DeepPseudo: Pseudo Value Based Deep Learning Models for Competing Risk Analysis

Md Mahmudur Rahman[1], **Koji Matsuo**[2], **Shinya Matsuzaki**[3], **Sanjay Purushotham**[1]
[1] Department of Information Systems, University of Maryland, Baltimore County, Baltimore, Maryland, USA
[2] University of Southern California, Los Angeles, California, USA
[3] Osaka University, Osaka, Japan
mrahman6@umbc.edu, Koji.Matsuo@med.usc.edu, zacky_s@gyne.med.osaka-u.ac.jp, psanjay@umbc.edu

Abstract

Competing Risk Analysis (CRA) aims at the correct estimation of the marginal probability of occurrence of an event in the presence of competing events. Many of the statistical approaches developed for CRA are limited by strong assumptions about the underlying stochastic processes. To overcome these issues and to handle censoring, machine learning approaches for CRA have designed specialized cost functions. However, these approaches are not generalizable and are computationally expensive. This paper formulates CRA as a cause-specific regression problem and proposes **DeepPseudo** models, which use simple and effective feed-forward deep neural networks, to predict the cumulative incidence function (CIF) using Aalen-Johansen estimator-based pseudo values. DeepPseudo models capture the time-varying covariate effect on CIF while handling the censored observations. We show how DeepPseudo models can address covariate dependent censoring by using modified pseudo values. Experiments on real and synthetic datasets demonstrate that our proposed models obtain promising and statistically significant results compared to the state-of-the-art CRA approaches. Furthermore, we show that explainable methods such as Layer-wise Relevance Propagation can be used to interpret the predictions of our DeepPseudo models.

Introduction

Competing Risk Analysis (CRA) is a special type of survival analysis - *time-to-event analysis* - that aims to correctly estimate the marginal probability of occurrence of an event in the presence of competing events (Putter, Fiocco, and Geskus 2007). In standard survival analysis, it is assumed that a subject can experience only one event of interest during the follow-up period. However, in real life, a subject can experience an event from multiple causes. The occurrence of a certain event precludes the individual from experiencing other events. For example, a patient may experience death (event) at a time t from one of the following causes; cardiovascular disease, breast cancer, or kidney damage. Individuals who die of cardiovascular disease are no longer at risk of dying of breast cancer or kidney damage. These causes of failure are referred to as *competing events*, and the probability of these events are referred to as *competing risks*. CRA is common in medical settings (Fine and Gray 1999; Mogensen and Gerds 2013) since a patient can experience more than one type of a certain event. According to the multi-morbidity trend in the USA during 2013-14, the overall prevalence of more than 2 morbidities was 59.6% (King, Xiang, and Pilkerton 2018). Among the older adults over 65 years, that prevalence was around 92%. Therefore, multi-morbidity is a crucial problem, which should be seriously taken into consideration for patient diagnosis. Ignoring multi-morbidity or competing risks while predicting the marginal risk of an event may lead to overestimation and inaccurate risk predictions. CRA can make a more accurate prediction of the risks, and thus, help clinicians to provide better treatment to the patients. Moreover, CRA can also help to allocate health resources more efficiently. Thus, CRA is an important problem in the medical domain, and it has received substantial attention in statistics (Aalen and Johansen 1978; Fine and Gray 1999; Parner and Andersen 2010), and machine learning literature (Ishwaran et al. 2014; Alaa and van der Schaar 2017; Bellot and Schaar 2018; Bellot and van der Schaar 2018; Lee et al. 2018; Ren et al. 2019; Lee, Yoon, and Van Der Schaar 2019). The statistical CRA approaches are popular, interpretable, and can be easily implemented; however, they are limited by the underlying parametric, linearity, and/or proportional hazards assumptions. On the other hand, many machine learning approaches overcome these limitations by capturing nonlinear relationships between covariates and the risk of an event, and thus, have achieved better results. Among all these models, *DeepHit* (Lee et al. 2018) model - a deep learning approach that makes no underlying assumption on the stochastic process - has shown the state-of-the-art performance for CRA. However, it relies on a specialized objective function to handle censoring and uses a large model (i.e., a large number of parameters) to achieve good predictive performance. Thus, there is a need to develop new approaches for CRA which can address these drawbacks.

In this paper, we formulate CRA as a cause-specific regression analysis problem and propose **DeepPseudo** models, which use simple and effective deep feed-forward neural networks, to estimate cumulative incidence function (CIF) (Kalbfleisch and Prentice 2011) using pseudo values, derived from the *Aalen-Johansen* estimate of the CIF (Klein and Andersen 2005). DeepPseudo models model the nonlinear time-varying covariate effect on CIF and handle the

complexity of censored data using pseudo values. To the best of our knowledge, there is no existing work that has studied pseudo value based deep models for the CRA. DeepPseudo models are well-suited for medical applications because of their accurate predictive-ability as well as explainability. These models can help clinicians to identify which patients are at more risk (of experiencing an event) from a specific cause and to study the important features that influence those predictions. Our contributions include the following:

- We propose pseudo value based feed-forward deep neural networks, called DeepPseudo models for CRA, which model the non-linear covariate effect on CIF and easily handle censored observations.
- We propose four variants of the DeepPseudo model to predict marginal and conditional CIFs for CRA. Our models can predict subject-specific risks for different causes, and we show that conditional DeepPseudo model has good theoretical properties.
- We compare and contrast the performance of DeepPseudo models with respect to the state-of-the-art CRA models on multiple real-world and synthetic datasets and with varying censoring settings. In addition, we show that DeepPseudo models consistently obtain good performance in predicting risks of an event that can occur from more than 2 causes.
- We show that DeepPseudo models with a *small* number of parameters obtain similar or better performance compared to the state-of-the-art deep learning based CRA models, which use a large number of parameters.
- We demonstrate that off-the-shelf explainable AI methods such as Layer-wise Relevance Propagation techniques (Montavon et al. 2019) can provide a global interpretation of the covariate influence on DeepPseudo model predictions.

