A Hierarchical Approach to Multi-Event Survival Analysis *

Donna Tjandra,[1] **Yifei He,** [1] **Jenna Wiens** [1]
[1]Computer Science and Engineering, University of Michigan, Ann Arbor MI, USA
dotjandr, heyifei, wiensj@umich.edu

Abstract

In multi-event survival analysis, one aims to predict the probability of multiple different events occurring over some time horizon. One typically assumes that the timing of events is drawn from some distribution conditioned on an individual's covariates. However, during training, one does not have access to this distribution, and the natural variation in the observed event times makes the task of survival prediction challenging, on top of the potential interdependence among events. To address this issue, we introduce a novel approach for multi-event survival analysis that models the probability of event occurrence hierarchically at different time scales, using coarse predictions (e.g., monthly predictions) to iteratively guide predictions at finer and finer grained time scales (e.g., daily predictions). We evaluate the proposed approach across several publicly available datasets in terms of both intra-event, inter-individual (global) and intra-individual, inter-event (local) consistency. We show that the proposed method consistently outperforms well-accepted and commonly used approaches to multi-event survival analysis. When estimating survival curves for Alzheimer's disease and mortality, our approach achieves a C-index of 0.91 (95% CI 0.88-0.93) and a local consistency score of 0.97 (95% CI 0.94-0.98) compared to a C-index of 0.75 (95% CI 0.70-0.80) and a local consistency score of 0.94 (95% CI 0.91-0.97) when modeling each event separately. Overall, our approach improves the accuracy of survival predictions by iteratively reducing the original task to a set of nested, simpler subtasks.

Introduction

Survival analysis, commonly used in fields such as healthcare, aims to estimate the probability of an event occurring over some time horizon (Lanza et al. 2020; Wongvibulsin, Wu, and Zeger 2020). In contrast to single-event analysis, multi-event survival analysis is more challenging due to the potentially complex interdependence among events. This relationship is further complicated by the natural variation in time-to-events, as one typically assumes that they are non-deterministic and are drawn from some distribution. In light of these challenges, researchers commonly make the simplifying assumption that the events are independent and apply existing single-event analysis techniques (Roe et al. 2012). However, when this assumption does not hold, it can lead to poor performance (Castañeda and Gerritse 2010). Recently, Lee et al. (Lee et al. 2018) used multitask learning to jointly model events. While this improved performance over existing approaches that assume independence, it remains limited in that it does not directly account for the natural variation in time-to-events. To address this, we propose a novel hierarchical approach that jointly models events. Our approach improves performance by solving a series of simpler tasks building up to the original target task.

Given a complex target task, we first divide it into a series of simpler tasks that serve as guides to help complete the target task. Our hierarchical approach *iteratively predicts survival at multiple granularities of time* (e.g., months, weeks, then days). Predictions at coarser time scales (e.g., will the event happen during the prediction horizon or not) are used to guide predictions at finer time scales (a more complex task), improving the overall accuracy on the target task. Our approach applies to the general multi-event setting. We do not assume any constraints on the order in which events may occur (Hsieh and Wang 2018), though we outline ways to account for such constraints if the setting calls for it.

Applied to publicly available data, we present *a comprehensive evaluation of our proposed approach* relative to existing baselines. To date, evaluation of survival analysis models in the multi-event setting has been limited. Common evaluation techniques like the C-index focus on discriminative performance at the global level (i.e., comparing the risk of two different individuals for the same event) but not at the local level (i.e., comparing the risk for two different events for the same individual). We propose new evaluation metrics that address these limitations.

Overall, we introduce a novel, hierarchical approach for multi-event survival analysis. Across several datasets, both synthetic and real, we show that the proposed method consistently outperforms several well-accepted baselines. More specifically, our contributions are as follows:

*Data used in preparation of this article were in part obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- We propose a novel, hierarchical, deep learning approach for multi-event survival analysis that models survival hierarchically along the horizon, using predictions at coarser time scales to guide predictions at finer time scales
- We highlight the shortcomings of current evaluation approaches for multi-event survival analysis, present a novel evaluation framework, and use it for additional supervision during training

Background and Related Work

Here, we set up the problem of multi-event survival analysis and review the current approaches as well as the existing evaluation methods, highlighting limitations.

Problem Setup and Definitions

We aim to predict the probability of survival (an event *not* occurring) over time for multiple events based on an individual's covariates. At training time, we assume that we are provided with the covariates, observed times, and censoring statuses for each individual. At test time, we are provided with only the covariates. For each event, an individual's *observed time* is the time at which the event occurred or the time at which the individual was last observed, depending on the censoring status.

Multi-Event Survival Analysis

The majority of past work in multi-event survival analysis focuses on extending the widely used Cox proportional hazards model (Cox 1972). In short, a Cox model learns the linear contribution of each covariate to the hazard (i.e., the instantaneous rate of event occurrence at a time point given survival until that time point). It assumes that the ratio between the hazards of two individuals is the same over time (hence the term proportional hazards).

Among the most common Cox-based approaches to multi-event survival analysis is that proposed by Larson (Larson 1984), in which each event is modeled separately. This approach is widely accepted in practical applications of multi-event survival analysis (Armstrong et al. 2014; Solomon et al. 2017), but it fails to capture the dependency among events (e.g., if disease onset is likely, the probability for death increases). Alternatively, one can use the multi-state model, which learns a model for each event transition (e.g., a model is learned for experiencing death conditioned on previously experiencing disease onset) (Andersen and Keiding 2002). However, by separately modeling each event transition, this approach also fails to leverage shared information among the events. This limitation has been addressed for Cox models by using a shared frailty term across models (i.e., an individual specific variable whose effect on the hazard is multiplicative) (Jiang and Haneuse 2017), additional forms of regularization (Wang et al. 2017), or by modeling the joint survival distribution (Hsieh and Wang 2018). However, these approaches require assumptions on the distribution of the frailty term, the proportional hazards assumption, or assumptions on the form of the joint survival distribution, respectively. Although recent approaches (Lee et al. 2018; Engelhard et al. 2020) address issues with respect to these restrictive assumptions, they still suffer in terms of accuracy due to the natural variation of time-to-events. We tackle this issue via our hierarchical approach.

