Neurological Prognostication of Post-Cardiac-Arrest Coma Patients Using EEG Data: A Dynamic Survival Analysis Framework with Competing Risks

Xiaobin Shen XIAOBINS@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University

Jonathan Elmer ELMERJP@UPMC.EDU
Department of Emergency Medicine
University of Pittsburgh

George H. Chen GEORGECHEN@CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University

Abstract

Patients resuscitated from cardiac arrest who enter a coma are at high risk of death. Forecasting neurological outcomes of these patients (i.e., the task of *neurological prognostication*) could help with treatment decisions: which patients are likely to awaken from their coma and should be kept on life-sustaining therapies, and which are so ill that they would unlikely benefit from treatment? In this paper, we propose, to the best of our knowledge, the first *dynamic* framework for neurological prognostication of post-cardiac-arrest comatose patients using EEG data: our framework makes predictions for a patient over time as more EEG data become available, and different training patients' available EEG time series could vary in length. Predictions themselves are phrased in terms of either time-to-event outcomes (time-to-awakening or time-to-death) or as the patient's probability of awakening or of dying across multiple time horizons (e.g., within the next 24, 48, or 72 hours). Our framework is based on using any dynamic survival analysis model that supports competing risks in the form of estimating patient-level cumulative incidence functions. We consider three competing risks as to what happens first to a patient: awakening, being withdrawn from life-sustaining therapies (and thus deterministically dying), or dying (by other causes). For some patients, we do not know which of these happened first since they were still in a coma when data collection stopped (i.e., their outcome is censored). Competing risks models readily accommodate such patients. We demonstrate our framework by benchmarking three existing dynamic survival analysis models that support competing risks on a real dataset of 922 post-cardiac-arrest coma patients. Our main experimental findings are that: (1) the classical Fine and Gray model which only uses a patient's static features and summary statistics from the patient's latest hour's worth of EEG data is highly competitive, achieving accuracy scores as high as the recently developed Dynamic-DeepHit model that uses substantially more of the patient's EEG data; and (2) in an ablation study, we show that our choice of modeling three competing risks results in a model that is at least as accurate while learning more information than simpler models (using two competing risks or a standard survival analysis setup with no competing risks).

1. Introduction

Cardiac arrest is one of the leading causes of death and disability worldwide, resulting in approximately 300,000 to 450,000 deaths annually in the U.S. alone (NIH[1], 2022) and a 43% reported rate of suffering from cognitive impairment (Byron-Alhassan et al., 2021). In this paper, we specifically focus on cardiac arrest patients who enter a coma and are admitted to the ICU, where they are placed on life-sustaining therapies (e.g., mechanical ventilation, cardiac support devices). Here, forecasting the neurological outcome of patients (i.e., the task of *neurological prognostication*) is important: if physicians perceive a patient to have poor neurological prognosis, then they may discontinue life-sustaining therapies for the patient, which deterministically ends the patient's life. Withdrawal of life-sustaining therapies accounts for 48% of all nonsurvivors, with 31% occurring in medically unstable patients and 17% in medically stable patients (Matthews et al., 2017). This raises the possibility that some patients may have survived if different decisions had been made regarding their care. Neurological prognostication, therefore, is crucial in determining the appropriate treatment plan for each patient.

In recent years, a promising direction for neurological prognostication has been to take advantage of brain activity measurements using electroencephalography (EEG) (Abend et al., 2010; Glass et al., 2013; Friberg et al., 2015). A number of recent studies have demonstrated these EEG signals are predictive of patients' neurological outcomes (e.g., Thenayan et al. 2010; Rittenberger et al. 2012; Rossetti et al. 2012; Søholm et al. 2014; Oh et al. 2015; Elmer et al. 2016b; Westhall et al. 2016; Shekhar et al. 2023). In this paper, we use both EEG data recorded over time and some patient characteristics collected upon hospital admission.

While prognostication is essential, how it has been modeled in existing literature has some major limitations. First, binary classification has been the most common approach for modeling prognostication for post-cardiac arrest coma patients, where the two classes are poor neurological recovery or death (taken to be the "positive" class) and favorable neurological recovery with certain levels of consciousness at hospital discharge (the "negative" class) (e.g., Rossetti et al. 2016; Admiraal et al. 2021; Moseby-Knappe et al. 2020). The goal is to achieve a high true positive rate (TPR) with a false positive rate (FPR) close to 0 (see, for instance, the overview by Geocadin et al. (2019)). However, the existence of patients for whom life-sustaining therapies were withdrawn complicates this binary classification setup: including these particular patients in training data is problematic because we do not know what their neurological outcomes would have been if they had been kept on life-sustaining therapies (i.e., we do not know which of the two classes they belong to). Some existing studies simply exclude such patients (e.g., De-Arteaga et al. 2019). However, by excluding these patients, we ignore potentially useful information: these patients likely have characteristics that led to physicians withdrawing them from life-sustaining therapies.

Another major limitation of existing work is ignoring the dynamic nature of both the EEG signals as well as physicians' decision-making and only focusing on using EEG data of a fixed period of time to make a prediction at a single point in time. For example, De-Arteaga et al. (2019) only use EEG data between hours 34 to 36 after ICU admission to make a single prediction of the chance of favorable recovery for each patient; any patient with missing EEG data between hours 34 to 36 would be excluded, and EEG information outside of

1. https://www.nhlbi.nih.gov/health/cardiac-arrest

this time window would not be used by their prediction model. Meanwhile, Admiraal et al. (2021) make a single prediction using EEG data 24 hours after cardiac arrest. In practice, physicians make decisions regarding treatment plans at different times after the patients have been admitted to the ICU (e.g., adjusting medications such as the dosage of vasopressors over time, or deciding whether to withdraw life-sustaining therapies), potentially using all information collected of the patient up until present time. To the best of our knowledge, no existing method has been developed for neurological prognostication of these coma patients that is truly *dynamic*, where the training patients' time series can vary in length, and where we make predictions at any point in time.

Our main contribution in this paper is to propose a framework for neurological prognostication of post-cardiac-arrest coma patients that addresses both of the above major limitations. In particular, our framework is dynamic and directly models specific outcomes of interest as *competing risks* in the sense that one outcome happening (e.g., withdrawal from life-sustaining therapies) prevents other outcomes from happening (e.g., awakening, dying of other causes). Moreover, our framework allows for outcomes to be *censored*, meaning that for some patients, we do not get to see their eventual outcome since they were still in a coma by the time data collection stopped. That they were still alive at the end of data collection could be due to specific patient characteristics that make them more likely to be alive rather than to have been pulled off life-sustaining therapies or to have died of other causes.

Our framework builds on a class of existing dynamic survival analysis models that support competing risks; we refer to this class of models as *dynamic competing risks (DCR) models* (we precisely define this class of models in Section 2). An example of such a model is Dynamic-DeepHit (Lee et al., 2019). Roughly, a DCR model predicts, for any test patient with measurements up to time t, the probability of the patient experiencing each of the different competing events at any time after time t.

Even though we learn a DCR model with three competing risks for the neurological prognostication problem, at prediction time, one of these competing risks is often not of primary interest: whether a patient will be withdrawn from life-sustaining therapies. Even if we did predict who would be withdrawn from life-sustaining therapies, we would effectively be predicting how past decisions were made by physicians and *not* what the true neurological outcomes of these patients would have been. Ideally, predictions of the latter should be what we use to assist with treatment decisions. For this reason, we show how to derive a binary classifier from the DCR model's predicted output that aims to predict whether a patient will awaken or die (of causes aside from withdrawal from life-sustaining therapies) within any user-specified time horizon (Section 3.1). Conceptually, whereas some existing work on neurological prognostication excluded patients who were withdrawn from life-sustaining therapy altogether from their analysis (e.g., De-Arteaga et al. 2019), we are including such patients when training a DCR model, and only when using our classifier, we condition on (in a probabilistic sense) the test patient not being withdrawn from life-sustaining therapies in the future. Especially as this classifier is meant to help with decisions such as whether a patient should be withdrawn from life-sustaining therapies, a reasonable assumption is to condition on this event not happening yet. This binary classifier that we derive from the DCR model can classify variable-length input time series using any user-specified time horizon without needing to re-train the DCR model. We further develop a patient-specific heat map visualization for the classifier that is straightforward to interpret (Section 3.2).

In experiments on real data (cohort selection and other dataset details are in Section 4), we benchmark three DCR models to compare how accurate they are, and we also conduct an ablation study to show why our choice of modeling the competing risks setting with three competing risks is better than using two or one instead (Section 5). Note that the one competing risk setting reduces to a dynamic survival analysis setup without competing risks.

Generalizable Insights about Machine Learning in the Context of Healthcare

For neurological prognostication of post-cardiac-arrest coma patients using EEG data, our paper is, to the best of our knowledge, the first to consider a dynamic problem setup. We believe that we have framed this prognostication problem in a manner that is more useful for clinical decision support compared to how it has been framed in existing literature.

As our dynamic problem setup and accompanying evaluation metrics have not previously appeared in literature for our specific clinical application, our two main experimental findings are novel: (1) in benchmarking three DCR models, we find that the classical competing risks model by Fine and Gray (1999) that only uses static patient features and the summary information from the last hour of a patient's EEG data is highly competitive, achieving accuracy scores as high as the recently developed Dynamic-DeepHit model (Lee et al., 2019) that uses substantially more EEG data; and (2) our ablation study shows that our choice of modeling three competing risks results in a model that is at least as accurate while providing more information than a model that uses two competing risks or a dynamic survival analysis model without competing risks. These findings suggest that researchers working on the same clinical application may want to also consider modeling at least the three competing risks we consider (or even finer-grain versions of some of them, such as accounting for more causes of death), and also trying the classical Fine and Gray model as a baseline.

From a technical standpoint, our paper does not introduce a new model. Instead, our paper demonstrates how to effectively use any existing DCR model to address a specific clinical problem. The novelty is thus in the application of existing DCR models and also in our proposal of a classifier (derived from a DCR model) and an accompanying heat map visualization for this classifier. The crucial insight of the classifier that we develop is to condition on a particular event—an action taken by a physician that inevitably ends the patient's life—*not* happening in the future of the patient because the output of the classifier is meant to help with deciding on whether to take this action (where a reasonable assumption is that we have not taken the action as doing so has a permanent consequence). We suspect that this same idea would be relevant in various other clinical problems.

2. Background

Our paper builds on a specific class of existing dynamic survival analysis models, which we refer to as dynamic competing risks (DCR) models. We review this class of models in Section 2.1, where we also state the dynamic problem setting that our framework uses. We briefly give a concrete example of a DCR model (Dynamic-DeepHit by Lee et al. (2019)) in Section 2.2. Note that throughout this paper, for any positive integer m, we frequently use the notation $[m] := \{1, 2, \ldots, m\}$. We typically use uppercase variables to refer to random variables whereas lowercase variables refer to constants, realized values of random variables, or dummy indices (e.g., training data indices).

2.1. Dynamic Problem Setup and DCR Models

Training data We assume that we have a training dataset consisting of n patients. For each training patient $i \in [n]$, we observe a times series with a total of L_i time steps, where at each time step, we observe d features. Specifically, we observe the feature vectors $X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(L_i)} \in \mathbb{R}^d$, where $X_i^{(\ell)} \in \mathbb{R}^d$ is the i-th patient's feature vector at time step $\ell \in [L_i]$ (time steps are sorted chronologically, so time step L_i is the last time step observed for training patient i). Moreover, time step $\ell \in [L_i]$ happens at a time that is recorded as a real number $T_i^{(\ell)} \in \mathbb{R}$, meaning that the amount of time that elapses between time steps ℓ and $\ell + 1$ is $T_i^{(\ell+1)} - T_i^{(\ell)}$. A common assumption is to set the initial time step's time to be $T_i^{(1)} = 0$. Note that for the d features that are tracked over time, it is possible that some always stay the same (e.g., age upon hospital admission). For ease of exposition, we do not introduce additional notation that separates static from time-varying features although separating these two types of features could be done in practice.

