DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks

Changhee Lee,[1] William R. Zame, [2] Jinsung Yoon, [1] Mihaela van der Schaar [3, 1, 4]
[1] Department of Electrical and Computer Engineering, University of California, Los Angeles, USA
[2] Department of Economics, University of California, Los Angeles, USA
[3] Department of Engineering Science, University of Oxford, UK
[4] Alan Turing Institute, London, UK
chl8856@ucla.edu, zame@econ.ucla.edu, jsyoon0823@ucla.edu, mihaela.vanderschaar@oxford-man.ox.ac.uk

Abstract

Survival analysis (time-to-event analysis) is widely used in economics and finance, engineering, medicine and many other areas. A fundamental problem is to understand the relationship between the covariates and the (distribution of) survival times (times-to-event). Much of the previous work has approached the problem by viewing the survival time as the first hitting time of a stochastic process, assuming a specific form for the underlying stochastic process, using available data to learn the relationship between the covariates and the parameters of the model, and then deducing the relationship between covariates and the distribution of first hitting times (the risk). However, previous models rely on strong parametric assumptions that are often violated. This paper proposes a very different approach to survival analysis, *DeepHit*, that uses a deep neural network to learn the distribution of survival times directly. DeepHit makes no assumptions about the underlying stochastic process and allows for the possibility that the relationship between covariates and risk(s) changes over time. Most importantly, DeepHit smoothly handles *competing risks*; i.e. settings in which there is more than one possible event of interest. Comparisons with previous models on the basis of real and synthetic datasets demonstrate that DeepHit achieves large and statistically significant performance improvements over previous state-of-the-art methods.

Introduction

Survival analysis – also called *time-to-event analysis* – is fundamental in many areas, including economics and finance, engineering and medicine. A long and diverse literature approaches survival analysis by viewing the event of interest as the first hitting time of an underlying stochastic process; i.e. the first time at which the stochastic process reaches a prescribed boundary. Depending on the context, the first hitting time may represent the time until a stock option can profitably be exercised, the time to failure of a mechanical system or the length of time a patient survives following treatment (or non-treatment); see (Lee and Whitmore 2006) for many other examples. A fundamental problem of survival analysis in all of these areas is to understand the relationship between the (distribution of) hitting times and the covariates, such as the characteristics of the stock on which the option is written, the physical environment in which the mechanical system must operate, and the features of the individual patient. Especially in medical setting, the survival analysis is further applied to discovering risk factors affecting the survival (Koene et al. 2016), comparison among risks of different subjects at a certain time of interest (Yoon et al. 2017), decision of a cost-efficient sensing period (e.g. screening for cancer) (Ahuja, Zame, and van der Schaar 2017).

Most of the previous work in this area has approached the problem by assuming a specific form for the underlying stochastic process, using available data to learn the relationship between the covariates and the parameters of the model, and then deducing the relationship between covariates and the distribution of first hitting times – the *risk* of the event. (In the medical setting, this is typically the risk of death or onset of a certain disease.) The Cox proportional hazards model (Cox 1972) is the most widely-used model in the medical setting but it makes many strong assumptions about the underlying stochastic process and about the relationship between the covariates and the parameters of that process. Other models allow for various other specific forms of the underlying stochastic process and for more general relationships between covariates and the parameters, but still maintain strong parametric assumptions (especially that the relationship between covariates and parameters of the stochastic process are time-invariant).

This paper proposes a very different approach to survival analysis: we construct and use a deep neural network that learns the distribution of first hitting times *directly*. An important aspect of our method, which we call *DeepHit*, is that it smoothly handles situations in which there is a single underlying risk (cause) and situations in which there are multiple *competing risks* (causes). DeepHit employs a network architecture that consists of a single shared sub-network and a family of cause-specific sub-networks. We train the network by using a loss function that exploits both survival times and relative risks. DeepHit makes *no assumptions* about the form of the underlying stochastic process; it therefore allows for the possibility that, even for a fixed cause or causes (e.g. a disease or diseases), both the parameters and the form of the stochastic process *depend on the covariates*.

Although our approach is quite general and applies to all the settings mentioned above, and many others, we focus here on the medical setting (and so we will use medical language, and speak of patients rather than instances, etc.). In the med-

ical context, competing risks are extremely common. (For example, patients suffering from a particular disease, such as cancer, frequently have co-morbidities, such as cardiovascular disease.) With the exception of the Fine-Gray model (Fine and Gray 1999), existing work on survival analysis either cannot be applied or is inadequate in the presence of competing risks except under the assumption that the risks are independent, which is very seldom the case. (To refer to the same example: studies (Koene et al. 2016) have shown that various treatments for breast cancer increase the risk of a cardiovascular event; the risks are not at all independent.) Survival analysis with competing risks is a challenging problem, and made all the more important because the choice of treatment must take account of these competing risks. We note that right-censoring of data is extremely common in the medical setting: patients are frequently lost to follow-up (often for unknown reasons).[1]

We are not the first to apply neural networks to time-to-event analysis; for example, (Faraggi and Simon 1995; Katzman et al. 2016; Luck et al. 2017) have employed neural networks for modeling non-linear representations for the relation between covariates and the risk of a clinical event. However, these studies have maintained the basic assumptions of the Cox model, weakening only the assumption of the form of the relationship between covariates and the hazard rate. In particular, the time-dependent influence of covariates on time-to-event cannot be addressed by these models.

