nature machine intelligence

Supplementary information

https://doi.org/10.1038/s42256-024-00932-5

Stable Cox regression for survival analysis under distribution shifts

In the format provided by the authors and unedited

Appendix A Implementation detail

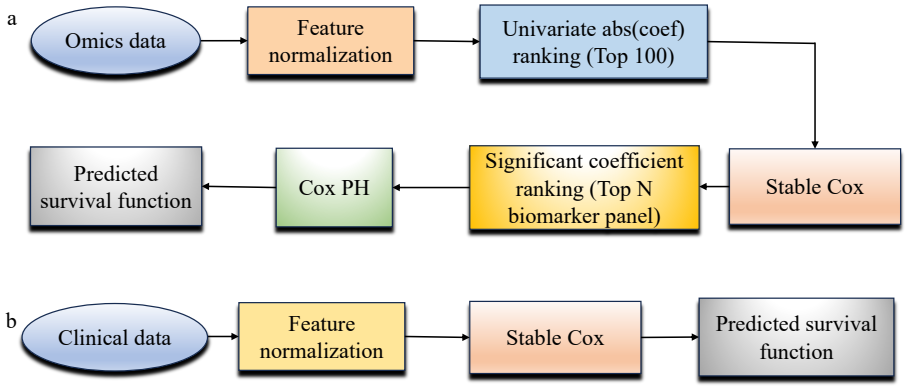

Fig. A1: **Pipelines for experiments. a,** The pipeline of the biomarker panel discovery for omics data. **b,** The pipeline of clinical data.

A.1 Data preprocessing

Omics data. To account for varying methodologies in quantifying gene expression in tumor tissues across different cohorts, we first apply a $\log_2$ transformation to omics data. This is followed by z-score feature normalization, standardizing the expression values of proteins or mRNA across different cohorts to a uniform scale [1–3]. Additionally, we utilize a univariate Cox PH model to calculate the coefficient value of each gene one by one, where the input covariate is the subgroup assignment by a median cutoff of gene expression. The absolute value of the coefficient will reflect the hazard value of each gene on survival outcome. We select the top 100 genes with the highest absolute value of the coefficients as the input of the following model. Subsequently, the Stable Cox model and corresponding baseline methods are employed to identify the top N (5, 10, 15, 20) significant proteins/genes as the biomarker panel. To validate that the screened biomarker panel is sufficient to predict the survival outcome, we train a new Cox PH model or corresponding baseline models to predict the survival function of samples in independent test cohorts. The pipeline of biomarker panel discovery for omics data is shown in Fig. A1a. The number of samples for each subgroup in Fig.5b are summarized in Table A2.

Clinical data. Clinical data typically comprises both categorical variables, such as cancer stage, and continuous variables, like blood pressure. For categorical variables with more than two categories, we utilize one-hot encoding to transform them. Continuous variables are standardized using unit-scale normalization to ensure consistency in data measurement. Given that clinical data usually involves a manageable number of features, often in the dozens, and that these values are readily accessible, we incorporate all features as inputs in our model to enhance prediction accuracy. The pipeline

of feature selection results for clinical data is shown in Fig. A1b. The information for all the real-world datasets used in the paper is summarized in Table A1.

A.2 Data generation model of simulated data

Table A3 lists the survival time function and survival time and hazard function. The generation models are from [16, 17], which are common underlying survival time generation processes that may occur in the real world.

A.3 Survival outcomes

In this study, we used the survival outcomes which are pivotal in prognostic research as the prediction targets. These outcomes are common time-to-event data, including time to death or time to disease occurrence. Particularly,

- **Overall Survival (OS)**: Overall survival refers to the length of time from either the date of diagnosis or the start of treatment for a disease, such as cancer, that patients diagnosed with the disease are still alive. In a clinical context, OS is considered a gold standard endpoint and is a direct measure of the therapeutic effect of a treatment.
- **Disease-Free Survival (DFS)**: DFS typically refers to the length of time after primary treatment for a cancer that the patient survives without any signs or symptoms of that cancer. This measure includes not just the absence of relapse, but also the absence of any new tumors at the original site or elsewhere (i.e., no secondary cancers).
- **Relapse-Free Survival (RFS)**: RFS is generally considered the time during which a patient remains free from the return of the disease symptoms after achieving remission. In other words, it is the period that the original cancer has not come back (relapsed) after the primary treatment.

A.4 Hyperparameter

For our method, the sample reweighting module requires an MLP classifier to discriminate between the training distribution and the weighted distribution. We adopt an MLP with two hidden layers and adjust the width of the hidden layer based on the performance of the validation set. The first hidden layer is selected from (10, 100) and the second hidden layer is selected from (1, 20) according to validation performance. To further restrict the range of learned sample weights and increase stability, we clip the weights to (γ_1, γ_2), where $\gamma_1 \in (1e-3, 0.3)$ and $\gamma_1 \in (2, 5)$. For the sparse penalty λ, we select it from $(1e-5, 0.1)$ based on the performance of Cox PH on the validation set and keep the same for all other methods to keep fair comparison. Note that for omics datasets, we select the best performance according to the validation performance of the Top 10, and apply the same hyperparameter for the Top 5, 15 and 20 scenarios. For all the baseline methods, we utilize the implementation in Python package lifelines (https://lifelines.readthedocs.io/en/latest/). Our weighted Cox regression module is also based on the implementation of Cox PH in lifelines. And it provides the p-value of each coefficient so that it can be used to select the most significant features. The hyperparameters for Stable Cox are reported in Table A4.

Table A1: Summary of datasets used in this paper. The #covariates is the number of covariate inputting methods after preprocessing. The number in $(\cdot)$ represents the number of samples in the corresponding cohort.

Datasets	#covariates	training set	validation set	testing set
HCC transcriptome	100	[4] (351) Data link	[5] (115) Data link	[6] (203) Data link [7] (209) Data link [8] (80) Data link
Breast transcriptome	100	[9] Data link Cohort id = 3 (763)	Cohort id= 5 (170)	Cohort id = 1 (521) Cohort id = 2 (288) Cohort id = 3 (238)
Melanoma transcriptome	100	[10] (120) Data link	[11] (26) Data link	[12] (91) Data link [13] (54) Data link [14] (41) Data link
Lung cancer clinical	51	[15] Data link random 40% (240)	female (4) male (152)	age<=60 (43), age>60 (113) loc (central) (47), loc (pph) (107) OA (present) (80), OA (absent) (76)
Breast cancer clinical	29	[9] Data link Cohort id = 3 (394)	Cohort id = 5 (14)	Cohort id = 1 (273) Cohort id = 2 (177)

Table A2: The number of samples for each subgroup in Fig.5b. 'All' means the number of samples in the corresponding clinical subgroup. 'High' and 'Low' mean the number of samples of high-risk and low-risk subgroups separated by optimal cutoff value, respectively.

Method	Subgroup	Cohort1(All=High+Low)	Cohort2	Cohort3
Cox PH	Age			
	≤60	243 = 104 + 139	179 = 61 + 118	100 = 47 + 53
	>60	278 = 184 + 94	109 = 35 + 74	138 = 78 + 60
	ER Status			
	Negative	123 = 83 + 40	71 = 23 + 48	63 = 45 + 18
	Positive	398 = 259 + 139	217 = 149 + 68	175 = 99 + 76
	HER2 Status			
	Negative	464 = 227 + 237	247 = 162 + 85	208 = 102 + 106
	Positive	57 = 19 + 38	41 = 21 + 20	30 = 10 + 20
	PR Status			
	Negative	257 = 138 + 119	133 = 41 + 92	117 = 77 + 40
	Positive	264 = 171 + 93	155 = 64 + 91	121 = 63 + 58
Stable Cox	Age			
	≤60	243 = 141 + 102	179 = 54 + 125	100 = 43 + 57
	>60	278 = 117+161	109 = 54 + 55	138 = 49 + 89
	ER Status			
	Negative	123 = 44 + 79	71 = 35 + 36	63 = 19 + 44
	Positive	398 = 139 + 259	217 = 73 + 144	175 = 53 + 122
	HER2 Status			
	Negative	464 = 215 + 249	247 = 81 + 166	208 = 65 + 143
	Positive	57 = 36 + 21	41 = 28 + 13	30 = 9 + 21
	PR Status			
	Negative	257 = 180 + 77	133 = 73 + 60	117 = 37 + 80
	Positive	264 = 87 + 177	155 = 48 + 107	121 = 65 + 56

Table A3: The data generation mechanism for simulated data. U is a variable following a uniform distribution on the interval from 0 to 1. ϕ is a zero-mean normal variable with 0.5 standard deviation. $g(S) = S_1 \cdot S_2 \cdot S_3$ is a non-linear model misspecification term. λ of Cox-Exp is set as 0.5. λ and v of Cox-Weibull is set as 0.5 and 0.25 respectively. For Poly, we only select the samples whose survival time is larger than 0 and set λ as 1.