Background and Notations

A survival dataset with K competing events is a collection of time-to-event information of the patients along with their corresponding event status during a follow-up period. For an individual i, competing risk data is a tuple $\{(T_i, \delta_i, X_i)\}_{i=1}^{n}$. T_i is the survival time for i^{th} individual. By the definition of competing risks, there exist K latent or unobserved survival times $(T^1, ..., T^j, ..., T^K)$, one for each of the K competing events. Under the competing risk settings, it is assumed that patients eventually experience only one of the events. We observe only the first event, and so the survival time of i^{th} individual is the earliest survival time defined as $T_i = min(T_i^1, ..., T_i^K)$. For an individual i, $\mathbf{X_i} = (X_{i1}, ..., X_{ip})$ is a p dimensional vector of observed covariates and δ_i is the event indicator, where

$$\delta_i = \begin{cases} j & \text{, if the } i^{th} \text{ individual is uncensored and event} \\ & \text{occurred due to cause } j; j = 1, 2, ..., K \\ 0 & \text{, if the } i^{th} \text{ individual is censored} \end{cases}$$

Cumulative Incidence Function (CIF): Cumulative incidence function is the probability that an event; such as death, will occur at or before time t. When we have only one event of interest, CIF is defined as $F(t) = P(T \leq t) = 1 - S(t)$, where, $S(t)$ is the survival function. If the event occurs due to more than one cause, the CIF for the event due to cause k at time t is defined as

$$F_k(t) = P(T \leq t, \delta = k) = E(I(T \leq t, \delta = k)) \quad (1)$$

E and I stand for expectation and indicator function respectively. To estimate the CIF, a non-parametric Aalen-Johansen estimator is used as $\hat{F}_k(t) = \int_0^t \hat{S}(u^-) d\hat{A}_k(u)$ where, $\hat{A}_k(t)$ is the Nelson–Aalen estimator of the hazard for cause k and $\hat{S}(t)$ is the Kaplan–Meier estimator of overall survival function using all causes of the event (Binder, Gerds, and Andersen 2014).

Pseudo values: Suppose, we are interested in a function of survival time, $f(T)$, in the presence of censoring, then $\theta = E(f(T))$ is the parameter of interest, such as; a survival function or a CIF at time t. Let $\hat{\theta}$ be an unbiased estimator of θ. Then the pseudo values for an individual i , i.e, $\hat{\theta}_i$, is given by

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i} \quad (2)$$

where, $\hat{\theta}^{-i}$ is a leave-one-out unbiased estimator of θ. Pseudo values are calculated for both censored and uncensored subjects for all causes of an event at a given time point. For the i^{th} subject, a Jackknife pseudo value, based on the Aalen–Johansen (AJ) estimate of the CIF (Klein and Andersen 2005), is computed for cause k at time horizon t^* as

$$\hat{F}_{ik}(t^*) = n\hat{F}_k(t^*) - (n-1)\hat{F}_k^{-i}(t^*) \quad (3)$$

where, $\hat{F}_k(t^*)$ is the AJ estimate of the CIF for cause k based on a sample with n subjects and $\hat{F}_k^{-i}(t^*)$ is the AJ estimate of the CIF for cause k based on leave-one-out sample with $(n-1)$ subjects, obtained by omitting the i^{th} subject. The non-parametric AJ estimate of the CIF depends on a full sample, while pseudo values are subject-specific and reflect each subject's contribution to the AJ estimate of the CIF.

Related Works

CRA has been well studied in statistics and machine learning literature. Popular statistical approaches for survival analysis (Kaplan and Meier 1958; Cox 1972) have been extended to CRA (Aalen and Johansen 1978; Fine and Gray 1999). Many of these statistical approaches (Putter, Fiocco, and Geskus 2007) model the effect of covariates on the CRA outcomes through the cause-specific hazard function. In the cause-specific hazard model, k independent Cox-proportional hazard models (Cox 1972) are modeled for k causes of the event, considering competing events as censoring. The independent censoring assumption for the competing risks used in cause-specific hazard models do not hold in general, and hence, researchers have used cumulative incidence function (CIF) (Kalbfleisch and Prentice 2011) for CRA. There is a one-to-one relationship between hazard rate and CIF when there is only one event of interest. In the presence of competing risks, there is no such direct relationship as the CIF depends on the crude hazard rate of all the causes of the event. Therefore, directly modeling the effect of covariates on the CIF is needed. Fine et al. (Fine and Gray 1999) introduced a proportional hazard model to achieve this by using a sub-distribution hazard function. However, CRA based on hazard function might be difficult to model and understand (Koller et al. 2012). An alternative regression ap-

proach based on pseudo values was proposed by Klein et al. (Klein and Andersen 2005) for directly modeling the effect of covariates on CIF. The idea behind using the pseudo values is to replace the function of incompletely observed survival time by pseudo observations (Binder, Gerds, and Andersen 2014). However, these statistical approaches are limited by the underlying parametric, linearity, and/or proportional hazards assumptions.

Recently, machine learning and deep learning models have been developed for CRA (Ishwaran et al. 2014; Alaa and van der Schaar 2017; Bellot and Schaar 2018; Bellot and van der Schaar 2018; Lee et al. 2018). These models, designed to overcome the drawbacks of statistical approaches, can capture nonlinear relationships between covariates and the risk of an event. Among all these models, *DeepHit* (Lee et al. 2018) model - a deep learning approach that makes no underlying assumption on the stochastic process - has achieved state-of-the-art performance for CRA. However, it relies on a specialized objective function to handle censoring. Moreover, it uses a large model (i.e., a large number of parameters) to achieve good predictive performance, and as a result, does not provide easily explainable results. More recently, pseudo value based deep neural networks (Zhao and Feng 2019) have been developed for the standard survival analysis problem but not for the challenging CRA problem. Inspired by these works and to address their limitations, we formulate CRA as a cause-specific regression analysis problem and develop pseudo value based deep learning models called **DeepPseudo** models for CRA.

