Explainable Censored Learning: Finding Critical Features with Long Term Prognostic Values for Survival Prediction

Xinxing Wu\textsuperscript{a}, Chong Peng\textsuperscript{b}, Richard Charnigo\textsuperscript{c}, and Qiang Cheng\textsuperscript{a,1}

\textsuperscript{a}Institute for Biomedical Informatics, University of Kentucky, Lexington, KY 40506, USA
\textsuperscript{b}Department of Computer Science and Engineering, Qingdao University, Shandong 266071, China
\textsuperscript{c}Department of Statistics, University of Kentucky, Lexington, KY 40536, USA

Abstract

Interpreting critical variables involved in complex biological processes related to survival time can help understand prediction from survival models, evaluate treatment efficacy, and develop new therapies for patients. Currently, the predictive results of deep learning (DL)-based models are better than or as good as standard survival methods, they are often disregarded because of their lack of transparency and little interpretability, which is crucial to their adoption in clinical applications. In this paper, we introduce a novel, easily deployable approach, called EXplainable CEnsored Learning (EXCEL), to iteratively exploit critical variables and simultaneously implement (DL) model training based on these variables. First, on a toy dataset, we illustrate the principle of EXCEL; then, we mathematically analyze our proposed method, and we derive and prove tight generalization error bounds; next, on two semi-synthetic datasets, we show that EXCEL has good anti-noise ability and stability; finally, we apply EXCEL to a variety of real-world survival datasets including clinical data and genetic data, demonstrating that EXCEL can effectively identify critical features and achieve performance on par with or better than the original models. It is worth pointing out that EXCEL is flexibly deployed in existing or emerging models.

\textsuperscript{1}Correspondence should be addressed to: qiang.cheng@uky.edu. Address: RM 230, Multidisciplinary Science Building, KY 40506, USA. Phone number: 859-323-7238.
for explainable survival data in the presence of right censoring.

**Keywords:** Censored Learning, deep learning, feature selection, generalization error bound

---

1. Introduction

Different from common prediction tasks in machine learning, survival analysis focuses on modeling censored data. The Cox proportional-hazards (CPH) model [1] is a standard regression model widely used in epidemiological studies and clinical trials, e.g., for analyzing gene expression- or clinical variable-related time-to-event survival data. Recently, deep learning (DL)-based techniques have been adapted for CPH, such as DeepSurv [2], Cox-nnet [3], and PASNet [4]. Although the predictive results of these DL-based models are better than or as good as standard survival methods, they are often disregarded because of their lack of transparency and little interpretability [5], which is crucial to their adoption in clinical applications. In survival analysis, an important goal is to identify potential risk factors from thousands of variables to evaluate treatment effects [7]. Kaplan-Meier (KM) curves [8] and Log-rank tests [9, 10] are two typical methods for univariate factor analysis, describing survival in terms of one factor but ignoring the impact of the other variables [7]. The CPH model, as a representative survival analysis method, can assess survival prediction based on risk variables. To identify informative variables in the CPH model, variable selection via penalization like employing an $\ell_0$-norm is statistically preferred and has attracted much attention in recent years. Because the exact $\ell_0$ optimization needs to search the space of all possible subsets, which is an NP-hard problem [11,12,13], a variety of convex surrogate functions for $\ell_0$-norm have been developed. The Lasso [14,15] penalty, i.e., $\ell_1$-norm regularization, is a popular

---

2KM estimation is a non-parametric method which uses data from subjects at one level of the univariate factor to construct a non-increasing step function which approximates the survival probability in relation to time. Log-rank test is a non-parametric method which uses data from independent subjects at different levels of the univariate factor to test a null hypothesis of identical survival probabilities in relation to time.
Figure 1: Schematic illustration of our EXCEL approach embedded in the standard CPH model (denoted as Maxk-CPH) and its comparison with the baseline model on dataset Worcester Heart Attack Study (WHAS500). (a) The statistic of dataset WHAS500. We directly import function `sksurv.datasets.load_whas500()` (for more details, see https://scikit-survival.readthedocs.io/en/latest/api/generated/sksurv.datasets.load_whas500.html) to download WHAS500 [6]; (b) and (c) are the architectures of Maxk-CPH and the baseline, respectively; (d) and (e) are the loss and CI comparisons of Maxk-CPH with the baseline model with different epochs on the test set; (f) The comparison of variable selection of our method with the baseline; (g)-(h) The performance comparison of our method with the baseline model in CI and IBS with 5 random training/test set splits; (i) The comparison of the number of used variables of our method with the baseline.
one; however, [16] pointed out that convex surrogate functions, such as \( \ell_1 \)-norm, generally bring non-negligible estimation biases and thus an adverse effect on the predictive ability. Lately, [17] used two convex programs to optimize the CPH model with a non-convex penalty for variable selection; [18] generalized the primal-dual active set algorithm for general convex loss functions to solve the best subset constraint problem, such as the sparse CPH model with an \( \ell_0 \)-norm-based constraint, and developed the corresponding R package, BeSS.

Although these existing works can select useful variables in the baseline CPH model, effective ways to identify critical factors in DL-based survival analysis is evasive; in particular, due to their being generally non-convex, it is impossible to directly apply these existing techniques for the CPH model to the DL-based models.

In this paper, we aim to find critical features with long-term prognostic values and help establish necessary explainability for the prognosis of survival time in DL-based survival analysis over censored data. To overcome the bias issue of the \( \ell_1 \)-norm based sparse CPH model, we develop an approach for DL-based survival models by leveraging a direct counterpart of the \( \ell_0 \)-norm regularization. Inspired by the recent development of feature selection algorithms in unsupervised learning [19, 20], we adopt an \( \ell_1 \)-norm regularization to score all input variables globally, and then determine the \( k \) variables with the maximum scores as a counterpart of \( \ell_0 \) norm, to iteratively revamp the learning of our target model. We dub this approach EXplainable CEnsored Learning (EXCEL). Notably, EXCEL has a high stability advantage in that its identified factors do not change much with different subsets of training examples, in sharp contrast to other methods; see Figure 3(k). The schematic of EXCEL is illustrated in Figure 1 with a toy dataset, WHAS500. The EXCEL-extended CPH model for selecting \( k \) features, denoted by Max\(_k\)-CPH, is shown in Figure 1(b). Compared to the standard CPH model shown in Figure 1(c), our proposed method has one more layer of slack variables and a Max\(_k\) embedded module (marked in green) as our target model. The slack variables are used to score the importance of all input survival variables globally, while the Max\(_k\) module implements learning.
based on $k$ selected variables with the maximum scores in each iteration. Figure 1 (d) and (e) show that, on the test set of WHAS500 with $k = 6$, with the increase of epoch, our method achieves a smaller loss and better concordance index (CI) than the standard CPH model. Figure 1 (f) and (g) indicate that our method performs better in CI and integrated Brier score (IBS) than the standard CPH model with 5 random training/test set splits. Figure 1(h) shows that our model uses 57.1% fewer variables than the standard CPH model.