As the above notation suggests, patients' time series can vary in length (e.g., one patient could have EEG data recorded every second for 6 hours, whereas another could have EEG data recorded every second for 12 hours). For the real data we consider later, the time series are regularly sampled (i.e., the amount of time that elapses between consecutive time steps is the same) but in general, DCR models can handle irregularly sampled time series.

In terms of ground truth information, for each training patient $i \in [n]$, we assume that we observe two quantities:

- (Event indicator) We observe which of k different competing events happened first to the i-th patient, or alternatively we could also observe that none of the competing events happened by the time training data collection stopped. This information is stored in the event indicator $K_i \in \{0, 1, 2, \ldots, k\}$. For example, in the neurological prognostication problem, we have $k = 3$ and the competing events are awakening ($K_i = 1$), dying of causes aside from withdrawal from life-sustaining therapies ($K_i = 2$), and withdrawal from life-sustaining therapies ($K_i = 3$). The special value of $K_i = 0$ means that by the time data collection stopped, none of the competing events happened.
- (Event time) We also observe the time $Y_i \in \mathbb{R}$ for when the first competing event happened or, if none of them happened, then Y_i is the time when data collection stopped for the i-th patient (i.e., the time of "censoring"). Note that at any time step $\ell \in [L_i]$, the time until the first competing event or censoring happens is $Y_i - T_i^{(\ell)}$.

In summary, for training patient $i \in [n]$, we observe an event indicator $K_i \in \{0, 1, \ldots, k\}$, an event time $T_i \in \mathbb{R}$, and an input time series of feature vectors $X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(L_i)} \in \mathbb{R}^d$ at corresponding times $T_i^{(1)}, T_i^{(2)}, \ldots, T_i^{(L_i)} \in \mathbb{R}$. As shorthand notation for referring to the entire observed time series, we write $Z_i := \big((X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(L_i)}), (T_i^{(1)}, T_i^{(2)}, \ldots, T_i^{(L_i)})\big)$.

We model the training data $(Z_1, K_1, Y_1), \ldots, (Z_n, K_n, Y_n)$ to be i.i.d. To formally state how each training point is generated, we define a few probability distributions: $\mathbb{Q}_{\mathcal{T}}$ denotes an underlying probability distribution over variable-length time series, $\mathbb{Q}_{\mathcal{E}}(Z)$ denotes an underlying conditional probability distribution over nonnegative time durations across all k competing events given a specific time series Z (i.e., a random sample from $\mathbb{Q}_{\mathcal{E}}(Z)$ yields a vector in $\mathbb{R}^k$ where the j-th entry of the vector is a random time duration until competing event $j \in [k]$ happens), and $\mathbb{Q}_{\mathcal{C}}(Z)$ denotes the underlying conditional probability

distribution over nonnegative time durations until censoring. Specifically, the i-th training point (Z_i, K_i, Y_i) is generated as follows:

1. We sample time series Z_i (with L_i time steps and last time $T_i^{(L_i)}$ using our earlier notation) from $\mathbb{Q}_{\mathcal{T}}$.
2. We sample the true nonnegative time durations $(\Xi_{i,1}, \Xi_{i,2}, \ldots, \Xi_{i,k})$ from $\mathcal{Q}_{\mathcal{E}}(Z_i)$ (so that $\Xi_{i,1}$ is the time until the 1st competing event happens, $\Xi_{i,2}$ is the time until the 2nd competing event happens, etc).
3. We sample the true nonnegative time duration $\Xi_{i,0}$ (time until censoring) from a conditional distribution $\mathbb{Q}_{\mathcal{C}}(Z_i)$.
4. We set $K_i := \arg\min_{j=0,1,\ldots,k} \Xi_{i,j}$, and $Y_i := T_i^{(L_i)} + \Xi_{i,K_i}$.

Note that the competing events are "exhaustive" in the sense that with probability 1, either one of them happens or censoring happens.

Prediction target We model a test patient's data using the same distributions that we introduced for training data. However, for the test patient, our goal will never be to predict whether the test patient is censored. In particular, we model a test patient with time series Z, event indicator K (always a value in $\{1, \ldots, k\}$), and event time Y as follows:

1. We sample time series $Z = \big((X^{(1)}, \ldots, X^{(L)}), (T^{(1)}, \ldots, T^{(L)})\big)$ (using notation similar to that of training data) from $\mathbb{Q}_{\mathcal{T}}$. Note that Z has L time steps.
2. We sample the true nonnegative time durations $(\Xi_1, \Xi_2, \ldots, \Xi_k)$ from $\mathbb{Q}_{\mathcal{E}}(Z)$.
3. We set $K := \arg\min_{j=1,\ldots,k} \Xi_j$, and $Y := T^{(L)} + \Xi_K$.

Even though time series Z is generated so that it has L time steps, in how we set up the prediction task next, we do not observe all time steps immediately. Instead, we progressively observe more of Z over time, similar to what would happen in a real clinical context. In particular, we state our prediction task to depend on time $t \in \mathbb{R}$ and use the random variable $Z^{(\leq t)}$ to denote time series Z limited to information up until time t. Specifically, we aim to predict the so-called *cumulative incidence function* (CIF) of event $j \in [k]$, which is the probability of event j happening within time duration $\Delta \geq 0$ starting from time $t \in \mathbb{R}$, given a time series observed up until time t. Formally, we write the CIF as

$$F_j(\Delta \mid z, t) := \mathbb{P}(Y \leq t + \Delta, K = j \mid Z^{(\leq t)} = z^{(\leq t)}, Y > t) \qquad \text{for } \Delta \geq 0, \tag{1}$$

where t is the time that we are making a prediction at, and z is any specific realization of random variable Z (again, the superscript "$^{(\leq t)}$" restricts time to be up until t).

Note that we have intentionally stated the CIF in the dynamic setting with variable-length time series, where we can make predictions at different points in time. The classical version of the CIF (Gray, 1988; Fine and Gray, 1999) is stated in the "static" setting without time series and can be viewed as a special case of the dynamic setup we have described, where all time series sampled from $\mathbb{Q}_{\mathcal{T}}$ have exactly one time step, the recorded time of this first time step is always just taken to be 0, and we only ever evaluate equation (1) at $t = 0$. Separately, if the number of competing risks is equal to $k = 1$, then the entire setup we have described would instead be for dynamic survival analysis without competing risks. In fact, one could show that the static setting with one competing risk simply reduces to the classical right-censored survival analysis setup (e.g., see the random censoring setup described in Section 3.2 of the textbook by Kalbfleisch and Prentice (2002)).

Dynamic competing risks (DCR) models The class of DCR models that our framework for neurological prognostication builds on is any model that can predict CIFs as given in equation (1). For example, Dynamic-DeepHit (Lee et al., 2019) and SurvLatent ODE (Moon et al., 2022) are DCR models. Note that any classical competing risks model (e.g., Fine and Gray 1999) that does not actually handle variable-length time series could be made into a DCR model in a simple manner: simply only use the last time step's feature vector to predict. Some existing dynamic survival analysis models (such as DDRSA (Venkata and Bhattacharyya, 2022)) that were not originally developed to support competing events could be modified to estimate CIFs as well. To give a sense of how a DCR model works, we review Dynamic-DeepHit next. Note that we specifically review a DCR model that directly models variable-length time series without manual feature engineering or resorting to, for instance, only ever using the last time step of an input time series.

2.2. Example of a DCR Model: Dynamic-DeepHit

We provide an overview of Dynamic-DeepHit (Lee et al., 2019), deferring details to the original paper.[2] Importantly, Dynamic-DeepHit discretizes possible values for time duration Δ in equation (1) into m unique values $\Delta_1 < \Delta_2 < \cdots < \Delta_m$. We assume that $\Delta_1 > 0$ and that Δ_m is an upper bound on possible durations encountered across all events $j \in [k]$ (i.e., all k events happen within time duration Δ_m). In what follows, we regularly use the variables $u, v \in [m]$ to denote indices of the discretized values of Δ. Dynamic-DeepHit estimates a probability mass function variant of the CIF for event $j \in [k]$ given by

$$O_j(\Delta_u \mid z, t) := \mathbb{P}(\Delta_{u-1} < Y - t \leq \Delta_u, K = j \mid Z^{(\leq t)} = z^{(\leq t)}, Y > t) \quad \text{for } u \in [m], \tag{2}$$

where we define $\Delta_0 := 0$ to handle the case when we plug in $u = 1$. Before explaining why the above function behaves like a probability mass function, we point out that one can readily verify that the CIF for event j from equation (1) satisfies the equality

$$F_j(\Delta_u \mid z, t) = \sum_{v=1}^{u} O_j(\Delta_v \mid z, t) \qquad \text{for } u \in [m]. \tag{3}$$

Thus, so long as we can estimate $O_j(\Delta_u \mid z, t)$ in equation (2), then we obtain an estimate of the CIF in equation (1) albeit only along a discrete time grid $\Delta \in \{\Delta_1, \ldots, \Delta_m\}$.

As for why $O_j(\Delta_u \mid z, t)$ behaves like a probability mass function, note that

$$\sum_{j=1}^{k} \sum_{u=1}^{m} O_j(\Delta_u \mid z, t) = \sum_{j=1}^{k} F_j(\Delta_m \mid z, t) = 1, \tag{4}$$

where we have used equation (3) and the assumption that Δ_m is chosen as an upper bound on possible durations across all k competing events.

With this motivation, Dynamic-DeepHit models $O_j(\Delta_u \mid z, t)$ in equation (2) using a neural network. For the i-th training time series $Z_i := \big((X_i^{(1)}, \ldots, X_i^{(L_i)}), (T_i^{(1)}, \ldots, T_i^{(L_i)})\big)$, we specifically estimate $O_j(\Delta_u \mid Z_i, T_i^{(L_i)})$ to be equal to $O_{i,j,u}$ (with $i \in [n], j \in [k], u \in [m]$),

2. Note that Lee et al. (2019) explicitly keep track of a separate vector per time step indicating which of the d features are missing. Instead of introducing notation for such a "missingness" boolean vector, we can augment our original feature vector to include such missingness indicator variables.

where $O_{i,j,u}$ is shown on the right side of Figure 1; we collect all the $O_{i,j,u}$ values specific to the i-th patient in the vector $O_i \in \mathbb{R}^{k \cdot m}$. We compute O_i from Z_i as follows:

1. We first feed the input time series Z_i into a user-specified RNN (with p output features per time step, where the choice of p is up to the user), where we slightly transform what the input looks like per time step. Specifically at time step $\ell \in [L_i]$, the input to the RNN is taken to be $(X_i^{(\ell)}, T_i^{(\ell+1)} - T_i^{(\ell)})$, i.e., we provide both a feature vector and a time duration to get to the next time step. However, the last time step's RNN input is taken to be $(X_i^{(L_i)}, 0)$. The RNN's output at time step $\ell \in [L_i]$ is denoted as $H_i^{(\ell)} \in \mathbb{R}^p$. This first step is shown on the left side of Figure 1.
2. The next step is to summarize the variable-length time series $(H_i^{(1)}, \ldots, H_i^{(L_i)})$ into a fixed-length vector $C_i \in \mathbb{R}^p$. To do this, we set $C_i = \sum_{\ell=1}^{L_i} a_\ell H_i^{(\ell)}$, where
$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{L_i} \end{bmatrix} := \text{softmax}\left(\begin{bmatrix} f_{\text{attention}}((H_i^{(1)}, X_i^{(L_i)})) \\ f_{\text{attention}}((H_i^{(2)}, X_i^{(L_i)})) \\ \vdots \\ f_{\text{attention}}((H_i^{(L_i)}, X_i^{(L_i)})) \end{bmatrix} \right) \in [0,1]^{L_i},$$
and $f_{\text{attention}}$ is a user-specified feed-forward neural network, such as a multilayer perceptron (MLP), that outputs a single real number. This second step is shown in the middle of Figure 1.
3. We use vector C_i as an input to k different MLPs (one per competing risk) that each outputs m numbers, and the overall output is passed through a softmax layer to produce the final output O_i (the softmax enforces the constraint in equation (4)). Note that this last step corresponds to the original DeepHit model (Lee et al., 2018) that is meant for handling input data that are fixed-length feature vectors rather than variable-length time series. This third step is shown on the right side of Figure 1.