Our proposed approach leverages a series of simpler subtasks to guide the predictions of the original task. To date, the concept of using simpler tasks to aid with predictions at more complex, finer-grained tasks has not been explored in survival analysis. However, such an approach has been used for classification (Kowsari et al. 2017; Meng et al. 2019; Malakouti and Hauskrecht 2019; Seo, Kim, and Han 2019). For example, in text classification, Kowsari et al. showed how training a classifier to learn to discriminate among super-classes (e.g., novels and research papers) can aid in learning to discriminate among sub-classes (e.g., fiction novels and review papers). By breaking the overall task into a series of nested subtasks, the model can more effectively accomplish the original task.

Evaluation with Multiple Events

Methods for evaluating approaches for multi-event survival analysis have been limited. Many report the learned coefficients of the Cox model (Andersen and Gill 1982; Peng, Xiang, and Wang 2018). However, this requires knowledge of how all possible combinations of the features contribute to the hazard. In the context of multiple events, alternative metrics, like the C-index (Harrell et al. 1982), have their own issues. Here, we focus on two shortcomings of existing metrics for multi-event survival analysis: 1) they can result in biased estimates with respect to portions of the curve, and 2) they consider only inter-individual rankings.

The C-index is perhaps the most commonly used evaluation metric in survival analysis (Harrell et al. 1982). For a single event, it summarizes the accuracy of ranking individuals by calculating the probability that a comparable, randomly chosen pair of individuals is correctly ranked. For two individuals to be comparable, at least one must have experienced the event, and the event must have occurred prior to when the second individual was last observed. The ranking is correct if the probability of survival for the individual with the earlier time-to-event is lower at some fixed time point. The C-index is a weighted average of the AUROC at different time points (Heagerty and Zheng 2005).

The C-index assumes that rankings are *consistent* over time (i.e, if two individuals are correctly ranked at time 1, then they are correctly ranked at all other time points). However, for deep approaches to survival analysis, there is no guarantee that the curves will not cross. To address this assumption, Antolini et al. proposed an extension that compares the probabilities of survival at the *earlier* point of observation within each pair instead of some fixed time point for all pairs (Antolini, Boracchi, and Biganzoli 2005). However, this comparison can result in an incomplete estimate of performance, since it does not consider the *later* points of observation within the pair, i.e., the time between when the first and second individuals experienced the event. We propose an alternative consistency score that compares individuals at *all* relevant time points.

Within the multi-event setting, the C-index is typically computed for each event separately. As a result, this com-

pares individuals within the context of a specific event. While this captures discriminative performance *among* individuals (inter-individual, intra-event) at the *global* level, it does not capture discriminative performance *within* an individual (inter-event, intra-individual) at the *local* level (i.e., it does not evaluate the survival curves among events within the same individual). We go beyond the notion of global rankings and consider local rankings within individuals.

Methods

We introduce a novel approach for multi-event survival analysis that improves the accuracy of previous work. Our architecture contains a hierarchical component that uses predictions at coarser time scales (a simpler task) to guide predictions at finer time scales (a more complex task). During training, we rely on a composite loss function composed of both established and novel ranking penalties that provide supervision across the entire prediction horizon. Unless otherwise mentioned, we assume censoring is non-informative.

Notation

Let N be the number of individuals, K be the number of events, and T be the size of the discrete time horizon over which we consider survival. We define the 'original task' as predicting the probability of event occurrence at each time point $t = 1, 2, ..., T$.

Each individual $i = 1, 2, ..., N$ is associated with three variables: $\{\mathbf{x}^{(i)}, \mathbf{c}^{(i)}, \mathbf{o}^{(i)}\}$. Let $\mathbf{x}^{(i)} \in \mathbb{R}^D$ be the vector of covariates. Let $\mathbf{c}^{(i)} \in \{0, 1\}^K$ be the vector of censoring statuses for each event, where $c_k^{(i)}$ is 1 event k is censored and 0 otherwise. Let $\mathbf{o}^{(i)} \in \{1, 2, ..., T\}^K$ be the vector of observed times for each event. Let $\mathbf{e}^{(i)} \in \{1, 2, ..., T\}^K$ be the vector of time-to-events, where $o_k^{(i)} = e_k^{(i)}$ if $c_k^{(i)} = 0$, and $o_k^{(i)} < e_k^{(i)}$ otherwise. $\mathbf{e}^{(i)}$ is not provided in the data. Let $\mathcal{S}^{(i)} = \{\mathbf{s}_1^{(i)}, \mathbf{s}_2^{(i)}, ..., \mathbf{s}_K^{(i)}\}$ be a matrix of survival curves for individual i. $\mathbf{s}_k^{(i)} = (P(e_k^{(i)} > 1), P(e_k^{(i)} > 2), ..., P(e_k^{(i)} > T))$ describes the probability of individual i *not* having event k until t for $t = 1, 2, ..., T$. Let $\mathcal{P}^{(i)} = \{\mathbf{p}_1^{(i)}, \mathbf{p}_2^{(i)}, ..., \mathbf{p}_K^{(i)}\}$ be a set of probability distributions for individual i. $\mathbf{p}_k^{(i)} = (P(e_k^{(i)} = 1), P(e_k^{(i)} = 2), ..., P(e_k^{(i)} = T))$ describes the probability of i experiencing event k at $t = 1, 2, ..., T$.