	Characteristic	
Generation model	Survival Time	Hazard Function
Cox-Exp	$T = -\frac{\log(U)}{\lambda \exp(\beta^T(S)S + g(S))}$	$h(t\|S) = \lambda \exp(\beta^T(S)S + g(S))$
Cox-Weibull	$T = \left(-\frac{\log(U)}{\lambda \exp \beta^T(S)S + g(S)}\right)^{\frac{1}{v}}$	$h(t\|S) = \lambda \exp(\beta^T(S)S + g(S))vt^{v-1}$
Poly	$T = -\frac{\log(U)}{\lambda(\beta^T(S)S + g(S))}$	$h(t\|S) = \lambda(\beta^T(S)S + g(S))$
Log T	$\log T = \beta^T(S)S + g(S) + \phi$	Unavailable

A.5 Peseudocode of Stable Cox

The training process of our Stable Cox model is summarised in Algorithm 1. Line 1-7 are the procedure of the independence-driven sample reweighting module, which aims

Table A4: Hyperparameters for Stable Cox. Baselines set the same λ value for the l_2 penalty. 'N/A' means we do not clip the weights.

Datasets	Stable Cox				
	λ	γ_1	γ_2	hidden1	hidden2
Simulated data	1e-5	N/A	N/A	100	5
HCC transcriptome (OS)	5e-4	4e-1	4	98	11
Melanoma transcriptome (OS)	3e-4	3e-2	3	30	9
Breast transcriptome (OS)	3e-2	3e-2	3	30	2
Lung cancer clinical (OS)	2e-3	3e-1	3	37	4
Lung cancer clinical (DFS)	4e-4	1e-1	4	38	9
Breast cancer clinical (OS)	3e-2	1e-3	2	15	13
Breast cancer clinical (RFS)	2e-3	2e-2	2	69	15

to learn a set of sample weights w to decorrelate covariates $\mathbf{X}$. After obtaining w, line 8-9 are the procedure of the weighted Cox regression module, learning the stable coefficients β and baseline function $\lambda_0(t)$.

Algorithm 1 Training process of Stable Cox

Require: $\left\{x^{(i)}, t^{(i)}, \delta^{(i)}\right\}_{i=1}^{n}$

1: Initialize a new covariate matrix $\hat{\mathbf{X}} \in \mathbb{R}^{n \times p}$
2: **for** $j = 1 \cdots p$ **do**
3: Randomly permutate the j-th column of covariates $\mathbf{X}$;
4: Assign the permutated column of $\mathbf{X}_j$ to the j-th column of $\hat{\mathbf{X}}$
5: **end for**
6: Set samples of $\hat{\mathbf{X}}$ as positive samples and samples of $\mathbf{X}$ as negative samples, then train a binary MLP classifier.
7: Set $w(x) = \frac{P'(Z=1|x)}{P'(Z=0|x)}$ for each sample $x^{(i)}$ in $\mathbf{X}$, where $P'(Z=1|x)$ is the probability of sample x been drawn from decorrelated dataset estimated by the trained classifier.
8: Using the learned sample weights $w(x)$ to reweight the event in a Cox PH model as Eq. (11) in the main text.
9: Optimize the weighted partial log-likelihood loss to estimate the β and $\lambda_0(t)$

Return: Learned Stable Cox model

Appendix B Omitted proofs

B.1 Details in Remark 2

We can get that the derivative of $\tilde{L}_w(\beta)$ on β is

$$\frac{\partial \tilde{L}_w}{\partial \beta} = D(\beta) \triangleq \int_0^\tau q_w^{(1)}(u)\mathrm{d}u - \int_0^\tau \frac{q_w^{(1)}(\beta,u)}{q_w^{(0)}(\beta,u)} \cdot q_w^{(0)}(u)\mathrm{d}u \tag{B1}$$

and the Hessian matrix is given by

$$\text{Hessian}\left(\tilde{L}_w\right) = \int_0^\tau \frac{q_w^{(1)}(\beta,u)\left(q_w^{(1)}(\beta,u)\right)^T - q_w^{(2)}(\beta,u)q_w^{(0)}(\beta,u)}{\left(q_w^{(0)}(\beta,u)\right)^2} \cdot q_w^{(0)}(u)\mathrm{d}u. \tag{B2}$$

Note that the term $\left(q_w^{(1)}(\beta,u)\left(q_w^{(1)}(\beta,u)\right)^T - q_w^{(2)}(\beta,u)q_w^{(0)}(\beta,u)\right) / \left(q_w^{(0)}(\beta,u)\right)^2$ can be considered as the covariance matrix of X under a weighted distribution of P^{tr} with weighting functions $w'(X) = w(X)Y(u)\exp(\beta^T X)$. As a result, this term is positive semi-definite. Therefore, Hessian$(\tilde{L}_w)$ is also positive semi-definite.

B.2 Details in Remark 3

Proposition B.1. *$T^{failure} \perp V \mid S$ if and only if there exists a function $\lambda'(u;S)$ such that for any u and X, $\lambda'(u;S) = \lambda(u;X)$.*

Proof. On the one hand, when $T^{\text{failure}} \perp V \mid S$, we have

$$\begin{aligned}
\lambda(u;X) &= \lim_{h\to 0^+} \frac{P^{tr}\left(T^{\text{failure}} \le u+h | T^{\text{failure}} \ge u, X\right)}{h} \\
&= \lim_{h\to 0^+} \frac{P^{tr}\left(u \le T^{\text{failure}} \le u+h, X\right)}{hP^{tr}\left(T^{\text{failure}} \ge u, X\right)} \\
&= \lim_{h\to 0^+} \frac{P^{tr}(S)P^{tr}\left(u \le T^{\text{failure}} \le u+h, V|S\right)}{hP^{tr}(S)P^{tr}\left(T^{\text{failure}} \ge u, V|S\right)} \\
&= \lim_{h\to 0^+} \frac{P^{tr}\left(u \le T^{\text{failure}} \le u+h|S\right)P^{tr}(V|S)}{hP^{tr}\left(T^{\text{failure}} \ge u|S\right)P^{tr}(V|S)} \\
&= \lim_{h\to 0^+} \frac{P^{tr}\left(T^{\text{failure}} \le u+h | T^{\text{failure}} \ge u, S\right)}{h}.
\end{aligned} \tag{B3}$$

This means that $\lambda(u;X)$ depends only on S and there exists a function $\lambda'(u;S)$ such that for all $X \in \mathcal{X}$, we have $\lambda'(u;S) = \lambda(u;X)$.

On the other hand, consider the scenario that there exists a function $\lambda'(u;S)$ such that for all $X \in \mathcal{X}$, we have $\lambda'(u;S) = \lambda(u;X)$. Then we have for any u and X,

$$\lim_{h\to 0^+} \frac{P^{tr}\left(T^{\text{failure}} \le u+h | T^{\text{failure}} \ge u, X\right)}{h} = \lim_{h\to 0^+} \frac{P^{tr}\left(T^{\text{failure}} \le u+h | T^{\text{failure}} \ge u, S\right)}{h}. \tag{B4}$$

As a result, $P^{tr}\left(T^{\text{failure}}|S\right)$ and $P^{tr}\left(T^{\text{failure}}|X\right)$ have the same probability density function. Therefore, $T^{\text{failure}} \perp V \mid S$. Now the claim follows. □

Proposition B.2. *Suppose in the training distribution P^{tr}, we have $(T^{failure}, T^{censored}) \perp V \mid S$. In addition, the test distribution differs from the training distribution in covariate shift only, i.e., $P^{tr}(T^{failure}, T^{censored}|X) = P^{te}(T^{failure}, T^{censored}|X)$. Then in the test distribution P^{te}, we also have $(T^{failure}, T^{censored}) \perp V \mid S$. In addition, $P^{tr}(T^{failure}, T^{censored}|S) = P^{te}(T^{failure}, T^{censored}|S)$*

Proof. Let $\mathcal{V}$ denote the support of V. By the conditions, we have

$$\begin{aligned}
&P^{te}\left(T^{\text{failure}}, T^{\text{censored}}|S\right) \\
=& \int_{\mathcal{V}} P^{te}(V|S) P^{te}\left(T^{\text{failure}}, T^{\text{censored}}|S, V\right) \mathrm{d}V \\
=& \int_{\mathcal{V}} P^{te}(V|S) P^{tr}\left(T^{\text{failure}}, T^{\text{censored}}|S, V\right) \mathrm{d}V \\
=& \int_{\mathcal{V}} P^{te}(V|S) P^{tr}\left(T^{\text{failure}}, T^{\text{censored}}|S\right) \mathrm{d}V \\
=& P^{tr}\left(T^{\text{failure}}, T^{\text{censored}}|S\right) \int_{\mathcal{V}} P^{te}(V|S) \mathrm{d}V \\
=& P^{tr}\left(T^{\text{failure}}, T^{\text{censored}}|S\right) \\
=& P^{tr}\left(T^{\text{failure}}, T^{\text{censored}}|X\right) = P^{te}\left(T^{\text{failure}}, T^{\text{censored}}|X\right).
\end{aligned} \tag{B5}$$

