IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 13, 2021, accepted December 28, 2021, date of publication January 11, 2022, date of current version January 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3142032

Regularizing the Deepsurv Network Using Projection Loss for Medical Risk Assessment

PHAWIS THAMMASORN[1], STEPHANIE K. SCHAUB[2], DANIEL S. HIPPE[3], MATTHEW B. SPRAKER[4], JAN C. PEEKEN[5], LANDON S. WOOTTON[2], PAUL E. KINAHAN[2,6], (Fellow, IEEE), STEPHANIE E. COMBS[5], WANPRACHA A. CHAOVALITWONGSE[1], AND MATTHEW J. NYFLOT[2]

[1]Department of Industrial Engineering, University of Arkansas, Fayetteville, AR 72701, USA
[2]Department of Radiation Oncology, University of Washington, Seattle, WA 98195, USA
[3]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
[4]Department of Radiation Oncology, Washington University, St. Louis, MO 63130, USA
[5]Department of Radiation Oncology, Klinikum rechts der Isar, Technical University of Munich (TUM), 81675 Munich, Germany
[6]Department of Radiology, University of Washington, Seattle, WA 98195, USA

Corresponding author: Phawis Thammasorn (pthammas@uark.edu)

This work was supported in part by the National Science Foundation (NSF) under Award 1734913, and in part by the Biostatistics Shared Resource of the Fred Hutch/University of Washington Cancer Consortium under Grant P30 CA015704.

ABSTRACT State-of-the-art deep survival prediction approaches expand network parameters to accommodate performance over a fine discretization of output time. For medical applications where data are limited, the regression-based Deepsurv approach is more advantageous because its continuous output design limits unnecessary network parameters. Despite the practical advantage, the typical network lacks control over the feature distribution causing the network to be more prone to noisy information and occasional poor prediction performance. We propose a novel projection loss as a regularizing objective to improve the time-to-event Deepsurv model. The loss formulation maximizes the lower bound of the multiple-correlation coefficient between the network's features and the desired hazard value. Reducing the loss also theoretically lowers the upper bound on the likelihood of discordant pair and improves C-index performance. We observe superior performances and robustness of regularized Deepsurv over many state-of-the-art approaches in our experiments with five public medical datasets and two cross-cohort validation tasks.

INDEX TERMS Machine learning, pattern recognition, supervised learning, medical expert systems, biomedical computing.

I. INTRODUCTION

Survival analysis, also known as time-to-event analysis, is a crucial task in medical prognosis and risk assessment. Fundamentally, the objective is to create a predictive function mapping from relevant input data to time until event occurrence, or "hitting time." A wide range of medical applications is enabled once the hitting times are predicted [14]. Unlike the typical regression tasks, the model has to account for right-censored outcomes (e.g., unknown time-to-death due to patients dropping out of studies.) The problem has been widely investigated in the Statistics community [18]. The typical approaches are developed from regression models with the additional consideration of outcome censoring. For example, Cox proportional hazard model (CPH) [5] and Survival Support Vector Machine (SVM) [20] are formulated from linear-based regressors, which allow applications of classical regularization and analysis, such as L-1 and L-2 norm, into the learning. Survival Tree [21] and Survival Forest [22] are early attempts to utilize non-linear regressors for the survival task.

Recent approaches utilize deep learning architectures for the prediction ("deep survival"). The introduction of deep survival networks not only opens opportunities for developing non-linear prediction models but also allows prediction using other data forms instead of handcrafted features, such as medical 3D-MRI scans, and sequence data for survival tasks [3], [7], [19]. Without the tedious work of defining relevant features, recent emphases of many works have shifted toward identifying relevant data and developments of

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

suitable deep survival network architectures. Our study focuses on the development of deep survival approaches that have previously adapted to data forms without changing objective functions or introducing drastic preprocessing steps.

A prominent trend in developing deep survival approaches is discretizing possible survival outcomes for classification architecture in modeling hitting time distribution. The strategy trains deep networks to classify or maximize likelihoods corresponding to pre-defined intervals of discretized hitting time. Variations of the networks are differently characterized based on assumptions and outcomes about censored data. For example, partial-logistic regression [1] and NNet-survival [6] are feed-forward classification networks trained by increasing likelihood over output nodes corresponding to the period after censoring with the assumption that the unobserved event is likely to occur after the censored time point. The survival probability mass function (PMF) network [10] is trained to output a discretized cumulative density function of survival time instead of the likelihood of occurrence. DeepHit [12] is the state-of-the-art method that combines the discretized density training similar to that of the PMF with a multi-task network for the prediction. Notice that these discrete-time network can adapt to new input forms by simply changing the input layer architecture (e.g., Convolutional Neural Network [3], [40].) The drawback of this approach is the requirement of hitting time discretization, which inconveniently introduces a trade-off between the number of parameters and the granularity of output time. Specifically, fine-grain discretization helps the network distinguish cases with slight differences in hitting time at the price of more network parameters. Optimally discretizing the data for the deep-learning approach is non-trivial and requires consensus between engineers and medical experts. On the broader picture, the medical context often involves the lack of gold-standard outcomes and several difficulties in data acquisition, including privacy laws, lack of patient enrollment on research studies, and low frequency of events. Training a large-scale neural network under these data circumstances often causes overfitting and poor prediction performance. Unlike some other domains, solving the insufficient data using data augmentation is also limited as the learning from invalid data introduces risk in subsequent medical practice. The discretization leads to more difficulties in network design and training.

In Contrast to the discretization strategy, regression-based networks avoid unnecessary difficulties by modeling the hitting time density with continuous hazard output. By mapping the input to a single-dimension hazard value instead of a discrete outcome, the architecture development allows more flexible control over the number of parameters regardless of the survival outcome range, which is favorable in the small-data medical context. Variations of the regression networks are defined on how the output values relate to survival likelihood. Deepsurv [9] is a well-established network under this approach with the assumption of proportional hazards such that the difference between survival likelihood for a given time is proportional to the difference in feature or hazard values. The model is developed with intuitions behind the Cox proportional hazard (CPH) model, a widely applied statistical method for survival analysis. The Cox-time network [11] improves this approach by dropping the CPH's proportionality assumption. Many survival prediction works adapted the regression approach for various health-related data [7], [19]. There are also hybrid regression approaches, such as Survival-net [3], that train the network to classify censored and uncensored data before using the features from penultimate layers for survival regressions. However, the regression-based and the hybrid networks have inferior performance to many discretized networks, as reported in [6], [10], and [12].

Inspired by the advantages offered by the deep survival regression approach, we investigate the lack of feature distribution control during the regular Deepsurv training that exposes the network to suboptimal performance. Our contributions in this work are three-fold.

- Theoretical foundation to enhance network on the perspective of representation learning without scaling the number of network parameters.
- The projection loss regularization based on the theory. The regularization loss is applicable to survival prediction on various input forms implying that the proposed improvements are applicable to the other networks with similar organizations to Deepsurv, such as DeepConvSurv [19] networks, which we used in our experiments.
- Demonstration of the improvement generality through experiments on many expert-defined medical prognosis datasets derived from clinical outcomes, histology, cellularity, immune markers, and volumetric imaging.

Unlike many previous deep survivals works, we present thorough qualitative and quantitative investigations covering both theorems and empirical experiments. We also set up experiments to mimic cross-cohort validations, a crucial practice in developing practical medical decision support systems. Our experiments also show that the regularized Deepsurv outperform discrete-time state-of-the-art baselines in the small-data medical context.

It is noteworthy that the discrete-time and regression-based survival approaches focused in this study are not the entire spectrum of deep learning techniques for survival analysis. We consider approaches under a similar medical context and experimented with small-sample survival datasets. To consider applicable approaches to various data forms, we exclude networks that are difficult to adjust to new forms of data (e.g., focusing on specific modality [33] or specific architecture [34], and relying on time-varying information [35]) because there are increasingly diverse types of input for survival prediction. Because medical data are difficult to model, we also avoid approaches that attempt to model the distribution of data or distribution of outcome or outcome censoring (e.g., sampling from generative modeling [36], Perturbation [37], assuming distributions of outcome [38], and training with artificial target responses [39]) because data modeling and augmentation are difficult to validate and

introduce risks of learning from invalid data. Deepsurv and our baselines under our investigations fit all these criteria. We discuss the advantages of Deepsurv over the other baselines in further sections.

II. PARAMETRIC SURVIVAL REGRESSION

Let $\{(I_1, \delta_1, t_1), (I_2, \delta_2, t_2), \ldots, (I_N, \delta_N, t_N)\}$ be the dataset for survival analysis where I_i is input data for case i, t_i is the latest elapsed time recorded for the case, and N is the total number of data cases. δ_i indicates whether the outcome event of case i is observed where $\delta_i = 0$ means that the case outcome is censored after t_i, and $\delta_i = 1$ means that the event takes place at time t_i. The general goal is to model survival function $S(t) = P(T \geq t)$ where $S(t)$ value is the probability of hitting time happening after time t. The model can be parameterized such that adjusting parameters gives flexibility in learning to predict new cases.

A. COX PROPORTIONAL HAZARD REGRESSION

One of the earliest methods for the parametric approach is Cox proportional hazard regression. Cox proportional hazard (CPH) model define $S(t)$ as

$$S(t) = \exp(-H_0(t)\exp(X \cdot \beta)), \tag{1}$$

where X is a vector of input covariates or features extracted from I, and β is the vector of CPH regression parameters. $h_0(t)$ and $H_0(t)$ are baseline hazard function and cumulative baseline hazard function. $h_0(t)$ is often constructed using the Breslow estimator [13]. The cumulative hazard function is defined as

$$H_0(t) = \int_0^t h_0(u)du \tag{2}$$

Along with $S(t)$, the method also defines hazard function $h(t, X)$ as

$$h(t, X) = h_0(t) \exp(X \cdot \beta) \tag{3}$$

Notice that taking a ratio between $h(t, X_1)$ and $h(t, X_2)$ cancels out $h_0(t)$, resulting in $exp((X_i - X_j) \cdot \beta)$. This built-in model property without $h_0(t)$ is referred to as the proportional hazard assumption.

Without hitting time censoring, β can be obtained with a least-squares approach explored in [8] by rearranging the target variable as a linear function with an $X \cdot \beta$ term. With the censoring, however, the regression exploits the proportional hazard assumption and be solved by minimizing the negative log-partial-likelihood (NLPL) loss objective formed by all comparable pairs of X_i and X_j. The NLPL loss is defined as

$$\mathcal{L}_{NLPL}(\beta) = -\sum_{i=1}^{N} \delta_i \left\{ X_i \cdot \beta - \log\left(\sum_{j \in \mathcal{R}_i} exp\left(X_j \cdot \beta\right) \right) \right\}, \tag{4}$$

where $\mathcal{R}_j = \left\{\forall i, t_i > t_j\right\}$ is a set of data case indices that are still at risk at time t_j. Even though the loss itself is convex, it has no lower bound and causes $\|\beta\|$ to be unnecessarily large or difficult to obtain after optimization runs, especially when the amount of data is less than that of parameters. Therefore, the training often reduces the loss with norm-based regularizations, such as Ridge, Lasso, or their variations and combinations [23], to avoid the degenerate solution and to select suitable features for improving performance.

