Dynamic Survival Analysis with Individualized Truncated Parametric Distributions

Preston Putzel PPUTZEL@UCI.EDU
Padhraic Smyth SMYTH@ICS.UCI.EDU
Department of Computer Science, University of California, Irvine, CA, USA

Jaehong Yu JHYU@INU.AC.KR
Department of Industrial and Management Engineering, Incheon National University, 119 Academy-Ro, Yeonsu-Gu, Songdo-dong Incheon 22012, South Korea

Hua Zhong JUDY.ZHONG@NYULANGONE.ORG
Department of Population Health, NYU Grossman School of Medicine, New York, NY, USA

Abstract

Dynamic survival analysis is a variant of traditional survival analysis where time-to-event predictions are updated as new information arrives about an individual over time. In this paper we propose a new approach to dynamic survival analysis based on learning a global parametric distribution, followed by individualization via truncating and renormalizing that distribution at different locations over time. We combine this approach with a likelihood-based loss that includes predictions at every time step within an individual's history, rather than just including one term per individual. The combination of this loss and model results in an interpretable approach to dynamic survival, requiring less fine tuning than existing methods, while still achieving good predictive performance. We evaluate the approach on the problem of predicting hospital mortality for a dataset with over 6900 COVID-19 patients.

Keywords: Dynamic Survival Analysis, Personalized Predictions, Parametric Survival Analysis

1. Introduction

Survival analysis focuses on the analysis and modeling of time-to-event data. Traditional approaches to survival modeling in statistics typically construct global time-to-event distributions, or else make restrictive assumptions about the effect of an individual's covariates on their risk of an event, such as in proportional hazard models (Aalen et al., 2008). In the past few years new machine learning approaches relaxing these assumptions have been applied to survival analysis problems (Spooner et al., 2020; Nemati et al., 2020; Wang et al., 2019).

In this paper we describe a novel type of dynamic survival model which learns a global distribution over a continuous time-to-event, while using covariate information to locate an individual on the time-axis relative to this global density. This shifting procedure acts as a natural regularization to the model since it reduces the family of possible distributions to truncated (and renormalized) versions of the global density. This connection between global and individualized densities also allows interpretation of each individualized density

as the average density of individuals who have survived to the same shifted time along the global time-axis. Our approach also makes fuller use of the entire history of covariate information than existing methods by including predictions at every step within an individual's trajectory in the loss instead of only including a term at the last available measurement per individual. We also introduce a novel form of the dynamic C-index to more carefully evaluate a model's expected prognostic performance when used in a true clinical setting. We apply our model to electronic health record (EHR) data for patients hospitalized with COVID-19, learning individualized time-to-death distributions after hospitalization.

2. Related Work

Common statistical approaches to dynamic survival modelling include landmarking and joint modeling. Landmarking involves constructing a nested set of datasets at 'landmark' times and fitting a static survival model at each time point (Van Houwelingen, 2007; Parast et al., 2014). It can, however, be difficult from an interpretability viewpoint to understand the connection of the predicted risks to one another at the different landmark times. In joint modeling approaches, the probability distribution of the covariate trajectory, and the event of interest are modelled with a joint distribution (Rizopoulos et al., 2017; Wei et al., 2018). However, doing so usually requires strong simplifying assumptions about the covariate trajectories, which can be particularly problematic with high-dimensional covariate data.

In recent years, traditional survival analysis modeling approaches have been augmented with deep learning techniques to produce individualized predictions over time. For example, Lee et al. (2020), Ren et al. (2019), Deasy et al. (2020), and Singh et al. (2020) all use the outputs of a recurrent neural network (RNN) to make personalized risk predictions. Lee et al. (2020) discretize time, and predict the probabilities in each time window. Ren et al. (2019) predict the hazard function (i.e. density of having the event at time t given that the event hasn't occurred yet at time t) at each measurement time, and connect the hazards together using the probability chain rule to predict the survival function. Singh et al. (2020) use a proportional hazards assumption (that the hazard per individual can be broken down into a baseline hazard times a covariate contribution) and predict the hazard ratio directly. Deasy et al. (2020) learns an embedding of high-dimensional ICU data, and predicts the probability of event in the next time window. Other deep learning approaches for dynamic survival include using temporal convolutions as in Jarrett et al. (2020), and transformer-based architectures as in Horn et al. (2020).

