ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230410

Towards Flexible Time-to-Event Modeling: Optimizing Neural Networks via Rank Regression

Hyunjun Lee[a,*], Junhyun Lee[b,*], Taehwa Choi[c], Jaewoo Kang[b,†] and Sangbum Choi[d,†]

[*]Equal contribution
[a]Digital Tech. Center, SK Inc. C&C
[b]Department of Computer Science and Engineering, Korea University
[c]Department of Biostatistics and Bioinformatics, Duke University
[d]Department of Statistics, Korea University
[†]Corresponding authors. Email: kangj@korea.ac.kr, choisang@korea.ac.kr
ORCiD ID: Jaewoo Kang https://orcid.org/0000-0001-6798-9106,
Sangbum Choi https://orcid.org/0000-0001-6983-5821

Abstract. Time-to-event analysis, also known as survival analysis, aims to predict the time of occurrence of an event, given a set of features. One of the major challenges in this area is dealing with censored data, which can make learning algorithms more complex. Traditional methods such as Cox's proportional hazards model and the accelerated failure time (AFT) model have been popular in this field, but they often require assumptions such as proportional hazards and linearity. In particular, the AFT models often require pre-specified parametric distributional assumptions. To improve predictive performance and alleviate strict assumptions, there have been many deep learning approaches for hazard-based models in recent years. However, representation learning for AFT has not been widely explored in the neural network literature, despite its simplicity and interpretability in comparison to hazard-focused methods. In this work, we introduce the Deep AFT Rank-regression model for Time-to-event prediction (*DART*). This model uses an objective function based on Gehan's rank statistic, which is efficient and reliable for representation learning. On top of eliminating the requirement to establish a baseline event time distribution, *DART* retains the advantages of directly predicting event time in standard AFT models. The proposed method is a semiparametric approach to AFT modeling that does not impose any distributional assumptions on the survival time distribution. This also eliminates the need for additional hyperparameters or complex model architectures, unlike existing neural network-based AFT models. Through quantitative analysis on various benchmark datasets, we have shown that *DART* has significant potential for modeling high-throughput censored time-to-event data.

1 Introduction

Time-to-event analysis, also known as survival or failure time analysis, is a widely used statistical method in fields such as biostatistics, medicine, and economics to estimate either risk scores or the distribution of event time, given a set of features of subjects [43, 8, 11, 31]. While assessing risk and quantifying survival probabilities have benefits, time-to-event analysis can be challenging due to the presence of censoring. In real-world studies, subjects (e.g. patients in medical research) may be dropped out before the event of interest (e.g. death) occurs, which can prevent full follow-up of the data [30]. The presence of censoring in survival data can create a serious challenge in applying standard statistical learning strategies. In general, the censoring process is assumed to be non-informative in that it is irrelevant of the underlying failure process given features, but their relationship should be properly accounted for, otherwise leading to biased results.

Cox's proportional hazards (CoxPH) model is the most well-known method for time-to-event data analysis. It specifies the relationship between a conditional hazard and given features in a multiplicative form by combining the baseline hazard function with an exponentiated regression component, allowing for the estimation of relative risks. However, this model requires so-called the proportional hazards assumption and time-invariant covariate-effects, which can be difficult to verify in many applications [2]. Statistical testing procedures, such as Schoenfeld's test, are typically conducted to examine the PH assumptions, as they are often vulnerable to violation of underlying assumptions [3, 23].

The accelerated failure time (AFT) model, also known as the accelerated life model, relates the logarithm of the failure time to features in a linear fashion [28]. This model has been used as an attractive alternative to the CoxPH model for analyzing censored failure time data due to its natural physical interpretation and connection with linear models. Unlike the CoxPH models, the classical parametric AFT model assumes the underlying time-to-event distribution can be explained with a set of finite-dimensional parameters, such as Weibull or log-normal distribution. However, such assumption on the failure time variable can be restrictive and may not accurately reflect real-world data. This can decrease performance of the AFT model compared to Cox-based analysis, making it less attractive for practical use [9, 23]. Recently, researchers have been exploring a range of time-to-event models that leverage statistical theories and deep learning techniques to circumvent the necessity of assumptions such as linearity, single risk, discrete time, and fixed-time effect [22, 26, 37, 24, 5, 41, 36].

In particular, *Cox-Time* [25] and *DATE* [7] alleviate some of the strict assumptions of the CoxPH and parametric AFT models by allowing non-proportional hazards and non-parametric event-time distribution, respectively. *Cox-Time* utilizes neural networks as a relative risk model to access interactions between time and features. The authors also show that their loss function serves as a good approximation of the Cox partial log-likelihood. *DATE* is a conditional generative adversarial network that implicitly specifies a time-to-event distribution within the AFT model framework. It does not require a pre-specified distribution in the form of a parameter, instead the generator can learn from the data using an adversarial loss function. Incidentally, various deep learning-based approaches have been proposed to improve the performance by addressing issues such as temporal dynamics and model calibration [27, 34, 13, 20, 17]. These approaches have highlighted the importance of utilizing well-designed objective functions that not only take into account statistical properties but also optimize neural networks.

In this paper, we introduce a Deep AFT Rank-regression for Time-to-event prediction model (*DART*), a deep learning-based semiparametric AFT model, trained with an objective function originated from Gehan's rank statistic. We eliminate the need for specifying a baseline event time distribution while still preserving the advantages of AFT models that directly predict event times. By constructing comparable rank pairs in the simple form of loss functions, the optimization of *DART* is efficient compared to other deep learning-based event time models. Our experiments show that *DART* not only calibrates well, but also competes in its ability to predict the sequence of events compared to risk-based models. Furthermore, we believe that this work can be widely applied in the community while giving prominence to the advantages of AFT models which are relatively unexplored compared to the numerous studies on hazard-based models.