Let $(b_1, b_2, ..., b_M)$ be a vector of size M, where M is the number of temporal granularities. At granularity m, for $m = 1, 2, ..., M$, we split the horizon into T/b_m non-overlapping time bins, where the number of time points from the original task contained in each time bin is b_m. $b_1 = T$ is the coarsest granularity (i.e., it has one time bin containing all time points from the original task), and $b_M = 1$ is the granularity of the original task. b_m are monotonically decreasing, and $b_{m-1} mod b_m \equiv 0$. The time bin in which $o_k^{(i)}$ falls at granularity m is then $o_{km}^{(i)}$. Let $\tau_z^{(m)}$ be a set of time points from the original task, regrouped, under granularity m, for $z = 1, 2, ..., T/b_m$. For example, if $T = 6$ and $b_m = 3$, then we split the horizon in half with $\tau_1^{(m)} = \{1, 2, 3\}$ and $\tau_2^{(m)} = \{4, 5, 6\}$. $p_{km}^{(i)}[z] = \sum_{u \in \tau_z^{(m)}} p_k^{(i)}[u]$ is then the probability of event occurrence in $\tau_z^{(m)}$. We denote indexes into probability vectors with square brackets (e.g. $p[t]$) and use a 'hat' for the model's estimates (e.g., $\hat{\mathbf{p}}$).

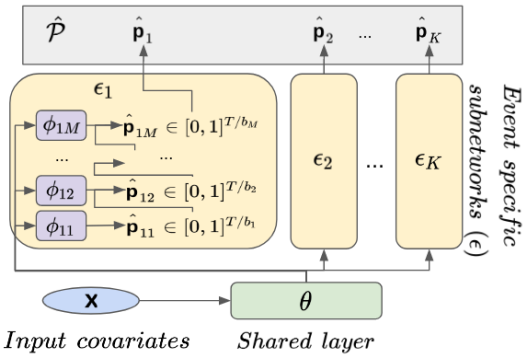


Figure 1: Proposed architecture. A multitask network that takes $\mathbf{x}$ as input and outputs $\hat{\mathcal{P}}$, a set of probability distributions for each event. Event specific subnetworks, ϵ_k, estimate the probability distribution, $\hat{\mathbf{p}}_k$, hierarchically. At granularity m, $\hat{\mathbf{p}}_{km}$ is constructed from $\phi_{km}(\theta(\mathbf{x}))$ and $\hat{\mathbf{p}}_{km-1}$.

Overall Architecture

The components of our approach are incorporated into a multitask architecture (**Figure 1**) with a shared layer θ to account for dependencies among events. It contains K event specific subnetworks, $\epsilon_1, \epsilon_2, ..., \epsilon_K$, where ϵ_k takes the output from θ as input and constructs $\hat{\mathbf{p}}_k$ in a hierarchical manner, as described below. From $\hat{\mathcal{P}}$, the survival curves, $\hat{\mathcal{S}}$, can be derived as $s_k^{(i)}[t] = 1 - \sum_{u=1}^{t-1} p_k^{(i)}[u]$ for time point t.

Hierarchical Component. We iteratively predict whether the event will occur at increasingly finer time scales and hypothesize that it improves performance through the decreased variance of the coarser scales. In the example from **Figure 2** ($T = 4$, $M = 3$), we aim to predict the discrete probability of event occurrence at $t = 1$ (i.e., $P(e = 1)$). Instead of directly predicting $P(e = 1)$, we begin at the coarsest granularity ($m = 1$, $b_1 = 4$ and $\tau_1^{(1)} = \{1, 2, 3, 4\}$), and predict $P(e \in \tau_1^{(1)})$. Then, we move to $m = 2$ ($b_2 = 2$, $\tau_1^{(2)} = \{1, 2\}$, and $\tau_2^{(2)} = \{3, 4\}$). We predict the probability that the event occurs in the first half of the horizon, $\tau_1^{(2)}$, conditioned on it occurring within the horizon (i.e., $P(e \in \tau_1^{(2)} | e \in \tau_1^{(1)})$). Next, we move to $m = 3$ (the original task), where $b_3 = 1$. We then predict $P(e \in \tau_1^{(3)} | e \in \tau_1^{(2)}) = P(e = 1 | e \in \tau_1^{(2)})$ and recover $P(e = 1) = P(e = 1 | e \in \tau_1^{(2)}) P(e \in \tau_1^{(2)} | e \in \tau_1^{(1)}) P(e \in \tau_1^{(1)})$.

Claim. Predicting the guiding task is a lower variance task compared to predicting the original task directly.

Justification. Assume for the original task that an individual's time-to-event, e, is a random variable drawn from some discrete distribution conditioned on their covariates $\mathbf{x}$, where the probability of each time point occurring is non-zero. Let $\mathbb{E}_O(e)$ and $Var_O(e)$ be the expected value and variance of e, respectively, under the original task, O. Next, consider

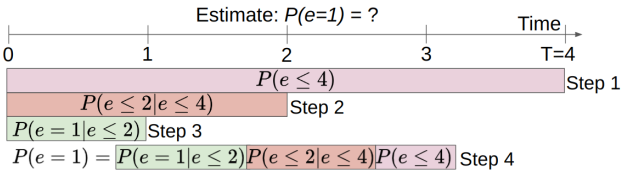


Figure 2: Example prediction. We predict if the event will occur within the horizon, and then iteratively predict at finer time scales until reaching the original time scale.

a guiding task, G, that uses coarser grained time bins. For coarse grained time bin τ_z, $z = 1, 2, ..., T/b$, $P(e \in \tau_z) = \sum_{u \in \tau_z} P(e = u)$. Time steps from the original task that fall into the same coarse grained time bin are equivalent under the guiding task, and $Var_G(e) < Var_O(e)$, where $Var_G(e)$ is the variance of e under G.

$$\begin{aligned} Var_G(e) &= \sum_{z=1}^{T/b} P(e \in \tau_z)(z - \mathbb{E}_G(e))^2 \\ &= \sum_{t=1}^{T} P(e = t)(\lfloor t/b \rfloor - \lfloor \mathbb{E}_O(e)/b \rfloor)^2 \\ &< \sum_{t=1}^{T} P(e = t)(t - \mathbb{E}_O(e))^2 = Var_O(e) \end{aligned}$$

We prove the inequality in the Supplement.

Claim. Predicting the original task conditioned on a guiding task is a lower variance task compared to predicting the original task directly in expectation.

