Dynamic Survival Analysis for EHR Data with Personalized Parametric Distributions

Preston Putzel[1], **Hyungrok Do**[2], **Alex Boyd**[3], **Hua Zhong**[2], **and Padhraic Smyth**[1]

[1]Department of Computer Science, University of California, Irvine, CA, USA
[2]Department of Population Health, NYU Grossman School of Medicine, New York, NY, USA
[3]Department of Statistics, University of California, Irvine, CA, USA

Abstract

The widespread availability of high-dimensional electronic healthcare record (EHR) datasets has led to significant interest in using such data to derive clinical insights and make risk predictions. More specifically, techniques from machine learning are being increasingly applied to the problem of dynamic survival analysis, where updated time-to-event risk predictions are learned as a function of the full covariate trajectory from EHR datasets. EHR data presents unique challenges in the context of dynamic survival analysis, involving a variety of decisions about data representation, modeling, interpretability, and clinically meaningful evaluation. In this paper we propose a new approach to dynamic survival analysis which addresses some of these challenges. Our modeling approach is based on learning a global parametric distribution to represent population characteristics and then dynamically locating individuals on the time-axis of this distribution conditioned on their histories. For evaluation we also propose a new version of the dynamic C-Index for clinically meaningful evaluation of dynamic survival models. To validate our approach we conduct dynamic risk prediction on three real-world datasets, involving COVID-19 severe outcomes, cardiovascular disease (CVD) onset, and primary biliary cirrhosis (PBC) time-to-transplant. We find that our proposed modeling approach is competitive with other well-known statistical and machine learning approaches for dynamic risk prediction, while offering potential advantages in terms of interepretability of predictions at the individual level.

1. Introduction

Survival analysis focuses on the analysis and modeling of time-to-event data. Traditional approaches to survival modeling in statistics, such as proportional hazard models (Aalen et al., 2008), typically construct global time-to-event distributions or make simplifying assumptions about the effect of an individual's covariates on their risk of an event. The past few years have seen the development of a number of different machine learning methods applied to survival modeling (Ishwaran et al., 2008; Wang et al., 2019; Spooner et al., 2020; Nemati et al., 2020). The flexibility of these methods makes it possible to relax the parametric assumptions used in previous approaches for survival analysis.

Survival analysis models can be static or dynamic. For static models the predicted risk of an event occurring is modeled purely as a function of baseline covariates. For dynamic models the predicted risk can change as a function of time-varying covariates and predictions are made using the full history of information, from the time of the initial patient measurement to the most recent measurement available. Thus, dynamic survival models can, in principle, better represent changing risks over time and determine which patients are in the greatest need of treatment at any time, instead of only at the time of baseline measurements (Van Houwelingen and Putter, 2011). For example, in clinical applications such as COVID-19 treatment, the current risk of an adverse event can be valuable in determining how best to treat a patient, such as deciding whether or not to put a patient on ventilation. For cardiovascular disease (CVD) prediction, it has also been shown that the visit-to-visit variability of risk factors, including body weight, triglycerides, and HDL and LDL cholesterol, predicts CVD independent of their mean values, a factor which would be missed in static modeling (Bangalore et al., 2017).

In this paper we introduce a novel type of dynamic survival model which learns a global density model over a continuous time-to-event and then as measurements are dynamically updated locates an individual on the time-axis relative to this global density. This shifting procedure acts as a natural regularization for the model since it reduces the family of possible individualized distributions to truncated (and renormalized) versions of the global density. This connection between global and individualized densities also allows for the interpretation of each individualized density as the average density of individuals who have survived to the same shifted time along the global time-axis. This shifted time then represents an 'effective time' for the individual, i.e. the time for which the density of the individual would look the same as that of the average individual who had survived until then.

We also introduce a novel evaluation metric in this paper: a new form of the dynamic C-Index that more accurately reflects a model's expected prognostic performance in a true clinical setting compared to the more commonly used standard dynamic C-Index. We apply and evaluate our methods using EHR data for three real-world cohorts of patients. The first cohort consists of patients hospitalized with COVID-19 where the problem is to predict individualized distributions over time-to-severe outcome after COVID-19 diagnosis. The second cohort consists of patients diagnosed with diabetes mellitus where the goal is to predict individualized distributions over time-to-onset of CVD. For our third dataset we use the PBC2 dataset which contains data collected over the course of a ten year period for a randomized controlled trial of a treatment for primary biliary cirrhosis of the liver. For this dataset the goal is to predict individualized distributions over time-to-transplant.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our experience in building different predictive risk models, ranging from traditional statistical approaches to recurrent deep network models, has led to a number of generalizable which we discuss below. Several of these insights reinforce well-known points in medical informatics. The primary insights are:

- Evaluation metrics for machine learning methods in healthcare should mirror, to the extent possible, how a model would be used in practice if deployed. While it is natural for researchers to tend to use metrics adopted in earlier literature (e.g, to allow for standardized comparison with the same metrics across different sets of results), it is also worth taking a critical look at whether evaluation metrics used in the past are both realistic and clinically relevant. In addition, as more complex deep learning approaches are applied to dynamic survival analysis, care must be taken to ensure that standard evaluation metrics are not arbitrarily inflated by exploiting clinically irrelevant information in ways that more standard statistical methods cannot.

- It is important when evaluating the efficacy of complex machine learning models (such as recurrent neural networks) to compare them to simpler traditional models (such as linear models) to determine if the additional modeling complexity is worthwhile. The results for the three real-world datasets in our study illustrated that off-the-shelf deep learning methods are not necessarily going to produce better predictions than simpler linear modeling approaches.

- While electronic health record (EHR) data is a natural target for dynamic survival analysis, such data also presents multiple challenges in terms of interpretation and modeling (Yadav et al., 2018; Ghassemi et al., 2020). For example, covariate measurements at each patient encounter are often both irregularly sampled in time as well as highly sparse, with only a small subset of possible measurements being taken. The missing measurements are also often not missing at random where for example, healthier patients may be given less extensive lab tests and monitoring. Complications in data interpretation also arise from billing and reimbursement policies, e.g., diagnosis codes need not necessarily reflect the underlying health state of a patient

- While both statistical and machine learning predictive models based on EHR data show promise for dynamic risk prediction in all three of our datasets, the models' performances (particularly for the non-hospitalized multi-year CVD data) are not necessarily at the level where they provide enough of an improvement to augment or displace current practices in clinical risk prediction.