2 Related Works

We first overview time-to-event modeling focusing on the loss functions of *Cox-Time* and *DATE* models to highlight the difference in concepts before introducing our method. The primary interest of time-to-event analysis is to estimate survival quantities like survival function $S(t) = P(T \geq t)$ or hazard function $h(t) = \lim_{\delta \to 0} P(t \leq T \leq t+\delta | T \geq t)/\delta$, where $T \in \mathbb{R}^+$ denotes time-to-event random variable. In most cases, due to censored observations, those quantities cannot be directly estimated with standard statistical inference procedure. In the presence of right censoring, Kaplan and Meier and Aalen provided consistent nonparametric survival function estimators, exploiting right-censoring time random variable $C \in \mathbb{R}^+$. Researchers then can get stable estimates for survival quantities with data tuples $\{y_i, \delta_i, X_i\}_{i=1}^N$, where $y_i = \min(T_i, C_i)$ is the observed event time with censoring, $\delta_i = I(T_i \leq C_i)$ is the censoring indicator, and a vector of features $X_i \in \mathbb{R}^P$. Here, N and P denote the number of instances and the number of features, respectively. While those nonparametric methods are useful, one can improve predictive power by incorporating feature information in a way of regression modeling. Cox proportional-hazards (CoxPH) and accelerated-failure-time (AFT) frameworks are the most common approaches in modeling survival quantities utilizing both censoring and features.

2.1 Hazard-Based Models

A standard CoxPH regression model [10] formulates the conditional hazard function as:

$$h(t|X_i) = h_0(t)\exp(\beta^T X_i),\ (i = 1, \dots, N), \tag{1}$$

where $h_0(\cdot)$ is an unknown baseline hazard function which has to be estimated nonparametrically, and $\beta \in \mathbb{R}^P$ is the regression coefficient vector. It is one of the most celebrated models in statistics in that β can be estimated at full statistical efficiency while achieving nonparametric flexibility on h_0 under the proportionality assumption. Note the model is semiparametric due to the unspecified underlying baseline hazard function h_0. Letting $\mathcal{R}_i$ be the set of all individuals "at risk", meaning that are not censored and have not experienced the event before T_i, statistically efficient estimator for regression coefficients can be obtained minimizing the loss function with respect to β:

$$L_{\text{CoxPH}}(\beta) = \sum_i \delta_i \log \left(\sum_{j \in \mathcal{R}_i} \exp\left[\beta^T X_j - \beta^T X_i\right] \right) \tag{2}$$

which is equivalent to the negative partial log-likelihood function of CoxPH model.

Based on this loss function, Kvamme et al. proposed a deep-learning algorithm for the hazard-based predictive model, namely *Cox-Time*, replacing $\beta^T X_j$ and $\beta^T X_i$ with $g(y_j, X_j; \theta)$ and $g(y_i, X_i; \theta)$, respectively. Here, $g(\cdot)$ denotes the neural networks parameterized by θ, and $\mathcal{R}_i$ would be replaced by $\tilde{\mathcal{R}}_i$, representing the sampled subset of $\mathcal{R}_i$. With a simple modification of the standard loss function in Eq. (2), *Cox-Time* can alleviate the proportionality for relative risk, showing empirically remarkable performance against other hazard-based models in both event ordering and survival calibration.

2.2 Accelerated-Failure-Time Models

The conventional AFT model relates the log-transformed survival time to a set of features in a linear form:

$$\log T_i = \beta^T X_i + \epsilon_i,\ (i = 1, \dots, N), \tag{3}$$

where ϵ_i is an independent and identically distributed error term with a common distribution function $F_0(\cdot)$ that is often assumed to be Weibull, exponential, log-normal, etc. As implied in Eq. (3), AFT model takes a form of linear modeling and provides an intuitive and physical interpretation on event time without detouring via the vague concept of hazard function, making it a powerful alternative to hazard-based analysis. However, imposing a parametric distributional assumption for ϵ_i is a critical drawback of the model, for which model in Eq. (3) could be a subclass of the hazard-based models.

To alleviate linearity and parametric distributional assumptions, several works brought the concept of generative process and approximated the error distribution via neural networks like generative adversarial networks (GANs) [33, 7]. Especially, Chapfuwa et al. proposed a deep adversarial time-to-event (*DATE*) model, which specifies the loss function as:

$$\begin{aligned} L_{\text{DATE}}(\theta, \phi) = & \ \mathbb{E}_{(X,y)\sim F_{nc}}[D_\phi(X, y)] \\ & + \mathbb{E}_{X\sim F_{nc}, \xi\sim F_\xi}[1 - D_\phi(X, G_\theta(X, \xi; \delta = 1))] \\ & + \lambda_2 \mathbb{E}_{(X,y)\sim F_c, \xi\sim F_\xi}[\max(0, y - G_\theta(X, \xi; \delta = 0))] \\ & + \lambda_3 \mathbb{E}_{(X,y)\sim F_{nc}, \xi\sim F_\xi}[\|t - G_\theta(X, \xi; \delta = 1)\|_1] \end{aligned} \tag{4}$$

where θ and ϕ denote the parameter set associated with a generator G_θ and a discriminator D_ϕ, respectively, (λ_2, λ_3) are hyperparameters to tune censoring trade-off, $F_{nc}(X, y)$ and $F_c(X, y)$ are empirical joint distributions for non-censored cases and censored cases, respectively, and F_ξ is the simple distribution, such as uniform distribution. The generator G_θ implicitly defines event time distribution. Despite *DATE* achieves prominent survival calibration via the sample-generating process, the objective function is quite complicated and the GAN framework is inherently prone to mode collapse, i.e., the generator learns only a few modes of the true distribution while missing other modes [40]. Also, when optimizing neural networks with multiple loss functions, it is difficult to balance and there might be conflicts (i.e. trade-off) with each term [12]. Therefore, their loss function might be difficult to be optimized as intended and requires a burdening training time, and consequently not be suitable for large-scale time-to-event analysis.