Technically, our designed algorithm trains a (DL-based) survival model to identify potentially informative variables for survival learning globally; simultaneously, it leverages an embedded module to select a subset locally from the globally informative variables to examine their diversity, which is efficiently measured by their abilities to fit the survival time. In this way, the embedded module enables us to effectively screen out the critical variables for predicting the survival time. The global component turns out to play a crucial role in regularization to stabilize the variable selecting process, similar to [19]. By capitalizing on such a regularization, we find that the resulting embedded model can select representative variables and simultaneously capture the complex non-linear relationship between these variables and survival time. In Section Algorithm development and analysis, we derive and prove a tight generalization error bound for the difference between the solution of the lead model and that of the induced model, thereby providing theoretical certificates of using our proposed regularization for variable selection. Notably, our method can be deployed as a lightweight module, easily plugged in, and efficiently optimized along with the existing standard survival methods, including DeepSurv, Coxnet, and PASNet. In Section Experimental results, on two semi-synthetic datasets and three real-world survival datasets, including clinical and genetic data, with varying numbers of subjects and variables and portions of censored observations, we demonstrate that our proposed EXCEL approach can identify critically important risk factors from thousands of variables potentially related to the survival time and, with only a small subset of identified variables, it can achieve prognostic performance comparable to or better than the original
2. Algorithm development and analysis

Description of the algorithm. Set \([N] = \{1, 2, \ldots, N\}\). Let \(\mathcal{D} = \{x_i, E_i, T_i\}_{i=1}^N\) be a survival dataset consisting of \(N\) subjects, where \(x_i \in \mathbb{R}^{d \times 1}, i \in [N]\); \(E_i\) is the censoring indicator: \(E_i = 0\) if the survival time of subject \(i\) was (right) censored, and \(E_i = 1\) otherwise; \(T_i\) is the censored survival time for subject \(i\) as indicated by \(E_i\). Usually, a maximum likelihood-based approach is adopted to model the distribution \(S(t|x) = \mathbb{P}(T > t|x)\) from survival data [1]. And we assume all samples are bounded, i.e., \(\forall x_i, i \in [N], \exists C_1 \geq 0, \text{ such that} \sum_{i \in [N]} \|x_i\|_2 \leq C_1\).

Here, we use the average negative log partial likelihood function as the loss function, and we give a more general form as follows:

\[
\mathcal{L}(f, \mathcal{D}) \triangleq -\frac{1}{N_{E=1}} \sum_{i : E_i = 1} \left( f(x_i) - \log \sum_{T_j \geq T_i} \exp f(x_j) \right) + \lambda_1 \cdot \text{Reg}(f), \tag{1}
\]

where \(N_{E=1}\) is the number of patients of nonzero \(E\) (i.e., observable events), \(f(\cdot)\) is a generally a non-convex function, e.g., a neural network, and \(\text{Reg}(f)\) represents the regularizations on \(f(\cdot)\). The above model is a generalization of a number of existing survival models: 1) It reduces to DeepSurv if \(f\) is parameterized as a neural network and \(f \in \mathbb{R}\). 2) It reduces to Cox-net if \(f\) can be written as \(\beta^T g\), where \(g\) is a neural network, \(g \in \mathbb{R}^{m \times 1}\), \(\beta\) denotes the weights in a standard CPH model that need to be trained, \(\beta \in \mathbb{R}^{m \times 1}\), and \(m\) is an integer. 3) It becomes PASNet if the first layer of \(g\) can be further rewritten as \(W_{g_1} \odot M\), where \(\odot\) denotes element-wise multiplication and the elements of \(M\) are either one or zero, determining which associated weights are dropped during training. And \(W_{g_1}, M \in \mathbb{R}^{d \times k_1}\), \(k_1\) denotes the number of neurons in the

\[\text{For original models with all variables, we denote that the training of models directly uses all of the potentially variables of the sample, rather than a subset of variables of the input sample.}\]
fist layer of $g$.

Our proposed EXCEL approach introduces a sub-architecture to existing models and \( \L(f,W, D) \) becomes

\[
L(f, W_1, D) \triangleq -\frac{\lambda_0}{N}\sum_{i:E_i=1} \left( f(W_1 x_i) - \log \sum_{T_j \geq T_i} \exp f(W_1 x_j) \right) - \frac{\lambda_2}{N}\sum_{i:E_i=1} \left( f(W_i^{\text{max}_k} x_i) - \log \sum_{T_j \geq T_i} \exp f(W_i^{\text{max}_k} x_j) \right)
+ \lambda_1 \cdot \text{Reg}(f) + \lambda_3 \cdot \text{Reg}(W_i),
\]

where $W_i^{\text{max}_k} = \text{Diag}(w^{\text{max}_k}) (\in \mathbb{R}^{d \times d})$ and $w^{\text{max}_k} (\in \mathbb{R}^{d \times 1})$ is an operation to keep the $k$ largest entries of $w$ while making other entries 0. We generally take $\text{Reg}(W_i)$ as $\|W_i\|_1$. In practical computation, we constrain the elements in $W_1$ to be non-negative, to simplify the processing of variable scores; otherwise, during optimization, we need to take the absolute value of the scores first before performing $\text{Max}_k$ operation.

The optimization of (3) is an efficient extension of the DL-based survival models of (1), where the main difference comes from the second term of (3). In each iteration of back propagation, the second term of (3) would require the gradients of the variables corresponding to the maximal $k$ weights in magnitude during this iteration while not affecting the gradients of other variables. Note that the EXCEL-extended model in the second term essentially has the same architecture as the DL-based model, i.e., (1). Thus, the gradients of the second term can reuse (up to a multiplicative factor) those corresponding to the maximal $k$ weights from the first term, with only an additional ranking operation. Finding the maximal $k$ out of $d$ weights in each iteration has a worst-case efficiency $O(d \min\{\log d, k\})$ that is independent of $N$. Therefore, the efficiency of optimizing (3) is essentially the same as (1).

Notably, the EXCEL approach can be easily used for many existing survival models, and we illustrate the schematic of the EXCEL-extended models for
DeepSurv, Cox-nnet, and PASNet in Figure 2.

![Figure 2: Schematic illustration of the EXCEL-extended models for various DL-based survival models.](image)

**Evaluation metrics.** Two main metrics are adopted for evaluating the performance of survival models: 1) Concordance index (CI) \[21, 22\] is widely used to evaluate the ability of models to rank individuals by their risk for the time-to-event predictions. The larger the CI value, the better the performance. 2) Integrated Brier Score (IBS) \[23\] extends Brier score \[24\] to right-censored data for evaluating the accuracy of an estimated survival function at time \(t\). The smaller the IBS value, the better the performance. In our experiments, we directly import function `sksurv.metrics.concordance_index_censored()` and `sksurv.metrics.integrated_brier_score()` for calculating CI and IBS.

**Experimental setting.** In our experiments, we set the maximum number of epochs to be 150 for two semi-synthetic datasets NSurvUSPS_3vs8 and SurvUSPS_3vs8 and 1,000 for real-world datasets, including WHAS500, Breast cancer, SUPPORT, and GBM. We initialize the weights of variable selection layer by sampling uniformly from \(U[0.999999, 0.999999]\) and the other layers with the Xavier normal initializer \[25\]. We adopt the Adam optimizer \[26\] with an initialized learning rate of 0.0001. For the hyper-parameter setting, we perform a grid search on the validation set, and then choose the optimal one. Taking (3) as an example, we mainly need to optimize four hyper-parameters. On the validation set split from training set, we optimize \(\lambda_0\) and \(\lambda_2\) over the search space:
[0.4, 0.8, 1.2, 1.6], and λ_1 and λ_3 over the search space: [0.0001, 0.0005, 0.001, 0.05, 0.01, 0.05, 0.1, 0.5].

For all datasets, we randomly split them into training and testing sets by a ratio of 80:20; the results are averaged over 10 runs of 10 different random splits for the results on the three datasets. All experiments are implemented with Python 3.7.8. The codes will be made publicly available upon acceptance.

