Supplement for: A Hierarchical Approach to Multi-Event Survival Analysis

Donna Tjandra,[1] **Yifei He,** [1] **Jenna Wiens** [1]
for in part the Alzheimer's Disease Neuroimaging Initiative*
[1]Computer Science and Engineering, University of Michigan, Ann Arbor MI, USA
dotjandr, heyifei, wiensj@umich.edu

Here we describe additional details to aid in the reproducibility of this paper. We begin by providing further justification for the hierarchical component. We then describe how our data were preprocessed as well as implementation details. This includes hyperparameter values considered and architecture details.

1 Extended Hierarchical Justification

We show that $\sum_{t=1}^{T} P(o = t)(\lfloor t/b \rfloor - \lfloor \mathbb{E}_O(e)/b \rfloor)^2 < \sum_{t=1}^{T} P(e = t)(t - \mathbb{E}_O(e))^2$. Here, $d_1, d_2 \in \mathbb{R}^+$ and $0 \leq d_1, d_2, < 1$. Let $Y = \lfloor t/b \rfloor - \lfloor \mathbb{E}_O(e)/b \rfloor$.

$$\begin{aligned} t - \mathbb{E}_O(e) &= b(t/b - \mathbb{E}_O(e)/b) \\ &= b(\lfloor t/b \rfloor - \lfloor \mathbb{E}_O(e)/b \rfloor + (d_1 - d_2)) \\ &\in (b(Y-1), b(Y+1)) \end{aligned} \tag{1}$$

By **Eq.** 1, $t - \mathbb{E}_O(e) > b(Y-1)$ and $t - \mathbb{E}_O(e) < b(Y+1)$. Therefore,

$$\begin{aligned} & t - \mathbb{E}_O(e) > b(Y-1) \\ & \quad \implies Y < ((t - \mathbb{E}_O(e))/b) + 1 \\ & t - \mathbb{E}_O(e) < b(Y+1) \\ & \quad \implies Y > ((t - \mathbb{E}_O(e))/b) - 1 \end{aligned} \tag{2}$$

The last line from **Eq.** 2 breaks into three cases.

- Case 1: $t - \mathbb{E}_O(e) \geq 2$. Then Y is of smaller magnitude since $0 \leq Y < ((t - \mathbb{E}_O(e))/b) + 1 \leq t - \mathbb{E}_O(e)$.
- Case 2: $-1 \leq t - \mathbb{E}_O(e) \leq 1$. Then the magnitude of Y is smaller or the same since $0 \leq Y, t - \mathbb{E}_O(e) \leq 1$.
- Case 3: $t - \mathbb{E}_O(e) < -2$. Then the magnitude of Y is smaller since $0 > Y > ((t - \mathbb{E}_O(e))/b) + 1 \geq t - \mathbb{E}_O(e)$.

As a result, $|\lfloor t/b \rfloor - \lfloor \mathbb{E}_O(e)/b \rfloor| \leq t - \mathbb{E}_O(e)$. To finish the justification, note that there are b distinct values of t for which $\lfloor t/b \rfloor = \lfloor \mathbb{E}_O(e) \rfloor$. On the other hand, there is only one value of t for which $t = \mathbb{E}_O(e)$. This results in more zero terms in the calculation for $Var_G(e)$. Combined with $|\lfloor t/b \rfloor - \lfloor \mathbb{E}_O(e)/b \rfloor| \leq |t - \mathbb{E}_O(e)|$, we have verified the claim from the beginning of this section.

2 Data Preprocessing Details

Here we more thoroughly describe our preprocessing procedure for all of our datasets.

2.1 Synthetic

The distributions from which the time to events were drawn were chosen such that the skew of the time to event distributions could be controlled, since time to event in real data is often skewed. Given our continuous valued time to events, we discretized all time to events into 20 equally spaced temporal segments. We included more variability in the last 10 covariates of $\mathbf{X}$ to allow for more variation in the time-to-event distribution. The squared terms introduce non-linear relationships between the covariates and time to event. The dataset was also constructed such that the time to the second event is conditioned on the first event, introducing dependencies among events. More specifically, the second time to event is taken as the corresponding v value instead of the corresponding u value had it occurred first.

2.2 ADNI

Covariates, including cognitive test scores, AD-specific biomarkers, medical imaging data, and demographics were pre-processed such that continuous covariates were discretized based on quintiles, discrete covariates were encoded as one-hot vectors, and infrequent (frequency=1) or overly-frequent (frequency=1,603) covariates were removed.