In the statistical literature, there have been many attempts to directly estimate regression coefficients in the semiparametric AFT model, where the error distribution F_0 is left unknown, rather than imposing specific parametric distribution or exploiting generative models. In this work, we bridge non-linear representation learning and an objective function for estimation of semiparametric AFT model, which is originated from Gehan's rank statistic. By extensive quantitative analysis, we have shown the beauty of simplicity and compatibility of rank-based estimation, along with outstanding experimental performance.

3 Method

In this section, we introduce the concept of *DART*, followed by predictive analysis for survival quantities. The conceptual differences with the other neural network-based AFT models are illustrated in Figure 1. The semiparametric AFT is distinct from a parametric version in that the error distribution function F_0 is left completely unknown like the baseline hazard function in the CoxPH. By further exploiting neural networks, we propose *DART* model that can be formulated as a generalization of model in Eq. (3):

$$\log T_i = g(X_i;\theta) + \epsilon_i,\ (i = 1, \dots, N), \tag{5}$$

where $g(X_i;\theta)$ denotes arbitrary neural networks with input feature vector X_i and a parameter set θ, outputting single scalar value as predicted log-scaled time-to-event variable. With this simple and straightforward modeling, *DART* entails several attractive characteristics over existing AFT-based models. First, the semiparametric nature of *DART* enables flexible estimation of error distribution, allowing improved survival prediction via neural network algorithms for F_0. Second, the restrictive log-linearity assumption of AFT model can be further alleviated by exploiting deep neural networks. Specifically, while standard AFT model relates time-to-event variable to feature variable in linear manner, deep learning is able to approximate any underlying functional relationship, lessening linearity restriction. Although *DART* still requires log-transformed time as a target variable, its deep neural network redeems the point with powerful representative performance supported by universal approximation theorems, enabling automated non-linear feature transformation [29, 38, 45].

3.1 Parameter Estimation via Rank-based Loss Function

In statistical literature, many different estimating techniques have been proposed for fitting semiparametric AFT model [42, 18, 19, 44]. Among them, we shall adopt the l_1-type rank-based loss function by taking into account the censoring information, which is efficient and suitable for stably fitting neural networks. Letting a residual term $e_i \equiv e_i(\theta) = \log y_i - g(X_i;\theta)$, the objective loss function for *DART* can be formulated as:

$$L_{\text{Rank}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_i(e_i - e_j)I\{e_i \geq e_j\}, \tag{6}$$

where $I(\cdot)$ is the indicator function that has value 1 when the condition is satisfied, otherwise 0. The estimator $\hat{\theta}$ can be obtained by minimizing the loss function with respect to model parameter set θ. Optimization of model parameters can be conveniently conducted via batched stochastic gradient descent (SGD). Notice that the loss function in Eq. (6) involves model parameter θ only, without concerning estimation of the functional parameter F_0, enabling flexible time-to-event regression modeling.

Strength of the loss function is theoretical consistency of optimization without requiring any additional settings. Let the neural network be expressed: $g(X_i;\phi,\beta) = \beta^T W_i$, where $W_i \in \mathbb{R}^K$ is transformed feature vector through hidden layers with parameter set ϕ, and $\beta \in \mathbb{R}^K$ is a parameter set of linear output layer. Then, it is easy to see that the following estimating function is the negative gradient of the loss function with respect to β:

$$\begin{aligned} U_{\text{Rank}}(\beta) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_i(W_i - W_j)I(\tilde{e}_i \leq \tilde{e}_j) \overset{\text{set}}{=} 0 \\ \tilde{e}_i &= \log y_i - \beta^T W_i. \end{aligned} \tag{7}$$

Eq. (7) is often called the form of Gehan's rank statistic [18], testing whether β is equal to true regression coefficients for linear model $\log T_i = \beta^T W_i + \epsilon_i$, and the solution to the estimating equation $\hat{\beta}$ is equivalent to the minimizer of Eq. (6) with respect to β. This procedure entails nice asymptotic results such as $\sqrt{n}$-consistency and asymptotic normality of $\hat{\beta}$ under the counting processes logic, assuring convergence of $\hat{\beta}$ towards true parameter β as the number of instances gets larger [42, 18]. Although these asymptotic results might not be directly generalized to the non-linear predictor function, we expect that hidden layers would be able to assess effective representations W_i with non-linear feature transformation, as evidenced by extensive quantitative studies. Note that, to keep theoretical alignment, it is encouraged to set the last layer as a linear transformation with an output dimension of 1 to mimic the standard linear model following non-linear representation. In addition, a robust estimation against outlying instances can be attained, depending rank of residual terms along with their difference.