**Generalization error bounds.** Now, we study the error bound for the difference between the solution of the lead model and that of the induced model. For analytical tractability, we focus on shallow neural networks (NNs) with the \( \ell_2 \)-norm regularization here and our analysis may be extended to more general NNs with more layers and we will leave the extension as a line for future study. The optimization consists of minimizing the following average negative log partial likelihood objective function:

\[
\mathcal{L}(w, D) = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left( x_i^T w - \log \sum_{T_j \supset T_i} \exp(x_j^T w) \right) - \frac{\lambda_2}{N_{E=1}} \sum_{i:E_i=1} \left( x_i^T w_{\text{max}k} - \log \sum_{T_j \supset T_i} \exp(x_j^T w_{\text{max}k}) \right) + \frac{\lambda_3}{2} \|w\|_2^2.
\]

In practical computation, we constrain the elements in \( w \) to be non-negative, to simplify the processing of variable scores; otherwise, during optimization, we need to take the absolute value of the scores first before performing Max_k operation.

Let \( \hat{w} = \arg \min_{w \in \mathbb{R}^{d \times 1}} \mathcal{L}(w, D) \), we have

**Theorem 2.1.**

\[
\|\hat{w} - \hat{w}_{\text{max}k}\|_2^2 \leq \frac{2\hat{w}^T(I_d - I_d(k))}{\max \{\lambda_2 N_{E=1}, \lambda_3 N_{E=1}\}} \sum_{i:E_i=1} \left( \frac{\sum_{T_j \supset T_i} (x_i - x_j) \exp(x_j^T \hat{w}_{\text{max}k})}{\sum_{T_j \supset T_i} \exp(x_j^T \hat{w}_{\text{max}k})} \right),
\]

(4)
where $I_d$ is a $d \times d$ identity matrix, and $I_d(k)$ denotes a $d \times d$ identity matrix with $k$ ones on its diagonal.

Proof. Let $\hat{w}^{\text{max}_k} = I_d(k)\hat{w}$, and the gradient of the original loss function $\nabla L$ computed at $\hat{w}^{\text{max}_k}$ is as follows:

$$
\nabla L(\hat{w}^{\text{max}_k}) = -\frac{1}{N}\sum_{i:E_i=1} \left( x_i - \frac{\sum_{T_j \geq T_i} x_j \exp(x_j^T \hat{w}^{\text{max}_k})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w}^{\text{max}_k})} \right) - \frac{\lambda_2}{N}\sum_{i:E_i=1} \left( I_d(k)x_i - \frac{\sum_{T_j \geq T_i} I_d(k)x_j \exp(x_j^T \hat{w}^{\text{max}_k})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w}^{\text{max}_k})} \right) + \lambda_3 \hat{w}^{\text{max}_k}.
$$

(5)

Note that $\hat{w}$ is the the optimal solution to (3), so we have

$$
\nabla L(\hat{w}) = 0.
$$

(6)

Then, we get

$$
\begin{align*}
\nabla L(\hat{w}^{\text{max}_k}) &= \nabla L(\hat{w}^{\text{max}_k}) - I_d(k)\nabla L(\hat{w}) \\
&= -\frac{1}{N}\sum_{i:E_i=1} \left( x_i - \frac{\sum_{T_j \geq T_i} x_j \exp(x_j^T \hat{w}^{\text{max}_k})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w}^{\text{max}_k})} \right) \\
&= -\frac{1}{N}\sum_{i:E_i=1} \left( x_i - \frac{\sum_{T_j \geq T_i} x_j \exp(x_j^T \hat{w}^{\text{max}_k})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w}^{\text{max}_k})} \right).
\end{align*}
$$

10
\[
- \frac{\lambda_2}{N_E} \sum_{i : E_i = 1} \left( I_d(k) x_i - \frac{\sum_{T_j \geq T_i} I_d(k) x_j \exp(x_j^T \hat{w}_{\text{max}})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w}_{\text{max}})} \right) \\
+ \lambda_3 \hat{w}_{\text{max}} - \lambda_3 I_d(k) \hat{w} + \frac{I_d(k)}{N_E} \sum_{i : E_i = 1} \left( x_i - \frac{\sum_{T_j \geq T_i} x_j \exp(x_j^T \hat{w})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w})} \right) \\
+ \frac{\lambda_2 I_d(k)}{N_E} \sum_{i : E_i = 1} \left( I_d(k) x_i - \frac{\sum_{T_j \geq T_i} I_d(k) x_j \exp(x_j^T I_d(k) \hat{w})}{\sum_{T_j \geq T_i} \exp(x_j^T I_d(k) \hat{w})} \right)
\]

(7)

Let
\[
\theta = \frac{\hat{w}_{\text{max}} - \hat{w}}{\| \hat{w}_{\text{max}} - \hat{w} \|_2}
\]

(8)

Notice that \( \mathcal{L} \) is \( \mu \)-strongly convex, thus, \( \exists \mu > 0, \forall t > 0 \), we obtain
\[
\mathcal{L}(\hat{w}_{\text{max}} - t \theta) - \mathcal{L}(\hat{w}_{\text{max}}) \geq -t \theta^T \nabla \mathcal{L}(\hat{w}_{\text{max}}) + \frac{\mu t^2}{2}
\]

(9)

Noting that \( \hat{w} \) is the optimal solution of \( \mathcal{L}(\hat{w}) \), we have
\[
-\theta^T \nabla \mathcal{L}(\hat{w}_{\text{max}}) + \frac{\mu t}{2} \leq 0
\]

(10)

Then we get
\[
\| \hat{w}_{\text{max}} - \hat{w} \|_2 \leq t \leq \frac{2}{\mu} \theta^T \nabla \mathcal{L}(\hat{w}_{\text{max}}).
\]

(11)
Plugging (7) and (8) into (11), we have

\[ \parallel \hat{w}_{\text{max}} - \hat{w} \parallel_2^2 \leq -\frac{2(\hat{w}_{\text{max}} - \hat{w})^T}{\mu N_{E=1}} \sum_{i:E_i=1} \left( x_i - \frac{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w}_{\text{max}})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w})} \right) \]

\[ + 2(\hat{w}_{\text{max}} - \hat{w})^T \sum_{i:E_i=1} \left( I_d(k)x_i - \frac{\sum_{T_j \geq T_i} I_d(k)x_j \exp(x_j^T \hat{w})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w})} \right) \]  

\[ = \frac{2\hat{w}^T(I_d - I_d(k))}{\mu N_{E=1}} \sum_{i:E_i=1} \left( x_i - \frac{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w}_{\text{max}})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w})} \right). \]  

(12)

Taking \( \mu = \max \{ \lambda_2, \lambda_3 \} \), the proof follows. \( \square \)

**Lemma 2.1.** \( \mathcal{L}(\hat{w}) \) has a Lipschitz continuous gradient with Lipschitz constant \( L \).

**Proof.** We compute the Hessian matrix of \( \mathcal{L}(\hat{w}) \) in (3) as follows:

\[ \nabla^2 \mathcal{L}(\hat{w}) = -\partial \left[ \frac{1}{N_{E=1}} \sum_{i:E_i=1} \left( x_i - \frac{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w})} \right) / \partial \hat{w} \right] \]

\[ - \partial \left[ \frac{\lambda_2}{N_{E=1}} \sum_{i:E_i=1} \left( I_d(k)x_i - \frac{\sum_{T_j \geq T_i} I_d(k)x_j \exp(x_j^T \hat{w})}{\sum_{T_j \geq T_i} \exp(x_j^T \hat{w})} \right) / \partial \hat{w} \right] \]

\[ + \lambda_3. \]  