Our proposed DeepPseudo models

Before we describe our proposed DeepPseudo models, we briefly highlight the properties and advantages of the pseudo values for estimating a survival quantity such as CIF.

Why Pseudo values? Pseudo values can be derived from an asymptotically unbiased estimator such as Aalen-Johansen estimator, and they can be obtained for both *uncensored* and *censored* observations. Graw et al. (Graw, Gerds, and Schumacher 2009) showed that when all the observations are uncensored, the expectation of the pseudo values is equivalent to the true CIF, $E\{\hat{F}_{ik}(t)\} = F_k(t)$; and the conditional expectation of the pseudo values given covariates is equivalent to the CIF conditioned on covariates, $E\{\hat{F}_{ik}(t)|Z_i\} = F^*_k(t|Z_i)$ at all time points t. In the right censoring condition, *et al.*, (Lemma 2) showed that the pseudo values derived from the AJ estimate of the CIF satisfy the conditional unbiasedness under the assumption of covariate independent censoring. Thus, the pseudo values are good at handling right-censored data (Mogensen and Gerds 2013; Zhao and Feng 2019). Moreover, pseudo values are applicable in survival data with interval censoring (Do and Kim 2017), and they are easier to compute using publicly available R packages such as "pseudo" (Klein et al. 2008) and "prodlim" (Andersen and Pohar Perme 2010).

We formulate the complex CRA as a cause-specific regression analysis problem using pseudo values as a quantitative response variable and propose deep learning-based models called **DeepPseudo** models to perform this regression analysis. DeepPseudo models use simple feed-forward deep neural networks to predict the pseudo values for CIF and thus inherently handle the censoring without needing a specialized objective function (as required by other deep learning models such as DeepHit). Moreover, DeepPseudo models capture the non-linearity in the data without making any underlying assumptions such as linearity or proportional hazards (as required by statistical models such as cause-specific Cox proportional hazard model).

In DeepPseudo models, covariates are inputs, and the target variables (y) are the subject-specific pseudo values. We can calculate subject-specific CIF, i.e., the probability that a subject will experience an event at or before a time point t using these predicted subject-specific pseudo values. Depending on how the pseudo values are calculated and how the cause-specific events are modeled, we propose four variants of our DeepPseudo models.

Marginal DeepPseudo Model: The outcome of interest in CRA is the probability of occurrence of an event at or before time t for cause k, which is given by the marginal CIF, $F_k(t) = P(T \leq t, \delta = k)$. We compute the Aalen-Johansen estimate of marginal CIF and then estimate the pseudo values using equation 3. The goal of this model is to estimate the marginal CIF for each patient using the pseudo values given the covariates. In particular, we feed the covariates as input and treat the pseudo values for the marginal CIF as the output of the deep feed-forward neural network, as shown in Figure 1 (a). The outputs of this model are the cause-specific prediction of pseudo values at M evaluation time points $(\tau_1, \tau_2, ..., \tau_M)$. The main advantage of this model is that we use the same deep neural network to compute the marginal CIFs for different time points and for different causes for each patient (covariates at t=0).

Cause-specific (CS) Marginal DeepPseudo Model: Inspired by (Lee et al. 2018), we extend the Marginal DeepPseudo Model with cause-specific sub-networks to predict the pseudo values for each cause separately. As shown in Figure 1 (b), this model uses a shared sub-network to learn the shared representation of the competing events. The cause-specific sub-networks are used to predict the pseudo values (and thus the marginal CIFs) for a specific cause at M evaluation time points.

Conditional DeepPseudo Model: The risks of experiencing an event changes over time (Jung, Lee, and Chow 2018), and so both patients and clinicians want to know the risks of the patients at different time intervals given that the patients survived the previous time intervals to monitor the progress of the treatment and disease. For instance, they may be interested in estimating the probability that a patient will experience an event at an interval $(\tau_Q, \tau_{Q+1}]$ who already survived the previous interval $(\tau_{Q-1}, \tau_Q]$. For this case, we should include only the patients who were event free at the interval $(\tau_{Q-1}, \tau_Q]$ into the risk set for $(\tau_Q, \tau_{Q+1}]$ interval and then estimate the CIF conditioned on the risk set, which we denote as conditional CIF and define it as $F_k(\tau_{Q+1}|R_Q) = P(T \leq \tau_{Q+1}, \delta = k|T \geq \tau_Q)$. Here, R_Q is the risk set at time interval $(\tau_Q, \tau_{Q+1}]$ and it corresponds to the patients who survived the time interval $(\tau_{Q-1}, \tau_Q]$. After estimating this conditional CIF, we can use the fol-

