Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis with Competing Risks based on Longitudinal Data

Changhee Lee, Jinsung Yoon, *Member, IEEE*, and Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*—Currently available risk prediction methods are limited in their ability to deal with complex, heterogeneous, and longitudinal data such as that available in primary care records, or in their ability to deal with multiple competing risks. This paper develops a novel deep learning approach that is able to successfully address current limitations of standard statistical approaches such as landmarking and joint modeling. Our approach, which we call Dynamic-DeepHit, flexibly incorporates the available longitudinal data comprising various repeated measurements (rather than only the last available measurements) in order to issue dynamically updated survival predictions for one or multiple competing risk(s). Dynamic-DeepHit learns the time-to-event distributions without the need to make any assumptions about the underlying stochastic models for the longitudinal and the time-to-event processes. Thus, unlike existing works in statistics, our method is able to learn data-driven associations between the longitudinal data and the various associated risks without underlying model specifications. We demonstrate the power of our approach by applying it to a real-world longitudinal dataset from the UK Cystic Fibrosis Registry which includes a heterogeneous cohort of 5,883 adult patients with annual follow-ups between 2009-2015. The results show that Dynamic-DeepHit provides a drastic improvement in discriminating individual risks of different forms of failures due to cystic fibrosis. Furthermore, our analysis utilizes post-processing statistics that provide clinical insight by measuring the influence of each covariate on risk predictions and the temporal importance of longitudinal measurements, thereby enabling us to identify covariates that are influential for different competing risks.**

***Index Terms*—Dynamic survival analysis, competing risks, longitudinal measurements, time-to-event data, deep learning, cystic fibrosis**

I. Introduction

SURVIVAL analysis informs our understanding of the relationships between the (distribution of) first hitting times of events of interest (such as death, onset of a certain disease, etc.) and the covariates, and enables us to issue corresponding risk assessments for such events. Clinicians use survival analysis to make screening decisions or to prescribe treatments, while patients use the information about their clinical risks to adjust their lifestyles in order to mitigate such risks. Since the Cox proportional hazard model [2] was first introduced, a variety of methods have been developed for survival analysis, ranging from statistical models to deep learning techniques [3]–[9].

A key limitation of existing survival models is that they utilize only a small fraction of the available longitudinal (repeated) measurements of biomarkers and other risk factors. In particular, even though biomarkers and other risk factors are measured repeatedly over time, survival analysis is typically based on the last available measurement. This represents a severe limitation, since the evolution of biomarkers and risk factors has been shown to be informative in predicting the onset of disease and various risks. For example, Cystic Fibrosis (CF), which is the most common genetic disease in Caucasian populations [10], gives rise to different forms of dysfunction involving the respiratory and gastrointestinal systems, which primarily lead to progressive respiratory failure [11], [12]. Forced expiratory volume (FEV_1), and its development, is a crucial biomarker in assessing the severity of CF as it allows clinicians to describe the progression of the disease and to anticipate the occurrence of respiratory failures [12], [13]. Therefore, to provide a better understanding of disease progression, it is essential to incorporate longitudinal measurements of biomarkers and risk factors into a model. Rather than discarding valuable information recorded over time, this allows us to make better risk assessments on the clinical events.

This paper presents a deep neural network, which we call *Dynamic-DeepHit* (demonstrator available in [1]), that extends our previous work in [5] to dynamic survival analysis. Dynamic-DeepHit learns, on the basis of the available longitudinal measurements, a data-driven distribution of first hitting times of competing events. Thus, the proposed method completely removes the need for explicit model specifications (i.e., no assumption about the form of the underlying stochastic processes are made) and learns the complex relationships between trajectories and survival probabilities. An important aspect of our method is that it naturally handles situations in which there are multiple competing risks where more than one type of event plays a role in the survival setting. (Competing risks are not independent and must be treated jointly; for example, [14] has shown that various treatments for breast cancer increase the risk of a cardiovascular event. See [3], [5] for details of existing survival models that address competing risks.)

To enable dynamic survival analysis with longitudinal time-

C. Lee is with the Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA; (correspondence e-mail: chl8856@ucla.edu).

J. Yoon is with the Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA.

M. van der Schaar is with the Department of Engineering, University of Cambridge, UK, the Alan Turing Institute, London, UK, and the Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA.

to-event data, Dynamic-DeepHit employs a shared subnetwork and a family of cause-specific subnetworks. The shared subnetwork encodes the information in longitudinal measurements into a fixed-length vector (i.e., a context vector) using a recurrent neural network (RNN), which has achieved a great success in various applications handling time-series data (e.g, machine translation [15], image caption generation [16], and speech recognition [17]). We employ a temporal attention mechanism [18] in the hidden states of the RNN structure when constructing the context vector. This renders Dynamic-DeepHit to access the necessary information, which has progressed along with the trajectory of the past longitudinal measurements, by paying attention to relevant hidden states across different time stamps. Then, the cause-specific subnetworks take the context vector and the last measurements as an input and estimate the joint distribution of the first hitting time and competing events that is further used for risk predictions.

To demonstrate the usefulness of our method, we compare its performance with that of competing approaches using a longitudinal data which was collected by the UK Cystic Fibrosis Registry. This data contains a cohort of 5,883 adult patients (from age 18 onwards) suffering from CF, who had annual follow-ups between 2009-2015. Throughout the evaluation, we define two competing events: death from respiratory failures and that from other causes. It is essential to jointly account for competing risks to take preventative steps for CF patients: CF is a systemic disease which gives rise to different forms of dysfunctions in multiple systems and organs – CF-associated liver disease has been reported as the third most frequent cause of death [19]. We show that our method achieves significant improvements on the discriminative performance over the state-of-the-art methods and provides the calibration performance that was comparable to the best performing benchmarks. Particularly, Dynamic-DeepHit achieved improvements of 4.36% and 9.67% over the best benchmark (6.26% and 14.97% over the joint model) on average in terms of discriminative performance for death from respiratory failure and death from other causes, respectively. In addition, while the vast majority of clinical literature has focused on spirometric biomarkers, e.g., $FEV_1\%$ predicted[1], as the main CF risk factors, Dynamic-DeepHit confirmed the importance of the history of intravenous antibiotic treatments and nutritional status in the risk assessment of CF patients. Our demonstrator is available in [1]

II. Related Work

We start by noting that in this work we focus on dynamic survival analysis with competing risks outside the hospital, where the measurements are sparse and irregular, and a disease develops or progresses over the duration of months or even years. Hence, our work differs from existing work on predicting risks in the hospital setting, where numerous measurements are available and a patient is recovering or deteriorating over the course of a few hours or possibly days. For instance, with chronic diseases such as CF, patients are followed up over the span of years, usually as part of regular physical examinations. The clinical status of the patient also evolves slowly, allowing for the development of related comorbidities (e.g. CF-induced diabetes), which in turn affect key biomarkers that reflect a patient's clinical status and rate of deterioration, such as lung function scores (e.g., $FEV_1\%$ predicted) in CF. Thus, we examine related work on dynamic survival analysis that utilizes measurements collected repeatedly, but infrequently, outside the hospital.

The most widely used dynamic survival methods in this setting are joint models which jointly describe both longitudinal and survival processes [20]–[26]. In particular, a joint model comprises two sub-models – one for repeated measurements of the longitudinal process and the other for the time-to-event data (e.g., typically, a linear mixed model and a Cox model) – linking them using a function of shared random effects. Overall, joint models find to learn a full representation of the joint distribution of the longitudinal time-to-event data. From a dynamic prediction perspective, the full representation of joint models leads to a reduced bias in estimation [21] providing flexibility to make predictions at any time points of interest. However, learning such full representation requires an optimization of the joint likelihood and relies on fixed model specifications for both processes. Thus, model misspecifications (e.g., the assumption on longitudinal process and proportional hazard assumption on time-to-event) will limit the overall performance and the optimization of the joint likelihood requires severe computational challenges when applied to high-dimensional datasets [24]. Nonparametric specification of the longitudinal process was previously explored in [22] and [23], which models the longitudinal process via individual-level penalized splines and cubic B-splines, respectively, at the cost of higher computational complexity. Joint models integrating latent classes [25], [26] have been recently developed to account for heterogeneous population. However, these approaches still maintain a proportional hazard assumption which we refrain from doing by adopting deep learning.

Landmarking is another approach for dynamic survival analysis on the basis of longitudinal data [27]–[30]. The basic idea behind landmarking is to build a survival model (e.g., a Cox model), fitted to the subjects from the original dataset who are still at risk at the landmarking time (usually, the prediction time of the interest). Thus, landmarking is "partially conditional" since each survival model is conditioned on the available information accrued by the corresponding landmarking time, rather than incorporating the entire longitudinal history, and predictions on survival probabilities are issued using the last measurements as an estimate of biomarkers at the landmarking time. Even though longitudinal measurements are not fully explored, it is shown that, in practice, landmarking is competitive with joint models and significantly easier to implement [30]. However, landmarking is not fully dynamic; survival predictions are only available at the predefined landmarking times, not at times at which new measurements are obtained. Moreover, it makes assumptions about the underlying stochastic process for the survival model, which may not be true in practice, thereby limiting the model

[1]$FEV_1\%$ predicted is a ratio of the maximum volume of air blown out during lung function test to the predicted value for a 'normal' person of the similar age, sex, and body composition in percentage.

in terms of learning the relationships between the covariates and events of interest. Lastly, it only incorporates a subset of the longitudinal history up to the landmarking time, which may result in information loss when making predictions.

