Subdistribution-Based Imputation for Deep Survival Analysis with Competing Events

Shekoufeh Gorgi Zadeh (shekoufeh.gorgizadeh@imbie.uni-bonn.de)
Department of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Sigmund-Freud-Str: 25, D-53127 Bonn

Charlotte Behning
Department of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Sigmund-Freud-Str: 25, D-53127 Bonn

Matthias Schmid
Department of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Sigmund-Freud-Str: 25, D-53127 Bonn

Research Article

Keywords: event, competing, analysis, survival, networks

Posted Date: September 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-912257/v1

Subdistribution-Based Imputation for Deep Survival Analysis with Competing Events

Shekoufeh Gorgi Zadeh[1,*], **Charlotte Behning**[1], **and Matthias Schmid**[1]

[1]Department of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn,
Sigmund-Freud-Str: 25, D-53127 Bonn, Germany
[*]shekoufeh.gorgizadeh@imbie.uni-bonn.de

ABSTRACT

With the popularity of deep neural networks (DNNs) in recent years, many researchers have proposed DNNs for the analysis of survival data (time-to-event data). These networks learn the distribution of survival times directly from the predictor variables without making strong assumptions on the underlying stochastic process. In survival analysis, it is common to observe several types of events, also called competing events. The occurrences of these competing events are usually not independent of one another and have to be incorporated in the modeling process in addition to censoring. In classical survival analysis, a popular method to incorporate competing events is the subdistribution hazard model, which is usually fitted using weighted Cox regression. In the DNN framework, only few architectures have been proposed to model the distribution of time to a specific event in a competing events situation. These architectures are characterized by a separate subnetwork/pathway per event, leading to large networks with huge amounts of parameters that may become difficult to train. In this work, we propose a novel imputation strategy for data preprocessing that incorporates the subdistribution weights derived from the classical model. With this, it is no longer necessary to add multiple subnetworks to the DNN to handle competing events. Our experiments on synthetic and real-world datasets show that DNNs with multiple subnetworks per event can simply be replaced by a DNN designed for a single-event analysis without loss in accuracy.

Introduction

In the recent years deep networks have become the state-of-the-art method in various applications, for instance in object detection[1], image captioning[2], image classification[3,4], speech recognition[5], and many other areas. One key advantage of deep neural networks is their capacity to learn specific intermediate representations/features of the data in a hierarchical manner[6] in order to create a mapping from the input predictor variables onto the outcome. In addition to the novel machine learning methods developed for survival analysis[7], recently, there has been a growing interest in using deep neural networks for this purpose, for example, the work by Giunchiglia et al.[8], Lee et al.[9], Zafar Nezhad et al.[10] and many other[11–16].

In "survival data" analysis the outcome is the time duration until one or more events occur[17]. For instance in the medical field this event could be recurrence of a disease or patient's death. A multitude of examples can e.g. be found in the work by Lee et al.[18]. Since survival data (also called data with *time-to-event* outcome) is collected over time, they are often subject to right censoring, which means that the event times of some instances are only known up to a minimum survival time. The real event times of these instances remain unknown as they are no longer observed, e.g., when a patient drops out of a study.

Many observational studies track more than one event. Often these so-called *competing events* do not occur independently, and therefore require to be analyzed together in order to avoid bias. For instance, in the CRASH-2 trial[19], which is a large randomized study on hospital death in adult trauma patients, there are multiple recorded causes for death throughout the study. The causes include death due to bleeding, head injury, multi-organ failure or other. Obviously, the occurrences of these causes are not independent. More examples on competing risks data can be found in the work by Lau et al.[20], and Austin et al.[21].

For modeling the time span T until a specific event of type $j \in \{1,...,J\}$ occurs, multiple approaches have been proposed. For example, Wolbers et al.[22] consider methods for prognostic models with right-censored data that incorporate information on the occurrence of competing events. Prentice et al.[23] model the cause-specific hazard functions of each event separately as $\xi_j(t) = \lim_{\Delta t \longrightarrow 0}\{P(t < T \leq t + \Delta t, \varepsilon = j \,|\, T > t, x)/\Delta t\}$, where $x = (x_1,...,x_p)^T$ is the vector of predictor variables and ε is a random variable indicating the type of the event that occurred at the first observed event time T. In their approach for each event j, a separate model is used for ξ_j by regarding the observations that experience the competing events as random drop-outs. To calculate cumulative incidence probabilities for any of the J events, all ξ_j need to be considered together. Another approach, on which the methods considered in this paper are based, is the subdistribution model by Fine and Gray[24]. For any event j of interest, this model considers a *subdistribution hazard* function defined by $\lambda_j(t) = \lim_{\Delta t \longrightarrow 0}\{P(t < \vartheta \leq t + \Delta t \,|\, T > t, x)/\Delta t\}$, where ϑ is a subdistribution time defined by $\vartheta = T$ if $\varepsilon = j$ and $\vartheta = \infty$ otherwise. It can be shown[24] that specifying a regression model for $\lambda_j(t)$ allows for modeling cumulative incidences of type-j events without having to model the hazard functions of the other events. Thus, only one hazard model is required.

To analyze competing events data using deep neural networks, Lee et al.[9] proposed the DeepHit network that directly learns the distribution of survival times for an event of interest while handling the competing events at the same time. In their architecture, a separate subnetwork is added for each competing event. Similarly, Gupta et al.[11] use separate subnetworks per event. In another work, Nagpal et al.[25] proposed a Deep Survival Machine (DSM), to learn a mixture of primitive distributions in order to estimate the conditional survival function $S(t|x) = P(T > t)$. Again, in this model an additional set of parameters are added in order to describe the event distribution for each competing risk.