(13)
Then, by Minkowski’s inequality [27], we easily get

\[ \| \nabla^2 \mathcal{L}(\hat{w}) \|_2 \]

\[ \leq \left( 1 + \lambda_2 \right) \max_{T_j \geq T_i} \| x_j^T \|_2 + \lambda_3 \]

(14)

Take \( L = (1 + \lambda_2) \max_{T_j \geq T_i} \{ \| x_j^T \|_2 \} + \lambda_3 \), the proof follows. \( \square \)

**Lemma 2.2** ([28]). If a function \( f(\cdot) \) is differentiable on \( \mathbb{R}^{d \times 1} \) and has an \( L \)-Lipschitz gradient, then \( \forall x, y \in \mathbb{R}^{d \times 1} \), we have

\[ |f(y) - f(x) - (y - x)^T \nabla f(x)| \leq \frac{L}{2} \| y - x \|_2. \]

(15)

**Lemma 2.3.**

\[ \frac{\theta^T \nabla \mathcal{L}(\hat{w}_{\text{max}})}{L} \leq \| \hat{w} - \hat{w}_{\text{max}} \|_2. \]

(16)
Proof. By Lemma 2.2, ∀δ ⩾ 0, we have the following inequality

\[ L(\hat{w}^{\text{max}} - \delta \theta) - L(\hat{w}^{\text{max}}) \leq -\delta \theta^T \nabla L(\hat{w}^{\text{max}}) + \frac{L\delta^2}{2}. \]  

(17)

By \|\theta\| = 1 and Lemma 2.1, we have

\[
\frac{\partial(-\delta \theta^T \nabla L(\hat{w}^{\text{max}}) + L\delta^2)}{\partial \delta}
= -\theta^T \nabla L(\hat{w}^{\text{max}}) + L\delta
\geq -\theta^T L(\hat{w}^{\text{max}} - \delta \theta)
= \frac{\partial(L(\hat{w}^{\text{max}} - \delta \theta) - L(\hat{w}^{\text{max}}))}{\partial \delta}
\]

(18)

And note that \( \hat{w} \) is the optimal solution, so \( -\theta^T \nabla L(\hat{w}^{\text{max}}) \leq 0 \). Then we can easily get

\[
\arg \min_{\delta > 0} \left\{ -\delta \theta^T \nabla L(\hat{w}^{\text{max}}) + \frac{L\delta^2}{2} \right\}
\leq \arg \min_{\delta > 0} \{ L(\hat{w}^{\text{max}} - \delta \theta) - L(\hat{w}^{\text{max}}) \}.
\]

(19)

Finally, we calculate the minima on the left and right sides of (19) as follows:

\[
\arg \min_{\delta > 0} \left\{ -\delta \theta^T \nabla L(\hat{w}^{\text{max}}) + \frac{L\delta^2}{2} \right\} = \frac{\theta^T \nabla L(\hat{w}^{\text{max}})}{L},
\]

(20)

and

\[
\arg \min_{\delta > 0} \{ L(\hat{w}^{\text{max}} - \delta \theta) - L(\hat{w}^{\text{max}}) \} = \|\hat{w}^{\text{max}} - \hat{w}\|_2,
\]

(21)

then we complete the proof. □

Theorem 2.2. 

\[
\|\hat{w} - \hat{w}^{\text{max}}\|_2^2
\geq \frac{\hat{w}^T (I_d - I_d(k))}{((1 + \lambda_2)C_1 + \lambda_3)N_{E=1}} \sum_{i:E_i=1} \left( \sum_{T_j \ni T_i} \frac{(x_i - x_j) \exp(x_j^T \hat{w}^{\text{max}})}{\sum_{T_j \ni T_i} \exp(x_j^T \hat{w}^{\text{max}})} \right). 
\]

(22)
Proof. By (7) and (8), we have
\[
\frac{\theta^T \nabla L(\hat{w}^{\text{max}}_k)}{L ||\hat{w}^{\text{max}}_k - \hat{w}||_2} = - \left( \hat{w}^{\text{max}}_k - \hat{w} \right)^T \frac{\hat{w}^{\text{max}}_k - \hat{w}}{L ||\hat{w}^{\text{max}}_k - \hat{w}||_2} \sum_{i, E_i = 1} I_d(k) x_i - \sum_{i, E_i = 1} x_i \exp(x_j^T \hat{w}^{\text{max}}_k) \right) \]

By Lemmas 2.1 and 2.3, we get
\[
\frac{\hat{w}^T (I_d - I_d(k))}{L ||\hat{w}^{\text{max}}_k - \hat{w}||_2} \sum_{i, E_i = 1} \left( x_i - \sum_{T_j \not \geq T_i} x_j \exp(x_j^T \hat{w}^{\text{max}}_k) \right) \]

By Theorems 2.1 and 2.2, it is observed that the proved error bound of \( ||\hat{w} - \hat{w}^{\text{max}}_k||^2 \) tight up to constant factors. Further, it implies that there is a close relationship between \( ||\hat{w} - \hat{w}^{\text{max}}_k||^2 \) and \( ||I_d(k) - I_d||_2 \). Based on this intuition, we have the following corollary:
Corollary 2.1. Let \( \sup_{w \in \mathbb{R}^{d \times 1}} \| w \|_2 \leq C_0 \), then we have

\[
\| \hat{w} - \hat{w}_{\text{max}}^k \|_2^2 \leq \frac{4C_0C_1 \sqrt{d-k}}{\max \{\lambda_2, \lambda_3\}}.
\] (25)

Proof. By applying the non-negative property of exponential function and the sub-multiplicativity, the proof follows. \( \square \)

From Corollary 2.1, it is clear that the error of \( \| \hat{w} - \hat{w}_{\text{max}}^k \|_2^2 \) can be reduced by increasing the regularization parameters \( \lambda_2 \) and \( \lambda_3 \) appropriately; meanwhile, it can also make the error smaller by increasing \( k \), which can be observed from Supplementary Figures C.6, D.8, and E.10, and they show a general upward trend with the increasement of \( k \). Despite the useful insights obtained from the bound in (25), it is noted that the upper bound in (25) might not be very tight because of the uniform upper bound for \( w \in \mathbb{R}^{d \times 1} \). When there is additional information on \( w \), such as sparsity or group-level sparsity, we may further improve the upper found in (25). We leave this as a line of future study while presenting a possible improvement below.

Potential for improvement. Next, we mathematically analyze that our algorithm has room for further improvement in some cases.

If we use the selected variables to re-train the Max_k part in the EXCEL-extended models, the performance might be improved. Such a fact is revealed by the following Theorem. For simplicity, let

\[
A(f) = -\frac{\lambda_0}{N_E=1} \sum_{i: \bar{E}_i=1} \left( f(W_1 x_i) - \log \sum_{T_j \ni T_i} \exp f(W_1 x_i) \right),
\]

and

\[
B(f) = -\frac{\lambda_2}{N_E=1} \sum_{i: \bar{E}_i=1} \left( f(W_{1 \text{max}}^k x_i) - \log \sum_{T_j \ni T_i} \exp f(W_{1 \text{max}}^k x_i) \right).
\]

Theorem 2.3. Let \( (f^*, W_1^*) \) be an optimal solution of (3). If \( \text{Pos}(W_1^* - (W_1^*)_{\text{max}}^k) > \)
0, then there exists \( f^{*'} \), such that

\[
B(f^{*'}) \leq B(f^{*}),
\]

where \( \text{Pos}(W) \) counts the number of positive values in \( W \).