Deep networks have been shown to achieve significantly improved performance in survival analysis [5]–[9] owing to the ability to represent complicated associations between features and outcomes. Authors in [6], [7] have employed deep neural networks for modeling non-linear representations of the relationships between covariates and the risk of a single clinical event. However, these networks are limited to the conventional Cox proportional hazard assumption without addressing time-dependent influences of covariates on the time-to-event. Recently, deep networks have been utilized to develop a nonparametric Bayesian model using the Gaussian process [8], to construct the tree-based Bayesian mixture model [9], and to directly learn the distribution of survival times [5] for survival analysis with competing risks. However, all of these methods provide only static survival analysis: they use only the current information to perform the survival predictions and most of the works focus on a single risk rather than multiple risks. To our best knowledge, this paper is the first to investigate a deep learning approach for dynamic survival analysis with competing risks on the basis of repeated measurements (longitudinal data).

III. Problem Formulation

A. Time-to-Event Data

Time-to-event (survival) data provides three pieces of information for each subject: i) observed covariates, ii) time-to-event(s), and iii) a label indicating the type of event (e.g., death or adverse clinical event) including right-censoring. Observed covariates include static (time-invariant) and time-varying covariates that are recorded for a period of time. We suppose that the longitudinal measurement times, event times, and censoring times are aligned based on a synchronization event, such as the entry to a clinical trial, the date of an intervention, and the onset of a condition.

Formally, for each subject i, a sequence of longitudinal observations until time t is described as a d_x-dimensional multivariate time-series $\mathcal{X}^i(t) = \{\mathbf{x}^i(t^i_j) : 0 \leq t^i_j \leq t \text{ for } j = 1, \cdots, J^i\}$, where $\mathbf{x}^i(t_j)$ can be simplified as $\mathbf{x}^i_j = [x^i_{j,1}, \cdots, x^i_{j,d_x}]$ which includes both static and time-varying covariates recorded at time t_j. Covariates are not necessarily measured at regular time intervals and not every covariate is observed at each measurement (i.e., partially missing). Thus, we i) distinguish notations between time stamps $j = 1, \cdots, J^i$ and the corresponding actual times $t^i_j = t^i_1, \cdots, t^i_{J^i}$ and ii) set $x^i_{j,d} = *$ to denote that the d-th element of $\mathbf{x}^i_j$ was not measured (otherwise, $\mathbf{x}^i_j \in \mathbb{R}$). For notational simplicity, we use $\mathcal{X}^i = \mathcal{X}^i(t^i_{J^i})$ to denote a whole set of longitudinal observations available for subject i until the last measurement time $t^i_{J^i}$ of that subject.

We treat survival time as discrete (e.g., a temporal resolution of one month) and the time horizon as finite (e.g., no patients lived longer than 100 years). Thus, a set of possible survival times is denoted as $\mathcal{T} = \{0, 1, \cdots, T_{\max}\}$ where $T_{\max}$ is a predefined maximum time horizon. Discretization is performed by transforming continuous-valued times into a set of contiguous time intervals, i.e., $T = \tau$ implies $T \in [\tau, \tau + \delta t)$ where δt implies the temporal resolution. We assume that every subject experiences exactly one event among $K \geq 1$ possible events of interest within $\mathcal{T}$. (We cannot observe the occurrence of the other events once one event is observed.) For instance, a patient eventually dies, but can die from only one cause [31]. This includes cause-specific deaths due to CF, where deaths from other causes are competing risks for death due to respiratory failure. Survival data is frequently right-censored because events of interest are not always observed (i.e., subjects are lost to follow-up). The set of possible events is $\mathcal{K} = \{\varnothing, 1, 2, \cdots, K\}$, with $\varnothing$ denoting right-censoring. Throughout this paper, we assume that censoring is *uninformative*. This assumption is common in the survival literature and implies that whether a subject withdraws from the study depends only on the observed history but not on the clinical outcomes [20], [22], [27], [29], [32].

We consider a dataset $\mathcal{D} = \{(\mathcal{X}^i, \tau^i, k^i)\}_{i=1}^N$ comprising survival data for N subjects who have been followed up for a certain amount of time. Here, $\tau^i = \min(T^i, C^i)$ is the time-to-event with $T^i \in \mathcal{T}$ and $C^i \in \mathcal{T}$ indicating the event and the censoring times, respectively, and $k^i \in \mathcal{K}$ being the event or censoring that occurred at time τ^i. Note that τ is either the time at which an event (e.g., death) occurred or the time at which the subject was censored (e.g., disappeared from follow-up); in either case, the subject was known to experience no event at times prior to τ. Fig. 1 depicts a survival dataset comprising histories of longitudinal measurements with different numbers of measurements at irregular time intervals, where each subject experiences either event type 1 or type 2, or has its endpoint censored.

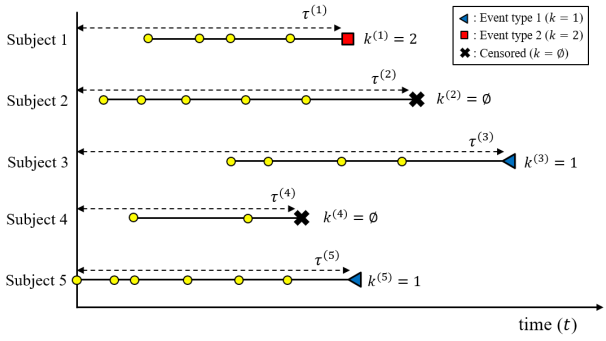
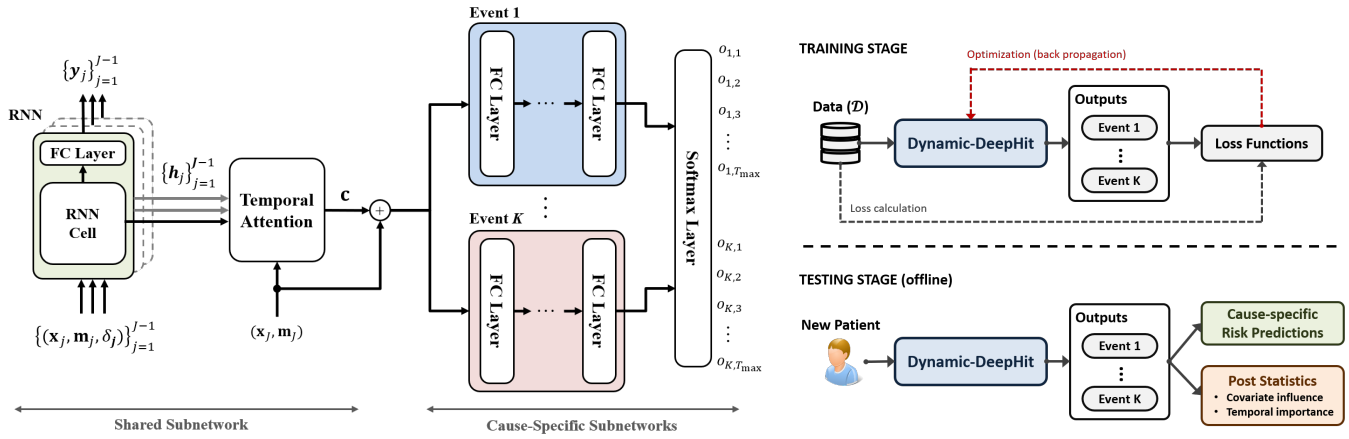


Fig. 1: Illustration of survival data with longitudinal measurements where subjects are aligned based on the synchronization event. Colored dots indicate the times at which longitudinal measurements are observed.

B. Cumulative Incidence Function

Our goal is to analyze the cause-specific risk given the history of observations over time and to issue dynamic risk predictions when new measurements are available. To do so, we use the cause-specific cumulative incidence function (CIF) which is key to survival analysis under the presence of competing risks. As defined in [3], the CIF expresses



(a) The network architecture with K competing risks. (b) A schematic depiction

Fig. 2: An illustration of (a) the network architecture of Dynamic-DeepHit with K competing risks and (b) a schematic depiction of the network at training/testing stages.

the probability that a particular event $k^* \in \mathcal{K}$ occurs on or before time τ^* conditioned on the history of longitudinal measurements $\mathcal{X}^*$. The fact that longitudinal measurements have been recorded up to $t^*_{J^*}$ implies survival of the subject up to this time point. Thus, the CIF is defined as follows:

$$\begin{aligned} F_{k^*}(\tau^*|\mathcal{X}^*) &\triangleq P(T \leq \tau^*, k = k^*|\mathcal{X}^*, T > t^*_{J^*}). \\ &= \sum_{\tau \leq \tau^*} P(T = \tau, k = k^*|\mathcal{X}^*, T > t^*_{J^*}). \end{aligned} \quad (1)$$

Whenever a new measurement is recorded for this subject at time $t > t^*_{J^*}$, we can update (1) accounting for that information in a dynamic fashion.

Similarly, the survival probability of a subject at time τ^* given $\mathcal{X}^*$ can be derived by

$$\begin{aligned} S(\tau^*|\mathcal{X}^*) &\triangleq P(T > \tau^*|\mathcal{X}^*, T > t^*_{J^*}) \\ &= 1 - \sum_{k \neq \varnothing} F_k(\tau^*|\mathcal{X}^*). \end{aligned} \quad (2)$$

However, the *true* CIF, $F_{k^*}(\tau^*|\mathcal{X}^*)$, is not known; we utilize the *estimated* CIF, $\hat{F}_{k^*}(\tau^*|\mathcal{X}^*)$, in order to perform dynamic risk prediction of event occurrences and to assess how models discriminate between cause-specific risks among subjects. The estimated CIF will be described in the next section.