Justification. By the law of total variance, where $\mathbb{E}_{O|G}$ and $Var_{O|G}$ denote the expectation and variance of e with respect to the original task conditioned on the guiding task,

$$\begin{aligned} &Var_O(e) = \mathbb{E}(Var_{O|G}(e)) + Var(\mathbb{E}_{O|G}(e)) \\ &\implies \mathbb{E}(Var_{O|G}(e)) < Var_O(e) \end{aligned}$$

since $Var_O(e) \neq 0$ from our assumption on the distribution of e. We later demonstrate empirically that using predictions from the guiding task and the original task conditioned on the guiding task improve the predictions of the original task.

The hierarchical component is implemented within each ϵ subnetwork (**Figure 1**) as described in **Algorithm 1**. For event k, we first predict at the coarsest granularity (i.e., the horizon) via ϕ_{k1} (line 1). We then consider grains $m = 2, 3, ..., M$ iteratively (line 2). For granularity m, we compute the conditional probability distribution (line 3; e.g., $P(e_k \in \tau_1^{(m)} | e_k \in \tau_1^{(m-1)})$), and recover the marginal probabilities for each time bin (e.g., $P(e_k \in \tau_1^{(m)})$) via multiplication with the corresponding marginal probability from granularity $m - 1$ (e.g., $P(e_k \in \tau_1^{(m-1)})$) (lines 4-6). After reaching granularity M, we will have computed corresponding the probabilities of the original task (lines 7-8).

For each event e, we account for individuals who do not experience event e by time T by predicting $\hat{P}(e \in \tau_1^{(1)}) = \hat{P}(e \leq T)$. We implement this within the ϵ subnetworks as part of the probability predictions.

Training Loss

We learn the model parameters by minimizing a composite loss $L = L_{TTE} + \alpha L_g$. Scalar hyperparameter $\alpha \in \mathbb{R}^+$ controls the tradeoff between the losses.

Algorithm 1 Hierarchical prediction for event k. $\hat{\mathbf{p}}_{km}$ is the predicted probability distribution at granularity m.

Input: $\theta(\mathbf{x})$, learned representation from θ
Output: $\hat{\mathbf{p}}_k$, estimated distribution of occurrence at the original time scale

Hierarchical_Prediction($\theta(\mathbf{x})$)

1: $\hat{\mathbf{p}}_{k1} \leftarrow \phi_{k1}(\theta(\mathbf{x}))$ ▷ $\hat{P}(e \leq T)$
2: **for** $m = 2$ to M **do**
3: $\quad \hat{\mathbf{p}}_{km} \leftarrow \phi_{km}(\theta(\mathbf{x}))$ ▷ conditional probabilities
4: $\quad$ **for** $z = 1$ to T/b_m **do**
5: $\quad\quad z_{m-1} \leftarrow \lfloor (z-1)b_m/b_{m-1} \rfloor + 1$ ▷ time index
6: $\quad\quad \hat{\mathbf{p}}_{km}[z] \leftarrow \hat{\mathbf{p}}_{km}[z] \times \hat{\mathbf{p}}_{km-1}[z_{m-1}]$ ▷ marginal
7: $\hat{\mathbf{p}}_k \leftarrow \hat{\mathbf{p}}_{kM}$
8: **return** $\hat{\mathbf{p}}_k$

$L_{TTE} = \left(\sum_{m=1}^{M} -L_{TTE}^m\right)$ maximizes the likelihood of each event when it occurs, for uncensored individuals, and the likelihood of survival for each individual at each event until their last observation time, for censored individuals. Unlike previous work (Lee et al. 2018; Ren et al. 2019), *we compute L_{TTE} over multiple time scales* (**Eq. 1**). $\mathbb{1}(a)$ is an indicator for whether condition a is true.

$$\begin{aligned} L_{TTE}^m = \sum_{i=1}^{N} \sum_{k=1}^{K} &\mathbb{1}\left(c_k^{(i)} == 0\right) log\left(\hat{p}_{km}^{(i)}[o_{km}^{(i)}]\right) \\ &+ \mathbb{1}\left(c_k^{(i)} == 1\right) log\left(1 - \left(\sum_{t=1}^{o_{km}^{(i)}} \hat{p}_{km}^{(i)}[t]\right)\right) \end{aligned} \tag{1}$$

$L_g = \sum_{k=1}^{K} L_g^k$ maximizes global consistency and further improves performance by encouraging the model to correctly rank the relevant pairs of individuals. For event k, let $\mathcal{C}_k$ be the set of comparable individuals (defined in the Supplement). *We compute L_g at the finest time scale for each event* (**Eq. 2**) where $F(s_1, s_2, \sigma) = exp((s_1 - s_2)/\sigma)$, $o*_k^{(j)} = o_k^{(j)}$ if $c_k^{(j)} = 1$ and $o*_k^{(j)} = o_k^{(j)} - 1$ otherwise (explained below). σ_g is a scalar hyperparameter.

$$\begin{aligned} L_g^k = \sum_{i,j \in \mathcal{C}_k} &F\left(\hat{s}_k^{(i)}[o_k^{(i)}], \hat{s}_k^{(j)}[o_k^{(i)}], \sigma_g\right) \\ &+ F\left(\hat{s}_k^{(i)}[o*_k^{(j)}], \hat{s}_k^{(j)}[o*_k^{(j)}], \sigma_g\right) \end{aligned} \tag{2}$$

Previous work (Lee et al. 2018) only uses the forward consistency (C-index) term of L_g^k (i.e., the first term of L_g^k). When evaluating global rankings, comparable individuals are compared at the earlier observation point only. This ignores the time steps between the two observation points, and could lead to inconsistent rankings (**Figure 3**).

Claim. Let $i = 1, 2$ be a comparable pair of individuals with observed times, $o^{(1)}$ and $o^{(2)}$, respectively, such that $o^{(1)} < o^{(2)}$. Comparing at time $o^{(1)}$ only is insufficient.