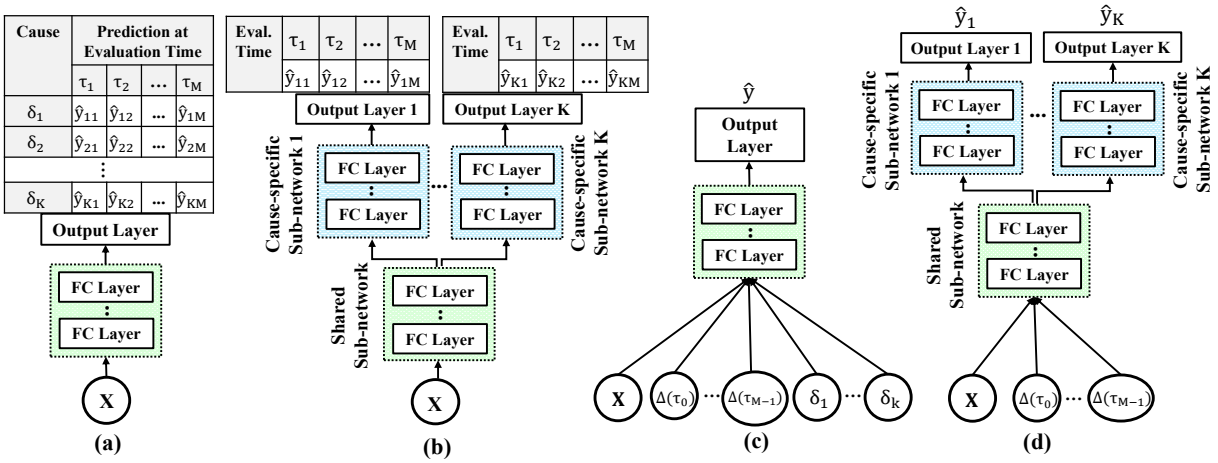

Figure 1: Model architecture of 4 variants of our proposed DeepPseudo models. From the left to right in order (a) Marginal DeepPseudo Model ($\hat{y}_{ki}$ are the predicted pseudo values for cause k and time point τ_i), (b) Cause-specific Marginal DeepPseudo Model (each cause-specific "k" branch is predicting $\hat{y}_{ki}$) , (c) Conditional DeepPseudo Model ($\Delta(\tau_i)$ is the indicator of time-interval $(\tau_i, \tau_{i+1}]$ and δ_k is the event indicator for cause k. $\hat{y}$ is the cause-specific and interval-specific predicted pseudo value), (d) Cause-specific Conditional DeepPseudo Model ($\Delta(\tau_i)$ is the indicator of time-interval $(\tau_i, \tau_{i+1}]$ and each cause-specific "k" branch is predicting the pseudo values for cause k; $\hat{y}_k$). FC Layer means Fully Connected Layer with dropout. $\mathbf{X}$ is a p dimensional vector of covariates for all subjects.

lowing equation (4) to estimate pseudo values by replacing n in equation (3); with R_Q:

$$\begin{aligned}\hat{F}_{ik}(\tau_{Q+1}|R_Q) &= R_Q * \hat{F}_k(\tau_{Q+1}|R_Q) \\ &-(R_Q - 1) * \hat{F}_k^{-i}(\tau_{Q+1}|R_Q)\end{aligned} \tag{4}$$

Note that the above equation is analogous to the pseudo conditional survival probability defined in (Zhao and Feng 2019) but our equation (4) is defined to estimate the pseudo values for conditional CIF.

Lemma 1. *The pseudo values for conditional CIFs are conditionally independent given the risk set at different time intervals.*

Theorem 1. *The pseudo values for the marginal CIF at a time point τ_M is the product of the pseudo values for conditional CIFs of the previous intervals up to time point τ_M.*

Proof of Theorem 1 follows from Lemma 1 and is provided in the supplementary materials.

Similar to Marginal DeepPseudo models, in this Conditional DeepPseudo model, we can use the pseudo values for conditional CIF as a quantitative response variable of the feed-forward deep neural networks, as shown in Figure 1 (c). The inputs to this deep model include the covariates, and one-hot encoded vectors of the time intervals and the causes of the event. The model's output layer has a single node (neuron), which predicts the pseudo value for the conditional CIF for a cause at a particular time interval, and thus can predict the pseudo values for all of the causes at each of the intervals. The pseudo values of the marginal CIF can be obtained from this model using Theorem 1.

Cause-specific (CS) Conditional DeepPseudo Model: We can modify the Conditional DeepPseudo model by using cause-specific sub-networks to predict the conditional CIF pseudo values for each cause separately. As shown in Figure 1 (d), the input to this model includes the covariates, along with one-hot encoded vectors of the time intervals. The output layer of each cause-specific network is a single neuron, which predicts the pseudo value at an input time interval.

We trained all the above variants by minimizing the mean squared error loss, which minimizes the squared differences between the true and predicted pseudo values.

Handling covariate dependent censoring: The Aalen-Johansen estimator is consistent for CIF when the censoring distribution is independent of the covariates (Binder, Gerds, and Andersen 2014), and in the presence of covariate dependent censoring, the Aalen-Johansen estimator produces a large-sample bias for CIF. A simple way to check the dependence between censoring and covariates is to model the conditional censoring survival distribution using the Cox-PH model. If there is significant dependence, we can use modified estimators of the CIF proposed by *Binder et al.* (Binder, Gerds, and Andersen 2014). The pseudo values based on the modified estimators are marginally (approximately) unbiased in the presence of covariate-dependent censoring. So to handle the covariate dependent censoring, our DeepPseudo model variants use the modified pseudo values as: $\tilde{F}_{ik}(t) = n\tilde{F}_k(t) - (n-1)\tilde{F}_k^{-i}(t)$, where, $\tilde{F}_k(t)$ is defined as

$$\tilde{F}_k(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{N_{ik}(t)}{\hat{C}_i(\tilde{T}_{i^-}|Z_i)} \tag{5}$$

$N_{ik}(t)$ is the indicator of observing event from cause k for subject i and $\hat{C}_i(.|Z_i)$ is the censoring survival distribution. We compute $\tilde{F}_k^{-i}(t)$, the leave-i-out estimator for CIF as suggested in (Binder, Gerds, and Andersen 2014). In our experiments, we have used both Jackknife pseudo values (assuming covariate independent censoring) and modified pseudo values (assuming covariate dependent censoring) and compared their performance. We observed that our DeepPseudo models obtain similar results irrespective of the dependence of censoring on the covariates. Please refer to supplementary materials for more details.

Experiments

We conducted experiments to answer the following questions: (a) how do our proposed models compare against state-of-the-art CRA models? (b) how do different censoring settings affect our model's performance? (c) what is the impact of model capacity on discriminative ability, and (d) how to explain the predictions of our models?

Datasets: We conducted experiments on two real-world datasets and one synthetic dataset.