IV. Dynamic-DeepHit

In this section, we describe our novel Dynamic-DeepHit architecture for survival analysis with competing risks on the basis of longitudinal measurements. We seek to train the network to learn an estimate of the joint distribution of the first hitting time and competing events given the longitudinal observations. This representation is then used to estimate the cause-specific CIFs (1) and survival probability (2).

Before describing the network architecture in detail, we redefine the history of longitudinal measurements in order to provide the information on measurement times and missing observations to the network as described in the previous section. Let $\mathcal{X}^i = (\mathbf{X}^i, \mathbf{M}^i, \Delta^i)$ where $\mathbf{X}^i = \{\mathbf{x}^i_1, \cdots, \mathbf{x}^i_{J^i}\}$, $\mathbf{M}^i = \{\mathbf{m}^i_1, \cdots, \mathbf{m}^i_{J^i}\}$ which is a sequence of mask vectors that indicate which covariates are missing, and $\Delta^i = \{\delta^i_1, \delta^i_2 \cdots, \delta^i_{J^i}\}$ which is a sequence of time intervals between two adjacent measurements. Here, $\mathbf{m}^i_j = [m^i_{j,1}, \cdots, m^i_{j,d_x}]$ with $m^i_{j,d} = 1$ if $x^i_{j,d} = *$ and $m^i_{j,d} = 0$ otherwise, and δ^i_j implies the actual amount of time that has elapsed until the next measurements are collected, i.e., $\delta^i_j = t^i_{j+1} - t^i_j$ for $1 \leq j < J^i$, and $\delta^i_{J^i} = 0$. Then, the entire training set can be given as a set of tuples $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{M}^i, \Delta^i, \tau^i, k^i)\}_{i=1}^N$.

A. Network Architecture

Dynamic-DeepHit is a multi-task network, which consists of two types of subnetworks: a shared subnetwork that handles the history of longitudinal measurements and predicts the next measurements of time-varying covariates, and a set of cause-specific subnetworks which estimates the joint distribution of the first hitting time and competing events. As the multi-task learning has been successful across different applications [33]–[36], we jointly optimize the two subnetworks to help the overall network capture associations between the time-to-event under competing risks and i) the static covariates and ii) the progression of underlying process that governs the time-varying covariates. Fig. 2 illustrates (a) the overall architecture of Dynamic-DeepHit which comprises a shared subnetwork and K cause-specific subnetworks and (b) the conceptual framework of the proposed network at training/testing stages. Throughout this subsection, we omit the dependence on i for ease of notation.

1) Shared Subnetwork: The shared subnetwork consists of two components: i) a **RNN structure** to flexibly handle the longitudinal data with each subject having different numbers of measurements, that are captured at irregular time intervals and are partially missing and ii) an **attention mechanism** to unravel the temporal importance of the history of measurements in making risk predictions. For each time stamp $j = 1, \cdots J-1$, the RNN structure takes a tuple of $(\mathbf{x}_j, \mathbf{m}_j, \delta_j)$ as an input and outputs $(\mathbf{y}_j, \mathbf{h}_j)$, where $\mathbf{y}_j$ is the estimate of

time-varying covariates after time δ_j has elapsed, i.e., $\mathbf{x}_{j+1}$[2] and $\mathbf{h}_j$ is the hidden state at time stamp j. Utilizing the Gated Recurrent Unit (GRU) RNN [37], $\mathbf{h}_j$ can be derived as follows:

$$\begin{aligned}
\mathbf{z}_j &= \sigma(W_z\mathbf{h}_{j-1} + U_z[\mathbf{x}_j \ \mathbf{m}_j \ \delta_j] + \mathbf{b}_z),\\
\mathbf{r}_j &= \sigma(W_r\mathbf{h}_{j-1} + U_r[\mathbf{x}_j \ \mathbf{m}_j \ \delta_j] + \mathbf{b}_r),\\
\tilde{\mathbf{h}}_j &= \tanh(W_h(\mathbf{r}_j \odot \mathbf{h}_{j-1}) + U_h[\mathbf{x}_j \ \mathbf{m}_j \ \delta_j] + \mathbf{b}_h),\\
\mathbf{h}_j &= (1-\mathbf{z}_j)\odot \mathbf{h}_{j-1} + \mathbf{z}_j \odot \tilde{\mathbf{h}}_j,
\end{aligned} \tag{3}$$

where W, U, and $\mathbf{b}$ are weight matrices and vectors which parameterize the shared subnetwork, $\odot$ is element-wise multiplication, and $\sigma(\cdot)$ is the sigmoid function. Note that we illustrate the subnetwork with GRUs but other RNNs, such as vanilla RNNs, LSTMs [38], and bidirectional RNNs [39], can be also utilized.

The temporal attention mechanism [18] on the hidden states helps our network decide which parts of the previous longitudinal measurements to pay attention to. Formally, it outputs a context vector, $\mathbf{c}$, as an weighted sum of the previous hidden states as follows:

$$\mathbf{c} = \sum_{j=1}^{J-1} a_j \mathbf{h}_j, \tag{4}$$

where $a_j = \frac{\exp(e_j)}{\sum_{\ell=1}^{J-1}\exp(e_\ell)}$ represents the importance of the j-th measurements. Here, $e_j = f_a(\mathbf{h}_j, \mathbf{x}_J, \mathbf{m}_J)$ is used to score the importance of the j-th measurement by referencing on the last measurement, $(\mathbf{x}_J, \mathbf{m}_J)$. We set $f_a(\cdot)$ as a two-layer feed-forward network that takes the hidden state at time stamp j, $\mathbf{h}_j$, and the tuple of $(\mathbf{x}_J, \mathbf{m}_J)$ as the input and outputs a scalar e_j for $j = 1, \cdots, J-1$. The temporal mechanism is jointly trained with all the other components of our network.

2) Cause-specific Subnetworks: Each cause-specific subnetwork utilizes a feed-forward network composed of fully-connected layers to capture relations between the cause-specific risk and the history of measurements. The inputs to these subnetworks is the context vector of the shared subnetwork. This gives the subnetworks access to the learned common representation of the longitudinal history, which has progressed along with the trajectory of the past longitudinal measurements, by paying attention to relevant hidden states across the time stamps. Overall, each cause-specific subnetwork captures the latent patterns that are distinct to each competing event. Formally, the k-th cause-specific subnetwork takes as input the vector $\mathbf{c}$ and the last measurement $(\mathbf{x}_J, \mathbf{m}_J)$ and outputs a vector, $f_{c_k}(\mathbf{c}, \mathbf{x}_J, \mathbf{m}_J)$.

3) Output Layer: Dynamic-DeepHit employs a soft-max layer in order to summarize the outcomes of each cause-specific subnetwork, $f_{c_1}(\cdot), \cdots, f_{c_K}(\cdot)$, and to map into a proper probability measure. Overall, the network produces an estimated joint distribution of the first hitting time and competing events. In particular, given a subject with $\mathcal{X}^*$, each output node represents the probability of having event k at time τ, i.e., $o^*_{k,\tau} = \hat{P}(T=\tau, k=k|\mathcal{X}^*)$. Therefore, we can define the estimated CIF for cause k^* at time τ^* as follows:

$$\hat{F}_{k^*}(\tau^*|\mathcal{X}^*) = \frac{\sum_{t^*_{J^*}<\tau\le\tau^*} o^*_{k^*,\tau}}{1-\sum_{k\neq\varnothing}\sum_{n\le t^*_{J^*}} o^*_{k,n}}. \tag{5}$$

Note that (5) is built upon the condition that this subject has survived up to the last measurement time.

B. Training Dynamic-DeepHit

To train Dynamic-DeepHit, we minimize a total loss function $\mathcal{L}_{\text{total}}$ that is specifically designed to handle longitudinal measurements and right-censoring. The total loss function is the sum of three terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3, \tag{6}$$

where $\mathcal{L}_1$ is the negative log-likelihood of the joint distribution of the first hitting time and events, which is necessary to capture the first hitting time in the right-censored data, and $\mathcal{L}_2$ and $\mathcal{L}_3$ are utilized to enhance the overall network. More specifically, $\mathcal{L}_2$ combines cause-specific ranking loss functions to concentrate on discriminating estimated individual risks for each cause, and $\mathcal{L}_3$ incorporates the prediction error on trajectories of time-varying covariates to capture the hidden representations of the longitudinal history and to regularize the network.

1) Log-likelihood Loss: The first loss function is the negative log-likelihood of the joint distribution of the first hitting time and corresponding event considering the right-censoring [40], which is extended to the survival setting where the history of longitudinal measurements and K competing risks are available. More specifically, for a subject who *is not censored*, it captures both the event that occurs and the time at which the event occurs; for a subject who *is censored*, it captures the time at which the subject is censored (lost to follow-up) in both cases conditioned on the longitudinal measurements recorded until the last observation. We define $\mathcal{L}_1$ as follows:

$$\begin{aligned}
\mathcal{L}_1 = -\sum_{i=1}^{N}\Bigg[&\mathbb{1}(k^i \neq \varnothing)\cdot\log\Big(\frac{o^i_{k^i,\tau^i}}{1-\sum_{k\neq\varnothing}\sum_{n\le t^i_{J^i}} o^i_{k,n}}\Big)\\
&+ \mathbb{1}(k^i=\varnothing)\cdot\log\Big(1-\sum_{k\neq\varnothing}\hat{F}_k(\tau^i|\mathcal{X}^i)\Big)\Bigg],
\end{aligned} \tag{7}$$

where $\mathbb{1}(\cdot)$ is the indicator function. The first term captures the information provided by uncensored subjects. The second term follows from the knowledge that they are alive at the censoring time, and so the first hitting time of each event $k \in \mathcal{K}$ occurs after the given censoring time; see [41].