To demonstrate the usefulness of our approach, we compare its predictive performance with that of competing approaches using three medical datasets and one synthetic dataset. For all these datasets, we compare the performance of DeepHit with previous state-of-the art competing methods, using as the metric of performance the time-dependent concordance index C^{td} (Antolini, Boracchi, and Biganzoli 2005). (C^{td} measures the extent to which the ordering of actual survival times of pairs agrees with the ordering of their predicted risk; it is the most-widely-used metric for evaluating the performance of survival models (Harrell et al. 1982).) DeepHit provides large and statistically significant performance improvements over previous state-of-the-art methods. (Detailed descriptions of these datasets, the competing methods, and the performance comparisons are presented in the following sections.)

Related Work

The survival model most widely used in the statistical and medical research literature is the Kaplan-Meier estimator (Kaplan and Meier 1958), which has the advantage of being able to learn very flexible survival curves, but the disadvantage of not incorporating patients' covariates. Hence it is useful at the population level but not useful at the individual level. As we have noted already the Cox proportional hazard model (Cox 1972) (CPH) is capable of incorporating patients' covariates, but assumes that the hazard rate is constant and that the log of the hazard rate is a linear function of covariates. Other models make different assumptions about the underlying stochastic processes and about the relationship between the covariates and the parameters of the assumed process. For instance (Lee and Whitmore 2010; Doksum and Hyland 1992) assume a Wiener process, while (Longini et al. 1989) assumes a Markov Chain; see (Lee and Whitmore 2010) for other examples and discussion of the literature. An advantage of these models is that, because they formulate survival analysis as the problem of determining the distribution of the first time at which the prescribed stochastic process hits a prescribed boundary, they are able to incorporate competing risks. The disadvantage of these models is that they are tied to the specific form of stochastic process that they assume. Put differently: the models are of limited use unless we have already learned the underlying stochastic process. In the medical setting this means learning the underlying disease process, which would seem to be an even more complicated problem than survival analysis itself – especially since the states of the disease or diseases are typically hidden and not directly observable. An alternative to this family of models is the one offered by (Fine and Gray 1999), which modifies the traditional proportional hazard model by direct transformation of the cumulative incidence function, but the Fine-Gray model is also severely limited by strong assumptions on the form of the hazard rates and on the way in which the parameters depend on covariates.

The problem of survival analysis has also received substantial recent attention in the machine learning literature. Recently developed survival models include random survival forests (Ishwaran et al. 2008), deep exponential families (R. Ranganath and Blei 2016), dependent logistic regressors (Yu et al. 2011), and semi-parametric Bayesian models based on Gaussian processes (Fernndez, Rivera, and Teh 2016). All of these methods are capable of incorporating the individual patient's covariates, but none of them has considered the problem of competing risks, and none of them seems readily adaptable to this problem. (In principle, these models could be applied to the problem of competing risks by fixing a single event and simply treating all other events right-censoring, but this approach is inadequate unless the competing risks are independent, which is frequently not the case.). Recently, deep multi-task Gaussian process was used to develop a nonparametric Bayesian model for survival analysis with competing risks (Alaa and van der Schaar 2017) while still relying on assumption that the latent stochastic process follows Gaussian process.

(Faraggi and Simon 1995) represents the first application of neural networks to survival analysis. In contrast to the standard CPH model, this work uses a feed-forward network to *learn* the relationship of the covariates to the hazard function. More recently, (Katzman et al. 2016) and (Luck et al. 2017) have followed the same general approach, although using more sophisticated network architectures and loss functions. These works have improved on the CPH model by relaxing the specific functional relationship between covariates and the hazard function in the standard CPH model while maintaining the other central assumption– that the hazard rate is constant over time. As a result, these works do not fully exploit the potential capacity of deep neural networks to learn complex representations of risk and in particular to capture

[1] Throughout this paper, we follow the literature and assume that right-censoring occurs completely at random.

the time-dependent influence of covariates on survival.

DeepHit improves on existing models because it suffers from none of the difficulties identified above. Because DeepHit learns the (joint) distribution of survival times and events directly, it avoids the problems inherent in assuming a particular form for the underlying stochastic process or a particular form for the relationship of covariates to the underlying stochastic process or any kind of time-invariance. As we shall see, the performance of DeepHit improves dramatically on the performance of previous models in the setting of competing risks and significantly even in the (simpler) setting of a single risk.

Survival Analysis

In this Section we describe our formal model.

Survival Data

Survival data provides three pieces of information for each instance/patient: 1) observed covariates, 2) time elapsed since covariates were first collected, and 3) a label indicating the type of event (e.g. adverse clinical event or death) that occurred.[2] We treat survival time as discrete and the time horizon as finite (e.g. no patients lived longer than 100 years) so the time set is $\mathcal{T} = \{0, \dots, T_{\max}\}$ for a predefined maximum time horizon $T_{\max}$. We consider $K \geq 1$ possible events of interest; we assume that at exactly one event eventually occurs for each instance/patient (e.g. a patient eventually dies, but can die from only one cause (Gooley et al. 1999)).[3] Because events of interest are not always observed (e.g. patients may be lost to follow-up), survival data are frequently right-censored; handling this difficulty will be a crucial aspect of the analysis. We indicate right-censoring as the "event" $\varnothing$ and therefore represent the set of possible events – including right-censoring – as $\mathcal{K} = \{\varnothing, 1, \cdots, K\}$. Each data point/instance (e.g. patient history) is therefore a triple $(\mathbf{x}, s, k)$ where $\mathbf{x} \in X$ is a D-dimensional vector of covariates, $s \in \mathcal{T}$ is the time at which the (unique) event or censoring occurred, and $k \in \mathcal{K}$ is the event or censoring that occurred at time s. Note that s is either the time at which an event (death) occurred or the time at which the patient was censored (disappeared from follow-up), but in either case the patient was known to be alive at times prior to s. We are given a dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, s^{(i)}, k^{(i)})\}_{i=1}^{N}$ that describe a finite set of observed instances/patients.