SEER: The Surveillance, Epidemiology, and End Results (SEER)[1] Program provides information on cancer statistics to reduce the cancer burden in the United States. We extracted a cohort of 28366 patients out of which 23.2% died due to cervical cancer (cause 1), 8.4% died of other causes (Cause 2), and 68.4% patients are right-censored. We considered 13 features/covariates, including age at diagnosis, race, marital status, histology record, Grade, tumor size, cancer stages (TNM staging system), surgery record, cancer therapies, histology, etc., for our analysis. We also consider the SEER dataset with multiple causes (6 causes), where 23.2% patients died of cervical cancer (cause 1), 2.6% died due to other cancers (cause 2), 2.4% died of cardiovascular disease (cause 3), 1.1% died due to chronic medical disease (cause 4), 0.6% died of infectious disease (cause 5), and 1.8% died due to other causes (cause 6).

WIHS: We selected a cohort of 1164 women enrolled in WIHS (Bacon et al. 2005) study who were alive, infected with HIV, and free of clinical AIDS during the study period December 1995 - September 2006. The dataset contains two competing risks (Highly active antiretroviral therapy (HAART) initiation (Cause 1) & AIDS/Death before HAART (Cause 2) as well as right censoring. The dataset included the following features: the history of injection drug use at enrollment, race, age, and CD4 nadir prior to baseline.

Synthetic data: We constructed a synthetic dataset with two competing risks similar to *Lee et al.* (Lee et al. 2018), by generating the hitting times from an exponential distribution with a mean parameter depending on both linear and non-linear functions. This dataset consists of 12 features that follow a standard normal distribution. We generated 30000 observations, out of which 15000 were right-censored. We use this dataset to examine the models' performance in the presence of non-linearity in the dataset.

Implementation details: We performed stratified 5-fold cross-validation so that a constant censoring ratio[2] is maintained in each fold. We used one-hot encoding for representing categorical variables. We obtain the (ground-truth) pseudo values for CIF using the 'jackknife' function of R package 'prodlim' for each cause and evaluation time point (separately for training and validation sets). For DeepPseudo models, early stopping was performed, and the best model was chosen based on the performance on the validation fold. We fine-tuned hyperparameters by random search over the number of layers, number of neurons, learning rate, batch size. We used dropout, Adam optimizer, and 'selu' activation in training. For performance metrics, we used (a) cause-specific time-dependent concordance index (Gerds et al. 2013) for evaluating the discriminative-ability, and (b) Brier Score (Mogensen et al. 2012) metric for evaluating the predictive-ability. We used Tukey's HSD test (Abdi et al. 2010) -a pairwise statistical significant test- to measure the statistical significance of our best DeepPseudo models' results over other models. We ran experiments on a 128GB RAM Intel Xeon dual 10-core processor with 3 GPUs. Our codes and supplementary materials are at this link[3].

Censoring settings: We examine and compare our models' performance in different censoring settings. (a) *Incremental censoring*: Incrementally, we add censored observations to a fixed number of uncensored observations. This helps us to study the impact of censoring on uncensored observations in an increasing dataset. (b) *Induced censoring*: Starting with uncensored observations, we gradually induce censoring by changing (flipping) the label of the uncensored observations. This helps us to study the impact of increasing censoring ratio in a fixed-sized dataset. In our experiments on SEER dataset (see Table 2), for incremental censoring setup, we incrementally add 1k or 5k observations to a fixed number of uncensored observations; and for the induced censoring, we change labels for 1k uncensored observations in each setting for a fixed-sized dataset.

Models Comparison: We compared the following models

- **Statistical models**: Cause-specific hazard model **[CSH]** (Cox 1972); Fine & Gray subdistribution hazard model **[FG]** (Fine and Gray 1999), GEE approach **[GEE]** (Klein and Andersen 2005)
- **Machine Learning models**: Random Survival Forest **[RSF]** (Ishwaran et al. 2014); Deep Multi-task Gaussian Processes **[DMGP]** (Alaa and van der Schaar 2017)
- **Deep learning models**: DeepHit (Lee et al. 2018); Our proposed models: Marginal DeepPseudo **[M-DP]**, Cause-specific Marginal DeepPseudo **[CS-M-DP]**, Conditional DeepPseudo **[C-DP]**, Cause-specific Conditional DeepPseudo **[CS-C-DP]**

Results and Discussion

The model comparison results are shown in Table 1. For the SEER dataset, our Marginal DeepPseudo model showed statistically significant performance over all the other models except the DeepHit model in almost all the cases. All the DeepPseudo model variants perform similar or better than the DeepHit model in most cases. For WIHS dataset, our Conditional DeepPseudo model gave significantly better performance than all the other baseline models, especially for 5 years of evaluation time. On the Synthetic dataset, our Conditional DeepPseudo model showed statistically significant results compared to all the other models. It is clear that the statistical models showed the worst performance on Synthetic data as these models are limited by the linearity assumptions between covariates and risks, whereas the synthetic data was generated considering the non-linear relationships. On the other hand, our proposed models and the other ML/deep models capture both the linear and non-linear

[1] https://seer.cancer.gov/

[2] (Censoring ratio=No. of censored observations/Total no. of observations)(Berberidis, Kekatos, and Giannakis 2016)