Despite being dynamic, the majority of these approaches do not make predictions at multiple timesteps per individual during training. Instead they only make one prediction per individual at the last available measurement in the dataset for that individual. In addition, all but Lee et al. (2020), and Singh et al. (2020) treat survival prediction as a binary classification problem and use a cross-entropy loss setup during training which, as shown in Gorgi Zadeh and Schmid (2020), can produce poorly calibrated survival probabilities. An additional issue is that these models also tend to be blackbox and difficult to interpret, which is a significant limitation in the context of potential adoption for use in clinical applications (Miotto et al., 2018; Rudin, 2019).

3. Notation

A dataset $\mathcal{D}$ containing N individuals can be represented as

$$\mathcal{D} = \{(\mathcal{H}_i, \tau_i, c_i), \;\; i \in \{1, \ldots, N\}\} \tag{1}$$

where $\mathcal{H}_i$ represents the full history of covariate measurements for individual i, τ_i represents the censored event time, and c_i is the censoring indicator which takes a value of 0 if τ_i is the true time until event, and takes a value of 1 if τ_i is right-censored. In more detail, $\mathcal{H}_i$ consists of a collection of measurement times, measurement values, and missing indicators. Missing indicators are needed since for many applications not all covariates will be available for each individual i at every measurement time. The history for individual i can be represented as

$$\mathcal{H}_i = \{(\mathbf{x_{ij}}, \; \mathbf{m_{ij}}, \; t_{ij}), \;\; j \in \{1, \ldots, l_i\}\} \tag{2}$$

Letting M represent the total number of covariate measurements, $\mathbf{x_{ij}}$ is an $M \times 1$ vector representing the values for each measurement at time t_{ij}, $\mathbf{m_{ij}}$ is a $M \times 1$ vector with a zero if a value is missing otherwise taking value 1, and l_i is the number of total measurement times for individual i. We assume that there is a synchronizing event (such as diagnosis with COVID-19) across individuals, i.e., that the first time point for each individual i is defined as $t_{i1} = 1$. Note that the times, t_{ij}, are not restricted to discretized values in principle, i.e., they can occur continuously in time. In this paper, however, for convenience we discretize time (into days) to reduce the amount of missingness. In training and evaluating models we only include the history for an individual up to and including the measurement before the true event time, since for unseen new data we are interested only in making predictions for individuals for whom the event has yet to occur.

4. Model Description

For convenience define $\mathcal{H}_{i,j}$ to be a partial history including measurements up until time t_{ij}:

$$\mathcal{H}_{i,j} = \{(\mathbf{x_{ik}}, \; \mathbf{m_{ik}}, \; t_{ik}), \;\; k \in \{1, \ldots, j\}\} \tag{3}$$

We make predictions at each step of an individual i's history using $H_{i,j}$, the history up until time t_{ij} allowing the model to update it's predictions as new covariate measurements arrive. The predictive densities at each step j in an individual's history must be conditioned on the event of interest occurring after the measurement time t_{ij} since the fact that a measurement was taken implies survival until that time. Therefore, the model output for individual i at each time step j is the predictive density $f^{(i)}(t|\mathcal{H}_{i,j}, T > t_{ij})$ over the event time. Let θ_g be the parameters of a parametric global density (e.g., Rayleigh, Weibull) on time-to-event. Then, conditioned on θ_g, at each time-step in an individual's history an individualized time-shift Δ_{ij}, is used to define an individualized predictive density. The range of Δ_{ij} is restricted to be greater than $-t_{ij}$ to avoid shifting to a negative time. The global density, $f^{(g)}(t|\theta_g)$, and the individualized density, $f^{(i)}(t|\mathcal{H}_{i,j}, T > t_{ij})$, are then connected as follows:

$$\begin{aligned} f^{(i)}(t|\mathcal{H}_{i,j}, T > t_{ij}) \;\; &= f^{(g)}(t + \Delta_{ij}|\theta_g, T > t_{ij}) \\ &= \tfrac{f^{(g)}(t+\Delta_{ij}|\theta_g)}{S^{(g)}(t_{ij}+\Delta_{ij}|\theta_g)} \quad \text{for} \quad t > t_{ij}, \end{aligned} \tag{4}$$