2. Methods

2.1. Notation

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

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

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 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

$$\mathcal{H}_i = \{(\mathbf{x}_{ij},\ \mathbf{m}_{ij},\ t_{ij}),\ \ j \in \{1, \dots, l_i\}\} \tag{2}$$

Letting M represent the total number of different covariates, $\mathbf{x}_{ij}$ 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 where an entry is 0 if the value is missing and 1 otherwise, 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

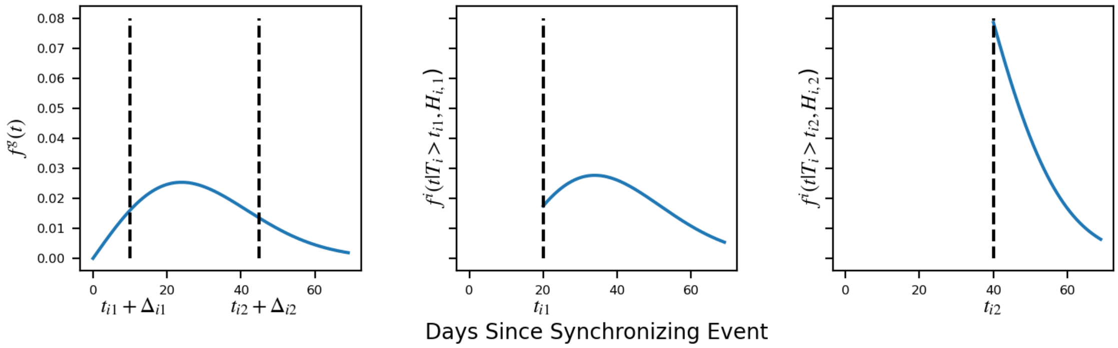

Figure 1: The connection between the global density (left plot) and individual densities at two different measurement times (center and right plots). For each time, the model generates a shift Δ_{ij} which locates the individual density along the timeline of the global density. The individualized density 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.

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 they can occur continuously in time. In this paper, however, for convenience we discretize time (into days, or months depending on the dataset) 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.

2.2. Model Description

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

$$\mathcal{H}_{ij} = \{(\mathbf{x}_{ik},\ \mathbf{m}_{ik},\ t_{ik}),\ \ k \in \{1, \ldots, j\}\} \tag{3}$$

We make predictions at each step of an individual i's history using $\mathcal{H}_{ij}$, the history up until time t_{ij}, allowing the model to update its 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 $f^{(i)}(t|\mathcal{H}_{ij}, T > t_{ij})$ 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, $f^{(i)}(t|\mathcal{H}_{ij}, T > t_{ij})$, are then connected as follows:

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

where $S^{(g)}(t)$ is the survival function for the global model (one minus the cumulative distribution function 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 $\mathcal{H}_{ij}$. This function in general can be split into two (deterministic) pieces. The first function, z, maps $\mathcal{H}_{ij}$ to a state h_{ij} in an autoregressive fashion (see below). The second function, g, then maps the state h_{ij} to Δ_{ij}. This gives:

$$h_{ij} = z\big(h_{i(j-1)}, (\mathbf{x}_{ij}, \mathbf{m}_{ij}, \Delta t_{ij}); \phi\big) \tag{5}$$

$$\Delta_{ij} = g\big(h_{ij}; \alpha\big) \qquad j \in \{1, \ldots, l_i\} \tag{6}$$

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. One option is to parameterize z by an recurrent neural network (RNN) and g by a feedforward network, which we will refer to as the RNN-Δ model. The second option is to let z be the identity over the covariates, and g simply be a linear layer, which we will refer to as the Linear-Δ model. 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.

This parameterization is still flexible enough for individualization since the model can produce a different Δ_{ij} for each individual i at step j and, thus, different risk predictions across individuals. At the same time the constraint imposed by using a global distribution results in the model having a stronger inductive bias than more flexible approaches such as deep learning approaches, restricting the individualized densities to a family of parametric truncated distributions. This inductive bias effectively acts as a form of built-in regularization of the model.

2.3. Loss Function and Model Training

As described above the model makes a prediction at each step in the covariate history for an individual. The likelihood, which is a function over the unknown parameters θ, ϕ, and α, therefore takes the following form for a single individual i:

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

If individual i is uncensored, each term in this loss is the predicted probability density of the time-to-event at time t_{ij}, and if individual i is censored, it is the survival probability. This represents that the exact time-to-event is known for uncensored individuals, while for censored individuals it is only known that the individual has survived past τ_i. 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. A similar loss was used in prior work in the context of time-series clustering (temporal disease phenotyping) in Lee et al. (2021). 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, potentially due to the introduction of local optima. We also 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 would also remove the possibility of interpreting the global density as representing the average risk at time t for the average person. In contrast, using a fixed global density model allows interpretation of the learned Δ_{ij} parameters as 'locating' an individual along that 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.

Code Availability: Python code implementing our model can be found at the github link: https://github.com/pjputzel/dyn_surv_global_with_shifts

3. 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 to understand and interpret the connection between predicted risks at different landmark times. In contrast, the predictions of the Linear-Δ version of our model at different time points are all related to a single shared global density, and predictions are issued at every step along the covariate history of an individual rather than only at landmark times. The RNN-Δ model shares these differences, and differs even further with landmarking approaches by making full use of the entire covariate history to make it's predictions.

In joint modeling approaches to dynamic survival modeling, the probability distribution of the covariate history $\mathcal{H}$ and time-to-event T are modelled with a joint distribution (Rizopoulos et al., 2017; Wei et al., 2018). However, doing so often requires strong parametric assumptions about the shape of the covariate trajectories, which can be particularly problematic with high-dimensional covariate data.

In recent years, traditional survival analysis modeling approaches have also 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 an 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., the density of the event occurring 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. Making the proportional hazards assumption (that the hazard per individual can be broken down into a baseline hazard times a covariate contribution), Singh et al. (2020) predict the hazard ratio directly. In Deasy et al. (2020) an embedding of high-dimensional ICU data is used to predict the probability of event in the next time window. The approach in Lee et al. (2021) focuses on clustering time series based on the distribution of health outcomes, and could potentially be modified in order to make dynamic survival predictions. 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 difficult to interpret, which is a significant limitation when these models are being considered for use in clinical applications (Miotto et al., 2018; Rudin, 2019).

4. Datasets

To evaluate our proposed approach we use two EHR datasets. The datasets include a variety of dynamic categorical variables such as hospitalization status and whether or not a patient received a particular type of medication on a certain day. The datasets also include numeric covariates such as results of lab tests and vital measurements, as well as static demographic information about each patient. We represent medications by their pharmacy-subclass to reduce the number of possible types of medications. We filter out lab tests based on their total amount of missingness across all encounters—specific lab tests which are missing in more than 75% of encounters are dropped.

We also evaluate our approach on the publicly-available PBC2 clinical trial (non-EHR) dataset which contains the results of various laboratory tests, both discrete and numeric, over a ten year-period of followup after the start of study.

At each timestep in an individual's history we augment their vector of covariate measurements x_{ij} with an additional $M \times 1$ vector of missing indicators $\mathbf{m}_{ij}$ representing which of the covariate value for individual i are missing at measurement time t_{ij} (as also implemented for example in Lee et al. (2020)). The corresponding missing values in the covariate measurements vector $\mathbf{x}_{ij}$ are then replaced with the means across all times and individuals of the missing covariates.