2) Ranking Loss: The second loss function incorporates estimated CIFs calculated at different times (i.e., the time at which an event actually occurs) in order to fine-tune the network to each cause-specific estimated CIF. To do so, we utilize a ranking loss function which adapts the idea of concordance [42]: a subject who dies at time τ should have a higher risk at time τ than a subject who survived longer than τ. However, the longitudinal measurements of subjects can begin at any point in their lifetime or disease progression

[2]The time elapsed until the next-time measurements is available since the shared subnetwork only takes the past measurements as inputs.

[43], and this makes direct comparison of the risks at different time points difficult to assess. Thus, we compare the risks of subjects at times elapsed since their last measurements, that is, for subject i, we focus on $s^i = \tau^i - t^i_{J^i}$ instead of τ^i. Define a pair (i, j) an *acceptable pair* for event k if subject i experiences event k at time s^i while the other subject j does not experience any event until s^i (i.e., $s^j > s^i$).[3]

Then, the estimated CIF satisfies the concordance if $\hat{F}_k(s^i + t^i_{J^i}|\mathcal{X}^i) > \hat{F}_k(s^i + t^j_{J^j}|\mathcal{X}^j)$. We define the ranking loss among acceptable pairs of subjects having different histories of measurements as follows:

$$\mathcal{L}_2 = \sum_{k=1}^{K} \alpha_k \sum_{i \neq j} A_{kij} \cdot \eta\Big(\hat{F}_k(s^i + t^i_{J^i}|\mathcal{X}^i), \hat{F}_k(s^i + t^j_{J^j}|\mathcal{X}^j)\Big), \tag{8}$$

where $A_{kij} \triangleq \mathbb{1}(k^i = k, s^i < s^j)$ is an indicator for acceptable pairs (i, j) for event k, $\alpha_k \geq 0$ is a hyper-parameter chosen to trade off ranking losses of the k-th competing event, and $\eta(\cdot)$ is a differentiable loss function. For convenience, we choose here that the coefficients α_k are all equal (i.e., $\alpha_k = \alpha$ for $k = 1, \cdots, K$), and the loss function $\eta(a, b) = \exp(-\frac{a-b}{\sigma})$. Incorporating $\mathcal{L}_2$ into the total loss function penalizes incorrect ordering of pairs and encourages correct ordering of pairs with respect to each event.

3) Prediction Loss: Longitudinal measurements on time-varying covariates, such as the trajectory of biomarkers and the presence of comorbidities over time, may be highly associated with the occurrence of clinical events. Thus, we introduce an auxiliary task in the shared subnetwork, which makes predictions, $\mathbf{y}_j$, on the step-ahead covariates, $\mathbf{x}_{j+1}$, of our interest, to regularize the shared subnetwork such that the hidden representations preserve information for the step-ahead predictions. Taking account missing measurements into consideration, the prediction loss is defined as follows:

$$\mathcal{L}_3 = \beta \cdot \sum_{i=1}^{N} \sum_{j=0}^{J^i - 1} \sum_{d \in \mathcal{I}} (1 - m^i_{j+1,d}) \cdot \zeta(x^i_{j+1,d}, y^i_{j,d}), \tag{9}$$

where $\beta \geq 0$ is a hyper-parameter and $\zeta(a, b) = |a - b|^2$ for continuous covariates and $\zeta(a, b) = -a \log b - (1-a) \log(1-b)$ for binary covariates. By incorporating the missing indicators, the loss is calculated for the step-ahead predictions whose actual measurements are not missing. We select $\mathcal{I}$ as a set of time-varying covariates (e.g., biomarkers or comorbidities) on which we aim to focus the network to be regularized.

C. Discussion on the Scalability

For an accurate estimation of CIFs in (5), it is desirable to have the time interval resolution for discretizing the time horizon (i.e., $\mathcal{T}$ in Section III) to be fine rather than coarse to maintain more information on time-to-event/censoring. However, Dynamic-DeepHit might become over-fitted as it requires the number of output nodes equivalent to $|\mathcal{T}|$ (i.e., inversely proportional to the resolution of the time horizons). To prevent this, we utilize i) early stopping based on the performance metric of our interest (i.e., discriminative performance) and ii) L1 regularization over weights in the cause-specific sub-networks and the output layer. Throughout the experiments, we discretized the time with a resolution of one month that is a fine resolution for longitudinal data with regular follow-ups on a yearly basis, since the time information in the data was mostly available in month format. We show that Dynamic-DeepHit achieves a significant gain in terms of the discriminative performance and provides the calibration performance comparable to the best performing benchmark. We provide more details in the subsequent sections.

[3]An acceptable pair (i, j) naturally captures the right-censoring of subject j since it only considers subjects who lived longer than s^i.

V. Dataset

Experiments were conducted using retrospective longitudinal data from the UK Cystic Fibrosis Registry; this database is sponsored and hosted by the UK Cystic Fibrosis Trust[4]. The registry comprises a cohort of 10,995 patients during annual follow-ups between 2008-2015 with covariates for individual CF patients including demographics, genetic mutations, bacterial infections, comorbidities, hospitalization, lung function scores and therapeutic management. Lung transplantation (LT) is recommended for patients with end-stage respiratory failure as a means to improve life expectancy [44], [45]. Unfortunately, there are more LT candidates than available lung donors, and in addition, the LT procedure is accompanied with serious risks of subsequent post-transplant complications [46].

Meanwhile, complications due to organ transplantation and CF-associated liver disease have been reported as the most frequent causes of death among CF patients after lung-related disease, which share a number of risk factors with respiratory failure [19]. Hence, it is important that patients who are at risk of respiratory failure and other causes be provided with a joint prognosis in order to properly manage LT. More specifically, an effective LT referral policy should efficiently allocate the scarce donor lungs by identifying high-risk patients as candidates for transplant, without overwhelming the LT waiting list with low-risk patients for whom a LT might be an unnecessary exposure to the risk of post-transplant complications or be at risk of other CF-associated diseases [47].

In this paper, we focused on follow-up variables that are available from 2009 – this was due to covariate mismatch between measurements recorded in 2008 and those recorded in the rest of the years. Since transplantation decisions are mostly relevant for adults and deaths in children with CF are now very rare in developed countries [48], we excluded pediatric patients, and included only patients who were aged 18 years or older. Overall, out of 10,995 patients, experiments were conducted on 5,883 adult patients with total of 90 features (11 static covariates and 79 time-varying covariates). For each patient, longitudinal measurements were conducted roughly every year; the time interval between two adjacent measurements ranges from 0 to 69 months with mean of 9.20 months. Here, we discretized the time with a resolution of one month since the date information in the data was mostly available in month format. The number of yearly follow-ups was from 1 to 7 with mean of 5.34 measurements per patients.

[4]https://www.cysticfibrosis.org.uk/the-work-we-do/uk-cf-registry

Among the total of 5,883 patients, 605 patients (10.28%) were followed until death and the remaining 5,278 patients (89.72%) were right-censored (i.e., lost to follow-up). We divided the mortality cause into: i) 491 (8.35%) deaths due to respiratory failures and ii) 114 (1.94%) deaths due to other causes including complications due to organ transplantation and CF-associated liver failure. A full description of the CF dataset, including the data assembly process, and how the missing values were handled in our experiments are available in Appendix A of the supplementary material.

VI. Experiments

The usefulness of a survival model should be assessed primarily by how well the model discriminates among predicted risks and secondarily by how well the model is calibrated. As an illustration in CF, lung transplant is the treatment of last resort for patient with end-stage respiratory failure. Successful transplant can mean many additional years of life for such patients, but there are many more patients in need of transplants than there are available donor lungs. Therefore, it is important to correctly discriminate/prioritize recipients on the basis of risk. However, if the risk predictions of a given model are not well calibrated to the truth (i.e., if there is poor agreement between predicted and observed outcomes), then the model will have little prognostic value for clinicians. As discussed above, we assess the risk predictions of Dynamic-DeepHit with respect to how well the predictions discriminate among individual risks and how accurate the predictions are. In this section, we only reported the discriminative performance; please refer to Appendix H of the supplementary material for the calibration performance in terms of Brier score [49] that was comparable to the best performing benchmarks.

Throughout the experiments, all patients are aligned based on their date of birth to synchronize the time for comparing risk predictions made at different points. More specifically, the time at which measurements are recorded and that at which events or censoring occur is defined as the amount of time elapsed since births. We define the set of possible survival times to be up to 100 years with a monthly time interval, i.e., $T_{\text{max}} = 1200$). Our results are obtained using 5 random 80/20 train/test splits: we randomly separated the data into a training set (80%) and a testing set (20%) and then reserved 20% of the training set as a validation set for hyper-parameter optimization and for early-stopping to avoid over-fitting. The hyper-parameters, such as the coefficients, the activation functions, and the number of hidden layers and nodes of each subnetwork, are chosen utilizing Random Search [50]. The permitted values of the hyper-parameters are listed in Appendix B of the supplementary material. For the prediction loss in (9), we considered two scenarios: i) $\mathcal{I} = \{\text{FEV}_1\% \text{ predicted}\}$ for a fair comparison with the joint models, where $\text{FEV}_1\%$ predicted is a well-known biomarker of the respiratory failure and ii) $\mathcal{I}$ includes all the time-varying covariates including lung function scores, nutritional status, and comorbidities.