In this work, instead of extending the network's architecture by multiple subnetworks to handle competing events, we follow the approach by Fine and Gray and propose to employ deep network architectures for a *single* event of interest[8, 26–28]. To incorporate competing events, our method works on input data that have been preprocessed using a imputation strategy based on subdistribution weights. As will be demonstrated, this strategy allows analysts to benefit from the advantages of existing implementations for time-to-event data while being able to avoid a possible bias caused by ignoring competing events. In our experiments on simulated and real-world datasets, we show that approximately the same performance can be gained without the need for specifying a complex network architecture.

Methods

Notations and Definitions

To be able to use single-event DNN architectures like DeepSurv[26], SurvivalNet[27], RNN-Surv[8], DRSA[28], etc., continuous survival times have to be grouped. To this end, we define time intervals $[0,a_1),[a_1,a_2),...,[a_{k-1},\infty)$, where k is a natural number. Further denote by $T_i \in \{1,...,k\}$ and $C_i \in \{1,...,k\}$ the event and censoring times, respectively, of an individual contained in an i.i.d. sample of size n, $i = 1,\dots,n$. In this definition, $T_i = t$ means that the event has happened in time interval $[a_{t-1},a_t)$. It is assumed that T_i and C_i are independent random variables ("random censoring"). Furthermore, it is assumed that the censoring time does not depend on the parameters used to model the event time, i.e. the censoring mechanism is non-informative for T_i[23, 29]. For right-censored data, the observed time is defined by $\tilde{T}_i = \min(T_i,C_i)$, i.e. $\tilde{T}_i$ corresponds to true event time if $T_i \le C_i$, and to the censoring time otherwise. In addition, the random variable $\Delta_i := I(T_i \le C_i)$ indicates whether $\tilde{T}_i$ is right-censored ($\Delta_i = 0$) or not ($\Delta_i = 1$). In addition to the event of interest (defined without loss of generality by $j = 1$), we assume that each individual can experience one out of $J-1$ competing events, $j \in \{2,\dots,J\}$. The type of event that the i-th individual experiences at T_i is represented by $\varepsilon_i \in 1,...,J$[30]. Analogous to the work by Fine and Gray[24] and Berger et al.[31], we are interested in modeling the cumulative incidence function $F_1(t|x) = P(T \le t, \varepsilon = 1\,|x)$ of a type-1 event using the time-constant predictor variables $x_i = (x_{i1},\cdots,x_{ip})^T$.

Imputation Strategy

Next, we describe the imputation strategy to preprocess the input data, which have to be prepared such that it is possible to train a single-event DNN focusing on the cumulative incidence function of the event of interest, $\varepsilon_i = 1$. The single-event DNN requires the input values $\min(\vartheta_i,C_i)$ and $I(\vartheta_i \le C_i)$ in addition to the values of the predictor variables x_i. In a competing risk setting, these parameters are partly unknown.

First, consider those individuals i with $\Delta_i\varepsilon_i \in \{0,1\}$. Clearly, it is not necessary to preprocess the input data of these individuals, since both $\min(\vartheta_i,C_i)$ and $I(\vartheta_i \le C_i)$ are known in these cases. Next, consider those individuals who experience a competing event first, i.e. $\Delta_i\varepsilon_i > 1$. For these individuals $\vartheta_i = \infty$, so that $I(\vartheta_i \le C_i) = 0$ is known. However, $\min(\vartheta_i,C_i) = \min(\infty,C_i) = C_i$ is unknown in these cases due to the fact that the values of the censoring times C_i are unobserved.

The main idea of our approach is to impute the missing values of C_i by sampling a censoring time for any individual i who experiences a competing event $\varepsilon_i \neq 1$. This is done as follows:

Following Berger et al.[31], we define the set of discrete subdistribution weights $w_{it} = I(t \le \tilde{T}_i)$, $i = 1,...,n$, $t = 1,...,k-1$, indicating whether individual i belongs to the *risk set* $r(t)$ or not. For each timepoint t, the set $r(t)$ includes all individuals who have neither experienced the event of interest nor have been censored before t. For individuals who experience a competing event first, $r(t)$ is not fully known. These individuals remain at risk beyond $\tilde{T}_i$ until eventually they experience censoring event.

In line with Fine and Gray[24], Berger et al.[31] suggested to set $w_{it} = 1$ if $t \le \tilde{T}_i$, knowing that individuals are at risk (i.e. belong to $r(t)$) until $\tilde{T}_i$. Following the definitions in Berger et al.[31], the probability of belonging to $r(t)$ for time $t > \tilde{T}_i$ can be estimated by

$$w_{it} := \frac{\hat{G}(t-1)}{\hat{G}(\tilde{T}_i - 1)}, \quad \tilde{T}_i < t \le k-1, \tag{1}$$

where $\hat{G}(t)$ is an estimation of the censoring survival function $G(t) = P(C_i > t)$. With this definition, w_{it} becomes an estimate of the conditional probability of individual i being part of $r(t)$, given the knowledge that it is part of $r(\tilde{T}_i)$. For the experiments in this paper, we used the R package *discSurv*[32] to define subdistribution weights w_{it}. The *discSurv* package implements a life table estimator to obtain a nonparametric estimate of $G(t)$. Note that our method bears some similarities to the work by Ruan and Gray[33], who suggested a multiple imputation approach to model continuous-time survival data in a non-DNN context.