3.2 Prediction of Survival Quantities

Predicted output $g(X_i;\hat{\theta})$ from trained *DART* model represents estimated expectation of $\log T_i$ conditional on X_i, i.e. mean log-transformed survival time with given feature information of ith instance. However, estimating survival quantities (e.g. conditional hazard function) cannot be directly done for AFT-based models. Instead, we utilize the Nelson-Aalen estimator [1], verified to be consistent under the rank-based semiparametric AFT model [35]. Define $N(t;\theta) = \sum_{i=1}^{N} N_i(t)$ and $Y(t) = \sum_{i=1}^{N} Y_i(t)$, where $N_i(t) = I(e_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(e_i > t)$ are the counting and the at-risk processes, respectively. Then the Nelson-Aalen estimator of $H_0(t)$ is defined by

$$\hat{H}_0(t) = \int_0^t \frac{I\{Y(u) > 0\}}{Y(u)} dN(u). \tag{8}$$

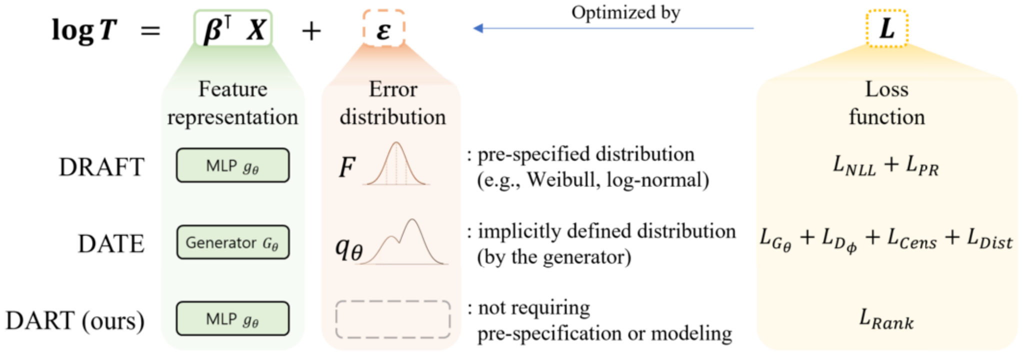
$\log T = \beta^{\top} X + \varepsilon$
Optimized by
L
Feature representation
Error distribution
Loss function
DRAFT
MLP g_θ
F
: pre-specified distribution (e.g., Weibull, log-normal)
$L_{NLL} + L_{PR}$
DATE
Generator G_θ
q_θ
: implicitly defined distribution (by the generator)
$L_{G_\theta} + L_{D_\phi} + L_{Cens} + L_{Dist}$
DART (ours)
MLP g_θ
: not requiring pre-specification or modeling
L_{Rank}

Figure 1. Illustration of conceptual differences between deep learning-based AFT models in terms of their respective contributions and required assumptions with a format of the standard AFT. To alleviate the parametric distribution assumption, which *DRAFT* has, *DATE* exploits the GAN framework and learns the implicit underlying distribution q_θ through the generator parameterized by θ. For *DRAFT*, L_{NLL} and L_{PR} denote negative log-likelihood and partial ranking likelihood, respectively. *DATE* basically requires four loss functions: L_{G_θ}, L_{D_ϕ} for the generator and the discriminator, L_{Cens} for adjusting censoring distribution, and L_{Dist} for the distortion penalty. Compared to the others, *DART* does not require pre-specification or modeling for error distribution and it is trained with a simple loss function supported by statistical theory.

The resulting conditional hazard function given X_i is defined by

$$\hat{h}(t|X_i) = \hat{h}_0[t\exp\{-g(X_i;\hat{\theta})\}]\exp\{-g(X_i;\hat{\theta})\}, \qquad (9)$$

where $\hat{h}_0(\cdot) = d\hat{H}_0(\cdot)$ is pre-trained baseline hazard function using Nelson-Aalen estimator. Consequently, conditional survival function can be estimated by relationship $\hat{S}(t|X_i) = \exp\{-\int_0^t \hat{h}(t|X_i)dt\}$, providing comparable predictions to other time-to-event regression models. In practice, training set is used to get pre-trained Nelson-Aalen estimator.

4 Evaluation Criteria

In this section, we evaluate models with two metrics for quantitative comparison: concordance index (CI) and integrated Brier score (IBS).

Concordance Index. Concordance of time-to-event regression model represents the proposition: if a target variable of instance i is greater than that of instance j, then the predicted outputs of i should be greater than that of j. By letting target variable y and predicted outcome $\hat{y}$, concordance probability of survival model can be expressed as $P(\hat{y}_i > \hat{y}_j | y_i > y_j)$, and concordance index measures the probability with trained model for all possible pairs of datasets [16]. With non-proportional-hazards survival regression models like *Cox-Time* or Lee et al. [26], however, Harrell et al. [16] cannot be used to measure discriminative performance properly. For fair comparison of survival regression models, time-dependent concordance index [4], or C^{td} was used for those baseline models proposed by Kvamme et al. [25] to account for tied events. $C^{\text{td}} \in [0,1]$ can be regarded as AUROC curve for time-to-event regression model, denoting better discriminative performance for a value close to 1. Note that standard concordance index yields identical results with C^{td} for AFT-based models.