**Proof.** Assume that, for any \( f^{*''} \),

\[
B(f^{*''}) > B(f^{*}).
\]

If \( f^{*} \) is the optimal solution of \( A \), noting that \( \text{Pos}(W^*_I - (W^*_I)^{\text{max}}) > 0 \), then there exists \( f^{*'''} \), such that

\[
B(f^{*'''}) \leq B(f^{*}).
\]

This contradicts our assumption. \( \square \)

### 3. Experimental results

Next, we apply EXCEL to two semi-synthetic datasets and three real-world survival datasets, including clinical and genetic data, with varying numbers of subjects and variables and portions of censored observations. We will demonstrate that our proposed EXCEL approach can identify critically important risk factors from thousands of variables potentially related to the survival time and, with only a small subset of identified variables, it can achieve prognostic performance comparable to or better than the original models with all variables.

**Semi-synthetic data.** We first benchmark our EXCEL approach by embedding it as a pluggable module into the standard CPH model with two semi-synthetic datasets. These datasets allow for straightforward visual inspection of selected features. Also, we compare the applications of our approach with two baselines, the standard CPH model and the more recent algorithm, BeSS. We will use
R function `bess.one()`, which aims to solve the best subset selection problem with a specified cardinality.

1) We use the pixels of handwritten digits 3 and 8 in USPS [29] as variables and generate synthetic survival times for them using the exponential distribution, following the way of [30, 31, 32] for constructing `survMNIST`. The resultant data is called SurvUSPS_{3vs8}.

2) For digits 3 and 8 in USPS, firstly, we resize the original 16 × 16 images into 28 × 28 images by extending each boundary by 6 pixels and filling in uniform noise. Then, we follow the way in 1) to assign survival times for them. The resultant data is called NSurvUSPS_{3vs8}.

On NSurvUSPS_{3vs8}, we compare the standard CPH model, BeSS for the best subset selection in CPH model, and Max_{k}-CPH. Randomly sampled images are illustrated in Figure 3 (a). It is found that many variables selected by the baseline CPH model are located in the noise boundary regions (see Figure 3 (b) and (c)), while those selected by BeSS and Max_{k}-CPH are salient points distinguishing the digits 3 and 8 (see Figure 3 (d)-(g)). These results indicate that BeSS and our proposed method can identify important variables and eliminate noise.

On SurvUSPS_{3vs8}, we randomly split all samples 10 times by a ratio of 80:20 for training and test sets, and then we use the three models to perform variable selection. The experimental results are demonstrated in Figure 3 (i)-(k). It is clear that the variables selected by Max_{k}-CPH essentially overlap for different splits, whereas those by the baselines do not overlap well. These results reveal that the features selected by our EXCEL method are more stable compared to those by the baselines.

Next, we will implement a variety of experiments on real-world datasets, and relevant statistics of these datasets are shown in Supplementary Table A.2.

---

4We randomly assigned each class label to a risk group, so that one digit would correspond to better survival and the other to worse survival. Then, we generate a risk score that indicates how risky it is to experience an event, relative to others. For more details, see the links of Semi-synthetic data in Section Data availability.
Figure 3: Experiments on two semi-synthetic datasets. (a)-(g) Noise elimination analysis of different methods with $k = 5$ and 10 on NSurvUSPS3vs8; (h) Illustration of random splitting for training and test sets; (i)-(k) Stability analysis of variable selection by different methods with 5 (left in each pair of images) and 10 (right in each pair of images) random splits for $k = 10$. 
Breast cancer dataset. We directly import the function `sksurv.datasets.load_breast_cancer()` to download this dataset. It contains the expression levels of 76 genes, age, estrogen receptor status (denoted by er), tumor size, and grade for 198 subjects. First, we perform one-hot encoding for categorical variables of er and grade; then, we apply CPH, DeepSurv, Cox-nnet, and their EXCEL-embedded counterparts to analyze this dataset. The results for \( k \) from 5 to 40 with a step size of 5 are shown in Supplementary Figure C.6(a)-(c). Using the number of selected variables corresponding to the optimal CI, we apply the EXCEL-extended models by setting \( k \) to the best numbers of selected variables to calculate IBS in Supplementary Figure C.6(e)-(g). It is observed that the EXCEL-extended survival models, with fractions of selected variables, can achieve comparable or better performance than their counterparts without EXCEL. Moreover, the EXCEL-extended models can identify a subset of critical variables for achieving maximal performance. We provide the top 10 variables selected by the best model, i.e., Max\(_{15}\)-DeepSurv, in Table 1.

Table 1: Top 10 variables selected by EXCEL-extended survival models on three datasets.

| Rank | Dataset       | Breast cancer | SUPPORT                     | GBM       |
|------|---------------|---------------|-----------------------------|-----------|
| 1    |               | X204540.at    | slo                                       | AGE       |
| 2    |               | X202240.at    | sfdm2.<2 mo. follow-up                  | HIST3H2A |
| 3    |               | X203306.s.at  | sex_female                             | PRODH    |
| 4    |               | X218883.s.at  | sex_male                               | CCL25    |
| 5    |               | er            | ca_metastatic                          | ATP10B   |
| 6    |               | X202239.at    | sfdm2_no (M2 and SIP pres)              | FZD7     |
| 7    |               | X202687.s.at  | sfdm2_adl>=4 (>5 if sur)               | TPM4     |
| 8    |               | X204014.at    | avtisst                                | NMB      |
| 9    |               | X208180.s.at  | dzgroup_Lung Cancer                     | RPS4Y1   |
| 10   |               | X207118.s.at  | sfdm2_SIP>=30                         | PACS1    |

For quantifying the significance of selected variables, we use \( k \)-means clustering to cluster the subjects into two groups according to the selected variables, and then we use the KM estimator to estimate the survival functions and the

---

5See the link: https://scikit-survival.readthedocs.io/en/latest/api/generated/sksurv.datasets.load_breast_cancer.html
Log-rank test to calculate p-values for the groups according to each of the individual variables. The top 10 variables identified by this clustering-then-test approach are as follows (the values in the parentheses are p-values): $X_{202240, at}$ (6.1E-05), $X_{202418, at}$ (2.3E-03), $X_{202687, s, at}$ (9.5E-03), $X_{203306, s, at}$ (5.8E-03), $X_{204014, at}$ (4.5E-03), $X_{204015, s, at}$ (9.8E-03), $X_{208180, s, at}$ (1.8E-03), $X_{211762, s, at}$ (3.8E-03), er (9.6E-03), and $X_{205034, at}$ (1.5E-02). In [31], two methods were proposed, C-based splitting and log-rank splitting, to select variables on this dataset. Four variables ($X_{202240, at}$, $X_{204014, at}$, $X_{204015, s, at}$, and er) selected by its former method and two variables ($X_{202240, at}$ and $X_{204014, at}$) by the latter are of statistical significance. In Table 1, it is seen that six of the top 10 variables identified by Max35-DeepSurv have statistical significance. We depict the KM survival curves for the top 3 variables identified by Max35-DeepSurv in Supplementary Figure C.7 (a)-(c) and the scatter plots with 95% confidence intervals for pairwise groups for different individual variables in Supplementary Figure C.7 (d)-(f). These figures indicate that the two groups obtained from each of the top 3 variables are well separated in survival times, suggesting these variables’ high predictive ability. Here, it is noted that the KM curve of Group 1 for the top feature in Supplementary Figure C.7 (a) plummets sharply after about time 7,000 (indicating the deaths of all subjects in this group) while Group 2 does not, which illustrates the close association of this feature with the patent difference between the 2 groups’ survival times. It is worth pointing out that, nonetheless, KM curves in Supplementary Figure C.7 (a) have a larger p-value than the two lower-ranked features in Supplementary Figure C.7 (b)-(c), which appears to be a limitation of the Log-rank test (that is, it can only indicate whether the 2 groups are of the same distribution but not the size of the difference [35]) yet a strength of EXCEL.

Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) dataset. To analyze models on real clinical data, we use a common benchmark dataset in the survival analysis SUPPORT [36]. This
dataset contains demographic, score data acquired from patients diagnosed with cancer, chronic obstructive pulmonary disease, cirrhosis, acute renal failure, multiple organ system failure, sepsis, and so on. We preprocess this dataset following [32] and obtain the resultant dataset with 9,105 subjects and 59 variables. Then we compare different models and show the corresponding results for \( k \) from 5 to 40 with a step size of 5 in Supplementary Figure D.8 (a)-(c), and the IBS by the EXCEL-embedded models with the best numbers of selected variables in Supplementary Figure D.8 (d)-(g). It is observed that EXCEL-extended models achieve comparable or better performance when a certain number of variables are selected, and the subset of top 10 critical variables for the best model, Max\(_{25}\)-Cox-nnet, is given in Table 1.

Also, we adopt \( k \)-means clustering to cluster the subjects into two groups according to each of the top 3 selected variables; then, on each resultant group, we perform the KM estimation and Log-rank test. The results are shown in Supplementary Figure D.9. Each of these variables clearly exhibits high predictive ability for survival time.

Glioblastoma multiforme (GBM) cancer dataset. Obtained from the Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov), this dataset contains gene expression and clinical data. We preprocess the data following [4] to obtain 5,567 genes, 860 pathways, and clinical data of ages from 522 subjects for our analysis. The pathway data are mainly used for Cox-PASnet [4] and its EXCEL-extended counterpart. For other models, we only use the combined data of gene variables and one clinical variable, AGE, as input without any pathway information. We compare different models with \( k \) ranging from 100 to 1000 with a step size of 100, showing the results in Supplementary Figure E.10 (a)-(c) and (g) and the corresponding IBS by the embedded models with the best numbers of selected variables in Supplementary Figure E.10 (d)-(f) and (h). It is observed that Max\(_{800}\)-DeepSurv achieves the best performance when 800 variables are selected, from which the top 10 critical variables are given in Table 1.

Further, we adopt \( k \)-means clustering to cluster the subjects into two groups...
according to each of the top 3 selected variables; then, we perform the KM estimation and Log-rank test for each obtained group. The results shown in Supplementary Figure E.11 indicate that each of these selected variables is well separable for the subjects and of high predictive ability for the survival time. More detailed results about CI and IBS for these datasets are provided in Supplementary Tables B.3 and B.4.

4. Discussion

Real-world survival data, such as gene expression, is often high-dimensional with a much larger number of covariates than the subjects. For computational feasibility and explainability of learning, we propose a plug-and-play approach that can be easily used with the existing DL-based survival models for variable selection.

Reducing the number of variables for learning. Structurally, our EXCEL approach uses the operator $\text{Max}_k$ to pinpoint $k$ most predictive variables, which introduces a sub-architecture into the architecture of existing models. The induced sub-architecture has a fitting error term similar to the existing models, and it is trained cooperatively with the architecture of existing models. By regarding the existing model as a regularization term, we observe that the resulting sub-model can select critical variables while capturing the complex nonlinear relationship between input variables and survival time. As empirically shown above and summarized in Figure 4, this concise approach for regularization is able to select the most predictive $k$ variables to enhance the learning ability of the survival model, achieving prediction performance on par with or better than strong baseline methods. By discarding noise or less informative variables and identifying useful variables of long-term prognostic values, our easily deployable approach can help facilitate explainability to the DL-based models and facilitate their prediction, especially on high-dimensional data.

explainability of model learning. Our EXCEL approach brings explainability to model learning. Taking the Breast cancer dataset as an example, the clinical
variable “er” is identified as one of the top 5 important variables; in the literature, the status of estrogen receptor has been confirmed to associate with the risk of breast cancer \cite{37, 38}. Also, take the GBM dataset as another example. GBM cancer is one of the most aggressive, malignant brain tumors with a poor prognosis \cite{39}. During learning, we use the clinical variable “AGE” together with more than 5,000 gene expressions as input. Our \texttt{Max800-DeepSurv} identifies 800 important variables out of 5,568 variables, with the most important 3 variables being AGE, \texttt{HIST3H2A}, and \texttt{PRODH}. In the literature, age was identified as a significant covariate for the prognosis of GBM \cite{40, 41}; \texttt{HIST3H2A} was reported as a differentially expressed gene in many studies, e.g., \cite{42, 40, 43}; \texttt{PRODH} was also found to be linked with GBM cancer in \cite{44}.

Figure 5: Survival curves by KM estimator with Log-rank tests for all selected variables by \texttt{Max25-DeepSurv}, \texttt{Max25-Cox-net}, and \texttt{Max800-DeepSurv}. The shaded areas in (a)-(c) show the 95% confidence interval.
Predictability of the selected subset of variables. For further analyzing the significance of selected variables, we use $k$-means clustering to cluster the subjects into two groups according to the subsets of variables selected by Max$_{35}$-DeepSurv, Max$_{25}$-Cox-nnet, and Max$_{800}$-DeepSurv, respectively; then we apply the KM estimator and Log-rank test to the groups, with the results shown in Figure 5. The p-values for Breast cancer and SUPPORT datasets are $9.6E-03$ and $2.5E-35$, indicating significant differences between the KM curves based on EXCEL-selected variables. For the dataset GBM, the number of the original variables is large, out of which 800 variables are selected as the most predictive for survival times. Subsequently, we use $k$-means clustering to cluster the subjects into four groups according to these variables. The p-values of two pairwise groups are significant, i.e., $7.9E-03$ for group 1 vs. group 2 and $1.9E-02$ for group 1 vs. group 4, each of which indicates a significant difference between the KM curves for these groups.

5. Conclusion

While crucial to explainability, it is challenging to reveal the associations between highly nonlinear, high-dimensional, and low-sample size data, such as genetic and clinical dataset, and the survival time in survival analysis. Nonlinear survival methods, including DL-based models, can well model high-level interactions; nonetheless, they typically work as a black-box, lacking interpretation and needed explainability. Thus, transparent learning and accurate prediction of survival time based on patients’ critical variables are urgent, unmet needs. A solution can help identify new intervention targets and potentially improve patient care and treatment practice. In this paper, we are interested in understanding how the survival models yield predictions on right censored data. To this end, we propose a novel approach, EXCEL, to identify critical variables of long term prognostic values and simultaneously implement deep model training based on these identified variables. In a wide variety of experiments, including two semi-synthetic datasets, one toy dataset, and three clinical and/or genetic...
datasets, we demonstrate the effectiveness of our proposed explainable censored learning or finding critical features and making survival predictions.

This paper focuses on a practical algorithm that can plug and play on existing DL-based survival models, and we will further study its theoretical properties in our future work.

Funding

This work was partially supported by the NIH grants R21AG070909, R56NS117587, R01HD101508, P30 AG072496, and ARO W911NF-17-1-0040.