Figure 1 illustrates survival data of the SEER dataset (see Experiment section for more details) for 6 patients and two possible events (causes of death); patient 2 and 5 died from cause 1, patient 1 and 6 died from cause 2; patient 3 and 4 were lost to follow-up (right-censored).

For each tuple $(\mathbf{x}^*, s^*, k^*)$ with $k^* \neq \varnothing$, we are interested in the true probability $P(s = s^*, k = k^* | \mathbf{x} = \mathbf{x}^*)$; i.e. the true *ex-ante* probability that a (new) patient with covariates $\mathbf{x}^*$ will experience the event k^* at time s^*. Of course the true probability cannot be known on the basis of any finite dataset, so our task is to find *estimates* $\hat{P}$ of the true probabilities.

[2] We use medical terms for convenience but we emphasize that our framework and results are quite general.

[3] We leave for later work the more complicated setting in which several events – e.g. the onsets of various diseases – might occur.

		k	s	$\mathbf{x}$							
ID	Death Cause	Survival Time (m)	Lymphod Node	Age	Gender	Married	...	Benign Tumors	Malignant Tumors	Histology ICD	EOD
1	2	57	0.4061	53	1	1		0	2	65	0.0080
2	1	71	0.1382	56	1	1		0	2	64	0
3	ø	135	0.1600	60	1	1	...	0	2	65	0.0620
4	ø	120	0.2195	50	1	0		0	1	65	0
5	1	29	0.7998	55	1	1		0	2	64	0
6	2	71	0.7998	55	1	0		0	2	64	0

Figure 1: Illustration of survival data (SEER dataset).

Model Description

Our goal is to train the network to learn $\hat{P}$, the estimate of the joint distribution of the first hitting time and competing events. As illustrated in Figure 2, DeepHit is a multi-task network (Collobert and Weston 2008) which consists of a shared sub-network and K cause-specific sub-networks. Our architecture, differs from that of conventional multi-task network in two ways. First, we utilize a single softmax layer as the output layer of DeepHit in order to ensure that the network learns the *joint distribution* of K competing events not the *marginal distributions* of each event. Second, we maintain a residual connection (He et al. 2016) from the input covariates into the input of each cause-specific sub-network.

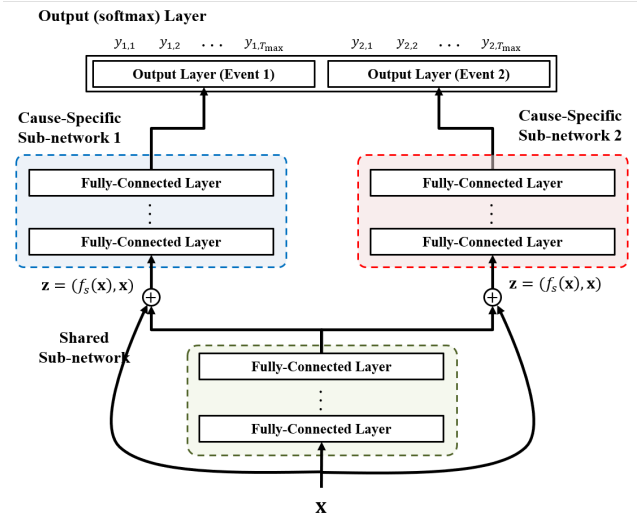

Figure 2: The architecture of DeepHit with two competing events.

The shared sub-network and the k-th cause-specific sub-network for $k = 1, \cdots, K$ are comprised of L_S and $L_{C,k}$ fully-connected layers, respectively. The shared sub-network takes as inputs the clinical covariates $\mathbf{x}$ and produces as output a vector $f_s(\mathbf{x})$ that captures the (latent) representation that is common to the K competing events.

Each cause-specific sub-network takes as inputs the pairs $\mathbf{z} = (f_s(\mathbf{x}), \mathbf{x})$ and produces as output a vector $f_{c_k}(\mathbf{z})$, which corresponds to the probability of the first hitting time of a specific cause k. More specifically, the inputs to the sub-networks include *both* the output of the shared network *and* the original covariates; this gives the sub-networks access to the learned common representation $f_s(\mathbf{x})$ while still allowing them to learn non-common part of the representation as well. (If only the learned common representation were used as an input to the sub-networks, the non-common part of the representation would be lost.) The totality of these outputs is a joint probability distribution on the first hitting time and event so the cause-specific sub-networks are learning the distribution for the first hitting time for each cause in parallel. The output of the softmax layer is a probability distribution $\mathbf{y} = [y_{1,1}, \cdots, y_{1,T_{\max}}, \cdots, y_{K,1}, \cdots, y_{K,T_{\max}}]$: given a patient with covariates $\mathbf{x}$, an output element $y_{k,s}$ is the (estimated) probability $\hat{P}(s, k|\mathbf{x})$ that the patient will experience the event k at time s. This architecture drives the network to learn potentially non-linear, even non-proportional, relationships between covariates and risks.