For all datasets, the measurement times are discretized. This is in order to have more dense inputs for all of the models we evaluate, and avoid using mostly empty covariate vectors for many timepoints, which could negatively affect training.

COVID-19 Severe Outcome Dataset The COVID-19 dataset consists of 6,999 individuals diagnosed with COVID-19 at New York University Langone Hospital (NYULH) during March to July 2020. We synchronize the covariate trajectories for each individual to time of COVID-19 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%) individuals have a severe health outcome while the rest were censored at the end of followup. In total we used 345 dynamic covariates (212 lab tests and vitals, and 123 medication types) and 11 static covariates (including age, sex, race, ethnicity, body mass index, and tobacco usage). 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 in units of minutes.

Diabetes Mellitus Cardiovascular Disease Dataset This dataset consists of 16,335 individuals diagnosed with diabetes mellitus at NYULH during January 2010 to December 2019. We synchronize the covariate trajectories for each individual to time of diabetes mellitus diagnosis, and define the event of interest as *time until onset of cardiovascular disease*. To identify CVD onset we use a list of 155 ICD-9 codes corresponding to CVD diagnosis or a health event associated with CVD such as heart attack. In total 29% of individuals experienced the onset of CVD during observation while the rest were censored at the end of followup. We used 185 dynamic covariates (124 lab tests and vitals, and 61 medications types) and 5 static covariates (age, sex, race, ethnicity, and smoking status). As with the COVID-19 data we discretize the measurement times (this time to months instead of days), allow arbitrary gaps between between measurement months, and represent the time-to-event in days.

PBC2 Dataset We also evaluate our approach on a non-EHR publicly-available dataset which was collected during a ten year period from 1974-1984 for a randomized control trial of a treatment for PBC. For this dataset we define the event of interest as *time to liver transplant* with about 45% of individuals having a transplant. The dataset includes the three static covariates of sex, age at start-of-study, and whether or not the patient received placebo or treatment. We use a total of 12 dynamic covariates including 7 labtests such as albumin and serum bilirubin, and 5 categorical diagnostic evaluations such as the presence of an enlarged liver (hepatomegaly). We discretize measurement times to months, while representing the time-to-event itself in days. We used the (lightly) preprocessed version of this dataset contained in the code base for Lee et al. (2020).

5. Evaluation

For evaluating dynamic model performance we introduce a modified dynamic form of the C-Index, which we call the **at-risk dynamic C-Index** to differentiate it from the **standard dynamic C-Index**, which is evaluated over all individuals, at risk or not (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 strictly prognostic, i.e., when making predictions at time t, the model's rankings are only evaluated going forwards in time from t, rather than also evaluating model performance before t. Furthermore, the model's rankings are evaluated only for individuals who are still at risk for the event to occur after time t. Intuitively this evaluates how well the model's rankings made at t for at-risk individuals will hold up in the future, which is a more relevant metric for clinical practice than the approach of the standard C-Index. 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 we 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 valid pairs $\mathcal{P}_v(t)$ and concordant pairs $\mathcal{P}_c(t)$ at time t as follows:

$$\mathcal{P}_v(t) = \Big\{(i,k) : \tau_i \leq \tau_k,\ \tau_i \in [t,\infty),\ c_i = 0\Big\}, \tag{8}$$

$$\mathcal{P}_c(t) = \Big\{(i,k) : (i,k) \in \mathcal{P}_v(t),\ F(\tau_i|\mathcal{H}_i(t)) > F(\tau_i|\mathcal{H}_k(t))\Big\}, \tag{9}$$

where $\mathcal{H}_i(t)$ is the most recent partial history of individual i available at time t. Intuitively the valid pairs represent the 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 the ratio of the number of concordant pairs to the valid pairs, which estimates the probability of a random valid pair being ordered correctly. Modifying the derivation in Van Houwelingen and Putter (2011) for time-varying risks, we compute the 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), for any pair (i,k). This probability is then used to rank individuals as shown in equation 9. Unlike the standard dynamic C-Index, the at-risk dynamic C-Index does not require specification of a prediction window, and instead represents the model's averaged performance over all future times for the current set of at-risk individuals.

Following the same notation used before, we can define valid and concordant pairs at time t with prediction window Δt for the standard dynamic C-Index as follows:

$$\mathcal{P}_v(t,\Delta t) = \Big\{(i,k) : \tau_i \leq \tau_k,\ \ \tau_i \in [0, t+\Delta t],\ \ c_i = 0\Big\}, \tag{10}$$

$$\mathcal{P}_c(t,\Delta t) = \Big\{(i,k) : (i,k) \in \mathcal{P}_v(t,\Delta t),\ \ F(t+\Delta t|\mathcal{H}_i(t)) > F(t+\Delta t|\mathcal{H}_k(t))\Big\}. \tag{11}$$

Here the risks from the previous equations are replaced with the cumulative densities evaluated at time $t+\Delta t$ conditioned on survival until time t. Notice that the standard dynamic C-Index requires specification of a prediction time t and window Δt.

Since the standard dynamic C-Index is calculated for all individuals with $\tau_i \leq t + \Delta t$, it includes evaluation over individuals who have already had the event at the prediction