Regardless of regularization methods, $\hat{\beta}$ is a β estimate obtain from minimizing the loss. The loss prefers that the score term $X_i \cdot \hat{\beta}$ should be higher than $X_j \cdot \hat{\beta}$ corresponding to an event at the time later than t_i. Minimizing the loss is the attempt to make all pairwise comparisons of $X \cdot \hat{\beta}$ as concordant with the risk ordering as possible. Specifically, the goal is to make $X_i \cdot \hat{\beta} > X_j \cdot \hat{\beta}$ if $t_i < t_j$ when case j is not censored. $X \cdot \hat{\beta}$ is often referred to as hazard score or risk value and used instead of $h(t, X)$ for performance evaluation. One performance measure of survival regression is the concordance index [17] or C-Index. For CPH, the measure can be calculated using $X \cdot \hat{\beta}$ as

$$C = \frac{1}{P} \sum_{i:\delta_i=1} \sum_{j:t_j>t_i} \mathbb{I}\left[X_i \cdot \hat{\beta} > X_j \cdot \hat{\beta}\right], \tag{5}$$

where P is the total number of comparable pairs in the dataset. $i : \delta_i = 1$ represents any index i belongs to the set where the $\delta_i = 1$ or not censored. $j : t_j > t_i$ means any index j belongs to the set where recorded time $t_j > t_i$. $\mathbb{I}$ is an indicator function. The measure gauges the proportion of pairs with concordance between the event times and hazard scores, (i.e., the higher hazard score corresponds to the shorter event time within the pair.) Notice that the predicted hazard score $X \cdot \beta$ is subject to uncertainty and error. C value closer to 1.0 means less error such that the predicted score mostly following the expected ordering.

B. DEEPSURV NETWORK

Deepsurv network uses a similar formulation for $S(t)$ as the CPH model. However, the approach replaces $X \cdot \beta$ with the output from the deep neural network. Specifically, the Deepsurv define the hazard score or risk score value as $g(I|\theta, \beta_{net}) = f(I|\theta) \cdot \beta_{net}$ or $X_{net} \cdot \beta_{net}$ and the loss in Eq. (4) become

$$\begin{aligned} &\mathcal{L}_{NLPL}(\theta, \beta_{net}) \\ &= -\sum_{i=1}^{N} \delta_i \left\{ g(I_i|\theta, \beta_{net}) - \log\left(\sum_{j \in R_i} exp\left(g\left(I_j|\theta, \beta_{net}\right)\right) \right) \right\} \\ &= -\sum_{i=1}^{N} \delta_i \left\{ f(I_i|\theta) \cdot \beta_{net} - \log\left(\sum_{j \in R_i} exp\left(f\left(I_j|\theta\right) \cdot \beta_{net}\right) \right) \right\} \end{aligned} \tag{6}$$

where θ is a parameter set for Deepsurv network until the penultimate layer $f(I|\theta)$, and β_{net} is the last layer's parameter of the network. Notice that the output $f(I|\theta)$ is a feature vector from the network. Its linear combination with β_{net} is the risk score $g(I|\theta, \beta_{net})$. From the perspective of deep representation learning, the network simultaneously digests a new set

of covariates $X_{net} = f(I|\theta)$ and regression parameters β_{net} as they are tuned in an end-to-end manner with a gradient-based backpropagation. Adjusting $f(I|\theta)$ as opposed to a manually defined X gives flexibility in lowering NLPL.

Despite the network being relatively dated since its first conception compared to other deep learning alternatives, the major advantage of the approach is its ability to cope with wide ranges of outcome value without increasing network parameters while allowing flexibility of adjusting the neural network architectures with modern mechanisms (e.g., batch-normalization) to improve prediction performance [27]. Moreover, network variations can be created by changing to accommodate other forms of input data, such as a convolutional neural network for images input [19], and a recurrent neural network or similar mechanism for time-dependent data [32]. With the benefits, Deepsurv remains widely employed recently to make survival predictions under small-sample medical and health-related contexts [28]–[31].

III. PROPOSED METHOD

We propose applying regularization to NLPL loss for the Deepsurv network. The regularization is designed to induce a correlation between features from the Deepsurv network and the desired hazard value trend as well as to reduce the probability of lower C-Index values caused by noise from the network's features. This section will first discuss the motivation and the quality measure that reveals the drawback of training the Deepsurv network under NLPL loss. Afterward, details on the regularization as well as its theoretical support are provided.

A. IMPROVING QUALITY MEASUREMENT OF DEEPSURV FEATURE

Our regularization objective is to ensure that the Deepsurv learns suitable feature distribution for $X_{net} = f(I|\theta)$ during the optimization of pairwise concordance. Intuitively, a good set of features lead to well-perform β estimation and a higher C-Index. For the typical CPH model, poor feature quality is the most likely cause of underperformance because optimizing β is simpler with the convex loss. For Deepsurv, however, the loss function provides no explicit goal on improving the feature. Consequently, it is unclear during training whether X_{net}, β, or the other aspects of the network architecture cause poor performance in models with inferior C-index. To clarify such questions, we seek to include the suitability of X_{net} into the objective function to ensure that the best β for X_{net} is easier to obtain, and efforts could be put elsewhere to improve the performance.

Despite the Deepsurv network being a non-linear regression approach, the risk score output $X_{net} \cdot \beta_{net}$ at the last layer is linear in nature. For linear predictions in general, the multiple-correlation coefficient R is a suitability measure for the feature. However, the typical calculation of R requires knowing the exact values of the target prediction variable, which are not available in our case due to censoring. We propose an alternative measure Λ as the substitute measure R for the feature quality goal. Λ is defined as

$$\Lambda = \sqrt{\frac{Var\left(X_{net} \cdot \vec{\beta}_{net}\right)}{Var\left(X_{net}\right)}} \tag{7}$$

$\vec{\beta}_{net}$ is a directional unit vector calculated from normalizing the β_{net} vector of parameters. *Var* is variance calculated from random variables in the parenthesis. We further simplify the calculation in Eq. (7) with principal component analysis (PCA), resulting in the following

$$\Lambda = \sqrt{\frac{\sum_{i=1}^{p} \lambda_i \left(\vec{W}_i \cdot \vec{\beta}_{net}\right)_+}{\sum_{i=1}^{p} \lambda_i}} \tag{8}$$

where each $\vec{W}_i$ and λ_i are eigen vector of the principal components (PCs) and its corresponding eigenvalue. The PCs are the result of applying PCA to the interested data of which p is the total number of dimensions. $\vec{W}_i$ is a directional unit vector calculated from normalizing W_i. The numerator is the summation of variance captured in the direction of $\vec{\beta}_{net}$ proportional to the directional similarity between the PCs and β_{net}. The operator $(\_)_+$ is the element-wise absolute value to prevent any negative cosine value. The measure is bounded in the range of [0, 1]. According to Eq. (7), the Λ value close to 0 means the network captures only a minor fraction of the total data variance for the prediction, which could entail more noises and ungeneralizable data trends. On the other hand, the value close to 1 implies utilizing a significant part of the variance for the prediction. Even though the fraction of variance does not directly lead to poor C-index performance, it semantically hints at the magnitude of feature information that supports $\vec{\beta}_{net}$ and the efficiency of the prediction.

The purpose of Λ is two-fold. The first purpose is to establish that maximizing Λ increases the multiple-correlation coefficient and decreases the discordance probability. We provide theoretical discussions on the derivation of measure Λ in Section III-C. In brief summary, the theory reveals that $\Lambda \propto \lfloor R \rfloor$ and $\Lambda^2 \propto \frac{1}{\lceil \mathrm{P(disc)} \rceil}$ where $\lfloor R \rfloor$ is the lower bound on the multiple-correlation coefficient and $\lceil \mathrm{P(disc)} \rceil$ is the upper bound on discordant pairs occurring probability. These properties are the foundation of our loss formulation. The second purpose is to gauge relative changes in the relevant feature distribution across many regularization strategies. Even though Λ is relative to a lower bound measure and does not imply poor prediction quality, it is calculated from distributional variance measures, directly reflecting the feature distribution changes. It is surmisable that different regularization strategies alter the network's features, although the effect is unclear and difficult to gauge through C-index performance. Thus, we use Λ to observe whether there is feature training emphasis among regularization strategies and to inspect whether the feature change improves C-index performances.

B. PROJECTION LOSS REGULARIZATION

To increase $\mathbf{\Lambda}$, we propose the following regularization term in addition to the NLPL loss

$$\begin{aligned}&\mathcal{L}_{NLPL\_reg}(\theta, \beta_{net})\\&= \mathcal{L}_{NLPL}(\theta, \beta_{net})\\&\quad + w_{reg}\frac{1}{N}\sum_{i=1}^{N}\left\{\left(\|f(I_i|\theta)\|_2 \cdot \|\beta_{net}\|_2\right)^2 - \left(f(I_i|\theta)\cdot\beta_{net}\right)^2\right\}\end{aligned} \tag{9}$$

where w_{reg} is the weight of our proposed regularization term and $\mathcal{L}_{NLPL}(\theta, \beta_{net})$ is the NLPL loss in Eq. (6). The loss in Eq. (9) is the primary objective function for our regularized Deepsurv network. With geometrical consideration, the loss is equivalent to the following rearrangement

$$\begin{aligned}&\mathcal{L}_{NLP\mathcal{L}_{reg}}(\theta, \beta_{net})\\&= \mathcal{L}_{NLPL}(\theta, \beta_{net})\\&\quad + w_{reg}\frac{1}{N}\sum_{i=1}^{N}\left\{\left(\|f(I_i|\theta)\|_2 \cdot \|\beta_{net}\|_2\right)^2 \cdot \left(1-\cos^2\alpha\right)\right\}\\&\mathcal{L}_{NLP\mathcal{L}_{reg}}(\theta, \beta_{net})\\&= \mathcal{L}_{NLPL}(\theta, \beta_{net})\\&\quad + w_{reg}\frac{1}{N}\sum_{i=1}^{N}\left\{\left(\|f(I_i|\theta)\|_2 \cdot \|\beta_{net}\|_2 \cdot sin\,\alpha\right)^2\right\}\end{aligned} \tag{10}$$

where α is an angle formed by β_{net} and a features vector $f_i = f(I_i|\theta)$. The $\|f(I_i|\theta)\|_2 \cdot \|\beta_{net}\|_2 \cdot sin\,\alpha$ is the magnitude of an f_i component that is perpendicular to the β_{net}.

The regularization aims to simultaneously control both the distribution of f_i and the β_{net}. By decreasing $\left(\|f(I_i|\theta)\|_2 \cdot \|\beta_{net}\|_2\right)^2$ and increasing $(f(I_i|\theta)\cdot\beta_{net})^2$, the loss reduces the perpendicular components, which has a similar effect to geometrical projection and makes $f(I_i|\theta)$ or X_{net} more distributed along the direction of β_{net}. The effect is illustrated graphically in Fig 1. In other words, the regularization reduces feature variance that does not contribute to the hazard score and improves the $Var\left(X\cdot\vec{\beta}\right)/Var(X)$ term value on Λ.

The proposed loss force of the Deepsurv network training to filter out the irrelevant information from the input I, and effectively predict with information in β_{net} and X_{net} that matter to the hazard score. It is noteworthy that the loss does not limit its applicability to only the typical Deepsurv network. The proposed method applies to any network designs with outputs with β_{net} and X_{net}, especially network with variations of $f(I_i|\theta)$ explained in section II-B trained with NLPL. However, the proposed loss is not without a drawback. If the loss value is reduced to 0, then that potentially makes $\|\beta_{net}\| = 0$ and renders the prediction useless. Thus, the strength of the regularization w_{reg} must not overwhelm the main objective of minimizing L_{NLPL}. The projection loss is proposed as a complementary objective such that it means to be reduced but not entirely minimized.

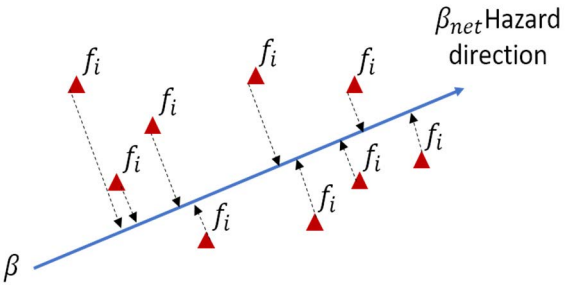

FIGURE 1. Geometric view of representation space. f_i is the penultimate-layer features $f(I_i|\theta)$ from the Deepsurv network. Projection loss regularization prevents the diverging by anchoring the feature toward linear β_{net} plain.

