Proceedings of Machine Learning Research 248:540–557, 2024 Conference on Health, Inference, and Learning (CHIL) 2024

Dynamic Survival Analysis for Early Event Prediction

Hugo Yèche[*] HYECHE@ETHZ.CH
ETH Zürich, Department of Computer Science, Switzerland
Manuel Burger[*] MANUEL.BURGER@ETHZ.CH
ETH Zürich, Department of Computer Science, Switzerland
Dinara Veshchezerova DINARA.VESHCHEZEROVA@ETHZ.CH
ETH Zürich, Department of Computer Science, Switzerland
Gunnar Rätsch RAETSCH@INF.ETHZ.CH
ETH Zürich, Department of Computer Science, Switzerland

Abstract

This study advances Early Event Prediction (EEP) in healthcare through Dynamic Survival Analysis (DSA), offering a novel approach by integrating risk localization into alarm policies to enhance clinical event metrics. By adapting and evaluating DSA models against traditional EEP benchmarks, our research demonstrates their ability to match EEP models on a time-step level and significantly improve event-level metrics through a new alarm prioritization scheme (up to 11% AuPRC difference). This approach represents a significant step forward in predictive healthcare, providing a more nuanced and actionable framework for early event prediction and management.

Data and Code Availability This paper uses the MIMIC-III dataset (version 1.4) (Johnson et al., 2016b) available on PhysioNet (Johnson et al., 2016a) and the HiRID dataset (version 1.1.1) (Faltys et al., 2021b) also available on Physionet (Faltys et al., 2021a). We provide a code repository[1].

Institutional Review Board (IRB) This research does not require IRB approval in the country it was performed in.

1. Introduction

Early event prediction (EEP) on time series is concerned with determining whether an event will occur within a fixed time horizon. It is highly relevant to a wide range of monitoring applications in fields such as environment (Di Giuseppe et al., 2016) or healthcare (Sutton et al., 2020). Using machine learning for EEP has gained particular interest in Intensive Care Unit (ICU) patient monitoring (Harutyunyan et al., 2019; Hyland et al., 2020; Yèche et al., 2021; van de Water et al., 2024), where large quantities of medical data are collected automatically. Existing works train such models through maximum likelihood estimation (MLE) of the cumulative failure function for a fixed horizon. However, to be usable by clinicians at an event scale, one needs to design an alarm mechanism leveraging the time-step level failure estimates. If existing works have proposed various ways of evaluating EEP models at event scale (Tomašev et al., 2019; Hyland et al., 2020), the design of the alarm policy based on time-step prediction has been overlooked. Current approaches (Tomašev et al., 2019; Hyland et al., 2020; Hüser et al., 2024; Lyu et al., 2024) rely on a simple fixed threshold mechanism on the time-step prediction to raise alarms at the event scale. One limitation to more advanced policies is that due to their cumulative nature current EEP models do not provide information concerning the imminence of the risk within the considered horizon as depicted in Figure 3.

Parallelly, in statistics, survival analysis (SA), also known as time-to-event analysis, considers the highly related problem of predicting the *exact* time of a future event given a set of covariates. With deep learning emergence, the field has also recently pivoted to discrete-time methods using neural networks (Tutz et al., 2016; Gensheimer and Narasimhan, 2019; Kvamme et al., 2019; Lee et al., 2018; Ren et al., 2019) to fit hazard or probability mass functions (PMF). The extension of SA to longitudinal covari-

[*] First co-authors
[1] https://github.com/ratschlab/dsa-for-eep

ates, namely dynamic survival analysis, also gained popularity in the deep learning field (Lee et al., 2019; Jarrett et al., 2019; Damera Venkata and Bhattacharyya, 2022; Maystre and Russo, 2022). As opposed to models trained to maximize EEP likelihood, DSA models estimate the event PMF at any horizon. Thus, in theory, such a model can also provide an estimate of the cumulative failure function as required in EEP tasks while additionally providing a decomposition of where such a risk lies within the considered horizon.

In this work, we study the usage of deep learning models trained with a DSA likelihood for EEP tasks to design more advanced alarm policies. Our contribution can be summarized as follows: (**i**) We formalize and propose how to train and use DSA models to match EEP models' timestep-level performance on three established benchmarks (**ii**) To this end, we propose *survTLS*, a non-trivial extension to temporal label smoothing Yèche et al. (2023) (TLS) for DSA. (**iii**) At the event level, we propose a simple yet novel scheme, leveraging the risk localization provided by DSA models to prioritize imminent alarms, resulting in further performance improvement over EEP models.

2. Related Work

Early event prediction from EHR data As previously mentioned, early warning systems (EWS) using deep learning models have recently gained traction in the literature. Indeed, over the years multiple large publicly available EHR databases (Johnson et al., 2016b; Pollard et al., 2018; Faltys et al., 2021b; Thoral et al., 2021) and benchmarks including EEP tasks (Harutyunyan et al., 2019; Wang et al., 2020; Reyna et al., 2020; Yèche et al., 2021; van de Water et al., 2024) were released. Using these, existing works proposed new architecture designs (Horn et al., 2020; Tomašev et al., 2019), imputation methods (Futoma et al., 2017), and more recently objective functions Yèche et al. (2023). However, in all these works, the backbone model is trained via MLE on the cumulative failure function at the horizon of prediction. Thus, to our knowledge, our work is the first to investigate the DSA models for these EEP tasks.

Survival analysis in the era of deep learning With deep learning emergence, SA quickly moved away from proportional linear hazard models, as originally proposed by Cox (1972). A stream of work uses neural networks to parameterize the hazard function (Yousefi et al., 2017; Katzman et al., 2018), Gensheimer and Narasimhan (2019) additionally remove the proportional hazard assumption. Simultaneously, other works focus on parameterizing the PMF (Kvamme et al., 2019; Lee et al., 2018; Ren et al., 2019), while adding regularization terms to their negative log-likelihood objective. Among these works, Lee et al. (2018) went in the opposite direction to our work by fitting the PMF with MLE on the cumulative function, similar to EEP. In DSA, based on the landmarking idea (Van Houwelingen, 2007; Parast et al., 2014), advances followed a similar trend with works parametrizing the PMF (Damera Venkata and Bhattacharyya, 2022) and fitting an MLE on the cumulative failure function (Jarrett et al., 2019; Lee et al., 2019).

Survival analysis and event classification Prior work has investigated the use of static survival analysis for event classification problems such as early detection of fraud (Zheng et al., 2019). This work however remains restricted to non-dynamic application, thus not applicable to DSA or EEP. On the other hand, Shen et al. (2023) propose a model for EEP applications trained with a vanilla DSA likelihood to classify neurological prognostication. However, they do not provide any comparison to EEP MLE nor propose tailored alarm policies to the localized risk estimation.

3. Methods

In this section, we describe our DSA approach to EEP tasks starting from the timestep estimator training to the alarm policy design. An overview of the pipeline and how it compares to EEP can be found in Figure 1.