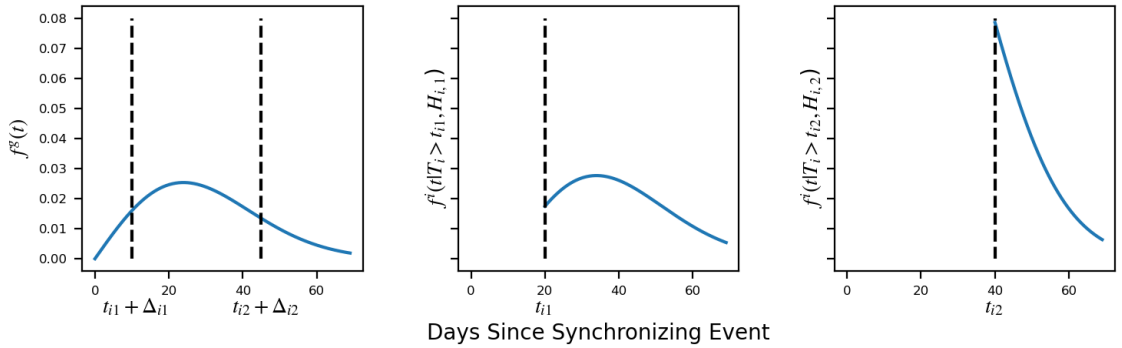

Figure 1: The connection between the global density, shown on the left, and the individual densities at two different measurement times. For each time, the model generates a shift Δ_{ij} which locates the individual density along the timeline of the global density. The individualized density is then determined by truncating the global density at the shifted time $t_{ij} + \Delta_{ij}$ and renormalizing as shown in the middle and right plots. For the first measurement time $t_{i1} = 20$ the model predicts a negative shift, indicating that the individual is doing better than average at the first measurement time. For the second time $t_{i2} = 40$ the model predicts a positive shift indicating that the individual is doing worse than the average individual at the second measurement time.

where $S^{(g)}(t)$ is the survival function for the global model (one minus the cdf of the global density). The first equality represents shifting the time for individual i to their effective time/age, given their covariate history. The second equality involves truncating at time t_{ij} and renormalizing to account for $t > t_{ij}$ as shown in Figure 1.

The Δ_{ij}'s are defined as a parametric function of $\mathcal{H}_{i,j}$. This function in general can be split into two (deterministic) pieces. The first function, z, maps $\mathcal{H}_{i,j}$ to a state $h_{i,j}$ in an autoregressive fashion (see below). The second function, g, then maps the state $h_{i,j}$ to Δ_{ij}. This gives:

$$h_{i,j} = z\big(h_{i,j-1}, (\mathbf{x_{ij}}, \mathbf{m_{ij}}, \Delta t_{ij}); \phi\big) \quad (5)$$

$$\Delta_{ij} = g\big(h_{i,j}; \alpha\big) \qquad j \in [1, l_i] \quad (6)$$

where $\Delta t_{ij} = t_{ij} - t_{i(j-1)}$ and ϕ and α are the parameters of z and g respectively. The function z can be thought of as a transition function taking the previous state, time elapsed since that state and the current measurements to evolve the previous state to the current one. In this work, we consider two options for z and g. In the first z is parametrized by an RNN, and g by a feedforward network, In the second option we let z be the identity over the covariates, and g simply be a linear layer. This amounts to ignoring the full history and only using the measurements at the current time to make predictions. In principle any parametric functions (with g having an appropriate range for Δ_{ij}) could be used.