C. IMPROVE MULTIPLE-CORRELATION COEFFICENT AND REDUCE DISCORDANCE FROM NOISE

This section establishes the theoretical foundation and derivation of the proposed method. Our development strategy is to simply find the lower bound on the coefficient R or measure R^2 then seek to maximize them as the secondary learning objective for the Deepsurv network. However, we discover that maximizing the lower bound also lowers the upper bound of the discordance pair occurrence likelihood. In this section, we first state our assumptions surrounding our investigations. Then we provide statements, propositions, and theorems that we used to derive Λ and the projection loss. Finally, we discuss the theorem on how our method reduces the upper bound on the discordance probability.

Consider a simplified survival regression problem in which all survival outcomes are available. Without the censoring, the simplified problem is to derive suitable hazard values and regress for the prediction model. Reference [8] explored the problem by deriving desired hazard value d and the linear prediction model where $d = log(H_0(t)) = X\cdot\beta^* + e$. The formulation converts the problem into a least-square estimation of β using $X - d$ covariance. Notice that d can be re-centered and re-scaled to eliminate the intercept and derive a valid estimation of β^*. The evaluation with the C-Index mostly concerns $\vec{\beta}^*$ and the ordering of $X\cdot\beta^*$ such that estimates of β^* in different magnitudes result in the same performance level. In survival prediction, it is common to assume that probability distributions of the outcome and the censoring are independent. If the model can capture the general trends and distribution of patient survival, the model from uncensored data can still be used to predict the censored cases. The relationship between the features and survival outcome would then apply to the censored cases as well.

We then made the following assumptions.

- First, we assume that $\vec{\beta}^* \approx \vec{\hat{\beta}}$ such that the NLPL solution $\hat{\beta}$ is a valid answer for the least-square β^* formulations that capture the underlying patient survival trend.
- Second, magnitudes of d and β^* are small such that $Var(X) \geq Var(d) \geq \|\beta^*\|_2$ where $Var(X)$

is total variance in the features. This assumption is in-line with many regularizations approach for survival prediction [23] and regression in general, which try to keep parameters small and simple for better generalization. It also implies that d can be rescaled (e.g., z-normalization) such that $Var\,(X) \geq Var\,(d)$ even if the variance is originally more significant than that of input features.

- Third, the noise term in $d \approx X \cdot \beta^* + e$ is a zero-mean error term independent of the covariates X. This assumption is a general assumption of linear prediction, implying that the prediction error is the primary source of discordance.

With the assumptions, we state the following statements and propositions.

Statement 1: $\beta^* = \Sigma_{XX}^{-1} \cdot C_{Xd}$ where Σ_{XX}^{-1} is the inverse of covariance matrix from covariates X, and C_{Xd} is a vector of $Cov\,(X, d)$.

Statement 2: Multiple correlation coefficient is the square root of $R^2 = \rho_{Xd}^T P_{XX}^{-1} \rho_{Xd}$ where P_{XX}^{-1} is the inverse of the $Corr\,(X, X)$ correlation matrix, and $\rho_{Xd} = Corr\,(X, d) = \frac{C_{Xd}}{\sigma_X \sigma_d}$.

Statement 1 is the standard closed-form solution for the least-square regression, which can be done when all the desired values are known. In statistical literature, Statement 2 defines the calculation of R. We use these statements to posit the following propositions.

Proposition 3: Given $\mathbb{E}[d] = 0$, $\mathbb{E}[e] = 0$, and z-normalized X, the coefficient $R^2 = \frac{Var(X^T \cdot d)}{Var(X^T \cdot d) + Var(e)}$.

Proof: With z-normalization, we can consider all elements of σ_X vector equal to 1 and $\Sigma_{XX}^{-1} = P_{XX}^{-1}$. R^2 can then be rearranged according to the above statements as

$$\begin{aligned} R^2 &= \left[\frac{C_{Xd}}{\sigma_d}\right]^T \Sigma_{XX}^{-1} \left[\frac{C_{Xd}}{\sigma_d}\right] \\ &= \left[\frac{C_{Xd}^T \cdot \beta^*}{\sigma_d^2}\right] \\ &= \frac{\left(\mathbb{E}\left[X^T \cdot d\right] - \mathbb{E}\left[X^T\right]\mathbb{E}[d]\right) \cdot \beta^*}{\mathbb{E}\left[d^2\right] - (\mathbb{E}[d])^2} \end{aligned} \tag{11}$$

Consider the following facts. $\mathbb{E}[d] = 0$. Each element of vector $\mathbb{E}[X]$ is 0 due to the z-normalization. Also, the linear formulation states that $d = X \cdot \beta^* + e$. Then,

$$\begin{aligned} R^2 &= \frac{\mathbb{E}\left[X^T \cdot (X \cdot \beta^* + e)\right] \cdot \beta^*}{\mathbb{E}\left[(X \cdot \beta^* + e)^2\right]} \\ &= \frac{\mathbb{E}\left[(X \cdot \beta^*)^2\right] + \mathbb{E}[X \cdot e] \cdot \beta^*}{\mathbb{E}\left[(X \cdot \beta^*)^2\right] + 2\mathbb{E}[(X \cdot \beta^*)]\,\mathbb{E}[e] + \mathbb{E}\left[e^2\right]} \end{aligned} \tag{12}$$

According to the third assumption, the noise term e is independent of X. Thus, $\mathbb{E}[X \cdot e] = \mathbb{E}[X]\,\mathbb{E}[e]$ and $\mathbb{E}[e] = 0$. Then,

$$\begin{aligned} R^2 &= \frac{\mathbb{E}\left[(X \cdot \beta^*)^2\right] + \mathbb{E}[X]\,\mathbb{E}[e] \cdot \beta^*}{\mathbb{E}\left[(X \cdot \beta^*)^2\right] + 2\mathbb{E}[(X \cdot \beta^*)]\,\mathbb{E}[e] + \mathbb{E}\left[e^2\right]} \\ &= \frac{\mathbb{E}\left[(X \cdot \beta^*)^2\right]}{\mathbb{E}\left[(X \cdot \beta^*)^{\mathbf{2}}\right] + \mathbb{E}\left[e^2\right]} \end{aligned} \tag{13}$$

Given that $\mathbb{E}[X] = 0$ and $\mathbb{E}[e] = 0$. Then,

$$\begin{aligned} R^2 &= \frac{\mathbb{E}\left[(X \cdot \beta^*)^2\right] - \left(\mathbb{E}\left[(X \cdot \beta^*)\right]\right)^2}{\mathbb{E}\left[(X \cdot \beta^*)^2\right] - (\mathbb{E}[(X \cdot \beta^*)])^2 + \mathbb{E}\left[e^2\right] - (\mathbb{E}[e])^2} \\ R^2 &= \frac{Var\,(X \cdot \beta^*)}{Var\,(X \cdot \beta^*) + Var\,(e)} \end{aligned} \tag{14}$$

Thus, the equality above proves the proposition. ■

Proposition 4: Given zero-centered rescaled d such that $Var\,(X) \geq Var\,(d)$, then $\frac{Var(X)}{Var(X \cdot \beta^*)} \geq \frac{Var(e)}{Var(X \cdot \beta^*)} + 1$.

Proof: consider the formulation $d = X \cdot \beta^* + e$. As d is a random variable. The variance of d can be calculated as

$$\begin{aligned} Var\,(d) &= Var\left(X \cdot \beta^* + e\right) \\ Var\,(d) &= Var\left(X \cdot \beta^*\right) + Var\,(e) + 2Cov\left(X \cdot \beta^*, e\right) \end{aligned} \tag{15}$$

Consider $Var\,(X) \geq Var\,(d)$ and $Cov\,(X \cdot \beta^*, e) = 0$ from the second and the third assumptions respectively. Then,

$$\begin{aligned} Var\,(d) &= Var\left(X \cdot \beta^*\right) + Var\,(e) \\ Var\,(X) &\geq Var\left(X \cdot \beta^*\right) + Var\,(e) \\ \frac{Var\,(X)}{Var\,(X \cdot \beta^*)} &\geq \frac{Var\,(e)}{Var\,(X \cdot \beta^*)} + 1 \end{aligned} \tag{16}$$

Thus, the equality above proves the proposition. ■

These statements and propositions give some clues on how to grasp the value of R from the available data. The apparent difficulty is d is not entirely available due to censoring. Therefore, we derive a corollary to proposition 3 and the theorem 6 to establish relationships between X, $\vec{\beta}^*$, and the lower bound on R.

Corollary 5: Given $\mathbb{E}[d] = 0$, $\mathbb{E}[e] = 0$, and z-normalized X, then $R \geq R^2 \geq \frac{\mathbb{E}\left[(X \cdot \beta^*)^2\right]}{\mathbb{E}\left[\left(\|X\|_2 \|\beta^*\|_2\right)^2\right] + \mathbb{E}[e^2]}$.

The corollary is merely a derivation from Eq. (13) with Cauchy-Schwarz inequality in the denominator. Specifically, the dot product $(X \cdot \beta^*)^2 \leq \|X\|_2^2 \|\beta^*\|_2^2$. As R is bounded in range [0,1], then $R \geq R^2$.

Theorem 6: Given z-normalized X, $R \geq \|\beta^*\|_2 \sqrt{\frac{Var(X \cdot \vec{\beta}^*)}{Var(X)}}$

Proof: Consider Eq. (14) in Proposition 3. The equation can be re-arranged as

$$\begin{aligned} \frac{1}{R^2} &= \frac{Var\,(X \cdot \beta^*) + Var\,(e)}{Var\,(X \cdot \beta^*)} \\ \frac{1}{R^2} &= 1 + \frac{Var\,(e)}{Var\,(X \cdot \beta^*)} \\ \frac{Var\,(e)}{Var\,(X \cdot \beta^*)} &= \frac{1}{R^2} - 1 \end{aligned} \tag{17}$$

Plugging in the value in Eq. (17) into Eq. (16) in Proposition 4 leads to the following

$$\begin{aligned}\frac{Var\,(X)}{Var\,(X\cdot\beta^*)} &\geq \frac{1}{R^2}+1-1\\ \frac{Var\,(X)}{Var\,(X\cdot\beta^*)} &\geq \frac{1}{R^2}\\ R^2 &\geq \frac{Var\,(X\cdot\beta^*)}{Var\,(X)}\end{aligned} \tag{18}$$

Let $\vec{\beta}^*$ be a unit vector representing the direction of β^*. Then,

$$\begin{aligned}R^2 &\geq \frac{Var\left(X\cdot\vec{\beta}^*\,\|\beta^*\|_2\right)}{Var\,(X)}\\ R^2 &\geq \frac{Var\left(X\cdot\vec{\beta}^*\right)\|\beta^*\|_2^2}{Var\,(X)}\\ R &\geq \|\beta^*\|_2\sqrt{\frac{Var\left(X\cdot\vec{\beta}^*\right)}{Var\,(X)}}\end{aligned} \tag{19}$$

Thus, the equality above proves the theorem. ■

Both Corollary 5 and Theorem 6 give helpful information about the lower bound of R related to $X\cdot\beta^*$,in which we use them to derive Λ and the projection loss. We substitute the direction $\vec{\beta}^*$ with $\vec{\hat{\beta}}$ from the NLPL formulation due to the first assumption. The Deepsurv formulation substitute X and $\vec{\beta}^*$ with X_{net} and $\vec{\beta}_{net}$. Therefore, the fraction of variance term can be substitute with Λ defined in Eq. (7). Eq. (19) establish the lower bound $\lfloor R\rfloor$ that

$$\begin{aligned}\lfloor R\rfloor &= c\Lambda\\ \lfloor R\rfloor &\propto \Lambda\end{aligned} \tag{20}$$

and vice versa,

$$\Lambda \propto \lfloor R\rfloor \tag{21}$$

where $c = \|\beta^*\|_2$ is from by solving for β^* using d or its scaled version according to the second assumption. The purpose of this relationship between $\lfloor R\rfloor$ and Λ is not to measure the exact value of $\lfloor R\rfloor$ via Λ because $\|\beta^*\|_2$ is unknown due to the d censoring. The property shows that maximizing Λ is equivalent to maximizing the lower bound $\lfloor R\rfloor$ and encourages using Λ as the feature quality measure.