There is one last neural network component that is not shown in Figure 1 as it is not used to compute the output vector O_i: Dynamic-DeepHit also requires that at time step $\ell \in [L_i]$, the RNN on the left side of Figure 1 can output an estimate $\widehat{X}_i^{(\ell+1)}$ of the next time step's feature vector $X_i^{(\ell+1)}$. There are different ways to achieve this. For example, at time step $\ell \in [L_i]$, we can feed $H_i^{(\ell)}$ (along with the time duration to get to the next time step) into a user-specified MLP $f_{\text{next-time-step}}$ with d output features to produce the estimate $\widehat{X}_i^{(\ell+1)}$. Alternatively, for the RNN in Figure 1, we could choose it to be a type of RNN that already distinguishes between hidden state vectors and output state vectors (e.g., LSTMs (Hochreiter and Schmidhuber, 1997)), in which case we let the hidden state vectors be what we denoted as the $H_i^{(\ell)}$ variables, and we use the output state vectors to predict the next steps' feature vectors (the output state vector would need to consist of d entries).

Training The final loss used to train a Dynamic-DeepHit model is the sum of three terms, two of which make up the original DeepHit loss (a negative log likelihood term and a ranking loss term), and the last term asks that each next feature vector estimate $\widehat{X}_i^{(\ell+1)}$ is close to $X_i^{(\ell+1)}$ (using, for instance, squared Euclidean distance).

Prediction At test time, we could feed any observed test time series (of arbitrarily nonzero length) as input to the neural network in Figure 1 to produce an estimate of the probability

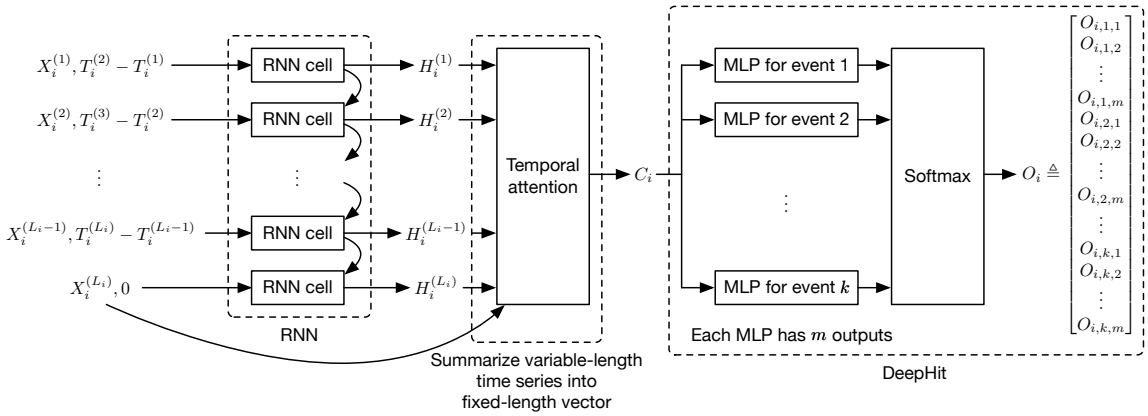


Figure 1: Dynamic-DeepHit network architecture.

mass function of the CIF in equation (2) that we can then use to estimate the CIF with using equation (3). We could trivially accommodate the setting where we see more of a test time series over time since the neural network accepts a variable-length input time series.

3. Framework for Neurological Prognostication

Any DCR model could be applied to the problem of neurological prognostication for post-cardiac-arrest coma patients. As stated in Section 2, we can take the number of competing risks to be $k = 3$ corresponding to awakening ($K = 1$), dying (not of withdrawal from life-sustaining therapies) ($K = 2$), or withdrawal from life-sustaining therapies (and thus dying as a result) ($K = 3$). A key goal of our framework is to help clinicians interpret the information contained in the CIFs (equation (1)) predicted by a DCR model.

Using the three competing risks stated above, we derive a binary probabilistic classifier from an already trained DCR model (Section 3.1). The resulting classifier can then be used to produce a patient-specific prediction heat map visualization that aims to be straightforward for a clinician to interpret (Section 3.2). Separately, standard binary classification evaluation metrics could be used for the derived classifier, which supplement survival analysis evaluation metrics that already exist for DCR models.

3.1. A Derived Binary Classifier

As discussed in Section 1, predicting whether a patient will be withdrawn from life-sustaining therapies is often not of primary interest since this is a human-made decision that deterministically ends a patient's life, and we do not actually know for sure whether the patient would have instead awakened or died of other causes in the ICU. For any test patient's time series up to time t, we now derive a binary probabilistic classifier that conditions on the event that *the test patient is never withdrawn from life-sustaining therapies* (we refer to this event as the "non-withdrawal event"). As we aim to develop a decision support tool to help physicians decide on whether to withdraw life-sustaining therapies, a reasonable assumption is that the patients are kept on these therapies, especially since withdrawal of these therapies has a permanent effect (in Section 6, we comment on whether this conditioning makes sense).

After conditioning on the non-withdrawal event, we then compute the probabilities of the remaining two competing events happening within a time duration $\Delta > 0$. This conditional

probability can be computed as follows, reusing notation from Section 2:

$$\begin{aligned} P_{\text{awaken}}(\Delta \mid z, t) &:= \mathbb{P}(Y \leq t + \Delta, \overbrace{K = 1}^{\text{awaken}} \mid \overbrace{K \in \{1, 2\}}^{\substack{\text{awaken or death (not} \\ \text{by withdrawal from} \\ \text{life-sustaining therapies)}}}, Z^{(\leq t)} = z^{(\leq t)}, Y > t) \\ &= \frac{\mathbb{P}(Y \leq t + \Delta, K = 1 \mid Z^{(\leq t)} = z^{(\leq t)}, Y > t)}{\mathbb{P}(K \in \{1, 2\} \mid Z^{(\leq t)} = z^{(\leq t)}, Y > t)} \\ &= \frac{\mathbb{P}(Y \leq t + \Delta, K = 1 \mid Z^{(\leq t)} = z^{(\leq t)}, Y > t)}{\mathbb{P}(K = 1 \mid Z^{(\leq t)} = z^{(\leq t)}, Y > t) + \mathbb{P}(K = 2 \mid Z^{(\leq t)} = z^{(\leq t)}, Y > t)} \\ &= \frac{F_1(\Delta \mid z, t)}{F_1(\infty \mid z, t) + F_2(\infty \mid z, t)}. \end{aligned} \tag{5}$$

Thus, if we have trained a DCR model that has an estimate $\widehat{F}_j(\Delta \mid z, t)$ for each CIF $F_j(\Delta \mid z, t)$, then we can directly plug in these CIF estimates into the right-hand side above to yield the estimated conditional probability

$$\widehat{P}_{\text{awaken}}(\Delta \mid z, t) := \frac{\widehat{F}_1(\Delta \mid z, t)}{\widehat{F}_1(\infty \mid z, t) + \widehat{F}_2(\infty \mid z, t)}. \tag{6}$$

We could similarly estimate the probability for death (not by withdrawal from life-sustaining therapies) by

$$\widehat{P}_{\text{death (not withdrawal)}}(\Delta \mid z, t) := \frac{\widehat{F}_2(\Delta \mid z, t)}{\widehat{F}_1(\infty \mid z, t) + \widehat{F}_2(\infty \mid z, t)}. \tag{7}$$

We thus have a binary probabilistic classifier between the two classes "awaken" and "death (not by withdrawal from life-sustaining therapies)" defined for a specific time t and time duration Δ: if the ratio $\frac{\widehat{P}_{\text{death (not withdrawal)}}(\Delta|z,t)}{\widehat{P}_{\text{awaken}}(\Delta|z,t)}$ is below a threshold value of 1, then we predict "awaken". The threshold of 1 could of course be tuned (e.g., to achieve some desired tradeoff between TPR and FPR on some validation set). Note that it only makes sense to plug in times t that are at least the earliest time encountered in time series z.

Importantly, the binary classifier we just described was derived using estimated CIFs. Existing DCR models like Dynamic-DeepHit estimate CIFs in a manner that would include patients of all three competing risks as well as those who were censored. In particular, we do not have to, for example, exclude patients who are censored or who were withdrawn from life-sustaining therapies from the analysis. Moreover, this classifier can be constructed for any time t after the earliest time in the observed test time series $Z = z$ and for any choice of user-specified time duration $\Delta > 0$, without any re-training of the underlying DCR model.

Technical remark The estimated probabilities $\widehat{P}_{\text{awaken}}(\Delta \mid z, t)$ in equation (6) and $\widehat{P}_{\text{death (not withdrawal)}}(\Delta \mid z, t)$ in equation (7) do not, in general, sum to 1. This is intentional. Again, each probability is derived based on equation (5), where the final denominator is the probability that a patient with time series z up to time t experiences an eventual outcome that is either $K = 1$ or $K = 2$. From an interpretation standpoint, an appealing aspect of how $\widehat{P}_{\text{awaken}}(\Delta \mid z, t)$ (and similarly $\widehat{P}_{\text{death (not withdrawal)}}(\Delta \mid z, t)$) is defined is as follows. For a fixed time t, consider two time durations Δ and Δ', where $\Delta < \Delta'$ (for example, $\Delta = 24$ hours and $\Delta' = 48$ hours). Then intuitively it makes sense that the probability

of someone awakening within time duration Δ' should be larger than the probability of someone awakening within the time duration Δ, since $\Delta' > \Delta$. In other words, we would like $\widehat{P}_{\text{awaken}}(\Delta \mid z, t) \leq \widehat{P}_{\text{awaken}}(\Delta' \mid z, t)$. This property indeed holds for how we have defined equations (6) (and a similar result holds for $\widehat{P}_{\text{death (not withdrawal)}}$). This property would *not* be guaranteed to hold if instead we had changed the denominators of equations (6) and (7) to $\widehat{F}_1(\Delta \mid z, t) + \widehat{F}_2(\Delta \mid z, t)$, which would ensure that the probabilities sum to 1.