Now the claim follows. □

B.3 Proof of Theorem 1

Proof. For $r = 0, 1, 2$, define

$$\begin{aligned}
Q_w^{(r)}(u) &= \frac{1}{n}\sum_{i=1}^n w\left(x^{(i)}\right) Y^{(i)}(u) \lambda\left(u; x^{(i)}\right) \left(x^{(i)}\right)^{\otimes r}, \\
Q_w^{(r)}(\beta, u) &= \frac{1}{n}\sum_{i=1}^n w\left(x^{(i)}\right) Y^{(i)}(u) \exp\left(\beta^T x^{(i)}\right) \left(x^{(i)}\right)^{\otimes r}, \\
J_w^{(r)}(u) &= \frac{1}{n}\sum_{i=1}^n w^2\left(x^{(i)}\right) Y^{(i)}(u) \lambda\left(u; x^{(i)}\right) \left(x^{(i)}\right)^{\otimes r}.
\end{aligned} \tag{B6}$$

We further define the following random processes.

$$
\begin{aligned}
C(\beta, u) &= \sum_{i=1}^{n} w\left(x^{(i)}\right) \int_0^u \left(\beta^T x^{(i)} - \log \left(\sum_{j=1}^{n} w\left(x^{(j)}\right) Y^{(j)}(v) \exp\left(\beta^T x^{(j)}\right) \right) \right) \mathrm{d}N^{(i)}(v) \\
&= \sum_{i=1}^{n} w\left(x^{(i)}\right) \int_0^u \left(\beta^T x^{(i)} - \log Q_w^{(0)}(\beta, v) \right) \mathrm{d}N^{(i)}(v) \\
G(\beta, u) &= \frac{1}{n} \left(C(\beta, u) - C(\beta_w, u) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} w\left(x^{(i)}\right) \int_0^u \left((\beta - \beta_w)^T x^{(i)} - \log \left(\frac{Q_w^{(0)}(\beta, v)}{Q_w^{(0)}(\beta_w, v)} \right) \right) \mathrm{d}N^{(i)}(v).
\end{aligned}
\tag{B7}
$$

And then we define

$$
\begin{aligned}
& A(\beta, u) \\
&= \frac{1}{n} \sum_{i=1}^{n} w\left(x^{(i)}\right) \int_0^u \left((\beta - \beta_w)^T x^{(i)} - \log \left(\frac{Q_w^{(0)}(\beta, v)}{Q_w^{(0)}(\beta_w, v)} \right) \right) Y^{(i)}(v) \lambda\left(v; x^{(i)}\right) \mathrm{d}v \\
&= \int_0^u \left((\beta - \beta_w)^T Q_w^{(1)}(v) - \log \left(\frac{Q_w^{(0)}(\beta, v)}{Q_w^{(0)}(\beta_w, v)} \right) Q_w^{(0)}(v) \right) \mathrm{d}v.
\end{aligned}
\tag{B8}
$$

Then for each β, the process $G(\beta, \cdot) - A(\beta, \cdot)$ is a local square integrable martingale with

$$
\langle G(\beta, \cdot) - A(\beta, \cdot), G(\beta, \cdot) - A(\beta, \cdot) \rangle = B(\beta, \cdot),
\tag{B9}
$$

Here

$$
\begin{aligned}
& B(\beta, u) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \left(w\left(x^{(i)}\right) \right)^2 \int_0^u \left((\beta - \beta_w)^T x^{(i)} - \log \left(\frac{Q_w^{(0)}(\beta, v)}{Q_w^{(0)}(\beta_w, v)} \right) \right)^2 Y^{(i)}(v) \lambda\left(v; x^{(i)}\right) \mathrm{d}v \\
&= \frac{1}{n} \int_0^u \left((\beta - \beta_w)^T J_w^{(2)}(v) (\beta - \beta_w) - 2(\beta - \beta_w)^T J_w^{(1)}(v) \log \left(\frac{Q_w^{(0)}(\beta, v)}{Q_w^{(0)}(\beta_w, v)} \right) \right) \mathrm{d}v \\
&\quad + \frac{1}{n} \int_0^u \log^2 \left(\frac{Q_w^{(0)}(\beta, v)}{Q_w^{(0)}(\beta_w, v)} \right) J_w^{(0)}(v) \mathrm{d}v.
\end{aligned}
\tag{B10}
$$

Note that under Assumption 1, when $n \to \infty$,

$$
\begin{aligned}
J_w^{(r)}(v) &\xrightarrow{\mathrm{P}} \mathbb{E}\left[w^2(X) Y(v) \lambda(v; X) X^{\otimes r} \right] < \infty. \\
Q_w^{(r)}(\beta, v) &\xrightarrow{\mathrm{P}} \mathbb{E}\left[w(X) Y(v) \exp\left(\beta^T X\right) X^{\otimes r} \right] < \infty.
\end{aligned}
\tag{B11}
$$

As a result, $nB(\beta, u)$ converges in probability to some finite value based on the value of β. In addition, when $n \to \infty$,

$$A(\beta, \tau) \xrightarrow{\mathrm{P}} \int_0^\tau \left((\beta - \beta_w)^T q_w^{(1)}(u) - \log \left(\frac{q_w^{(0)}(\beta, u)}{q_w^{(0)}(\beta_w, u)} \right) q_w^{(0)}(u) \right) \mathrm{d}u. \tag{B12}$$

Therefore, according to Lemma B.3, we have that $G(\beta, \tau)$ converges in probability to the same limit as $A(\beta, \tau)$. Note that when $n \to \infty$,

$$\frac{\partial A(\beta, \tau)}{\partial \beta} \xrightarrow{\mathrm{P}} \int_0^\tau \left(q_w^{(1)}(u) - \frac{q_w^{(1)}(\beta, u)}{q_w^{(0)}(\beta, u)} \cdot q_w^{(0)}(u) \right) \mathrm{d}u = \frac{\partial \tilde{L}_w}{\partial \beta}. \tag{B13}$$

According to Assumption 2, β_w is the unique solution to the right-hand side of the above equation. As a result, when $n \to \infty$,

$$\hat{\beta}_w = \arg\min_\beta C(\beta, \tau) = \arg\min_\beta G(\beta, \tau) \xrightarrow{\mathrm{P}} \arg\min_\beta A(\beta, \tau) = \beta_w. \tag{B14}$$

Now the claim follows. □

B.4 Proof of Theorem 2

Proof. According to Theorem 1, we have that when $n \to \infty$, $\hat{\beta}_w \xrightarrow{\mathrm{P}} \beta_w$. As a result, to prove Theorem 2, it suffices to show that $\beta_w(V) = 0$ where $\beta_w(V)$ denotes the vectors of the coefficients β_w *w.r.t.* V.

We slightly abuse the notion here to use $q_w^{(1,S)}(\beta, u)$ and $q_w^{(1,V)}(\beta, u)$ to denote the vectors $q_w^{(1)}(\beta, u)$ *w.r.t.* S and V, respectively. And we use $q_w^{(1,S)}(u)$ and $q_w^{(1,V)}(u)$ to denote the vectors $q_w^{(1)}(u)$ *w.r.t.* S and V, respectively. Similarly, we use $D^{(S)}(\beta)$ and $D^{(V)}(\beta)$ to denote the vectors $D(\beta)$ (defined in Equation (B1)) *w.r.t.* S and V, respectively.

According to Assumption 3, since $Y(u) = \mathbb{I}[T > u] = \mathbb{I}[(T^{\text{failure}} > u) \wedge (T^{\text{censored}} > u)]$, we have $Y(u) \perp V \mid S$ in the training distribution. According to Proposition B.2, we also have $Y(u) \perp V \mid S$ in the test distribution. In addition, we have $S \perp V$ in the weighted distribution $\tilde{P}_w$ since $w \in \mathcal{W}_\perp$. As a result, we have

$$\begin{aligned}
q_w^{(1,S)}(\beta, u) &= \mathbb{E}\left[w(X) Y(u) \exp\left(\beta^T X\right) S \right] = \mathbb{E}_w\left[Y(u) \exp\left(\beta_w^T(S) S\right) S \exp\left(\beta_w^T(V) V\right) \right] \\
&= \mathbb{E}_w\left[\mathbb{E}_w\left[Y(u) \exp\left(\beta_w^T(S) S\right) S \exp\left(\beta_w^T(V) V\right) \mid S \right] \right] \\
&= \mathbb{E}_w\left[\mathbb{E}_w\left[Y(u) \exp\left(\beta_w^T(S) S\right) S \mid S \right] \mathbb{E}_w\left[\exp\left(\beta_w^T(V) V\right) \mid S \right] \right] \\
&= \mathbb{E}_w\left[Y(u) \exp\left(\beta_w^T(S) S\right) S \right] \mathbb{E}_w\left[\exp\left(\beta_w^T(V) V\right) \right].
\end{aligned} \tag{B15}$$