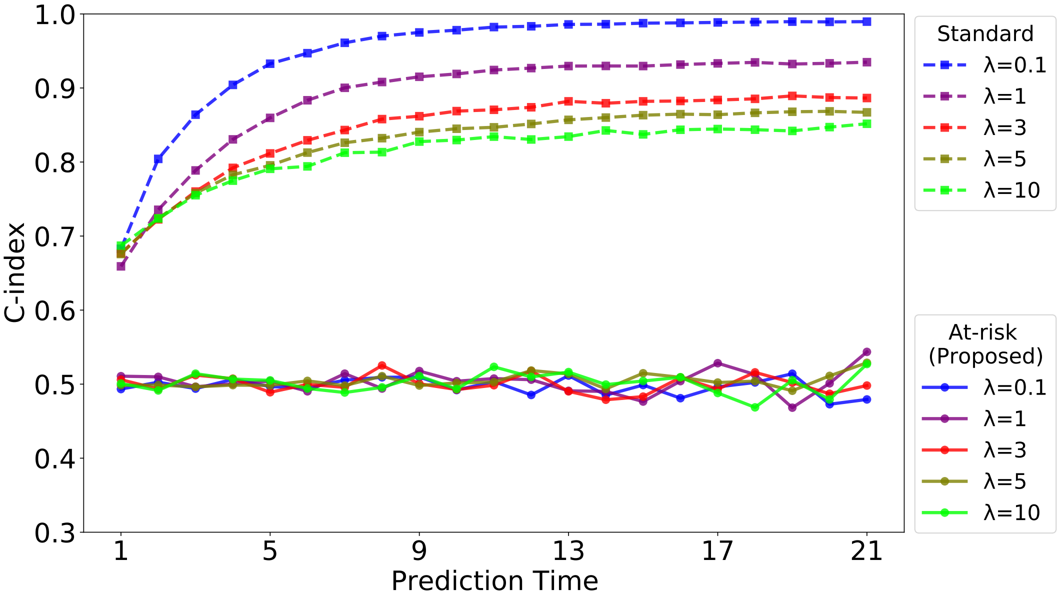

Figure 2: Risk prediction for simulated individuals evaluated using both (a) standard C-Index (dotted lines), and (b) proposed at-risk dynamic C-Index (solid lines), as a function of prediction time. Individuals are ranked using the 1/(N-Days) ranking approach (see text). The evaluation metric (y-axis) should indicate no predictive power (C-Index at 0.5). The results confirm that the standard dynamic C-Index (upper curves) can be inflated by inclusion of individuals who have already had events ($\tau_i < t$). On the other hand, the ranking using 1/(N-Days) (lower curves) is centered around 0.5, i.e., is no better (as it should be) than random guessing for at-risk individuals.

time t, i.e., it includes individuals who are not currently at-risk. This presents an issue for evaluating models that use time-since-synchronization as a predictor. To illustrate this particular issue, consider ranking by 1/(N-Days) (instead of the CDF in equations 9 and 11), where N-Days represents the time since synchronization. Given a valid pair of *not* at-risk individuals (i, j), with $\tau_i < \tau_j < t$, the pair is *not* a concordant pair only if individual j does not have an encounter between t_{il_i} and τ_j. The probability of this occurring is $P(t_{jl_j} < t_{il_i})$ and if we assume for example that each individual's frequency of encounters follow a Poisson(λ) distribution, the probability of occurrence is approximately $1-e^{-(\tau_j-\tau_i)\lambda}$. Thus, as the frequency of encounters increases (larger λ), the concordance probability for not at-risk individuals becomes higher using the 1/(N-Days) ranking. On the other hand, for the set of at-risk individuals with $\tau_i > t$, the time since synchronization is not strongly correlated with τ_i and thus we expect the 1/(N-Days) ranking not to produce any predictive performance above random chance.

To illustrate this issue, we generated simulated data and evaluated the performance when ranking by 1/(N-Days). For the evaluation of the ranking we used both versions of the dynamic C-Index: the standard dynamic C-Index and our proposed at-risk dynamic

C-Index Results from Ranking with 1/(N-Days)

	COVID-19	**DM-CVD**	**PBC2**
At-Risk Dynamic C-Index	0.50	0.49	0.49
Standard Dynamic C-Index	0.58	0.67	0.69

Table 1: Averaged performance of ranking by 1/(N-Days) across different prediction windows on the three real-world datasets used in the paper. For the at-risk version, performance is near random (as it should be), while the standard dynamic C-Index is well above chance for all three, i.e., it significantly inflates performance as it did on the simulated data.

C-Index. The simulated data has Poisson-distributed encounter times, with time-to-events drawn from an exponential distribution, and censoring applied randomly. We show the results of evaluating rankings for different prediction times in Figure 2, using both the standard dynamic C-Index and the at-risk dynamic C-Index. As expected, as the Poisson parameter increases, the probability of a valid pair for not at-risk patients decreases, and the performance of the standard dynamic C-Index drops. In contrast, the performance of the at-risk dynamic C-Index remains near chance on average across all settings of the Poisson parameter. These results are in agreement with our hypothesis that the standard dynamic C-Index can have a significant optimistic bias in estimating true predictive performance due to artificial correlations with the frequency of visits or encounters. In addition, the standard dynamic C-Index will be more inflated for datasets with more frequent encounters.

Based on this result, ranking using 1/(N-Days) will potentially have better than chance performance for the standard dynamic C-Index on real-world datasets (notice the performance is inflated for all λ settings). Any model which makes use of time-since-synchronization as a predictor could therefore have inflated performance on the standard dynamic C-Index by exploiting this. This is despite the fact that ranking by 1/(N-Days) does not represent a clinically meaningful predictor for many real-world applications. Our proposed at-risk C-Index does not have this potential issue since it doesn't include individuals with events before the prediction time. To further confirm this we ranked individuals by 1/(N-Days) for all three of our real clinical datasets, with results in Table 1, using both variants of the C-Index. The results confirm that the standard dynamic C-Index is inflated for our three real world datasets as well, i.e., only the at-risk dynamic C-Index results in performance that is at random chance (0.5).

6. Results on Real Data

6.1. Model Fitting and 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 implemented both of our approaches using PyTorch (Paszke et al., 2019). For training the model in practice we discretize time into suitable units. However, despite this discretization, the model is still capable of issuing risk predictions at any desired time. For example, in the COVID-19 application, despite the measurement time granularity being in days, predicting risks in the next 6-12 hours would allow clinicians to identify higher risk patients to be moved into the ICU quickly.

For each dataset in our experiments we use grid search on validation data to select hyperparameter settings for each model being evaluated. We then retrain each model on the development and validation data combined (a 70% random subset of patients split once more into 60% development and 40% validation), and evaluate on the test data to generate metrics (dynamic C-Index values and Brier scores) for out-of-sample data (the other 30% of patients).

For the Linear-Δ model we included L1 regularization to encourage sparsity and using the validation data searched over a grid of $[0, 0.001, 0.01, 0.1]$ to set the regularization strength. For the RNN-Δ model, we used L2 regularization searching over a grid of $[0, 0.0001, 0.001]$, and a grid of $[64, 128, 256]$ for the hidden dimension. For both Δ models, we additionaly performed a grid search over learning rates of $[0.001, 0.01, 0.1]$.

We also compare the performance of our time-shift Δ models with several baselines. The baselines include the Dynamic DeepHit (Dyn-DeepHit) model of Lee et al. (2020), in order to include comparison to a competitive deep recurrent model. For Dyn-DeepHit we performed our hyper parameter grid search using the validation data over values centered on the default settings. This resulted in a search over $[0.00001, 0.0001, 0.001]$ for the learning rate, and $[50, 100, 200]$ for the RNN hidden dimension. For all other parameters we used default settings.

In addition we also evaluate a standard Landmark-Cox model as traditional statistical baseline, and a Landmarked Random Survival Forest (Landmark-RF) model to compare to a non-neural machine learning approach. For the Landmark-RF model we performed a grid search on the validation data for the number of trees over $[10, 50, 250]$. Both landmarked models were implemented using the open-source python package pysurvival (Fotso et al., 2019). For the diabetes mellitus data we also include the risk score from the Framingham heart study to compare our results with a commonly used clinical risk assessment (D'Agostino et al., 2008).