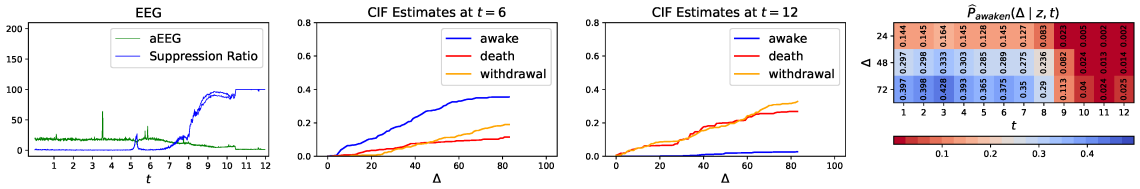
3.2. Patient-Specific Heat Map Visualization

We propose a heat map visualization specific to any patient that shows how the predicted probability of awakening ($\widehat{P}_{\text{awaken}}(\Delta \mid z, t)$ in equation (6)) changes for the patient as we observe more of the patient's time series, as shown in the fourth column plot of each row in Figure 2. The experimental setup that led to this figure is explained in more detail in Section 5. For now, focusing on any one of the fourth-column heat maps of Figure 2, we have the horizontal axis correspond to the prediction time t while the vertical axis is the time duration Δ. We specifically choose the Δ's to be equivalent of the next 24 to 72 hours, which is of interest according to clinician feedback we have received. Using Patient 1 in Figure 2 as an example, we observe a drastic decrease in the probability of awakening starting at $t = 9$ across all Δ values evaluated.

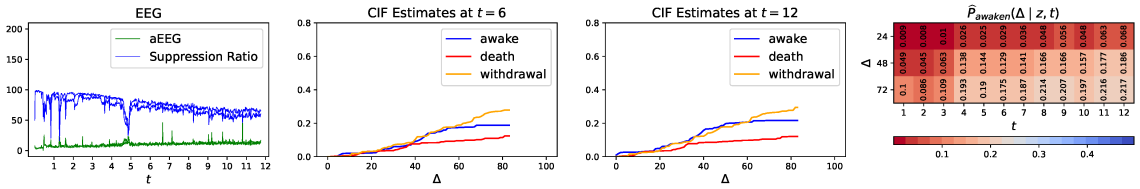
4. Cohort

We examine proprietary hospital data collected in a single medical center from 2010 to 2019 (the specific medical center has been blinded for reviewing purposes). The dataset includes patients who suffered sudden cardiac arrest, were successfully resuscitated, and survived to hospital admission at the medical center. EEG is initiated and monitored for these patients continuously as a routine standard of care for several days after cardiac arrest. As stated in the previous sections, we focus on three outcomes: awakening from the coma, dying (not from withdrawal of life-sustaining therapies, such as from brain death or rearrest) and withdrawal from life-sustaining therapies (due to perceived poor neurological prognosis, leading to death). We remark that there is a fourth outcome that is possible but that we exclude from analysis: patients could be withdrawn from life-sustaining therapies for non-neurological reasons such as a do-not-resuscitate order. As our focus is on neurological prognostication, we ignore this outcome in which a decision is made disregarding neurological prognosis. Note that in this study, we only care about the first occurrence of the event that ceases the coma status of a patient, i.e. if a patient awakened at some point and then still died shortly afterward, we would only consider awakening as the outcome of the patient.

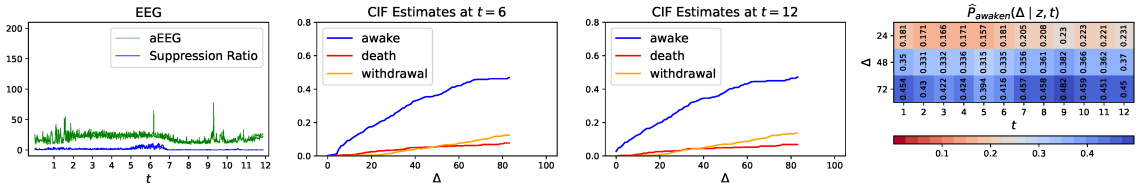
Dataset characteristics After excluding patients whose cause of death is withdrawal for non-neurological reasons, we have a dataset consisting of 922 patients. For the 922 patients, summary statistics of their characteristics and the time-to-event are shown in an appendix (Table A.1). Out of the 922 patients, 271 (29.4%) of them awakened at some point, 189 (20.5%) of them died (not from withdrawal from life-sustaining therapies), 432 (49.6%) of them died from withdrawal of life-sustaining therapies due to poor perceived neurological prognosis, and 30 (3.3%) of them were still in a coma when data collection stopped.



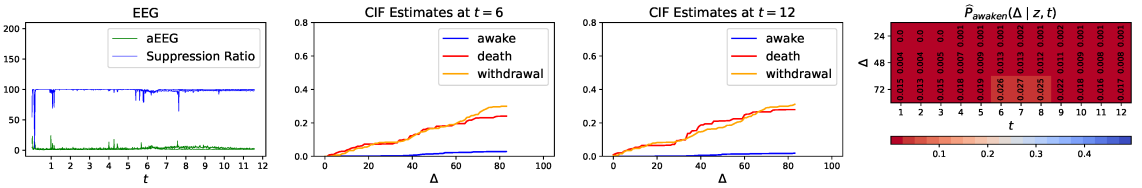
(a) Example Patient 1: died (not from withdrawal of life-sustaining therapies) at hour 17



(b) Example Patient 2: was still in a coma at hour 118



(c) Example Patient 3: awakened at hour 54



(d) Example Patient 4: died (not from withdrawal of life-sustaining therapies) at hour 71

Figure 2: For each example patient (panels (a)-(d)), we show time series of two summary EEG features *aEEG* and *suppression ratio* (first column plot), estimated CIFs at hours $t = 6$ (second column plot) and $t = 12$ (third column plot), and our proposed heat map visualization (fourth column plot). Note that aEEG values for normal brain activity should be within a certain range (constantly being lower than 5 or higher than 25 is usually considered abnormal), and higher suppression ratio values are a sign of more severe dysfunction or injury.

EEG time series data Raw EEG waveform data, typically recorded at 256Hz from 22 electrodes distributed on the brain according to standard clinical practice, are processed using FDA-approved clinical software (https://www.persyst.com/) to quantify more than 2,500 clinically-understood features every second (e.g., amplitude, frequency decompositions, suppression ratio, etc). For ease of exposition, here we focus on 12 features (see Appendix A for details) that have been proven to be useful in this domain: *suppression ratio* and *amplitude-integrated EEG (aEEG)* (Oh et al., 2015; Elmer et al., 2016a). We preprocess and downsample the raw second-by-second data to have one measurement (per feature) per hour in two steps: (1) For each of the 12 EEG features used in this study, we downsample the data by taking the average value of each consecutive non-overlapping block of 60 seconds to get a

single value (i.e., we first downsample the data so that each time step corresponds to one minute). (2) We then further downsample the minute-resolution EEG signals so that each time step corresponds to an hour, and for each of the 12 EEG features, we take 6 summary statistics (minimum, maximum, mean, 25% percentile, median, and 75% percentile) as the final features to be used (i.e., for each time step corresponding to 1 hour, we end up with a total of $12 \times 6 = 72$ summary features). We remark that the data after downsampling still captures most of the variation from the raw data and does not have much impact on prediction accuracy. Downsampling is mainly to reduce computation time.

Note that in how we curated the data, we only use at most 12 hours of EEG data per patient (so that if a patient has more than 12 hours worth of EEG data, we ignore the EEG data after 12 hours). This is a limitation of the dataset curation process that could be changed in future work. From a modeling perspective, DCR models could in principal use arbitrarily long EEG time series, subject to hardware memory constraints in practice.

Static features Upon hospital admission, a number of features are collected for the patient that we treat as static features, such as demographic information (e.g. age, gender), characteristics of the patient's initial cardiac arrest collapse (e.g., arrest location, initial arrest rhythm, category of the cardiac arrest), initial coma status, and medical history. A full list of the 43 static features we use can be found in Appendix A.

In summary, accounting for both the time-varying EEG features and the static features, we use a total of $d = 72 + 43 = 115$ final features per time step when we train a DCR model.

5. Experiments

In this section, we run experiments on the dataset described in the previous section using three DCR models: a classical competing risks model by Fine and Gray (1999) using only the last observed time step of each input time series (details are in Appendix B), Dynamic-DeepHit (Lee et al., 2019) and DDRSA (Venkata and Bhattacharyya, 2022). Also, note that the original DDRSA model by Venkata and Bhattacharyya (2022) does not support competing risks and we modify its network structure to accommodate competing risks (details are in Appendix C). For simplicity, we refer to the competing-risk-adapted version of DDRSA as DDRSA despite the modification we make. We specifically aim to:

- (Section 5.1) examine how accurate these models using the standard survival analysis metric of concordance index (abbreviated c-index; this is a value between 0 and 1 where higher is better) (Harrell et al., 1982) as well as AUROC of binary classifiers derived using our approach in Section 3.1,
- (Section 5.2) provide examples of patient-specific visualizations (time series of two summary EEG signals, estimated CIFs, and the heat map visualization we described in Section 3.2)
- (Section 5.3) show that using a setup with fewer than the three competing risks we consider result in models that are not as good.

Experimental setup We repeat the following basic experiment five times with different random training/validation/test splits of the data. For each experimental repeat, we randomly select 80% of the 922 patients to be in the training set and the remaining 20% in the testing set. For Dynamic-DeepHit and DDRSA, within the training set, a random

Table 1: Test set c-indices (average ± standard deviation across five experimental repeats) for three DCR models. The entries with bold values represent the highest average c-index for each (event, t, Δ) combination.

Model	Prediction time	Event	Evaluation time horizon		
			$\Delta = 24$ hrs	$\Delta = 48$ hrs	$\Delta = 72$ hrs
Fine and Gray	$t = 6$	awakening	**0.853 ± 0.017**	**0.874 ± 0.012**	**0.875 ± 0.012**
		death	0.633 ± 0.171	0.673 ± 0.081	0.684 ± 0.061
		withdrawal	0.691 ± 0.063	**0.634 ± 0.050**	**0.652 ± 0.039**
	$t = 12$	awakening	0.831 ± 0.032	0.851 ± 0.023	0.854 ± 0.023
		death	**0.751 ± 0.110**	**0.709 ± 0.040**	0.713 ± 0.044
		withdrawal	0.709 ± 0.082	**0.675 ± 0.035**	**0.681 ± 0.024**
Dynamic-DeepHit	$t = 6$	awakening	0.851±0.018	0.867±0.012	0.864±0.017
		death	**0.702±0.080**	**0.684±0.096**	**0.697±0.071**
		withdrawal	**0.742±0.103**	0.612±0.062	0.621±0.031
	$t = 12$	awakening	**0.847±0.038**	**0.859±0.020**	**0.858±0.024**
		death	0.701±0.078	0.699±0.042	**0.722±0.044**
		withdrawal	**0.739±0.076**	0.640±0.028	0.666±0.017
DDRSA	$t = 6$	awakening	0.821±0.032	0.845±0.028	0.836±0.030
		death	0.599±0.072	0.615±0.059	0.633±0.075
		withdrawal	0.677±0.151	0.626±0.121	0.637±0.100
	$t = 12$	awakening	0.825±0.021	0.818±0.025	0.798±0.027
		death	0.681±0.095	0.651±0.086	0.651±0.074
		withdrawal	0.663±0.126	0.652±0.092	0.670±0.068

20% of the training points are held out as a validation set for hyperparameter tuning. The random splits are stratified as the preserve the fraction of data experiencing each event indicator value $\{0, 1, \ldots, k\}$. The hyperparameter grid we use is given in Appendix D. Note that the Fine and Gray model has no hyperparameters.