Similarly, we have

$$q_w^{(1,V)}(\beta,u) = \mathbb{E}\left[w(X)Y(u)\exp\left(\beta^T X\right)V\right] = \mathbb{E}_w\left[Y(u)\exp\left(\beta_w^T(S)S\right)\exp\left(\beta_w^T(V)V\right)V\right]$$
$$= \mathbb{E}_w\left[Y(u)\exp\left(\beta_w^T(S)S\right)\right]\mathbb{E}_w\left[\exp\left(\beta_w^T(V)V\right)V\right]. \tag{B16}$$

In addition, since $T^{\text{failure}} \perp V \mid S$, according to Proposition B.1, there exists a function $\lambda'(u;S)$ such that for all $X \in \mathcal{X}$, we have $\lambda'(u;S) = \lambda(u;X)$. Therefore,

$$q_w^{1,S}(u) = \mathbb{E}\left[w(X)Y(u)\lambda(u;X)S\right] = \mathbb{E}_w\left[Y(u)\lambda'(u;S)S\right]$$
$$q_w^{1,V}(u) = \mathbb{E}\left[w(X)Y(u)\lambda(u;X)V\right] = \mathbb{E}_w\left[Y(u)\lambda'(u;S)V\right] = \mathbb{E}_w\left[Y(u)\lambda'(u;S)\right]\mathbb{E}_w[V]. \tag{B17}$$

In addition,

$$q_w^{(0)}(u) = \mathbb{E}\left[w(X)Y(u)\lambda(u;X)\right] = \mathbb{E}_w\left[Y(u)\lambda'(u;S)\right]$$
$$q_w^{(0)}(\beta,u) = \mathbb{E}\left[w(X)Y(u)\exp\left(\beta^T X\right)\right] = \mathbb{E}_w\left[Y(u)\exp\left(\beta_w^T(S)S\right)\right]\mathbb{E}_w\left[\exp\left(\beta_w^T(V)V\right)\right] \tag{B18}$$

Now we can get that

$$\begin{aligned}
&D^{(S)}(\beta_w)\\
&= \int_0^\tau q_w^{(1,S)}(u)\mathrm{d}u - \int_0^\tau \frac{q_w^{(1,S)}(\beta,u)}{q_w^{(0)}(\beta,u)}\cdot q_w^{(0)}(u)\mathrm{d}u\\
&= \int_0^\tau \mathbb{E}_w\left[Y(u)\lambda'(u;S)S\right]\mathrm{d}u\\
&\quad - \int_0^\tau \frac{\mathbb{E}_w\left[Y(u)\exp\left(\beta_w^T(S)S\right)S\right]\mathbb{E}_w\left[\exp\left(\beta_w^T(V)V\right)\right]}{\mathbb{E}_w\left[Y(u)\exp\left(\beta_w^T(S)S\right)\right]\mathbb{E}_w\left[\exp\left(\beta_w^T(V)V\right)\right]}\cdot\mathbb{E}_w\left[Y(u)\lambda'(u;S)\right]\mathrm{d}u\\
&= \int_0^\tau \mathbb{E}_w\left[Y(u)\lambda'(u;S)S\right]\mathrm{d}u - \int_0^\tau \frac{\mathbb{E}_w\left[Y(u)\exp\left(\beta_w^T(S)S\right)S\right]}{\mathbb{E}_w\left[Y(u)\exp\left(\beta_w^T(S)S\right)\right]}\cdot\mathbb{E}_w\left[Y(u)\lambda'(u;S)\right]\mathrm{d}u = 0
\end{aligned} \tag{B19}$$

and

$$\begin{aligned}
&D^{(V)}(\beta_w)\\
&= \int_0^\tau q_w^{(1,V)}(u)\mathrm{d}u - \int_0^\tau \frac{q_w^{(1,V)}(\beta,u)}{q_w^{(0)}(\beta,u)}\cdot q_w^{(0)}(u)\mathrm{d}u\\
&= \int_0^\tau \mathbb{E}_w\left[Y(u)\lambda'(u;S)\right]\mathbb{E}_w[V]\mathrm{d}u\\
&\quad - \int_0^\tau \frac{\mathbb{E}_w\left[Y(u)\exp\left(\beta_w^T(S)S\right)\right]\mathbb{E}_w\left[\exp\left(\beta_w^T(V)V\right)V\right]}{\mathbb{E}_w\left[Y(u)\exp\left(\beta_w^T(S)S\right)\right]\mathbb{E}_w\left[\exp\left(\beta_w^T(V)V\right)\right]}\cdot\mathbb{E}_w\left[Y(u)\lambda'(u;S)\right]\mathrm{d}u\\
&= \int_0^\tau \mathbb{E}_w\left[Y(u)\lambda'(u;S)\right]\mathbb{E}_w[V]\mathrm{d}u - \int_0^\tau \frac{\mathbb{E}_w\left[\exp\left(\beta_w^T(V)V\right)V\right]}{\mathbb{E}_w\left[\exp\left(\beta_w^T(V)V\right)\right]}\cdot\mathbb{E}_w\left[Y(u)\lambda'(u;S)\right]\mathrm{d}u = 0.
\end{aligned} \tag{B20}$$

It is easy to verify that $D^{(V)}(\beta_w) = 0$ when $\beta_w(V) = 0$. In addition, $D^{(S)}$ only depends on $\beta_w(S)$ and D_V only depends on $\beta_w(V)$. According to Assumption 2, we must have $\beta_w(V) = 0$. Now the claim follows. □

B.5 Important Lemmas

Lemma B.3 (Lenglart's inequality [18, 19]). *For each $n = 1, 2, \ldots$, let $N^{(n)}$ be a multi-variate counting process with n components. Let $H^{(n)}$ be a $p \times n$ ($p \geq 1$ is fixed) matrix of locally bounded predictable processes. Suppose that $N^{(n)}$ has an intensity process $\lambda^{(n)}$, and define local square integrable martingale $W^{(n)} = \left(W_1^{(n)}, W_2^{(n)}, \ldots, W_p^{(n)}\right)$ by*

$$W_i^{(n)}(t) = \int_0^t \sum_{l=1}^{n} H_{il}^{(n)}(u) \left(\mathrm{d}N_l^{(n)}(u) - \lambda_l^{(n)}(u)\mathrm{d}u\right). \tag{B21}$$

Let A be a $p \times p$ matrix of continuous functions on $[0, 1]$ which form the covariance functions of a continuous p-variate Gaussian martingale $W^{(\infty)}$, with $W^{(\infty)}(0) = 0$; i.e., $\mathrm{Cov}\left(W_i^{(\infty)}, W_j^{(\infty)}\right) = A_{ij}(t \wedge u)$ for all i, j, t, and u. Suppose that for all i, j, and t,

$$\left\langle W_i^{(n)}, W_j^{(n)} \right\rangle (t) = \int_0^t \sum_{l=1}^{n} H_{i,l}^{(n)}(s) H_{j,l}^{(n)}(s) \lambda_l^{(n)}(s)\mathrm{d}s \xrightarrow{P} A_{ij}(t) \tag{B22}$$

as $n \to \infty$ and that for all i and $\epsilon > 0$

$$\int_0^1 \sum_{l=1}^{n} H_{il}^{(n)}(u)^2 \lambda_l^{(n)}(u) \mathbb{I}\left[\left|H_{il}^{(n)}(u)\right| > \epsilon\right] \mathrm{d}u \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty. \tag{B23}$$

Then $W^{(n)} \xrightarrow{D} W^{(\infty)}$ as $n \to \infty$ in $D([0, 1]^p)$.

Appendix C Extended results

C.1 The impact of the number of the genes in the biomarker panel

Here we present the performance of the screened biomarker panel with varying numbers of genes for HCC transcriptome dataset in Fig. C2. As we can see, when the number of genes is few (1-7) the biomarker panel cannot achieve satisfied results. Nevertheless, as the number of genes selected in the biomarker panel increases, we note an improvement in average performance alongside a reduction in standard error, with peak performance attained at a gene number of approximately 10. Beyond this point, adding more genes leads to a slight decline in performance. The phenomenon indicates the importance of including essential genes within the biomarker panel to ensure robust generalization capabilities. Identifying the optimal gene number for the biomarker panel remains a challenging yet crucial task, potentially guided by prior knowledge.