A. Benchmarks

We compared Dynamic-DeepHit with state-of-the-art methods that account for dynamic survival analysis under the presence of longitudinal measurements including the joint model [20], the joint model based on latent classes [26], and survival methods under landmarking approaches [29].

In particular, the joint model (**JM**)[5] was implemented using a Bayesian framework that uses MCMC algorithms [51] by modeling the time-to-event data using a cause-specific Cox proportional hazards regression and the longitudinal process using a multivariate linear mixed model. (Due to the computational limitations of standard joint models [24], we selected only $\text{FEV}_1\%$ predicted for the longitudinal process.) To account for the competing risks setting, the cause-specific Cox was created by fixing an event (e.g., death from respiratory cause) and treating the other event (e.g., death from other causes) simply as a form of censoring; see [52]. The joint models integrating latent class (**JM-LC**)[6] to characterize the underlying heterogeneity of the cohort [26] was implemented with $G = 3$ latent classes whose parameters are associated with each class with the similar model specifications to JM.

For the landmarking approaches, we chose the landmarking times as the prediction times, which is age at 30, 40, and 50, and only patients who are at risk at these landmarking times (patients who have not experienced any event or been censored) are considered when we fit survival models at each landmarking time. Overall, the landmarking approaches are implemented utilizing the following survival models: the cause-specific version of the Cox proportional hazards model (**cs-Cox**)[7] and random survival forests under competing risks (**RSF**)[8] [4] with 1000 trees, as a non-parametric alternative of the Cox model.

Please refer to Appendix C of the supplementary material for more details on the benchmark implementations.

B. Discriminative Performance

In this subsection, we present the performance metric that is extended to the survival setting with competing risks and longitudinal measurements, and then we evaluate Dynamic-DeepHit in terms of this metric. To assess the discriminative performance of the various methods, we use a cause-specific time-dependent concordance index ($C_k(t, \Delta t)$), which is an extension of the time-dependent concordance index[9] in [53] adapted to the competing risks setting with longitudinal measurements; similar extensions[10] are made in [54], [55]. More specifically, $C_k(t, \Delta t)$ takes both prediction and evaluation

[5]https://cran.r-project.org/web/packages/JMbayes/

[6]https://cran.r-project.org/web/packages/lcmm/

[7]https://cran.r-project.org/web/packages/survival/

[8]https://cran.r-project.org/web/packages/randomForestSRC/

[9]This metric is suitable for evaluating discriminative performance at different time horizons once risk predictions are issued with the same condition. However, since the time horizon at which risk predictions are made is not considered, this metric cannot be directly used in the longitudinal setting.

[10]This metric provides area under ROC curve (AUC) considering both the prediction and evaluation times. However, it quantifies how well a survival model can order risks at given evaluation time, while our proposed metric quantifies how well a survival model can order risks up to that evaluation time, which better represents the time-to-event setting with right-censoring

times into account to reflect possible changes in risk over time compared to the ordinary concordance index [42], which is a widely used discriminative index in survival analysis.[11] Given the estimated CIF in (5), $C_k(t, \Delta t)$ for event k is defined as

$$C_k(t,\Delta t) = P\Big(\hat{F}_k(t+\Delta t|\mathcal{X}^i(t)) > \hat{F}_k(t+\Delta t|\mathcal{X}^j(t)) \Big| \tau^i < \tau^j, k^i = k, \tau^i < t+\Delta t\Big), \tag{10}$$

where t indicates the prediction time which is the time when the prediction is made to incorporate dynamic predictions and Δt denotes the evaluation time which is the time elapsed since the prediction is made. Throughout the evaluations, $\hat{F}_k(t + \Delta t|\mathcal{X}(t))$ implies the risk of event k occurring in Δt years, which is predicted at age t given the longitudinal measurements until that age.

The discriminative performance of Dynamic-DeepHit on the CF dataset is reported in Table I; means and standard deviations were obtained via 5 random splits. Throughout the evaluation, the tested prediction and evaluation times are in years. Dynamic-DeepHit outperformed the benchmarks for all evaluated prediction and evaluation times with respect to $C_k(t, \Delta t)$ for both causes. All the improvements over the benchmarks were statistically significant; we denoted $*$ for p-value < 0.01 and † for p-value < 0.05. More specifically, on average, Dynamic-DeepHit achieved improvements of 4.36% and 9.67% over the best benchmark (6.26% and 14.97% over JM) for death from respiratory failure and death from other causes, respectively. Please refer to Appendix G of the supplementary material for further comparison especially on the heterogeneous sub-populations.

To provide more fair comparison with JM, we also reported the discriminative performance of simplified versions of Dynamic-DeepHit: i) the proposed network (denoted as **FEV$_1$%**) whose $\mathcal{L}_3$ is computed only based on $\mathcal{I} = \{\text{FEV}_1\%$ predicted$\}$ and ii) the proposed network (denoted as **cause-spec.**) that is separately trained for each cause in a cause-specific manner (by fixing an event and treating the other event as right-censoring). As seen in Table I, the simplified versions still achieved significant performance improvements over JM. It is worth to highlight that, especially for predicting the risk of death from other causes, the full-fledged network achieved performance improvement over the cause-specific version by jointly learning latent representations that are common to competing events.

To further understand the source of gains, we compare Dynamic-DeepHit with the following variations: the network in [5] which performs risk predictions based only on the last available measurements (the dynamic-RNN in the shared subnetwork is replaced with a feed-forward network and the network is trained without $\mathcal{L}_3$) and a deep network utilizing the same architecture with that of Dynamic-DeepHit whose output layer is modified to model the time-to-event data via the Exponential distribution (denoted as **Exponential**; see Appendix C of the supplementary material for details). For the comparison, the same hyper-parameter optimization is applied. Dynamic-DeepHit leverages the RNN architecture to learn the associations between the longitudinal measurements and the time-to-events, and to incorporate the history of the measurements when making risk predictions. Hence, as expected, our method outperformed our previous work in [5], which discards the historical information and relies only on the last available measurements. In contrast to the network which specifies the underlying survival process as Exponential distribution and, thus, is limited to learn the complex interactions with the covariates, our network better discriminates individual risks by directly learning the joint distribution of the first hitting time and the competing events. More experiments on the source of gain –networks that are trained utilizing only parts of the loss functions– can be found in Appendix D of the supplementary material.

In clinical follow-up studies, it is often the case where only a small number of patients are available or where the missing rate of longitudinal measurements varies significantly. To further evaluate the robustness of the proposed network in these scenarios, we reported the discriminative performance by varying the number of training samples and the missing rate of FEV$_1$% predicted in Appendix E and F, respectively.

C. Interpreting Dynamic-DeepHit Predictions

Although deep networks offer tremendous success in predictive ability including survival analysis, low interpretability of the inference process has prevented them from being widely used in medicine. In this subsection, we utilize a post-processing statistic that can be used by clinicians to interpret predictions issued by Dynamic-DeepHit and to understand the associations of covariates and survival over time. It is worth drawing a distinction between interpreting a model, versus interpreting its decision [56], [57]. While interpreting complex models (e.g deep neural networks) may sometimes be infeasible, it is often the case that clinicians only want explanations for the prediction made by the model for a given subject. To help interpret predictions issued by Dynamic-DeepHit, we leverage the partial dependence introduced in [58] by extending it to the survival setting with longitudinal measurements.

Let $\mathcal{X}_d$ be a chosen target subset of the input covariates $\mathcal{X}$ and $\mathcal{X}_{\setminus d}$ be its complement, i.e., $\mathcal{X}_d \cup \mathcal{X}_{\setminus d} = \mathcal{X}$. Then, we can rewrite the estimated CIF in (5) as $\hat{F}_k(\tau|\mathcal{X}) = \hat{F}_k(\tau|\mathcal{X}_d, \mathcal{X}_{\setminus d})$ to explicitly denote the dependency on variables in both subsets. The partial dependence function at time Δt, which is the time elapsed since the last measurement, for event k can be defined as a function of $\mathcal{X}_d$ as follows:

$$\begin{aligned}\gamma_k(\Delta t, \mathcal{X}_d) &= \mathbb{E}_{\mathcal{X}_{\setminus d}}\Big[\hat{F}_k(t_J + \Delta t|\mathcal{X}_d, \mathcal{X}_{\setminus d})\Big] \\ &\approx \frac{1}{N}\sum_{i=1}^{N} \hat{F}_k(t^i_{J^i} + \Delta t|\mathcal{X}_d, \mathcal{X}^i_{\setminus d}),\end{aligned} \tag{11}$$

where t_J indicates the time of the last measurement. Thus, from (11), we can approximately assess how the estimated CIFs are affected by different values of $\mathcal{X}_d$ on average.

[11] The concordance index and its variations are based on the assumption that patients who experienced an event should be assigned a higher risk than those who lived longer (i.e., patients experienced event or was censored afterward). Thus, it naturally handles right-censoring – for example, if both patients are censored, we do not include this pair of patients as defined in (10).

TABLE I: Comparison of $C_k(t, \Delta t)$ (mean $\pm$ std) for various methods. Higher the better.