Integrated Brier Score. Graf et al. [14] introduced generalized version of Brier score [6] for survival regression model along with inverse probability censoring weight (IPCW), which can be described as:

$$\begin{aligned}\text{BS}(t) = &\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{S}(t|X_i)^2 I(y_i \le t, \delta_i = 1)}{\hat{G}(y_i)} \\ &+\frac{1}{N}\sum_{i=1}^{N}\frac{(1-\hat{S}(t|X_i))^2 I(y_i > t)}{\hat{G}(t)}\end{aligned} \qquad (10)$$

where $\hat{G}(t) = \hat{P}(C > t)$ is a Kaplan-Meier estimator for censoring survival function to assign IPCW. BS(t) measures both how well calibrated and discriminative is predicted conditional survival function: if a given time point t is greater than y_i, then $\hat{S}(t|X_i)$ should be close to 0. Integrated Brier score (IBS) accumulates BS for a certain time grid $[t_1, t_2]$:

$$\text{IBS} = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} BS(s)ds. \qquad (11)$$

If $\hat{S}(t|X_i) = 0.5$ for all instances, then IBS becomes 0.25, thus well-fitted model yields lower IBS. For experiments, time grids can practically be set to minimum and maximum of y_i of the test set, equally split into 100 time intervals.

5 Experiments

In this section, we describe our experiment design and results to validate performance of *DART* compared to other time-to-event regression models. We conduct experiments using four real-world survival datasets and baseline models provided by Kvamme et al. and Chapfuwa et al. with two evaluation metrics mentioned in previous section.

5.1 Datasets

We use three benchmark survival datasets and a single large-scale dataset provided by Kvamme et al.. The descriptive statistics are provided in Table 1. Specifically, three benchmark survival datasets include: the Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT), the Assay of Serum Free Light Chain (FLCHAIN), and the Rotterdam Tumor Bank and German Breast Cancer Study Group (GBSG). Of particular interest is the GBSG dataset, which includes an indicator variable for hormonal therapy, allowing us to evaluate the effectiveness of a treatment recommendation system built using survival regression models. In addition, we use the large-scale WSDM KKBox dataset containing more than two millions of instances for customer churn prediction, which was prepared for the 11th ACM International Conference on Web Search and Data Mining. With a large-scale dataset, we can clearly verify the consistency of predictive performance of time-to-event models.

Table 1. Summary of survival datasets.

DATASET	SIZE	# FEATURES	% CENSORED
WSDM KKBOX	2,646,746	15	0.28
SUPPORT	8,873	14	0.32
FLCHAIN	6,524	8	0.70
GBSG	2,232	7	0.43

Table 2. Hyperparameter search space for the WSDM KKBox dataset.

Hyperparameter	Values
# Layers	{4,6,8}
# Nodes per layer	{128, 256, 512}
Dropout	{0.0, 0.1, 0.5}

Table 3. Hyperparameter search space for GBSG, FLCHAIN, and SUPPORT datasets.

Hyperparameter	Values
# Layers	{1, 2, 4}
# Nodes per layer	{64, 128, 256, 512}
Dropout	{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7}
Weight decay	{0.4, 0.2, 0.1, 0.05, 0.02, 0.01, 0.001}
Batch size	{64, 128, 256, 512, 1024}
λ (CoxTime and CoxCC)	{0.1, 0.01, 0.001, 0.0}

5.2 Baseline Models

We select six neural network-based time-to-event regression models as our experimental baselines: *DRAFT* and *DATE* [7] as AFT-based models for direct comparison with our model, and *DeepSurv* [22], *Cox-CC* and *Cox-Time* [25] as hazard-based models.

For AFT-based models, *DRAFT* utilizes neural networks to fit log-normal parametric AFT model in non-linear manner. However, it should be noted that if the true error variable does not follow a log-normal distribution, this model may be misspecified. In contrast, *DATE* exploits generative-adversarial networks (GANs) to learn conditional time-to-event distribution and censoring distribution using observed dataset. The generator's distribution of *DATE* can be trained from data to implicitly encode the error distribution. One major benefit of this approach is that it eliminates the need to pre-specify the parameters of the distribution. Despite these advantages, the method is challenging to apply to real-world datasets due to the complexity of the training procedure and objective function.

In case of hazard-based models, *DeepSurv* fits Cox regression model whose output is estimated from neural networks. The model outperforms the standard CoxPH model in performance, not clearly exceeding other neural network-based models. Furthermore, the proportional hazards assumption still remains unsolved with *DeepSurv*. *Cox-CC* is another neural network-based Cox regression model, using case-control sampling for efficient estimation. While both *DeepSurv* and *Cox-CC* are bounded to proportionality of baseline hazards, *Cox-Time* relieves this restriction using event-time variable to estimate conditional hazard function.

In this study, we focus on neural network-based models and exclude other machine learning-based models from the comparison to avoid redundant analysis that was conducted in previous studies. Some neural network-based models are also excluded as we aim to alleviate fundamental assumptions such as proportionality and parametric distribution. Note that comparing hazard-based models and AFT-based models has been rarely studied due to their difference in concept and purpose: modeling hazard function and modeling time-to-event variable. While models can be evaluated using common metrics, it is important to conduct a thorough analysis when analyzing numerical experiments, particularly when comparing hazard-based models and AFT-based models. These two types of models have different underlying concepts and purposes, making it crucial to take into consideration their unique characteristics during the analysis.

5.3 Model Specification and Optimization Procedure

For a fair comparison, we apply neural network architecture used in Kvamme et al.: MLP with dropout and batch-normalization. Every dense blocks are set to have the equal number of nodes (i.e. the dimension of hidden representations), no output bias is utilized for output layer, and ReLU function is chosen for non-linear activation for all layers. Preprocessing procedure has also been set based on Kvamme et al. including standardization of numerical features, entity embeddings [15] for multi-categorical features. The dimension of entity embeddings is set to half size of the number of categories. In addition, due to the fact that parameters of AFT-based models tend to be influenced by scale and location of the target variable, y has been standardized and its mean and variance are separately stored to rescaled outputs.