In the final step, we use the weights w_{it} to sample the imputed censoring time $\hat{C}_i$. For this, we generate a random number $\hat{C}_i$ from a discrete distribution with support $(\tilde{T}_i + 1, \dots, k-1)$ that is defined by $P(\hat{C}_i = t) = \Delta w_{it}$, where $\Delta w_{it} = w_{it-1} - w_{it}$. In the next section we demonstrate that without loss of accuracy, the use of the imputed data simplifies the analysis of survival data with competing risks when single-event DNNs are used.

Experimental Analysis

DeepHit Network

To investigate the effectiveness of the proposed method, we use the DeepHit architecture by Lee et al.[9]. DeepHit consists of a "shared sub-network" that has two fully connected layers. (Note that in the work by Lee et al.[9], the authors use one fully connected layer for their experiments. However, empirically we found that using two fully connected layers improves the overall accuracy.) The shared sub-network creates an intermediate representation that is further combined with the input features and passed on to J "cause-specific sub-networks". As recommended by Lee et al.[9], we used two fully connected layers in each sub-network. The output of each cause-specific sub-network is a vector that estimates the probability of the first hitting time of a specific cause j at each time point t (see Figure 1). For training DeepHit, the authors use the log-likelihood of the joint distribution of the first hitting time as well as another loss term to incorporate a mixture of cause-specific ranking loss functions. They also modified the loss to handle right-censored data. In our experiments, we use the same loss term that was used to optimize DeepHit[9].

To assess the performance of our proposed method we compare three different setups: 1) *New approach using single-event DNN with preprocessed input data:* We train the DeepHit network with only one subnetwork (see Figure 1, DeepHit^1). Instead of the original input data, we use the modified version of the input data (with T_i replaced by ϑ_i), in which the censoring times corresponding to the observed competing events are imputed using the subdistribution weights. 2) *Original DeepHit approach with J subnetworks:* We train the DeepHit network with a separate cause-specific subnetwork per event (see Figure 1, DeepHit^2) 3) *Single-event DNN that ignores competing events:* Similar to the first setup, we train the DeepHit network with only one subnetwork. Instead of replacing T_i by ϑ_i, we ignore the competing events and treat all individuals with an observed competing event as censored (i.e., we treat the observed time to the occurrence of the competing event as the censoring time).

Each experiment was repeated 10 times per dataset in order to reduce the effect of random sampling and random initialization on the results.

Data Description

In this subsection, we describe the datasets that are used in the experiments. To show the effectiveness of the imputation strategy, we create three sets of simulated competing risk data. Additionally, to test our method in real-world scenarios, we use two datasets from clinical and epidemiological research: The first one was collected for the CRASH-2 clinical trial[19] mentioned above; the second one is the 2013 breast cancer dataset from the Surveillance, Epidemiology, and End Results (SEER) program[34].

Simulated Data

For generating simulated data, we use the discrete model by Berger et al.[35]. Their data generation approach is adopted from Fine and Gray[24] and Beyersmann et al.[36] schemes, and allows to create datasets from a discretized subdistribution hazard model with two competing events $\varepsilon_i \in \{1, 2\}$.

More specifically, Berger et al.[35] define the discretized subdistribution hazard model based on the continuous subdistribution hazard model

$$F_1(t|x_i) = P(T_{cont,i} \leq t, \varepsilon_i = 1\,|\,x_i) = 1 - (1 - q + q \cdot \exp(-t))^{\exp(x_i^\top \gamma_1)}, \tag{2}$$

where $T_{cont,i} \in \mathbb{R}^+$ is a continuous time variable and γ_1 is a set of regression coefficients for individual i, with predictor variables x_i. The parameter q can be used to tune the probability of having the event $\varepsilon_i = 1$ with $P(\varepsilon_i = 1|x_i) = 1 - (1-q)^{\exp(x_i^\top \gamma_i)}$ and the probability of having a competing event $\varepsilon_i = 2$ with $P(\varepsilon_i = 2|x_i) = 1 - P(\varepsilon_i = 1|x_i) = (1-q)^{\exp(x_i^\top \gamma_i)}$. Further, the continuous times for the second event are drawn from an exponential model $T_{cont,i}|\varepsilon_i = 2 \sim \mathrm{Exp}(\xi_2 = \exp(x_i^\top \gamma_2))$, with rate ξ_2 and regression parameters γ_2 for the predictor variables x_i. To obtain grouped data, we discretize the continuous event times into $k = 20$ time-intervals using empirical quantiles. Anologous to Berger et al.[31], discrete censoring times are drawn from the probability distribution $P(C_i = t) = b^{(k+1-t)}/\sum_{i=1}^{k} b^i$, where the parameter $b \in \mathbb{R}^+$ affects the overall censoring rate. Furthermore, we generate four predictor variables: two of them are normally distributed, $x_1, x_2 \sim N(0,1)$, and the other two follow a binomial distribution each, $x_3, x_4, \sim$ Binomial$(1, 0.5)$. The regression coefficients are the same as in the work by Berger et al.[35], with $\gamma_1 = c(0.4, -0.4, 0.2, -0.2)^\intercal$ and $\gamma_2 = c(-0.4, 0.4, -0.2, 0.2)^\intercal$. We simulate datasets of size $n = 30{,}000$ with different type-1 event rates $q \in \{0.2, 0.4, 0.8\}$ and a *medium* censoring rate of $b = 1$. In the simulated datasets the empirical censoring rates corresponding to $b = 1$ are $\{47.4\%, 47.6\%, 48.0\%\}$, the proportion of type-1 event rates corresponding to values of q are $\{11.5\%, 21.8\%, 38.6\%\}$, and consequently type-2 event rates are $\{41.1\%, 30.6\%, 13.4\%\}$.