Corollary 5 offers a simpler way to increase $\lfloor R\rfloor$. It suggests a simultaneous decrease in $\mathbb{E}\left[\left(\|X\|_2\,\|\beta^*\|_2\right)^2\right]$ and increase in $\mathbb{E}\left[(X\cdot\beta^*)^2\right]$, which we use in the projection loss formulation in Eq. (9). For the formulation, we substitute the constant β^* with β_{net} to control the variance direction and magnitude of β_{net} at the same time. Technically, the proposed regularization improves Λ by increasing $Var\left(X_{net}\cdot\vec{\beta}_{net}\right)$ and reduce $Var\,(X_{net})$ that does not contribute to hazard score.

Because many parts of the derivation use z-normalization to ensure the zero-mean property, the theorem encourages using batch-normalization on the penultimate layer at the minimum. Therefore, we include batch-normalization as a crucial part of our network designs and training approach due to this theoretical insight.

There are valuable interpretations on improving R and R^2 through the proposed projection loss. Achieving good C-index performance with β_{net} at a higher value of R means that the network finds concrete X_{net} that correlated well with the desire hazard score even if complete knowledge of the exact score d is not available. The higher value of R^2 means the β_{net} trend line has smaller fraction of unexplained variance, which implies that the network efficiently filters out the irrelevant information from the input I and effectively predicts with information in X_{net} that really matters.

The minimizing projection loss not only forces the Deepsurv network to digest better information from the input I. We discover that it also reduces the discordance rate caused by noise and improves C-index performance. Consider that the desired hazard score variable $d = X\cdot\beta + e$ is the variable that strictly follows the ordering. In other words, the linear model inherently establishes that the error or noise term e causes the discordance of some $X\cdot\beta$ pair comparisons such that $X\cdot\beta = d - e$ and $d - e$ diverge from the ordering of d, which follows the third assumption.

We establish the following proposition and corollary to proposition 4 to show the foundation of our upper bound on discordance probability.

Proposition 7: Given β^*, and (X_i, X_j) which are two samples of random variables X, then the upper bound $P\left(\left(e_i-e_j\right)^2 \geq \left(X_i\cdot\beta^* - X_j\cdot\beta^*\right)^2\right) \leq \frac{\mathbb{E}\left[\left(e_i-e_j\right)^2\right]}{\left(X_i\cdot\beta^*-X_j\cdot\beta^*\right)^2}$.

Proof: Consider a random variable$\Delta_e = \left(e_i-e_j\right)^2$. From the definition, Δ_e is non-negative. Then, the Markov inequality defines an upper-bound probability related to the quantity of Δ_e as

$$P\left(\Delta_e \geq k\right) \leq \frac{\mathbb{E}\left[\Delta_e\right]}{k} \tag{22}$$

where k is a non-negative constant. As values of β^*, X_i, and X_j are given, let $k = \left(X_i\cdot\beta^* - X_j\cdot\beta^*\right)^2$. Then,

$$\begin{aligned}P\left(\Delta_e \geq \left(X_i\cdot\beta^* - X_j\cdot\beta^*\right)^2\right) &\leq \frac{\mathbb{E}\left[\Delta_e\right]}{\left(X_i\cdot\beta^* - X_j\cdot\beta^*\right)^2}\\ P\left(\left(e_i-e_j\right)^2 \geq \left(X_i\cdot\beta^* - X_j\cdot\beta^*\right)^2\right) &\leq \frac{\mathbb{E}\left[\left(e_i-e_j\right)^2\right]}{\left(X_i\cdot\beta^* - X_j\cdot\beta^*\right)^2}\end{aligned} \tag{23}$$

Thus, the derivation of Eq. (20) proves the proposition. ■

Corollary 8: *Given zero-centered rescaled d such that* $Var\,(X) \geq Var\,(d)$, *then* $\frac{Var(e)}{Var(X\cdot\beta^*)} \leq \frac{Var(X)}{Var(X\cdot\beta^*)} - 1$.

The corollary is a rearrangement of Eq. (16) in proposition 4. The equation focuses on the upper bound of noise variance.

The stated propositions and corollary provide information about the error term. Specifically, proposition 7 provide a probability bound of events that the error difference exceeds

TABLE 1. Summary of datasets in experiments.

Dataset	# Feature	# Cases		Event Time			
		Total	Censored	Min	Max	Mean	Median
Wisconsin	35	198	151 (76.26%)	1	125	46.73	39.5
METABRIC	9	1904	801 (42.07%)	0	355.20	125.03	114.90
GBSG	7	2232	965 (43.23%)	0.26	87.36	44.49	40.22
NWTCO	8	4028	3457 (85.82%)	4	6209	2276.98	1939.0
FLCHAIN	10	6524	4562 (69.93%)	0	5166	3647.50	4303.0
Sarcoma-Rad-UW	45	200	157 (78.5%)	43	6139	1261.77	1111.0
Sarcoma-Rad-Munich	45	72	51 (70.83%)	69	2486	1136.47	1031.5
Sarcoma-3DMRI-UW	3D-MRI-Scans (64x64x64)	200	157 (78.5%)	43	6139	1261.77	1111.0
Sarcoma-3DMRI-Munich	3D-MRI-Scans (64x64x64)	72	51 (70.83%)	69	2486	1136.47	1031.5

that of the hazard score. Corollary 8 suggests a worse-case quantity of error variance relative to variance in input data and the predicted score. Then, we derive the discordance probability in the predicted score as in the following theorem.

Theorem 9: Given β^* and z-normalized X, then $P\left(discordantin X\cdot\beta^*\right) \leq \frac{Var(X)}{\|\beta^*\|_2^2 Var\left(\vec{X\cdot\beta}^*\right)} - 1$.

Proof: Let X_i, X_j and e_i, e_j be two random samples of features X and the error noise e, respectively. We define discordance events caused by the noise term as events where the ordering of $d = X\cdot\beta^* + e$ differs from that of the predicted score $X\cdot\beta^*$ for any two pair of samples i, j. In other words, orders for d and $X\cdot\beta^*$ should be equivalent if the magnitude of the error term does not change the results of any pairwise comparison. Consider the formulation of discordance events as the following joint cases.

$$\begin{aligned}\#\left(discd\right) = \#\Big(&\left(sign\left(X_i\cdot\beta^* - X_j\cdot\beta^*\right) \neq sign\left(e_i - e_j\right)\right) \\ &\wedge \left(\left|e_i - e_j\right| \geq \left|X_i\cdot\beta^* - X_j\cdot\beta^*\right|\right)\Big) \end{aligned} \tag{24}$$

where $\#\left(discd\right)$ is the number of discordant events. The function $sign\left(k\right) = 1$ if $k \geq 0$, and $sign\left(k\right) = -1$ otherwise. From the above definition, discordant pair occurs when pairwise the comparison of signs between $X_i\cdot\beta^*, X_j\cdot\beta^*$ pair is not the same as that of e_i, e_j pair and the noise pair magnitude difference is significant enough to interfere with the order. Otherwise, the comparison of $X_i\cdot\beta^* + e_i$ and $X_j\cdot\beta^* + e_j$ is the same as that of $X_i\cdot\beta^*$ and $X_j\cdot\beta^*$. Eq. (24) further establishes that

$$P\left(discd\right) = P\left(sign\left(\Delta_{X\cdot\beta^*}\right) \neq sign\left(\Delta_e\right), |\Delta_e| \geq \left|\Delta_{X\cdot\beta^*}\right|\right)$$

$$\begin{aligned} P\left(discd\right) &\leq P(|\Delta_e| \geq \left|\Delta_{X\cdot\beta^*}\right|) \\ P\left(discd\right) &\leq P\left(\left(\Delta_e\right)^2 \geq \left(\Delta_{X\cdot\beta^*}\right)^2\right) \end{aligned} \tag{25}$$

where $\Delta_{X\cdot\beta^*} = X_i\cdot\beta^* - X_j\cdot\beta^*$, and $\Delta_e = e_i - e_j$. The probability of $\left(\Delta_e\right)^2 \geq \left(\Delta_{X\cdot\beta^*}\right)^2$ is a $P\left(discd\right)$ upper bound. Using proposition 7, Eq. (23) can be elaborated as

$$\begin{aligned} P\left(discd\right) &\leq \frac{\mathbb{E}\left[\left(e_i - e_j\right)^2\right]}{\left(X_i\cdot\beta^* - X_j\cdot\beta^*\right)^2} \\ \left(X_i\cdot\beta^* - X_j\cdot\beta^*\right)^2 &\leq \frac{\mathbb{E}\left[\left(e_i - e_j\right)^2\right]}{P\left(discd\right)} \end{aligned} \tag{26}$$

We seek to gauge overall quantity for any pair comparison. Thus, we apply expectation on both sides of the inequality

$$\mathbb{E}\left[\left(X_i\cdot\beta^* - X_j\cdot\beta^*\right)^2\right] \leq \mathbb{E}\left[\frac{\mathbb{E}\left[\left(e_i - e_j\right)^2\right]}{P\left(discd\right)}\right] \tag{27}$$

As $P\left(discd\right)$ is an intrinsic constant on the population, the expectation operator does not change its value. Likewise, $\mathbb{E}\left[\left(e_i - e_j\right)^2\right]$ is a constant. Then,

$$\begin{aligned} &\mathbb{E}\left[\left(X_i\cdot\beta^* - X_j\cdot\beta^*\right)^2\right] \\ &\leq \frac{\mathbb{E}\left[\left(e_i - e_j\right)^2\right]}{P\left(discd\right)} \end{aligned}$$

$$P(discd) \leq \frac{\mathbb{E}\left[(e_i - e_j)^2\right]}{\mathbb{E}\left[\left(X_i \cdot \beta^* - X_j \cdot \beta^*\right)^2\right]} \leq \frac{\mathbb{E}\left[e_{\mathbf{i}}^2 + e_j^2 - 2e_i e_j\right]}{\mathbb{E}\left[(X_i \cdot \beta^*)^2 + \left(X_j \cdot \beta^*\right)^2 - 2\left(X_i \cdot \beta^*\right)\left(X_j \cdot \beta^*\right)\right]} \tag{28}$$