Unless otherwise mentioned, all data was taken from the 'ADNIMERGE' table. Individuals were labeled as follows. AD onset was taken as the first encounter where the value of the 'DX' column was 'Dementia', where the time to onset was taken as the value in the 'Month' column. Date of death was taken as the value of the 'NPDOD' column in the 'NEUROPATH_04_12_18' table, and the time to death was taken as the difference between the date of death and the date of the first visit, converted to months. The date of the first visit was taken as the value of the 'update_timestamp' column

*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

at which the corresponding value in the 'Month' column was zero. Individuals whose time to AD onset or death was greater than five years were considered as horizon censored at 60 months. For those who did not experience either event, the last point of observation was taken as the minimum of the largest corresponding value in the 'Months' column and 60 months. Additionally, we discarded the symbols '$<$' and '$>$' from all numerical covariates whose value included such a symbol. For example, if a value was recorded as '> 1.1', we took the value of 1.1. This was done before binning continuous covariates into quintiles as described earlier.

2.3 MIMIC-III

Data were extracted from the following tables: PATIENTS, ADMISSIONS, ICUSTAYS, CHARTEVENTS, LABEVENTS, MICROBIOLOGYEVENTS, OUTPUTEVENTS, DATETIMEEVENTS, INPUTEVENTS_MV, and PROCEDUREEVENTS_MV, and were processed using the FlexIble Data Driven pipeLinE (FIDDLE), (Tang et al. 2020), a publicly available pre-processing tool for electronic health record data. To preprocess the tables from MIMIC-III into input for FIDDLE, we used the procedure from previous work. We then used the default settings of FIDDLE (theta_1 = 0.001, theta_2 = 0.001, and theta_freq = 1.0). In addition, we used T = 5 hours, and dt = 0.5 hours. As model input, we used the first slice of the output corresponding to the first 30 minutes of the ICU visit.

2.4 SEER

An individual's last observed time was taken according to the 'Survival months' feature. Individuals whose value was 9999 were excluded, as this indicated an unknown observed time. Cause of death was recorded from the 'Cause of Death to SEER site recode' feature. Only individuals whose value was '00000', '22020', or '50130' were included, as these codes corresponded to being alive, experiencing death due to larynx cancer, and death due to pulumonary disease, respectively. (I.e., we excluded individuals who died from causes other than the ones we considered).

We used the 'yr1975_2016.seer9/RESPIR.TXT' file as our source of data. Similar to ADNI, covariates were preprocessed such that continuous covariates were discretized based on quintiles, discrete covariates were encoded as one-hot vectors.

3 Network Implementation/Training Details

Here, we describe our ranges of hyperparameters and implementation choices for the proposed network. All networks were trained on Intel(R) Xeon(R) CPUs, E7-4850 v3 @ 2.20GHz.

3.1 Hyperparameter Values Considered

Here, we show the range of values we considered for our random grid search. All random seeds were initialized with a value of 0. This includes the Pytorch random seed, the numpy random seed, and the seed from Python's random. More details are provided in **Table 1**. For α, we found that values of 0.0001, 0.05, 0.0005, and 0.001 were optimal for the synthetic, ADNI, MIMIC-III, and SEER tasks, respectively. We considered the different ranges based on the magnitudes of the two loss terms (L_g was usually much larger than L_{TTE}, hence the small values for alpha). For any hyperparameters associated with the Adam optimizer not mentioned above, we used the default values. We divide our training data into five batches during training.

Table 1: For each dataset, we list the range of hyperparameters considered for each dataset. The upper and lower rows for each dataset represent the lower and upper bounds, respectively.

Dataset	Learning Rate	L2	α	σ_g
Synthetic	0.005	0.005	0.0001	10
	0.05	0.05	0.001	100
ADNI	0.00001	0.001	0.01	10
	0.001	0.5	0.1	100
MIMIC-III	0.00001	0.01	0.0001	10
	0.001	0.5	0.005	1000
SEER	0.0001	0.001	0.0001	10
	0.01	0.1	0.01	100

3.2 Architecture Details

For the overall architecture, we used one layer for the θ component of **Figure 1**, whose size was 20 on the synthetic dataset, 500 on the ADNI dataset, 500 on the MIMIC-II dataset, and 100 on the SEER dataset. For the first time scale, we used a layer size of 20 (synthetic), 100 (ADNI and MIMIC-III), and 50 (SEER) within the corresponding ϕ subnetwork. For the second time scale, we used a layer size of 20 (synthetic), 100 (MIMIC-III), and 50 (ADNI and SEER) within the corresponding ϕ subnetwork. The ReLU activation function was used. The complete implementation can be found in the attached code.