The *(cause-specific) cumulative incidence function* (CIF) expresses the probability that a particular event $k^* \in \mathcal{K}$ occurs on or before time t^* conditional on covariates $\mathbf{x}^*$; as in the Fine-Gray model (Fine and Gray 1999), understanding the CIF is key to the analysis of survival under competing risks. By definition, the CIF for the event k^* is:

$$\begin{aligned} F_{k^*}(t^*|\mathbf{x}^*) &= P(s \leq t^*, k = k^*|\mathbf{x} = \mathbf{x}^*) \\ &= \sum_{s^*=0}^{t^*} P(s = s^*, k = k^*|\mathbf{x} = \mathbf{x}^*). \end{aligned} \tag{1}$$

However, since the *true* CIF, $F_{k^*}(s^*|\mathbf{x}^*)$, is not known, we utilize the *estimated* CIF, $\hat{F}_{k^*}(s^*|\mathbf{x}^*) = \sum_{m=0}^{s^*} y^*_{k^*,m}$, in order to compare the risk of event occurring and to assess how models discriminate across cause-specific risks among patients.

Loss Function

To train DeepHit, we minimize a total loss function $\mathcal{L}_{\text{Total}}$ that is specifically designed to handle censored data. This loss function is the sum of two terms $\mathcal{L}_{\text{Total}} = \mathcal{L}_1 + \mathcal{L}_2$; $\mathcal{L}_1$ is the log-likelihood of the joint distribution of the first hitting time and event; $\mathcal{L}_2$ incorporates a combination of cause-specific ranking loss functions.

$\mathcal{L}_1$ is the log-likelihood of the joint distribution of the first hitting time and corresponding event, modified to take account of the right-censoring of the data (Lee and Whitmore 2006) considering K competing risks. For patients who are not censored, it captures both the event that has occurred and the time at which the event has occurred; for patients who are censored, it captures the time at which the patient is censored (lost to follow-up) which provides the information that the patient was alive up to that time. We define $\mathcal{L}_1$ by

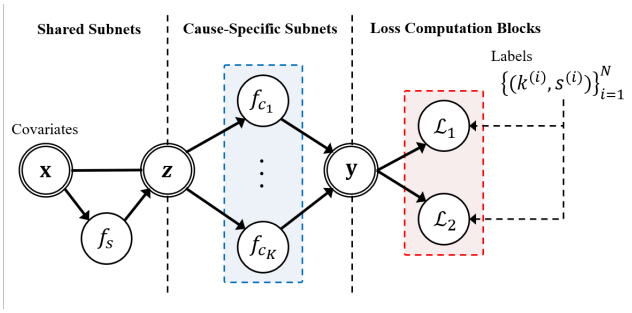

Figure 3: An illustration of a computational graph to compute the training loss of DeepHit.

$$\begin{aligned} \mathcal{L}_1 = -\sum_{i=1}^{N} \Big[&\mathbb{1}(k^{(i)} \neq \varnothing) \cdot \log\left(y^{(i)}_{k^{(i)}, s^{(i)}}\right) \\ &+ \mathbb{1}(k^{(i)} = \varnothing) \cdot \log\left(1 - \sum_{k=1}^{K} \hat{F}_k(s^{(i)}|\mathbf{x}^{(i)})\right) \Big], \end{aligned} \tag{2}$$

where $\mathbb{1}(\cdot)$ is an indicator function. The first term captures the information provided by uncensored patients; the second term captures the censoring bias by exploiting the knowledge that they are alive at the censoring time, so that that the first hitting event will occur among one of the K causes *after* the given censoring time); see (Lawless 2002).

$\mathcal{L}_1$ drives DeepHit to learn the general representation for the joint distribution of the first hitting time and events; $\mathcal{L}_2$ incorporates estimated CIFs calculated at different times (i.e. the time at which an event actualy occurs) in order to fine-tune the network to each cause-specific estimated CIF. To do so, we utilize a ranking loss function which adapts the idea of concordance (Harrell et al. 1982): a patient who dies at time s should have a higher risk at time s than a patient who survived longer than s. Write

$$A_{k,i,j} \triangleq \mathbb{1}(k^{(i)} = k, s^{(i)} < s^{(j)}), \tag{3}$$

for the indicator function of pairs (i, j) who experience risk k at different time, and whose risks for event k can therefore be directly compared; we call these pairs *acceptable for event* k. Now define

$$\mathcal{L}_2 = \sum_{k=1}^{K} \alpha_k \cdot \sum_{i \neq j} A_{k,i,j} \cdot \eta\left(\hat{F}_k(s^{(i)}|\mathbf{x}^{(i)}), \hat{F}_k(s^{(i)}|\mathbf{x}^{(j)})\right) \tag{4}$$

where the coefficients α_k are chosen to trade off ranking losses of the k-th competing event, and $\eta(x, y)$ is a convex loss function. For convenience, we assume here that the coefficients α_k are all equal (i.e. $\alpha_k = \alpha$ for $k = 1, \cdots, K$ and some α to be chosen), and we use the loss function $\eta(x, y) = \exp(\frac{-(x-y)}{\sigma})$. Incorporating $\mathcal{L}_2$ into the total loss function penalizes incorrect ordering of pairs (with respect to each event) and so minimizing the total loss encourages *correct ordering* of pairs (with respect to each event).