CRASH-2 Data

The first real-world dataset used in our experiments was collected for the randomized CRASH-2 (Clinical Randomisation of an Antifibrinolyticin Significant Haemorrhage 2) trial, which was conducted in 274 hospitals in 40 countries between 2005 and 2010[19]. The data provide information on hospital death in adult trauma patients with or at risk of significant haemorrhage. Death was recorded during hospitalization of the patients for up to 28 days after randomization. Up to this date, patients had either died, been discharged alive, transferred to another hospital, or were still alive in hospital. For our analysis we use the publicly available version of the study database at `https://hbiostat.org/data/`. Based on Table 1 in[19], we select eight variables for analysis: Categorical variables include the sex of the patient (male/female) and type of injury (blunt/penetrating/blunt and penetrating). Continuous and ordinal variables include total Glasgow Coma Score (range 3 to 15, median = 15), the estimated age of the patient (mean = 34.6 years, sd = 14.3 years), number of hours since injury (mean = 2.8, sd = 2.4), systolic blood pressure in mmHg (mean = 97.5, sd = 27.4), respiratory rate per minute (mean = 23.1, sd = 6.7), and heart rate per minute (mean = 104.5, sd = 21.0). After discarding patients with missing values, we analyze this dataset in two ways: 1) We specify *death due to bleeding* as the event of interest for analysis ($\varepsilon = 1$) and consider *discharge from the hospital or death due to other causes* as the competing event ($\varepsilon = 2$). In this scenario, the censoring rate is 16.8%, the type-1 event rate is 4.9% and the type-2 event rate is 78.3%. 2) We specify *death from any cause* as the event for interest for analysis ($\varepsilon = 1$) and consider *discharged from the hospital* as the competing event ($\varepsilon = 2$). In this scenario, the censoring rate is 16.8, the type-1 event rate is 14.9% and the type-2 event rate is 68.3%. Table 1 summarizes the percentage of patients experiencing each event first. These different analyses enable us to investigate the performance of different methods for varying event rates while censoring remains the same.

SEER Breast Cancer Data

The second real-world dataset used in our experiments is the 2013 breast cancer data from the Surveillance, Epidemiology, and End Results (SEER) program[34]. Here our focus is on female patients with breast cancer, aged 18-75 years at the time of diagnosis. We specify *patient's death due to breast cancer* as event of interest ($\varepsilon = 1$) and consider *death due to other causes* as the competing event ($\varepsilon = 2$). The predictor variables include TNM stage (twelve T stage and four N stage categories), tumor grade (I - IV), estrogen and progesterone receptor statuses (positive/negative), primary tumor site (nine categories), surgery of primary site (yes/no), type of radiation therapy and sequence (seven and six categories, respectively), laterality (right/left), ethnicity (white, black, American Indian/Alaska Native, Asian or Pacific Islander, unknown), Spanish origin (nine categories), and marital status at diagnosis (single, married, separated, divorced, widowed). In addition to these categorical variables, we selected the following continuous and ordinal features; patient's age at diagnosis (recorded in years, mean age = 55.6 years, standard deviation (sd) = 10.8 years), the number of positive and examined lymph nodes (0-84 and $1, 2, \ldots, 89$, >90, respectively), the number of primaries (1-6), and tumor size ($0, 1, \ldots, 988$, >989 mm). After discarding patients with missing values, $121{,}798$ patients remained. For this dataset the censoring rate is 88.4%, the type-1 event rate is 6.9% and the type-2 event rate is 4.7%. For a detailed explanation of the features, please see the SEER text data file description at http://seer.cancer.gov.

Training Setup

Simulated data. For our experiments we split the $30{,}000$ instances of each set of simulated data into train ($\mathscr{D}_{train}$) , test ($\mathscr{D}_{test}$) and validation ($\mathscr{D}_{validation}$) sets randomly, making sure that the event and censoring rates are the same across the three datasets. The size of the train, test and validation datasets are $15{,}000$, $10{,}000$ and $5{,}000$ respectively. Table 1 briefly summarizes the size of the datasets used in each experiment. Since in our method the censoring times for individuals with an observed competing event are randomly imputed, we repeat the experiments 10 times and report the average performance. For each repetition, all of the individuals in train, test, and validation sets remain unchanged, except for the censoring times that are re-imputed.

CRASH-2 data. For this dataset, we use the same training setup as for the simulated data. We randomly split the $19{,}836$ instances into the train, test, and validation sets, using a stratified sampling approach that ensures all have approximately the same censoring and competing event rates (see Table 1). The size of the train, test and validation datasets are $9{,}729$, $6{,}851$ and $3{,}256$ respectively.

SEER data. We use the same training setup as for the other datasets. We randomly split the $121,798$ instances into the train, test, and validation sets, making sure all have 88.4%, 6.9%, and 4.7%, of censoring, event of interest and competing event rates respectively (see Table 1). The size of the train, test and validation datasets are $60,898$, $36,539$ and $24,361$ respectively.

Evaluation Metrics

Calibration plots based on the cumulative incidence function (CIF)

To assess the calibration of the fitted models, we perform graphical comparisons of the estimated (model-based) CIF for type-1 events and a respective nonparametric estimate obtained from the Aalen-Johansen method[37].