DeepSurv, Cox-CC, Cox-Time, DART. The PyCox[1] python package provides the training codes for these models. For WSDM KKBox dataset, we repeat experiments 30 times with best configurations provided by Kvamme et al.. Because train/valid/test split of KKBox dataset is fixed, we don't perform a redundant search procedure. For the other datasets (SUPPORT, FLCHAIN, and GBSG), we perform 5-fold cross-validation as performed at Kvamme et al. because the size of datasets is relatively small. At each fold, the best configuration is selected among 300 combinations of randomly selected hyperparameters which are summarized in Table 3. We use AdamWR [32] starting with one epoch of an initial cycle and doubling the cycle length after each cycle. The batch size is set to 1024 and the learning rates are found by Smith as performed at Kvamme et al..

DRAFT, DATE. The implementation of *DATE*[2] by the authors includes the code of *DRAFT* as well. We utilize their official codes for all datasets. The batch size for KKBox dataset is set to 8192 because the experiments are not feasible with the batch size 1024 due to their training time. The best configurations of *DRAFT* and *DATE* also are founded by grid search with same hyperparameter search space

[1] https://github.com/havakv/pycox

[2] https://github.com/paidamoyo/adversarial_time_to_event

Table 4. Mean and standard deviation of C^{td}. HAZ and AFT denote hazard-based and AFT-based methods, repectively.

	MODEL	WSDM KKBox	SUPPORT	FLCHAIN	GBSG
HAZ	*DeepSurv*	0.841 (0.000)	0.619 (0.008)	0.797 (0.013)	0.685 (0.013)
	Cox-CC	0.836 (0.046)	0.618 (0.009)	0.797(0.013)	0.684 (0.012)
	Cox-Time	**0.853** (0.049)	**0.637** (0.009)	**0.800** (0.012)	**0.687** (0.012)
AFT	*DRAFT*	0.861 (0.005)	0.599 (0.018)	0.725 (0.057)	0.611 (0.016)
	DATE	0.852 (0.001)	0.608 (0.008)	0.784 (0.009)	0.598 (0.034)
	DART (ours)	**0.867** (0.001)	**0.624** (0.009)	**0.797** (0.014)	**0.687** (0.014)

Table 5. Mean and standard deviation of Integrated Brier Score (IBS).

	MODEL	WSDM KKBox	SUPPORT	FLCHAIN	GBSG
HAZ	*DeepSurv*	0.111 (0.000)	**0.190** (0.004)	**0.101** (0.006)	**0.174** (0.004)
	Cox-CC	0.115 (0.012)	0.191 (0.003)	0.122 (0.028)	0.177 (0.004)
	Cox-Time	**0.107** (0.009)	0.194 (0.006)	0.114 (0.016)	**0.174** (0.005)
AFT	*DRAFT*	0.147 (0.002)	0.314 (0.043)	0.144 (0.022)	0.310 (0.010)
	DATE	0.131 (0.002)	0.227 (0.004)	0.124 (0.012)	0.204 (0.004)
	DART (ours)	**0.108** (0.001)	**0.176** (0.005)	**0.068** (0.007)	**0.150** (0.023)

with others. We repeat experiments 30 times with the best configuration as mentioned above. For the other datasets, as same with other models, we perform 5-fold cross-validation and choose the best configuration among 300 random hyperparameter sets at each fold. The hyperparameter search space for WSDM KKBox dataset is summarized in Table 2. Our implementation for *DART* is publicly available at: https://github.com/teboozas/dart_ecai23.

5.4 *Performance Evaluation*

To measure discriminative performance of outputs, we exploit standard C-index [16] for AFT-based models while letting hazard-based models to utilize C^{td} since equivalent evaluation is possible for AFT-based models including *DART* since it outputs a single scalar value to evaluate ranks. In terms of survival calibration, we implement our own function to obtain IBS based on its definition, due to the fact that evaluation methods of the conditional survival function and IPCW provided by Kvamme et al. are not compatible with AFT-based models. Specifically, we first fit Kaplan-Meier estimator upon standardized training set, and subsequently evaluate conditional survival estimates and IPCW utilizing estimated residuals, following the definition of baseline hazard function of AFT framework rather than to use time-to-event variable directly. For numerical integration, we follow settings of time grid from Kvamme et al., and standardize the grid with mean and standard deviation stored with standardization procedure of training set. By doing so, IBS can be compared upon identical timepoints for both hazard-based models and AFT-based models.

5.5 *Summary of Results*

Experiment results are provided in Table 4 and 5. In summary, *DART* is competitive in both discriminative and calibration performance, especially for large-scale survival datasets. Specifically, *DART* yields consistent results for WSDM KKBox dataset compared to other baselines, maintaining competitive performance in terms of C^{td} and IBS. We point out that *DART* is the most powerful and AFT-based time-to-event model that can be a prominent alternative when hazard-based models might be not working.

6 Analysis

We provide analysis on experimental results, pointing out strengths of *DART* model in terms of performance metrics.