In Figure 3, we illustrate a computational graph to compute the training loss of the proposed network: the inputs

Table 1: Descriptive Statistics of Real-World Datasets for Competing Risks

Dataset		No. Uncensored	No. Censored	No. Features	Event Time			Censoring Time		
				(real, categorical)	min	max	mean	min	max	mean
SEER	CVD	903 (1.3%)	56,788 (83.1%)	23 (7,16)	0	176	79.8	0	179	144.6
	BC	10,634 (15.6%)			0	177	55.9			
UNOS		29,436 (48.7%)	30,964 (51.3%)	50 (17,33)	0	331	71.5	1	331	90.5
METABRIC		888 (44.8%)	1,093 (55.2%)	21 (6,15)	1	299	77.8	1	308	116.0

are the covariates $\mathbf{x}$ and the output is the vector $\mathbf{y}$. Double-circled nodes imply inputs or outputs of DeepHit or those of sub-networks, and single-circled nodes indicate calculation blocks (e.g. sub-networks or loss functions). In training stage, the network exploits $\{k^{(i)}, s^{(i)}\}_{i=1}^{N}$ in order to calculate the indicator functions, to find acceptable pairs, and, hence, to compute the loss function corresponding to input covariates. Based on this computational graph, we can obtain the gradient on the nodes (including hidden nodes of all the sub-networks) and parameters for training the proposed network.

Experiments

The prognostic performance of DeepHit was evaluated by comparing it with the performance of conventional benchmarks in analyzing three real-world clinical datasets and one synthetic dataset. We give brief descriptions of the datasets below; Table 1 gives more detail. Throughout the evaluations, we take 30 days = 1 month as the basic time interval.

UNOS The United Network for Organ Sharing (UNOS) database[4] consists of patients who underwent heart transplantation in the period 1985-2015. Of the total of 60,400 patients who received heart transplants, 29,436 patients (48.7%) were followed until death; the remaining 30,964 patients (51.3%) were right-censored. We used a total of 50 features (30 recipient-relevant, 9 donor-relevant and 11 donor-recipient compatibility). For details on selected features and pre-processing methods, see to (J. Yoon et al. 2017).

METABRIC The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset contains gene expression profiles and clinical features used to determine breast cancer subgroups. Of the total of 1,981 patients in the dataset, 888 patients (44.8%) were followed until death; the remaining 1,093 patients (55.2%) were right-censored. We restricted attention to 21 publicly available clinical features including tumor size, number of positive lymph nodes, etc.; for details see (Bilal et al. 2013). Missing values were replaced by the mean value for real-valued features and by the mode for categorical features. One-hot encoding was applied for categorical features.

SEER The Surveillance, Epidemiology, and End Results Program (SEER)[5] dataset provides information on breast cancer patients during the years 1992-2007. Among the 72,809 patients, we focused on 68,325 patients who died due to breast cancer or cardiovascular disease (CVD), or who were right-censored. (So we have two competing risks.) We have 23 patient features, including age, race, gender, morphology information, diagnostic information, therapy information, tumor size, tumor type, etc. Missing values were replaced by mean value for real-valued features and by the mode for categorical features.

SYNTHETIC We also created a synthetic dataset with two competing risks, in the spirit of (Alaa and van der Schaar 2017). To do this we constructed two stochastic processes with parameters and the hitting times described as follows:

$$\begin{aligned} &\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \mathbf{x}_3^{(i)} \sim \mathcal{N}(0, \mathbf{I}) \\ &T_1^{(i)} \sim \exp\left((\gamma_3^T \mathbf{x}_3^{(i)})^2 + \gamma_1^T \mathbf{x}_1^{(i)}\right) \\ &T_2^{(i)} \sim \exp\left((\gamma_3^T \mathbf{x}_3^{(i)})^2 + \gamma_2^T \mathbf{x}_2^{(i)}\right) \end{aligned} \tag{5}$$

where $\mathbf{x}^{(i)} = (\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \mathbf{x}_3^{(i)})$ is the vector of clinical covariates for patient i and consists of three 4-dimensional variables: for $k = 1, 2$, the covariates $\mathbf{x}_k$ only have an effect on the hitting time for event k while $\mathbf{x}_3$ has an effect on the hitting times of both events. Note that we assume hitting times are exponentially distributed with a mean parameter depending on both linear and non-linear (quadratic) function of covariates. For convenience, we set $\gamma_1 = \gamma_2 = \gamma_3 = 10$. Given the parameters, we first produced $30,000$ patients; among those, we randomly selected $15,000$ patients (50%) to be right-censored at a time $s_c^{(i)}$ randomly drawn from the uniform distribution on the interval $[0, \min\{T_1^{(i)}, T_2^{(i)}\}]$. (This censoring fraction was chosen to be roughly the same censoring fraction as in the real datasets, and hence to present the same difficulty as found in those datasets.) The data for each patient i is therefore $(\mathbf{x}^{(i)}, s^{(i)}, k^{(i)})$ where $s^{(i)} = \min\{T_1^{(i)}, T_2^{(i)}\}$ and $k^{(i)} = \arg\min T_k^{(i)}$ for patients who were not censored and $s^{(i)} = s_c^{(i)}$ and $k^{(i)} = \varnothing$ for patients who were censored.