5. Log-Likelihood

As described above the model makes prediction at each step in the covariate history for an individual. The likelihood for a single individual i therefore takes the form:

$$L((\tau_i, c_i)|\theta_{j=1:l_i}) = \prod_{j=1}^{l_i} f^{(g)}(\tau_i + \Delta_{ij}|\theta_g, \phi, \alpha, T > t_{ij})^{1-c_i} \times S^{(g)}(\tau_i + \Delta_{ij}|\theta_g, \phi, \alpha, T > t_{ij})^{c_i} \tag{7}$$

This likelihood can be interpreted as adding l_i independent psuedo-individuals, one for each time step in H_i. However, if the sequence lengths for some individuals are dramatically larger than others then the data for those individuals will bias the loss since they will have a larger contribution. To address this we average the log-likelihood over the l_i terms per individual, yielding smoother training and gradients when parameters are being learned by gradient descent. An alternative option would be to sample a single time step per individual for each step of training and only use that term to compute the gradient (as with SGD), which would save computation time at the cost of noisier gradients.

To train the model we first fit the parametric global density with parameter vector θ_g, $f^{(g)}(t|\theta_g)$. Then we fix θ_g and learn at each step in an individual i's history the time shift Δ_{ij}. For the results in this paper, we used a global Rayleigh distribution, although any parametric model could be used. In principle θ_g could be jointly learned with the Δ_{ij} parameters. However, we found this makes optimization more difficult, likely due to the increase in local optima introduced, and we conjecture that learning the θ_g parameters jointly with the Δ_{ij}'s causes non-identifiability of the parameters.

Interpretation: Joint learning of the θ_g and Δ_{ij} parameters also removes the possibility of interpreting the global density as representing the average risk at time t for the average person. Such an interpretation is desirable as it allows interpreting the learned Δ_{ij} parameters as 'locating' an individual along the global density. An individual i with predicted shift Δ_{ij} at time t_{ij} would have the same risk as the average individual at time $t_{ij} + \Delta_{ij}$. For example an individual at time 0 with learned shift of 5 days at time 0 would have similar risk to the average person 5 days after the synchronizing event.

6. COVID-19 Severe Outcome DataSet

We evaluate our approach on a COVID-19 dataset consisting of 6,999 individuals diagnosed with COVID-19 at New York University Langone Hospital (NYULH) during the period March to July 2020. We synchronize the covariate trajectories for each individual to time of COVID diagnosis, and define the event of interest as *time until severe outcome after diagnosis*, where severe outcomes are the first occurrence of any of five severe health events: ICU admission, stroke, dialysis, death, and ventilation. In total 882 (12%) of individuals have a severe health outcome with the rest being censored at the end of followup. The dataset includes a variety of dynamic categorical variables such as hospitalization status and whether or not a patient received a particular medication on a particular day. The data also includes numeric covariates such as results of lab tests and vital measurements, as well as static demographic information about each patient. In total we used 345 dynamic

covariates and 11 static covariates. We discretize the measurement times to days, although we allow arbitrary gaps in units of days between measurements, and represent the time to event itself down to the precision of minutes.

We partitioned the data at the patient level into development, validation, and testing subsets with 70% used for development and validation (with 42% and 28% in each respectively corresponding to a 60/40 split), and the remaining 30% for testing. The development and validation sets were used to perform a grid-based hyperparameter search over the learning rate (for all models), regularization, and the hidden dimension of the recurrent neural network (for the RNN-Δ model, and Dynamic DeepHit). We then fit a model to the full training dataset, using the hyperparameter settings that performed best on the validation data, making use of the event times, censoring indicators, and covariate trajectories in the training data. At test time we use only the covariate histories per individual to make predictions about the event times in the test data, mimicking the real-world use case of making predictions on new previously unseen patients.