6.2. Experimental Results

We evaluated our proposed models and baselines on the three clinical datasets described earlier in the paper. Table 2 shows test data results for all of the methods we evaluated for the new **at-risk dynamic C-Index** that we introduced in equation 9. Table 2(a) shows results for the COVID-19 data Table 2(b) shows results for the diabetes mellitus data, and Table 2(c) shows results for the PBC2 data. For completeness in **Appendix A** we also evaluate our model with the standard version of the dynamic C-Index as described in Lee et al. (2020).

Table 3 shows results for all three datasets with the Brier score. We compute dynamic Brier scores in the same manner as used in the Dyn-DeepHit codebase (Lee et al., 2020). As with the standard dynamic C-Index the dynamic Brier score requires both a prediction time and a time window to compute. To produce the results in our table, we take the average for each of the five prediction times (selected according to percentiles of the true event times) across four different time windows. For example, with the COVID-19 data

(a)

COVID-19 Data

Day	0	3	4	7	11	Avg (std)
Landmark-Cox	0.79	0.78	0.79	0.77	0.47	0.72 (0.13)
Landmark-RF	0.81	0.85	0.85	0.88	0.85	0.85 (0.02)
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)

Diabetes Mellitus CVD Data

Month	0	3	8	16	30	Avg (std)
Framingham (static)	0.58	0.59	0.59	0.58	0.59	0.59 (0.01)
Landmark-Cox	0.66	0.66	0.66	0.66	0.63	0.66 (0.01)
Landmark-RF	0.66	0.66	0.66	0.66	0.60	0.65 (0.02)
Linear-Δ (proposed)	0.67	0.68	0.67	0.67	0.66	0.67 (0.01)
RNN-Δ (proposed)	0.66	0.68	0.68	0.68	0.65	0.67 (0.01)
Dyn-DeepHit	0.61	0.61	0.60	0.55	0.49	0.57 (0.05)

(c)

PBC2 Data

Month	0	4	7	10	13	Avg (std)
Landmark-Cox	0.77	0.81	0.80	0.83	0.89	0.82 (0.04)
Landmark-RF	0.77	0.81	0.81	0.95	0.96	0.86 (0.08)
Linear-Δ (proposed)	0.78	0.80	0.79	0.89	0.87	0.83 (0.04)
RNN-Δ (proposed)	0.79	0.80	0.80	0.80	0.85	0.81 (0.02)
Dyn-DeepHit	0.78	0.81	0.68	0.57	0.74	0.72 (0.08)

Table 2: Results using the proposed At-Risk Dynamic C-Index for three datasets: (a) COVID-19 data, (b) diabetes mellitus CVD data, and (c) PBC2 data. Larger scores are better. Standard deviations reflect variability across the five times.

COVID-19 Data

(a)

Day	0	3	4	7	11	Avg (std)
Majority Class	0.076	0.090	0.085	0.060	0.027	0.067 (0.023)
Landmark-Cox	0.064	0.085	0.081	0.058	0.030	0.064 (0.020)
Landmark-RF	0.063	0.068	0.064	0.047	0.024	0.053 (0.016)
Linear-Δ (proposed)	0.067	0.071	0.068	0.049	0.026	0.056 (0.017)
RNN-Δ (proposed)	0.069	0.078	0.075	0.054	0.029	0.061 (0.018)
Dyn-DeepHit	0.065	0.070	0.067	0.050	0.025	0.055 (0.017)

Diabetes Mellitus CVD Data

(b)

Month	0	3	8	16	30	Avg (std)
Majority Class	0.170	0.134	0.114	0.090	0.070	0.115 (0.035)
Landmark-Cox	0.132	0.108	0.096	0.078	0.066	0.096 (0.023)
Landmark-RF	0.113	0.111	0.097	0.080	0.065	0.098 (0.024)
Linear-Δ (proposed)	0.146	0.132	0.124	0.112	0.106	0.124 (0.014)
RNN-Δ (proposed)	0.138	0.124	0.120	0.107	0.101	0.118 (0.013)
Dyn-DeepHit	0.137	0.112	0.098	0.080	0.063	0.098 (0.026)

PBC2 Data

(c)

Month	0	4	7	10	13	Avg (std)
Majority Class	0.160	0.197	0.137	0.105	0.071	0.134 (0.044)
Landmark-Cox	0.094	0.137	0.121	0.151	0.130	0.126 (0.019)
Landmark-RF	0.095	0.125	0.096	0.083	0.059	0.092 (0.021)
Linear-Δ (proposed)	0.093	0.127	0.130	0.142	0.100	0.118 (0.019)
RNN-Δ (proposed)	0.107	0.114	0.119	0.123	0.109	0.115 (0.060)
Dyn-DeepHit	0.118	0.105	0.125	0.102	0.051	0.100 (0.026)

Table 3: Results for Brier score on three datasets: (a) COVID-19 data, (b) diabetes mellitus CVD data, and (c) PBC2 data. Lower scores are better. Standard deviations reflect variability across the five times.

each of the columns representing prediction times 0, 3, 4, 7, and 11 days is averaged across four time windows of 3, 6, 9, and 12 days. See **Appendix A** for the time-windows used for each dataset. Each sub-table of Table 3 also contains an additional row showing base-rates for the Brier score. These base rates are computed by predicting the majority class on the test (not train) data itself. This row is intended to provide a scale for the reported Brier scores in order to help judge the difficulty of achieving better (lower) Brier scores on each dataset at different times, and is not intended as a baseline itself (a true less-optimistic baseline would instead pick the majority class from training and then test the results on unseen data).

For the COVID-19 data evaluated on the at-risk dynamic C-Index, the Linear-Δ model has better performance than the RNN-Δ and Dynamic-DeepHit models while being slightly outperformed by the Landmark-RF model. All models see a drop in performance at day 11 for the COVID-19 data, which especially affects the Landmark-Cox model causing it to perform the poorest on this data. For the Brier score, the Landmark-RF, Linear-Δ, and Dynamic-DeepHit all perform similarly. All models have better performance than majority class prediction on the test data.

For the diabetes mellitus CVD data evaluated on the at-risk dynamic C-Index, we also include the static clinical risk score from the Framingham heart study (D'Agostino et al., 2008). For this data, both of our Δ models perform similarly to the Landmark-Cox and Landmark-RF baselines while outperforming the (static) Framingham risk score. Here Dynamic-DeepHit has performance similar to the Framingham risk score, which may come from fitting on the not-at-risk patients. In general, differences between models' discriminative performance as evaluated by the at-risk dynamic C-Index are small for this dataset. On the Brier score the two Δ models have poorer performance compared to the others. Since the Δ models have slightly better discriminative performance, this implies their calibration is worse given that Brier score can be decomposed into a combination of calibration and discrimination. This poorer calibration of the Δ based models could come from using a Rayleigh distribution for the global parametric density which will have worse fit with longer timescales as those found in the DM-CVD data.