Experimental Setting

For evaluation, we applied 5-fold cross validation: we randomly separated the data into training set (80%) and testing set (20%). We reserved 20% of the training set as a validation set. (In all of these sets, we maintained a constant ratio of patients who experienced each event and patients who were censored.) The hyper-parameters for $\mathcal{L}_{\text{Total}}$, including α and σ, were selected based on the discriminative performance on

[4] https://www.unos.org/data/

[5] https://seer.cancer.gov/causespecific/

the validation set. Early stopping was performed based on the total loss. DeepHit is a 4-layer network consisting of 1 fully-connected layer for the shared sub-network and 2 fully-connected layers for each cause-specific sub-network and a softmax layer as the output layer. (Note that if there is a single event, this reduces to 3 fully-connected layers and a softmax layer as the output layer.) For hidden layers, the number of nodes were set as 3, 5, and 3 times of the covariate dimension for the layer 1, 2, and 3, respectively, with ReLu activation function. The network was trained by back-propagation via Adam optimizer with a batch size of 50 and a learning rate of 10^{-4}. Dropout probability of 0.6 and Xavier initialization was applied for all the layers (DeepHit was implemented in a Tensorflow environment).

Discriminative Performance

Performance Metric As our metric of performance, we use the *time-dependent concordance index* (C^{td}-index) (Antolini, Boracchi, and Biganzoli 2005). (Recall that the ordinary concordance index (C-index) (Harrell et al. 1982) is a widely used discriminative index based on the assumption that patients who lived longer should have been assigned a lower risk than patients who lived less long. However the ordinary C-index is computed only at the initial time of observation and hence cannot reflect the possible change in risk over time. The time-dependent concordance index takes time into account.) Given the estimated CIF in Eq. (1), the C^{td}-index for event k is defined as

$$C^{td} = P\left(\hat{F}_k(s^{(i)}|\mathbf{x}^{(i)}) > \hat{F}_k(s^{(i)}|\mathbf{x}^{(j)}) | s^{(i)} < s^{(j)}\right) \approx \frac{\sum_{i \neq j} A_{k,i,j} \cdot \mathbb{1}\left(\hat{F}_k(s^{(i)}|\mathbf{x}^{(i)}) > \hat{F}_k(s^{(i)}|\mathbf{x}^{(j)})\right)}{\sum_{i \neq j} A_{k,i,j}} \quad (6)$$

where, as before, $A_{k,i,j}$ is the indicator function for a pair (i, j) to be acceptable for an event k and the approximation comes from the empirical definition. Thus, the C^{td}-index for event k is derived from comparison of pairs in which one patient has experienced event k at a particular time while the other has not experienced any event nor been censored by that time. Because this discriminative index does not depend on a single fixed time, it provides an appropriate assessment for situations in which the influence of covariates on survival varies over time (in other words, risks are non-proportional over time). (Note that the C^{td}-index is equivalent to the usual C-index of (Harrell et al. 1982) in the case of a single event and a survival model for which the proportional hazards assumption holds.)

For the SEER and SYNTHETIC datasets, which have two events (competing risks), the discriminative performance of DeepHit was compared with the Fine-Gray proportional sub-distribution hazards model (**Fine-Gray**) (Fine and Gray 1999), deep multi-task Gaussian process (**DMGP**) (Alaa and van der Schaar 2017), and with a cause-specific version of the Cox Proportional Hazards Model (*cs*-**Cox**) that was created by fixing an event (e.g. death from CVD) and treating the other event (e.g. death from breast cancer) simply as a form of censoring; see (Haller, Schmidt, and Ulm 2013). The results are shown in Tables 2 and 3. (For completeness, we also compared with cause-specific versions of other models intended for single-event analysis; the results are shown in the Table 1 and 2 of the Supplementary Materials.)

For the UNOS and METABRIC datasets, which have a single event (risk), the discriminative performance of DeepHit was compared with two families of other survival models [6]. The first of these families consists of conventional survival regression models: including Cox Proportional Hazards (**Cox**) (Therneau 2015), Threshold Regression (**ThresReg**) (Lee and Whitmore 2006), and Random Survival Forests (**RSF**) with # of trees = 100 (Ishwaran and Kogalur 2017). The other family consists of survival models which are derived from mortality prediction performed by machine learning algorithms: Random Forest (**MP-RForest**), Logistic Regression (**MP-LogitR**), and AdaBoost (**MP-AdaBoost**) and with the cutting-edge deep neural network (**DeepSurv**), which is developed upon Cox proportional assumption (Katzman et al. 2016) [7]. (In order to make fair comparisons, the training of the MP based machine learning algorithms was adjusted for survival data; see the Supplementary Material for details.)

Comparisons of the performance of DeepHit with other models for the SEER and the SYNTHETIC datasets are shown in Table 2 and 3, respectively. In the SEER dataset, there are two events – competing risks: death from cardiovascular disease (CVD) and from Breast Cancer. As can be seen, DeepHit provides performance improvements over other models; with the exception of *cs*-Cox for death by CVD, the performance improvements were all statistically significant ($p < 0.05$ and often $p < 0.001$).[8] The comparisons for death by breast cancer are particularly striking. Fine-Gray and *cs*-Cox both perform poorly with respect to the risk of breast cancer, while DeepHit performs much better. Because Fine-Gray and *cs*-Cox assume linear proportional hazards and DMGP model assumes the underlying stochastic process to follow Gaussian process, while DeepHit makes no such assumption, the performance comparison strongly suggests that non-proportional and/or non-linear relationships between covariates and survival times is crucial for assessing the risk of breast cancer.