3.1. From Early Event Prediction to Dynamic Survival Analysis

Early event prediction In EEP, we consider a dataset of multivariate time series of covariates $\mathbf{X}_i$ and binary event labels $e_{i,t}$ representing the occurrence of an event at time t in trajectory i. Each sample i is a sequence $\{(\mathbf{x}_{i,0}, e_{i,0}), \dots, (\mathbf{x}_{i,T_i}, e_{i,T_i})\}$ of length T_i. For each timepoint t along a time series, the covariates observed up to this point are denoted $\mathbf{X}_{i,t} = [\mathbf{x}_{i,0}, \dots, \mathbf{x}_{i,t}]$ and the time of the next event is given by $T_e(t) = \arg\min_{\tau:\tau \geq t}\{e_\tau : e_\tau = 1\}$. If no event happened, we define $T_e(t) = +\infty$ and call this

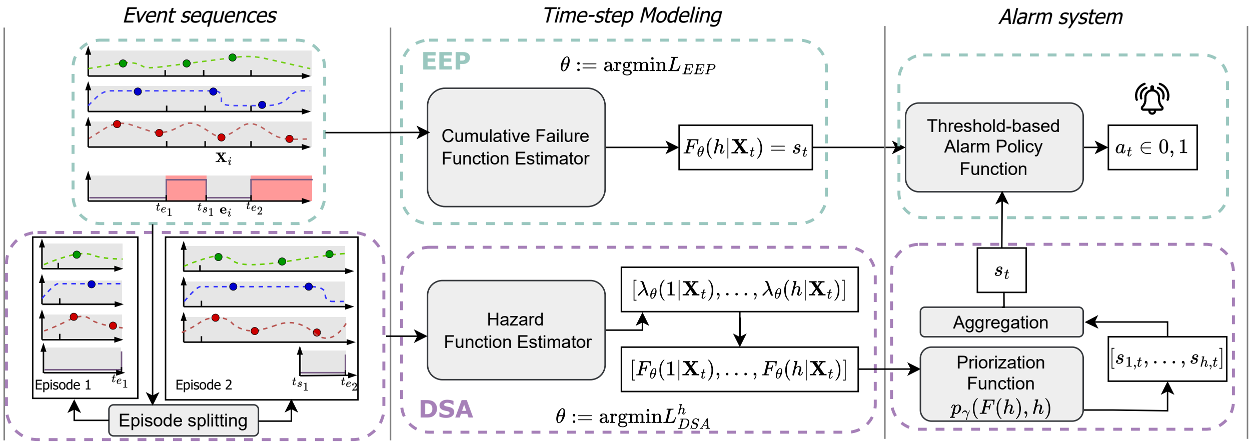

Event sequences
Time-step Modeling
Alarm system
EEP
$\theta := \mathrm{argmin} L_{EEP}$
Cumulative Failure Function Estimator
$F_\theta(h|\mathbf{X}_t) = s_t$
Threshold-based Alarm Policy Function
$a_t \in 0,1$
$\mathbf{X}_i$
t_{e_1} t_{s_1} $\mathbf{e}_i$ t_{e_2}
Episode 1
t_{e_1}
Episode 2
t_{s_1} t_{e_2}
Episode splitting
Hazard Function Estimator
$[\lambda_\theta(1|\mathbf{X}_t), \ldots, \lambda_\theta(h|\mathbf{X}_t)]$
$[F_\theta(1|\mathbf{X}_t), \ldots, F_\theta(h|\mathbf{X}_t)]$
DSA
$\theta := \mathrm{argmin} L^h_{DSA}$
s_t
Aggregation
Priorization Function $p_\gamma(F(h), h)$
$[s_{1,t}, \ldots, s_{h,t}]$

Figure 1: Overview of the pipeline for EEP (green) and our proposed DSA (purple) approach to EEP tasks. In the common EEP pipeline, a cumulative failure function estimator F_θ is trained by MLE on corresponding likelihood L_{EEP} and serves as a risk score for a threshold-based alarm policy. On the other hand, in our proposed DSA approach, using a re-organized train set into episodes, we fit a hazard function estimator λ_θ by MLE on a partial survival likelihood L^h_{DSA}. We obtain a unique risk after applying a prioritization function and aggregation function to $[F_\theta(1), ..., F_\theta(h)]$.

sample "right-censored" to align with DSA terminology. The EEP task consists of modeling the cumulative probability of this event occurring within a fixed prediction horizon h defined as $F(h|\mathbf{X}_t) = P(T_e \leq t+h|\mathbf{X}_t)$. Importantly, as EEP focuses on early warning, no prediction is carried out during events. The common approach in the EEP literature is to train models that directly parameterize the cumulative failure function F_θ by minimizing the following negative log-likelihood with labels $y_{i,t} = \mathbb{1}_{[\sum_{k=t}^{t+h} e_k \geq 1]}$:

$$L_{EEP} = \sum_i^N \sum_t^{T_i} -\Big((1-e_{i,t})[y_{i,t}\log(F_\theta(h|\,\mathbf{X}_{i,t})) + (1-y_{i,t})\log(1-F_\theta(h|\,\mathbf{X}_{i,t}))]\Big) \quad (1)$$

Dynamic survival analysis Conversely, DSA aims to model the precise time $T_e(t)$ until an event of interest occurs given observation up to t. Thus, when considering a discrete setting, to model for all horizons $k \in \mathbb{N}^*$, the mass function $f(k|\,\mathbf{X}_t) = P(T_e = t+k|\mathbf{X}_t)$. This statistical framework has the particularity of considering terminal events only. There can be at most one event per trajectory i and if not censored, it is at time $T_e = T_i + 1$. Right-censoring refers to sequences where an event hasn't been observed before last observation T_i. For this purpose, it is common to define $c_i = \mathbb{1}_{[T_i \neq T_e]}$ a right-censoring indicator. Then the DSA negative log-likelihood for a model parameterizing the PMF f_θ is defined as:

$$L_{DSA} = -\sum_{i=1}^N \sum_{t=0}^{T_i}\Big((1-c_i)f_\theta(T_i+1-t|\mathbf{X}_{i,t}) + c_i([1-\sum_{k=1}^{T_i+1-t} f_\theta(k|\mathbf{X}_i)])\Big) \quad (2)$$

It is known (Kalbfleisch and Prentice, 2011) that the survival likelihood can be re-written as binary cross-entropy over the hazard function $\lambda(k|\,\mathbf{X}_t) = P(T_e = t+k|\mathbf{X}_t, T_e > t+k-1)$. Thus, it is common in DSA to parameterize the hazard λ_θ and to minimize the following survival negative log-likelihood using binary labels $y_{i,t,k} = \mathbb{1}_{[T_i - t = k \wedge c_i = 0]}$ and sample weights $w_{i,t,k} = \mathbb{1}_{[k \leq T_i - t]}$:

$$L_{DSA} = \sum_{i=1}^N \sum_{t=0}^{T_i} \sum_{k=1}^{T_{\max}} w_{i,t,k} \Big([y_{i,t,k}\log(\lambda_\theta(k|\,\mathbf{X}_{i,t})) + (1-y_{i,t,k})\log(1-\lambda_\theta(k|\,\mathbf{X}_{i,t}))]\Big) \quad (3)$$

Given an estimate λ_θ, for a fixed h, we obtain estimates of the PMF $f_\theta(h|\,\mathbf{X}_t) = \prod_{k=1}^{h-1}(1 -$

$\lambda_\theta(k|\,\mathbf{X}_t))\lambda_\theta(h|\,\mathbf{X}_t)$ and the cumulative failure function $F_\theta(h|\,\mathbf{X}_t) = 1-\prod_{k=1}^{h}(1-\lambda_\theta(k|\,\mathbf{X}_t))$. Thus, from a hazard estimate, one can also perform EEP tasks.