For the PBC2 dataset evaluated on the at-risk dynamic C-Index, the two landmark baselines and the Linear-Δ model all have similar performance. The Landmark-RF model has the highest discriminative performance of the models for this data. On the Brier score all models have better performance than majority class prediction on the test data.

Overall the Landmark-RF and Linear-Δ model have the best performance on the at-risk dynamic C-Index across all three datasets. Dynamic-DeepHit and the Landmark-Cox models have poorer performance on this metric overall. For Brier score, overall the Δ methods had weaker performance likely due to miscalibration from the use of Rayleigh distributions as the parametric form, suggesting potential improvement by using more flexible modeling approaches for the global distribution in order to maintain both calibration and discriminative performance. It is also interesting to note that Dynamic-DeepHit's performance on the Brier score is much better than on the at-risk dynamic C-Index. This discrepancy is likely due to the inclusion of an additional loss term in their training loss which encourages good performance on the standard dynamic C-Index by penalizing incorrectly-ranked individuals. Such a loss term could cause the model to focus on non-at-risk individuals, and

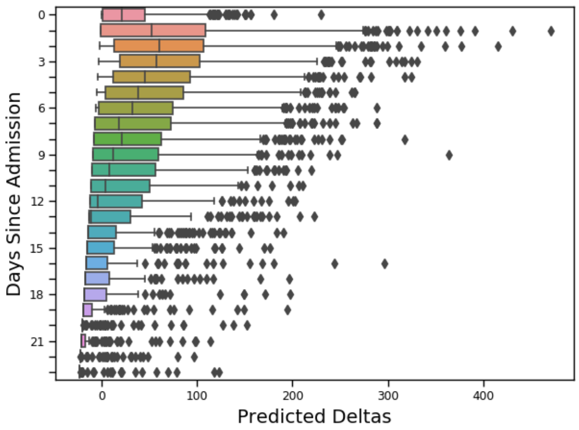

Figure 3: Boxplots of the predicted Δ's at different prediction times for the Linear-Δ model on the COVID-19 data. For each prediction time we collect all individuals who are still in the hospital, and plot a boxplot for the Δ's output by the model for that set of individuals. The minimum possible Δ at time t is negative t. As time passes the Δ's tend to become smaller.

therefore decrease performance on the at-risk dynamic C-Index (which only evaluates over at-risk individuals).

Figure 3 shows boxplots of the Δ's learned by the Linear-Δ model for different prediction times on the COVID-19 data. 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. The top row of Figure 4 shows the predicted individualized densities over time-until-severe-outcome for two hospitalized individuals from the COVID-19 dataset. Since our approach uses a simple parametric density around which to structure it's predictions, we are able to easily visualize the learned densities, whereas visualization is more challenging with other approaches. We also show the predicted hazards for the same two individuals in Figure 4. 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 Figure 4 starts off with slightly higher than average hazard after COVID-19 diagnosis, but after developing a fever, and a high difference between systolic and diastolic 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 day 4, and then after the hazard decreases to zero over time they leave the hospital with no severe outcome. We further explore interpretation of the the Linear-Δ model in **Appendix B** showing the top ten learned features, and the predicted hazards split by demographic.

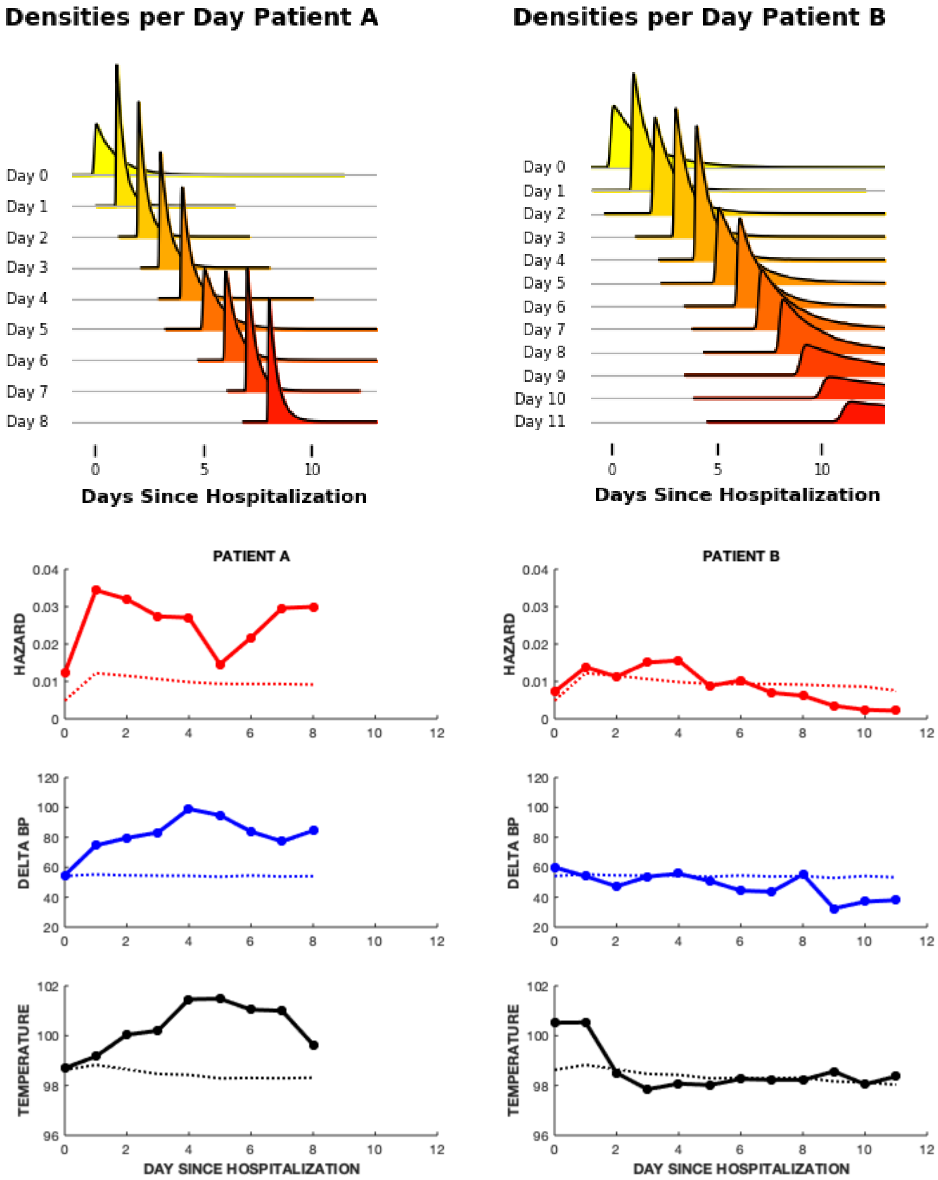

Figure 4: The top row shows the predicted individualized density for two hospitalized patients. The rows below show the results for the same individuals, illustrating the model's predicted hazard function evaluated at the start of each day, along with the difference between systolic blood pressure and diastolic blood pressure, and temperature. Average values, shown as dashed lines, are for comparison to the individual values shown as bolded lines.

