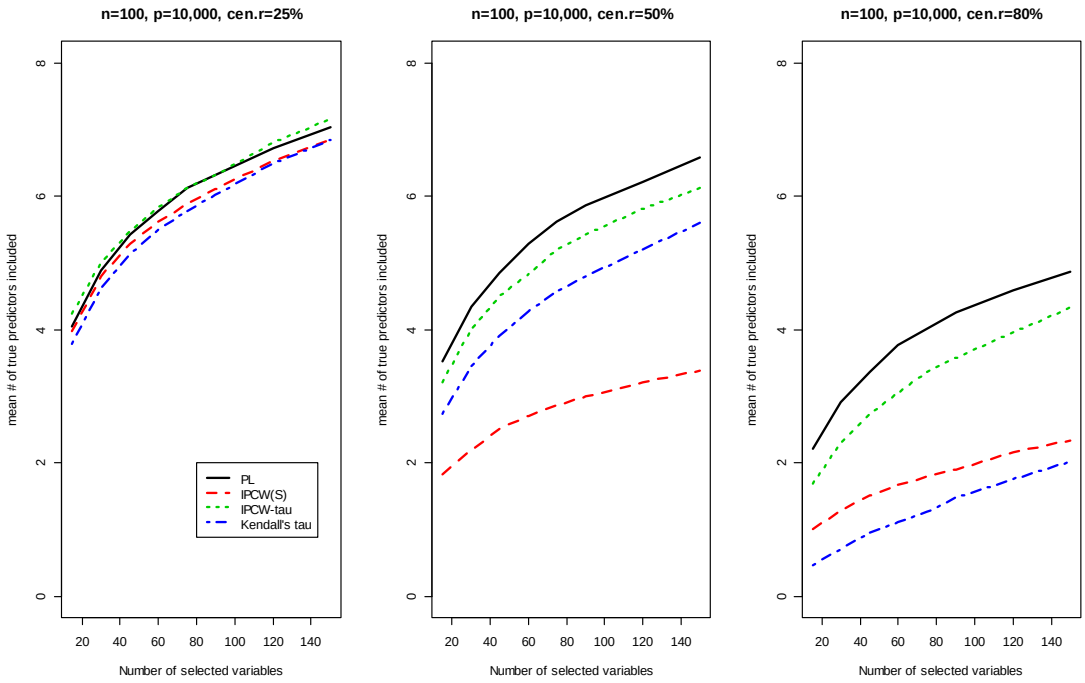


Fig. S3. Mean numbers of true predictors included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables under the scenario with cohort size equal to 100 and uncontaminated covariates.

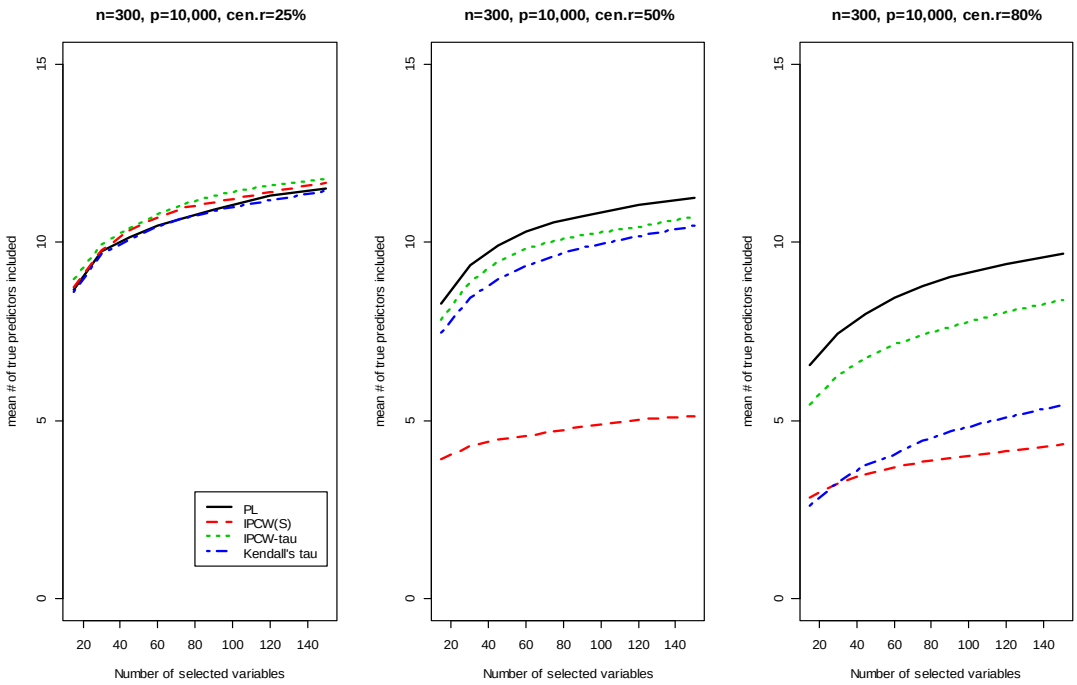


Fig. S4. Mean numbers of true predictors included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables under the scenario with cohort size equal to 300 and uncontaminated covariates.

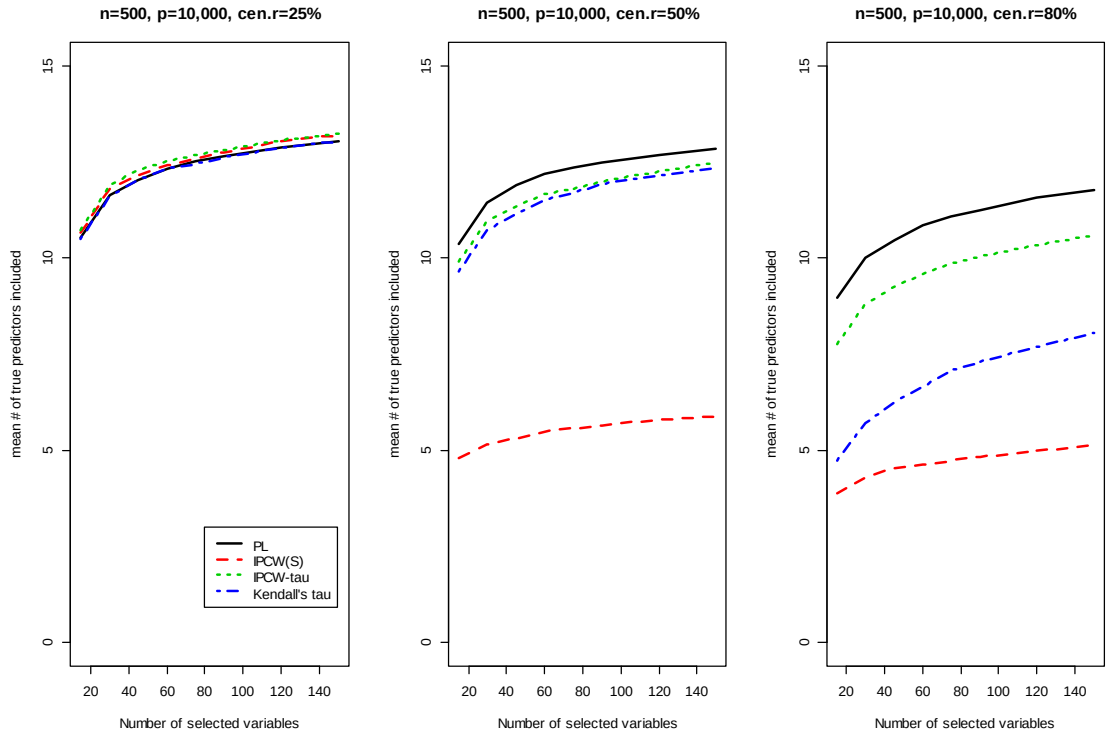


Fig. S5. Mean numbers of true predictors included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables under the scenario with cohort size equal to 500 and uncontaminated covariates.