[3] https://github.com/umbc-sanjaylab/DeepPseudo_AAAI2021

Dataset	Cause (C)	Eval. Time	Statistical Models			ML Models		Deep Learning Models				
			CSH	FG	GEE	RSF	DMGP	DeepHit	M-DP	CS-M-DP	C-DP	CS-C-DP
SEER	C1	1yr	0.868[a]	0.865[a]	0.868[a]	0.870[c]	0.871	0.876	**0.877**	**0.877**	0.866	0.865
		5yr	0.803[a]	0.805[b]	0.804[a]	0.804[a]	0.803[a]	0.808	0.812	**0.813**	0.806	0.803
	C2	1yr	0.838	0.779[a]	0.801[a]	0.816[a]	0.763[a]	0.846	**0.852**	0.841	0.838	0.845
		5yr	0.803	0.783[a]	0.785[a]	0.779[a]	0.768[a]	0.803	0.808	0.809	0.799	**0.814**
WIHS	C1	1yr	**0.734**	0.694	0.704	0.702	0.722	0.721	0.731	**0.734**	0.732	0.723
		5yr	0.646	0.635	0.643	0.602[a]	0.626[c]	0.608[a]	0.619	0.616	**0.654**	0.638
	C2	1yr	0.668	0.646[a]	0.672	0.691	0.681	0.680	**0.702**	**0.702**	0.698	0.700
		5yr	0.639[a]	0.637[a]	0.661[d]	0.659[c]	0.676	0.656[c]	0.671	0.668	**0.684**	0.665
Synthetic	C1	1yr	0.581[a]	0.581[a]	0.581[a]	0.628[a]	0.751[b]	0.753[d]	0.755	0.749	**0.761**	0.753
		5yr	0.557[a]	0.557[a]	0.558[a]	0.576[a]	0.676[a]	0.682[a]	0.671	0.681	**0.703**	0.690
	C2	1yr	0.584[a]	0.586[a]	0.585[a]	0.624[a]	0.748[a]	0.752[c]	0.754	0.749	**0.760**	0.757
		5yr	0.559[a]	0.559[a]	0.559[a]	0.573[a]	0.674[a]	0.679[a]	0.664	0.675	**0.699**	0.690

Table 1: Model performance comparisons using cause-specific time dependent C-index. Tukey's HSD test - statistically significant codes: 0 'a' 0.001 'b' 0.01 'c' 0.05 'd' 0.1 ' ' 1, (Read p 'a' as significant at p% level of significance)

	Induced censoring						Incremental censoring							
	0k	1k	2k	3k	4k	5k	0k	1k	2k	3k	4k	5k	10k	15k
M-DP	**0.61** (**0.02**)	**0.60** (**0.03**)	**0.62** (**0.02**)	**0.61** (**0.01**)	**0.60** (**0.01**)	**0.59** (**0.03**)	**0.61** (**0.03**)	**0.64** (**0.04**)	**0.67** (**0.01**)	**0.68** (**0.02**)	**0.69** (**0.02**)	**0.70** (**0.02**)	**0.74** (**0.02**)	**0.75** (**0.01**)
DeepHit	0.56 (0.03)	0.56 (0.03)	0.55 (0.02)	0.56 (0.01)	0.52 (0.04)	0.55 (0.04)	0.55 (0.03)	0.58 (0.07)	0.64 (0.03)	0.67 (0.02)	0.66 (0.02)	0.68 (0.03)	0.72 (0.03)	0.74 (0.02)
RSF	0.58 (0.03)	0.58 (0.02)	0.58 (0.03)	0.58 (0.02)	0.57 (0.01)	0.58 (0.05)	0.58 (0.02)	0.61 (0.03)	0.61 (0.01)	0.63 (0.03)	0.65 (0.02)	0.66 (0.01)	0.71 (0.02)	0.73 (0.03)
FG	0.51 (0.02)	0.52 (0.04)	0.53 (0.02)	0.54 (0.02)	0.55 (0.02)	0.57 (0.05)	0.51 (0.01)	0.52 (0.03)	0.50 (0.02)	0.51 (0.03)	0.53 (0.02)	0.54 (0.01)	0.60 (0.02)	0.65 (0.03)

Table 2: C-index (mean and SD) comparisons for different censoring settings. Dataset: SEER; Event due to cause 2; Evaluation time: 5 years. 0k means no censored observations, 1k corresponds to 1k censored observations in the dataset, and so on.

No. of Params	Cause (C)	Eval. Time	DeepHit	M-DP
∼5k	C1	1yr	0.829(0.042)	**0.868**(**0.017**)
		5yr	0.784(0.027)	**0.810**(**0.017**)
	C2	1yr	0.800(0.043)	**0.822**(**0.020**)
		5yr	0.763(0.031)	**0.787**(**0.011**)
∼50k	C1	1yr	0.871(0.014)	**0.872**(**0.013**)
		5yr	0.805(0.016)	**0.809**(**0.016**)
	C2	1yr	0.796(0.047)	**0.835**(**0.030**)
		5yr	0.766(0.039)	**0.799**(**0.016**)
∼100k	C1	1yr	0.871(0.012)	**0.875**(**0.012**)
		5yr	0.805(0.015)	**0.811**(**0.016**)
	C2	1yr	0.802(0.065)	**0.814**(**0.013**)
		5yr	0.777(0.031)	**0.804**(**0.016**)
∼1M	C1	1yr	**0.870**(**0.012**)	0.869(0.015)
		5yr	0.801(0.014)	**0.810**(**0.017**)
	C2	1yr	**0.814**(**0.043**)	0.668(0.162)
		5yr	**0.787**(**0.026**)	0.704(0.031)

Table 3: C-index (mean and SD) comparisons of DeepHit and Marginal DeepPseudo for different parameter settings on SEER data

relationships and thus perform much better. An interesting finding is that our DeepPseudo models obtain better results over the statistical GEE approach, which also uses pseudo values for CRA. Brier scores for our models (For example, 0.051 for cause 1 at 1 year for SEER data) is similar or better than other models Brier scores (0.0598 for DeepHit, 0.0512 for GEE). The model performance using Brier score, 95% confidence intervals for C-index (Table 1), and additional results are shown in the supplementary materials. Table 2 shows that performance of different CRA approaches for different censoring settings for the SEER dataset. From this table, we see that our Marginal DeepPseudo model outperforms all the models in different censoring settings which indicates that the pseudo value based deep models are good at handling censored observations. Table 3 shows the impact of model capacity on discriminative performance of deep learning based CRA models. We see that our *small* Marginal DeepPseudo model which use around $50k$ parameters obtains similar or better results than a *large* DeepHit model, which uses around 1 Million parameters. In Table 4, we compare the results of our Marginal DeepPseudo and Cause-specific Marginal DeepPseudo model on SEER dataset with multiple causes (6 causes). We see that our proposed models perform better in almost every cases (for all causes and time points) compared to the other approaches. Another interesting finding is that our models show consistently good performance even for the causes with small number of events.