Consider that X_i, X_j and e_i, e_j are randomly sampled from same variables X and e, respectively. Therefore, X_i, X_j and e_i, e_j are independent and identically distributed (iid). With zero-mean noise according to the third assumption and z-normalization X centered at zero means, then $\mathbb{E}[e] = \mathbb{E}[e_i] = \mathbb{E}[e_j] = 0$, $\mathbb{E}[X \cdot \beta^*] = \mathbb{E}[X_i \cdot \beta^*] = \mathbb{E}[X_j \cdot \beta^*] = 0$, $\mathbb{E}[e_i e_j] = \mathbb{E}[e_i]\mathbb{E}[e_j] = 0$, and $\mathbb{E}\left[(X_{\mathbf{i}} \cdot \beta^*)\left(X_j \cdot \beta^*\right)\right] = \mathbb{E}\left[(X_i \cdot \beta^*)\right]\mathbb{E}\left[\left(X_j \cdot \beta^*\right)\right] = 0$. With these facts, Eq. (28) can be simplified as

$$\begin{aligned} &P(discd) \\ &\leq \frac{\mathbb{E}\left[e_i^2\right] + \mathbb{E}\left[e_j^2\right] - 2\mathbb{E}[e_i]\,\mathbb{E}\left[e_j\right]}{\mathbb{E}\left[(X_i \cdot \beta^*)^2\right] + \mathbb{E}\left[\left(X_j \cdot \beta^*\right)^2\right] + 2\mathbb{E}\left[(X_i \cdot \beta^*)\right]\mathbb{E}\left[X_j \cdot \beta^*\right]} \\ &P(discd) \leq \frac{2\mathbb{E}\left[e^2\right]}{2\mathbb{E}\left[(X \cdot \beta^*)^2\right]} \\ &P(discd) \leq \frac{\mathbb{E}\left[e^2\right] - (\mathbb{E}[e])^2}{\mathbb{E}\left[(X \cdot \beta^*)^2\right] - (\mathbb{E}[X \cdot \beta^*])^2} \\ &P(discd) \leq \frac{Var(e)}{Var(X \cdot \beta^*)} \end{aligned} \tag{29}$$

Consider the upper bound in Corollary 8, Eq. (29) can be re-arranged as

$$\begin{aligned} P(discd) &\leq \frac{Var(X)}{Var(X \cdot \beta^*)} - 1 \\ P(discd) &\leq \frac{Var(X)}{Var\left(X \cdot \vec{\beta}^* \|\beta^*\|_2\right)} - 1 \\ P(discd) &\leq \frac{Var(X)}{\mathbb{E}\left[\left(X \cdot \beta^* \|\beta^*\|_2\right)^2\right] - \left(\mathbb{E}\left[X \cdot \beta^* \|\beta^*\|_2\right]\right)^2} - 1 \\ P(discd) &\leq \frac{Var(X)}{\|\beta^*\|_2^2\, Var\left(X \cdot \vec{\beta}^*\right)} - 1 \end{aligned} \tag{30}$$

Therefore, Eq. (30) prove the theorem. ■

From theorem 2, we establish that the upper bound of the discordance probability $\lceil P(discd) \rceil$ is inversely proportional to the fraction of the $X \cdot \vec{\beta}^*$ variance and the total variance in X. For the Deepsurv network, X is X_{net}, and $\vec{\beta}^*$ is substituted with $\vec{\beta}_{net}$ according to the second assumption. Eq. (30) establish the upper bound that

$$\begin{aligned} \lceil P(discd) \rceil &= c\frac{1}{\Lambda^2} - 1 \\ \lceil P(discd) \rceil &\propto \frac{1}{\Lambda^2} \end{aligned} \tag{31}$$

and vice versa,

$$\Lambda^2 \propto \frac{1}{\lceil \mathrm{P}(\mathrm{discd}) \rceil} \tag{32}$$

where $c = 1/\|\beta^*\|_2^2$.The relationship further expands the merit of increasing Λ such that maximizing Λ^2 is equivalent to minimizing the upper bound of the probability of discordance pair in the predicted score. The theorem reveals that our proposed regularization improves C-index performance by increasing Λ^2, which we investigate further in our experiments.

IV. EXPERIMENTS

Our investigations through the experiments aim to address the following aspect under the medical context with limited sample size.

- Impact of the regularization on C-index performance.
- Effect of the regularization on the Deepsurv's feature.
- Performance of the regularized model relative to that of the State-of-the-art.

A. DATASETS

Table 1 gives an overview of the survival datasets on the number of features and censoring. Our experiments employed five public datasets and four in-house developed datasets. We use two criteria in selecting the current public datasets in our experiments. First, the survival datasets should be related to healthcare, have limited sample sizes, and are highly censored. In this study, we posit that each dataset should have less than 8,000 cases, and more than 40% are censored. Second, the selected datasets have been previously used to demonstrate the performance of the included baselines (e.g., when the included baselines were proposed). Intuitively, we ensure fair comparisons such that all baselines are expected to perform well on some datasets.

In brief details, Wisconsin Breast Cancer prognosis (Wisconsin) is a dataset for breast cancer diagnosis which analyzes digital images of cells taken from breast lumps to time recurrence of cancer. Expert-defined features were extracted from cellular nuclei images. Each record represents follow-up data for a cancer case. The censored case is defined by no recurrence during the follow-up period. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) is a breast cancer survival prediction dataset based on gene expressions and clinical features. The target variable is the number of months until observed death. Rotterdam and German Breast Cancer Study Group (GBSG) dataset contains records of node-positive breast cancer patients with features related to effects from chemotherapy and hormone treatments. The recorded survival times are in the number of months. National Wilm's Tumor Study (NWTCO) [2] is a dataset to study the relationship between tumor histology on embryonal kidney cancer and treatment outcome. The features are clinical and histological variables. All the cases are associated with the time-to-death or survival time of patients. Patients who survived over the follow-up period are censored.

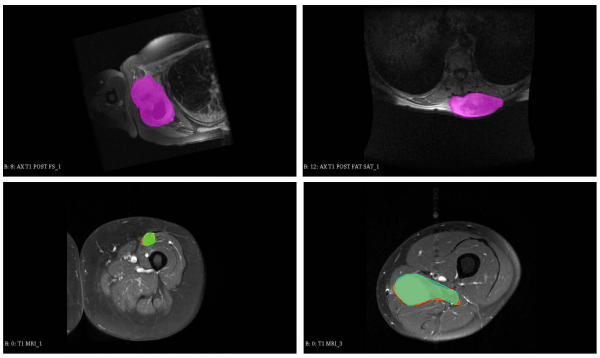

FIGURE 2. Example visualization in MRI dataset with segmented ROI defined by expert. Top: samples from the UW cohort patients. Bottom: samples from the Munich cohort patients.

Assay of Serum Free Light chain (FLCHAIN) is a dataset for studying the prevalence of the monoclonal gammopathy of undetermined significance (MGUS), an immune disregulation condition. We employ the same pre-processing as in [11]. Notably, the censored patients either survived or dropped out of the study during follow-up. The extracted variables are conditions related to the ailment. Wisconsin data is available at the UCI repository. METABRIC, GBSG, NWTCO, and FLCHAIN datasets are available either in the R Survival package or Pycox package.

We developed Sarcoma datasets from magnetic resonance imaging (MRI) scans of patients with sarcoma soft-tissue cancer. Soft tissue sarcoma is a heterogeneous cancer with severe outcomes for many patients. Pre-treatment contrast-enhanced T1-weighted 3D MRI scans were acquired from two independent cohorts of patients diagnosed with biopsy-proven soft tissue Sarcoma (STS) from the University of Washington (UW cohort) and the Technical University of Munich (Munich cohort). The acquisition was done with the institutional picture archiving and communication system (PACS) standard using a similar image matrix and resolutions to [15]. Using the similar protocol to [15], the selected patients had high-risk STS of various histologies of the extremity, trunk, or retroperitoneum. In both cohorts of Sarcoma datasets, radiologist and radiation oncologist experts manually segmented the gross tumor as ROI at the fixed resolution of $1mm^3$, which later resampled into the size of $64\times64\times64$ voxels. The segmentation and resampling were completed using MIM software (version 6.6, MIM Software Inc., Cleveland, OH) for the UW cohort and iPlan RT (version 4.1.2, Brainlab, Munich, Germany) for the Munich cohort. Fig. 2 visualizes some samples in our datasets.

Sarcoma-Rad-UW and Sarcoma-Rad-Munich are radiomics features derived from the MRI scans describing the tumor's textural appearances from each cohort. The relationship of these empirical image features sets to patient survival, pathologic response, and tumor grade has been described previously [15], [16], [24], [25]. The features are extracted using the PORTS software package and extraction protocol as in [15]. Sarcoma-3DMRI-UW and Sarcoma-3DMRI-Munich data are bounded MRI scans from each corresponding institute. The scans were normalized as in [3]. Sarcoma-3DMRI is not in the typical feature vector format. There have been many successful explorations in predicting severity and patient risk directly from 3D cancer scans [3], [19], [26]. We use these datasets to demonstrate our regularization applicable on DeepConvSurv network [19] and to perform comparisons with many other deep survival baselines for end-to-end prediction. More details are provided in experiment Section IV-D test scenario three.

B. EFFECT ON C-INDEX PERFORMANCE

The first experiment demonstrated effects on the Deepsurv network C-Index performance given different weights w_{reg} settings on Wisconsin, METABRIC, GBSG, NWTCO, and FLCHAIN datasets. In all datasets, all features were z-normalized before the experiments except for binary features. We randomly split each dataset into 60% training, 20% validation, and 20% testing. Our Deepsurv network architecture was (#feature-128-64) network which consisted of 2 hidden layers of size 128 and 64 nodes followed by a 1-nodes hazard output. Each of the hidden layers used Leaky-Rectified Linear unit (L-ReLU), defined as $max(0.01X, X)$, followed by batch-normalization. In prior experiments, we tried multiple activation functions and designated L-ReLU for its performance. We set w_{reg} as 1×10^m where integer $m \in [-5, 5]$. All the results were compared with no regularization and norm-based Ridge and Lasso regularization on β_{net} with the same regularization weight range. We set Adam optimizer with a 0.01 learning rate in all trainings. The networks were trained under loss function in equation (6) until no improvement on validation performance after 10 patience epochs from the best validation round. The experiments were repeated 1000 times with different random data splits. The estimated C-index from the testing set was averaged across all the repetitions. Then, the averaged C-index estimates were used to compare the performances of different regularization strategies. Notice that the splitting and result measurement reflect the scenario where the training, validation, and testing set belong to the identical distributions from well-explored datasets. We applied the same network architecture and the grid search on regularization weights of all the compared strategies.

Fig. 3 summarizes the results. Improvement in C-index performance under our regularization method can be observed in different magnitudes across the five datasets. In the smaller Wisconsin, METABRIC, and GBSG datasets (<3000 cases), all the regularization strategies with appropriate weight led to better performance. For the larger datasets (approximately 4000-6000 cases), however, we observe that the norm-based method performances were marginally worse than that of no regularization, whereas our method was improved from the no regularization baseline to a certain extent. It is also noteworthy that the appropriate weight for our approach tends to be more stable such that the peak C-Index performances were from w_{reg} values around [0.01, 1], whereas the suitable range

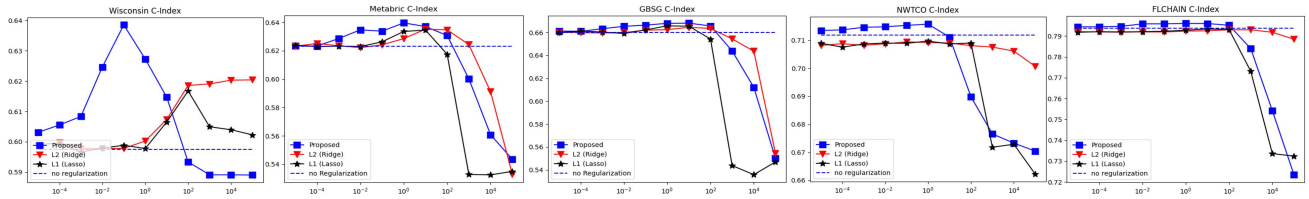

FIGURE 3. C-Index performances on 5 datasets using different regularization strategies. From left to right: Wisconsin, Metabric, GBSG, NWTCO, and FLCHAIN datasets, respectively. Square, triangle, and star markers denote entries from the proposed, Ridge, and Lasso strategies, respectively. The dash lines represent performance from no regularization. The Horizontal axis represents value of w_{reg} 1×10^{m} where integer $m \in [-5, 5]$. The vertical axis represents average C-Index value.

for other strategies varied depending on the dataset. Applying too large w_{reg} values can lead to poor performance worse than that of no regularization. Nevertheless, the performance trends show the potential of our proposed regularization on improving the C-index.