We also compared the discriminative performance of DeepHit with that of Fine-Gray and *cs*-Cox on the SYNTHETIC dataset where there are again two events/competing risks: death from Event 1 and from Event 2. As can be seen in Table 3, DeepHit outperformed all the benchmarks and the performance improvements were all statistically significant ($p < 0.001$). This is expected since the *cs*-Cox and Fine-Gray restrict the relationship between covariates and risks to be linear. Thus, they are not able to capture the quadratic relationship introduced when generating the synthetic data. However, DeepHit allows the network to learn the representation of the non-linear relation of covariates.

[6] We did not compare with (Luck et al. 2017) because that paper did not provide detailed information to permit implementation.

[7] https://github.com/jaredleekatzman/DeepSurv

[8] As noted earlier, for the sake of completeness we also compared the performance of DeepHit with cause-specific versions of other methods; see Tables 2 and 3 in the Supplementary Materials.

Table 2: Comparison of cause-specific C^{td}-index performance (mean and 95% confidence interval) tested on the SEER dataset

Algorithms	CVD	Breast Cancer
cs-Cox	0.672 (0.664 - 0.680)	0.639* (0.633 - 0.645)
Fine-Gray	0.663‡ (0.656 - 0.670)	0.639* (0.632 - 0.646)
DMGP	0.657 (0.632 - 0.682)	0.742‡ (0.738 - 0.746)
DeepHit ($\alpha = 0$)	0.674 (0.661 - 0.687)	0.736 (0.733 - 0.739)
DeepHit	**0.684 (0.674 - 0.694)**	**0.752 (0.748 - 0.756)**

$*$ indicates p-value < 0.001
‡ indicates p-value < 0.05

Table 3: Comparison of cause-specific C^{td}-index performance (mean and 95% confidence interval) tested on the SYNTHETIC dataset

Algorithms	Event 1	Event 2
cs-Cox	0.578* (0.570 - 0.586)	0.588* (0.584 - 0.593)
Fine-Gray	0.579* (0.572 - 0.586)	0.589* (0.585 - 0.593)
DMGP	0.663* (0.658 - 0.668)	0.666* (0.660 - 0.672)
DeepHit ($\alpha = 0$)	0.739 (0.735 - 0.744)	0.737 (0.732 - 0.742)
DeepHit	**0.755 (0.749 - 0.761)**	**0.755 (0.748 - 0.762)**

$*$ indicates p-value < 0.001

Single Event/Single Risk As we have noted in the Introduction, an important aspect of DeepHit is that it smoothly handles competing risks. However, it also provides improved performance when there is only a single risk. To show this, we compared the performance of DeepHit with other models for the UNOS and METABRIC (single event) datasets in Table 4. As can be seen, DeepHit consistently provided the best performance for both the UNOS and METABRIC datasets. For the UNOS dataset, the improvement of DeepHit over all the competing methods other than AdaBoost was highly statistically significant ($p < 0.01$ and often $p < 0.001$). For the METABRIC data set, the improvement of DeepHit over all the competing methods other than RSF was statistically significant ($p < 0.001$, $p < 0.05$, and often $p < 0.01$).

We suspect that for the single risk setting, the performance improvement of DeepHit comes from its capacity to capture the complicated relationship between covariates and risk, especially in the presence of many covariates. Because the other models make restrictive parametric assumptions, they are unable to capture this complicated relationship. In particular, when compared with DeepSurv, we suspect the performance improvement comes from not relying on the proportional assumption.

Table 4: Comparison of cause-specific C^{td}-index performance (mean and 95% confidence interval) tested on Single Event Datasets.

Algorithms	Datasets	
	UNOS	METABRIC
Cox	0.566* (0.563 - 0.569)	0.648† (0.634 - 0.662)
RSF	0.575† (0.571 - 0.579)	0.672 (0.655 - 0.689)
ThresReg	0.571* (0.568 - 0.574)	0.649† (0.633 - 0.665)
MP-RForest	0.552* (0.548 - 0.556)	0.650† (0.630 - 0.670)
MP-AdaBoost	0.582 (0.578 - 0.586)	0.633* (0.617 - 0.649)
MP-LogitR	0.571* (0.567 - 0.575)	0.661‡ (0.643 - 0.679)
DeepSurv	0.563* (0.555 - 0.571)	0.648† (0.636 - 0.660)
DeepHit ($\alpha = 0$)	0.573 (0.571 - 0.575)	0.646 (0.634 - 0.658)
DeepHit	**0.589 (0.586 - 0.592)**	**0.691 (0.679 - 0.703)**

$*$ indicates p-value < 0.001
† indicates p-value < 0.01
‡ indicates p-value < 0.05

In the Supplementary Material, we further investigate the performance gain of using $\mathcal{L}_2$ utilizing the definition of C^{td}-index: weighted average of the area under time-specific ROC curve (Antolini, Boracchi, and Biganzoli 2005).

Conclusion

This paper presents a novel approach, DeepHit, to the analysis of survival data. DeepHit trains a neural network to learn the estimated joint distribution of of survival time and event, while capturing the right-censored nature inherent in survival data. We train the network by using a loss function that exploits both survival times and relative risks. As a test, we compared the performance of DeepHit with the performance of previous models. In settings with competing risks, the performance of DeepHit is much better than that of previous models; even in settings with a single risk the performance of DeepHit is significantly better than that of previous models.