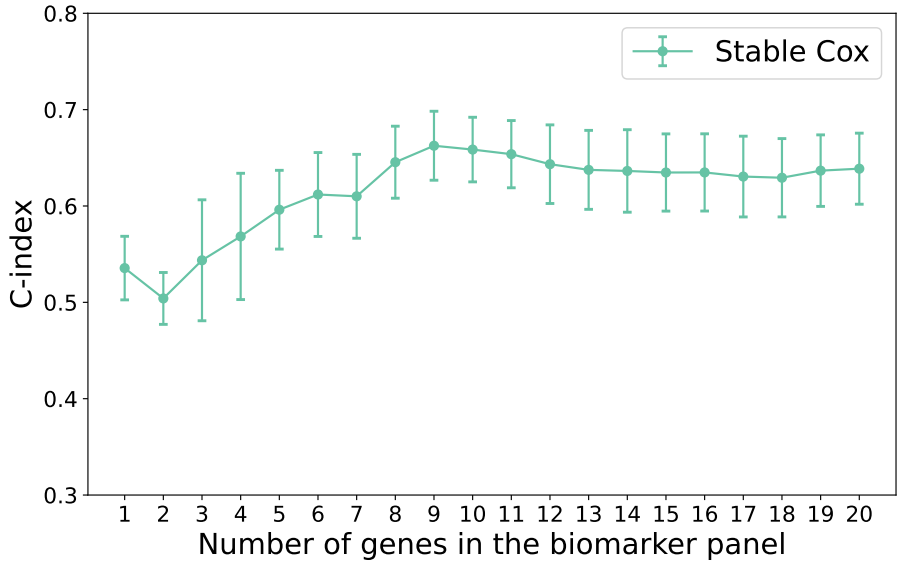

Fig. C2: **The performance of the different number of genes within the biomarker panel for HCC transcriptome dataset.** The dot represents the mean value of the C-index across three testing cohorts. The error bars represent the mean values +/- SD of three testing cohorts.

C.2 The sensitivity analysis of independence assumption

Here we would like to conduct experiments to examine the sensitivity of the independence assumption among covariates after reweighing. Specifically, we adjust the initial learning rate of the classifier in Eq.10 of main text to control its ability to generate accurate sample weights. We take 15 values as learning rates from the range [1e-4, 3e-6] with equal log-scale spacing. Subsequently, we assess the dependence between variables by summing pairwise HSIC [20] values of reweighted covariates, denoted as $\sum_{i,j}\text{HSIC}(w(X), X_i, X_j)$ [21]. We plot the points of the average C-index of testing sets and the corresponding $\sum_{i,j}\text{HSIC}(w(X), X_i, X_j)$ value in Fig. C3a. Then, we use a B-spline smoothing method [22] to fit a smoothed line for these points. Additionally, we display the C-index of the Cox PH with a horizontal black dotted line and the corresponding unweighted HSIC value with a vertical black dotted line. As illustrated, lower values of $\sum_{i,j}\text{HSIC}(w(X), X_i, X_j)$ correspond to a fully trained classifier where the learned weights have successfully minimized dependence between variables, thereby enhancing our model's performance and significantly outperforming the Cox PH model. However, as $\sum_{i,j}\text{HSIC}(w(X), X_i, X_j)$ increases, the performance of the Stable Cox model begins to deteriorate, aligning closely with that of the Cox PH model when the classifier is severely undertrained (near the horizontal line scenarios). The phenomenon indicates that the independence between variables could influence the performance of Stable Cox. Nevertheless, the Stable Cox model consistently demonstrates superior performance compared to the Cox PH model across a

broad range of independence values, particularly when the independence value of the reweighted dataset is lower than that of the original unweighted dataset, highlighting the robustness of our model.

Moreover, we present the sensitivity analysis of independence assumption on breast cancer transcriptome dataset. We take 15 values as learning rates from the range [1e-3, 3e-5] with equal log-scale spacing. As shown in Fig. C3b, lower summations of $\sum_{i,j} \text{HSIC}(w(X), X_i, X_j)$ are indicative of a classifier that is fully trained, wherein the learned weights effectively reduce the dependence among variables, thus boosting the performance of our model and substantially outshining the Cox PH model. Conversely, as $\sum_{i,j} \text{HSIC}(w(X), X_i, X_j)$ increases, the efficacy of the Stable Cox model declines, particularly in the extreme scenario where the classifier is markedly undertrained. Such observations suggest that the variable independence impacts the performance of the Stable Cox. Despite these fluctuations, the Stable Cox model maintains superior performance compared to the Cox PH model across a wide range of independence values (or initial learning rates), highlighting our model's robustness in practical applications.

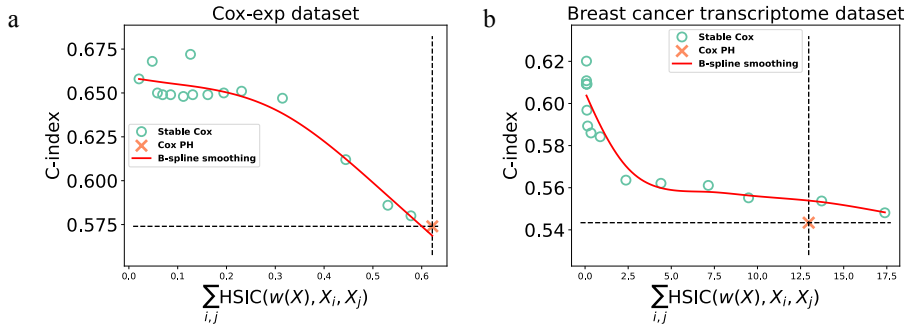

Fig. C3: **The sensitivity analysis of independence assumption. a,** Simulated Cox-exp dataset. The dot represents the mean value of the C-index across all testing cohorts. **b,** Breast cancer transcriptome dataset. The dot represents the mean value of the C-index across three testing cohorts. With the power of the classifier increased (lower $\sum_{i,j} \text{HSIC}(w(X), X_i, X_j)$), Stable Cox outperforms Cox PH with a larger margin.

C.3 The results of time-dependent Brier Score and AUC score metrics

Here we present the results of the time-dependent Brier Score [23] and time-dependent AUC score [24] on three cancer transcriptome datasets and two clinical datasets. The C-index, time-dependent Brier Score and time-dependent AUC score could reflect the accuracy of the survival model from different perspectives. Compared with the time-dependent Brier Score and AUC score which mainly measure the accuracy at a specific time point, C-index evaluates the proportion of all possible paired samples for

which the model correctly predicts their order. The C-index emphasizes the correct prediction of order, that is, the extent to which the model can distinguish between high-risk and low-risk individuals, which is crucial for our downstream task, i.e., subgroup stratification.

In clinical studies, metrics of primary interest often include the 1-year, 3-year, and 5-year survival rates. However, not all datasets, such as certain cohorts from the HCC cancer and melanoma datasets, have a follow-up duration that supports the calculation of 5-year or 3-year metrics. Therefore, we compute the 5-year, 3-year, and 1-year metrics for the breast transcriptome cancer, HCC transcriptome and melanoma transcriptome datasets, respectively, based on their longest recorded survival times. For clinical survival datasets, we calculate their 5-year metrics. To simplify, we report results from the Cox PH and Stable Cox models to validate the effectiveness of our proposed framework. Note that we utilize the same hyperparameters as those used in the C-index results.

The results of the time-dependent AUC of transcriptome datasets are summarized in Fig. C4. The AUC assesses model's discriminative performance at the given time point, measuring how well it separates those who have an event from those who do not at each time point. A higher AUC value indicates superior performance. Overall, Stable Cox significantly outperforms Cox PH on three datasets (from 16.1% to 22.5% overall improvements in the top 10 panel), demonstrating its enhanced capability to differentiate between patients at each time point. This advancement aids physicians in providing personalized medicine by tailoring treatment plans based on a patient's projected prognosis. The results of the time-dependent Brier Score of transcriptome datasets are summarized in Fig. C5. The Brier Score focuses more on the accuracy and calibration of predicted probabilities at the specific time point, with a lower score indicating better performance. The Stable Cox model shows a significantly lower Brier Score compared to traditional models across the three cancer transcriptome datasets (from 6.8% to 22.4% overall improvements in the top 10 panel). These findings suggest that our model can deliver precise survival probability estimates that are essential for guiding personalized treatment strategies at each crucial time point. Overall, the trends and conclusions drawn from the time-dependent AUC and Brier Score results align with those observed from the C-index, reinforcing the robustness of our findings in evaluating model performance. Additionally, the results of time-dependent AUC and Brier Score of clinical datasets are summarized in Fig. C6. Our model still outperforms Cox PH with a large margin, demonstrating not only superior calibration of probabilities but also an enhanced capacity to differentiate between patients at critical time points based on clinical covariates.