7. Discussion

In terms of limitations of this work, the experimental results described in this paper are only for three datasets (two non-public and one public). Further evaluation on additional datasets would be valuable to explore in more detail the robustness of the comparisons between baselines and the proposed models.

In addition, the choice of parametric form for the global distribution in our proposed Δ models may be a significant limiting factor in terms of the representational capacity of these models since it limits the global hazard function to an increasing linear function of time since the synchronizing event. Exploration of more flexible global representations is a natural direction for future investigation.

Another unresolved question is why the additional modeling power of the RNN-based models (Dyn-DeepHit and RNN-Δ) did not perform any better (and were often worse) than alternatives based on linear representations or random forests. A potential reason for this is that the deep models did not have sufficient data in our experiments to fully utilize their capacity. The largest dataset (DM-CVD) has 16,335 patient sequences, which is orders of magnitude less data than typical sequence prediction tasks in natural language processing (for example) where deep recurrent models have been particularly successful. A natural question is whether deep models can show systematic improvements over alternative methods for dynamic survival analysis on problems with significantly larger training datasets.

8. Conclusions

In this paper, we introduce a new class of personalized dynamic survival models which learn a simple global parametric density and perform individualization by locating individuals along the time axis of that global density. We also illustrated the importance of careful evaluation of the predictive performance of models for dynamic survival analysis in a healthcare context. In particular, evaluation metrics may need to be adapted to prevent inflated performance of such models caused by exploiting clinically irrelevant information. We introduced the at-risk dynamic C-Index to help bridge this gap since it is more suited for making predictions in a clinical context evaluating over only at-risk patients, i.e., patients for whom an intervention to prevent a negative health outcome of interest is still possible.

Acknowledgments

We thank the MLHC reviewers for their suggestions on improving the original version of this paper. This work was supported in part by the HPI Research Center in Machine Learning and Data Science at UC Irvine (PP, PS), by NIA grants 1R01AG065330-01A1 (HZ, HD) and 5R01AG054467-04 (HZ, HD), by NCATS grant 2UL1TR001445-06A1 (HZ), and by an NSF Graduate Fellowship (AB).

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. doi: 10.1002/sim.2427.

Sripal Bangalore, Rana Fayyad, Rachel Laskey, David A. DeMicco, Franz H. Messerli, and David D. Waters. Body-weight fluctuations and outcomes in coronary disease. *New England Journal of Medicine*, 376(14):1332–1340, 2017. doi: 10.1056/NEJMoa1606148.

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):22129, 2020. doi: 10.1038/s41598-020-79142-z.

Ralph B. D'Agostino, Ramachandran S. Vasan, Michael J. Pencina, Philip A. Wolf, Mark Cobain, Joseph M. Massaro, and William B. Kannel. General cardiovascular risk profile for use in primary care. *Circulation*, 117(6):743–753, 2008. doi: 10.1161/CIRCULATIONAHA.107.699579.

S Fotso et al. Pysurvival: open source package for survival analysis modeling, 2019.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, Irene Y. Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020. URL https://pubmed.ncbi.nlm.nih.gov/32477638.

Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. doi: 10.1109/TPAMI.2020.2979450.

Alpana Kumar Gupta, Suzanne Tanya Nethan, and Ravi Mehrotra. Tobacco use as a well-recognized cause of severe covid-19 manifestations. *Respiratory Medicine*, 176:106233, 2021. doi: 10.1016/j.rmed.2020.106233.

Jr Harrell, Frank E., 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. doi: 10.1001/jama.1982.03320430047030.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4353–4363. PMLR, Jul 2020. URL http://proceedings.mlr.press/v119/horn20a.html.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008. doi: 10.1214/08-AOAS169.

Daniel Jarrett, Jinsung Yoon, and Mihaela 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.

Changhee Lee, Jinsung Yoon, and Mihaela 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.

Changhee Lee, Jem Rashbass, and Mihaela van der Schaar. Outcome-oriented deep temporal phenotyping of disease progression. *IEEE Transactions on Biomedical Engineering*, 68 (8):2423–2434, 2021. doi: 10.1109/TBME.2020.3041815.

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. doi: 10.1093/bib/bbx044.

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, 2014. doi: 10.1080/01621459.2013.842488.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

Brandt D. Pence. Severe covid-19 and aging: are monocytes the key? *GeroScience*, 42(4): 1051–1061, 2020. doi: 10.1007/s11357-020-00213-0.

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, 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.

Dimitris Rizopoulos, Geert Molenberghs, and Emmanuel M.E.H. Lesaffre. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276, 2017. doi: 10.1002/bimj.201600238.

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. doi: 10.1038/s42256-019-0048-x.

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, 2020. doi: 10.1038/s41598-020-77220-w.

Lukasz Szarpak, Kurt Ruetzler, Kamil Safiejko, Michal Hampel, Michal Pruc, Luiza Kanczuga - Koda, Krzysztof Jerzy Filipiak, and Milosz Jaroslaw Jaguszewski. Lactate dehydrogenase level as a covid-19 severity marker. *The American Journal of Emergency Medicine*, 45:638–639, 2021. doi: 10.1016/j.ajem.2020.11.025.

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.

Hans C. Van Houwelingen and Hein Putter. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, Inc., USA, 1st edition, 2011. ISBN 1439835330.

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, 2018. doi: 10.1093/gerona/gly038.

Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (ehrs): A survey. *ACM Comput. Surv.*, 50(6), January 2018. doi: 10. 1145/3127881.

Bin Zhu, Xiaokai Feng, Chunguo Jiang, Song Mi, Liya Yang, Zhigang Zhao, Yong Zhang, and Liming Zhang. Correlation between white blood cell count at admission and mortality in covid-19 patients: a retrospective study. *BMC Infectious Diseases*, 21(1):574, 2021. doi: 10.1186/s12879-021-06277-3.