7. Evaluation

For evaluating dynamic model performance we introduce a modified dynamic form of the C-index, which we will call the *at-risk dynamic C-index* to differentiate it from the standard dynamic C-index (Harrell et al., 1982; Antolini et al., 2005) Unlike the standard dynamic C-index used in Lee et al. (2020), the at-risk dynamic C-index is prognostic, i.e., when making predictions at time t, the model's rankings are evaluated going forwards in time from t, rather than evaluating model performance before t. Furthermore, the model's rankings are evaluated only for individual's who are still at risk for the event to occur after time t. This intuitively represents evaluating how well the model's rankings made at t for at-risk individuals will hold up in the future, which we believe is a more relevant metric for clinicians. We follow the derivation in van Houwelingen and Putter (2011) which derives the C-index as a weighted average of incident dynamic AUC at each event time, and modify it for dynamic predictions by replacing the time varying covariates with the time varying risk predictions from the model. More precisely, we define a set of concordant pairs P_c, and a set of valid pairs P_v as follows:

$$\begin{aligned} P_v &= \Big\{(i,k) \;\; s.t. \;\; \tau_i \leq \tau_k, \;\; t \leq \tau_i, \;\; c_i = 0\Big\} \\ P_c &= \Big\{(i,k) \;\; s.t. \;\; (i,k) \in P_v, \;\; R_i > R_k\Big\} \end{aligned} \tag{8}$$

Intuitively the valid pairs are all pairs for which we know the correct ordering, and the concordant pairs are those for which the model gets that ordering correct. The at-risk dynamic C-index is then just the ratio of the number of concordant pairs, to the valid pairs, estimating the probability of a random valid pair being ordered correctly. Modifying the derivation in van Houwelingen and Putter (2011) for time-varying risks, R_i and R_k are computed for any pair (i,k) as the total probability that the event occurs from time t to time τ_i using the most recent predicted densities at time t for each individual (appropriately re-normalized for survival until time t). Notice that the at-risk dynamic C-index does not require a prediction window, and instead represents the model's averaged performance over all future times for the current set of at-risk individuals.

For completeness we also evaluate our model with the standard version of the dynamic C-index as described in Lee et al. (2020). Following the same notation used before, we can define valid and concordant pairs for the standard dynamic C-index as follows:

$$\begin{aligned} P_v &= \Big\{(i,k) \;\; s.t. \;\; \tau_i \leq \tau_k, \;\; \tau_i \leq t + \Delta t, \;\; c_i = 0\Big\} \\ P_c &= \Big\{(i,k) \;\; s.t. \;\; (i,k) \in P_v, \;\; F_i > F_k\Big\} \end{aligned} \tag{9}$$

Here the risks from the previous equation are replaced with the cumulative densities, F_i and F_k, evaluated at time $t + \Delta t$ conditioned on survival until time t. Notice that the standard dynamic C-index requires a prediction time t and window Δt. This version of the C-index can easily be falsely inflated when time is included as a predictor. This occurs because individuals with the event before the prediction time t will likely have their last measurement time prior to the event ,t_{il_i}, (which a model could use to make it's predictions) be close to the true event time. Therefore ranking by $\frac{1}{t_{il_i}}$ (which for time-discretized data is the frequency of encounters) will be better than a random ranking for these individuals, and a model could inflate the index by exploiting this, despite it not representing a clinically meaningful predictor. Our proposed at-risk C-index does not have this potential issue since it doesn't include individual's before the prediction time.

We compare the performance of our time-shift Δ models with several baselines. We evaluate our time-shift approach using both a linear model (Linear-Δ) which only makes use of the current covariate measurements without using the full history, and an RNN-based model (RNN-Δ) which uses the full history. We compare these models to the Dynamic DeepHit model of Lee et al. (2020), in order to include comparison to a competitive deep recurrent model. In addition we also evaluate a standard Landmark-Cox model in order to include a traditional statistical approach. Finally, we include for comparison a model free baseline that ranks patients by the the inverse of the total number of days (1/N-Days) they've been in hospital at the prediction time t—this is included to differentiate between the two versions of the C-index that we described above, i.e., we expect our proposed At-Risk Dynamic C-Index to be robust to timing effects and therefore for this baseline to produce C-Index values close to 0.5 that are no better than random ordering. Conversely on the standard C-Index, this baseline may produce spurious performance that is above random chance.