Table S2. Mean numbers of true predictors with the positive and negative coefficients included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables under the scenario with cohort size equal to 300 and uncontaminated covariates.

	size of the selected set							
	15	30	45	60	75	90	120	150
IPCW-tau,	cen.r=25%							
Beta>0	3.865	4.270	4.485	4.690	4.840	4.980	5.090	5.205
Beta<0	5.110	5.665	5.935	6.100	6.235	6.345	6.505	6.570
IPCW(S),	cen.r=25%							
Beta>0	3.980	4.420	4.695	4.925	5.070	5.125	5.265	5.395
Beta<0	4.755	5.370	5.645	5.780	5.905	6.010	6.160	6.285
	size of the selected set							
	15	30	45	60	75	90	120	150
IPCW-tau,	cen.r=50%							
Beta>0	3.455	3.900	4.120	4.275	4.380	4.440	4.570	4.685
Beta<0	4.365	5.020	5.335	5.550	5.680	5.755	5.880	6.035
IPCW(S),	cen.r=50%							
Beta>0	3.925	4.280	4.450	4.570	4.690	4.805	4.975	5.075
Beta<0	0.005	0.005	0.015	0.020	0.020	0.025	0.040	0.055
	size of the selected set							
	15	30	45	60	75	90	120	150
IPCW-tau,	cen.r=80%							
Beta>0	2.485	2.820	3.045	3.235	3.360	3.480	3.710	3.845
Beta<0	2.980	3.460	3.705	3.920	4.050	4.150	4.355	4.525
IPCW(S),	cen.r=80%							
Beta>0	2.855	3.230	3.495	3.705	3.845	3.960	4.165	4.350
Beta<0	0	0	0	0	0	0	0	0

cen.r: censoring rate

B.4 Simulations in Section 3.2 of the main text with small effect sizes

We examine the performance of the PC-IPCW-tau approach under the settings with small effect sizes. The simulations are conducted under the same settings as in Section 3.2 (Fig. 2 and table 2) of the main text, except that the true regression coefficient vector is now given as

$$(\beta_{10},\beta_{40},\beta_{80},\beta_{510},\beta_{580},\beta_{5485},\beta_{5545},\beta_{19760}) = (-1.2, 1.5, 1.8, 1.5, -1.8, 1.5, -1.8, 1.5).$$

The simulation results shown in Figure S6 and Table S3 reveal that, the PC-IPCW-tau approach still outperforms the other methods in the small effect settings, although small effect sizes do make interaction screening less efficient.

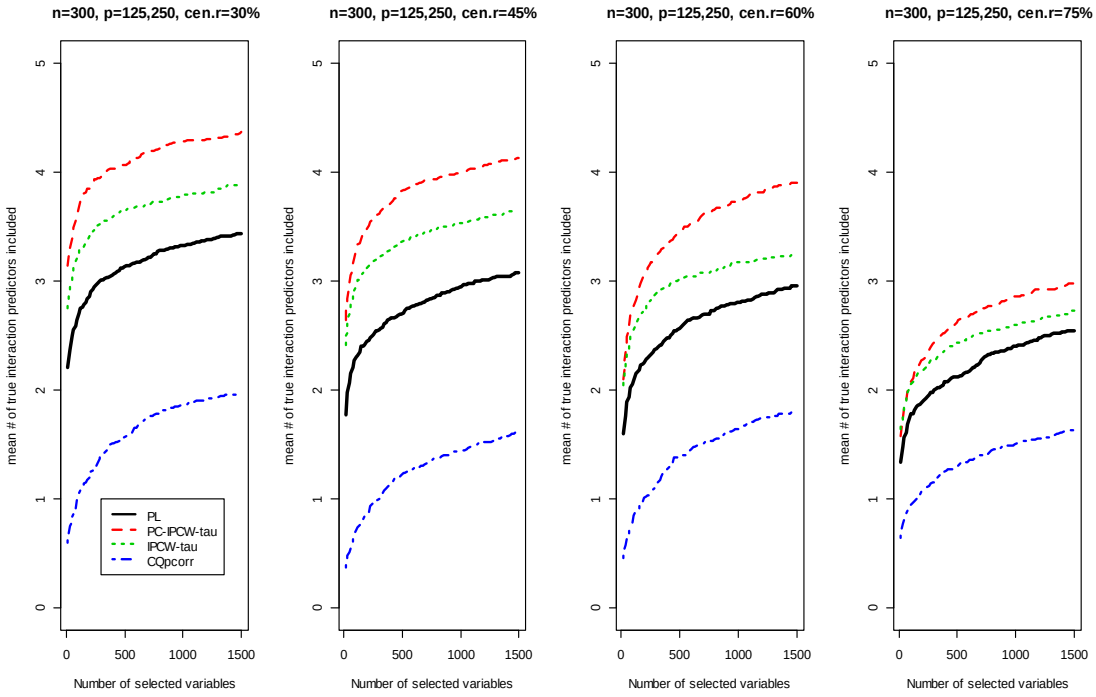


Fig. S6. Mean numbers of true second-order predictors included by the top (15, 30,…,1485,1500) selected variables under the scenario of Section 3.2 in the main text with cohort size equal to 300 and small effects.

Table S3. The median of the minimum model size out of 200 replications under the scenario of Section 3.2 in the main text with cohort size equal to 300 and small effects.

	PL	IPCW-tau	PC-IPCW-tau	CQpcorr
cen.r=30%	23675	10740	1745	76536
cen.r=45%	36872	17304	3178	82119
cen.r=60%	45939	30090	6120	89073
cen.r=75%	55440	65387	26321	86313

cen.r: censoring rate

B.5 Simulations with NSCLC gene expression dependence structure, and with log-normal gene expression distribution

We perform simulations under a gene dependence structure that is more reasonable in real gene expression data. Specifically, we directly use the gene expression data in the NSCLC data set (with n=125) as the covariates ($\boldsymbol{x}$), and simulate survival time data using the linear transformation model considered in Section 3.1 of the main text, with the "true predictors", i.e., the genes with non-zero regression coefficients, randomly selected from whole genes in the NSCLC dataset. Please see Fig. S7 for the mean numbers of the true predictors included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables among 200 replications for the PL, IPCW(S), IPCW-tau and Kendall's tau methods under different censoring rates. The proposed IPCW-tau method consistently selects more true variables over different censoring rates compared to the alternative methods.