Appendix A: Results with the Standard Dynamic C-Index

COVID-19 Data

(a)

Day	0	3	4	7	11	Avg (std)
Landmark-Cox	0.79	0.79	0.80	0.74	0.61	0.75 (0.07)
Landmark-RF	0.82	0.75	0.76	0.70	0.79	0.80 (0.02)
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 (0.02)

Diabetes Mellitus CVD Data

(b)

Month	0	3	8	16	30	Avg (std)
Framingham (static)	0.58	0.59	0.58	0.58	0.58	0.58 (0.003)
Landmark-Cox	0.67	0.66	0.66	0.65	0.61	0.65 (0.02)
Landmark-RF	0.67	0.66	0.66	0.67	0.63	0.66 (0.02)
Linear-Δ (proposed)	0.68	0.71	0.75	0.81	0.86	0.76 (0.07)
RNN-Δ (proposed)	0.67	0.64	0.67	0.79	0.86	0.73 (0.09)
Dyn-DeepHit	0.67	0.69	0.71	0.74	0.74	0.71 (0.03)

PBC2 Data

(c)

Day	0	4	7	10	13	Avg (std)
Landmark-Cox	0.81	0.85	0.84	0.77	0.78	0.81 (0.03)
Landmark-RF	0.81	0.83	0.81	0.78	0.77	0.80 (0.02)
Linear-Δ (proposed)	0.78	0.83	0.82	0.79	0.80	0.81 (0.02)
RNN-Δ (proposed)	0.81	0.86	0.84	0.83	0.82	0.83 (0.02)
Dyn-DeepHit	0.80	0.84	0.77	0.66	0.71	0.76 (0.07)

Table 4: Results on the three datasets on the Standard Dynamic C-Index, (a) COVID-19 data, (b) the Diabetes Mellitus CVD data, and (c) the PBC2 data.

Table 4 shows the performance on the three datasets of all models for the standard dynamic C-Index. The standard dynamic C-Index requires both a prediction time, and a time window. Therefore to generate each entry of this table, the performance at each prediction time is averaged over four time windows. These time windows were $[3, 6, 9, 12]$, $[7, 14, 21, 28]$, and $[2, 4, 6, 8]$ in the corresponding time units (days, months, and months) for the COVID-19 data, DM-CVD data, and PBC2 data respectively. For the COVID-19 data we see that Dynamic-DeepHit apparently has better performance (0.87) on average than any of the other methods. However this gap in performance disappears for the at-risk dynamic C-Index in Table 2(a), and ranking by the inverse of the number of days since COVID-19 diagnosis (1/(N-Days)) leads to performance well above chance as shown in Table 5. For the diabetes mellitus CVD data, the 1/(N-Days) ranking has even higher performance as shown in Table 5. The performance of all three machine learning based

1/(N-Days) Ranking

	At-Risk Dynamic C-Index						**Standard Dynamic C-Index**					
Day/Month	T0	T1	T2	T3	T4	Avg (std)	T0	T1	T2	T3	T4	Avg (std)
COVID-19	0.48	0.49	0.49	0.51	0.52	0.50 (0.01)	0.50	0.59	0.60	0.60	0.60	0.58 (0.04)
DM-CVD	0.49	0.49	0.49	0.48	0.48	0.49 (0.01)	0.50	0.59	0.67	0.75	0.82	0.67 (0.13)
PBC2	0.48	0.45	0.47	0.53	0.50	0.49 (0.03)	0.50	0.65	0.75	0.75	0.78	0.69 (0.10)

Table 5: Results per prediction time with the 1/(N-Days) ranking. Prediction times are represented with capital T's since they vary per dataset.

approaches (Linear-Δ, RNN-Δ, and Dynamic-DeepHit) outperform the other two baseline models on the dynamic C-Index for the diabetes mellitus dataset. However when compared to the at-risk dynamic performance in Table 2 the performance of the machine learning approaches is near identical to the landmarked Cox baseline. This clearly demonstrates the potential for inflated performance estimates when the standard dynamic C-Index is used for evaluation.

Appendix B: Interpretation of Linear-Δ Model for COVID-19 Data

The top feature, and two other features included in the top ten features with the highest weights for the Linear-Δ model are all related to tobacco usage which has been linked to severe COVID-19 outcome (Gupta et al., 2021). White blood cell count, monocytes percentage (among elderly in particular), and lactate dehydrogenase have also all been suggested as markers for COVID-19 disease severity (Zhu et al., 2021; Pence, 2020; Szarpak et al., 2021). Given this, the Linear-Δ version of our model seems to have recovered important biomarkers of COVID-19 severity with real clinical usefulness.

Name	Weight	Missing %	Static or Dyanmic
Smokeless Tobacco Use	11.45	0.5%	Static
White Blood Cell Count	11.31	53%	Dynamic
Lactate Dehydrogenase (LDH)	10.90	80%	Dynamic
Platelet Count	-10.75	53%	Dynamic
Monocytes %	-10.47	59%	Dynamic
Blood Urea Nitrogen	10.28	49%	Dynamic
Tobacco Use Missing	10.01	0%	Static
Smoking Tobacco Use	9.83	0.5%	Static
C-Reactive Protein	9.59	75%	Dynamic
Alanine Transaminase (ALT)	-9.58	70%	Dynamic

Table 6: The top ten features in terms of the linear weights produced by the Linear-Δ model. Missing percentages are computed across all encounters and all individuals. Note that the tobacco-use-missing feature is completely observed by definition.

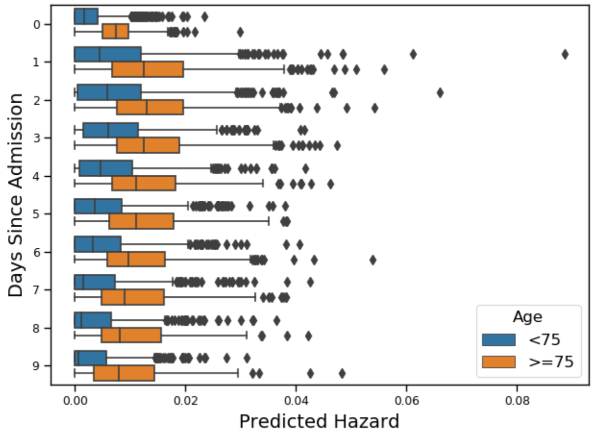

Figure 5: Hazard boxplots from the Linear-Δ model for the COVID-19 data for each day since hospital admission split by age. The more elderly patients have higher predicted risk as expected, and those younger than 75 see their risks decreasing at a faster rate each day than those over 75.

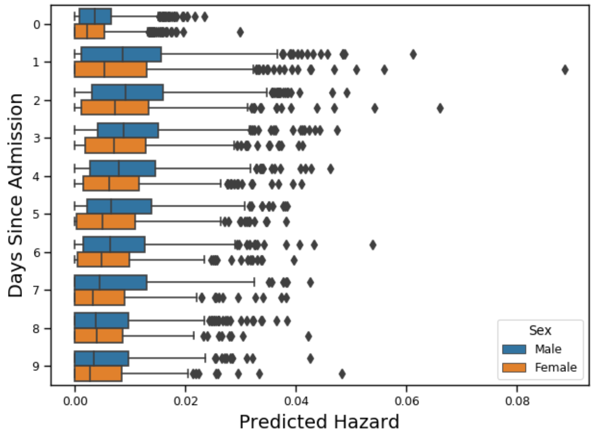

Figure 6: Hazard boxplots from the Linear-Δ model for the COVID-19 data at each day since hospital admission, split by sex. There is a small but consistent difference with males having higher risks for the first 6 days. Afterwards, the differences between the sexes disappears.