8. Results

To obtain the results in Table 1 we perform validation experiments training each model on the development data for different hyperparameter settings, and evaluating performance on the validation data. After selecting hyperparameters, we then retrain the model on the development and validation data combined (a 70% random subset of patients), and evaluate on the test data to generate C-Index values for out-of-sample data (the other 30% of patients). For the dynamic C-index results in Table 1 at each prediction time (in days, 0, 3, 4, 6, and 11) we average the C-index across four different time windows of 5, 10, 15, and 20 days. The at-risk dynamic C-index version does not require a time window.

Table 1(a) shows results for all of the methods we evaluated for the new **at-risk dynamic C-index** that we introduced in equation 8, while Table 1(b) shows results for the

(a)

At-Risk Dynamic C-Index

Day	0	3	4	7	11	Avg (std)
1/(N-Days)	0.48	0.49	0.49	0.51	0.52	0.50 (0.01)
Landmark-Cox	0.79	0.78	0.79	0.77	0.47	0.72 (0.13)
Linear-Δ (proposed)	0.77	0.85	0.85	0.86	0.78	**0.82** (0.04)
RNN-Δ (proposed)	0.78	0.83	0.83	0.82	0.71	0.78 (0.05)
Dyn-DeepHit	0.69	0.79	0.80	0.79	0.72	0.76 (0.05)

(b)

Dynamic C-Index

Day	0	3	4	7	11	Avg (std)
1/(N-Days)	0.50	0.59	0.60	0.60	0.60	0.58 (0.04)
Landmark-Cox	0.79	0.79	0.80	0.74	0.61	0.75 (0.07)
Linear-Δ (proposed)	0.76	0.80	0.82	0.85	0.86	0.82 (0.04)
RNN-Δ (proposed)	0.70	0.76	0.78	0.83	0.84	0.78 (0.05)
Dyn-DeepHit	0.82	0.87	0.87	0.89	0.88	**0.87** (0.01)

Table 1: Results with the two dynamic C-indices on the COVID-19 data, (a) our proposed At-Risk Dynamic C-Index, and (b) the standard Dynamic C-index.

standard dynamic C-index as defined in equation 9. Focusing initially on Table 1(b) we see that Dynamic-DeepHit apparently has better performance (0.87) on average than any of the other methods. However, we also see in Table 1(b) that ranking by the inverse of the number of days since COVID-19 diagnosis (1/N-Days) leads to performance well above chance. This clearly demonstrates the potential for inflated performance estimates when the standard dynamic C-index is used for evaluation.

However, when evaluating via the new at-risk dynamic C-index in Table 1(a), the performance of ranking by the inverse of the number of days is near random. This is due to the exclusion of individuals with event times before time t in the set of valid pairs, showing that ranking by the inverse of number of days is not useful for making predictions for at-risk patients. Thus, comparing the C-index values for 1/N-Days ranking method between Table 1(a) and (b), we see (as expected) that the at-risk dynamic C-index in (a) does not suffer from the optimism bias inherent to the standard dynamic C-index in (b). Under the at-risk dynamic C-index , in Table (a), the performance of Dynamic-Deephit is now considerably lower, implying that some of its performance on the standard dynamic C-index may have been due to exploiting the correlation for individuals with event times before the prediction time t. In contrast our proposed RNN-Δ and linear-Δ models have more stable performance across the two indices. In addition our Linear-Δ model has the highest performance over all models (0.82) in Table 1(a). We note that neither of the two RNN models (RNN-Δ or Dyn-DeepHit) were able to use the longer-term history of patient data to achieve the performance of the linear-Δ model. In addition, the two proposed Δ models and the Dyn-DeepHit model all outperformed the traditional statistical Landmark-Cox model. We conjecture that Dynamic DeepHit performed better under the standard dynamic C-index