Justification. During the time between $o^{(1)}$ and $o^{(2)}$, $i = 1$ experienced the event, while $i = 2$ did not. The last time point at which this is known to be true is $o^{(2)}$ if $c^{(2)} = 1$

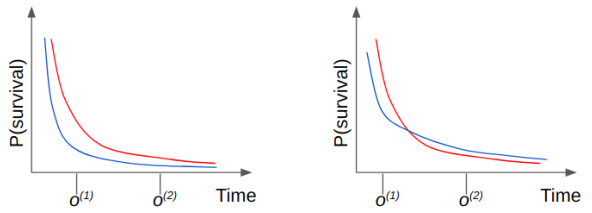


Figure 3: Consistency example. Individuals $i = 1, 2$ are shown by the blue and red curves, with time to event $o^{(1)}$ and observed time $o^{(2)}$, respectively. Left: curves are consistent at $o^{(1)}$ and $o^{(2)}$. Right: curves are inconsistent at $o^{(2)}$.

and $o^{(2)} - 1$ otherwise (i.e., $o*^{(2)}$). Thus, at $o*^{(2)}$, we expect that $s^{(1)}[o*^{(2)}] < s^{(2)}[o*^{(2)}]$. This is captured by the second term of **Eq. 2**. Thus, we encourage the network to correctly rank individuals throughout the horizon.

Censorship Due to (Semi-)Competing Events

Until now, we considered the general multi-event case, where there were no constraints on the ordering of events. This assumed that all events occur eventually (i.e., $s[t] \to 0$ as $t \to \infty$). However, this is not true in some cases of informative censoring, where the occurrence of one event can be prevented by another, such as with competing (Fine and Gray 1999) and semi-competing (Fine, Jiang, and Chappell 2001) events. For example, cancer relapse and death are semi-competing events, where death prevents the occurrence of relapse but not vice versa. For relapse, we account for this constraint by predicting the probability that it will never occur, even past $t = T$. For events that never occur, $e = null$.

We implement this by adding another level of granularity to our predictions. Instead of first predicting $P(e \leq T)$, we first predict $P(e \neq null)$ and $P(e \leq T | e \neq null)$. We then multiply to obtain $P(e \leq T)$. Within L_{TTE}, we maximize $1 - P(e \neq null)$ instead of $1 - \sum_{t=1}^{o} \hat{p}[t]$. Within L_g, $o*$ takes the value T instead of o or $o - 1$.

Experimental Setup

We evaluate our proposed approach across different publicly available datasets and compare to several baselines. We hypothesize that our hierarchical approach to survival analysis will lead to consistent improvements across a range of evaluation metrics, compared to existing approaches.

Datasets

We used four datasets. One was a multi-event synthetic dataset for which we know ground truth, facilitating evaluation. The remainder were real, publicly available datasets from the health domain. Of the real datasets, one of them contained competing risks while the other two were semi-competing risks. Outcome rates are reported in **Table 1**.

Synthetic: this dataset is based on a synthetic dataset from previous work (Lee et al. 2018) and contains two events, where both can occur in any order. It serves as a sanity check. We generated 5,000 individuals, each with 15 covariates. The covariates and time-to-events were drawn from the distributions described below. $\mathbf{X}_{y:z}$ denotes $\mathbf{X}$ from covariate indexes y to z, inclusive. U denotes a uniform distribution. Given the time to events, a time horizon was chosen at the 50^{th} percentile among the times of the first event experienced to allow for an equal proportion of individuals who experience no event and at least one event. We discretized all observed times into 20 evenly spaced bins.

$$\mathbf{X}_{1:5} \sim U(-5, 5)^5, \mathbf{X}_{6:15} \sim U(-10, 10)^{10}$$
$$u_1 \sim Lognormal((\mathbf{1}^T|\mathbf{X}_{1:5}|)^2 + (\mathbf{1}^T|\mathbf{X}_{6:10}|)^2, 0.4)$$
$$u_2 \sim Lognormal((\mathbf{1}^T|\mathbf{X}_{1:5}|)^2 + (\mathbf{1}^T|\mathbf{X}_{11:15}|)^2, 0.4)$$
$$v_1 \sim u_1 + Lognormal(0.5(\mathbf{1}^T|\mathbf{X}_{1:5}|)^2, 0.4)$$
$$v_2 \sim u_2 + Lognormal(0.5(\mathbf{1}^T|\mathbf{X}_{1:5}|)^2, 0.4)$$
$$e_1 = u_1 \ if \ u_1 < u_2, \ v_1 \ otherwise$$
$$e_2 = u_2 \ if \ u_2 < u_1, \ v_2 \ otherwise$$

Dataset	N	D	K	T	Event Sequence: %
Synthetic	5,000	15	2	20	o_1: 20.9, o_2: 20.4 o_1, o_2: 4.4, o_2, o_1: 4.3
ADNI	1,604	965	2	60	AD: 19.3, D: 0.5 AD, D 0.6
MIMIC -III	13,801	3,822	3	12	A: 5.2, S: 1.5, D: 4.3 A, S: 5.0, A, D: 0.7 A, S, D: 1.9 S, A: 0.3, S, D: 0.5 S, A, D: 0.2
SEER	11,374	689	2	120	DC: 6.9, DP: 8.2

Table 1: Dataset summaries. T is in months for ADNI and SEER, and hours for MIMIC-III. 'D', 'A', 'S', and 'DC'/'DP' mean death, ARF, shock, and death from larynx cancer/pulmonary disease, respectively.

ADNI: a publicly available dataset containing data on Alzheimer's disease (AD)[1]. We considered the semi-competing outcomes of AD onset and death as events. We selected a cohort of 1,604 participants, excluding left censored individuals. We considered a prediction horizon of five years, where time was measured relative to the individual's first encounter with the ADNI study, and we predicted at the monthly level ($T = 60$). After preprocessing (Supplement), each individual had 965 covariates. Overall, 28.7% of the population was event free by $t = 60$, and 50.8% of the population did not experience an event due to loss of followup.

MIMIC-III: a publicly available dataset of electronic health record data (Johnson et al. 2016). Unlike ADNI, the data cover a smaller time scale (e.g., hours rather than years).