C. EFFECT ON THE FEATURE DISTRIBUTION

The improved performance warranted deeper investigations on the suitability between the learned feature and the hazard direction, along with the effect of regularization on the feature distribution. The second experiment repeated previous settings with the highest performance from each regularization strategy 1000 times in all datasets. In each run, Λ measures were calculated to inspect whether there are increases from any regularization strategies. The values were averaged and then compared to observe the learned information from the trained network in each setting. In addition to the measure, testing $X = f(I|\theta)$ and β_{net} of each strategy were extracted from a data split with the highest C-Index. Then PCA dimensional reduction was applied to reduce the dimension of X_{net} to 2 for visual comparison.

Table 2 presents average Λ and C-index measures. Due to slight differences in C-index performances between some strategies, 95% confidence interval halfwidths are provided for all C-Index results. The unregularized Deepsurv resulted in poor Λ and learned β_{net} such that a tiny fraction of the variance contributed to the prediction. Even though low Λ does not necessarily result in a poor C-index, the network could have done better optimization to digest input I for relevant information. Otherwise, the network would predict using less suitable β_{net} and more likely to capture noise information. The results opened opportunity for the feature control and selection.

Norm-based regularization had varying effects on Λ and C-index performances. We observe increases in both Λ and C-index performance on Wisconsin, METABRIC, and GBSG data. On the other hand, the norm-based strategies somewhat altered the Λ in NWTCO and FLCHAIN datasets, which means the regularizations have some influence over features distribution, not just the β_{net} parameter. However, they offered no significant change on C-index, suggesting failure to select well-perform features information on both X_{net} and β_{net}. Thus, norm-based strategies are less effective for the Deepsurv network.

The proposed regularization remedies the problem by allowing more control on the distribution of the learned feature. Larger Λ values mean that the network learned to encode more information from X_{net} that the β_{net} line can capture. Fig. 4 also elaborates our observations from Table 2. The useful information that drives the prediction is the data projected toward the β_{net} line, with less variation in the other direction. Thus, network trainings with less effective control cause more data to be distributed more perpendicular to the line instead of the trend of increasing or decreasing hazards. From this perspective, the proposed regularization showed a more linearly oriented distribution toward the hazard prediction. Better C-index performance across all the datasets also supported the utility of our proposed method.

D. PERFORMANCE COMPARISON WITH OTHER STATE-OF-THE-ART APPROACHES

In the third experiment, we evaluate our proposed method against other state-of-the-art approaches. The tests are under 3 increasingly difficult data scenarios relevant to the prediction model development for medical prognosis and decision making. We design experiments to inspect whether the proposed method's performance can generalize.

The first scenario is when expert-defined variables are available, and distributions of training and testing datasets are almost identical. This scenario is expected at the early-stage model development to confirm that the prediction approach holds sufficient predictive power under the available data. The test is constructed to satisfy iid data environments such that errors are mainly caused by prediction approaches rather than the discrepancy between training and testing distributions. We use well-established public datasets to further ensure that tested models learn to predict only from the relevant feature information. We employ 5 datasets from previous experiments with the same training/validation/testing splits. For our proposed strategy, we tried various values of w_{reg} for each dataset among the [0.1, 1.0] interval. Then, we select configurations with the best validation performance and compare them with those of 6 deep learning and 4 non-deep learning approaches. The deep learning baselines were unregularized Deepsurv [9], Cox-time network [11], Survival Net [3] with CPH model regression, NNet [6], PMF network [10], and Deep Hit network [12]. All baselines used the

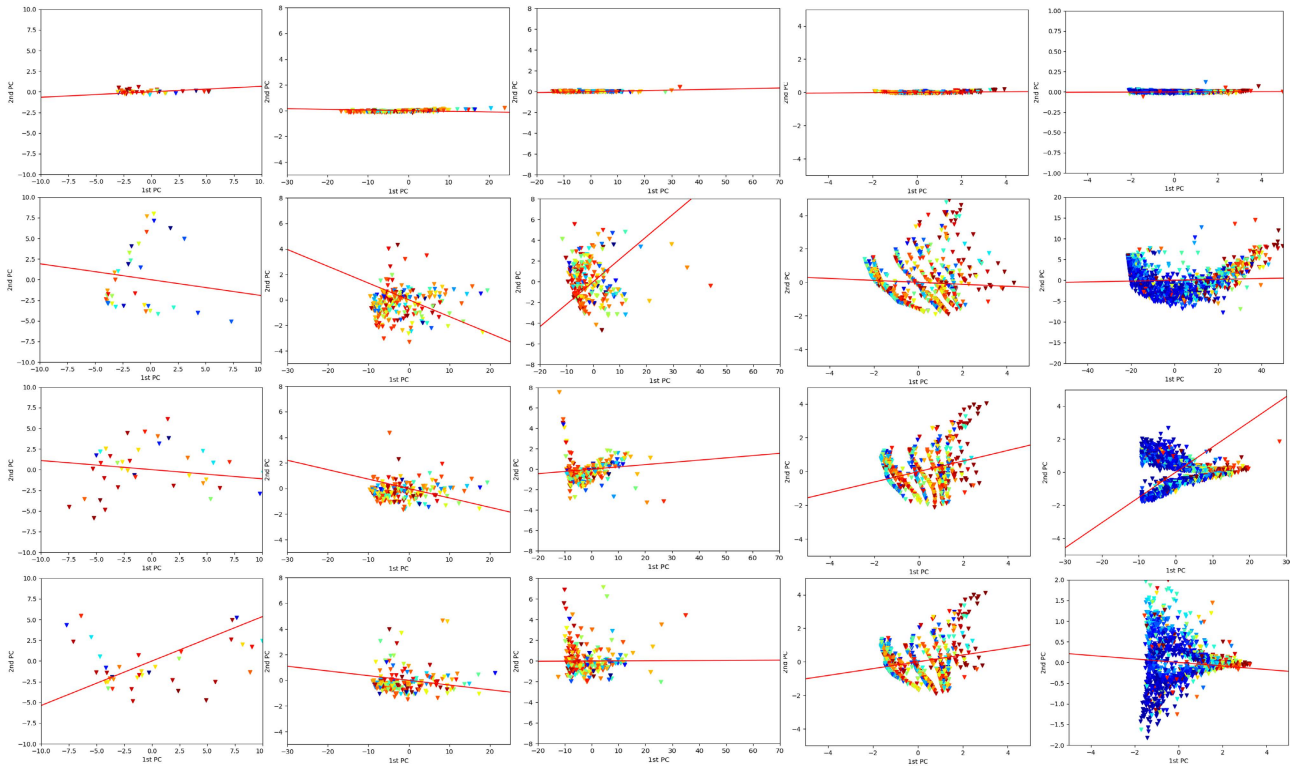

FIGURE 4. Scatterplots of the extracted feature $X = f(I|\theta)$ reduced with PCA to 2 dimensions. The horizontal and vertical axes are coordinate values of first and second PCs respectively. The data markers are non-censored cases color-coded in heatmap style according to decreasing value of non-censored event time. High-risk cases mark in red are data points closer to the minimum event time record. Low-risk cases marked in dark blue are data points that occurred near the maximum of the record time. Red lines are projected β_{net} of each setting. From Left to Right: Wisconsin, METABRIC, GBSG, NWTCO, and FLCHAIN datasets respectively. From Top to bottom: Proposed regularization, Ridge, Lasso, and no regularization strategies.

TABLE 2. C-index performances comparison and the amount of the projected variance Λ contribute to the hazard prediction as the results of different regularization strategies. 95% confidence interval halfwidths are provided to illustrate significant differences in performances.

Dataset	Proposed Regularization		Ridge (L-2 norm penalty)		Lasso (L-1 norm penalty)		No Regularization	
	Λ	C-Index	Λ	C-Index	Λ	C-Index	Λ	C-Index
Wisconsin	**0.9434**	**0.639** (±0.006)	0.6541	0.619 (±0.006)	0.4911	0.617 (±0.006)	0.3591	0.591 (±0.006)
METABRIC	**0.9026**	**0.640** (±0.001)	0.2284	0.635 (±0.001)	0.1480	0.635 (±0.001)	0.1332	0.623 (±0.001)
GBSG	**0.6426**	**0.668** (±0.001)	0.2577	0.665 (±0.001)	0.1264	0.665 (±0.001)	0.1488	0.660 (±0.001)
NWTCO	**0.8779**	**0.716** (±0.002)	0.3680	0.709 (±0.002)	0.4061	0.709 (±0.002)	0.3437	0.710 (±0.002)
FLCHAIN	**0.7871**	**0.797** (±0.001)	0.4964	0.793 (±0.001)	0.3499	0.793 (±0.001)	0.4538	0.793 (±0.001)

same core network as in experiment 1 with different last layers and losses depends on their respective training strategies. The non-deep learning baselines were a typical CPH model, survival SVM with linear kernel [20], Survival Regression Tree [21], and survival regression Forest [22]. For all baselines which require discretization of output time (NNet, PMF, Deep Hit), we tried varying coarse discretization by doubly increasing the grouping of survival time from 1 day, 2 days, 4 days, 8, days, and etc. After trials, we set the finest discretization defined as max(time) - min(time) + 1 because it consistently outputted the best C-Index across many of the discretization baselines. We designate the time data in the unit of days for Wisconsin, NWTCO, and FLCHAIN datasets, and in the unit of months for METABRIC and GBSG datasets. Intuitively, the fine-grain discretization assumed no prior expert knowledge on time-to-event distribution such that the discretization should not destroy comparability between cases. Similar to the proposed strategy, all the deep learning baselines were repeatedly trained using Adam Optimizer with the same learning rate and early stopping criteria. The repeated evaluations compared average testing C-Index with 95% confidence interval (CI) under 1000 runs. Notice that all networks had their number of parameters greater than the number of data cases ($\gg$8000). The large numbers reflect the usual scenario of limited data in survival analysis under the medical setting.

The second scenario is when distributions of training and testing sets are not necessarily identical. However, the

features that encoded some expert knowledge are available. Under practical circumstances, information learned from the expert-provided features must be applicable for predicting various new cases. This expectation is necessary and realistic, especially when the survival model undergoes external validation between different cohorts of patients (e.g., due to differences in patient or tumor characteristics) from different institutions. The test is designed to mimic the validation to ensure that medical-decision making is based on generalizable predictions. The proposed method and baselines were trained and validated using Sarcoma-Rad-UW with random 80% training and 20% validation data split. Instead of the same dataset, we tested the trained models on Sarcoma-Rad-Munich data. The experiment repeated 1000 times with different train-validate splits. All baselines from the first scenario were also subject to this experiment with the same architectural settings as in the first scenario. We then compared C-Index performances with 95% CI on the testing set across survival prediction models.