	Algorithms	Resp. Failure				Other Causes			
		$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 10$	$\Delta t = 1$	$\Delta t = 3$	$\Delta t = 5$	$\Delta t = 10$
$t = 30$	cs-Cox	0.840±0.09†	0.837±0.08†	0.837±0.08†	0.837±0.08†	0.667±0.10*	0.664±0.10*	0.665±0.10*	0.665±0.10*
	RSF	0.936±0.01†	0.932±0.01	0.931±0.02†	0.929±0.01†	0.798±0.04*	0.792±0.04*	0.773±0.05*	0.776±0.05*
	JM	0.882±0.03*	0.896±0.01*	0.896±0.01*	0.897±0.01*	0.760±0.02*	0.795±0.03*	0.802±0.02*	0.812±0.01*
	JM-LC	0.897±0.04†	0.894±0.05†	0.894±0.05†	0.894±0.05†	0.856±0.02*	0.855±0.02*	0.855±0.02*	0.855±0.02*
	[5]	0.910±0.02*	0.907±0.02*	0.907±0.02*	0.907±0.01*	0.819±0.07†	0.831±0.07†	0.834±0.07†	0.839±0.07†
	Exponential	0.895±0.03*	0.890±0.03*	0.890±0.03*	0.890±0.02*	0.824±0.05*	0.825±0.05*	0.824±0.05*	0.824±0.05*
	Proposed								
	$FEV_1\%$	0.948±0.01	0.939±0.01	0.938±0.01	0.937±0.01	0.924±0.02	0.922±0.02	0.921±0.02	0.921±0.02
	cause-spec.	0.946±0.01	0.937±0.02	0.936±0.02	0.933±0.02	0.875±0.04†	0.867±0.05†	0.862±0.05†	0.866±0.05†
	full-fledged	**0.949±0.01**	**0.941±0.01**	**0.942±0.01**	**0.941±0.01**	**0.929±0.02**	**0.927±0.02**	**0.925±0.02**	**0.926±0.02**
$t = 40$	cs-Cox	0.842±0.03*	0.842±0.03*	0.842±0.03*	0.842±0.03*	0.748±0.10*	0.749±0.10*	0.749±0.10*	0.749±0.10*
	RSF	0.888±0.01*	0.887±0.02*	0.886±0.03*	0.891±0.03*	0.803±0.06†	0.771±0.05*	0.749±0.05*	0.746±0.05*
	JM	0.906±0.01*	0.905±0.01*	0.908±0.01*	0.909±0.01*	0.818±0.03*	0.814±0.03*	0.813±0.02*	0.840±0.02*
	JM-LC	0.911±0.04†	0.910±0.04†	0.910±0.04†	0.910±0.04†	0.851±0.02*	0.851±0.02*	0.850±0.02*	0.850±0.02*
	[5]	0.913±0.02*	0.923±0.02*	0.923±0.01*	0.923±0.01*	0.837±0.07†	0.845±0.07†	0.846±0.07†	0.849±0.07†
	Exponential	0.883±0.03*	0.883±0.03*	0.882±0.03*	0.882±0.03*	0.816±0.04*	0.817±0.04*	0.816±0.04*	0.816±0.04*
	Proposed								
	$FEV_1\%$	0.956±0.01	0.958±0.01	0.957±0.01	0.957±0.01	0.934±0.02	0.931±0.02	0.931±0.02	0.931±0.02
	cause-spec.	0.955±0.01	0.957±0.01	0.957±0.01	0.958±0.01	0.907±0.02†	0.909±0.02†	0.906±0.03†	0.909±0.02†
	full-fledged	**0.961±0.01**	**0.963±0.01**	**0.963±0.01**	**0.963±0.01**	**0.939±0.01**	**0.938±0.01**	**0.939±0.01**	**0.939±0.01**
$t = 50$	cs-Cox	0.851±0.11†	0.851±0.11†	0.851±0.11†	0.851±0.11†	0.721±0.09*	0.720±0.09*	0.720±0.09*	0.720±0.09*
	RSF	0.898±0.01*	0.890±0.03*	0.892±0.02*	0.891±0.02*	0.741±0.05*	0.764±0.03*	0.763±0.03*	0.768±0.04*
	JM	0.900±0.01*	0.902±0.01*	0.908±0.01*	0.908±0.01*	0.824±0.03*	0.823±0.02*	0.826±0.01*	0.843±0.02*
	JM-LC	0.916±0.04*	0.916±0.04*	0.916±0.04*	0.916±0.04*	0.852±0.02*	0.852±0.02*	0.852±0.02*	0.853±0.02*
	[5]	0.929±0.01*	0.929±0.01*	0.929±0.01*	0.929±0.01*	0.851±0.07†	0.858±0.06†	0.859±0.06†	0.862±0.06†
	Exponential	0.875±0.02*	0.874±0.02*	0.874±0.02*	0.873±0.02*	0.806±0.04*	0.806±0.04*	0.806±0.04*	0.806±0.04*
	Proposed								
	$FEV_1\%$	0.962±0.01	0.962±0.00	0.962±0.00	0.961±0.00	0.926±0.03	0.935±0.02	0.930±0.02	0.934±0.02
	cause-spec.	0.962±0.01	0.961±0.01	0.944±0.03	0.954±0.02	0.896±0.04†	0.929±0.03	0.929±0.03	0.925±0.03
	full-fledged	**0.968±0.00**	**0.968±0.01**	**0.967±0.01**	**0.967±0.01**	**0.941±0.01**	**0.942±0.01**	**0.943±0.01**	**0.936±0.02**

$*$ indicates p-value < 0.01, † indicates p-value < 0.05

TABLE II: The top 15 most influential covariates with $\Delta t = 5$ year. The values indicate the amount of increase(+)/decrease(-) in the predicted risks on average and the covariates are ranked by the absolute values.

Rank	Death Cause	
	Resp. Failure	Other Causes
1	FEV_1 Predicted (-0.033)	IV ABX Days Hosp. (+0.014)
2	IV ABX Days Hosp. (+0.032)	Gram-Negative (-0.013)
3	Gram-Negative (-0.029)	FEV_1 Predicted (-0.012)
4	FEV_1 (-0.026)	FEV_1 (-0.012)
5	Weight (-0.026)	Weight (-0.011)
6	BMI (-0.025)	BMI (-0.010)
7	Colonic Stricture (-0.024)	Oral Hypo. Agents (-0.008)
8	Oral Hypo. Agents (-0.019)	Class IV Mutation (-0.008)
9	Class IV Mutation (-0.017)	IV ABX Days Home (+0.007)
10	B. Cepacia (+0.016)	Cancer (+0.007)
11	GI Bleed (non-var.) (-0.016)	GI Bleed (var.) (+0.007)
12	O_2 Continuous (+0.015)	HypertonicSaline (-0.006)
13	Drug Dornase (-0.015)	Bone Fracture (-0.006)
14	IV ABX Days Home (+0.014)	Colonic Stricture (-0.006)
15	O_2 Nocturnal (+0.013)	O_2 Nocturnal (+0.006)

IV: intravenous, ABX: antibiotics

To see the influence of covariates on risk predictions issued by Dynamic-DeepHit, we calculated the change in (11) for each covariate $\mathcal{X}_d$ for $d = 1, \cdots, d_{\mathbf{x}}$ by varying the value from its minimum, $x_{d,\min}$, to its maximum, $x_{d,\max}$:

$$\gamma_k(\Delta t, \mathcal{X}_d = x_{d,\min}) - \gamma_k(\Delta t, \mathcal{X}_d = x_{d,\max}). \qquad (12)$$

Table II illustrates the fifteen most influential covariates for the death from respiratory failure and the death from other causes, respectively. Here, we set $\Delta t = 5$ year and the amount of increase/decrease is used to rank the influence. Here, the values imply the averaged increase/decrease of the risk predictions (by varying the covariate from its minimum to maximum) and the signs indicate whether the increase of each covariate increases (+) or decreases (-) the risk predictions.

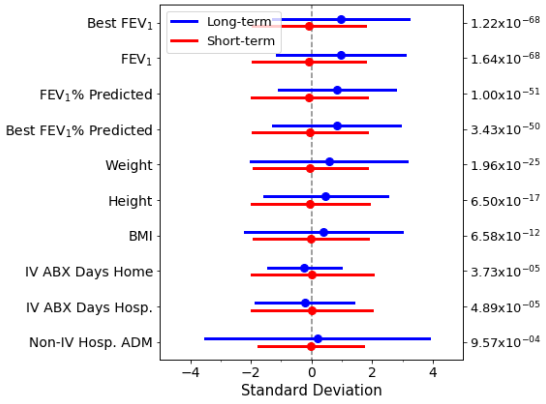
Previous studies in respiratory failures of CF patients have identified $FEV_1\%$ predicted as a strong surrogate for the survival, and have shown that a decrease in $FEV_1\%$ predicted severely increases the mortality of CF patients [10], [11]. Notably, the risk predictions on the respiratory failure made by Dynamic-DeepHit was highly influenced by $FEV_1\%$ predicted in a similar manner. In addition, days of intravenous (IV) antibiotics (ABX), which are used to treat severe bacterial infections, both in hospital and at home, and body mass index (BMI) and weight turned out to be highly influential covariates. This finding is consistent with the domain knowledge, which finds the IV ABX and hospitalization periods are often considered as key risk factors for CF patients [12] and the occurrence of malnutrition, which is often indicated by BMI, is associated with reductions in their survival [59]. More interestingly, the predicted risks for respiratory failure were significantly increased when a patient has Burkholderia cepacia (B. Cepacia), which is a rare but significant threat to CF patients colonizing in the lungs that causes infection and inflammation that deteriorates lung function [60].