Handling non-terminal events In general, events from EEP are not terminal, meaning that observations are carried out during and after an event and events can occur multiple times. To train a model using a survival analysis method for these cases, the DSA framework requires unique terminal events. To address this issue, we propose to instead only predict the occurrence of the closest event, if there is one.

In EEP tasks, timesteps within events are ignored in the likelihood. Thus, for a patient stay i experiencing v events at times $t_{e_1}, ..., t_{e_v}$, respectively ending at times $s_{e_1}, ..., s_{e_v}$, we proposed instead to consider distinct episodes $[\mathbf{X}_{i,0}, ..., \mathbf{X}_{i,t_{e_1}-1}]$,$[\mathbf{X}_{i,0}, ..., \mathbf{X}_{i,t_{e_2}-1}]$, ..., $[\mathbf{X}_{i,0}, ..., \mathbf{X}_{i,T_i}]$ associated to their respective labels $[\mathbf{y}_{i,0}, ..., \mathbf{y}_{i,t_{e_1}-1}]$,$[\mathbf{y}_{i,s_{e_2}}, ..., \mathbf{y}_{i,t_{e_2}-1}]$,...,$[\mathbf{y}_{i,s_{e_v}}, ..., \mathbf{y}_{i,T_i}]$. It is important to note that for any episode beyond the first one indexed by k, we provide the history of measurement between 0 and $s_{e_{k-1}}$ to ensure preserving the signal from previous occurrences. This procedure ensures, that in each sample, if not censored, the event occurrence is unique and at the end of the sequence, as in DSA.

3.2. Bridging the gap on timestep-level performance

In practice, as reported by Yèche et al. (2023), we show that fitting hazard deep learning models to a DSA likelihood on EEP data is unstable and underperforms on timestep metrics (see Figure 5 and Table 1). However, we find we can overcome this issue with two simple modifications to the training. First, we find that instability is due to extreme imbalance compared to EEP likelihood and propose a specific *logit bias initialization* to handle it. Second, to focus on the horizon of prediction used at inference, we propose to match EEP likelihood and *truncate DSA likelihood only until the horizon of prediction h*. Finally, we propose, survTLS our extension to TLS for survival analysis, allowing to further improve performance over DSA MLE for EEP.

Bias initialization As resolution is high in ICU data and horizons of prediction relatively short, associated tasks tend to already be imbalanced. Unfortunately, as shown in Eq 3, when fitting a hazard model, each positive label from the EEP is associated with $T_i - t - 1$ negative elements for a single positive label.

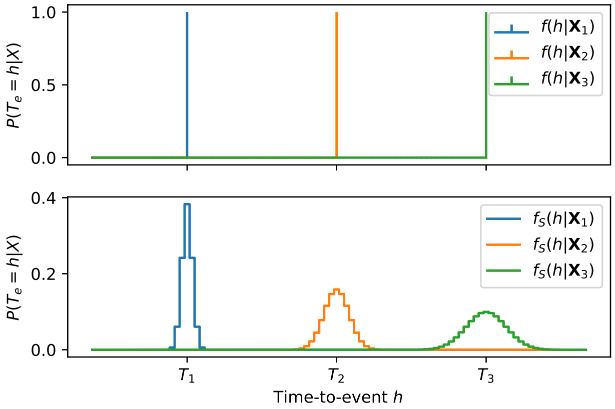


Figure 2: Illustration of **survTLS** for three non-censored samples. (Top) Ground-truth PMF; (Bottom) survTLS smoothed PMF. The further the event, the higher the entropy of the mass function around the ground-truth time-to-event.

Similarly, each negative is associated with $T_i - t$ negative labels. Hence the prevalence from the EEP task is divided by a factor $\bar{T}$ corresponding to the average sequence length. This becomes extreme in EHR data where sequences have thousands of steps. We found this to forbid convergence in certain cases.

To overcome this issue, inspired by Karpathy (2019), we propose to initialize the bias of the logit layer $\mathbf{b} = [b_1, ..., b_{T_{\max}}]$ such that the output probability $\frac{1}{1+e^{-b_k}} = \tilde{y}(k)$ with $\tilde{y}(k)$, the average hazard label value for horizon k. Thus we initialize the bias as follows:

$$b_k = \log(\frac{\tilde{y}(k)}{1-\tilde{y}(k)}), \forall k \leq T_{\max}$$

Survival likelihood truncating As shown in Table 1, correct initialization of biases already allows DSA models to be trainable for ICU data, however, they still lag behind EEP counterparts. As motivated by Yèche et al. (2023), further events are generally harder to predict due to their lower signal. We observe sequences to be longer in EEP datasets than in DSA. Indeed, PCB2 and AIDS, two commonly used datasets in DSA, have median lengths of 3 and 5 Maystre and Russo (2022), whereas MIMIC-III and HiRID have median sequence lengths of 50 and 275. Thus, when fitting a DSA likelihood, a model is trained to model event occurrence possibly much further than the fixed horizon of prediction h used in EEP. As empirically validated in Figure 5), we believe such events dominate the loss due to their hardness, forbidding the model to learn properly for events occurring within h steps. In the alarm policy, no prediction, whether it is cumulative or not, is required beyond h.

To solve this issue, because in EEP no prediction is carried beyond h, we propose to fit DSA models only until h. This translates into a truncated negative log-likelihood as follows:

$$L^h_{DSA} = \sum_{i=1}^{N}\sum_{t=0}^{T_i}\sum_{k=1}^{h} w_{i,t,k} \Big([y_{i,t,k}\log(\lambda_\theta(k|\,\mathbf{X}_{i,t})) + (1-y_{i,t,k})\log(1-\lambda_\theta(k|\,\mathbf{X}_{i,t}))]\Big) \quad (4)$$

survTLS – A temporal label smoothing approach for Survival Analysis In EEP, prior work showed the effectiveness of TLS (Yèche et al., 2023) for timestep-level performance. As a form of regularization, this method enforces EEP model certainty on their estimate $F_\theta(h|\,\mathbf{X}_t)$ to decrease with the distance to the next event, by modulating similarly label smoothing strength during training. Given its success for EEP, transferring TLS to DSA is sensible. Unfortunately, this is not straightforward to do, as in DSA we model the more granular hazard function λ_θ over all horizons with the constraint that $\sum_h f_\theta(h) = 1$.

In our extension survTLS, we propose to leverage this higher granularity in labels, not to control the certainty of event occurrence, as in TLS, but rather to control the certainty of the event localization based on its distance.