C.4 Survival risk subgroup stratification

We present the Kaplan–Meier plots of the HCC transcriptome dataset and the melanoma transcriptome dataset in Fig. C7. Overall, our model has demonstrated the capability to stratify subgroups with markedly distinct survival curves, suggesting that it can identify biomarker panels for prognosis more effectively than the Cox PH model. Moreover, the Kaplan–Meier plots of two identified subgroups in each test subpopulation of lung cancer clinical data with respective OS and DFS outcomes are

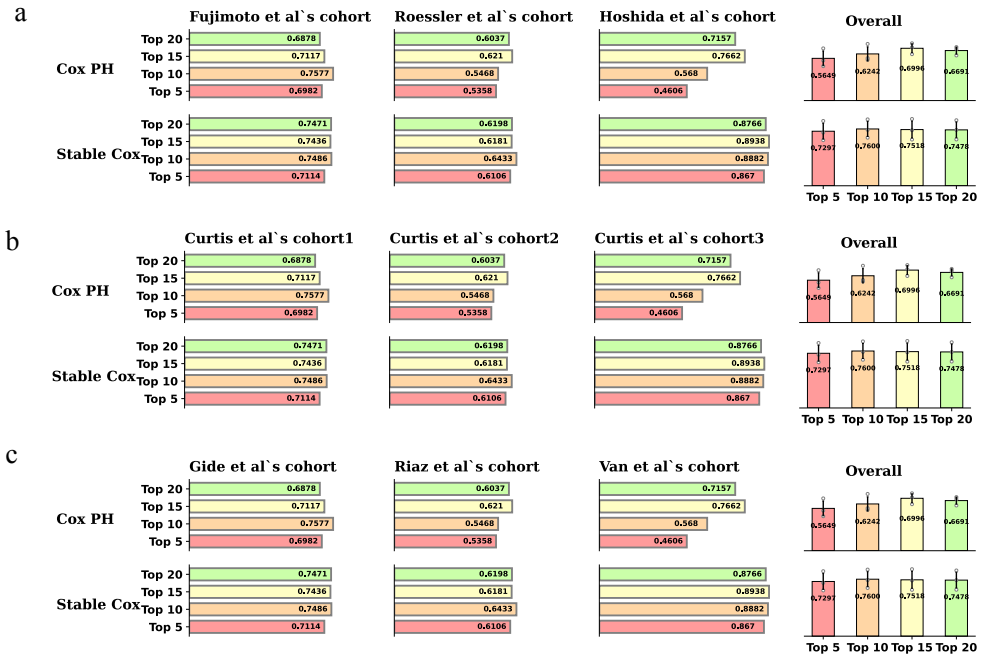

Fig. C4: **The time-dependent AUC score of Cox PH and Stable Cox. a,** HCC dataset (3-year). Training set is TCGA-LIHC cohort ($n = 351$). Fujimoto et al's cohort ($n = 203$), Roessler et al's cohort ($n = 209$) and Hoshida et al's cohort ($n = 80$) are testing cohorts. The mean value of the time-dependent AUC across three test cohorts, specifically for the top 5, 10, 15, and 20 features selected by each model, are reported in the Overall column. The error bars represent the mean values +/- SD. **b,** Breast cancer (5-year). Training set is one cohort of Curtis et al. ($n = 763$). Testing cohorts are cohort1 ($n = 521$), cohort2 ($n = 288$) and cohort3 ($n = 238$) from Curtis et al. The mean value of the time-dependent AUC across three test cohorts, specifically for the top 5, 10, 15, and 20 features selected by each model, are reported in the Overall column. The error bars represent the mean values +/- SD. **b,** Breast cancer (5-year). **c,** Melanoma dataset (1-year). Training set is Liu et al's cohort ($n = 120$). Gide et al's cohort ($n = 91$), Riaz et al's cohort ($n = 209$) and Van et al's cohort ($n = 80$) are testing cohorts. The mean value of the time-dependent AUC across three test cohorts, specifically for the top 5, 10, 15, and 20 features selected by each model, are reported in the Overall column. The error bars represent the mean values +/- SD. The AUC assesses model's discriminative performance at the given time point, measuring how well it separates those who have an event from those who do not at each time point. A higher value of AUC indicates better performance. The average improvement on three independent testing cohorts shows the stable performance of our method.

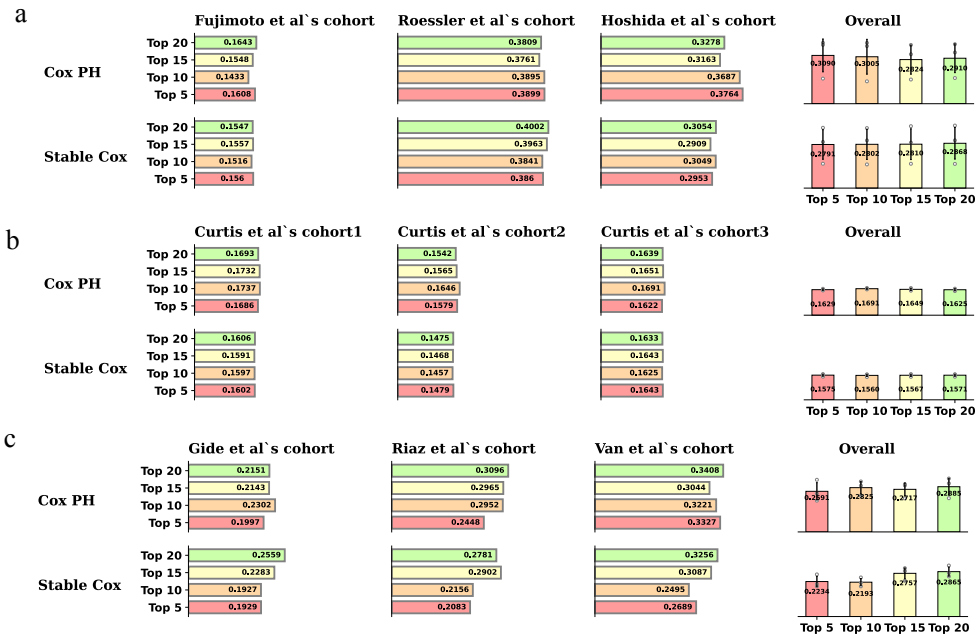

Fig. C5: **The Brier Score of Cox PH and Stable Cox. a,** HCC dataset (3-year). Training set is TCGA-LIHC cohort ($n = 351$). Fujimoto et al's cohort ($n = 203$), Roessler et al's cohort ($n = 209$) and Hoshida et al's cohort ($n = 80$) are testing cohorts. The mean value of the Brier Score across three test cohorts, specifically for the top 5, 10, 15, and 20 features selected by each model, are reported in the Overall column. The error bars represent the mean values +/- SD. **b,** Breast cancer (5-year). Training set is one cohort of Curtis et al. ($n = 763$). Testing cohorts are cohort1 ($n = 521$), cohort2 ($n = 288$) and cohort3 ($n = 238$) from Curtis et al. The mean value of the Brier Score across three test cohorts, specifically for the top 5, 10, 15, and 20 features selected by each model, are reported in the Overall column. The error bars represent the mean values +/- SD. **c,** Melanoma dataset (1-year). Training set is Liu et al's cohort ($n = 120$). Gide et al's cohort ($n = 91$), Riaz et al's cohort ($n = 209$) and Van et al's cohort ($n = 80$) are testing cohorts. The mean value of the Brier Score across three test cohorts, specifically for the top 5, 10, 15, and 20 features selected by each model, are reported in the Overall column. The error bars represent the mean values +/- SD. The Brier Score focuses more on the accuracy and calibration of predicted probabilities at the given time point. A lower value of the Brier Score indicates better performance. The average improvement on three independent testing cohorts shows the stable performance of our method.

shown in Fig. C8 and C9 respectively. As we can see, our model could identify subgroups with distinct survival curves in each clinical subpopulation, demonstrating our model could reduce the risk of applying the method to potential subpopulations.