For death from other causes, the partial dependence displayed the similar trend, while IV ABX days was more influential to the predicted risks than $FEV_1\%$ predicted was. In particular, the risk predictions for the death from other causes

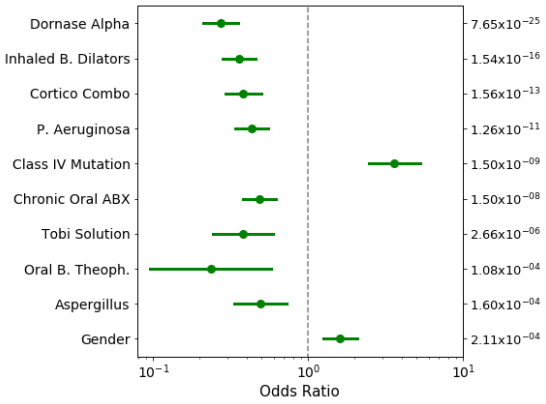
showed slightly different influences from other covariates, such as the indicators of cancer and GI bleeding in variceal source that is a strong sign of liver failure. Therefore, the risk factors and corresponding risk predictions issued by Dynamic-DeepHit need to be carefully interpreted with different priorities depending on the events.

D. Temporal Importance of Longitudinal Measurements

The temporal attention mechanism in the shared subnetwork renders Dynamic-DeepHit to pay special attention to time stamps at which the measurements are important for making risk predictions. To investigate the attention mechanism, we aim this subsection at finding to which patients the network focuses on the long-term (or short-term) dependency of the measurements. For ease of illustration, we define $j^* = \arg\max_{j\in\{1,\cdots,J-1\}} a_j$ as the time stamp at which the proposed network pays the most attention to.



(a) Mean and 95% CI for continuous covariates



(b) Odds ratio and 95% CI for binary covariates

Fig. 3: Forest plots on (a) continuous covariates and (b) binary covariates with the smallest p-values. The left column displays the covariate names and the right column denotes the corresponding p-values. (The covariates are ordered from smallest to largest.)

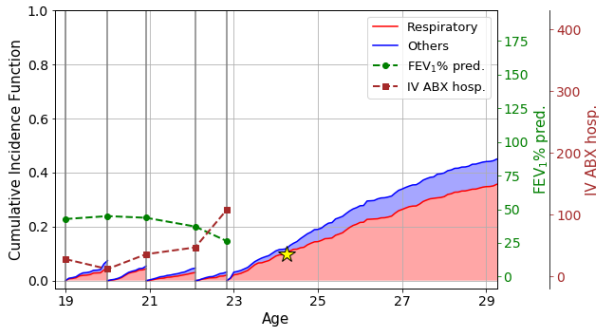
We divide the patients into two groups based on their temporal dependency: the long-term dependency group comprises patients having the highest attention weight to earlier measurements, i.e., $j^* < J-1$, and the short-term dependency group consists of patients having the highest attention weight to the most recent measurement, i.e., $j^* = J-1$. Among 3710 patients with at least three measurements (i.e., $J \geq 3$), our network focused on the long-term dependency of longitudinal measurements for 290 patients (7.82%) and on the short-term dependency for 3420 patients (92.18%). Then, the characteristics of the two groups were compared using independent two-sample t-test for continuous covariates and Fisher's exact test for discrete covariates.

In Fig. 3, we illustrated forest plots on twenty covariates (ten for the continuous and ten for binary covariates) with the smallest p-values, which implies strong evidence that their distributions are different in the two groups. More specifically, for each continuous covariate in Fig. 3(a), we aligned the mean and the 95% confidence interval (CI) of each group with the overall population – this implies that how much the distribution of each group is different from the mean of the overall population in terms of its standard deviation. For an example of Best FEV_1, the mean of the long-term dependency group (i.e., 3.37) was approximately a standard deviation (i.e., 0.92) larger than the overall mean (i.e., 2.44) while that of the short-term dependency group (i.e., 2.36) was very close to the overall mean. For each binary covariate in Fig. 3(b), we displayed the odds ratio (OR) and the 95% CI, which is the ratio of the odds of being in the long-term dependency group in the presence of the covariate and the odds of being in the long-term dependency group without the presence of the covariate – this statistic quantifies the strength of the association between each covariate and being in the long-term dependency group. For instance, if the OR is greater than 1, then the presence of the covariate raises the odds of being in the long-term dependency group.

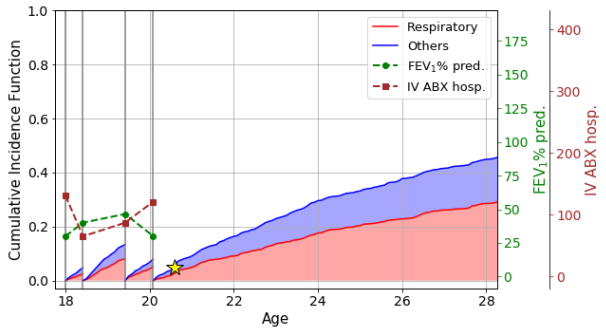
Interestingly, patients in the long-term dependency group displayed, on average, factors that mitigate the predicted risks compared to those in the short-term dependency group. For continuous covariates, as seen in Fig. 3(a), the factors include higher lung functions scores (i.e., FEV_1, FEV_1% predicted, Best FEV_1, and Best FEV_1% predicted), shorter IV ABX periods (i.e., IV ABX days at home and in hospital), richer nutritional status (i.e., weight and BMI), that decrease the predicted risks for both death from the respiratory failure and that from other causes as reported in Table II. For binary covariates, as seen in Fig. 3(b), the factors include lower bacterial infection rate (i.e., pseudomonas aeruginosa and aspergillus whose infection increases the risk of the respiratory failure [60]) and lower therapy/treatments rate (i.e., dornase alpha, cortico combo, chronic oral ABX, and tobi solution). Indeed, Dynamic-DeepHit issued lower risk predictions for patients in the long-term group; the predicted risks were 38.98% and 35.20% lower on average for respiratory failure and death from other causes, respectively.

E. Dynamic Survival Prediction

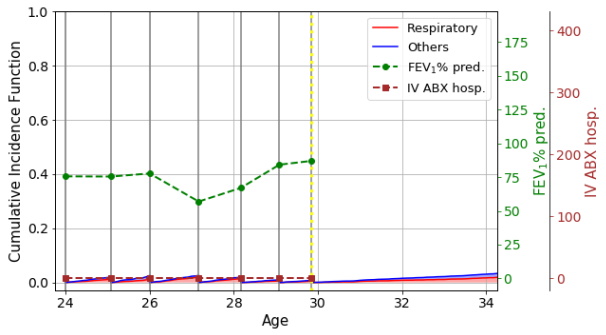
At run-time, Dynamic-DeepHit issues cause-specific risk predictions as defined in (5) for each subject incorporating his/her medical history. Owing to the RNN structure utilized in the shared subnetwork, whenever a new observation is made for that subject, the proposed method is easily able to integrate



(a) A patient died of respiratory failure ($k = 1$)



(b) A patient died of other causes ($k = 2$)



(c) A censored patient ($k = \varnothing$)

Fig. 4: Illustration of dynamic risk predictions issued by Dynamic-DeepHit for patients with (a) $k = 1$, (b) $k = 2$, and (c) $k = \varnothing$. Gray solid lines, yellow dotted lines, and stars indicate times at which measurement are taken, the time at which a patient is censored, and the time at which an event occurred, respectively.

this information into the history of measurements and to issue new risk predictions in a fully dynamic fashion. It is worth highlighting that the landmarking methods can only provide risk assessment at the predefined landmarking times [29]. In Fig. 4, we have illustrated the dynamic survival analysis for representative patients in order to show how Dynamic-DeepHit issues and updates risk predictions for different causes (including right-censoring) with new measurements being collected. Along with the predicted risks, trajectories of two highly influential covariates, $\text{FEV}_1\%$ predicted and IV ABX Days in Hospital, are illustrated to show their associations. As demonstrated in Fig. 4, Dynamic-DeepHit was able to flexibly update the cause-specific risks by incorporating new measurements in a dynamic fashion. For example, the predicted risks for the patient in Fig. 4(a) was relatively high compared to that of the patient in Fig. 4(c), presumably due to the high and increasing IV ABX days in hospital and the decreasing $\text{FEV}_1\%$ predicted. The importance of this dynamic approach can be seen in Fig. 4(a) when a sudden increase in the number of IV ABX days around at age 23 resulted in a steep increase in predicted risks.

VII. Conclusion

In this paper, we developed a novel approach, Dynamic-DeepHit, to perform dynamic survival analysis with competing risks on the basis of longitudinal data. Dynamic-DeepHit is a deep neural network which learns the estimated joint distributions of survival times and competing events, without making assumptions regarding the underlying stochastic processes. We train the network by leveraging a combination of loss functions that capture the right-censoring and the associations of longitudinal measurements with disease progression, both of which are inherent in time-to-event data. We demonstrated the utility of our proposed method through a set of experiments conducted on a cohort of 5,883 adult CF patients whose follow-ups have been recorded in the UK Cystic Fibrosis Registry. The experiments show that the proposed method significantly outperforms the cutting-edge benchmarks in terms of discriminative performance. Supported with a post-processing statistic to interpret risk predictions issued by the proposed method, the results suggest the possibility of improved dynamic analysis on disease progression that will result in more effective health care.