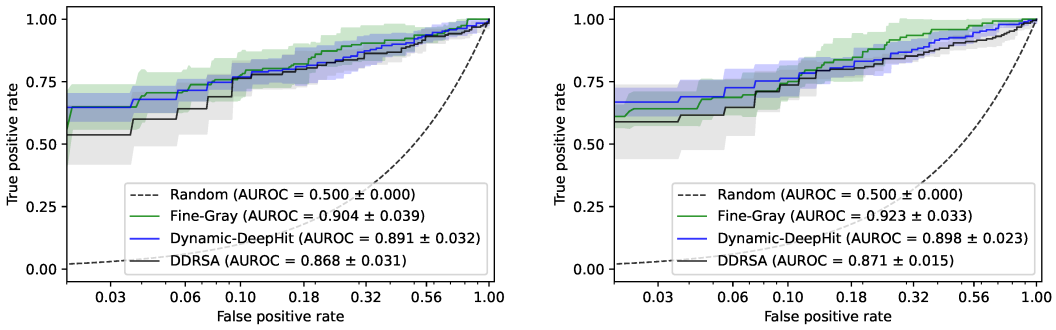
5.1. Accuracy Benchmark in the Dynamic Problem Setup

Survival analysis accuracy metric To evaluate the accuracy of different competing risks models, we compute c-indices per competing risk at different prediction times $t = 6, 12$ and different time horizons $\Delta = 24, 48, 72$; these are all in units of hours. For example, with the prediction time $t = 6$ and evaluation time $\Delta = 24$, it means at hour 6 after ICU admission, we are comparing the model's predicted risk of different events occurring in the next 24 hours. The results for the models are shown in Table 1. From this table, we see that in terms of c-indices per event, no model is uniformly the best across all prediction times t and time durations Δ, while DDRSA appears to have slightly lower accuracy scores compared to the other two models.

Binary classification accuracy metric By using the binary classifier derived from a DCR model as described in Section 3.1, we can use binary classification evaluation metrics such as the area under the ROC curve (AUROC). After restricting the test set cohort to patients that we either observe awakening or death (not from withdrawal from life-sustaining therapies), we can compute the AUROC scores shown in Table 2. Here, note that the Fine

Table 2: Test set AUROC (average ± standard deviation across five experimental repeats) for three DCR models. The entries with bold values represent the highest average AUROC for each (t, Δ) combination.

Model	Prediction time	Evaluation time horizon		
		$\Delta = 24$ hrs	$\Delta = 48$ hrs	$\Delta = 72$ hrs
Fine and Gray	$t = 6$	**0.904 ± 0.039**	**0.903 ± 0.039**	**0.903 ± 0.039**
	$t = 12$	**0.923 ± 0.033**	**0.921 ± 0.034**	**0.920 ± 0.034**
Dynamic-DeepHit	$t = 6$	0.891 ± 0.032	0.883 ± 0.030	0.885 ± 0.029
	$t = 12$	0.898 ± 0.023	0.891 ± 0.029	0.899 ± 0.025
DDRSA	$t = 6$	0.868 ± 0.031	0.863 ± 0.028	0.849 ± 0.031
	$t = 12$	0.871 ± 0.015	0.853 ± 0.019	0.822 ± 0.020



(a) $t = 6, \Delta = 24$ (b) $t = 12, \Delta = 24$

Figure 3: Test set ROC curve (average curve ± standard deviation intervals across five experimental repeats). The x-axis is on a log scale to emphasize the low FPR regime.

and Gray model achieves mean AUROC scores that are higher than those of Dynamic-DeepHit across all values of t and Δ evaluated. Accounting for the standard deviations of the AUROC scores, the two models do have AUROCs that are quite close. Again, DDRSA seems to perform worse than the other two models in terms of AUROCs.

We can also plot the ROC curve at different t and Δ values. Specifically, we show the ROC at $t = 6$ and $t = 12$ with the estimated ratio of the probability of awakening and probability of death (not from withdrawal) within the next $\Delta = 24$ hours in Figure 3, where we plot the x-axis on a log scale to focus on the low FPR regime. For example, we can see that at $t = 12, \Delta = 24$, Dynamic-DeepHit reaches an average AUROC of 0.898 and an average TPR of 0.668 with a small FPR of 0.020. The ROC curves for a few other t and Δ values can be found in Appendix E.

5.2. Patient-Specific Visualizations

After we train a DCR model, we can easily derive the estimated CIF for different events at different prediction times (e.g., $t = 6, 12$), and the estimated conditional probability of awakening using equation (6). We focus on using the trained Dynamic-DeepHit to derive all the visualizations in this part as an illustrative example (the same sorts of visualizations could be made for the Fine and Gray model and DDRSA). In Figure 2, we display two

summary EEG signals for four example patients, the estimated CIFs, and the heat map visualization we described in Section 3.2; each row/panel in the figure corresponds to one patient. Each entry in the heatmap is the estimated conditional probability of awakening occurring at different times ($t = 1, 2, \ldots, 12$) for the different durations ($\Delta = 24, 48, 72$). For the four patients shown:

- (Figure 2(a)) Patient 1 died at hour 17 (not from withdrawal of life-sustaining therapies). We observe a drastic change in the patient's EEG signals before and after $t = 6$, which is reflected in the two sets of estimated CIFs at $t = 6$ and $t = 12$ with a higher estimated CIF of the "awakening" event given the first 6 hours of EEG data but a lower "awakening" CIF curve after the first 12 hours of EEG.
- (Figure 2(b)) Patient 2 is still in a coma at hour 118. From the EEG signals of this patient, we can see that the probability of awakening is increasing gradually in the heat map (fourth column plot) as we go from $t = 1$ to $t = 12$.
- (Figure 2(c)) Patient 3 awakened at hour 54, and we do observe good EEG signals (we briefly describe some common patterns considered "good" or "bad" in the caption of Figure 2). In the heat map visualization, the predicted probability of awakening is high at all times.
- (Figure 2(d)) Patient 4 died (not from withdrawal of life-sustaining therapies) at hour 71, and we did observe very poor EEG signals over the entire 12-hour period. In the heat map, the predicted probability of awakening is low at all entries.

Note that the estimated CIF for withdrawal from life-sustaining therapies could be viewed as the model's prediction of how likely a physician (at least according to historical data) would make a decision to withdraw said therapies.

5.3. Ablation Study

We conduct an ablation study to show why including the three competing risks in how we framed the problem is better than had we used fewer competing risks. In particular, we repeat the same experiments as above but with only two competing risks (awakening and dying not by withdrawal from life-sustaining therapies), and a single event (awakening). For the purpose of this ablation study, we only focus on the Fine and Gray model and Dynamic-DeepHit as they achieved noticeably higher accuracy than DDRSA in our earlier experiments.

In the case when we only consider two competing risks, a patient with the outcome label of withdrawal from life-sustaining therapies would be viewed as being censored (along with those who stayed in a coma). Similarly, when we only model a single competing risk, patients with all other outcomes are viewed as being censored. The resulting c-indices are shown for the two-competing-risk case in Table 3 and for the one competing risk case in Table 4.

By comparing Tables 1 and 3, we see that c-indices for death (not by withdrawal of life-sustaining therapies) stay at a similar level for the Fine and Gray model considering the standard deviation interval when we model only two competing risks, while those of Dynamic-DeepHit are clearly higher when we model all three competing risks. While we do not observe a gain in the c-indices for the event of awakening when we move from the single risk setting to those of two or three competing risks, by incorporating more events, we are

Table 3: Test set c-indices (average ± standard deviation across five experimental repeats) for the Fine and Gray model and Dynamic-DeepHit with two competing events. The entries with bold values represent the highest average c-index for each (event, t, Δ) combination.

Model	Prediction time	Event	Evaluation time horizon		
			$\Delta = 24$ hrs	$\Delta = 48$ hrs	$\Delta = 72$ hrs
Fine and Gray	$t = 6$	awakening	0.835 ± 0.030	**0.870 ± 0.012**	**0.867 ± 0.008**
		death	**0.723 ± 0.040**	**0.697 ± 0.075**	**0.707 ± 0.061**
	$t = 12$	awakening	0.822 ± 0.020	0.855 ± 0.014	0.851 ± 0.014
		death	**0.664 ± 0.272**	**0.678 ± 0.146**	**0.706 ± 0.101**
Dynamic-DeepHit	$t = 6$	awakening	**0.852 ± 0.019**	0.858 ± 0.011	0.855 ± 0.007
		death	0.638 ± 0.074	0.622 ± 0.073	0.642 ± 0.050
	$t = 12$	awakening	**0.842 ± 0.017**	**0.860 ± 0.007**	**0.857 ± 0.007**
		death	0.599 ± 0.092	0.639 ± 0.095	0.649 ± 0.078

Table 4: Test set c-indices (average ± standard deviation across five experimental repeats) for the Fine and Gray model and Dynamic-DeepHit with a single event. The entries with bold values represent the highest average c-index for each (event, t, Δ) combination.

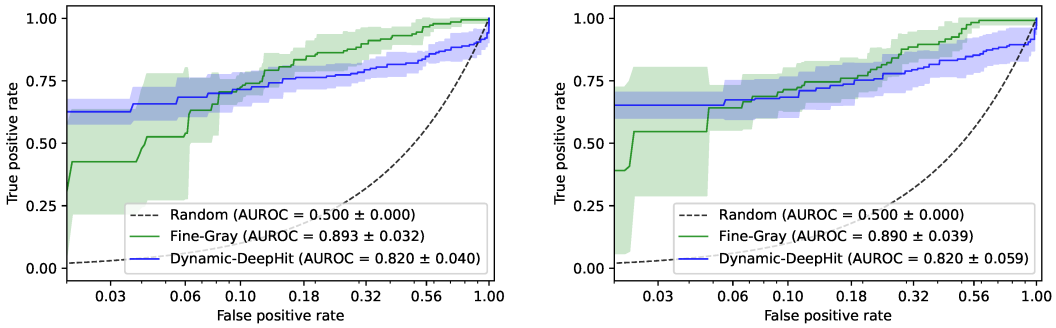
Model	Prediction time	Event	Evaluation time horizon		
			$\Delta = 24$ hrs	$\Delta = 48$ hrs	$\Delta = 72$ hrs
Fine and Gray	$t = 6$	awakening	**0.864 ± 0.021**	**0.883 ± 0.013**	**0.877 ± 0.010**
	$t = 12$	awakening	0.843 ± 0.023	**0.866 ± 0.018**	**0.861 ± 0.015**
Dynamic-DeepHit	$t = 6$	awakening	0.850 ± 0.031	0.861 ± 0.019	0.855 ± 0.021
	$t = 12$	awakening	**0.845 ± 0.021**	0.855 ± 0.014	0.855 ± 0.018

able to capture more information without decreasing the model's accuracy. For example, if we only trained the single competing risk model and if the patient is not likely to awaken from the coma, the model would not help us distinguish between any of the other possible outcomes of the patient.

We also derive the binary classifier under two competing risks. The resulting AUROC scores are shown in Table 5 and ROC plots at $t = 6, 12, \Delta = 24$ in Figure 4. By comparing Tables 2 and 5, average AUROCs drop for both the Fine and Gray model and Dynamic-DeepHit when we change from three events to two events, showing that there is a benefit to including the event of withdrawal from life-sustaining therapies. By comparing Figures 3 and 4, focusing on the low FPR regime, we observe the TPRs for Dynamic-DeepHit stay at a similar level under the two event setting while those for the Fine and Gray model become very unstable, again, suggesting the benefit of including the event of withdrawal. The ROC curves under the two event setting for a few other t and Δ values can be found in Appendix G.

Table 5: Test set AUROC (average ± standard deviation across five experimental repeats) for the Fine and Gray model and Dynamic-DeepHit with two events. The entries with bold values represent the highest average AUROC for each (t, Δ) combination.