[1]Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

We considered three semi-competing outcomes: 1) acute respiratory failure (ARF), 2) shock, and 3) in hospital mortality (death). MIMIC-III includes data pertaining to vital signs, medications, diagnostic and procedure codes, and laboratory measurements. Each event was defined as described by Oh et al. (Oh et al. 2019). We considered a prediction horizon of 12 hours, where time was measured relative to six hours after the start of an ICU (intensive care unit) encounter, and we predicted at the hourly level ($T = 12$). We used data from the first 30 minutes of the ICU encounter to avoid label leakage, excluding encounters where an event occurred within the first six hours. This resulted in a sample of 13,801 ICU encounters corresponding to 10,947 patients. After preprocessing (Supplement), each encounter was associated with 3,822 covariates. Overall, 76.3% of the population was event free by $t = 12$, and 4.1% of the population did not experience an event due to loss of followup.

SEER: (https://seer.cancer.gov/data/): a publicly available dataset containing cancer incidence data from population-based registries, used by previous work to study competing risks (Lee et al. 2018; Zhang and Zhou 2018). We considered the competing outcomes of death due to larynx cancer and death due to pulmonary disease. We selected a cohort of 11,374 participants between ages 60-65, excluding left censored individuals. We considered a prediction horizon of 10 years, where time was measured relative to the individual's age, and we predicted at the monthly level ($T = 120$). Covariates included features relating to demographics, tumor behavior, and tumor characteristics. After preprocessing (Supplement), each individual was associated with 689 covariates. Overall, 21.0% of the population was event free by $t = 120$, and 63.9% of the population did not experience an event due to loss of followup.

Baselines

We compare our approach with the following baselines. We also perform an ablation study as outlined in **Section 5.2**.

Independent: This approach learns a separate model for each event and assumes that the time-to-events among all events are independent. This baseline, though simplistic, is most commonly used in clinical research (Felker et al. 2017). Our implementation uses a series of separate feed forward networks. It is trained to predict at the given time scale only, minimizing the forward consistency term of L_g.

DeepHit: This approach (Lee et al. 2018) models all events simultaneously in a single model, but was designed for competing risks (where the occurrence of one event prevents the occurrence of the other events) in that it assumes $\sum_{t=1}^{T}\sum_{k=1}^{K} p_k[t] = 1$. It accounts for dependencies among events via θ but not for individuals who experience no event by time T. It is trained using the forward consistency term of L_g and predicts at the given time scale only. For the non-competing risks datasets, we adapt DeepHit by instead assuming $\sum_{t=1}^{T} p_k[t] = 1$ for $k = 1, 2, ..., K$. For the semi-competing risks datasets, we also predict $P(e \neq null)$. This baseline allows us to explore whether approaches for competing risks generalize to other multi-event settings.

Oracle: An 'approach' with access to the ground truth time-to-event distributions on the synthetic data. It provides an upper bound for performance.

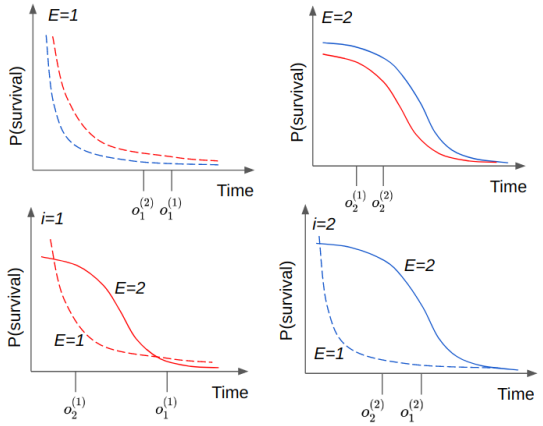


Figure 4: Local consistency. The red and blue curves represent individuals $i = 1, 2$, and the dashed and solid curves represent events $E = 1, 2$, respectively. The curves are globally (top), but not locally (bottom) consistent. $i = 1$: inconsistent at $o_2^{(1)}$. $i = 2$: inconsistent at $o_1^{(2)}$ and $o_2^{(2)}$.

Evaluation

We evaluate using measures of both global and local consistency. For global evaluation, we report the average C-index, as in previous work (Lee et al. 2018; Katzman et al. 2018), and average global consistency across all events. For event, k, we compute the global consistency as in **Eq. 3**. $|\mathcal{C}_k|$ is the size of $\mathcal{C}_k$ (i.e., the number of comparable individuals). Let $\mathcal{T}_k(i,j)$ (defined in the Supplement) be the set of comparable time points for individuals i and j with size $|\mathcal{T}_k(i,j)|$.

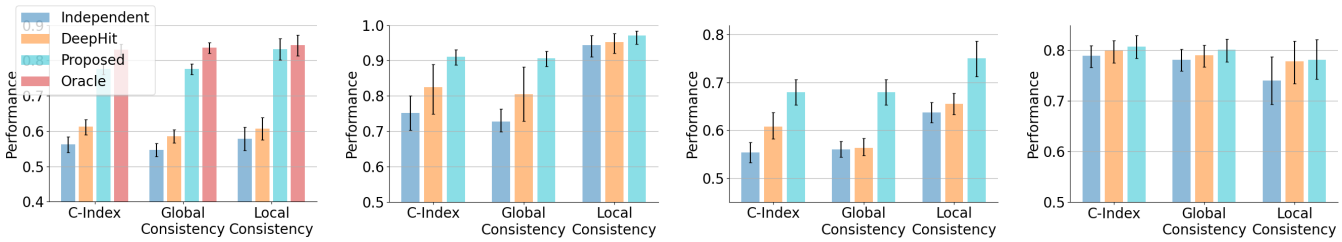
$$\frac{1}{|\mathcal{C}_k|}\sum_{i,j\in\mathcal{C}_k}\frac{1}{|\mathcal{T}_k(i,j)|}\sum_{t\in\mathcal{T}_k(i,j)}\mathbb{1}(\hat{s}_k^{(i)}[t] < \hat{s}_k^{(j)}[t]) \quad (3)$$

Unlike past work, we measure local consistency, as it is critical that multi-event models can discriminate among events *within* an individual (**Figure 4**). We compute it for individual i as in **Eq. 4**. Let $\mathcal{C}^i$ be the set of comparable events with size $|\mathcal{C}^i|$, and let $\mathcal{T}^i(j,k)$ be the set of comparable time points for events j and k with size $|\mathcal{T}^i(j,k)|$ (defined in the Supplement). We report the average among all individuals.