Acknowledgment

This work was supported by the National Science Foundation (NSF) (Grant Number: 1533983, ECCS1462245, and 1722516) and the UK Cystic Fibrosis Trust. We thank Dr. Janet Allen (Director of Strategic Innovation, UK Cystic Fibrosis Trust) for the vision and encouragement. We thank Rebecca Cosgriff and Elaine Gunn for the help with data access, extraction and analysis. We also thank Prof. Andres Floto and Dr. Tomas Daniels, our collaborators, for the very helpful clinical discussions.

References

[1] C. Lee et al., "Demonstrator of Dynamic-Deephit," [Online]. Available: http://www.vanderschaar-lab.com/NewWebsite/CF_Changhee_TBME_demonstrator.html (April, 2019).

[2] D. R. Cox, "Regression models and life tables (with discussion)," *Journal of the Royal Statistical Society. Series B*, vol. 34, pp. 187–220, 1972.

[3] J. P. Fine and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk," *Journal of the American Statistical Association*, vol. 94(446), pp. 496–509, June 1999.

[4] H. Ishwaran et al., "Random survival forests," *The Annals of Applied Statistics*, vol. 2(3), pp. 841–860, September 2008.

[5] C. Lee et al., "Deephit: A deep learning approach to survival analysis with competing risks," *In Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI 2018)*, 2018.

[6] M. Luck et al., "Deep learning for patient-specific kidney graft survival analysis," *arXiv preprint arXiv:1705.10245*, 2017.

[7] J. Katzman et al., "Deep survival: A deep cox proportional hazards network," *arXiv preprint arXiv:1606.00931*, 2016.

[8] A. M. Alaa and M. van der Schaar, "Deep multi-task gaussian processes for survival analysis with competing risks," *In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[9] A. Bellot and M. van der Schaar, "Tree-based bayesian mixture model for competing risks," *In Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018)*, 2018.

[10] S. D. Aarona et al., "A statistical model to predict one-year risk of death in patients with cystic fibrosis," *Journal of Clinical Epidemiology*, vol. 68, pp. 1336–1345, 2015.

[11] T. G. Liou et al., "Use of lung transplantation survival models to refine patient selection in cystic fibrosis," *American Journal of Respiratory and Critical Care Medicine*, vol. 171(9), pp. 1053–1059, 2005.

[12] L. Nkam et al., "A 3-year prognostic score for adults with cystic fibrosis," *Journal of Cystic Fibrosis*, vol. 16(6), pp. 702–708, November 2017.

[13] D. Li et al., "Flexible semiparametric joint modeling: an application to estimate individual lung function decline and risk of pulmonary exacerbations in cystic fibrosis," *Emerging Theme in Epidemiology*, vol. 14, December 2017.

[14] R. J. Koene et al., "Shared risk factors in cardiovascular disease and cancer," *Circulation*, vol. 133, pp. 1104–1114, March 2016.

[15] I. Sutskever et al., "Sequence to sequence learning with neural networks," *In Proceedings of the 27th Conference on Neural Information Processing Systems (NIPS 2014)*, 2014.

[16] A. Karpathy and F.-F. Li, "Deep visual-semantic alignments for generating image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39(4), pp. 664–676, April 2017.

[17] A. Graves et al., "Speech recognition with deep recurrent neural networks," *In Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, 2013.

[18] D. Bahdanau et al., "Neural machine translation by jointly learning to align and translate," *In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, 2015.

[19] N. Kobelska-Dubiel et al., "Liver disease in cystic fibrosis," *Prz Gastroenterol*, vol. 9(3), pp. 136–141, June 2014.

[20] R. Henderson et al., "Joint modelling of longitudinal measurements and event time data," *Biostatistics*, vol. 1(4), pp. 465–480, December 2000.

[21] J. G. Ibrahim et al., "Basic concepts and methods for joint models of longitudinal and survival data," *Journal of Clinical Oncology*, vol. 28(16), pp. 2796–2801, June 2010.

[22] J. Barrett and L. Su, "Dynamic predictions using flexible joint models of longitudinal and time-to-event data," *Statistics in Medicine*, vol. 36(9), pp. 1447–1460, April 2017.

[23] E. R. Brown et al., "A flexible b-spline model for multiple longitudinal biomarkers and survival," *Biometrics*, vol. 61(1), pp. 64–73, March 2005.

[24] G. L. Hickey et al., "Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues," *BMC Medical Research Methodology*, vol. 16, p. 117, September 2016.

[25] Y. Liu et al., "Joint latent class model of survival and longitudinal data: An application to cpcra study," *Computational Statistics & Data Analysis*, vol. 91, pp. 40–50, November 2015.

[26] E.-R. Andrinopoulou et al., "Integrating latent classes in the bayesian shared parameter joint model of longitudinal and survival outcomes," *arXiv preprint arXiv:1502.02072*, 2018.

[27] Y. Zheng and P. J. Heagerty, "Partly conditional survival models for longitudinal data," *Biometrics*, vol. 61, pp. 379–391, March 2005.

[28] P. J. Heagerty and Y. Zheng, "Survival model predictive accuracy and roc curves," *Biometrics*, vol. 61, pp. 92–105, March 2005.

[29] H. C. van Houwlingen, "Dynamic prediction by landmarking in event history analysis," *Scandinavian Journal of Statistics*, vol. 34(1), pp. 70–85, March 2007.

[30] H. C. van Houwelingen and H. Putter, "Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data," *Lifetime Data Analysis*, vol. 14(4), pp. 447–463, December 2008.

[31] T. A. Gooley et al., "Estimation of failure probabilities in the presence of competing risks: New representations of old estimators," *Statistics in Medicine*, vol. 18(6), pp. 695–706, March 1999.

[32] A. A. Tsiatis and M. Davidian, "Joint modeling of longitudinal and time-to-event data: an overview," *Statistica Sinica*, vol. 1(4), pp. 809–834, July 2004.

[33] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," *In Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, pp. 160–167, 2008.

[34] L. Deng et al., "New types of deep neural network learning for speech recognition and related applications: An overview," *In Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pp. 8599–8603, 2013.

[35] B. Ramsundar et al., "Massively multitask networks for drug discovery," *arXiv preprint arXiv:1502.02072*, 2015.

[36] H. Harutyunyan et al., "Multitask learning and benchmarking with clinical time series data," *arXiv preprint arXiv:1703.07771*, 2017.

[37] J. Chung et al., "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9(8), pp. 1735–1780, 1997.

[39] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45(11), pp. 2673–2681, November 1997.

[40] M.-L. T. Lee and G. A. Whitmore, "Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary," *Statistical Science*, vol. 21(4), pp. 501–513, November 2006.

[41] J. F. Lawless, *Statistical Models and Methods for Lifetime Data, 2nd Edition*. Wiley, 2002.

[42] F. E. Harrell et al., "Evaluating the yield of medical tests," *Journal of the American Medical Association*, vol. 247(18), pp. 2543–2546, May 1982.

[43] R. Ranganath et al., "Deep survival analysis," *In Proceedings of the 1st Machine Learning for Healthcare Conference (MLHC 2016)*, 2016.

[44] T. Liou et al., "Survival effect of lung transplantation among patients with cystic fibrosis," *JAMA*, vol. 286(21), pp. 2683–2689, December 2001.

[45] M. Hofer et al., "True survival benefit of lung transplantation for cystic fibrosis patients: the zurich experience," *The Journal of Heart and Lung Transplantation*, vol. 28(4), pp. 334–339, April 2009.

[46] N. Mayer-Hamblett et al., "Developing cystic fibrosis lung transplant referral criteria using predictors of 2-year mortality," *American Journal Respiratory Critical Care Medicines*, vol. 166, pp. 1550–1555, December 2002.

[47] T. G. Liou et al., "Use of lung transplantation survival models to refine patient selection in cystic fibrosis," *American Journal Respiratory Critical Care Medicines*, vol. 171, pp. 1053–1059, May 2005.

[48] D. S. Urquhart et al., "Deaths in childhood from cystic fibrosis: 10-year analysis from two london specialist centres," *Archives of Disease in Childhood*, vol. 98(2), p. 123–127, February 2013.

[49] U. B. Mogensen et al., "Evaluating random forests for survival analysis using prediction error curves," *Journal of Statistical Software*, vol. 50(11), 2012.

[50] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, February 2012.

[51] D. Rizopoulos, "The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc," *Journal of Statistical Software*, vol. 72(7), 2016.

[52] B. Haller et al., "Applying competing risks regression models: an overview," *Lifetime Data Analysis*, vol. 19, pp. 33–58, January 2013.

[53] T. A. Gerds et al., "Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring," *Statistics in Medicine*, vol. 32(13), pp. 2173–2184, 2013.

[54] D. Rizopoulos et al., "Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking," *Biometrical Journal*, vol. 59(6), pp. 1261–1276, November 2017.

[55] K. Suresh et al., "Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model," *Biometrical Journal*, vol. 59(6), pp. 1277–1300, November 2017.

[56] M. T. Ribeiro et al., ""Why should I trust you?": Explaining the predictions of any classifier," *In Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2016)*, 2016.

[57] A. Avati et al., "Improving palliative care with deep learning," *arXiv preprint arXiv:1711.06402*, 2017.

[58] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29(5), pp. 1189–1232, 2001.

[59] A. L. Stephenson et al., "Longitudinal trends in nutritional status and the relation between lung function and BMI in cystic fibrosis: a population-based cohort study," *The American Journal of Clinical Nutrition*, vol. 97(4), p. 822–827, April 2013.

[60] B. Fauroux et al., "Burkholderia cepacia is associated with pulmonary hypertension and increased mortality among cystic fibrosis patients," *Journal of Clinical Microbiology*.