For this purpose, as shown in Figure 2, we replace the (hard) ground truth PMF vector $\mathbf{f}_{i,t} = [f(h|\,\mathbf{X}_{i,t}) = \mathbb{1}_{[T_i-t=h\wedge c_i=0]}]_{h\geq 1}$ by a smooth version f_S. Given a continuous distribution $g_i \sim \mathcal{N}(T_i, (\frac{T_i}{l})^2)$ and G_i its cumulative distribution, we define f_S as its discretization :

$$f_S(h|\mathbf{X}_i) = \begin{cases} 0 & \text{if } c_i = 1 \\ G_i(h+\frac{1}{2}) & \text{elif } h = 1 \\ G_i(h+\frac{1}{2}) \\ \quad -G_i(h-\frac{1}{2}) & \text{elif } h \in [2, T_{\max}-1] \\ 1 - G_i(h-\frac{1}{2}) & \text{elif } h = T_{\max} \end{cases}$$

The lengthscale hyperparameter l control the strength of the smoothing. It was selected on validation metrics and more details can be found in Appendix A. Note that we preserve $\sum_h f_S(h|\mathbf{X_{i,t}}) =$ 1. Following discrete survival analysis definitions, we can define the smooth survival function $S_S(h|\mathbf{X}_{i,t}) = 1 - \sum_{k=1}^{h} f_S(h|\mathbf{X}_{i,t})$ and the hazard function $\lambda_S(h|\mathbf{X}_{i,t}) = \frac{f_S(h|\mathbf{X}_{i,t})}{S_S(h-1|\mathbf{X}_{i,t})}$.

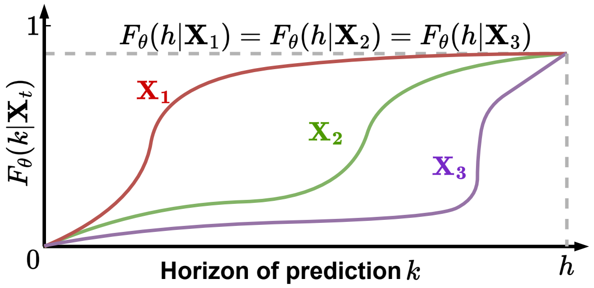


Figure 3: Localization problem when estimating failure function F_θ. If only provided with an estimate at h, as in EEP modeling, alarms would be raised similarly for $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_3$. If on the other hand provided with earlier estimates like depicted, as in DSA modeling, an event for $\mathbf{X}_1$ is more probable to happen earlier than $\mathbf{X}_2$ and $\mathbf{X}_3$. Because of the imminence of the risk, we argue that given that information, $\mathbf{X}_1$ should have a higher risk score for the alarm policy. We further develop this idea in Section 3.3

By identifying w_{ij} and y_{ij} in Eq. 3 as respectively the ground-truth survival $S(j|\mathbf{X})$ and hazard probability $\lambda(j|\mathbf{X})$ as proposed by Maystre and Russo (2022), we define our survTLS objective as follows:

$$L_{survTLS} = \sum_{i=1}^{N}\sum_{t=0}^{T_i}\sum_{k=1}^{h} S_S(k|\,\mathbf{X}_{i,t}) \Big([\lambda_S(k|\,\mathbf{X}_{i,t}))\log(\lambda_\theta(k|\,\mathbf{X}_{i,t})) + (1-\lambda_S(k|\,\mathbf{X}_{i,t}))\log(1-\lambda_\theta(k|\,\mathbf{X}_{i,t}))]\Big) \quad (5)$$

We show the new labels y and weights w obtained from survTLS in Figure 8 and Figure 9 that can be found in Appendix A.2

3.3. Leveraging Hazard Estimation for Alarm Policy

Threshold-based policy In the literature, though crucial to clinical adoption, most works do not explore model performance in terms of alarms for whole events and rather focus on timestep modeling. This leads the current state-of-the-art alarm policy to still be straightforward. First, Tomašev et al. (2019) proposed to define a working threshold $\tau \in [0,1]$ selected based on a timestep precision constraint. Then, the alarm policy raises alarm $a_t \in 0, 1$ at any timestep where risk score $s_t = F_\theta(h|\,\mathbf{X}_t)$ is above τ :

$$a_t = \mathbb{1}_{F_\theta(h|\,\mathbf{X}_t) \geq \tau} \tag{6}$$

Silencing policy Later, to reduce the false alarm rate, Hyland et al. (2020) introduced the concept of silencing. After a raised alarm following Eq. 6. the system *silences* all subsequent alarms until the duration σ of the silencing time has passed. This was then adopted in subsequent works on EWS (Hüser et al., 2024; Lyu et al., 2024) If we define d_t^a to be the distance to the last alarm at timestep t, the silenced alarm policy is defined as follows:

$$a_t = \mathbb{1}_{F_\theta(h|\,\mathbf{X}_t) \geq \tau} \mathbb{1}_{d_t^a \geq \sigma} \tag{7}$$

It is important to note that silencing is applied regardless of the correctness of the alarm. Indeed, EEP tasks are prognosis tasks, hence contrary to diagnosis tasks, the veracity of prediction is not verifiable until the event occurs.

A straightforward approach to using a DSA model is to extract the failure estimate $F_\theta(h|\,\mathbf{X}_t)$ and apply the same alarm policy. We refer to this approach as "Fixed horizon". However, our motivation to estimate the more challenging hazard function is to have as a counterpart an estimate on the localization of the risk within horizon h for the alarm policy design. Given a risk estimate vector $\mathbf{r}_t = [F_\theta(1|\,\mathbf{X}_t), ..., F_\theta(h|\,\mathbf{X}_t)]$, we formalize a mechanism to raise alarms depending on a unique functioning threshold compatible with the different event metric definitions.

Imminent prioritization policy We follow the same intuition as Yèche et al. (2023) during training and Heitz et al. (2021) for evaluation, to favor more imminent events given similar risk at horizon h. Indeed, similarly to these prior work, we argue that impending events should be acted on immediately and are less likely to be impacted by a competitive event, thus they should have a higher priority than events equally probable but at a further horizon.

We formalize this intuition by introducing a priority function p and transforming the output of the DSA model to create a score vector $\mathbf{s}_t \in [0,1]^h$ from the output vector $[[F_\theta(1|\,\mathbf{X}_t), ..., F_\theta(h|\,\mathbf{X}_t)]]$ as follows:

$$\begin{aligned} \mathbf{s}_t &= [s_1, \ldots, s_h] \\ &= [p(F_\theta(1|\,\mathbf{X}_t), 1), \ldots, p(F_\theta(h|\,\mathbf{X}_t), h)] \end{aligned} \tag{8}$$

This step allows us to rescale risk predictions to a joint scale, from which we aggregate risk scores into a single alarm, using a single working threshold τ, by defining a_t as follows:

$$a_t = \mathbb{1}_{(\sum \mathbb{1}_{s_k > \tau}) > 0} \mathbb{1}_{d_t^a \geq \sigma} \tag{9}$$

Additionally we can define an estimate of the time-to-event $d_t = \min_k [k \mid s_k > \tau]$.

To enforce a prioritization of the closer horizons of prediction, we simply have to enforce, p to be monotonically decreasing. We choose to implement the priority function with an exponential decay (Yèche et al., 2023):

$$p(F, k) = q^{exp}(k) \cdot F \tag{10}$$

where the exponential decay function $q^{exp}(k)$ is defined as follows:

$$q(t) = \begin{cases} 0 & \text{if } k > h_{\max} \\ e^{-\gamma(k-d)} + A & \text{if } k \leq h_{\max} \end{cases} \tag{11}$$

where

$$A = -e^{-\gamma(h_{\max} - d)} \tag{12}$$

$$d = -\frac{1}{\gamma} ln(1 - e^{-\gamma h_{\max}}) \tag{13}$$

As shown in Figure 4, $h_{\max}$ controls the intercept with 0 and γ the strength of the decay. Hence, for any prediction beyond $h_{\max}$, the risk score is scaled to 0. The two hyperparameters γ and $h_{\max}$ are tuned on the validation set Alarm/Event AuPRC and their value for different tasks is provided in Appendix A.