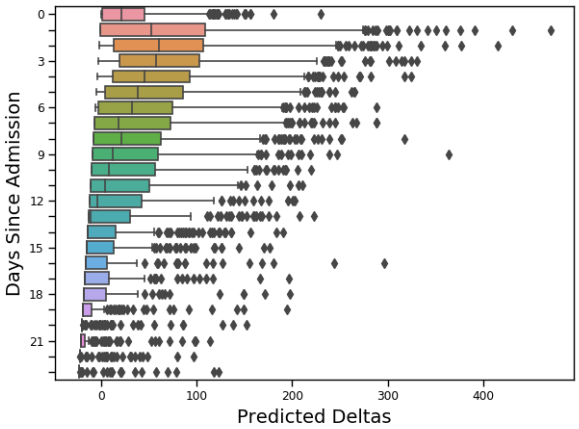

Figure 2: Boxplots of the predicted Δ's at different prediction times. For each prediction time we collect all individuals who are still in the hospital, and plot a boxplot for the Δ's output by the model for that set of individuals. The minimum possible Δ at time t is negative t. As time passes the Δ's tend to become smaller.

due to the inclusion of a ranking regularization term which closely mimics the dynamic C-index in its loss function. Nonetheless, Dynamic DeepHit performance, when measured by the at-risk C-index, while better than the landmark-Cox model, was poorer then both versions of our approach.

Figure 2 shows boxplots of the Δ's learned by the model for different prediction times. The model-predicted Δ's tend to decrease the longer a patient has been in the hospital, agreeing with the analysis in Rees et al. (2020) that length of stay is inversely correlated with serious (or adverse) outcomes. We also show the predicted hazards for two individuals in Figure 3. Here the plotted hazard represents the predicted hazard function, which is a continuous function of time, evaluated at the start of each day. The individual shown on the left plot of 3 starts off with slightly higher than average hazard after covid diagnosis, but after developing a fever, and high systolic blood pressure they then experience a severe outcome at day 8. Conversely, the individual shown on the right starts off with a better prognosis, their hazard rises slightly with their blood pressure at day 4, and then after the hazard decreases to zero over time they leave the hospital with no severe outcome.

9. Conclusion

In this paper we introduced a novel dynamic survival model based on individualized time-shifts relative to a global denisty. We also proposed a new at-risk dynamic C-index that evaluates a model's predictions only for individuals still at risk. Our model shows promising performance on a real-world COVID-19 hospitalization dataset. However, the current results described in this paper are only for a single dataset and further evaluation is needed on additional datasets (including public data such as MIMIC-Three). Another important question is why the additional modelling power of the RNN-based models do not yield im-

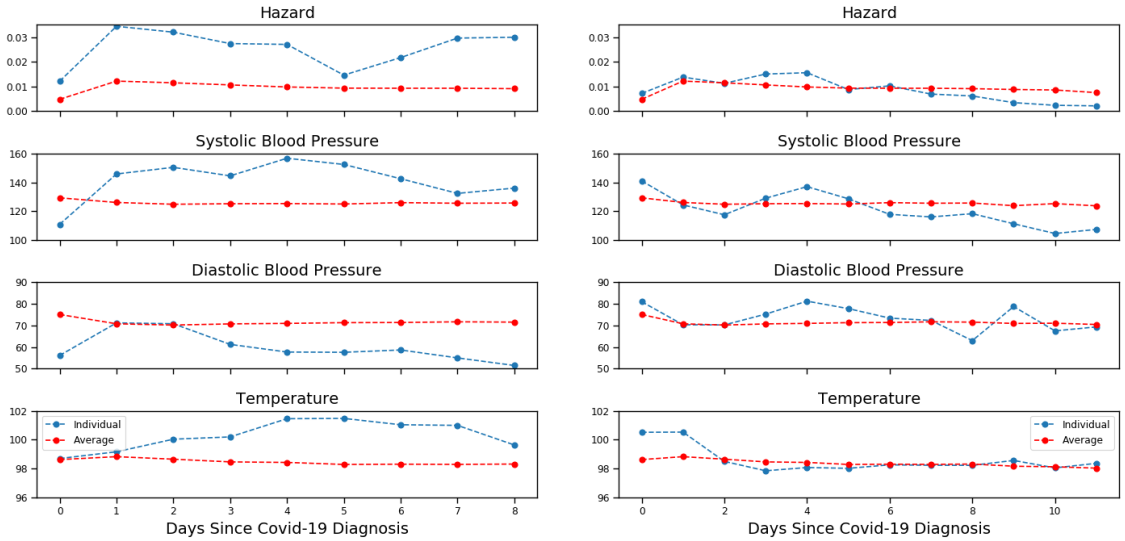

Figure 3: Examples of two different hospitalized individuals, illustrating the model's predicted hazard function evaluated at the start of each day, along with systolic blood pressure, diastolic blood pressure, and temperature. Average values, shown in red, are for comparison to the individual values shown in blue.