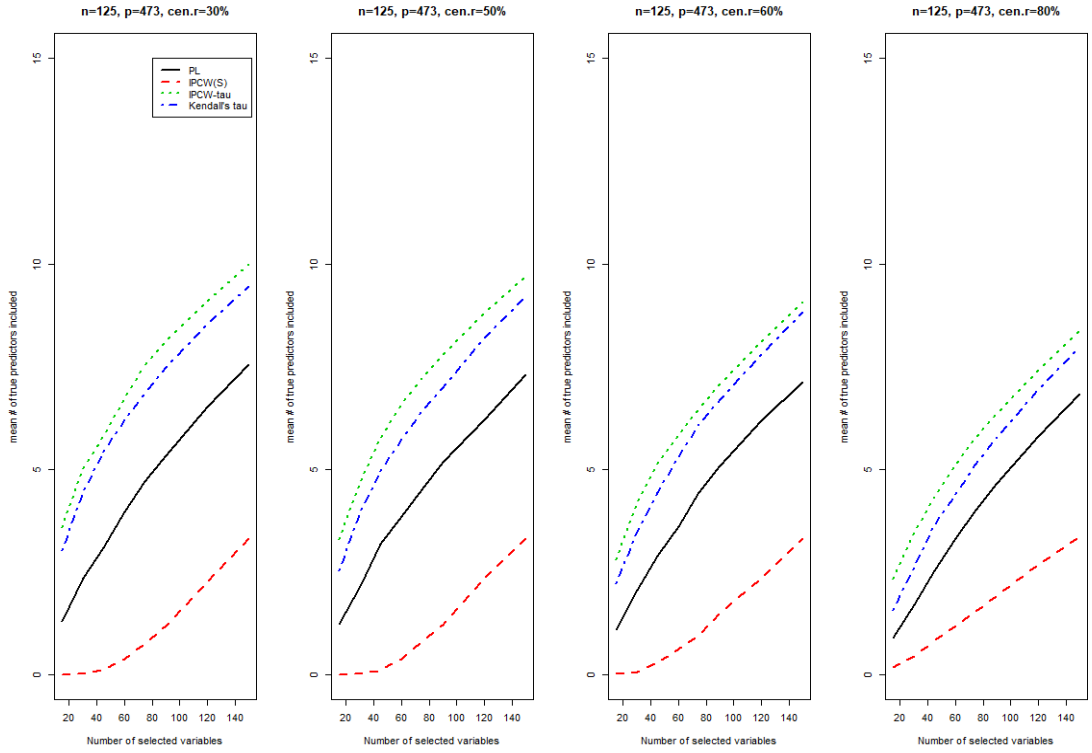


Fig. S7. Mean numbers of true predictors included by the top (15, 30, 45, 60, 75, 90, 120, 150) selected variables under the scenario of Section 3.1 in the main text with the covariates given by the NSCLC data (n=125, p=473).

Also, using the simulation setups in Section 3.2 but with the covariate data given by the NSCLC gene expression data (*n*=125, 473 primary covariates), Fig. S8 shows the mean numbers of the true second-order predictors included by the top 15, 30,…, 1485,1500 selected variables among 200 replications for the PL, IPCW-tau, PC-IPCW-tau, IPCW-tau and CQpcorr methods under different censoring rates (there are a total of 122,101 second-order predictors for 473 primary covariates.) The proposed PC-IPCW-tau method still performs best in this scenario for second-order predictor selection.

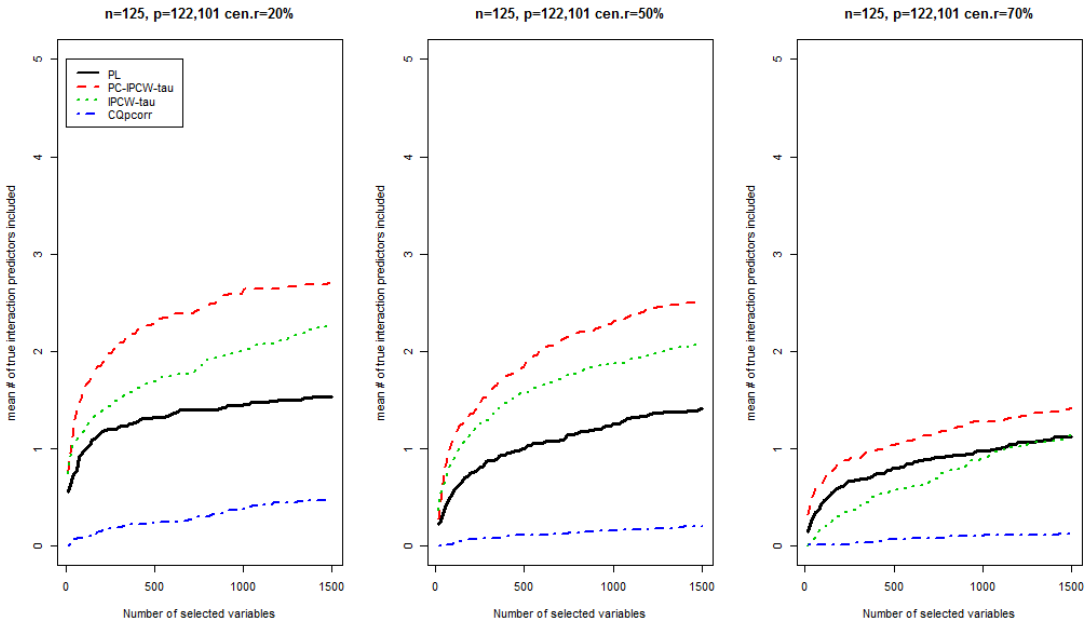


Fig. S8. Mean numbers of true second-order predictors included by the top (15, 30,…,1485,1500) selected variables under the scenario of Section 3.2 in the main text with the covariates given by the NSCLC data (n=125, p=473).

We also have performed simulations with the same setups in Section 3.1 with cohort size n=300, but with the covariate distribution changed to a multivariate log-normal distribution; namely we exponenciate the original multivariate normal covariates to obtain the new covariates in the simulations. Fig. S9 below shows the mean numbers of the true predictors included by the top 15, 30, 45, 60, 75, 90, 120 and 150 selected variables among 200 replications for the PL, IPCW(S), IPCW-tau and Kendall's tau methods under different censoring rates. The proposed IPCW-tau method still has nice performances in the case with non-normal gene expression distribution.