4. Experimental Set-up

Tasks We perform experiments on three EEP tasks, early prediction of *circulatory failure*, *mechanical ventilation* and *decompensation* on established benchmarks from HiRID (Yèche et al., 2021) and MIMIC-III (Harutyunyan et al., 2019). Both circulatory failure and ventilation are predicted at a 12-hour horizon with a 5-minute resolution, while decompensation is predicted at 24 hours with a 1-hour resolution. Further details about tasks and dataset can be found in Appendix A.1

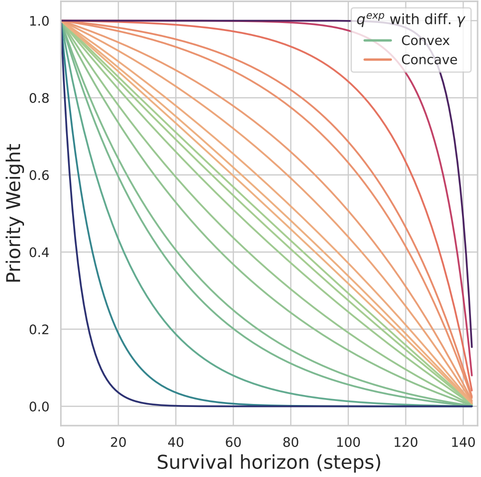


Figure 4: Visualizations of the q_{exp} exponential decay function used for prioritization of survival risk scores. The plot shows $h_{\max} = h = 144$, and different γ for the convex and a concave version, which we didn't find to work better.

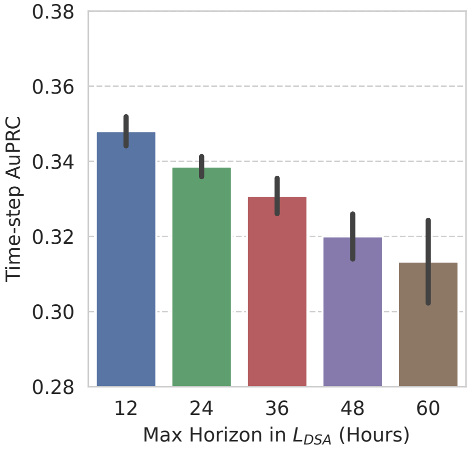


Figure 5: Ablation on the maximum considered horizon in L_{DSA} for the ventilation task. By not training the DSA model further than h, the model's estimation of $F_\theta(h)$ improves leading to a better timestep AuPRC.

Implementation details For all tasks, models are composed of a linear time-step embedding with $\mathcal{L}_1$-regularization (Tomašev et al., 2019) coupled to a GRU (Chung et al., 2014) backbone. All hyperparameters shared across methods were selected through validation performance (AuPRC) for the EEP model and then used for all methods. Specific parameters to each method, such as priority strength, are selected on individual validation performance. Further details about implementation can be found in Appendix A.2. As discussed in Section 2, existing works have proposed specific improvements to either EEP likelihood, with auxiliary regression (Tomašev et al., 2019) terms, or DSA likelihood with a ranking term (Lee et al., 2018; Jarrett et al., 2019). Both EEP and DSA likelihoods being versatile objectives, these extensions can be seamlessly incorporated for further applications. Hence, we focus on comparing directly likelihood objectives alone. We consider a model parameterizing the cumulative failure function at horizon h fitted by MLE over the EEP likelihood and a model parameterizing the hazard function fitted by MLE over the DSA likelihood.

Evaluation At a time-step level, to evaluate the goodness of the cumulative failure function estimate F_θ across models, we follow a common approach for highly imbalanced tasks, with the area under the precision-recall curve (AuPRC). We report the Alarm/Event AuPRC for event-level metrics, as defined by Hyland et al. (2020). To measure the timeliness of alarms, we also report the distance of the first alarm and event recall at high alarm precision thresholds. The high precision region represents a clinically relevant setting with a lower risk of alarm fatigue (Tomašev et al., 2019; Yèche et al., 2023). Unless stated otherwise, all results are reported in the form *mean* ± *95% CI* of the standard error across 10 runs. In the following section, we describe the event-level metrics we used.

4.1. Event-level Evaluation

Tomašev et al. (2019); Yèche et al. (2023); Moor et al. (2023) note the relevance of reporting event-oriented performance such as event recall or precision and distance to event at a fixed sensitivity level. To get a general event-level evaluation across different working thresholds, Hyland et al. (2020) define event-based precision-recall curve mapping event recall to an effective alarm precision. We detail the event metric definitions below

Event Recall Proposed by Tomašev et al. (2019), given binary timestep predictions a_t (for alarms), event recall R_{event} corresponds to the true positive rate over event predictions. An event E at time t_E is detected if any a_t is positive between $t_E - h$ and

$t_e - 1$. Hence the following definition:

$$R_{event} = \frac{\sum_E \mathbb{1}_{(\sum_{t_E-h}^{t_E-1} a_t)>0}}{\#Events}$$

Alarm Precision Proposed by Hyland et al. (2020), given binary timestep predictions a_t (for alarms), alarm precision P_{alarm} is the proportion of alarms raised located within h of a event. It is defined as follows:

$$P_{event} = \frac{\sum_E (\sum_{t_E-h}^{t_E-1} a_t)}{\sum_t a_t}$$

Note, that considering a single event, there can be multiple true alarms for it as long as they fall within the prediction horizon.

In our work, the first event metric we report is the **Event Recall @ Alarm Precision** for high precision thresholds corresponding to clinically applicable regions.

Distance to Event Because event recall does not capture the earliness of the alarm, Hyland et al. (2020) also proposes to report the distance for the first alarm for an event D_{fa}. Formally for an event event E at time t_E this is defined as:

$$D_{fa} = \max_{k \in [t_E-h, t_E-1]} [t_E - k | a_k = 1]$$

Alarm/Event AuPRC Finally, to not depend on binary predictions, Hyland et al. (2020) propose a curve score where alarms $\mathbf{a}$ directly depend on a working threshold $\tau \in [0,1]$ where a point $(x, y) = \big(R_{event}(\tau), P_{alarm}(\tau)\big)$ on the curve. The final score is the area under this curve. As for timestep predictions, this metric gives a global overview of the model performance at different thresholds but on the granularity of events.

5. Results

Time-step level As shown in Table 1, we find that vanilla survival models are significantly worse than their EEP counterparts going as far as not converging on the decompensation task. However, fixing bias initialization, as proposed in Section 3.2, allows convergence for all tasks. Additionally, by truncating the DSA likelihood up to h (Section 3.2), we managed to close the gap and even surpass in some cases MLE with EEP likelihood at a time-step level.