$$\frac{1}{|\mathcal{C}^i|}\sum_{j,k\in|\mathcal{C}^i|}\frac{1}{|\mathcal{T}^i(j,k)|}\sum_{t\in\mathcal{T}^i(j,k)}\mathbb{1}(\hat{s}_j^{(i)}[t] < \hat{s}_k^{(i)}[t]) \quad (4)$$

Implementation Details

For each dataset, we randomly split the data into 60/20/20% training/validation/test, and data from the same individual did not appear across splits. All models were trained in Python3.6 and Pytorch (Paszke et al. 2017), using Adam (Kingma and Ba 2014). Hyperparameters, including the learning rate, L2 regularization constant, and objective function scalars (e.g., α), were tuned using a random grid search, with a budget of 20. We used early stopping based on validation set performance, where we



(a) Synthetic. Multi-event. (b) ADNI. Semi-competing. (c) MIMIC-III. Semi-competing. (d) SEER. Competing risks.

Figure 5: Comparison to existing approaches. Our approach outperforms state-of-the-art approaches in different multi-event settings across different metrics. Error bars represent empirical 95% confidence intervals.

aimed to maximize the average of the proposed global and local consistencies. All network layers were initialized with Xavier initialization from a uniform distribution. We report results on the held-out test set, with error bars representing empirical 95% confidence intervals (CI) from 1,000 bootstrapped samples. More details about the code (https://gitlab.eecs.umich.edu/mld3/hierarchical-survival-analysis) are given in the Supplement.

Results and Takeaways

Here, we evaluate the proposed approach on four datasets.

How Does the Proposed Approach Perform?

As shown in **Figure 5**, the proposed method noticeably outperforms both baselines on all datasets. For SEER, the proposed method achieved a global and local consistency of 0.80 (95% CI=0.77-0.82) and 0.78 (95% CI=0.74-0.82), while using DeepHit achieved a global and local consistency of 0.79 (95% CI=0.77-0.82) and 0.77 (95% CI=0.73-0.81), respectively. The poorer performance of our adaptation of DeepHit to the non-competing risks datasets suggests that it does not generalize well to other settings.

Do Hierarchical Predictions Help?

To assess the extent to which the hierarchical component improved performance, we performed an ablation study. Our main ablations are as follows, with a more thorough evaluation in the Supplement. To control for the number of network parameters, the overall network size was kept constant.

DeepHit(A) (adapted) (DA): We generalized previous work to the multi-event setting (Lee et al. 2018) and implemented it as described in our adaptation of DeepHit to the non-competing risks datasets. **Proposed minus hierarchical** (-HA): We built from DA such that L_g penalizes at the first *and last* relevant time points, and we account for individuals who remain event free by time T. It assesses the proposed approach without the hierarchical component. Our **proposed method** (P) builds from -HA in that we utilize the hierarchical architecture and loss function.

Results are shown in **Table 2**. First, notice that the addition of *the hierarchical component improves* over the baselines across all datasets and all evaluation metrics. The magnitude of improvement is larger in MIMIC-III than the other real datasets. We hypothesize that this is due to the increased difficulty from modeling three events instead of two. The larger performance gains in local consistency compared to the other metrics on most datasets also show the that proposed approach is effective in modeling interdependencies among events. Second, increased supervision over L_g and accounting for individuals who remain event free by time T led to large improvements in performance. This is likely because over 20% of individuals in each dataset were event free throughout the horizon.

Dataset	Apr	C-Index	Global Consistency	Local Consistency
Synthetic	DA	.61(.59-.63)	.58(.56-.60)	.60(.57-.63)
	-HA	.73(.71-.75)	.73(.72-.75)	.77(.74-.81)
	P	**.77**(.76-.79)	**.77**(.76-.79)	**.83**(.80-.86)
ADNI	DA	.82(.74-.88)	.80(.72-.78)	.95(.92-.97)
	-HA	.90(.88-.92)	.89(.87-.91)	.95(.90-.98)
	P	**.91**(.88-.93)	**.90**(.88-.92)	**.97**(.94-.98)
MIMIC -III	DA	.60(.58-.63)	.56(.54-.58)	.65(.63-.67)
	-HA	.66(.63-.69)	.66(.63-.69)	.74(.71-.78)
	P	**.68**(.65-.70)	**.68**(.65-.70)	**.75**(.71-.78)
SEER	DA	.79(.76-.81)	.78(.75-.80)	.75(.71-.80)
	-HA	.79(.77-.81)	.79(.76-.81)	.76(.72-.80)
	P	**.80**(.78-.83)	**.80**(.77-.82)	**.78**(.74-.82)

Table 2: Ablation study. DA adapts DeepHit to the multi-event setting. -HA augments L_g and accounts for individuals who remain event free by time T. The highest values are bolded. Error bars show empirical 95% confidence intervals. 'Apr' means approach.

Conclusion

We introduced a novel approach for multi-event survival analysis. It utilizes a novel hierarchical structure that divides the task of predicting event occurrence over a time horizon T into a series of simpler, yet practically relevant, subtasks that iteratively build on one other. On both synthetic and real datasets, we showed that the proposed method led to improvements over well-accepted baselines across different performance metrics. Through ablations, we demonstrated the effectiveness of the hierarchical component. Going forward, one might consider extensions for recurrent events. Nonetheless, this work represents an important foundation for future work in multi-event survival analysis.

Acknowledgments

This work was supported by the National Science Foundation (NSF award no. IIS-1553146). The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation. We would also like to thank the anonymous reviewers for their valuable feedback.

For the ADNI dataset: Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Ethics Statement

With respect to the content and overall goals of this paper, we do not believe that there are any ethical concerns involved. If implemented carefully and correctly, our work has the potential to improve clinical care practices. For example, being able to correctly rank patients' risk for a condition may aid clinicians in resource allocation. Additionally, being able to correctly rank which conditions a patient is likely to experience may aid clinicians in deciding the best treatment plan. However, in the context of developing a model for the purpose of integration into clinical practice in general, we outline some ethical considerations.