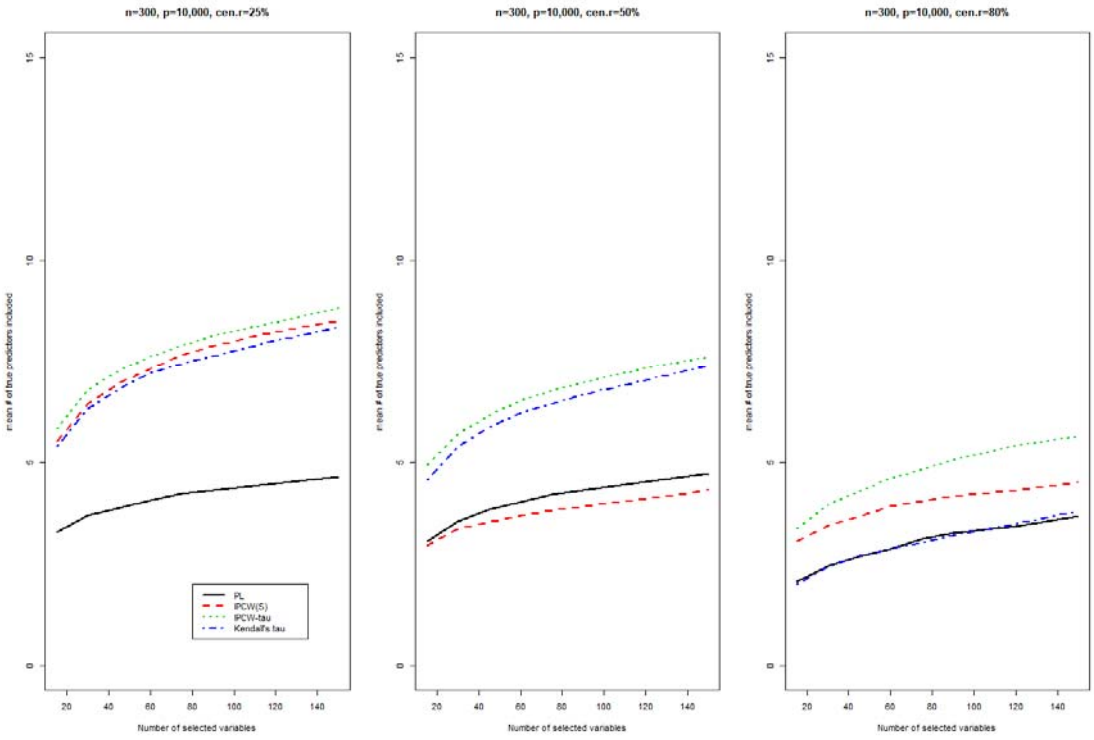


Fig. S9. Mean numbers of true predictors included by the top (15, 30, 45, 60, 75, 90, 120,150) selected variables under the scenario of Section 3.1 (n=300) in the main text with the covariate distribution given by multivariate log-normal distribution.

B.6 Simulations in Section 3.2 of the main text with alternative survival time distributions

We have performed simulations under the same setting as in Section 3.2, except that the distribution of the residual term (ε) in the linear transformation model for generating the survival time data is now given by the standard normal or standard logistic distribution, the dimension of the main covariates is 100, and the censoring rates considered are 30%, 50% and 75%. Fig. S10 (for normal distribution) and Fig. S11 (for logistic distribution) display the mean numbers of the true second-order predictors included by the top 15, 30, …, 1485, 1500 selected variables among 200 replications for different methods. We can see that in the settings when the survival time does not follow a Cox PH model, the proposed PC-IPCW-tau and IPCW-tau methods still outperform other methods, and the advantage is more apparent when the censoring rate is higher.

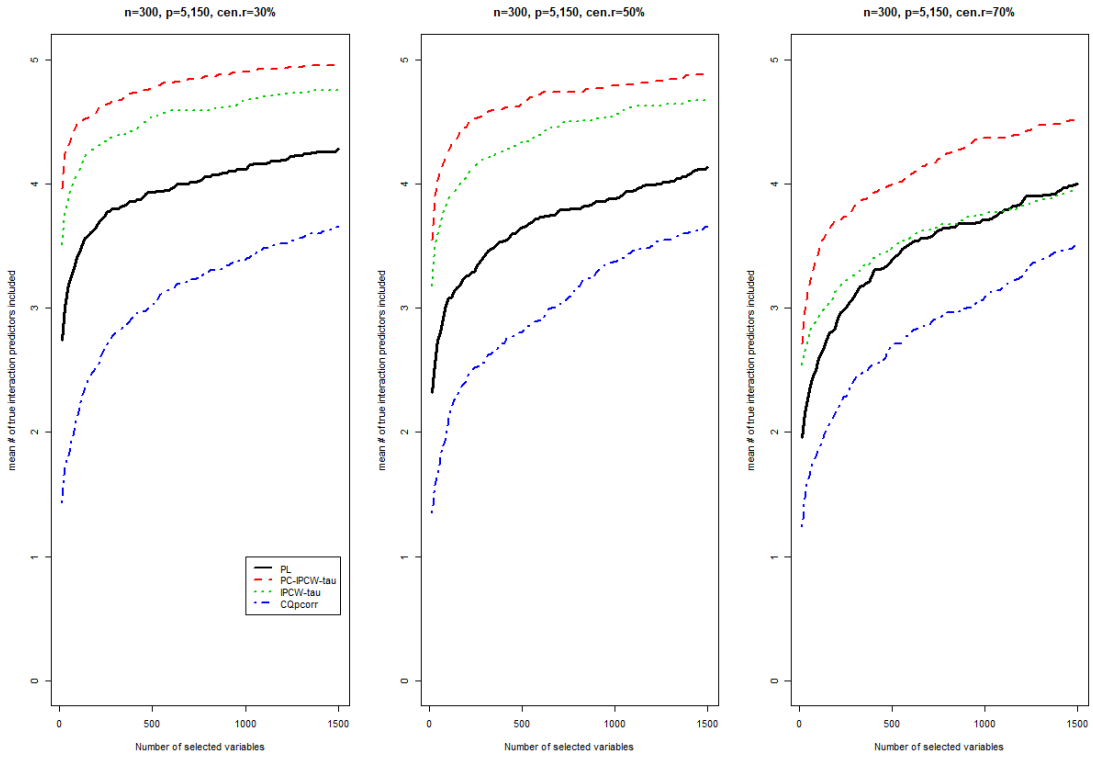


Fig. S10. Mean numbers of true second-order predictors included by the top (15, 30,…,1485,1500) selected variables under the scenario of Section 3.2 with cohort size equal to 300 and the survival time distribution in the transformation model given by standard normal.

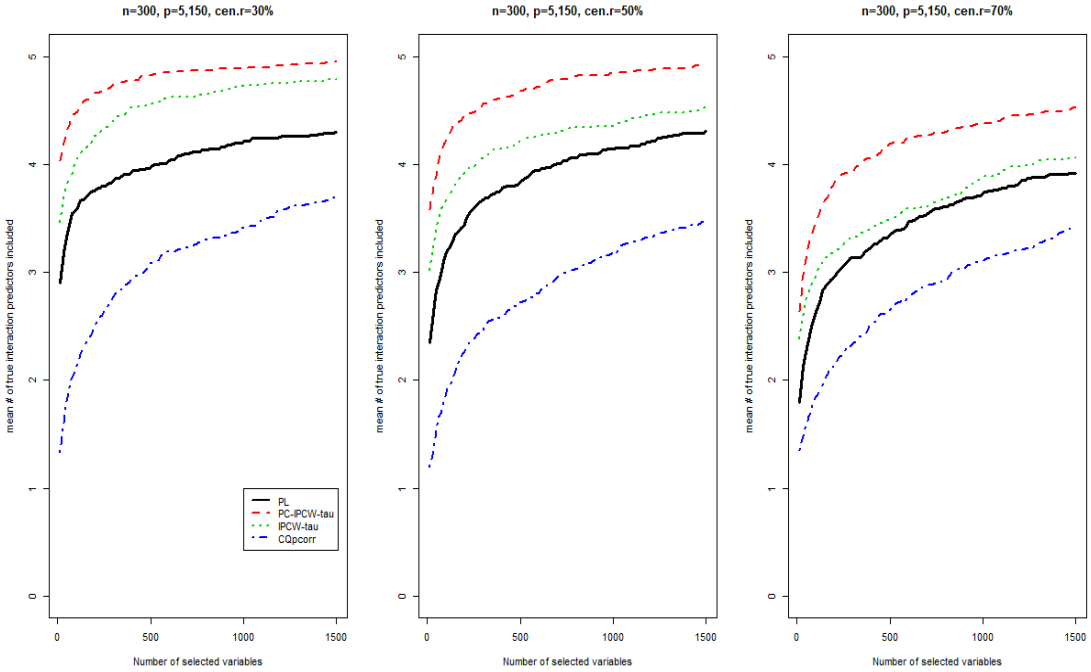


Fig. S11. Mean numbers of true second-order predictors included by the top (15, 30,…,1485,1500) selected variables under the scenario of Section 3.2 with cohort size equal to 300 and the survival time distribution in the transformation model given by standard logistic.