Cause (# of obs.)	Eval. Time	CSH	FG	RSF	DeepHit	M-DP	CS-M-DP
Cause 1 (6575)	**1 year**	0.868(0.002)	0.865(0.002)	0.870(0.005)	**0.877(0.003)**	0.872(0.002)	0.876(0.002)
	5 years	0.803(0.003)	0.805(0.004)	0.803(0.006)	0.810(0.005)	0.807(0.004)	**0.811(0.004)**
Cause 2 (725)	**1 year**	0.844(0.055)	0.775(0.077)	0.801(0.060)	**0.856(0.042)**	0.850(0.050)	**0.856(0.047)**
	5 years	0.782(0.031)	0.757(0.035)	0.732(0.028)	**0.799(0.026)**	0.793(0.029)	0.791(0.027)
Cause 3 (671)	**1 year**	0.857(0.047)	0.788(0.067)	0.835(0.014)	**0.873(0.037)**	0.849(0.050)	0.854(0.042)
	5 years	0.843(0.029)	0.824(0.027)	0.823(0.022)	0.856(0.034)	**0.859(0.025)**	0.850(0.028)
Cause 4 (322)	**1 year**	0.876(0.031)	0.789(0.068)	0.726(0.051)	0.877(0.021)	**0.903(0.010)**	0.901(0.017)
	5 years	0.819(0.050)	0.767(0.060)	0.734(0.045)	0.838(0.029)	0.845(0.024)	**0.858(0.026)**
Cause 5 (163)	**1 year**	0.697(0.151)	0.613(0.150)	0.598(0.208)	0.751(0.072)	**0.763(0.074)**	0.730(0.050)
	5 years	0.709(0.048)	0.677(0.073)	0.687(0.066)	**0.783(0.051)**	0.765(0.061)	0.765(0.080)
Cause 6 (500)	**1 year**	0.785(0.107)	0.733(0.110)	0.760(0.091)	0.790(0.053)	0.793(0.041)	**0.831(0.064)**
	5 years	0.747(0.054)	0.723(0.048)	0.695(0.032)	0.736(0.043)	0.755(0.053)	**0.760(0.044)**

Table 4: Model performance comparisons using C-index (mean and SD) on SEER data with multiple causes

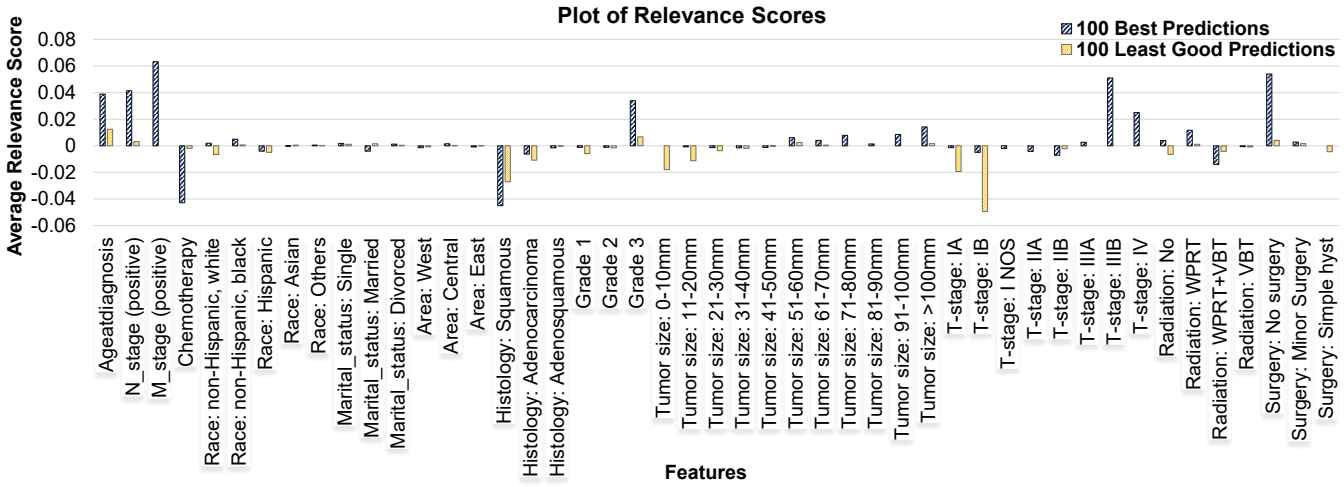

Figure 2: Explaining Marginal DeepPseudo predictions using LRP relevance score evaluated on SEER data

Explaining Our Model Predictions: Even though deep learning models provide accurate results, they are black-box models and thus, it is difficult to interpret their results. To address this issue, we employ Layer-wise Relevance Propagation (LRP) (Montavon et al. 2019) approach to explain the predictions of our proposed Marginal DeepPseudo model, i.e., we explain the covariates' contribution to the prediction of the pseudo values for CIF. LRP is typically used for classification problems, but here, we adapt it for our pseudo value based regression problem by calculating the relevance score of all the covariates based on the 100 best predictions and the 100 worse predictions as measured by the training mean squared errors. We used the python library "innvestigate" (Alber et al. 2019) and "epsilon LRP" to calculate the relevance scores of the features. Figure 1 shows the average relevance score for each of the features for these 200 predictions. In this figure, it is evident that our model can identify the important features for both good and bad predictions. For the best predictions, features such as *large tumor size, (no) surgery, T-stage (IIIB), M and N-stage (positive)* have a positive influence on prediction, which are correct as they are highly indicative of risk of dying due to cervical cancer. On the other hand, *chemotherapy* and *histology (squamous)* have a negative influence on the predictions, which shows these features reduce the risk for patients. Unlike the CIF bounded by the [0, 1], pseudo values can be < 0 and > 1 in the presence of censoring and, thus, are not directly interpretable. Therefore, we transformed the predicted pseudo values to [0, 1] range by using the clipping transform: $\hat{p} = min(1, max(0, p))$, where $\hat{p}$ and p are transformed and predicted pseudo values respectively. We can also use complementary log-log transform: $\hat{p} = 1 - e^{(-e^p)}$ or logit transform: $\hat{p} = \frac{e^p}{1+e^p}$ to make the predictions interpretable as CIF.