Acknowledgments

The authors would like to thank Kartik Ahuja and Anton Nemchenko for assistance in running simulations. The research presented in this paper was supported by Natural Science Foundation (NSF) under Grant Number 1407712 and 1533983.

References

Ahuja, K.; Zame, W. R.; and van der Schaar, M. 2017. Dpscreen: Dynamic personalized screening. *In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2017)*.

Alaa, A. M., and van der Schaar, M. 2017. Deep multi-task gaussian processes for survival analysis with competing risks. *In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2017)*.

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

Bilal, E.; Dutkowski, J.; Guinney, J.; Jang, I. S.; Logsdon, B. A.; Pandey, G.; Sauerwine, B. A.; Shimoni, Y.; Vollan, H. K. M.; Mecham, B. H.; Rueda, O. M.; Tost, J.; Curtis, C.; Alvarez, M. J.; Kristensen, V. N.; Aparicio, S.; Brresen-Dale, A.-L.; Caldas, C.; Califano, A.; Friend, S. H.; Ideker, T.; Schadt, E. E.; Stolovitzky, G. A.; and Margolin, A. A. 2013. Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS Computational Biology*.

Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. *In Proceedings of the 25th International Conference on Machine Learning (ICML 2008)* 160167.

Cox, D. R. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B* 34:187220.

Doksum, K. A., and Hyland, A. 1992. Models for variable-stress accelerated life testing experiments based on wiener processes and the inverse gaussian distribution. *American Statistical Association and American Society for Quality* 34(1):74–82.

Faraggi, D., and Simon, R. 1995. A neural network model for survival data. *Statistics in Medicine* 14:73–82.

Fernndez, T.; Rivera, N.; and Teh, Y. W. 2016. Gaussian processes for survival analysis. *In Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016)*.

Fine, J. P., and Gray, R. J. 1999. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94(446):496–509.

Gooley, T. A.; Leisenring, W.; Crowley, J.; and Storer, B. E. 1999. Estimation of failure probabilities in the presence of competing risks: New representations of old estimators. *Statistics in Medicine* 18(6):695–706.

Haller, B.; Schmidt, G.; and Ulm, K. 2013. Applying competing risks regression models: an overview. *Lifetime Data Analysis* 19:33–58.

Harrell, F. E.; Califf, R. M.; Pryor, D. B.; Lee, K. L.; and Rosati, R. A. 1982. Evaluating the yield of medical tests. *Journal of the American Medical Association* 247(18):25432546.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *In Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)* 770–778.

Ishwaran, H., and Kogalur, U. B. 2017. *Random Forests for Survival, Regression and Classification (RF-SRC)*. R package version 2.4.2.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; and Lauer, M. S. 2008. Random survival forests. *The Annals of Applied Statistics* 2(3):841–860.

J. Yoon, W. R. Z.; A. Banerjee, M. C.; Alaa, A. M.; and van der Schaar, M. 2017. Personalized survival predictions for cardiac transplantation via trees of predictors. *arXiv preprint arXiv:1704.03458*.

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

Katzman, J.; Shaham, U.; Bates, J.; Cloninger, A.; Jiang, T.; and Kluger, Y. 2016. Deep survival: A deep cox proportional hazards network. *arXiv preprint arXiv:1606.00931*.

Koene, R. J.; Prizment, A. E.; Blaes, A.; and Konety, S. H. 2016. Shared risk factors in cardiovascular disease and cancer. *Circulation* 133:1104–1114.

Lawless, J. F. 2002. *Statistical Models and Methods for Lifetime Data, 2nd Edition*. Wiley.

Lee, M.-L. T., and Whitmore, G. A. 2006. Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. *Statistical Science* 21(4):501–513.

Lee, M.-L. T., and Whitmore, G. A. 2010. Proportional hazards and threshold regression: Their theoretical and practical connections. *Lifetime Data Analysis* 16:196214.

Longini, I. M.; Clark, W. S.; Byers, R. H.; Ward, J. W.; Darrow, W. W.; Lemp, G. F.; and Hethcote, H. W. 1989. Statistical analysis of the stages of hiv infection using a markov model. *Statistics in Medicine* 8(7):831–843.

Luck, M.; Sylvain, T.; Cardinal, H.; Lodi, A.; and Bengio, Y. 2017. Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245*.

R. Ranganath, A. Perotte, N. E., and Blei, D. 2016. Deep survival analysis. *arXiv preprint arXiv:1608.02158*.

Steck, H.; Krishnapuram, B.; Dehing-oberije, C.; Lambin, P.; and Raykar, V. C. 2007. On ranking in survival analysis: Bounds on the concordance index. *In Proceedings of the 20th Conference on Neural Information Processing Systems (NIPS 2007)*.

Therneau, T. M. 2015. *A Package for Survival Analysis in S*. version 2.38.

Yoon, J.; Alaa, A. M.; Cadeiras, M.; and van der Schaar, M. 2017. Personalized donor-recipient matching for organ transplantation. *In Proceedings of the 31th AAAI Conference on Artificial Intelligence (AAAI 2017)*.

Yu, C. N.; Greiner, R.; Lin, H. C.; and Baracos, V. 2011. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *In Proceedings of the 24th Conference on Neural Information Processing Systems (NIPS 2011)*.