B.7 Comparing runtime among different methods

Table S4 shows the runtime (in seconds) over 200 simulation replications for different methods in the simulation in Section 3.1 of the main text.

Table S4. Runtime (in seconds) over 200 simulation replications for different methods in the simulation in Section 3.1 of the main text.

cen.r	PL	IPCW(S)	IPCW-tau	Kendall's tau
25%	20.92550	3.33890	6.81595	6.66965
50%	21.11420	3.37510	6.66615	6.57300
80%	20.89230	3.35325	6.86695	6.55040

cen.r = censoring rate

C. Additional Data Applications

C.1 Diffuse large B-cell lymphoma (DLBCL) data

The DLBCL data (Lenz et al., 2008) can be downloaded from the R package "bujar" (Wang, 2015), which consist of two sets of gene expression data, the CHOP and R-CHOP data sets; see Wang and Chen (2018) for detailed descriptions about the data. The CHOP and R-CHOP data sets contain censored survival outcomes from 181 and 233 patients, respectively, with gene expression data from the same 3833 genes after the filtering process. There are no significant differences in clinical survival outcome between subjects in the two data sets, and the censoring rates in the CHOP and R-CHOP datasets are 42% and 74%, respectively. Following Wang and Chen (2018), in the current analysis we randomly separate the pool of R-CHOP and CHOP patients into 207:207 training/test data sets.

As in Section 3.3 of the main text, we apply 4 screening methods ("IPCW(S)", "IPCW-tau", "PL", "PC-IPCW-tau") to the DLBCL data. After a grid search from 20 to 210 with step size 5, the number of candidate covariates, including both first- and second-order covariates, that yields the best overall performance for all methods is 190, so the top 190 predictors ranked by each method are selected as the candidate covariates, 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. In this way, we finally identify the main and second-order predictors by the PL, IPCW(S), IPCW-tau, and PC-IPCW-tau approaches, respectively, together with the MCP penalized regression. In addition, the Cox model with the whole 501,500 main and second-order predictors (formed by the top 1,000 genes selected by the univariate log-rank test), and the MCP penalty is applied directly to the training data to build the prediction model. Please see Table S10 for the lists of selected main and second-order predictors for different methods).

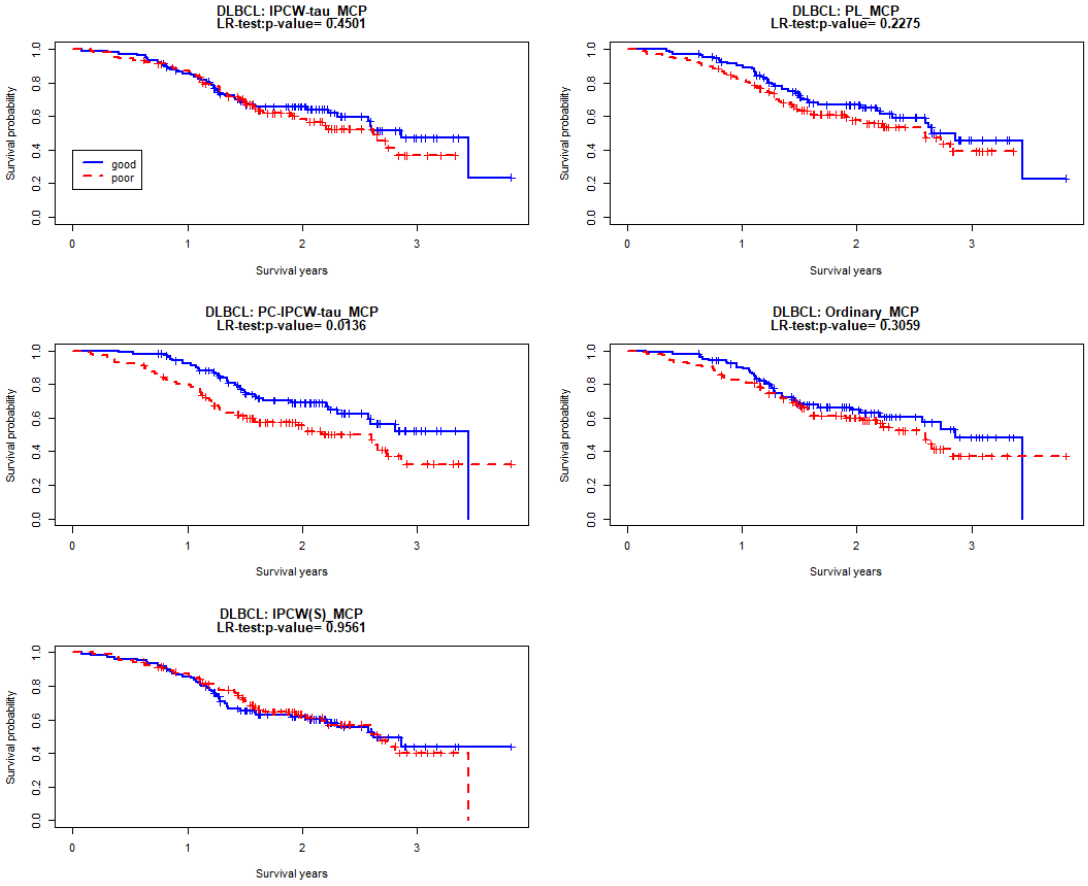
The prediction accuracy performances for different methods are evaluated in the same way as in Section 3.3 of the main text. The results are provided in Table S5. We can see that the proposed PC-IPCW-tau method outperforms other methods in the DLBCL test sample data. Fig. S12 displays the Kaplan-Meier survival curves for the two prognosis groups, "poor" (red) and "good" (blue) prognosis groups classified according to whether the PI value, the linear combination of the selected covariates with the coefficients given by the MCP penalized Cox model, exceeds the median PI value, in the test sample

of the DLBCL data.

Table S5. Results of prediction accuracies of different methods in DLBCL data (using the training and test sets as in Wang and Chen (2018)); all methods are applied together with the MCP penalized Cox regression.

	PL	IPCW(S)	IPCW-tau	PC-IPCW-tau	Ordinary
Cox-test	0.5149	0.7993	0.5860	0.0022	0.0760
LR-test	0.2275	0.9561	0.4501	0.0136	0.3059
C-index	0.5519	0.4899	0.5108	0.6165	0.5615

Fig. S12. Kaplan-Meier survival curves for the two prognosis groups ("good" (blue), "poor"(red) groups according to the median of the PI values) in the test sample of the DLBCL data.



C.2 The Cancer Genome Atlas lung adenocarcinoma (TCGA LUAD) data

The TCGA LUAD RNA-Seq expression data, together with the phenotype data containing the survival time and censoring status data can be downloaded from the R package "TCGAbiolinks" (Colaprico et al., 2016) or "UCSCXenaTools" (Wang et al., 2019). After excluding patients with missing survival time data, our analysis is focused on the subset of the TCGA LUAD data with 502 patients and 20531 gene expression variables. The censoring rate in the data is 64%. We randomly divide this subset into 251:251 training/test datasets.