The third scenario is when training and testing data distributions are not identical, and expert-defined features are unavailable. Unlike the second scenario, this scenario is more difficult as there is no expert guidance on specific information to capture from the raw data. The test is conducted to demonstrate the proposed method's applicability and to observe the generalizability of various deep survival approaches under non-typical input instead of the handcrafted feature vector. The proposed method and baselines were trained and validated using the Sarcoma-3DMRI-UW dataset and tested on the Sarcoma-3DMRI-Munich dataset. Similar to the second scenario, the UW cohort cases were randomly split into 80% train and 20% validate data. The experiment repeated 100 times with different train-validate splits. Due to no expert-defined feature variables, non-deep learning baselines were excluded for Sarcoma-3DMRI experiments. We used the same deep learning baselines as the first and second scenarios with the core network replaced with a convolutional neural network (CNN). CNN version of Deepsurv is also called Deepconvsurv, which has been explored in [19]. The CNN architecture setting was (img-conv16 conv32-conv64-conv128-flatten-512-128) consisted of 4 convolution layers with an increasing number of filters from 16-128 followed by a feed-forward network with 2 hidden layers of size 512 and 128. All convolutional filters have a size of $3 \times 3 \times 3$. All convolutions and feed-forward layers used the L-ReLU activations function and followed by batch normalization. Random flipping augmentation was tried but later dropped due to validity concerns and no significant improvement in all baselines. All training used Adam optimizer with a learning rate set to 0.0001. C-Index performances with 95% CI on the testing set were compared.

Table 3 summarizes the result comparisons of the proposed regularization with all the baselines. We observe that the proposed method performed on par or better than state-of-the-art methods and non-deep learning baselines in 4 out of 5 datasets. The outperformance demonstrates that our regularization applies well when data distributions are similar across training, validation, and testing sets. It also shows the utility of the proposed method that does not require discretization of output time, which can be difficult to define without expert knowledge. In larger GBSG NWTCO and FLCHAIN datasets, inferior performances of NNet, PMF, and Deep hit may be due to the simpler core network, unlike bigger networks in their original works that greatly increase the number of parameters in the networks. In a separate experiment, we experienced increased performance using larger networks with more time-consuming training. However, the larger network performed poorly on the small Wisconsin dataset. Interestingly, Deephit outperformed all other baselines in the METABRIC dataset despite being trained with a relatively small amount of data and underperformances of other discrete-time approaches. This discrepancy is a good reminder that the approach is not a bad survival baseline and should be considered in our following scenarios. Nevertheless, unregularized regression-based networks (Cox-time and Deepsurv) mostly performed on par or better than many discretization approaches at the current core network setting. The underperformance demonstrates limitations of the discretization approach under the current network setting and scenario.

Table 4 and Table 5 illustrate performances comparisons using Sarcoma-Rad datasets, whose expert features guide the prediction when training and testing datasets are not the same. The proposed method outperformed both deep learning and non-deep learning baselines, achieving the average C-Index of 0.657. In this small dataset, NNet, PMF, and Survival Net failed to outperformance many non-deep learning baselines, suggesting that this approach could not learn generalizable information for the prediction under such a data scenario. DeepHit, Deepsurv, Cox-time, and the proposed networks outperformed the non-deep learning baseline in this scenario. Most of these well-perform networks are from the regression-based approach.

Similar trends in performance also exist in experiments with more difficult Sarcoma-3DMRI in which the training and testing dataset may not have the same data distribution and no expert variable to guide the prediction. In Table 6, all the deep learning approaches had lower performances without the expert information, especially DeepHit network whose performance dropped the most from the previous scenario. All the discrete-time baselines performed significantly poorly compared to the regression-based networks. Despite the weaker performance, the proposed method outperformed the baselines with the average C-Index of 0.630.

Performance across all the scenarios show the utility of our approach under small-data cross-cohort environments. Even though it is arguable that improved performances under the first scenario in Table 3 were marginal, the regularization prevented significant performance drops due to data changes in Table 5 and Table 6. The minor performance decrease means that our regularized network was successful in learning more generalizable predictions. The compatibility

TABLE 3. C-index results comparison with states-of-the-art method in 5 datasets.

Methods	Performances				
	Wisconsin	Metabric	GBSG	NWTCO	FLCHAIN
Proposed	**0.649 (±0.006)**	0.640 (±0.001)	**0.668 (±0.001)**	**0.716 (±0.002)**	**0.797 (±0.001)**
CPH	0.606 (±0.006)	0.590 (±0.001)	0.655 (±0.001)	0.708 (±0.002)	0.696 (±0.005)
SVM	0.603 (±0.006)	0.587 (±0.001)	0.645 (±0.001)	0.710 (±0.002)	0.793 (±0.001)
Survival Tree	0.526 (±0.006)	0.576 (±0.001)	0.588 (±0.001)	0.613 (±0.002)	0.703 (±0.001)
Survival Forest	0.615 (±0.006)	0.640 (±0.001)	0.664 (±0.001)	0.681 (±0.002)	0.783 (±0.001)
Deepsurv	0.597 (±0.006)	0.623 (±0.001)	0.660 (±0.001)	0.708 (±0.002)	0.791 (±0.001)
Cox-Time	0.629 (±0.006)	0.637 (±0.001)	**0.667 (±0.001)**	0.709 (±0.002)	0.792 (±0.001)
NNet	0.561 (±0.006)	0.528 (±0.001)	0.606 (±0.001)	0.599 (±0.005)	0.643 (±0.003)
PMF	0.601 (±0.006)	0.604 (±0.001)	0.657 (±0.001)	0.693 (±0.002)	0.773 (±0.001)
DeepHit	0.619 (±0.006)	**0.668** (±0.001)	0.655 (±0.001)	0.703 (±0.002)	0.784 (±0.001)
SurvivalNet	0.561 (±0.007)	0.628 (±0.001)	0.661 (±0.001)	0.692 (±0.002)	0.790 (±0.001)

TABLE 4. Results comparison with non-deep learning survival prediction approaches in Sarcoma-Rad dataset.

Proposed	CPH	SVM	Survival Tree	Survival Forest
0.657 (±0.006)	0.592 (±0.007)	0.599 (±0.006)	0.526 (±0.006)	0.5900 (±0.006)

TABLE 5. Projection loss regularized network comparison with states-of-the-art deep survival baselines in Sarcoma-Rad dataset.

Proposed	Deepsurv	Cox-Time	NNet	PMF	DeepHit	SurvivalNet
0.657 (±0.006)	0.631 (±0.006)	0.628 (±0.007)	0.568 (±0.007)	0.573 (±0.007)	0.640 (±0.006)	0.558 (±0.007)

TABLE 6. Projection loss regularized network comparison with states-of-the-art deep survival baselines in Sarcoma-3DMRI dataset.

Proposed	Deepsurv	Cox-Time	NNet	PMF	DeepHit	SurvivalNet
0.630 (±0.008)	0.573 (±0.009)	0.584 (±0.009)	0.532 (±0.007)	0.555 (±0.009)	0.549 (±0.014)	0.522 (±0.013)

between training data and network parameters sizes is the primary factor in the failed generalization of unregularized deep learning baselines. Unlike public datasets in previous experiments, many real-world medical datasets such as our Sarcoma datasets are small and highly censored, which is detrimental to highly parameterized neural network training. In such a scenario, it is more likely for larger discrete-time networks to fail to generalize and capture proper information for the cross-cohort prediction. On the other hand, regression-based networks with fewer parameters outperformed the discretized alternatives, demonstrating the flexibility and robustness of our regularization to generalize across different data scenarios. It can be concluded that our regularized Deepsurv model further improves the regression-based survival prediction performance, especially for small datasets.

Nevertheless, this work focuses on C-index survival prediction performance with one type of input and one outcome. In real-world scenarios, there are needs for considering and integrating multiple input types simultaneously into making decisions, such as imaging scans to time-series signals. Adapting the prediction network to these grand challenges is the subject of our future work.

V. CONCLUSION

We successfully developed a novel regularization strategy that theoretically upgrades features quality and practically improves the robust performance of the Deepsurv network for survival prediction. The experiment results demonstrate the advantage of deep survival regression approaches over discrete-time networks and the generalizability of our method across datasets in medical applications. The success of our work demonstrates that deep survival regression performance can be further improved with applied insight from representation space instead of more parameterization and expansion of deep neural network architectures.

REFERENCES

[1] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, "Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach," *Statist. Med.*, vol. 17, no. 10, pp. 1169–1186, 1998.
[2] N. E. Breslow and N. Chatterjee, "Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis," *J. Roy. Stat. Soc., C (Appl. Statist.)*, vol. 48, no. 4, pp. 457–468, 1999.
[3] P. Ferdinand Christ, F. Ettlinger, G. Kaissis, S. Schlecht, F. Ahmaddy, F. Grun, A. Valentinitsch, S.-A. Ahmadi, R. Braren, and B. Menze, "SurvivalNet: Predicting patient survival from diffusion weighted magnetic resonance images using cascaded fully convolutional and 3D convolutional neural networks," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 839–843.