Characteristic of *DART* for large-scale dataset. As provided in Table 4 and 5, *DART* generally yields prominent survival calibration performance with small variance in terms of IBS. Especially for large-scale dataset (KKBox), *DART* shows state-of-the-art performance with the smallest variance in evaluated metrics. This result comes from the characteristic of rank-based estimation strategy. Specifically, on the basis of asymptotic property of Eq. (7), estimated model parameters get stable and close to true parameter set, when the size of dataset gets larger. Thus, once the trained model attains effective representation (W_i in Eq. (7)) from hidden layers via stochastic optimization methods, *DART* is able to provide stable outputs with strong predictive power, without sophisticated manipulation upon time-to-event distribution.

Comparison with AFT-based models. In case of *DRAFT*, model does not generally perform well for both C^{td} and IBS for most datasets. This is attributed to the fact that *DRAFT* is a simple extension of the parametric AFT model with log-normality assumption. Thus, this approach is quite sensitive to true underlying distribution of dataset. On the other hand, *DATE* yields clearly improved performance against *DRAFT* especially for survival calibration in terms of IBS. Unlike *DRAFT*, *DATE* utilizes GAN to learn conditional error distribution without parametric assumption, allowing the model to yield more precise survival calibration. However, time-to-event distribution is trained with divided loss functions by optimizing two tuning hyperparameters in Eq. (4). This approach can be significantly affected by well-tuned hyperparameters and heavy computation is required to this end, resulting insufficient performance. Meanwhile,

Table 6. Comparison of the training time (seconds) per epoch over the KKBox dataset.

	DeepSurv	*Cox-CC*	*Cox-Time*	*DRAFT*	*DATE*	*DART (ours)*
Time	27.81	44.86	42.60	759.04	2024.19	29.93

as illustrated in Figure 1, *DART* has advantages of simplicity in theoretical and practical points compared to the other AFT-based models.

Comparison with hazard-based models. As previously reported by Kvamme et al., *Cox-Time* shows competitive performance against other hazard-based models, directly utilizing event-time variable to model conditional hazard function. However, we found out that *Cox-Time* requires precise tuning of additional hyperparameters (λ and Log-durations) largely affecting predictive performance.

In contrast, *DART* shows smaller variance in evaluation metrics as the size of data increases, ensuring stable output for large-scale dataset with asymptotic property which is crucial for practical application. In addition, while *Cox-Time* showed better performance against *DART* with respect to C-index in some cases, *DART* outperformed in *IBS*; gaining equivalent mean IBS scores against *Cox-Time* with smaller variance indicates our method dominant others in comprehensive survival metrics.

Comparison of the required time for optimizing each model. To verify the compatibility for large-scale data, we measure the training time of each model. We strictly bound the scope of the target process for a fair comparison, as from data input to parameter update excluding other extra steps. The specifications of all models are set equally: the number of nodes 256, the number of layers 6, and the batch size 1024. With the consumed time of 1000 iterations, we calculate the training time for a single epoch. We exclude the first iteration that is an outlier in general. All experiments were run on a single NVIDIA Titan XP GPU. Table 6 shows that the simplicity of *DART* leads to practical efficiency, while *DATE* is computationally expensive due to the generator-discriminator architecture.

Notes on practical impact of performance gain We acknowledge that practically interpreting IBS might be less intuitive and challenging. Thus we provide a simplified example below. Consider that a random-guess model, which estimates all conditional survival functions at 0.5, would result in an IBS score of 0.250. In this context, an improvement from 0.174 to 0.150, which are close to IBS of *Cox-Time* and *DART* for GBSG case respectively, of is indeed substantial.

Perfect estimation	Random guess	IBS
0	100	0.250
30	70	0.175
40	60	0.150

While the analogy below might not be entirely suitable, one can infer the practical improvement in survival estimation precision. It is worth to mention that the decreased IBS from 0.175 to 0.150 is still significant improvement in model accuracy, even though this comparison is somewhat rough. It represents an increase in *Perfect* estimations (i.e. $\hat{S}(t|X_i) = 0 \ \forall i$ s.t. $t \geq y_i, \delta_i = 1$) from 30 to 40 occurrences (+33.3%) out of 100 estimations. Considering the fact that $\hat{S}(t|X_i)$ ranges from 0 to 1, it is hard to achieve 0.025 points improvements in real settings. Consequently, while the differences in the metrics might appear moderate, we would like to emphasize that they are practically significant. Regarding the C-index, our model showed a 1.64% improvement compared to Cox-Time on the KKBox dataset, with scores rising from 0.853 to 0.867. Considering that a perfect score is 1.00, this implies that our model shows noticeable performance and exhibits high consistency In summary, *DART* would be the attractive alternative to existing time-to-event regression frameworks by ensuring remarkable performance and fast computation

7 Conclusion

In this work, we propose flexible time-to-event regression model, namely *DART*, utilizing the semiparametric AFT rank-regression method, coupled with deep neural networks, to alleviate strict assumptions and to attain practical usefulness in terms of high and stable predictive power. Extensive experiments have shown that our approach is prominent in discrimination and correction performance even on large-scale survival datasets. Although we do not yet address more complex censoring data such as competing risks and interval censoring, our approach can provide a stable baseline to handle these tasks in the near future with simple modifications of the loss function.

Acknowledgements

The research conducted by Sangbum Choi was supported by grants from the National Research Foundation of Korea (2022M3J6A1063595, 2022R1A2C1008514) and the Korea University research grant (K2305261). Jaewoo Kang and Junhyun Lee's research was supported by grants from the National Research Foundation of Korea (NRF-2023R1A2C3004176), the Korea Health Industry Development Institute (HR20C0021(3)), and the Electronics and Telecommunications Research Institute (RS-2023-00220195).