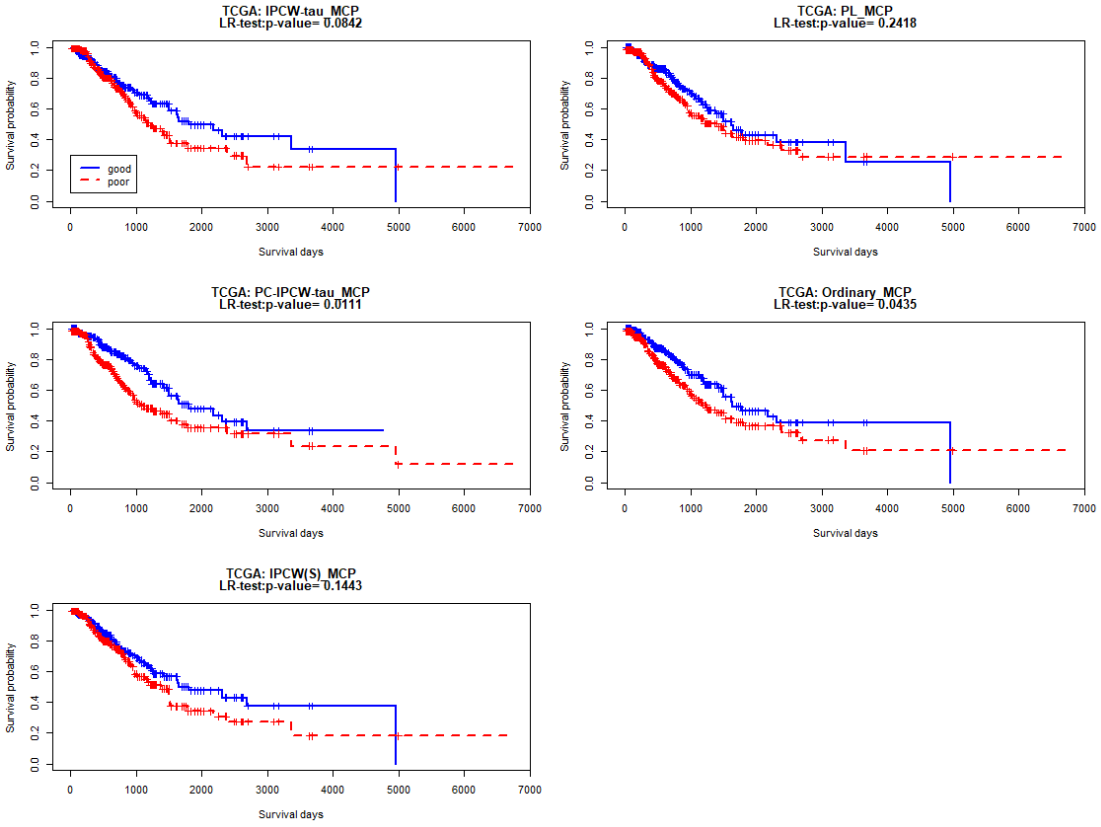
As in Section 3.3 of the main text, we apply 4 screening methods ("IPCW(S)", "IPCW-tau", "PL", "PC-IPCW-tau") to the TCGA LUAD data. After a grid search from 20 to 150 with step size 5, the number of candidate covariates, including both first- and second-order covariates, that yields the best overall performance for all methods is 130, so the top 130 predictors ranked by each method are selected as the candidate covariates, 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. In this way, we finally identify the main and second-order predictors by the PL, IPCW(S), IPCW-tau, and PC-IPCW-tau approaches, respectively, together with the MCP penalized regression. In addition, the Cox model with the whole 501,500 main and second-order predictors (formed by the top 1,000 genes selected by the univariate log-rank test) and the MCP penalty is applied directly to the training data to build the prediction model. Please see Table S11 for the lists of selected main and second-order predictors for different methods.

The prediction accuracy performances for different methods are evaluated in the same way as in Section 3.3 of the main text. The results are provided in Table S6. We can see that the proposed PC-IPCW-tau method performs best for survival prediction in the TCGA LUAD test data, followed by the IPCW-tau method, among all the methods considered. Fig. S13 displays the Kaplan-Meier survival curves for the two prognosis groups, "poor" (red) and "good" (blue) prognosis groups classified according to whether the prognostic index (PI) value, the linear combination of the selected covariates with the coefficients given by the MCP penalized Cox model, exceeds the median PI value, in the test sample of the TCGA LUAD data.

Table S6. Results of prediction accuracies of different methods in TCGA LUAD data (using the randomly selected 251:251 training /test sets); all methods are applied together with the MCP penalized Cox regression.

	PL	IPCW(S)	IPCW-tau	PC-IPCW-tau	Ordinary
Cox-test	0.1339	0.1687	0.0043	0.0047	0.4304
LR-test	0.2418	0.1443	0.0842	0.0111	0.0435
C-index	0.5568	0.5634	0.5764	0.6102	0.5706

Fig. S13. Kaplan-Meier survival curves for the two prognosis groups ("good" (blue), "poor"(red) groups according to the median of the PI values) in the test sample of the TCGA LUAD data.



C.3 Further results on the non-small cell lung cancer (NSCLC) data: the prediction performance results switching the original training and test data sets, and the prediction performance results using the SCAD penalty

We have performed an additional NSCLC analysis which switches the training and test data sets in the original analysis. In this additional analysis, the number of candidate covariates, including first and second-order covariates, is still chosen as 140, since it gives the best overall performances for different methods after the grid search from 20 to 160 with step size 5. Table S7 shows the prediction accuracy performances of different methods from this additional analysis, which still reveal that the proposed PC-IPCW-tau and IPCW-tau methods leads to higher prediction accuracies than the other method.

Table S7. Results of prediction accuracies of different methods in NSCLC data (switching the training and test sets in Chen et al. (2007)); all methods are applied together with the MCP penalized Cox regression.

	PL	IPCW(S)	IPCW-tau	PC-IPCW-tau	Ordinary
Cox-test	0.8908	0.7029	0.3286	0.0015	0.4369
LR-test	0.2026	0.0899	0.1154	0.0710	0.5925
C-index	0.5549	0.5970	0.5993	0.6717	0.5678

Using the same variable selection procedures as in Section 3.3 of the main text, the prediction models for the NSCLC data are now established by the Cox's regression model with the candidate covariates and the SCAD penalty (Fang and Li, 2011). The prediction performance results of the prediction models so constructed by different variable screening methods, including IPCW(S), IPCW-tau, PL, PC-IPCW-tau, and Ordinary SCAD (the ordinary Cox models with the whole 112,574 first- and second-order covariates from the 473 main covariate and the SCAD penalty), are shown in Table S8. We can see that, overall, the proposed PC-IPCW-tau has the best prediction accuracy performances.

Table S8. Results of prediction accuracies of different methods in NSCLC data (using the training and test sets as in Chen et al. (2007)); all methods are applied together with the SCAD penalized Cox regression.

	PL	IPCW(S)	IPCW-tau	PC-IPCW-tau	Ordinary
Cox-test	0.3333	0.2342	0.5202	0.3451	0.2403
LR-test	0.7626	0.2677	0.5410	0.0027	0.0480
C-index	0.5515	0.5540	0.5527	0.6029	0.6023

C.4 Complete lists of the predictors selected by different methods in the NSCLC, DCBCL, and TCGA LUAD data sets

Table S9. Lists of the main and second-order predictors selected by different methods in the NSCLC data.