[4] A. L. Edwards, *Multiple Regression and the Analysis of Variance and Covariance*. New York, NY, USA: W. H. Freeman, 1985.
[5] B. Efron, "The efficiency of Cox's likelihood function for censored data," *J. Amer. Stat. Assoc.*, vol. 72, no. 359, pp. 557–565, Sep. 1977.
[6] M. F. Gensheimer and B. Narasimhan, "A scalable discrete-time survival model for neural networks," *PeerJ*, vol. 7, p. e6257, Jan. 2019.
[7] E. Giunchiglia, A. Nemchenko, and M. V. D. Schaar, "RNN-SURV: A deep recurrent model for survival analysis," in *Proc. Int. Conf. Artif. Neural Netw.*, Cham, Switzerland: Springer, 2018, pp. 23–32.
[8] P. L. Gradowska and R. M. Cooke, "Least squares type estimation for Cox regression model and specification error," *Comput. Statist. Data Anal.*, vol. 56, no. 7, pp. 2288–2302, 2012.
[9] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 24, Dec. 2018.
[10] H. Kvamme and Ø. Borgan, "Continuous and discrete-time survival prediction with neural networks," 2019, *arXiv:1910.06724*.
[11] H. Kvamme, Ø. Borgan, and I. Scheel, "Time-to-event prediction with neural networks and Cox regression," *J. Mach. Learn. Res.*, vol. 20, no. 129, pp. 1–30, 2019.
[12] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 2314–2321.
[13] D. Y. Lin, "On the breslow estimator," *Lifetime Data Anal.*, vol. 13, no. 4, pp. 471–480, Dec. 2007.
[14] L. Ohno-Machado, "Modeling medical prognosis: Survival analysis techniques," *J. Biomed. Inform.*, vol. 34, no. 6, pp. 428–439, 2001.
[15] J. C. Peeken, M. Bernhofer, M. B. Spraker, D. Pfeiffer, M. Devecka, A. Thamer, M. A. Shouman, A. Ott, F. Nüsslin, N. A. Mayr, B. Rost, M. J. Nyflot, and S. E. Combs, "CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy," *Radiotherapy Oncol.*, vol. 135, pp. 187–196, Jun. 2019.
[16] M. B. Spraker, L. S. Wootton, D. S. Hippe, K. C. Ball, J. C. Peeken, M. W. Macomber, T. R. Chapman, M. N. Hoff, E. Y. Kim, S. M. Pollack, S. E. Combs, and M. J. Nyflot, "MRI radiomic features are independently associated with overall survival in soft tissue sarcoma," *Adv. Radiat. Oncol.*, vol. 4, no. 2, pp. 413–421, Apr. 2019.
[17] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei, "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statist. Med.*, vol. 30, no. 10, pp. 1105–1117, May 2011.
[18] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, 2017.
[19] X. Zhu, J. Yao, and J. Huang, "Deep convolutional neural network for survival analysis with pathological images," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 544–547.
[20] S. Pölsterl, N. Navab, and A. Katouzian, "Fast training of support vector machines for survival analysis," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, Sep. 2015, pp. 243–259.
[21] M. Leblanc and J. Crowley, "Survival trees by goodness of split," *J. Amer. Stat. Assoc.*, vol. 88, no. 422, pp. 457–467, Jun. 1993.
[22] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 841–860, 2008.
[23] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *J. Stat. Softw.*, vol. 39, no. 5, p. 1, 2011.
[24] J. C. Peeken, M. B. Spraker, C. Knebel, H. Dapper, D. Pfeiffer, M. Devecka, A. Thamer, M. A. Shouman, A. Ott, R. von Eisenhart-Rothe, F. Nüsslin, N. A. Mayr, M. J. Nyflot, and S. E. Combs, "Tumor grading of soft tissue sarcomas using MRI-based radiomics," *EBioMedicine*, vol. 48, pp. 332–340, Oct. 2019.
[25] J. C. Peeken, R. Asadpour, K. Specht, E. Y. Chen, O. Klymenko, V. Akinkuoroye, D. S. Hippe, M. B. Spraker, S. K. Schaub, H. Dapper, C. Knebel, N. A. Mayr, A. S. Gersing, H. C. Woodruff, P. Lambin, M. J. Nyflot, and S. E. Combs, "MRI-based delta-radiomics predicts pathologic complete response in high-grade soft-tissue sarcoma patients treated with neoadjuvant therapy," *Radiotherapy Oncol.*, vol. 164, pp. 73–82, Nov. 2021.
[26] F. Navarro, H. Dapper, R. Asadpour, C. Knebel, M. B. Spraker, V. Schwarze, S. K. Schaub, N. A. Mayr, K. Specht, H. C. Woodruff, P. Lambin, A. S. Gersing, M. J. Nyflot, B. H. Menze, S. E. Combs, and J. C. Peeken, "Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using MR imaging," *Cancers*, vol. 13, no. 12, p. 2866, Jun. 2021.
[27] J.-B. Chen, H.-S. Yang, S.-H. Moi, L.-Y. Chuang, and C.-H. Yang, "Identification of mortality-risk-related missense variant for renal clear cell carcinoma using deep learning," *Therapeutic Adv. Chronic Disease*, vol. 12, Jan. 2021, Art. no. 204062232199262.
[28] R. W. Oei, Y. Lyu, L. Ye, F. Kong, C. Du, R. Zhai, T. Xu, C. Shen, X. He, L. Kong, C. Hu, and H. Ying, "Progression-free survival prediction in patients with nasopharyngeal carcinoma after intensity-modulated radiotherapy: Machine learning vs. traditional statistics," *J. Personalized Med.*, vol. 11, no. 8, p. 787, 2021.
[29] S.-S. Byun, T. S. Heo, J. M. Choi, Y. S. Jeong, Y. S. Kim, W. K. Lee, and C. Kim, "Deep learning based prediction of prognosis in nonmetastatic clear cell renal cell carcinoma," *Sci. Rep.*, vol. 11, no. 1, pp. 1–8, Dec. 2021.
[30] B. Kim, Y. J. Jang, H. R. Cho, S. Y. Kim, J. E. Jeong, M. K. Shim, and M. G. Kim, "Predicting completion of clinical trials in pregnant women: Cox proportional hazard and neural network models," *Clin. Transl. Sci.*, early access, Nov. 4, 2021, doi: 10.1111/cts.13187. [Online]. Available: https://ascpt.onlinelibrary.wiley.com/doi/pdfdirect/10.1111/cts.13187
[31] N. Bice, N. Kirby, T. Bahr, K. Rasmussen, D. Saenz, T. Wagner, N. Papanikolaou, and M. Fakhreddine, "Deep learning-based survival analysis for brain metastasis patients with the national cancer database," *J. Appl. Clin. Med. Phys.*, vol. 21, no. 9, pp. 187–192, 2020.
[32] M. Shu, R. S. Bowen, C. Herrmann, G. Qi, M. Santacatterina, and R. Zabih, "Deep survival analysis with longitudinal X-rays for COVID-19," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 4046–4055.
[33] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101789.
[34] L. Wei, D. Owen, B. Rosen, X. Guo, K. Cuneo, T. S. Lawrence, R. Ten Haken, and I. El Naqa, "A deep survival interpretable radiomics model of hepatocellular carcinoma patients," *Phys. Medica*, vol. 82, pp. 295–305, Feb. 2021.
[35] C. Nagpal, V. Jeanselme, and A. Dubrawski, "Deep parametric time-to-event regression with time-varying covariates," in *Proc. Mach. Learn. Res. (PMLR)*. May 2021, pp. 184–193. [Online]. Available: http://proceedings.mlr.press/v146/nagpal21a/nagpal21a.pdf
[36] D. Feng and L. Zhao, "BDNNSurv: Bayesian deep neural networks for survival analysis using pseudo values," 2021, *arXiv:2101.03170*.
[37] Z. Zhang, H. Chai, Y. Wang, Z. Pan, and Y. Yang, "Cancer survival prognosis with deep Bayesian perturbation cox network," *Comput. Biol. Med.*, Nov. 2021, Art. no. 105012, doi: 10.1016/j.compbiomed.2021.105012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521008064?casa_token=8Pb9bZwX8woAAAAA:yo2WzBDmSQ3bGa-ARntDo_mvNBVQexEJOyZKP4VELJ27N0H12SgO-uNQLd7g3EumVdzNCZDQyg
[38] C. Nagpal, X. Li, and A. Dubrawski, "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 8, pp. 3163–3175, Aug. 2021.
[39] L. Zhao and D. Feng, "Deep neural networks for survival analysis using pseudo values," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 11, pp. 3308–3314, Nov. 2020.
[40] Y. Zhang, E. M. Lobo-Mueller, P. Karanicolas, S. Gallinger, M. A. Haider, and F. Khalvati, "CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging," *BMC Med. Imag.*, vol. 20, no. 1, pp. 1–8, Dec. 2020.

PHAWIS THAMMASORN received the B.E. degree in computer engineering from Chiang Mai University, Chiang Mai, Thailand, in 2013, and the M.S. degree in computer science from the University of Southern California, Los Angeles, CA, USA, in 2016. He is currently pursuing the Ph.D. degree with the Department of Industrial Engineering, University of Arkansas, Fayetteville, AR, USA. His primary research interest includes deep learning approach for predictive analytics on multi-modal data. The topics of interest include, but are not limited to, machine learning, artificial intelligence, data mining, computer vision, and multimedia processing.

STEPHANIE K. SCHAUB received the M.D. degree from Florida Atlantic University. She is currently a Board-Certified Radiation Oncologist, who joined the Department of Radiation Oncology, School of Medicine, University of Washington, as an Assistant Professor, in 2020. She specializes in the treatment of adult sarcoma patients and pediatric tumors of all types. She completed her residency in radiation oncology at the University of Washington, where she was the Chief Resident during her final year, and Harvard's Brigham and Women's Hospital. Her research interests include developing novel, non-invasive biomarkers (such as imaging or blood sample signals) that can predict cancer outcomes or treatment-related toxicity to advance toward precision radiation therapy. She is a member of the Connective Tissue Oncology Society, the Children's Oncology Group, and the American Society of Radiation Oncology.

DANIEL S. HIPPE received the B.S. degree in electrical engineering and the M.S. degree in statistics from the University of Washington, in 2004 and 2011, respectively. He is currently a Statistician with the Clinical Biostatistics Group, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. He enjoys working with investigators across a wide variety of disciplines and continually learning about different areas of medicine.

MATTHEW B. SPRAKER received the bachelor's degree in biology from Indiana University, in 2005, and the M.D. and Ph.D. degrees from the University of Illinois at Chicago. He joined the Department of Radiation Oncology, School of Medicine, Washington University, as an Assistant Professor, in 2018, after completing a residency in radiation oncology at the University of Washington. His clinical interests include sarcoma, clinical informatics, biomedical imaging, patient safety, and quality improvement.

JAN C. PEEKEN received the M.D. degree from the University of Freiburg, Germany, with intermittent studies abroad at the University of Nice, France, and the Havard Medical School, USA, and the Habilitation degree from the Medical Faculty, Technical University of Munich, in 2020. After pursuing a residency in radiation oncology at the Klinikum rechts der Isar, university hospital of the Technical University of Munich, he is currently serving as an Attending Physician. He is also serving as the Group Leader with the Institute of Radiation Medicine, Helmholtz Zentrum München. In 2019, he visited the Department of Radiation Oncology, University of Washington, Seattle, as a Research Fellow. His research interest includes the applications of artificial intelligence techniques in radiation oncology.

LANDON S. WOOTTON was born in Austin, TX, USA, in 1987. He received the B.S. degree in physics from The University of Texas at Austin, in 2008, and the Ph.D. degree in medical physics from the Graduate School of Biomedical Sciences, The University of Texas, in 2014. From 2015 to 2017, he was a Medical Physics Resident with the University of Washington, where he stayed on as a Faculty, until 2020. He is currently a Medical Physicist with the Baylor Scott and White Health Center, Round Rock, TX, USA. His research interests include scintillation dosimetry, neutron therapy, automation, and advanced image analysis.

PAUL E. KINAHAN (Fellow, IEEE) is currently a member of the UW Imaging Research Laboratory. He was a part of the group that built the first prototype combined PET/CT scanner and has also contributed to the current class of data processing image reconstruction algorithms used in PET/CT oncology imaging. He moved to the University of Washington, in 2001, where he continues his research in PET/CT imaging. He has served on committees for RSNA, AAPM, SNM, NIH, and IEEE.

STEPHANIE E. COMBS received the Doctorate degree, with preclinical work in field of neuroanatomy, where she studied the sympathoadrenal system and the impact of growth factors on development and formation of the neuronal and non-neuronal networks. She studied medicine in Heidelberg; Norfolk, USA; and San Antonio, USA. After her graduation and promotion in 2003, she worked as a Research Associate in Heidelberg. Following her postdoctoral lecturer qualification in 2009, she was promoted to the Vice Chair of the Radiation Oncology Department in Heidelberg, in 2011. In 2014, she was appointed as a Professor and the Chair of the Department of Radiation Oncology, Technical University of Munich (TUM). In 2015, she also took over the Institute of Radiation Medicine, Helmholtz Zentrum. Since 2019, she has been heading the TUM Senate and Serves as the Vice Dean for diversity and talent management. Her key expertise is highly conformal radiation therapy (stereotactic treatment, IMRT/IGRT/ART, protons, and carbon ions). Her scientific work includes areas, including treatment optimization for brain and skull base tumors, biomarkers in radiation oncology, pediatric oncology, gastrointestinal oncology, uro-oncology, gynecological oncology, radiochemotherapy, and radioimmunotherapy. She is serving as a Board Member for the Neurooncology Working Group of the German Cancer Society (DKG). She received several scientific awards, such as the Hermann Holthusen Prize of the German Radiation Oncology Society.

WANPRACHA A. CHAOVALITWONGSE received the Ph.D. degree in industrial and systems engineering from the University of Florida, Gainesville, FL, USA, in 2003. He was a Professor of industrial and systems engineering and radiology (joint) with the University of Washington, Seattle, WA, USA. He served as an Associate Director for the Integrated Brain Imaging Center, University of Washington Medical Center, Seattle. He currently serves as a Research Professor with the Department of Industrial Engineering, University of Arkansas, Fayetteville, AR, USA. He was a recipient of several awards for his research, such as the NSF CAREER Award in 2005 and the William Pierskalla Best Paper for Research Excellence in Operations Research and Health Care Applications by the Institute of Operations Research and the Management Sciences twice in 2004 and 2008.

MATTHEW J. NYFLOT received the Ph.D. degree in medical physics from the University of Wisconsin–Madison, in 2011. He was certified as a Diplomate of the American Board of Radiology in therapeutic medical physics, in 2014. He is currently a Medical Physicist and an Associate Professor with the Department of Radiation Oncology, with an adjunct appointment with the Department of Radiology, University of Washington, Seattle, WA, USA. His current research interests include the application of data science for precision cancer therapy, and safety and quality in radiation therapy.