Recent work has shown that machine learning models developed from clinical data, including those from electronic health records, can enforce harmful systemic biases from the clinical setting if not carefully implemented (Obermeyer et al. 2019). As a result, it is important to consider not only overall model performance, but also how the outcome is measured (i.e., choice of labels) and how model performance varies across different sub-groups of the population. This is especially important for under-represented minorities and individuals who may not have adequate access to healthcare services.

References

Andersen, P. K.; and Gill, R. D. 1982. Cox's regression model for counting processes: a large sample study. *The annals of statistics* 1100–1120.

Andersen, P. K.; and Keiding, N. 2002. Multi-state models for event history analysis. *Statistical methods in medical research* 11(2): 91–115.

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

Armstrong, G. T.; Kawashima, T.; Leisenring, W.; Stratton, K.; Stovall, M.; Hudson, M. M.; Sklar, C. A.; Robison, L. L.; and Oeffinger, K. C. 2014. Aging and risk of severe, disabling, life-threatening, and fatal events in the childhood cancer survivor study. *Journal of clinical oncology* 32(12): 1218.

Castañeda, J.; and Gerritse, B. 2010. Appraisal of several methods to model time to multiple events per subject: modelling time to hospitalizations and death. *Revista Colombiana de Estadística* 33(1): 43–61.

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

Engelhard, M.; Berchuck, S.; D'Arcy, J.; and Henao, R. 2020. Neural Conditional Event Time Models. In *MLHC*.

Felker, G. M.; Anstrom, K. J.; Adams, K. F.; Ezekowitz, J. A.; Fiuzat, M.; Houston-Miller, N.; Januzzi, J. L.; Mark, D. B.; Piña, I. L.; Passmore, G.; et al. 2017. Effect of natriuretic peptide–guided therapy on hospitalization or cardiovascular mortality in high-risk patients with heart failure and reduced ejection fraction: a randomized clinical trial. *JAMA* 318(8): 713–720.

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.

Fine, J. P.; Jiang, H.; and Chappell, R. 2001. On semi-competing risks data. *Biometrika* 88(4): 907–919.

Harrell, F. E.; Califf, R. M.; Pryor, D. B.; Lee, K. L.; and Rosati, R. A. 1982. Evaluating the yield of medical tests. *JAMA* 247(18): 2543–2546.

Heagerty, P. J.; and Zheng, Y. 2005. Survival model predictive accuracy and ROC curves. *Biometrics* 61(1): 92–105.

Hsieh, J.-J.; and Wang, J.-L. 2018. Quantile residual life regression based on semi-competing risks data. *Journal of Applied Statistics* 45(10): 1770–1780.

Jiang, F.; and Haneuse, S. 2017. A semi-parametric transformation frailty model for semi-competing risks survival data. *Scandinavian Journal of Statistics* 44(1): 112–129.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3: 160035.

Katzman, J. L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; and Kluger, Y. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology* 18(1): 24.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kowsari, K.; Brown, D. E.; Heidarysafa, M.; Meimandi, K. J.; Gerber, M. S.; and Barnes, L. E. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, 364–371. IEEE.

Lanza, E.; Muglia, R.; Bolengo, I.; Poretti, D.; D'Antuono, F.; Ceriani, R.; Torzilli, G.; and Pedicini, V. 2020. Survival analysis of 230 patients with unresectable hepatocellular carcinoma treated with bland transarterial embolization. *PloS one* 15(1): e0227711.

Larson, M. G. 1984. Covariate analysis of competing-risks data with log-linear models. *Biometrics* 459–469.

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*.

Malakouti, S.; and Hauskrecht, M. 2019. Hierarchical adaptive multi-task learning framework for patient diagnoses and diagnostic category classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 701–706. IEEE.

Meng, Y.; Shen, J.; Zhang, C.; and Han, J. 2019. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6826–6833.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464): 447–453.

Oh, J.; Wang, J.; Tang, S.; Sjoding, M. W.; and Wiens, J. 2019. Relaxed Parameter Sharing: Effectively Modeling Time-Varying Relationships in Clinical Time-Series. In *MLHC*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch .

Peng, M.; Xiang, L.; and Wang, S. 2018. Semiparametric regression analysis of clustered survival data with semi-competing risks. *Computational Statistics & Data Analysis* 124: 53–70.

Ren, K.; Qin, J.; Zheng, L.; Yang, Z.; Zhang, W.; Qiu, L.; and Yu, Y. 2019. Deep recurrent survival analysis. In *Proc. AAAI*, 1–8.

Roe, M. T.; Armstrong, P. W.; Fox, K. A.; White, H. D.; Prabhakaran, D.; Goodman, S. G.; Cornel, J. H.; Bhatt, D. L.; Clemmensen, P.; Martinez, F.; et al. 2012. Prasugrel versus clopidogrel for acute coronary syndromes without revascularization. *New England Journal of Medicine* 367(14): 1297–1309.

Seo, P. H.; Kim, G.; and Han, B. 2019. Combinatorial Inference against Label Noise. In *Advances in Neural Information Processing Systems*, 1173–1183.

Solomon, S. D.; Rizkala, A. R.; Gong, J.; Wang, W.; Anand, I. S.; Ge, J.; Lam, C. S.; Maggioni, A. P.; Martinez, F.; Packer, M.; et al. 2017. Angiotensin receptor neprilysin inhibition in heart failure with preserved ejection fraction: rationale and design of the PARAGON-HF trial. *JACC: Heart Failure* 5(7): 471–482.

Wang, L.; Li, Y.; Zhou, J.; Zhu, D.; and Ye, J. 2017. Multi-task survival analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, 485–494. IEEE.

Wongvibulsin, S.; Wu, K. C.; and Zeger, S. L. 2020. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Medical Research Methodology* 20(1): 1–14.

Zhang, Q.; and Zhou, M. 2018. Nonparametric Bayesian Lomax delegate racing for survival analysis with competing risks. In *Advances in Neural Information Processing Systems*, 5002–5013.