Appendix A. Datasets

| Dataset     | #Sample | #Variable | %Censored |
|-------------|---------|-----------|-----------|
| Breast cancer | 198     | 80        | 74.24     |
| SUPPORT     | 9,105   | 59        | 31.89     |
| GBM         | 522     | 5,568     | 14.37     |
Appendix B. Appendix

Appendix B.1. CI of Different Methods

Table B.3: CI of compared methods.

| Dataset Method | Breast cancer | SUPPORT | GBM          |
|----------------|---------------|---------|--------------|
| CPH            | 0.653±0.072   | 0.842±0.004 | 0.614±0.021 |
| Max_k-CPH      | k:40, 0.638±0.077 | k:35, 0.842±0.004 | k:300, 0.623±0.015 |
| DeepSurv       | 0.672±0.065   | 0.828±0.003 | 0.625±0.018  |
| Max_k-DeepSurv | k:35, 0.679±0.072 | k:30, 0.830±0.003 | k:800, 0.633±0.021 |
| Cox-nnet       | 0.672±0.066   | 0.871±0.005 | 0.526±0.043  |
| Max_k-Cox-nnet | k:15, 0.678±0.057 | k:25, 0.876±0.005 | k:800, 0.534±0.049 |
| PASNet         | \             | \       | 0.574±0.054  |
| Max_k-PASNet   | \             | \       | k:100, 0.580±0.019 |
## Appendix B.2. IBS of Different Methods

Table B.4: IBS of compared methods.

| Method            | Dataset       | Breast cancer | SUPPORT | GBM     |
|-------------------|---------------|---------------|---------|---------|
| CPH               | 0.174±0.021   | 0.260±0.006   | 0.392±0.028 |
| Max$_k$-CPH       | 0.176±0.023   | 0.260±0.006   | 0.359±0.023 |
| DeepSurv          | 0.174±0.017   | 0.266±0.004   | 0.337±0.019 |
| Max$_k$-DeepSurv  | 0.173±0.018   | 0.265±0.005   | 0.340±0.020 |
| Cox-nnet          | 0.174±0.018   | 0.269±0.007   | 0.346±0.019 |
| Max$_k$-Cox-nnet  | 0.174±0.021   | 0.264±0.007   | 0.346±0.019 |
| PASNet            | \             | \             | 0.351±0.018 |
| Max$_k$-PASNet    | \             | \             | 0.362±0.023 |
Appendix C. Different Methods on Breast Cancer

Figure C.6: Prediction performance, in CI (higher, better) and IBS (lower, better), of different methods on breast cancer dataset. The shaded areas in (a)-(c) show the fluctuations in standard error; the orange lines in (d)-(f) show the medians. For (a)-(c), the horizontal axis is only meaningful for the EXCEL-extended models.

Figure C.7: KM survival curves and pairwise groups for the top 3 variables selected by Max45-DeepSurv on the breast cancer dataset. The shaded areas in (a)-(c) show the 95% confidence intervals; the orange lines in (d)-(f) show the medians.
Appendix D. Different Methods on SUPPORT

Figure D.8: Prediction performance, in CI (higher, better) and IBS (lower, better), of different methods on SUPPORT. The shaded areas in (a)-(c) show the fluctuations in standard error; the orange lines in (d)-(f) show the medians. For (a)-(c), the horizontal axis is only meaningful for the EXCEL-extended models.

(a) CPH vs. Max\(_k\)-CPH
(b) Cox-nnet vs. Max\(_k\)-Cox-nnet
(c) DeepSurv vs. Max\(_k\)-DeepSurv
(d) CPH vs. Max\(_{35}\)-CPH
(e) Cox-nnet vs. Max\(_{25}\)-Cox-nnet
(f) DeepSurv vs. Max\(_{30}\)-DeepSurv

Figure D.9: KM survival curves and pairwise groups for the top 3 variables selected by Max\(_{25}\)-Cox-nnet on SUPPORT. The shaded areas in (a)-(c) show the 95% confidence intervals; the orange lines in (d)-(f) show the medians.

(a) KM curves for “slos”
(b) KM curves for “sfdm2<2 mo. follow-up”
(c) KM curves for “sex_female”
(d) Two groups of “slos”
(e) Two groups of “sfdm2<2 mo. follow-up”
(f) Two groups of “sex_female”
Appendix E. Different Methods on GBM

Figure E.10: Prediction performance, in CI (higher, better) and IBS (lower, better), of different methods on GBM. The shaded areas in (a)-(c) and (g) show the fluctuations in standard error; the orange lines in (d)-(f) and (h) show the medians. For (a)-(c) and (g), the horizontal axis is only meaningful for the EXCEL-extended models.
Figure E.11: KM survival curves and pairwise groups for the top 3 variables selected by Max800-DeepSurv on GBM. The shaded areas in (a)-(c) show the 95% confidence intervals; the orange lines in (d)-(f) show the medians.
Appendix F. Details of Datasets WHAS500 and SUPPORT

| Variable | Description                         | Codes / Values          |
|----------|-------------------------------------|-------------------------|
| afb      | Atrial Fibrillation                 | 0 = No, 1 = Yes         |
| age      | Age at Hospital Admission           | Years                   |
| avd      | Complete Heart Block                | 0 = No, 1 = Yes         |
| bmi      | Body Mass Index                     | kg/m²                   |
| chf      | Congestive Heart Complications      | 0 = No, 1 = Yes         |
| cvd      | History of Cardiovascular Disease   | 0 = No, 1 = Yes         |
| diasbp   | Initial Diastolic Blood Pressure    | mmHg                    |
| gender   | Gender                              | 0 = Male, 1 = Female    |
| hr       | Initial Heart Rate                  | Beats per minute        |
| los      | Length of Hospital Stay             | Days from Hospital Adm   |
| miord    | MI Order                            | 0 = First, 1 = Recurrent|
| mitype   | MI Type                             | 0 = non Q-wave, 1 = Q-wave|
| sho      | Cardiogenic Shock                   | 0 = No, 1 = Yes         |
| sysbp    | Initial Systolic Blood Pressure     | mmHg                    |
| Variable               | Description                                                                 | Codes / Values                                                                 |
|------------------------|-----------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| slos                   | Days from Study Entry to Discharge                                          | Days                                                                            |
| sfm2_<2 mo. follow-up  | Patient died before 2 months after study entry                              | 1 = 2 mo. follow-up                                                           |
| sex_female             | \                                                                            | 1 = Female                                                                      |
| sex_male               | \                                                                            | 1 = male                                                                        |
| ca_metastatic          | Cancer status                                                               | 1= metastatic                                                                  |
| sfm2_no (M2 and SIP pres) | Patient lived 2 months to be able to get 2 month interview, and from this interview there were no signs of moderate to severe functional disability | 1=no (M2 and SIP pres)                                                        |
| sfm2_adl>=4 (>5 if sur) | Patient was unable to do 4 or more activities of daily living at month 2 after study entry. If the patient was not interviewed but the patient’s surrogate was, the cutoff for disability was ADL | 1= adl>=4 (>5 if sur)                                                          |
| avtisst                | Average TISS, Days 3-25                                                      | Days                                                                            |
| dzgroup_Lung Cancer    | The current group is lung cancer                                             | 1=Lung Cancer                                                                   |
| sfm2_SIP>=30           | Sickness Impact Profile total score at 2 months                              | 1=SIP>=30                                                                      |

Here, for more details of SUPPORT, it can be found in the links [https://hbiostat.org/data/repo/Csupport2.html](https://hbiostat.org/data/repo/Csupport2.html) and [https://biostat.app.vumc.org/wiki/Main/SupportDesc](https://biostat.app.vumc.org/wiki/Main/SupportDesc)
References

[1] D. R. Cox, Regression models and life-tables, Journal of the Royal Statistical Society: Series B (Methodological) 34 (2) (1972) 187–202.