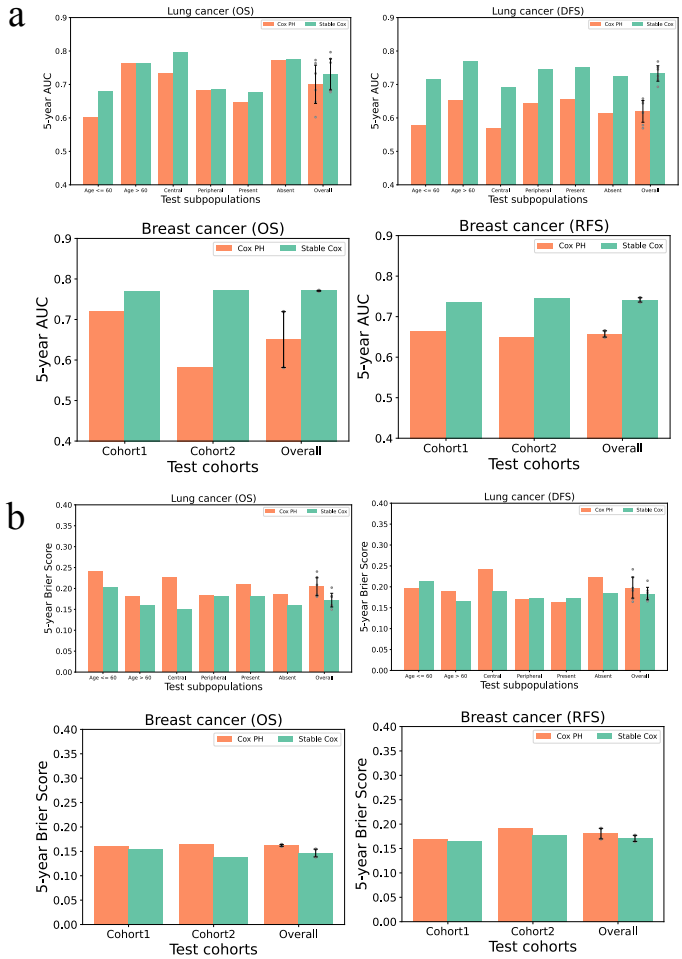

Fig. C6: **The AUC score and Brier Score of Cox PH and Stable Cox in clinical survival data. a,** 5-year AUC score of lung cancer and breast cancer clinical dataset. The higher value of the AUC score indicates better performance. For lung cancer clinical dataset, we randomly select 40% patients ($n = 240$) as training set and age<=60 ($n = 43$) or age>60 ($n = 113$), tumor location was classified as 'central' ($n = 47$) or 'peripheral' ($n = 107$), obstructive pneumonitis/atelectasis (OA) is 'present' ($n = 80$) or 'absent' ($n = 76$) as testing sets. For breast cancer clinical dataset, the methods are trained on one cohort ($n = 394$) from Curtis et al. and tested on two testing cohort1 ($n = 273$) and cohort2 ($n = 177$). Overall is the average performance across testing cohorts. The error bars represent the mean values +/- SD. **b,** 5-year Brier Score of lung cancer and breast cancer dataset. The datasets used are the same as the legend a. Overall is the average performance across testing cohorts. The error bars represent the mean values +/- SD. A lower value of the Brier Score indicates better performance. The average improvement on multiple independent testing cohorts/subpopulation shows the stable performance of our method.

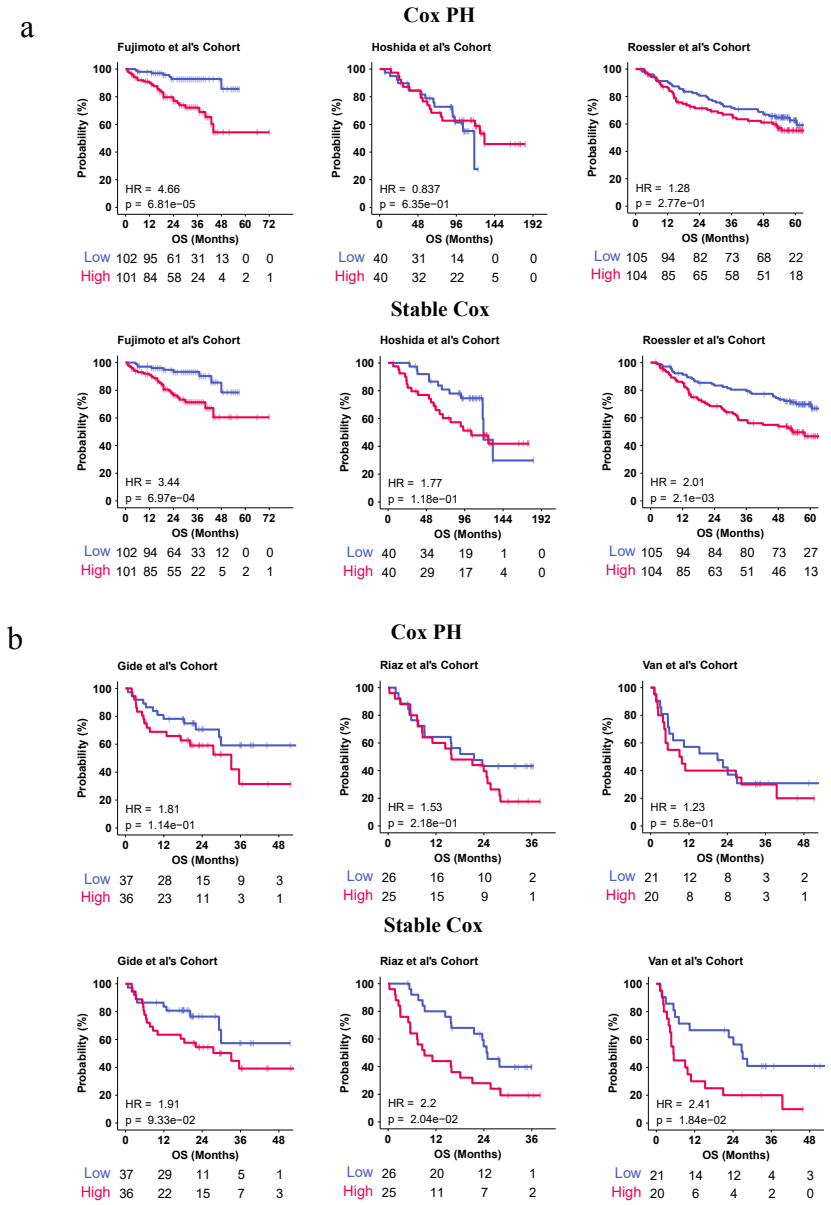

Fig. C7: **Kaplan–Meier plots of HCC transcriptome dataset and melanoma transcriptome dataset. a,** Kaplan–Meier plot of two subgroups (above and below the median of $\beta^T X$) separated by the top 10 genes identified by Cox PH (above) and Stable Cox (below) in the HCC transcriptome dataset. **b,** Kaplan–Meier plot of two subgroups (above and below the median of $\beta^T X$) separated by the top 10 genes identified by Cox PH (above) and Stable Cox (below) in the melanoma transcriptome dataset. The p-value is computed using a two-sided log-rank test. HR is calculated by a univariate Cox regression of the subgroup assignment variable on outcomes.

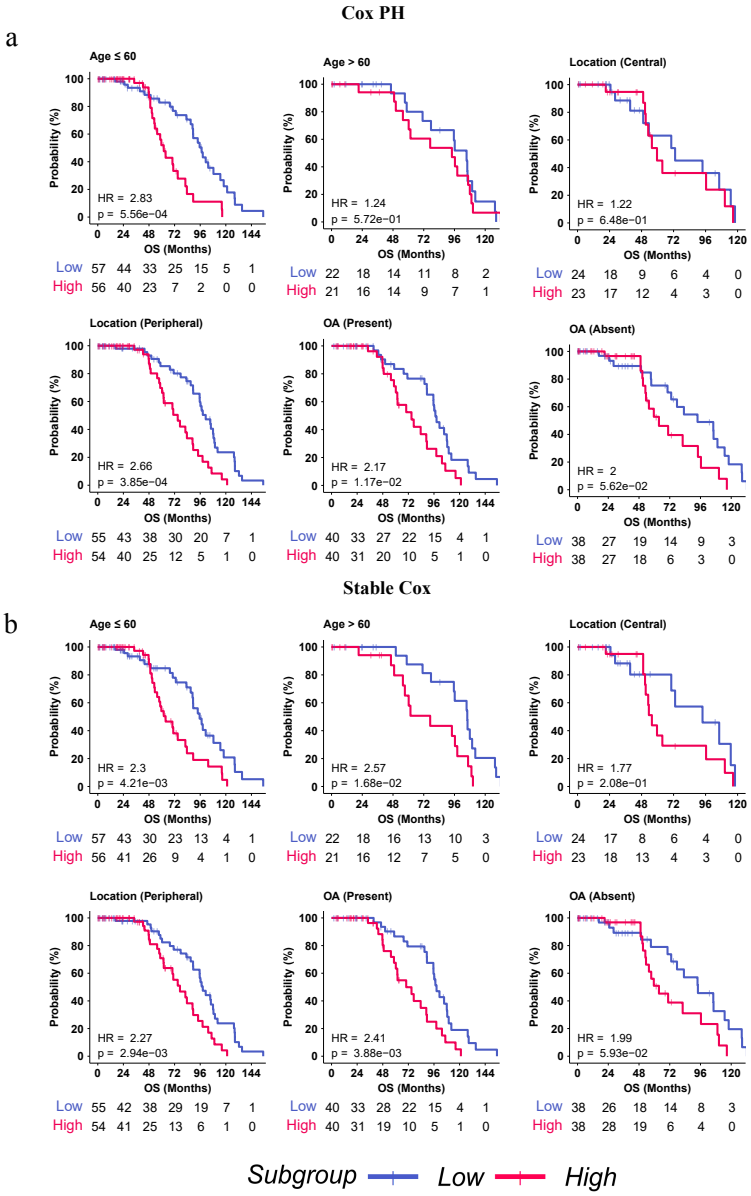

Fig. C8: **Kaplan–Meier plots of two risk-stratified subgroups in each test subpopulation of lung cancer clinical data (OS). a,** Kaplan–Meier plot of two subgroups (above and below the median of $\beta^T X$) separated by the learned coefficients of Cox PH. **b,** Kaplan–Meier plot of two subgroups (above and below the median of $\beta^T X$) separated the learned coefficients of Stable Cox. The risk subgroups identified by Stable Cox are more significantly different than Cox PH. The p-value is computed using a two-sided log-rank test. HR is calculated by a univariate Cox regression of the subgroup assignment variable on outcomes.