Table 1: Time-Step AuPRC. "EEP" and "Survival" refer to training with the respective vanilla likelihoods. Time-step level metrics are useful to assess a machine learning predictor's generalization performance on the trained task but do not explicitly quantify performance on clinically meaningful entities such as alarms and events.

Task	HiRID		MIMIC
	Circ.	*Vent.*	*Decomp.*
EEP	39.0 ± 0.4	34.3 ± 0.3	37.1 ± 0.6
+ TLS	**40.5** ± **0.4**	**34.9** ± **0.4**	37.2 ± 0.3
Survival	37.4 ± 2.0	21.5 ± 7.5	N.A
+ Bias Init.	39.2 ± 0.3	31.3 ± 0.8	36.2 ± 0.4
+ Limit Horizon	**40.7** ± 0.2	34.5 ± 0.4	37.4 ± 0.6
+ survTLS	**40.7** ± **0.2**	**34.9** ± **0.3**	**38.4** ± **0.2**

Table 2: Area under the Alarm Precision / Event Recall Curve (Hyland et al., 2020). Different from timestep metrics (Table 1), this metric evaluates the output of the alarm policy used over the model's timestep risk estimates. "Survival" refers to training with bias initialization and truncated survival likelihood L^h_{DSA}. We observe that an alarm policy prioritizing imminent events can notably improve the event-level performance.

Task	HiRID		MIMIC
	Circ.	*Vent.*	*Decomp.*
EEP	66.0±2.3	61.6±0.3	67.4±1.2
+ TLS	76.4±1.1	64.7±0.3	68.4±0.5
Survival	75.2±0.3	64.2±0.8	70.0±0.1
+ survTLS	77.5±0.6	64.8±1.5	70.9±0.3
+ Imminent prio.	**79.5±0.1**	**66.7±1.4**	**71.3±0.5**

Event level As shown in Table 2, we find that survival models, given the prior improvements, already outperform EEP when used with a similar alarm policy with a fixed threshold over $F_\theta(h)$. When additionally introducing a priority function favoring imminent events, this gap further increases by 1 to 3%. Finally, as shown in Table 3 for decompensation, our prioritization of more imminent events does not come at the cost of event recall nor distance to the event as our policy still matches or outperforms the base policy in both metrics. Similar conclusions can be drawn for HiRID tasks (Appendix B).

Finally, in Figures 6 and 7 we show the full performance curve over all operating thresholds $\tau \in [0, 1]$.

Table 3: Event Recall and Mean Distance (in hours) of the first alarm to the event start on MIMIC-III Decompensation at fixed alarm precisions of 60%, 70%, 80%. In Table 2 we provide area under the curve results to assess the event-level performance across operating thresholds τ. Here, to mimic deployment scenarios, we use a high precision constraint to fix a threshold on the risk score that has to be chosen for the alarm policy. Our method can provide noticeable improvements in event recall while maintaining a similar detection distance to the event.

Metric	Event Recall			Mean Dist. (h)		
	@ 60% P.	*@ 70% P.*	*@ 80% P.*	*@ 60% P.*	*@ 70% P.*	*@ 80% P.*
EEP	66.7±0.0	60.9±0.1	52.4±0.0	6.2±0.1	4.9±0.1	3.4±0.1
+ TLS	68.1±0.0	62.4±0.0	54.8±0.0	6.5±0.0	**5.2±0.0**	3.7±0.0
Survival	68.5±0.0	63.2±0.0	57.5±0.0	6.4±0.1	5.1±0.1	**4.0±0.1**
+ survTLS	68.8±0.1	64.1±0.1	58.5±0.0	6.4±0.1	**5.2±0.1**	**4.0±0.1**
+ Imminent prio.	**70.3±0.1**	**64.6±0.0**	**59.9±0.0**	**6.6±0.1**	5.1±0.1	**4.0±0.1**

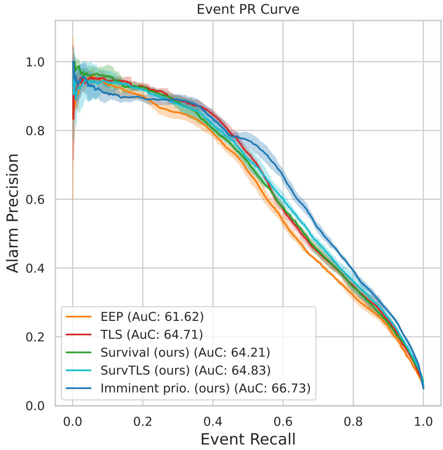


Figure 6: Alarm Precision (Hyland et al., 2020) per Event Recall on HiRID Ventilation. We show that our proposed DSA-based approach paired with a prioritizing alarm policy maintains a higher alarm precision for higher event recall regions.

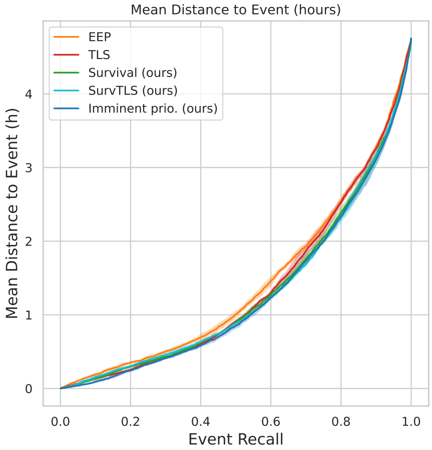


Figure 7: Mean distance of the first alarm to the event per Event Recall on HiRID Ventilation. It highlights that the improved event recall and alarm precision performance does come at a very small to no cost in terms of detection distance to the event.

The curves show alarm precision and mean distance of the first alarm over event recall on the HiRID Ventilation task, the two other benchmarked tasks are shown in Appendix B.

6. Limitation and Further Work

The focus of our study is to compare two distinct, yet related, modeling approaches on their ability to predict the likelihood of a future event. In an isolated comparison of the probabilistic modeling approach, our research establishes that by training a richer DSA model, one can successfully perform established EEP tasks. However, it also has some clear limitations. As mentioned in Section 2, multiple works have proposed to complement both objectives with regularization terms. If interested in absolute performance for a specific application, one should ensure that DSA improvements remain compatible with such extensions. Also, DSA likelihood suffers from a much higher class imbalance that grows with resolution. While we manage to tackle this issue in our tasks of interest, it is still likely that for other EEP tasks, this instability might remain. Also note that, as we focused on EEP applications, the notion of competitive risk is not considered (a setting where we aim to predict different types of events with a joint modeling approach). Extending to multiple risks constitutes a non-trivial, but highly relevant future work.

Furthermore, we highlight how the richer DSA output can be leveraged to enhance clinically relevant and deployment-oriented event-level performance metrics by prioritizing imminent events. However, our approach is straightforward and remains a first step. We hypothesize that this richer output can be further utilized to create more sophisticated alarm policies, which can be tuned to be more patient-specific by incorporating the shape of the individual estimated survival distributions at a given time point for a given patient. Additionally, this approach can provide clinicians with richer insights and explanations. We reserve this direction for further work.

7. Conclusion

In this work, we investigate the usage of DSA models for EEP tasks motivated by the additional localization of the risk they provide. We show that even though it is more challenging to train, with careful initialization and partial survival likelihood fitting, DSA models can be competitive at a time-step level. Further, we show that our simple prioritization scheme for alarms allows DSA models to outperform EEP counterparts even further. Our proposed prioritization transformation is a first step towards tailored alarm policies. Future work remains to further leverage risk localization to design more sophisticated alarm policies and provide a richer output to clinicians.