PL (11 interaction)	IPCW(S) (11 interaction)	IPCW-tau (11 interaction)	PC-IPCW-tau (3 interaction)	Ordinary (1 main, 12 interaction)
"PIK3CA" "ERCC3"	"FGR" "TUSC3"	"CENTB2" "MRPL1"	"PDCD2" "EMP1"	HLF.1
"ABCC2" "TUSC3"	"TCF8" "MMP13"	"PPT2" "ADK"	"IRF4" "WDTC1"	"PRKCA" "COX11"
"CSF3R" "DOPEY1"	"CNOT4" "NR2F6"	"SP2" "DMPK"	"STAT2" "JMJD1A	"TCF8" "CGRRF1"
"IRF1" "COX11"	"ANXA1" "ERCC3"	"NR4A1" "ZNF250"		"PLAU" "LILRA2"
"LGMN" "WDTC1"	"BAZ2B" "TGFBR3"	"DHPS" "MTIF2"		"ARMET" "CTTN"
"SP2" "DMPK"	"MEN1" "DOPEY1"	"IRF4" "WDTC1"		"BRCA1" "SCP2"
"KIF23" "NR2F6"	"IRF1" "RNF4"	"CREB3L1" "CMAS"		"CSF3R" "PSD3"
"FLAD1" "TM7SF2"	"TM4SF18" "MMP11"	"FLAD1" "TM7SF2"		"TNFRSF10B" "PRRX1"
"MYH11" "ABL1"	"FLAD1" "TM7SF2"	"ME3" "BCR"		"TNNI2" "ELAC1"
"ME3" "BCR"	"ME3" "BCR"	"PAX2" "F8"		"KLHL22" "RPL5"
"FCGR2B" "RIPK1"	"CDK4" "SCP2"	"FCGR2B" "RIPK1"		"LOC285086" "F8"
				"IRF4" "WDTC1"
				"FLAD1" "TM7SF2"

Table S10. Lists of the main and second-order predictors selected by different methods in the DLBCL data

PL (3 main, 1 quadratic, 45 interaction)	IPCW(S) (62 interaction)	IPCW-tau (4 main, 50 interaction)	PC-IPCW-tau (2 main, 1 quadratic, 51 interaction)	Ordinary (1 main, 16 interaction)
EEA1	"SCIN" "WT1"	ADAMTS5	LOC101927069	PDPN
PPP4R4	"LINC00161" "SDC1"	EDNRA	PDPN	"TMEM67" "FCRL4"
C3orf80	"PTPRC" "LOXL2"	POSTN	"PDPN" "PDPN"	"M1AP" "PTPN22"
"GLIS3" "GLIS3"	"SPEF2" "LOXL2"	C3orf80	"GABRG1" "SSMEM1"	"CNOT6L" "TET2"
"NLRP11 "BHLHE41"	"CADM2" "GLRB"	"PTPRC" "LOXL2"	"CCDC7" "GAD1"	"ABCA13.1" "NME8"
"LINC02551" "UPB1"	"HDAC9" "RNF128"	"LINC02551" "UPB1"	"C17orf77" "ZNF876P"	"TTBK2.1" "PLEKHS1.1"
"CNOT6L" "COBLL1"	"C17orf77" "ZNF876P"	"EFCAB13" "CFTR"	"DUSP5P1" "CACTIN"	"LINC02363" "STYK1"
"C17orf77" "ZNF876P"	"DUSP5P1" "CACTIN"	"C18orf54" "CNTN3"	"LINC00314" "CXorf57"	"ZNF678" "OFCC1.1"
"GPR82" "SNX20"	"LINC00314" "CXorf57"	"TTBK2" "DPY19L1P1"	"KIAA0825" "CTNNA3"	"FBXO28" "KLHDC1"
"SSMEM1" "WDR41"	"WDR78" "LOC285889"	"TTBK2.1" "PLEKHS1.1"	"TTBK2.1" "PLEKHS1.1"	"GEN1" "LINC02099.1"
"ARHGAP42" "LINC01725"	"TTBK2.1" "PLEKHS1.1"	"RPS6KA5" "GABPB1.IT1"	"RPS6KA5" "GABPB1.IT1"	"GBP6" "UBASH3A"
"LINC00314" "CXorf57"	"NT5C1B" "PCK1"	"RPS6KA5" "GAREM1"	"NT5C1B" "PCK1"	"FBXO9" "TMEM259"
"C18orf54" "FBXO9"	"NT5C1B" "ROBO2"	"LINC02363" "NFIB"	"LINC02363" "NFIB"	"RGS13" "SSX4"
"TTBK2.1" "PLEKHS1.1"	"KIAA1217" "LOC100507387"	"FGF7" "LOC105375172"	"PRKCA.AS1" "ANGPTL1"	"LOC105375172" "WT1"
"CYP19A1" "ZNF407"	"MYT1L" "IGK"	"ZNF396" "LIMA1"	"TTC6" "ONECUT2"	"INHBA" "ATP11C"
"RPS6KA5" "GAREM1"	"FGF7" "LOC105375172"	"GABRG3" "LOC283922"	"LRRC77P" "LRRC77P.1"	"NPY1R" "PDCL2"
"NT5C1B" "ROBO2"	"CA6" "ADGRF1"	"PRKCA.AS1" "SLC4A7"	"FAM201A" "STYK1"	"ANKRD7" "LINC01808"
"NT5C1B" "WDR41"	"UNC13C" "DAZ1"	"FBXO28" "LGR5"	"C7orf57" "LINC01220"	
"SNX29P1" "CHIT1"	"TTC6" "ONECUT2"	"TTC6" "ONECUT2"	"TMED5" "LOC101928909"	
"ZNF85" "LINC00323"	"FAM201A" "STYK1"	"FAM216B" "MIR124.2HG"	"MCCC2" "F2RL1"	
"LINC02363" "PNMA8A"	"SPATA22" "LOC101927069"	"FAM201A" "STYK1"	"SYDE2" "NRG4"	
"EPB41L4A" "PODXL2"	"GRTP1.AS1" "RBM46"	"ADAMTS5" "LOXL2"	"LOC645485" "HECTD2"	
"LOC101927870.1" "ZNF226"	"SYDE2" "NRG4"	"LOC101928307.1 "B3GALT2.1"	"GRK4" "SLC4A4"	