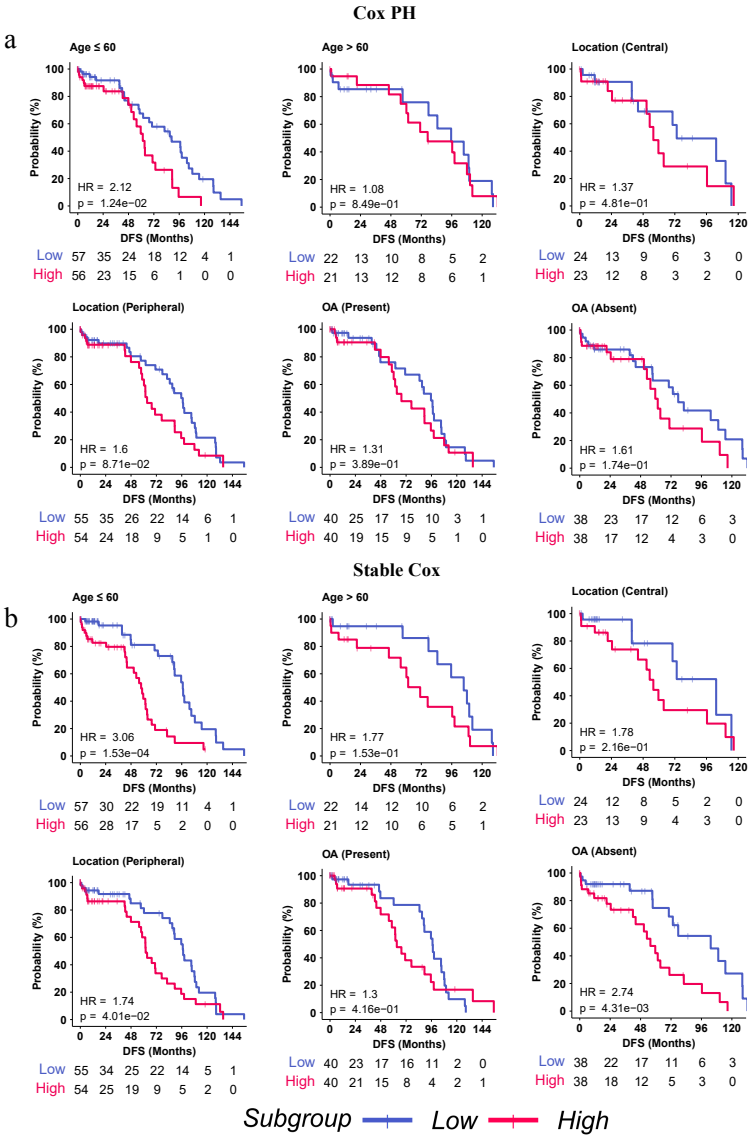

Fig. C9: **Kaplan–Meier plots of two risk-stratified subgroups in each test subpopulation of lung cancer clinical data (DFS). a,** Kaplan–Meier plot of two subgroups (above and below the median of $\beta^T X$) separated by the learned coefficients of Cox PH. **b,** Kaplan–Meier plot of two subgroups (above and below the median of $\beta^T X$) separated by the learned coefficients of Stable Cox. The risk subgroups stratified by Stable Cox are more significantly different than Cox PH. The p-value is computed using a two-sided log-rank test. HR is calculated by a univariate Cox regression of the subgroup assignment variable on outcomes.

References

[1] Jiang, Y., Sun, A., Zhao, Y., Ying, W., Sun, H., Yang, X., Xing, B., Sun, W., Ren, L., Hu, B., *et al.*: Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature **567**(7747), 257–261 (2019)

[2] Gao, Q., Zhu, H., Dong, L., Shi, W., Chen, R., Song, Z., Huang, C., Li, J., Dong, X., Zhou, Y., *et al.*: Integrated proteogenomic characterization of hbv-related hepatocellular carcinoma. Cell **179**(2), 561–577 (2019)

[3] Lian, Q., Wang, S., Zhang, G., Wang, D., Luo, G., Tang, J., Chen, L., Gu, J.: Hccdb: a database of hepatocellular carcinoma expression atlas. Genomics, proteomics & bioinformatics **16**(4), 269–275 (2018)

[4] Zhu, Y., Qiu, P., Ji, Y.: Tcga-assembler: open-source software for retrieving and processing tcga data. Nature methods **11**(6), 599–600 (2014)

[5] Grinchuk, O.V., Yenamandra, S.P., Iyer, R., Singh, M., Lee, H.K., Lim, K.H., Chow, P.K.-H., Kuznetsov, V.A.: Tumor-adjacent tissue co-expression profile analysis reveals pro-oncogenic ribosomal gene signature for prognosis of resectable hepatocellular carcinoma. Molecular oncology **12**(1), 89–113 (2018)

[6] Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., Tanaka, H., Taniguchi, H., Kawakami, Y., Ueno, M., *et al.*: Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nature genetics **48**(5), 500–509 (2016)

[7] Roessler, S., Jia, H.-L., Budhu, A., Forgues, M., Ye, Q.-H., Lee, J.-S., Thorgeirsson, S.S., Sun, Z., Tang, Z.-Y., Qin, L.-X., *et al.*: A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer research **70**(24), 10202–10212 (2010)

[8] Hoshida, Y., Villanueva, A., Kobayashi, M., Peix, J., Chiang, D.Y., Camargo, A., Gupta, S., Moore, J., Wrobel, M.J., Lerner, J., *et al.*: Gene expression in fixed tissues and outcome in hepatocellular carcinoma. New England Journal of Medicine **359**(19), 1995–2004 (2008)

[9] Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., *et al.*: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature **486**(7403), 346–352 (2012)

[10] Liu, D., Schilling, B., Liu, D., Sucker, A., Livingstone, E., Jerby-Arnon, L., Zimmer, L., Gutzmer, R., Satzger, I., Loquai, C., *et al.*: Integrative molecular and clinical modeling of clinical outcomes to pd1 blockade in patients with metastatic melanoma. Nature medicine **25**(12), 1916–1927 (2019)

[11] Hugo, W., Zaretsky, J.M., Sun, L., Song, C., Moreno, B.H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., *et al.*: Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. Cell **165**(1), 35–44 (2016)

[12] Gide, T.N., Quek, C., Menzies, A.M., Tasker, A.T., Shang, P., Holst, J., Madore, J., Lim, S.Y., Velickovic, R., Wongchenko, M., *et al.*: Distinct immune cell populations define response to anti-pd-1 monotherapy and anti-pd-1/anti-ctla-4 combined therapy. Cancer cell **35**(2), 238–255 (2019)

[13] Riaz, N., Havel, J.J., Makarov, V., Desrichard, A., Urba, W.J., Sims, J.S., Hodi, F.S., Martín-Algarra, S., Mandal, R., Sharfman, W.H., *et al.*: Tumor and microenvironment evolution during immunotherapy with nivolumab. Cell **171**(4), 934–949 (2017)

[14] Van Allen, E.M., Miao, D., Schilling, B., Shukla, S.A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Geukes Foppen, M.H., Goldinger, S.M., *et al.*: Genomic correlates of response to ctla-4 blockade in metastatic melanoma. Science **350**(6257), 207–211 (2015)

[15] Gu, K., Lee, H.Y., Lee, K., Choi, J.Y., Woo, S.Y., Sohn, I., Kim, H.K., Choi, Y.S., Kim, J., Zo, J.I., *et al.*: Integrated evaluation of clinical, pathological and radiological prognostic factors in squamous cell carcinoma of the lung. PLoS One **14**(10), 0223298 (2019)

[16] Lin, D.Y., Wei, L.-J.: The robust inference for the cox proportional hazards model. Journal of the American statistical Association **84**(408), 1074–1078 (1989)

[17] Bender, R., Augustin, T., Blettner, M.: Generating survival times to simulate cox proportional hazards models. Statistics in medicine **24**(11), 1713–1723 (2005)

[18] Andersen, P.K., Gill, R.D.: Cox's regression model for counting processes: a large sample study. The annals of statistics, 1100–1120 (1982)

[19] Lenglart, É.: Relation de domination entre deux processus. In: Annales de L'institut Henri Poincaré. Section B. Calcul des Probabilités et Statistiques, vol. 13, pp. 171–179 (1977)

[20] Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: International Conference on Algorithmic Learning Theory, pp. 63–77 (2005). Springer

[21] Fan, S., Wang, X., Shi, C., Cui, P., Wang, B.: Generalizing graph neural networks on out-of-distribution graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)

[22] Wahba, G.: Estimating the smoothing parameter. Spline Models for Observational

Data, 45–65 (1990)

[23] Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine **18**(17-18), 2529–2545 (1999)

[24] Hung, H., Chiang, C.-T.: Estimation methods for time-dependent auc models with survival data. Canadian Journal of Statistics **38**(1), 8–26 (2010)