Generally, for input predictor variables x_i from $\mathscr{D}_{\text{test}}$, the model-based CIF at timepoint t for the event of interest is estimated by

$$\hat{F}_1(t|x_i) = \hat{P}(T \leq t, j=1|x_i) = \sum_{s=1}^{t} \hat{P}(T=s, j=1|x_i)\,, \tag{3}$$

where the probability estimates $\hat{P}(\cdot)$ in (3) are taken from the output of the DeepHit network (for details, see Lee et al.[9]). Details on the Aalen-Johansen estimator, which is a covariate-free estimator of the CIF, have been given in the book by Klein et al.[37]. In our experiments, we consider a fitted DNN model to be well calibrated if the model-based and covariate-free CIF estimates agree closely.

Concordance index (C-index[38,39])

To evaluate the discriminatory power of each method for the event of interest we use the C-index as defined by Wolbers et al.[40]. For a pair of independent individuals i and j in the $\mathscr{D}_{\text{test}}$, this measure compares the ranking of a *risk marker* $M(t,x_i)$ at timepoint t with the ranking of the survival times of the event of interest. More specifically, summarizing all competing events by $\varepsilon = 2$, the C-index is defined by

$$C_1(t) := P\left(M(t,x_i) > M(t,x_j)\,|\,\varepsilon_i = 1 \text{ and } T_i \leq t \text{ and } (T_i < T_j \text{ or } \varepsilon_j = 2)\right). \tag{4}$$

In our experiments we define $M(t,x)$ by the cumulative incidence function (Equation 3). Ideally, the C-index takes value 1 if the rankings of the risk marker and the type-1 survival times are in perfect disagreement (i.e., larger marker values are associated with smaller survival times). For our experiments, we used the inverse-probability-weighted estimator by Wolbers et al.[40] (Equation 4) that is implemented in the R package **pec**.

Results

The calibration plots for the various model fits are presented in Figure 2. It is seen that despite the smaller learning capacity of the imputation-based DeepHit^1 approach, this network results in similarly well-calibrated models as the DeepHit^2 with two sub-networks. Note that in all cases, using the sub-distribution weights for imputing the censoring times leads to a better calibration compared to the single-event DeepHit architecture that treats individuals with an observed competing event as censored (thus ignoring the competing events).

Generally, the calibration of the overall average CIF estimate improved with our method when the rate type-1 events became larger. This is seen from the last row of Figure 2. For the same censoring rates and predictive variables (for CRASH-2), DeepHit^2 resulted in an underestimation of the CIF when the rate of type-1 events was high. This is also evident in the results from our experiments on simulated data. On the other hand, our proposed method shows an overall less sensitivity to the type-1 event rate. This effect could possibly be due overfitting issues, as adding an additional sub-network for each competing event to the architecture increases the learning capacity of the network without providing enough data to train each pathway.

Analogous to the results from the calibration plots, the C-indices obtained from our imputation-based method showed a discriminatory power that was similar to respective performance of the other methods (see Table 2). In a number of settings the discriminatory power even improved when using our method. For instance, in the experiments with the simulated data, the estimated mean C-index was highest for the DeepHit^1 method with imputed censoring times. For CRASH-2 with a type-1 event rate of 4.9% the observed difference (0.01%) between imputation-based DeepHit^1 and DeepHit^2 is small. For the type-1 event rate of 14.9% our proposed method performed slightly better. For the SEER breast cancer data, however, DeepHit^1 without imputation had the best average performance with regard to the C-index. This could be due to the fact that the rate of observed competing events is low to the degree that treating the respective event times as censoring times might not have substantially affected the censoring survival function.

In terms of execution time, we observed that the average time needed for training the deep networks has reduced by 21% for the simulated data, 10% for the SEER, and 37% for the CRASH-2 dataset using our method. This time reduction is possibly due to the reduced number of parameters involved in the training of DeepHit^1 compared to DeepHit^2 (see Table 3). Consequently, in applications with more than one competing event, where third or more subnetworks are added to the architecture, the reduced time using our algorithm is expected to be even greater. The average number of iterations, however, was on the same order of magnitude for both DeepHit^1 and DeepHit^2. For all datasets on average DeepHit^1 took $15,022$ iterations and DeepHit^2 $15,277$. Note that the stopping criterion for all of the networks was the performance of the validation data.

Discussion

Even though deep neural networks are increasingly used for survival analysis, it is still relatively complicated to adapt the available methodology to situations with competing events. This is in contrast to the classical statistical literature, in which a wide variety of methods are available[20–24], and in which it is widely agreed that a properly conducted competing-risks analysis is often necessary to avoid biased estimation results and/or predictions[36]. Although several adaptations to DNN architectures have been proposed recently[9,11,25], these adaptions rely on a huge number of parameters, making network training and regularization a challenging task. In this work, we showed that an imputation strategy based on subdistribution weights could convert the competing risks survival data into a dataset that has only one event in the presence of censoring. This conversion makes the analysis of the data available to any of the much simpler deep survival network architectures that are designed to handle a single event of interest in the presence of right censoring. Our experiments on simulated and real-world datasets illustrated that this preprocessing step not only simplifies the training in terms of number of parameters and running time but also preservers the accuracy in terms of discriminatory power and calibration. The method could be further stabilized by implementing a multiple imputation approach (analogous to the continuous-time method by Ruan and Gray[33]); however, such an approach would dramatically increase the run time and would be infeasible in the context of training DNN architectures. Further, in our experiments we observed that multiple imputations did not have a major effect on predictive performance in our datasets containing several thousands of instances with event rates larger than $\sim 5\%$. Our codes for simulated data generation, censoring time imputation, and the experiments are available at `https://github.com/shekoufeh/Deep-Survival-Analysis-With-Competing-Events`.