The network architecture was the same across all approaches for each dataset to control for the number of parameters. For the proposed hierarchical approach, we supervise at the 'yearly' and 'monthly' levels on ADNI and SEER. For MIMIC-III, we supervise at the 'two hour' and 'hour' levels. For the synthetic dataset, we split our time horizon into four non-overlapping bins, and then split each of those bins into five non-overlapping time bins.

3.3 Source Code

Our source code can be found in the attached zip file.

4 Additional Definitions

Here, we define additional sets from the manuscript.

- The set of comparable pairs of individuals i, j within event k. $\mathcal{C}_k := \{(i,j) | c_k^{(i)} = 0 \wedge (o_k^{(i)} < o_k^{(j)} \vee (o_k^{(i)} = o_k^{(j)} \wedge c_k^{(j)} = 1))\}$
- The set of comparable time points for individuals i, j within event k. $\mathcal{T}_k(i,j) := \{t | c_k^{(i)} = 0 \wedge (o_k^{(i)} \leq t < o_k^{(j)} \vee (o_k^{(i)} = t = o_k^{(j)} \wedge c_k^{(j)} = 1))\}$

Table 2: Extended ablation summary showing the benefits of including the hierarchical architecture and supervision along multiple time scales. The highest values are bolded. Error bars show empirical 95% confidence intervals.

Dataset	Method	C-Index	Global Consistency	Local Consistency
Synthetic	-hierarchical	.73 (.71-.75)	.73 (.72-.75)	.77 (.74-.81)
	+architecture	.76 (.74-.77)	.76 (.74-.77)	.82 (.79-.85)
	+supervision	.73 (.71-.74)	.73 (.72-.75)	.79 (.76-.82)
	Proposed	**.77** (.76-.79)	**.77** (.76-.79)	**.83** (.80-.86)
ADNI	-hierarchical	.90 (.88-.92)	.89 (.87-.91)	.95 (.90-.98)
	+architecture	**.91** (.89-.93)	**.90** (.89-.92)	.95 (.90-.98)
	+supervision	.90 (.88-.92)	**.90** (.88-.92)	.96 (.93-.98)
	Proposed	**.91** (.88-.93)	**.90** (.88-.92)	**.97** (.94-.98)
MIMIC-III	-hierarchical	.66 (.63-.69)	.66 (.63-.69)	.74 (.71-.78)
	+architecture	.67 (.65-.70)	.67 (.65-.70)	**.75** (.71-.78)
	+supervision	.67 (.64-.70)	.67 (.64-.70)	**.75** (.71-.78)
	Proposed	**.68** (.65-.70)	**.68** (.65-.70)	**.75** (.71-.78)
SEER	-hierarchical	.79 (.77-.81)	.79 (.76-.81)	.76 (.72-.80)
	+architecture	**.80** (.78-.82)	.79 (.77-.82)	.77 (.72-.81)
	+supervision	**.80** (.78-.82)	**.80** (.77-.82)	.77 (.73-.81)
	Proposed	**.80** (.78-.83)	**.80** (.77-.82)	**.78** (.74-.82)

- The set of comparable pairs of events j, k within individual i. $\mathcal{C}^i := \{(j,k)|c_j^{(i)} = 0 \wedge (o_j^{(i)} < o_k^{(i)} \vee (o_j^{(i)} = o_k^{(i)} \wedge c_k^{(i)} = 1))\}$
- The set of comparable time points for events j, k within individual i. $\mathcal{T}^i(j,k) := \{t|c_j^{(i)} = 0 \wedge (o_j^{(i)} \leq t < o_k^{(i)} \vee (o_j^{(i)} = t = o_k^{(i)} \wedge c_k^{(i)} = 1))\}$

5 Extended Ablation Study

We include an extension of the ablation study from **Table 2**. We build from -hierarchical to examine the benefits of the proposed architecture and additional supervision to L_{TTE}.

+architecture: This ablation builds from -hierarchical in that it utilizes the proposed architecture as opposed to a vanilla feed forward network. It serves to assess the effectiveness of the hierarchical architecture.

+supervision: This ablation builds from -hierarchical in that we supervise along multiple time scales instead of that from the original task. It serves to assess the effectiveness of the proposed approach without hierarchical supervision.

Our proposed method combines +architecture and +supervision in that we utilize the hierarchical architecture and supervise along all specified time scales instead of the original time scale. As shown in **Table 2**, *both the hierarchical architecture and supervision along multiple time scales improve performance*.

References

Tang, S.; Davarmanesh, P.; Song, Y.; Koutra, D.; Sjoding, M. W.; and Wiens, J. 2020. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association* .