Model	Prediction time	Evaluation time horizon		
		$\Delta = 24$ hrs	$\Delta = 48$ hrs	$\Delta = 72$ hrs
Fine and Gray	$t = 6$	**0.893 ± 0.032**	**0.892 ± 0.033**	**0.891 ± 0.033**
	$t = 12$	**0.890 ± 0.039**	**0.887 ± 0.038**	**0.886 ± 0.038**
Dynamic-DeepHit	$t = 6$	0.820 ± 0.040	0.827 ± 0.034	0.832 ± 0.035
	$t = 12$	0.820 ± 0.059	0.824 ± 0.053	0.835 ± 0.057



(a) $t = 6, \Delta = 24$ (b) $t = 12, \Delta = 24$

Figure 4: Test set ROC curve in the case of modeling two competing events (average curve ± standard deviation intervals across five experimental repeats). The x-axis is on a log scale to emphasize the low FPR regime.

6. Discussion

Our paper proposes a dynamic formulation of the neurological prognostication problem for post-cardiac-arrest coma patients. The modeling solution we propose uses any existing DCR model and, using specific structure in clinical outcomes of our clinical application, derives a classifier from the DCR model that aims to be helpful to clinicians for decision support. We believe that our dynamic formulation better models the specific clinical problem we focus on compared to what has been proposed in existing literature.

We now discuss an alternative to our conditioning strategy in Section 3.1 that we believe is worth investigating in future work, and we also point out various other limitations.

An alternative solution for deriving a binary classifier In how we derived our binary classifier in Section 3.1, we conditioned on the event that *the test patient never gets withdrawn from life-sustaining therapies* (the "non-withdrawal event"). Our classifier uses the probability of awakening and, separately, of dying both conditioned on the non-withdrawal event. However, we have found that these conditional probabilities are not entirely straightforward to explain to practitioners. Part of the challenge is that it is unclear how conditioning on the non-withdrawal event should impact the probabilities estimated. Concretely, consider the probability of awakening. Prior to conditioning on the

non-withdrawal event vs after conditioning on the event, it is unclear whether the probability of awakening should increase or decrease.

Fundamentally, two technical hurdles make reasoning about the non-withdrawal event difficult. First, we do not know what would have happened to patients withdrawn from life-sustaining therapies had they been kept on these therapies instead. Second, we do not know how "good" past decisions on withdrawing life-sustaining therapies are. As an extreme example, suppose that 100% of patients who die from non-withdrawal causes are perfectly identified by physicians in advance that they would have no chance of awakening so they are pulled off life-sustaining therapies, and 100% of patients who would awaken are also perfectly identified by physicians so that they are kept on life-sustaining therapies. In this case, if we condition on the non-withdrawal event, then we would always just get that the conditional probability of awakening is 100%, so our binary classifier in Section 3.1 would not be useful. However, an alternative that would still be useful is to compute the probabilities of awakening and of dying (not of withdrawal from life-sustaining therapies) *without* conditioning on the non-withdrawal event. These probabilities would still be conditional probabilities since we would condition on the test patient's time series. However, we no longer condition on the non-withdrawal event. We now sketch an approach for computing these probabilities, which in turn leads to a binary classifier different from the one we derived in Section 3.1.

To begin with, let's consider the probability of awakening within time duration Δ for a patient with features observed until time t. We write this probability as

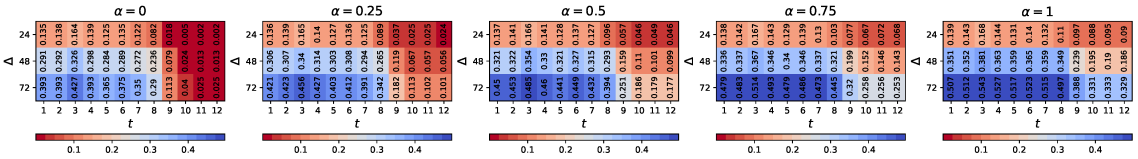
$$\begin{aligned}
&\widetilde{P}_{\text{awaken}}(\Delta \mid z, t) \\
&= \mathbb{P}(\text{earliest competing event that happens excluding withdrawal from life-sustaining} \\
&\quad \text{therapies is awakening}, Y \leq t + \Delta | Z^{(\leq t)} = z^{(\leq t)}, Y > t) \\
&= \mathbb{P}(\text{earliest competing event that happens is awakening}, Y \leq t + \Delta | Z^{(\leq t)} = z^{(\leq t)}, Y > t) \\
&\quad + \mathbb{P}(\text{earliest competing event that happens is withdrawal from life-sustaining} \\
&\quad \text{therapies followed by awakening}, Y \leq t + \Delta | Z^{(\leq t)} = z^{(\leq t)}, Y > t).
\end{aligned}$$

On the right-hand side above, the first term is just the CIF for awakening (i.e., $F_1(\Delta \mid z, t)$ using equation (1) and the same notation as in Section 3.1). As for the second term, we now make a major assumption that

$$\begin{aligned}
&\mathbb{P}(\text{earliest competing event that happens is withdrawal from life-sustaining} \\
&\quad \text{therapies followed by awakening}, Y \leq t + \Delta | Z^{(\leq t)} = z^{(\leq t)}, Y > t) \\
&\quad = F_3(\Delta \mid z, t) \times \alpha,
\end{aligned}$$

where $\alpha \in [0, 1]$ is a constant that does not depend on anything else, and as a reminder F_3 is the CIF for withdrawal from life-sustaining therapies. The above assumption says that among patients whose earliest competing event is being withdrawn from life-sustaining therapies, each of them has probability α (independent of everything else) of having their second earliest competing event be awakening (so that had withdrawal from life-sustaining therapies not been an option, they would have awaken). In practice, we have no way of estimating α but we can try different values for it. Putting together the pieces, we have

$$\widetilde{P}_{\text{awaken}}(\Delta \mid z, t) = F_1(\Delta \mid z, t) + F_3(\Delta \mid z, t) \times \alpha. \tag{8}$$



(*a*) Example Patient 1

Figure 5: For the same Example Patient 1 in panel (*a*) of Figure 2, we show the probability of awakening $\widetilde{P}_{\text{awaken}}(\Delta \mid z, t)$ (from equation (8)) for different values of α.

In Figure 5, we provide the heatmap visualization of the probability of awakening with $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ for the same Example Patient 1 as in Figure 2. We can see when α increases from zero to one, the estimation becomes more "optimistic" in terms of the probability of awakening being higher across different times t and durations Δ.

The same logic can be used to derive the probability of dying (of causes aside from withdrawal from life-sustaining therapies) within duration Δ; we can denote this probability as $\widetilde{P}_{\text{death (not withdrawal)}}(\Delta \mid z, t)$. Then we can derive a binary classifier in the same manner as how we derived the one in Section 3.1: e.g., we threshold on the ratio $\frac{\widetilde{P}_{\text{death (not withdrawal)}}(\Delta|z,t)}{\widetilde{P}_{\text{awaken}}(\Delta|z,t)}$ to decide on whether to predict between awakening or dying.

At present, we do not yet fully understand when our proposed solution in Section 3.1 should be used in practice vs the one stated in this discussion section which makes the major assumption involving the unknown probability α of patients withdrawn from life-sustaining therapies who would have awaken. Better understanding the pros and cons of these approaches would be interesting. Relaxing the factorization assumption involving the probability α would also be an interesting future research direction.

Other limitations At a high-level, our paper has proposed a framework in which to think about the neurological prognostication problem that works with any DCR model. However, at this point, we have not proposed a new DCR model and, as far as we know, there have simply not been many DCR models developed (where the model truly is using variable-length time series rather than how we set up the Fine and Gray baseline which actually only uses fixed-length feature vector inputs). Naturally, a future research direction would be the development of DCR models that are even more accurate than the ones we have tested, especially since for the accuracy metrics we use, state-of-the-art neural network DCR models do not appear to be significantly better than the classical Fine and Gray model that only uses the final time step's features (including static features).

In terms of evaluating the binary classifier, we simply use AUROC scores at different t, Δ with patients who we either observed awakening or death (not from the withdrawal of life-sustaining therapy), ignoring those who have been withdrawn or censored due to lost of follow-up, which could be biased. There are numerous work in the area of ROC analysis with the occurrence of censoring under time-dependent setting (e.g., Blanche et al. 2013; Kamarudin et al. 2017; Li et al. 2018), each with its own limitations. While it is not the focus of our current work, we encourage future work to probe how to better evaluate the binary classifier.

Another limitation of our framework is that we only consider the first "hitting time" of competing risks, i.e. the earliest event that happened. In reality, some patients may still

experience unfavorable outcomes after awakening, and it could be important to try to predict when this happens. One simple approach would be to split up the "awaken" event into different types of "awaken" events, although perhaps a better modeling framework would be to consider multiple critical events that could happen, one after another, rather than constraining the setting to only be on first hitting times.

As for the specific clinical problem we have focused on, there were a number of limitations related to the dataset we curated. First, the EEG data we have is limited to the first 12 hours after ICU admission. If we were to have more complete EEG data, we would potentially be able to provide more accurate prognoses. Second, for ease of computation, we downsample the EEG time series data. From some preliminary analyses, we found that this did not impact the resulting prediction accuracy much. However, more thorough experiments are needed to better understand the impact of our current downsampling procedure on information loss. We remark that our patient heat map visualization can help us also find when our DCR-derived classifier is actually wrong but it is very confident in its wrong answer (we provide some examples of this in Appendix F). In some such cases, from physician feedback, we have learned that the prognostication task even for physicians could be difficult given only EEG data and the static features we use. In particular, by collecting and using additional patient features (e.g., vitals, whether the patient experienced another cardiac arrest), more accurate predictions should be possible.

Ultimately, although we believe that our paper takes a step toward more realistically framing the problem of neurological prognostication of post-cardiac-arrest coma patients compared to existing literature, our work has not yet led to a "deployment-ready" solution. Moreover, at this point, it is unclear to what extent the patient-specific heat map visualization we proposed actually helps clinical decision support. By continuing to account for physician feedback, we hope that we can produce a decision support system that is practically useful. User studies with clinicians would be required to assess the impact of such a system on clinical decision making.

Acknowledgments

This work was supported by NSF CAREER award #2047981. The authors thank the anonymous reviewers for helpful feedback.

References

Nicholas S Abend, Dennis J Dlugos, Cecil D Hahn, Lawrence J Hirsch, and Susan T Herman. Use of EEG monitoring and management of non-convulsive seizures in critically ill patients: a survey of neurologists. *Neurocritical Care*, 12:382–389, 2010.

MM Admiraal, LA Ramos, S Delgado Olabarriaga, HA Marquering, J Horn, and AF van Rootselaar. Quantitative analysis of EEG reactivity for neurological prognostication after cardiac arrest. *Clinical Neurophysiology*, 132(9):2240–2247, 2021.

Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397, 2013.

Aziza Byron-Alhassan, Barbara Collins, Marc Bedard, Bonnie Quinlan, Michel Le May, Lloyd Duchesne, Christina Osborne, George Wells, Andra M Smith, and Heather E Tulloch. Cognitive dysfunction after out-of-hospital cardiac arrest: Rate of impairment and clinical predictors. *Resuscitation*, 165:154–160, 2021.

Maria De-Arteaga, Jieshi Chen, Peter Huggins, Jonathan Elmer, Gilles Clermont, and Artur Dubrawski. Predicting neurological recovery with canonical autocorrelation embeddings. *PLOS One*, 14(1):e0210966, 2019.

Jonathan Elmer, John J Gianakas, Jon C Rittenberger, Maria E Baldwin, John Faro, Cheryl Plummer, Lori A Shutter, Christina L Wassel, Clifton W Callaway, and Anthony Fabio. Group-based trajectory modeling of suppression ratio after cardiac arrest. *Neurocritical Care*, 25(3):415–423, 2016a.