References

1. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99 (2015).
2. Karpathy, A. & Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137 (2015).
3. Affonso, C., Rossi, A. L. D., Vieira, F. H. A., de Leon Ferreira, A. C. P. *et al.* Deep learning for biological image classification. *Expert. Syst. with Appl.* **85**, 114–122 (2017).
4. Abdel-Zaher, A. M. & Eldeib, A. M. Breast cancer classification using deep belief networks. *Expert. Syst. with Appl.* **46**, 139–144 (2016).
5. Graves, A., Mohamed, A. & Hinton, G. Speech recognition with deep recurrent neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649 (IEEE, New York, 2013).
6. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, Cambridge, MA, 2016).
7. Sonabend, R., Király, F. J., Bender, A., Bischl, B. & Lang, M. mlr3proba: An r package for machine learning in survival analysis. *arXiv preprint arXiv:2008.08080* (2020).
8. Giunchiglia, E., Nemchenko, A. & van der Schaar, M. RNN-SURV: A deep recurrent model for survival analysis. In *Proceedings of the 27th International Conference on Artificial Neural Networks*, 23–32 (Springer, Cham, 2018).
9. Lee, C., Zame, W. R., Yoon, J. & van der Schaar, M. DeepHit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2314–2321 (AAAI Press, Palo Alto, 2018).
10. Nezhad, M. Z., Sadati, N., Yang, K. & Zhu, D. A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer. *Expert. Syst. with Appl.* **115**, 16–26 (2019).
11. Gupta, G., Sunder, V., Prasad, R. & Shroff, G. Cresa: A deep learning approach to competing risks, recurrent event survival analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 108–122 (Springer, 2019).
12. Kvamme, H. & Borgan, Ø. Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724* (2019).
13. Zhu, X., Yao, J., Zhu, F. & Huang, J. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7234–7242 (2017).

14. Zhu, X., Yao, J. & Huang, J. Deep convolutional neural network for survival analysis with pathological images. In *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 544–547 (IEEE, New York, 2016).

15. Ren, K. *et al.* Deep recurrent survival analysis. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 4798–4805 (AAAI Press, Palo Alto, 2019).

16. Gorgi Zadeh, S. & Schmid, M. Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis Mach. Intell.* (2020).

17. Faraggi, D. & Simon, R. A neural network model for survival data. *Stat. Medicine* **14**, 73–82 (1995).

18. Lee, M.-L. T. & Whitmore, G. A. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Stat. Sci.* 501–513 (2006).

19. CRASH-2 Trial Collaborators. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): A randomised, placebo-controlled trial. *The Lancet* **376**, 23–32 (2010).

20. Lau, B., Cole, S. R. & Gange, S. J. Competing risk regression models for epidemiologic data. *Am. journal epidemiology* **170**, 244–256 (2009).

21. Austin, P. C., Lee, D. S. & Fine, J. P. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* **133**, 601–609 (2016).

22. Wolbers, M., Koller, M. T., Witteman, J. C. & Steyerberg, E. W. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology* 555–561 (2009).

23. Prentice, R. L. *et al.* The analysis of failure times in the presence of competing risks. *Biometrics* 541–554 (1978).

24. Fine, J. P. & Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. statistical association* **94**, 496–509 (1999).

25. Nagpal, C., Li, X. R. & Dubrawski, A. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE J. Biomed. Heal. Informatics* (2021).

26. Katzman, J. L. *et al.* Deep survival: A deep cox proportional hazards network. *stat* **1050**, 1–10 (2016).

27. Yousefi, S. *et al.* Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. reports* **7**, 1–11 (2017).

28. Ren, K. *et al.* Deep recurrent survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 4798–4805 (2019).

29. Kleinbaum, D. G. & Klein, M. *Survival analysis* (Springer, 2010).

30. Schmid, M. & Berger, M. Competing risks analysis for discrete time-to-event data. *Wiley Interdiscip. Rev. Comput. Stat.* e1529 (2020).

31. Berger, M., Schmid, M., Welchowski, T., Schmitz-Valckenberg, S. & Beyersmann, J. Subdistribution hazard models for competing risks in discrete time. *Biostatistics* **21**, 449–466 (2020).

32. Thomas Welchowski and Matthias Schmid. *R: Discrete Time Survival Analysis* (2019).

33. Ruan, P. K. & Gray, R. J. Analyses of cumulative incidence functions via non-parametric multiple imputation. *Stat. medicine* **27**, 5709–5724 (2008).

34. National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch. The surveillance, epidemiology and end results (SEER) research data. . (2013). Cases diagnosed in 1973–2010, follow up cutoff Dec 2010, released on April 2013, based on the November 2012 submission https://seer.cancer.gov/.

35. Berger, M. & Schmid, M. Semiparametric regression for discrete time-to-event data. *Stat. Model.* **18**, 1–24 (2018).

36. Beyersmann, J., Allignol, A. & Schumacher, M. *Competing risks and multistate models with R* (Springer Science & Business Media, 2011).

37. Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G. & Scheike, T. H. *Handbook of survival analysis* (CRC Press, 2016).

38. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *J. Am. Med. Assoc.* **247**, 2543–2546 (1982).

39. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. & Rosati, R. A. Regression modeling strategies for improved prognostic prediction. *Stat. Medicine* **3**, 143–152 (1984).