proved performance over the linear model for the at-risk dynamic C-index on the COVID-19 dataset. The choice of parametric form for the global distribution has also not yet been fully investigated, and the currently-used Rayleigh distribution may be a limiting factor since it limits the hazard function to an increasing linear function of time since synchronizing event.

Acknowledgments

Acknowledgments The work of PP and PS was supported in part by the HPI Research Center in Machine Learning and Data Science at UC Irvine.

References

Odd Aalen, Ørnulf Borgan, and Hakon Gjessing. *Survival and Event History Analysis: A Process Point of View*. 01 2008. ISBN 0-387-20287-0. doi: 10.1007/978-0-387-68560-1.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

Jacob Deasy, Pietro Liò, and Ari Ercole. Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or curation. *Scientific Reports*, 10(1):1–11, 2020.

S. Gorgi Zadeh and M. Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. *ICML*, 2020.

D. Jarrett, J. Yoon, and M. van der Schaar. Dynamic prediction in clinical survival analysis using temporal convolutional networks. *IEEE Journal of Biomedical and Health Informatics*, 24(2):424–436, 2020. doi: 10.1109/JBHI.2019.2929264.

C. Lee, J. Yoon, and M. van der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2020. doi: 10.1109/TBME.2019.2909027.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, 2018.

Mohammadreza Nemati, Jamal Ansary, and Nazafarin Nemati. Machine-learning approaches in covid-19 survival analysis and discharge-time likelihood prediction using clinical data. *Patterns*, 1(5):100074, 2020. doi: 10.1016/j.patter.2020.100074.

Layla Parast, Lu Tian, and Tianxi Cai. Landmark estimation of survival and treatment effect in a randomized clinical trial. *Journal of the American Statistical Association*, 109 (505):384–394, Jan 2014. doi: 10.1080/01621459.2013.842488.

Eleanor M. Rees, Emily S. Nightingale, Yalda Jafari, Naomi R. Waterlow, Samuel Clifford, Carl A. B. Pearson, CMMID Working Group, Thibaut Jombart, Simon R. Procter, and Gwenan M. Knight. Covid-19 length of hospital stay: a systematic review and data synthesis. *BMC Medicine*, 18(1):270, Sep 2020. doi: 10.1186/s12916-020-01726-3.

Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. Deep recurrent survival analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4798–4805, Jul. 2019. doi: 10.1609/aaai.v33i01.33014798.

D. Rizopoulos, G. Molenberghs, and E. M. E. H. Lesaffre. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biom J*, 59(6):1261–1276, Nov 2017.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Harvineet Singh, Moumita Sinha, Atanu R Sinha, Sahil Garg, and Neha Banerjee. An RNN-survival model to decide email send times. *arXiv preprint arXiv:2004.09900*, 2020.

Annette Spooner, Emily Chen, Arcot Sowmya, Perminder Sachdev, Nicole A. Kochan, Julian Trollor, and Henry Brodaty. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, 10 (1):20410, Nov 2020. doi: 10.1038/s41598-020-77220-w.

Hans C. Van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007. doi: 10.1111/j.1467-9469.2006.00529.x.

Johannes (Hans) van Houwelingen and Hein Putter. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, 2011.

Ping Wang, Yan Li, and Chandan K. Reddy. Machine learning for survival analysis: A survey. *ACM Comput. Surv.*, 51(6), February 2019. doi: 10.1145/3214306.

Melissa Y Wei, Mohammed U Kabeto, Andrzej T Galecki, and Kenneth M Langa. Physical functioning decline and mortality in older adults with multimorbidity: joint modeling of longitudinal and survival data. *The Journals of Gerontology: Series A*, 74(2):226–232, 02 2018. doi: 10.1093/gerona/gly038.