Jonathan Elmer, Jon C Rittenberger, John Faro, Bradley J Molyneaux, Alexandra Popescu, Clifton W Callaway, Maria Baldwin, and Pittsburgh Post-Cardiac Arrest Service. Clinically distinct electroencephalographic phenotypes of early myoclonus after cardiac arrest. *Annals of Neurology*, 80(2):175–184, 2016b.

Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

Hans Friberg, Tobias Cronberg, Martin W Dünser, Jacques Duranteau, Janneke Horn, and Mauro Oddo. Survey on current practices for neurological prognostication after cardiac arrest. *Resuscitation*, 90:158–162, 2015.

Romergryko G Geocadin, Clifton W Callaway, Ericka L Fink, Eyal Golan, David M Greer, Nerissa U Ko, Eddy Lang, Daniel J Licht, Bradley S Marino, and Norma D McNair. Standards for studies of neurological prognostication in comatose survivors of cardiac arrest: a scientific statement from the American Heart Association. *Circulation*, 140(9): e517–e542, 2019.

Hannah C Glass, Courtney J Wusthoff, and Renée A Shellhaas. Amplitude-integrated electro-encephalography: the child neurologist's perspective. *Journal of Child Neurology*, 28(10):1342–1350, October 2013.

Robert J Gray. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, pages 1141–1154, 1988.

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

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data (2nd ed.)*. John Wiley & Sons, 2002.

Adina Najwa Kamarudin, Trevor Cox, and Ruwanthi Kolamunnage-Dona. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, 17(1):1–19, 2017.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

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, 2019.

Liang Li, Tom Greene, and Bo Hu. A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Statistical Methods in Medical Research*, 27(8):2264–2278, 2018.

EA Matthews, J Magid-Bernstein, A Presciutti, A Rodriguez, David Roh, S Park, J Claassen, and S Agarwal. Categorization of survival and death after cardiac arrest. *Resuscitation*, 114:79–82, 2017.

Intae Moon, Stefan Groha, and Alexander Gusev. SurvLatent ODE: A neural ODE based time-to-event model with competing risks for longitudinal data improves cancer-associated Venous Thromboembolism (VTE) prediction. In *Machine Learning for Healthcare*, 2022.

Marion Moseby-Knappe, Erik Westhall, Sofia Backman, Niklas Mattsson-Carlgren, Irina Dragancea, Anna Lybeck, Hans Friberg, Pascal Stammet, Gisela Lilja, and Janneke Horn. Performance of a guideline-recommended algorithm for prognostication of poor neurological outcome after cardiac arrest. *Intensive Care Medicine*, 46:1852–1862, 2020.

Sang Hoon Oh, Kyu Nam Park, Young-Min Shon, Young-Min Kim, Han Joon Kim, Chun Song Youn, Soo Hyun Kim, Seung Pill Choi, and Seok Chan Kim. Continuous amplitude-integrated electroencephalographic monitoring is a useful prognostic tool for hypothermia-treated cardiac arrest patients. *Circulation*, 132(12):1094–1103, 2015.

Jon C Rittenberger, Alexandra Popescu, Richard P Brenner, Francis X Guyette, and Clifton W Callaway. Frequency and timing of nonconvulsive status epilepticus in comatose post-cardiac arrest subjects treated with hypothermia. *Neurocritical Care*, 16:114–122, 2012.

Andrea O Rossetti, Emmanuel Carrera, and Mauro Oddo. Early EEG correlates of neuronal injury after brain anoxia. *Neurology*, 78(11):796–802, 2012.

Andrea O Rossetti, Alejandro A Rabinstein, and Mauro Oddo. Neurological prognostication of outcome in patients in coma after cardiac arrest. *The Lancet Neurology*, 15(6):597–609, 2016.

Shubhranshu Shekhar, Dhivya Eswaran, Bryan Hooi, Jonathan Elmer, Christos Faloutsos, and Leman Akoglu. Benefit-aware early prediction of health outcomes on multivariate EEG time series. *Journal of Biomedical Informatics*, 139:104296, 2023.

Helle Søholm, Troels Wesenberg Kjær, Jesper Kjaergaard, Tobias Cronberg, John Bro-Jeppesen, Freddy K Lippert, Lars Køber, Michael Wanscher, and Christian Hassager. Prognostic value of electroencephalography (EEG) after out-of-hospital cardiac arrest in successfully resuscitated patients used in daily clinical practice. *Resuscitation*, 85(11): 1580–1585, 2014.

Eyad AL Thenayan, Martin Savard, Michael D Sharpe, Loretta Norton, and Bryan Young. Electroencephalogram for prognosis after cardiac arrest. *Journal of Critical Care*, 25(2): 300–304, 2010.

Niranjan Damera Venkata and Chiranjib Bhattacharyya. When to intervene: Learning optimal intervention policies for critical events. *Advances in Neural Information Processing Systems*, 35:30114–30126, 2022.

Erik Westhall, Andrea O Rossetti, Anne-Fleur van Rootselaar, Troels Wesenberg Kjaer, Janneke Horn, Susann Ullén, Hans Friberg, Niklas Nielsen, Ingmar Rosén, and Anders Åneman. Standardized EEG interpretation accurately predicts prognosis after cardiac arrest. *Neurology*, 86(16):1482–1490, 2016.

Appendix A. Data

Some summary statistics for the dataset corresponding to the cohort described in Section 4 are provided in Table A.1.

EEG Features After post-cardiac-arrest comatose patients are admitted to the intensive care unit, electroencephalography (EEG) is used to measure brain activity. For the data we use, EEG signals are typically recorded at 256Hz from 22 electrodes adhering to standard positions according to the 10-20 International System of electrode placement. They are then processed via the medical software Persyst (`https://www.persyst.com/`) to generate 6,037 features at 1Hz, which is the "raw" format of the data we start with (note that the actual raw data originally recorded by the EEG electrodes are not available, only this processed version from Persyst). The 6,037 features from Persyst include time, artifact intensity, electrode signal quality, seizure probability, FFT (fast Fourier transform) spectrogram, aEEG, peak envelope (0-25 Hz), rhythmicity spectrogram, asymmetry EASI/REASI, relative asymmetry spectrogram, spikes, and suppression ratio. Out of the above features, for simplicity, we focus on two types of EEG features that have been previously proven to be informative for neurological prognostication (e.g., Oh et al. 2015; Elmer et al. 2016a):

Suppression ratio is calculated as a 10-second summary and reflects the proportion of the amplitude of the EEG signals in that window that falls below some threshold compared to that which falls above the threshold. More suppressed EEG is a sign of more severe dysfunction or injury. Specifically, we use 2 features: the mean values of all the electrodes in the left and right hemispheres respectively.

Amplitude-integrated EEG (aEEG) is one way of describing the overall EEG amplitude. Very low amplitude EEGs are a sign of injury or dysfunction. Specifically, we use 10 features: the mean values of all the electrodes in the left and right hemispheres, each with 5 different statistical summaries (max, min, median, 25% percentile, 75% percentile).

In summary, we use the below 12 features in our study:

- Mean of suppression ratio of all the electrodes in the left hemisphere
- Mean suppression ratio of all the electrodes in the right hemisphere
- Mean of max aEEG values of all the electrodes in the left hemisphere
- Mean of min aEEG values of all the electrodes in the left hemisphere
- Mean of median aEEG values of all the electrodes in the left hemisphere
- Mean of 25% percentile aEEG values of all the electrodes in the left hemisphere
- Mean of 75% percentile aEEG values of all the electrodes in the left hemisphere
- Mean of max aEEG values of all the electrodes in the right hemisphere
- Mean of min aEEG values of all the electrodes in the right hemisphere
- Mean of median aEEG values of all the electrodes in the right hemisphere
- Mean of 25% percentile aEEG values of all the electrodes in the right hemisphere
- Mean of 75% percentile aEEG values of all the electrodes in the right hemisphere

Static Features A full list of the static features we use in the study and possible values they could take:

- Demographic
  - `age`: Age
  - `female`: Female gender

Table A.1: Summary statistics of patient characteristics and outcomes. All features from "Age (yr)" and below in the table have values reported as means except for the different "Presenting rhythm" features, which are all in raw counts. In the column names, "withdrawal" is an abbreviation for withdrawal from life-sustaining therapies (due to poor perceived neurological prognosis).

	Total	Awakened	Death (not from withdrawal)	Withdrawal	Coma
Number of patients	922	271	189	432	30
Percentage of patients	100%	29.4%	20.5%	46.9%	3.3%
Age (yr)	57.3	55.8	54.6	59.4	57.8
Female sex	0.37	0.35	0.44	0.36	0.40
Arrest out-of-hospital	0.85	0.79	0.85	0.87	0.90
Presenting rhythm					
Ventricular tachycardia/fibrillation	284	140	38	98	8
Pulseless electrical activity	317	83	72	153	9
Asystole	268	40	63	153	12
Unknown	53	8	16	28	1
Arrest duration (min)	20.4	13.3	26.4	22.4	18.1
Time to event (hr)	113.0	78.9	90.6	111.5	583.8

- Heart arrest
  - `oohca`: Arrest location. *Out-of-hospital, or In-hospital*
  - `edarrest`: Initial arrest occurred after ED arrival? *Yes, or No*
  - `rhythm`: Initial arrest rhythm. *No loss of pulse, VT/VF, PEA, Asystole, or Unknown*
  - `ca_type`: Pittsburgh Cardiac Arrest Category, *I, II, III, IV, or Unknown*
  - `transfer`: Referral from outside facility? *Yes, or No*
  - `witnessed`: Witnessed arrest? *No, Lay person witnessed, or EMS witnessed*
  - `bystander_cpr`: Bystander CPR? *No, Lay person, or Professional*
  - `shocks`: Number of AED and ALS shocks for the duration of the initial resuscitation
  - `duration`: Estimated cumulative duration of CPR in minutes
- Initial coma status
  - `four_r_0`: FOUR Score - Respiratory
  - `four_eye_0`: FOUR Score - Eyes
  - `four_m_0`: FOUR Score - Motor
  - `pupils_0`: Pupils status. *Both reactive, One reactive, Both nonreactive or Unable to determine*
  - `corneals_0`: Corneals. *Both present, One present, Neither present, or Unable to determine*
  - `cough_0`: Cough. *Present, Absent, or Unable to determine*
  - `gag_0`: Gag. *Present, Absent, or Unable to determine*
- Medical history: *Yes, or No*
  - `ccimi`: History of myocardial infarction
  - `ccipvd`: History of peripheral vascular disease

`ccidementia`: History of dementia
`ccicva`: History of stroke or TIA
`ccihemi`: History of hemiplegia
`ccichf`: History of congestive heart failure
`ccicvd`: History of cerebrovascular disease
`ccicld`: History of chronic lung disease/COPD
`ccictd`: History of connective tissue disease
`ccipud`: History of peptic ulcer disease
`cciaids`: History of AIDS
`ccickd`: History of moderate to severe chronic kidney disease
`ccielsd`: History of chronic liver disease
`ccidm`: History of diabetes
`ccica`: History of solid tumor
`ccileukemia`: History of leukemia
`ccilymphoma`: History of lymphoma

Appendix B. Training the Subdistribution Hazard Model

Since the subdistribution hazard model by Fine and Gray (1999) does not take vary-length time series data as inputs, we instead only use the static features and the EEG features of the last hour available for each patient in the training set to estimate the parameters of the subdistribution hazard model. After the parameters are estimated, we then plug in the static features and the EEG features at $t = 6$ to estimate the CIFs and calculate the c-indices corresponding to $t = 6$ and $\Delta = 24, 48, 72$. Similarly, we generate the c-indices corresponding to $t = 12$ and $\Delta = 24, 48, 72$, by using the same static features and EEG features at $t = 12$. This process is repeated five times with different random splits of training and testing sets to get the average and standard deviation of c-indices as in Table 1. Note that there is no validation set used for hyperparameter tuning as there are no hyperparameters.