In the past many application-oriented works have chosen to use classical EEP modeling approaches for their studies due to higher stability and ease of training. We hope our analysis paves the way for DSA, a more comprehensive modeling approach, to become the prominent choice.

References

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

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.

Francesca Di Giuseppe, Florian Pappenberger, Fredrik Wetterhall, Blazej Krzeminski, Andrea Camia, Giorgio Libertá, and Jesus San Miguel. The potential predictability of fire danger provided by numerical weather prediction. *Journal of Applied Meteorology and Climatology*, 55(11):2469–2491, 2016.

M. Faltys, M. Zimmermann, X. Lyu, M. Hüser, S. Hyland, G. Rätsch, and T. Merz. Hirid, a high time-resolution icu dataset, 2021a.

Martin Faltys, Marc Zimmermann, Xinrui Lyu, Matthias Hüser, Stephanie Hyland, Gunnar Rätsch, and Tobias Merz. Hirid, a high time-resolution icu dataset (version 1.1. 1). *Physio. Net*, 10, 2021b.

Joseph Futoma, Sanjay Hariharan, and Katherine Heller. Learning to detect sepsis with a multitask gaussian process rnn classifier. In *International conference on machine learning*, pages 1174–1182. PMLR, 2017.

Michael F Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

Jonathan Heitz, Joanna Ficek, Martin Faltys, Tobias M Merz, Gunnar Rätsch, and Matthias Hüser. Wrse-a non-parametric weighted-resolution ensemble for predicting individual survival distributions in the icu. In *Survival Prediction-Algorithms, Challenges and Applications*, pages 54–69. PMLR, 2021.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, pages 4353–4363. PMLR, 2020.

Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

Matthias Hüser, Xinrui Lyu, Martin Faltys, Alizée Pace, Marine Hoche, Stephanie Hyland, Hugo Yèche, Manuel Burger, Tobias M Merz, and Gunnar Rätsch. A comprehensive ml-based respiratory monitoring system for physiological monitoring & resource planning in the icu. *medRxiv*, 2024. doi: 10.1101/2024.01.23.24301516. URL https://www.medrxiv.org/content/early/2024/01/23/2024.01.23.24301516.

Daniel Jarrett, Jinsung Yoon, and Mihaela van der Schaar. Dynamic prediction in clinical survival analysis using temporal convolutional networks. *IEEE journal of biomedical and health informatics*, 24(2):424–436, 2019.

Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. MIMIC-III clinical database (version 1.4), 2016a.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016b. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data.* John Wiley & Sons, 2011.

Andrej Karpathy. A recipe for training neural networks. *Personal Blog*, 2019. URL https://karpathy.github.io/2019/04/25/recipe/.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger.

Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *arXiv preprint arXiv:1907.00825*, 2019.

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*, volume 32, 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.

Xinrui Lyu, Bowen Fan, Matthias Hüser, Philip Hartout, Thomas Gumbsch, Martin Faltys, Tobias M. Merz, Gunnar Rätsch, and Karsten Borgwardt. An empirical study on kdigo-defined acute kidney injury prediction in the intensive care unit. *medRxiv*, 2024. doi: 10.1101/2024.02.01.24302063. URL https://www.medrxiv.org/content/early/2024/02/03/2024.02.01.24302063.

Lucas Maystre and Daniel Russo. Temporally-consistent survival analysis. *Advances in Neural Information Processing Systems*, 35:10671–10683, 2022.

Michael Moor, Nicolas Bennett, Drago Plečko, Max Horn, Bastian Rieck, Nicolai Meinshausen, Peter Bühlmann, and Karsten Borgwardt. Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *EClinicalMedicine*, 62:102124, August 2023.

Layla Parast, Lu Tian, and Tianxi Cai. Landmark estimation of survival and treatment effect in a randomized clinical trial. *J Am Stat Assoc*, 109(505): 384–394, January 2014.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. Deep recurrent survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4798–4805, 2019.

Matthew A Reyna, Christopher S Josef, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Shamim Nemati, Gari D Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Critical care medicine*, 48(2):210–217, 2020.

Xiaobin Shen, Jonathan Elmer, and George H. Chen. Neurological prognostication of post-cardiac-arrest coma patients using eeg data: A dynamic survival analysis framework with competing risks. In Kaivalya Deshpande, Madalina Fiterau, Shalmali Joshi, Zachary Lipton, Rajesh Ranganath, Iñigo Urteaga, and Serene Yeung, editors, *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pages 667–690. PMLR, 11–12 Aug 2023. URL https://proceedings.mlr.press/v219/shen23a.html.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17, 2020.

Patrick J Thoral, Jan M Peppink, Ronald H Driessen, Eric JG Sijbrands, Erwin JO Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, et al. Sharing icu patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: the amsterdam university medical centers database (amsterdamumcdb) example. *Critical care medicine*, 49 (6):e563, 2021.

Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

Gerhard Tutz, Matthias Schmid, et al. *Modeling discrete time-to-event data*. Springer, 2016.

Robin van de Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another ICU benchmark: A flexible multi-center framework for clinical ML. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ox2ATRM90I.

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

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235, 2020.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. Hirid-icu-benchmark–a comprehensive machine learning benchmark on high-resolution icu data. *arXiv preprint arXiv:2111.08536*, 2021.

Hugo Yèche, Alizée Pace, Gunnar Ratsch, and Rita Kuznetsova. Temporal label smoothing for early event prediction. *ICML*, 2023.

Safoora Yousefi, Fatemeh Amrollahi, Mohamed Amgad, Chengliang Dong, Joshua E Lewis, Congzheng Song, David A Gutman, Sameer H Halani, Jose Enrique Velazquez Vega, Daniel J Brat, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1):11707, 2017.

Panpan Zheng, Shuhan Yuan, and Xintao Wu. Safe: A neural survival analysis model for fraud early detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1278–1285, Jul. 2019. doi: 10.1609/aaai.v33i01.33011278. URL https://ojs.aaai.org/index.php/AAAI/article/view/3923.

Appendix A. Experimental details

A.1. Datasets

Table 4: *Label and event prevalence statistics* computed on the training set for all tasks.

Task	Positive timesteps (%)	Patients undergoing event (%)	Number of events per positive patient
Circulatory Failure (HiRID)	4.3	25.6	1.9
Mechanical Ventilation (HiRID)	5.6	56.5	1.5
Decompensation (MIMIC)	2.1	8.3	1.0

Circulatory failure Online binary prediction of future circulatory failure events on the HiRID (Faltys et al., 2021b) dataset as defined by Yèche et al. (2021) every 5 minutes. The benchmarked prediction horizon for the EEP models and the survival model at a fixed horizon are set at 12 hours.

Ventilation Online binary prediction of future ventilator usage on the HiRID (Faltys et al., 2021b) dataset every 5 minutes. The ventilation status is extracted from the data as defined by Yèche et al. (2021) and the prediction horizon is 12 hours.

Decompensation Online binary prediction of patient mortality as defined by Harutyunyan et al. (2019). A label is positive if the patient dies within the horizon. Benchmarked and evaluated on MIMIC-III (Johnson et al., 2016b) at a 24-hour horizon.