"TMED5" "PDGFRL"	"DPH6" "ANKRD36BP2"	"LOC101929622" "RAB5A"	"CGAS" "F2RL1"	
"C9orf3" "LRCH3"	"CDKAL1" "LINC02099.1"	"GRTP1.AS1" "RBM46"	"C8orf34" "MEGF6"	
"CGAS" "ETV1"	"CDKAL1" "PCSK5"	"SYDE2" "NRG4"	"PHF21B"] "CTXN3"	
"PHF21B" "CTXN3"	"CDKAL1" "POU2F1"	"DPH6" "ANKRD36BP2"	"LOC285889" "DDX43"	
"LINC00514" "HERC1.1"	"S100B" "SAMD5"	"GRK4" "SLC4A4"	"TOP1P2" "ZBTB20"	
"KLHDC1" "RNF144A"	"SLC5A12" "MAF.1"	"CDKAL1" "POU2F1"	"EAF2" "RAVER2"	
"LMOD3" "ADRB1"	"LOC101927359" "SCARA5"	"TRIM64EP" "TEX14"	"RGS13.1" "PCNP"	
"C15orf62" "G6PC"	"LOC101927790" "HEMGN"	"PHF21B" "CTXN3"	"LOC728613" "CCDC169.2"	
"LINC02099.1" "CEP290"	"GLS2" "CCDC83"	"LOC101927790" "HEMGN"	"MYLK3" "ARFGEF3"	
"LINC02099.1" "PKHD1L1"	"RAB3IP" "DAZ1"	"KLHDC1" "RNF144A"	"SAMD5" "RASAL2"	
"SAMD5" "CLDN3"	"MAF" "RETREG1"	"METAP1D" "IGK"	"F11.1" "ATRNL1"	
"F11.1" "ATRNL1"	"OR5H1" "DNM1L"	"F11" "TMEM51.AS1"	"PDPN" "INHBA.1"	
"UGT2B4" "ARHGAP27"	"LOC728613" "CCDC169.2"	"NPAS3" "LINC01255"	"INHBA" "HSF5"	
"KLRC2" "CACTIN"	"SAMD5" "DAZ1"	"FRRS1" "ZNF674"	"PLN" "POU2F1"	
"PTGIS" "GABPB1.IT1"	"F11" "GNG8"	"GAD1" "TH"	"MYBPC2" "IGHA1"	
"CES1" "MAGEA9"	"LINC00630" "RAVER2"	"MOG" "SFTA3"	"UGT2B4" "PARVB"	
"CES1" "TC2N"	"INHBA" "HSF5"	"MYBPC2" "IGHA1"	"KLRC2" "BTF3"	
"INHBA.1" "ALKAL2"	"PLN" "GLRB"	"KLRC2" "BTF3"	"TH" "SPAG16.1"	
"COBLL1" "LIFR"	"PLN.1" "ADGRL3"	"CES1" "TC2N"	"CES1" "TC2N"	
"RFPL1S" "FAF1"	"COMP" "CLCA3P"	"MAGEA9" "HMGN2P46"	"INHBA.1" "ABCC4"	
"UBE2D1" "F2RL2"	"FABP1" "MATN3"	"INHBA.1" "ALKAL2"	"PHLPP1" "ABCG5"	
"UBE2D1" "ZBTB20.1"	"NOL4" "SFRP2"	"MAG" "MEGF6"	"FRRS1L" "NRG4"	
"HBA1" "CALM1"	"PLA2R1" "PAK6"	"MYO7B" "TMEM234"	"SPON1" "SOST"	
"SHROOM3" "CALM1"	"CES1" "TC2N"	"IRF4" "GKN1"	"UBE2D1" "F2RL2"	
"SCARA5" "CTTNBP2"	"INHBA.1" "ABCC4"	"MT1M" "SFTA3"	"MAG" "MEGF6"	
"ZBTB20.1" "CCDC169.2"	"RETREG1" "IL1RN"	"DPPA4" "TRDN"	"ZNF407" "HEMGN"	
	"HIST1H1B" "MEGF6"	"FEZF2" "CADM1"	"SCARA5" "WDR41"	

	"IRF4" "GKN1"	"GABPB1.IT1" "HBS1L"	"TUBE1" "EPPK1"	
	"MUC7" "SOST"	"GLIS3" "CSRNP3.1"	"KCNJ6" "GHSR"	
	"DPPA4" "FAF1"	"BEND6" "MKX"	"ANGPTL1.1" "CACTIN"	
	"CLCA3P" "MEGF6"	"KCNJ6" "GHSR"	"HMGN2P46" "TC2N"	
	"IL22" "CRNDE.1"			
	"SOST" "PDCL2"			
	"ZNF883" "ARFGEF3"			
	"CCDC110" "GPR158"			
	"BEND6" "MKX"			
	"RUNX2" "NOL4.1"			
	"FGF14.IT1" "RGS8"			
	"HMGN2P46" "TC2N"			

Table S11. Lists of the main and second-order predictors selected by different methods in the TCGA LUAD data

IPCW-tau (1 main, 9 interaction)	PL (1 main, 9 interaction)	PC-IPCW-tau (1 main, 9 interaction)	IPCW(S) (1 main, 10 interaction)	Ordinary (3 main, 32 interaction)
C1QTNF6	C6orf218	C1QTNF6	SLC22A8	C6orf218
"CD83" "NNT"	"ATP8B3" "PGPEP1"	"CD83" "NNT"	SNORA71A	CEACAM22P
"CUL4B" "PRRG1"	"CACNA1D" "ERLIN1"	"CDCP1" "GNMT"	"C20orf141" "C20orf141"	FKBP5
"DARS2" "TRIM7"	"CHST5" "FLNC"	"DARS2" "DNAJB4"	"SPANXE" "SPANXE"	"ABCA3" "MYO6"
"EFNB2" "FAM47E"	"EML4" "PDE9A"	"EFNB2" "FAM47E"	"SYT10" "SYT10"	"ABCC6P2" "STAM"
"FLNC" "TMEM178"	"JMJD7.PLA2G4B " "PCP4"	"FLNC" "TMEM178"	"ACCSL" "LOC650293"	"AKD1" "CPS1"
"GPC4" "SNORA71A"	"KYNU" "UCHL5"	"IFNA8" "PARM1"	"ACCSL" "OR1L4"	"BCAN" "MYOZ1"
"GUSBP1" "UNC13C"	"LRRC36" "ZNF502"	"LOC554202" "SSBP3"	"C8orf71" "SNORA1"	"BCL2L10" "LAMP3"
"MFSD2A" "SSBP3"	"MFSD2A" "ZNF737"	"MFSD2A" "SNORA1"	"DEFB103B" "SNORA71A"	"BIRC3" "KRT14"
"MTMR12" "TRIP10"	"NT5E" "ZNF552"	"SRCIN1" "SSBP3"	"DEFB103B" "TTTY10"	"BNC1" "SYNGR4"
			"GYPB" "OR52B4"	"BTG2" "EIF6"
				"C1orf114" "MYH16"
				"C1orf88" "CHRNA6"
				"C6orf218" "ZNF77"
				"CEACAM22P" "FLNC"
				"DKK1" "DPY19L1"
				"DSG2" "LOC202781"
				"E2F7" "SLC6A13"
				"EIF4E3" "TMEM168"
				"EML4" "PDE9A"
				"FUT1" "YWHAG"
				"GUSBL2" "GTF2E2"
				"HCN2" "LOC728989"
				"IL20RB" "TK1"
				"IRX3" "TLE1"

				"KIAA0562" "SLC16A3"
				"LZTFL1" "ZNF708"
				"MYH13" "OLFM1"
				"MYH16" "PATE4"
				"PNRC2" "SYT10"
				"SLC22A8" "SYT10"
				"TMEM168" "TOE1"
				"TP53I3" "ZNF345"
				"TRIP10" "ZNF185"
				"VANGL1" "VAX1"

References:

Colaprico, A. et al. (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016, 44, e71.

Fan, J., Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties, J. Am. Stat. Assoc., 96, 1348-1360.

Lenz G. et al. (2008) Stromal gene signatures in large-B-cell lymphomas, N Engl J Med., 359(22):2313–2323.

Wang, J. H., Chen, Y. H. (2018) Overlapping group screening for detection of gene-gene interactions: application to gene expression profiles with survival trait, BMC Bioinformatics, 19:335.

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

Wang, Z. (2015) bujar: Buckley-James regression for survival data with highdimensional covariates. R packages version 0.2–1.

Zhang, C. H. (2010) Nearly unbiased variable selection under minimax concave penalty, Ann. Stat., 38, 894:942.