Appendix C. Modifying DDRSA to Support Competing Risks

To make DDRSA (Venkata and Bhattacharyya, 2022) support competing risks, we modify the model structure of the original DDRSA, largely inspired by Dynamic-DeepHit (Lee et al., 2019). In the original DDRSA, the encoder RNN module takes in time-varying input features, and the decoder RNN module recurrently produces predicted hazards at different discretized time intervals. To make it support competing events, instead of having only one decoder RNN module, we have multiple decoder RNN modules to generate predictions for each of the competing events respectively. Instead of predicting hazards at different discretized time intervals, we make the model to predict the probability mass function variant of the CIF as in equation (2), where a softmax layer is applied at the end to make sure the constraint in equation (4) is satisfied. We also incorporate the attention mechanism as in Dynamic-DeepHit. The modified DDRSA can be trained in the same fashion as Dynamic-DeepHit.

Appendix D. Hyperparameter Grid

Note that for Dynamic-DeepHit, the loss is of the form:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \alpha \cdot \mathcal{L}_2 + \beta \cdot \mathcal{L}_3$$

where L_1 is the negative log likelihood loss term (this basically encourages first hitting times to be correctly predicted, accounting for censoring), L_2 is a ranking loss (especially as the standard survival analysis metric of concordance index is a ranking-based metric, having a ranking loss that correctly orders patients based on their predicted CIFs can be helpful), L_3 is the RNN prediction loss (recall that we ask the RNN to try to predict the next time step's feature vector), and lastly $\alpha > 0$ and $\beta > 0$ are hyperparameters.

We use the following hyperparameter grid for both Dynamic-DeepHit and DDRSA:

- learning rate $\in \{10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$
- weights of different loss terms: $\alpha \in \{0.5, 1, 5\}, \beta \in \{0.05, 0.1, 0.5\}$
- dropout rate $\in \{0.2, 0.4\}$

For all sets of hyperparameters, we train with the Adam optimizer and a batch size of 32. The maximum number of epochs we train the models is 100 while we stop training if the average concordance index (across all the events with $\Delta = 24, 48, 72$) on the validation set does not improve for 10 epochs. The best hyperparameter set is chosen to be the one with the highest average concordance index (across all the events with $\Delta = 24, 48, 72$) on the validation set.

Appendix E. Additional ROC Curves

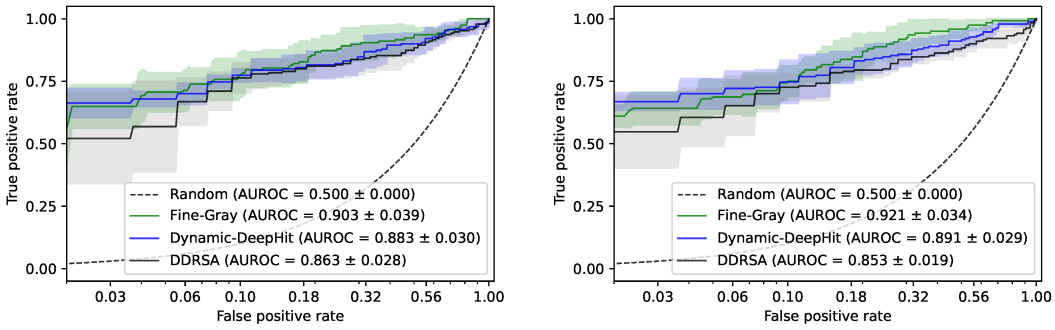
We provide the ROC derived based on the estimated conditional probability of awakening for each patient at $t = 6, 12$ and $\Delta = 48, 72$, as shown in Figure E.1.

Appendix F. Additional Patient-Specific Visualizations

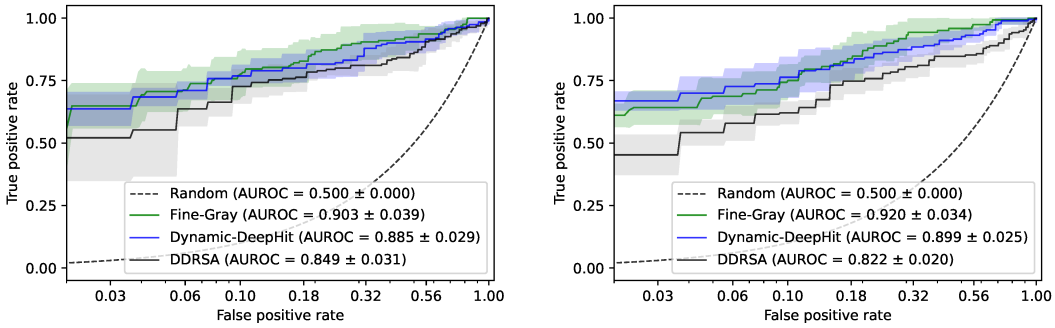
We show visualizations for three patients (Figure F.1) where the predictions are, in some sense, inconsistent with the EEG signals. In particular, for all three patients shown, their EEG signals would be perceived as corresponding to poor neurological activities, while the predictions of awakening are fairly high. While we do not understand what exactly leads to the overly high predicted probabilities of awakening for Patients A and C, we suspect the "rareness" of the EEG patterns is the reason for the inaccurate prediction for Patient B. In fact, the suppression ratio for one hemisphere and aEEG values are quite normal but the suppression ratio values for the other hemisphere are very high, indicating abnormal brain activities. This kind of inconsistency among different EEG features is very rare in the dataset that we curated.

Appendix G. Additional ROC in Ablation Study

We provide the ROC when we consider only two competing events, awakening and death not by withdrawal from life-sustaining therapies, at $t = 6, 12$ and $\Delta = 48, 72$, as shown in Figure G.1.

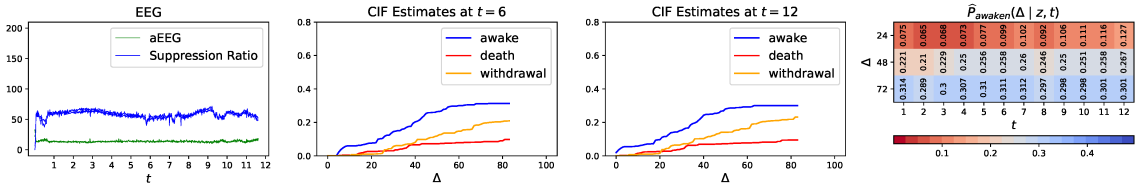


(a) $t = 6, \Delta = 48$ (b) $t = 12, \Delta = 48$

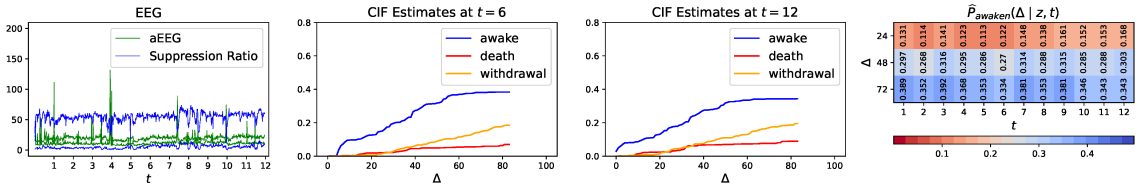


(c) $t = 6, \Delta = 72$ (d) $t = 12, \Delta = 72$

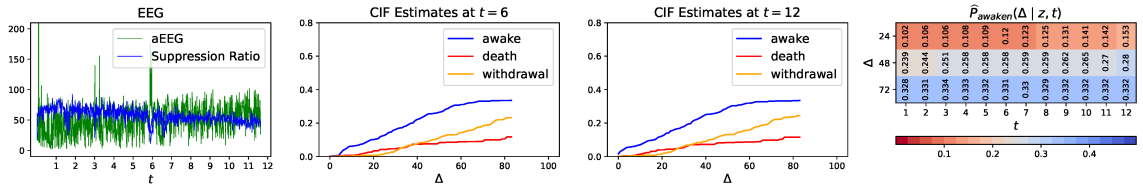
Figure E.1: Test set ROC curve (average curve ± standard deviation intervals across five experimental repeats). The x-axis is on a log scale to emphasize the low FPR regime.



(a) Example Patient A: died (not from withdrawal of life-sustaining therapies) at hour 106

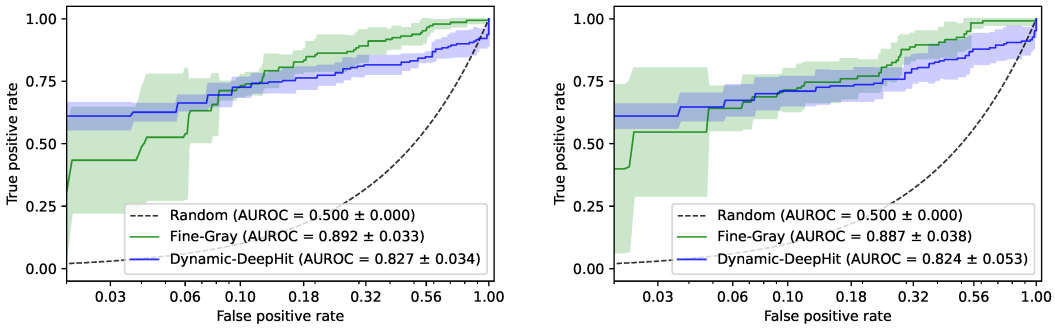


(b) Example Patient B: withdrawn from life-sustaining therapies at hour 101

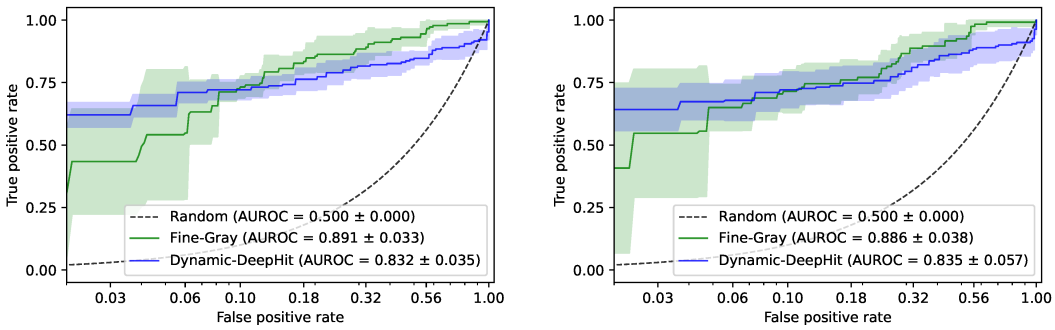


(c) Example Patient C: died (not from withdrawal of life-sustaining therapies) at hour 14

Figure F.1: For each example patient (panels (a)-(c)), we show time series of two summary EEG features *aEEG* and *suppression ratio* (first column plot), estimated CIFs at hours $t = 6$ (second column plot) and $t = 12$ (third column plot), and our proposed heat map visualization (fourth column plot). Note that aEEG values for normal brain activity should be within a certain range (constantly being lower than 5 or higher than 25 is usually considered abnormal), and higher suppression ratio values are a sign of more severe dysfunction or injury.



(a) $t = 6, \Delta = 48$ (b) $t = 12, \Delta = 48$



(c) $t = 6, \Delta = 72$ (d) $t = 12, \Delta = 72$

Figure G.1: Test set ROC curve with two events (average curve ± standard deviation intervals across five experimental repeats). The x-axis is on a log scale to emphasize the low FPR regime.