Conclusion

This paper formulates competing risk analysis as a pseudo value based regression problem and proposes simple deep feed-forward neural network based models, referred to as DeepPseudo models, to predict the cumulative incidence function. DeepPseudo models do not use any special cost functions or make any strong assumptions about the relationship between the covariates and the risks. Our proposed models achieve similar or better performance than the existing state-of-the-art CRA approaches and are apt at handling censoring. In addition, our models allow the use of off-the-shelf explainable AI methods to understand the positive and negative influence of covariates on the risks. For future work, we will investigate theoretical properties and study the effects of correlation between multiple risks.

Acknowledgements

This work is supported by grant CRII (IIS–1948399) from the National Science Foundation.

References

Aalen, O. O.; and Johansen, S. 1978. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 141–150.

Abdi, H.; et al. 2010. Tukey's honestly significant difference (HSD) test. *Encyclopedia of research design* 3: 583–585.

Alaa, A. M.; and van der Schaar, M. 2017. Deep multitask gaussian processes for survival analysis with competing risks. In *NeurIPS*, 2326–2334. Curran Associates Inc.

Alber, M.; Lapuschkin, S.; Seegerer, P.; Hägele, M.; Schütt, K. T.; Montavon, G.; Samek, W.; Müller, K.-R.; Dähne, S.; and Kindermans, P.-J. 2019. iNNvestigate neural networks! *J. Mach. Learn. Res.* 20(93): 1–8.

Andersen, P. K.; and Pohar Perme, M. 2010. Pseudo-observations in survival analysis. *Statistical methods in medical research* 19(1): 71–99.

Bacon, M. C.; Von Wyl, V.; Alden, C.; Sharp, G.; Robison, E.; Hessol, N.; Gange, S.; Barranday, Y.; Holman, S.; Weber, K.; et al. 2005. The Women's Interagency HIV Study: an observational cohort brings clinical sciences to the bench. *Clin. Diagn. Lab. Immunol.* 12(9): 1013–1019.

Bellot, A.; and Schaar, M. 2018. Tree-based bayesian mixture model for competing risks. In *AISTATS*.

Bellot, A.; and van der Schaar, M. 2018. Multitask boosting for survival analysis with competing risks. In *NeurIPS*, 1390–1399.

Berberidis, D.; Kekatos, V.; and Giannakis, G. B. 2016. Online censoring for large-scale regressions with application to streaming big data. *IEEE Transactions on Signal Processing* 64(15): 3854–3867.

Binder, N.; Gerds, T. A.; and Andersen, P. K. 2014. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime data analysis* 20(2): 303–315.

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

Do, G.; and Kim, Y.-J. 2017. Analysis of interval censored competing risk data with missing causes of failure using pseudo values approach. *Journal of Statistical Computation and Simulation* 87(4): 631–639.

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.

Gerds, T. A.; Kattan, M. W.; Schumacher, M.; and Yu, C. 2013. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine* 32(13): 2173–2184.

Graw, F.; Gerds, T. A.; and Schumacher, M. 2009. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* 15(2): 241–255.

Ishwaran, H.; Gerds, T. A.; Kogalur, U. B.; Moore, R. D.; Gange, S. J.; and Lau, B. M. 2014. Random survival forests for competing risks. *Biostatistics* 15(4): 757–773.

Jung, S.-H.; Lee, H. Y.; and Chow, S.-C. 2018. Statistical methods for conditional survival analysis. *Journal of biopharmaceutical statistics* 28(5): 927–938.

Kalbfleisch, J. D.; and Prentice, R. L. 2011. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Kaplan, E. L.; and Meier, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* 53(282).

King, D. E.; Xiang, J.; and Pilkerton, C. S. 2018. Multimorbidity trends in United States adults, 1988–2014. *The Journal of the American Board of Family Medicine* 31(4): 503–513.

Klein, J. P.; and Andersen, P. K. 2005. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* .

Klein, J. P.; Gerster, M.; Andersen, P. K.; Tarima, S.; and Perme, M. P. 2008. SAS and R functions to compute pseudovalues for censored data regression. *Computer methods and programs in biomedicine* .

Koller, M. T.; Raatz, H.; Steyerberg, E. W.; and Wolbers, M. 2012. Competing risks and the clinical community: irrelevance or ignorance? *Statistics in medicine* 31(11-12): 1089–1097.

Lee, C.; Yoon, J.; and Van Der Schaar, M. 2019. 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.

Lee, C.; Zame, W. R.; Yoon, J.; and van der Schaar, M. 2018. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mogensen, U. B.; and Gerds, T. A. 2013. A random forest approach for competing risks based on pseudo-values. *Statistics in medicine* 32(18): 3102–3114.

Mogensen, U. B.; et al. 2012. Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software* 50(11): 1.

Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; and Müller, K.-R. 2019. Layer-wise relevance propagation: an overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer.

Parner, E. T.; and Andersen, P. K. 2010. Regression analysis of censored data using pseudo-observations. *The Stata Journal* 10(3): 408–422.

Putter, H.; Fiocco, M.; and Geskus, R. B. 2007. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine* 26(11): 2389–2430.

Ren, K.; Qin, J.; Zheng, L.; Yang, Z.; Zhang, W.; Qiu, L.; and Yu, Y. 2019. Deep recurrent survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 4798–4805.

Zhao, L.; and Feng, D. 2019. DNNSurv: Deep Neural Networks for Survival Analysis Using Pseudo Values. *arXiv preprint arXiv:1908.02337* .