40. Wolbers, M., Blanche, P., Koller, M. T., Witteman, J. C. & Gerds, T. A. Concordance for prognostic models with competing risks. *Biostatistics* **15**, 526–539 (2014).

Acknowledgements

The authors thank Dr. Lothar Haberle (Department of Gynecology, Obstetrics and Mammology, University Hospital Erlangen, Germany) for valuable comments on the analysis of the SEER data. Support from the German Research Foundation (grant SCHM 2966/2-1), is gratefully acknowledged.

Author contributions statement

M.S. conceived the methodology with inputs from S.G. and C.B.. S.G. conceived and conducted the experiments, wrote the manuscript, and prepared Figures 1-2. C.B. performed the preprocessing of datasets. All authors analysed the results and reviewed the manuscript.

Additional information

The authors declare that there is no conflict of interest.

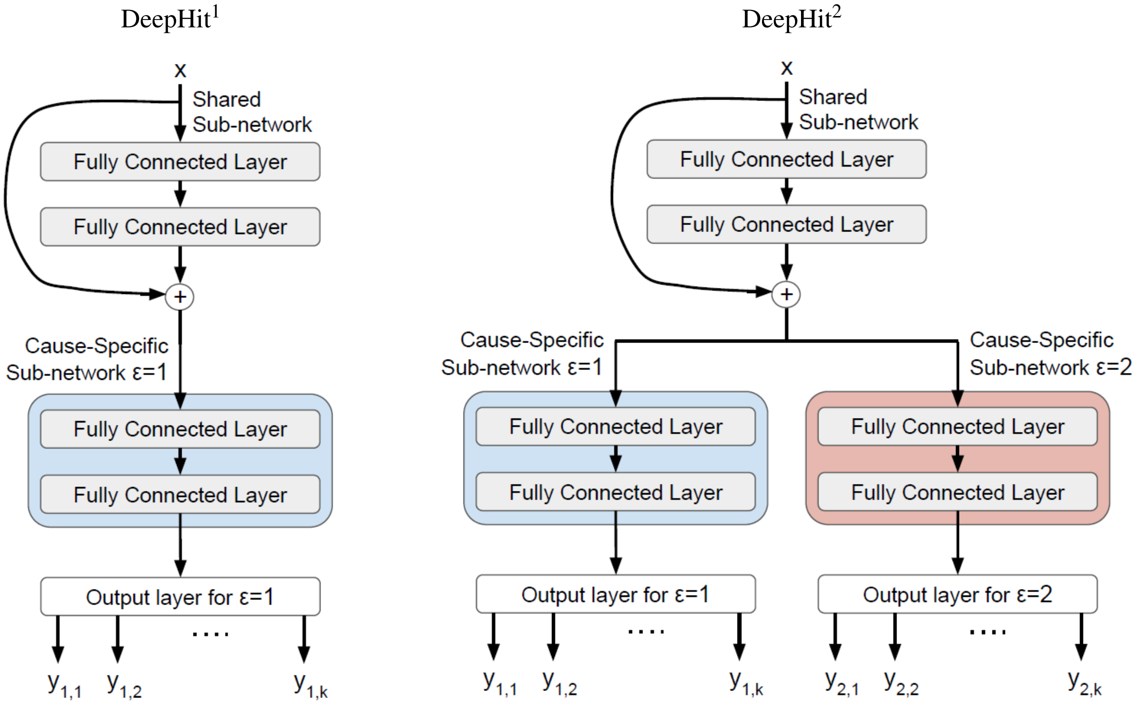
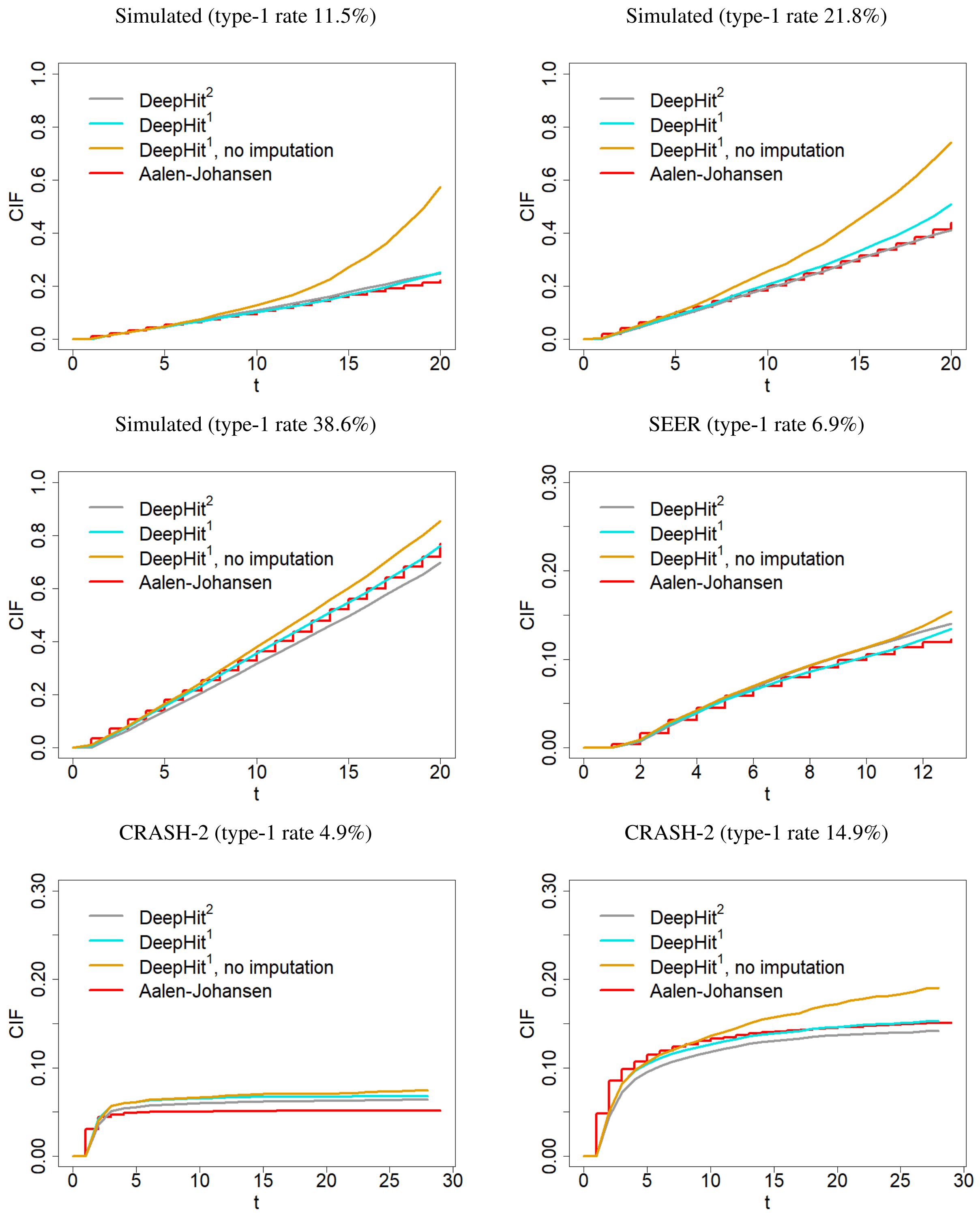