[2] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger, DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Medical Research Methodology 18 (1) (2018) 1–12.

[3] T. Ching, X. Zhu, L. X. Garmire, Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data, PLoS computational biology 14 (4) (2018) e1006076.

[4] J. Hao, Y. Kim, T.-K. Kim, M. Kang, PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data, BMC Bioinformatics 19 (1) (2018) 1–13.

[5] D. Gunning, D. W. Aha, DARPA’s explainable artificial intelligence (XAI) program, AI Magazine 40 (2) (2019) 44–58.

[6] D. W. Hosmer, S. Lemeshow, S. May, Applied Survival Analysis: Regression Modeling of Time-to-Event Data, 2nd Edition, 2008.

[7] J. Fan, Y. Feng, Y. Wu, High-dimensional variable selection for Cox’s proportional hazards model, Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown 6 (2010) 70–86.

[8] E. L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, Journal of the American Statistical Association 53 (282) (1958) 457–481.

[9] N. Mantel, Evaluation of survival data and two new rank order statistics arising in its consideration, Cancer Chemotherapy Reports 50 (3) (1966) 163–170.
[10] R. Peto, J. Peto, Asymptotically efficient rank invariant test procedures, Journal of the Royal Statistical Society. Series A (General) 135 (2) (1972) 185–207.

[11] B. K. Natarajan, Sparse approximate solutions to linear systems, SIAM Journal on Computing 24 (2) (1995) 227–234.

[12] J. Weston, A. Elisseeff, B. Schr"{e}lkopf, M. Tipping, Use of the zero norm with linear models and kernel methods, The Journal of Machine Learning Research 3 (2003) 1439–1461.

[13] Y. Hamo, S. Markovitch, The COMPSET algorithm for subset selection, in: IJCAI, 2005.

[14] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267–288.

[15] S. S. Chen, D. L. Donoho, M. A. Saunders, Atomic decomposition for basis pursuit, Society for Industrial and Applied Mathematics Journal on Scientific Computing 20 (1) (1998) 33–61.

[16] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American statistical Association 96 (456) (2001) 1348–1360.

[17] J. Fan, W. Gong, Q. Sun, A provable two-stage algorithm for penalized hazards regression, arXiv preprint arXiv:2107.02730v1.

[18] C. Wen, A. Zhang, S. Quan, X. Wang, BeSS: An R package for best subset selection in linear, logistic and Cox proportional hazards models, Journal of Statistical Software 94 (4) (2020) 1–24.

[19] X. Wu, Q. Cheng, Algorithmic stability and generalization of an unsupervised feature selection algorithm, in: NeurIPS, 2021.
[20] X. Wu, Q. Cheng, Fractal autoencoders for feature selection, in: AAAI, 2021.

[21] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, R. A. Rosati, Evaluating the yield of medical tests, Journal of the American Medical Association 247 (18) (1982) 2543–2546.

[22] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, V. C. Raykar, On ranking in survival analysis: Bounds on the concordance index, in: NIPS, 2007.

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

[24] G. W. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Review 78 (1) (1950) 1–3.

[25] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feed-forward neural networks, in: AISTATS, 2010.

[26] D. P. Kingma, J. L. Ba, Adam: A method for stochastic optimization, in: ICLR, 2015.

[27] Probability and Stochastics, 1st Edition, Erhan Çinlar, 2011.

[28] Y. Nesterov, Lectures on Convex Optimization, 2nd Edition, 2018.

[29] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, Feature selection: a data perspective, ACM Computing Surveys 50 (6) (2017) 1–45.

[30] S. Pölsterl, Survival analysis for deep learning, in: https://k-d-w.org/blog/2019/07/survival-analysis-for-deep-learning/, 2019.

[31] M. Goldstein, X. Han, A. Puli, A. J. Perotte, R. Ranganath, X-CAL: Explicit calibration for survival analysis, in: NeurIPS, 2020.
[32] L. Manduchi, R. Marcinkevičs, M. C. Massi, T. Weikert, A. Sauter, V. Gotta, T. Müller, F. Vasella, M. C. Neidert, M. Pfister, B. Stieltjes, J. E. Vogt, A deep variational approach to clustering survival data, in: ICLR, 2022.

[33] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d’Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse, C. Sotiriou, on behalf of the TRANSBIG Consortium, Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series, Clinical Cancer Research 13 (11) (2007) 3207–3214.

[34] M. Schmid, M. N. Wright, A. Ziegler, On the use of Harrell’s C for clinical risk prediction via random survival forests, Expert Systems with Applications 63 (2016) 450–459.

[35] J. M. Bland, D. G. Altman, The logrank test, British Medical Journal 328 (7447) (2004) 1073.

[36] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, P. Layde, R. K. Oye, P. E. Bellamy, R. B. Hakim, D. P. Wagner, The SUPPORT prognostic model: Objective estimates of survival for seriously Ill hospitalized adults, Annals of Internal Medicine 122 (3) (1995) 191–203.

[37] C.-H. Yip, A. Rhodes, Estrogen and progesterone receptors in breast cancer, Future oncology 10 (14) (2014) 2293–2301.

[38] A. M. Farcas, S. Nagarajan, S. Cosulich, J. S. Carroll, Genome-wide estrogen receptor activity in breast cancer, Endocrinology 16 (2) (2021) bqaa224.

[39] F. Hanif, K. Muzaffar, K. Perveen, S. M. Malhi, S. U. Simjee, Glioblastoma multiforme: A review of its epidemiology and pathogenesis through clinical
presentation and treatment, Asian Pacific Journal of Cancer Prevention 18 (1) (2017) 3–9.

[40] S. Bozdag, A. Li, G. Riddick, Y. Kotliarov, M. Baysan, F. M. Iwamoto, M. C. Cam, S. Kotliarova, H. A. Fine, Age-specific signatures of glioblastoma at the genomic, genetic, and epigenetic levels, PLOS ONE 8 (4) (2013) e62982.

[41] J. Lu, M. C. Cowperthwaite, M. G. Burnett, M. Shpak, Molecular predictors of long-term survival in glioblastoma multiforme patients, PLOS ONE 11 (4) (2016) e0154313.

[42] A. Yoshino, A. Ogino, K. Yachi, T. Ohta, T. Fukushima, T. Watanabe, Y. Katayama, Y. Okamoto, N. Naruse, E. Sano, K. Tsumoto, Gene expression profiling predicts response to temozolomide in malignant gliomas, International Journal of Oncology 36 (6) (2010) 1367–1377.

[43] N. K. Gerber, A. Goenka, S. Turcan, M. Reyngold, V. Makarov, K. Kannan, K. Beal, A. Omuro, Y. Yamada, P. Gutin, C. W. Brennan, J. T. Huse, T. A. Chan, Transcriptional diversity of long-term glioblastoma survivors, Neuro-Oncology 16 (9) (2014) 1186–1195.

[44] E. H. Panosyan, H. J. Lin, J. Koster, J. L. L. III, In search of druggable targets for GBM amino acid metabolism, BMC Cancer 17 (1) (2017) 1–12.