References

[1] O. Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978.

[2] O. O. Aalen. Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, 3(3):227–243, 1994.

[3] O. O. Aalen and H. K. Gjessing. Understanding the shape of the hazard rate: A process point of view (with comments and a rejoinder by the authors). *Statistical Science*, 16(1):1–22, 2001.

[4] L. Antolini, P. Boracchi, and E. Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

[5] A. Avati, T. Duan, S. Zhou, K. Jung, N. H. Shah, and A. Y. Ng. Countdown regression: sharp and calibrated survival predictions. In *UAI*, pages 145–155. PMLR, 2020.

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

[7] P. Chapfuwa, C. Tao, C. Li, C. Page, B. Goldstein, L. C. Duke, and R. Henao. Adversarial time-to-event modeling. In *ICML*, pages 735–744. PMLR, 2018.

[8] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen. Computer-aided diagnosis with deep learning architecture: applications to breast le-

sions in us images and pulmonary nodules in ct scans. *Scientific Reports*, 6(1):1–13, 2016.

[9] C. Cox. The generalized f distribution: an umbrella for parametric survival analysis. *Statistics in Medicine*, 27(21):4301–4312, 2008.

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

[11] L. Dirick, G. Claeskens, and B. Baesens. Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, 68(6):652–665, 2017.

[12] A. Dosovitskiy and J. Djolonga. You only train once: Loss-conditional training of deep networks. In *ICLR*, 2020.

[13] Y. Gao and Y. Cui. Multi-ethnic survival analysis: Transfer learning with cox neural networks. In *Survival Prediction-Algorithms, Challenges and Applications*, pages 252–257. PMLR, 2021.

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

[15] C. Guo and F. Berkhahn. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*, 2016.

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

[17] S. Hu, E. Fridgeirsson, G. van Wingen, and M. Welling. Transformer-based deep survival analysis. In *Survival Prediction-Algorithms, Challenges and Applications*, pages 132–148. PMLR, 2021.

[18] Z. Jin, D. Lin, L. Wei, and Z. Ying. Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353, 2003.

[19] Z. Jin, D. Lin, and Z. Ying. On least-squares regression with censored data. *Biometrika*, 93(1):147–161, 2006.

[20] F. Kamran and J. Wiens. Estimating calibrated individualized survival curves with deep learning. In *AAAI*, volume 35, pages 240–248, 2021.

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

[22] 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 Medical Research Methodology*, 18(1):24, 2018.

[23] D. G. Kleinbaum and M. Klein. *Survival Analysis*, volume 3. Springer, 2010.

[24] H. Kvamme and Ø. Borgan. Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*, 2019.

[25] H. Kvamme, Ø. Borgan, and I. Scheel. Time-to-event prediction with neural networks and cox regression. *JMLR*, 20(129): 1–30, 2019.

[26] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *AAAI*, pages 2314–2321, 2018.

[27] C. Lee, J. Yoon, and M. Van Der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2019.

[28] E. T. Lee and J. Wang. *Statistical Methods for Survival Data Analysis*, volume 476. John Wiley & Sons, 2003.

[29] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6): 861–867, 1993.

[30] K.-M. Leung, R. M. Elashoff, and A. A. Afifi. Censoring issues in survival analysis. *Annual Review of Public Health*, 18(1): 83–104, 1997.

[31] J. Li, H. Lu, C. Wang, W. Ma, M. Zhang, X. Zhao, W. Qi, Y. Liu, and S. Ma. A difficulty-aware framework for churn prediction and intervention in games. In *KDD*, pages 943–952, 2021.

[32] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[33] X. Miscouridou, A. Perotte, N. Elhadad, and R. Ranganath. Deep survival analysis: Nonparametrics and missingness. In *MLHC*, pages 244–256, 2018.

[34] C. Nagpal, S. Yadlowsky, N. Rostamzadeh, and K. Heller. Deep cox mixtures for survival regression. *MLHC*, 2021.

[35] Y. Park and L. Wei. Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika*, 90 (3):717–723, 2003.

[36] M. M. Rahman, K. Matsuo, S. Matsuzaki, and S. Purushotham. Deeppseudo: Pseudo value based deep learning models for competing risk analysis. In *AAAI*, volume 35, pages 479–487, 2021.

[37] K. Ren, J. Qin, L. Zheng, Z. Yang, W. Zhang, L. Qiu, and Y. Yu. Deep recurrent survival analysis. In *AAAI*, volume 33, pages 4798–4805, 2019.

[38] A. M. Schäfer and H. G. Zimmermann. Recurrent neural networks are universal approximators. In *ICANN*, pages 632–640. Springer, 2006.

[39] L. N. Smith. Cyclical learning rates for training neural networks. In *WACV*, pages 464–472. IEEE, 2017.

[40] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *NIPS*, pages 3310–3320, 2017.

[41] A. Tarkhan, N. Simon, T. Bengtsson, K. Nguyen, and J. Dai. Survival prediction using deep learning. In *Survival Prediction-Algorithms, Challenges and Applications*, pages 207–214. PMLR, 2021.

[42] A. A. Tsiatis. Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, pages 354–372, 1990.

[43] A. Viganò, M. Dorgan, J. Buckingham, E. Bruera, and M. E. Suarez-Almazor. Survival prediction in terminal cancer patients: a systematic review of the medical literature. *Palliative Medicine*, 14(5):363–374, 2000.

[44] D. Zeng and D. Lin. Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, 102(480):1387–1396, 2007.

[45] D.-X. Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2): 787–794, 2020.