Figure 1. DeepHit^1 and DeepHit^2 architectures used in the experiments.



Simulated (type-1 rate 11.5%)

Simulated (type-1 rate 21.8%)

Simulated (type-1 rate 38.6%)

SEER (type-1 rate 6.9%)

CRASH-2 (type-1 rate 4.9%)

CRASH-2 (type-1 rate 14.9%)

Figure 2. Calibration plots obtained from the test data in Table 1. Each plot presents the averaged type-1 cumulative incidence functions as obtained from i) training the DeepHit^1 with the preprocessed data (cyan), ii) training DeepHit^1 treating individuals with observed competing events as censored (orange), and iii) training DeepHit^2 for both the event of interest and the competing event (gray). Red curves refer to the nonparametric Aalen-Johansen reference curves.

censoring rate	type-1 rate	type-2 rate	train	validation	test
simulated data					
47.4%	11.5%	41.1%	15,000	5,000	10,000
47.6%	21.8%	30.6%	15,000	5,000	10,000
48.0%	38.6%	13.4%	15,000	5,000	10,000
CRASH-2 data					
16.8%	4.9%	78.3%	9,729	3,256	6,851
16.8%	14.9%	68.3%	9,729	3,256	6,851
SEER breast cancer data					
88.4%	6.9%	4.7%	60,898	24,361	36,539

Table 1. Characteristics of the datasets used in the experiments. The three leftmost columns represent the censoring, type-1 ($\varepsilon = 1$), and type-2 ($\varepsilon = 2$) rates in the train/validation/test datasets. The three rightmost columns represent the respective numbers of instances in the simulated, CRASH-2, and SEER breast cancer data. For CRASH-2, $\varepsilon = 1$ indicates either death due to bleeding event (upper row) and death due to any recorded cause (lower row).

data	type-1-rate	type-2-rate	DeepHit[1]	DeepHit[1], no imp.	DeepHit[2]
CRASH-2	4.9%	78.3%	78.17±1.04	76.80±4.96	**78.18**±0.94
CRASH-2	14.9%	68.3%	**80.14**±1.77	79.88±2.01	80.05±4.23
SEER	6.9%	4.7%	81.75±3.46	**81.80**±3.49	81.73±3.34
Simulated	11.5%	41.1%	**64.13**±0.75	62.58±2.17	63.71±0.96
Simulated	21.8%	30.6%	**65.90**±0.69	64.59±2.25	65.20±3.26
Simulated	38.6%	13.4%	**66.05**±0.47	64.97±2.51	64.39±6.26

Table 2. Mean estimated C-indices (averaged over time) with estimated standard deviations, as obtained from training the DeepHit architecture on the simulated, CRASH-2, and SEER breast cancer data. DeepHit[1] = DeepHit architecture with one sub-network trained with the preprocessed input data; DeepHit[2] = DeepHit architecture with two subnetworks; DeepHit[1], no imp. = DeepHit architecture with one sub-network trained on the original input data (treating individuals with observed competing events as censored individuals). Best-performing methods are marked bold. Note that the C-indices must be compared within each row, as the datasets used for training were different in terms of size, censoring, and event rates across the rows. For CRASH-2, in the upper and the lower rows $\varepsilon = 1$ indicates death due to bleeding and death due to any recorded cause, respectively. The numbers in this table are obtained from the test datasets.

	simulated	SEER	CRASH-2
	time | #itr	time | #itr	time | #itr
DeepHit[1]	184.78 | 10,666	827.97 | 22,600	116.47 | 11,800
DeepHit[2]	235.32 | 9,133	918.39 | 22,300	185.30 | 14,400

Table 3. Average time (in seconds) and number of iterations needed for training DeepHit[1] and DeepHit[2] per dataset. Performance on validation data was used as the stopping criterion.