A.2. Implementation

Training details. For all models, we set the batch size to 64 and the learning rate to $1e^{-4}$ using Adam optimizer Kingma and Ba (2017). We early-stop each model training according to their validation loss when no improvement was made after 10 epochs.

Libraries. An exhaustive list of libraries and their version we used is in the `environment.yml` file from the code release.

Infrastructure. We trained all models on a single `NVIDIA RTX2080Ti` with 8 `Xeon E5-2630v4` cores and 64GB of memory. Individual seed training took between 3 to 10 hours for each run.

Timestep modeling hyperparameters We used the same architecture and shared hyperparameters for both types of likelihood training. These were selected based on the validation performance of EEP likelihood training. It is possible further improvement can be achieved by selecting different hyperparameters for our DSA approach. However, we prefer to be conservative to ensure a fair comparison. Exact parameters are reported in Table 6, Table 7, Table 8.

Temporal label smoothing hyperparameters Similarly we selected hyperparameters for TLS and survTLS on validation set timestep AUPRC. We found similar hyperparameters as the original paper for TLS with $h_max = 2h$ and $h_min = 0$ and $\gamma_{circ} = 0.2$, $\gamma_{vent} = 0.1$, and $\gamma_{decomp} = 0.05$. For survTLS, we found for the lengthscale parameter that $l_{circ} = 10$, $l_{vent} = 50$, $l_{decomp} = 8$. We plot the obtained labels and weights from smoothing the groud-truth event PMF f corresponding to smooth hazard function λ_S and survival function S_S in Figure 8 and Figure 9.

Alarm policy hyperparameter We find that a short one-step silencing actually performs the best (5 minutes on HiRID and 1 hour on MIMIC-III) based on validation set performance for the EEP model and keep that constant across all experiments also for the survival models.

For the prioritized alarm policy we tune h_{max}, γ, and q^{exp} function type for each task on the validation set. Chosen values are shown in Table 5.

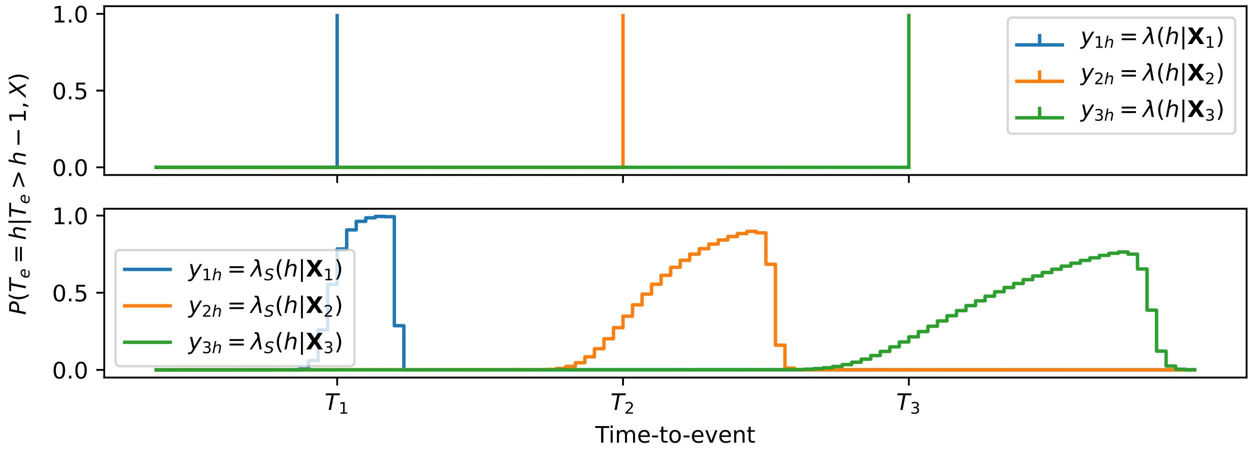


Figure 8: Illustration of the smooth hazard function λ_S serving as labels in **survTLS**.

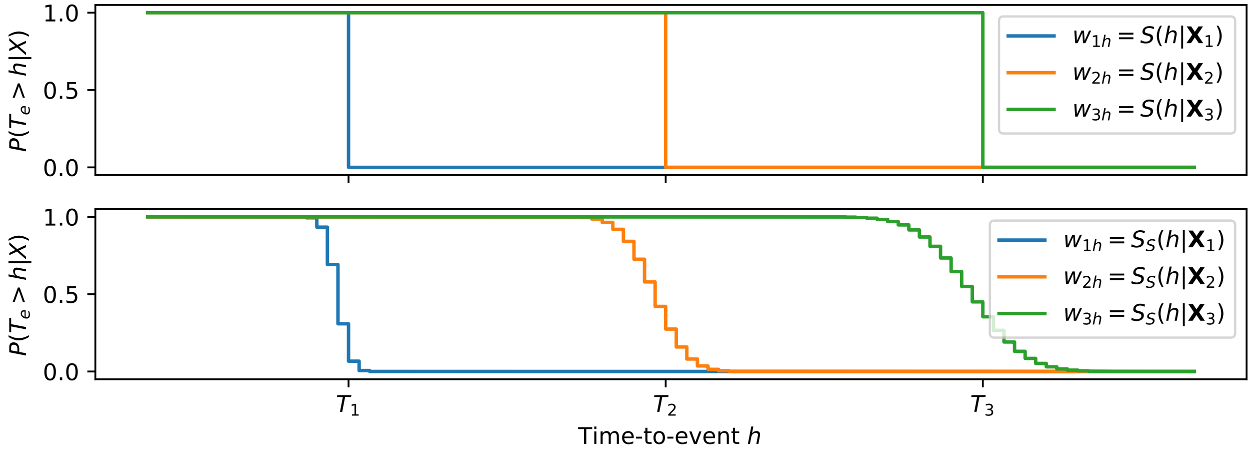


Figure 9: Illustration of the smooth survival function $\S_S$ serving as weights in **survTLS**.

Table 5: Hyperparameter search range for prioritized alarm policies on top of DSA models. In **bold** are parameters selected by grid search.

Hyperparameter	Decomp	Circ.	Vent.
h_{max}	(**12**, 24, 36, 48, 96)	**144**:[72,720]	**576**:[72,720]
γ	**0.1**:[0.01,2.0]	**0.5**:[0.01,2.0]	**2.0**:[0.01,2.0]
Function Type	(**Convex**, Concave)	(**Convex**, Concave)	(**Convex**, Concave)

Appendix B. Additional Results

Event Recall and Distance to Event In Tables 9 and 10 we show event recall and mean distance to events at different alarm precision levels for HiRID Ventilation and Circulatory Failure respectively. As noted before in the main manuscript, also on the HiRID dataset we can improve event performance while maintaining (or even slightly improving) on the distance of the first alarm to the event.

Table 6: Hyperparameter search range for *circulatory failure*, In **bold** are parameters selected by grid search.

Hyperparameter	Values
Drop-out	(**0.0**, 0.1, 0.2, 0.3)
Depth	(1, **2**, 3, 4)
Hidden Dimension	(64, 128, **256**, 512)
L1 Regularization	(1e-1, 1, **10**, 100)

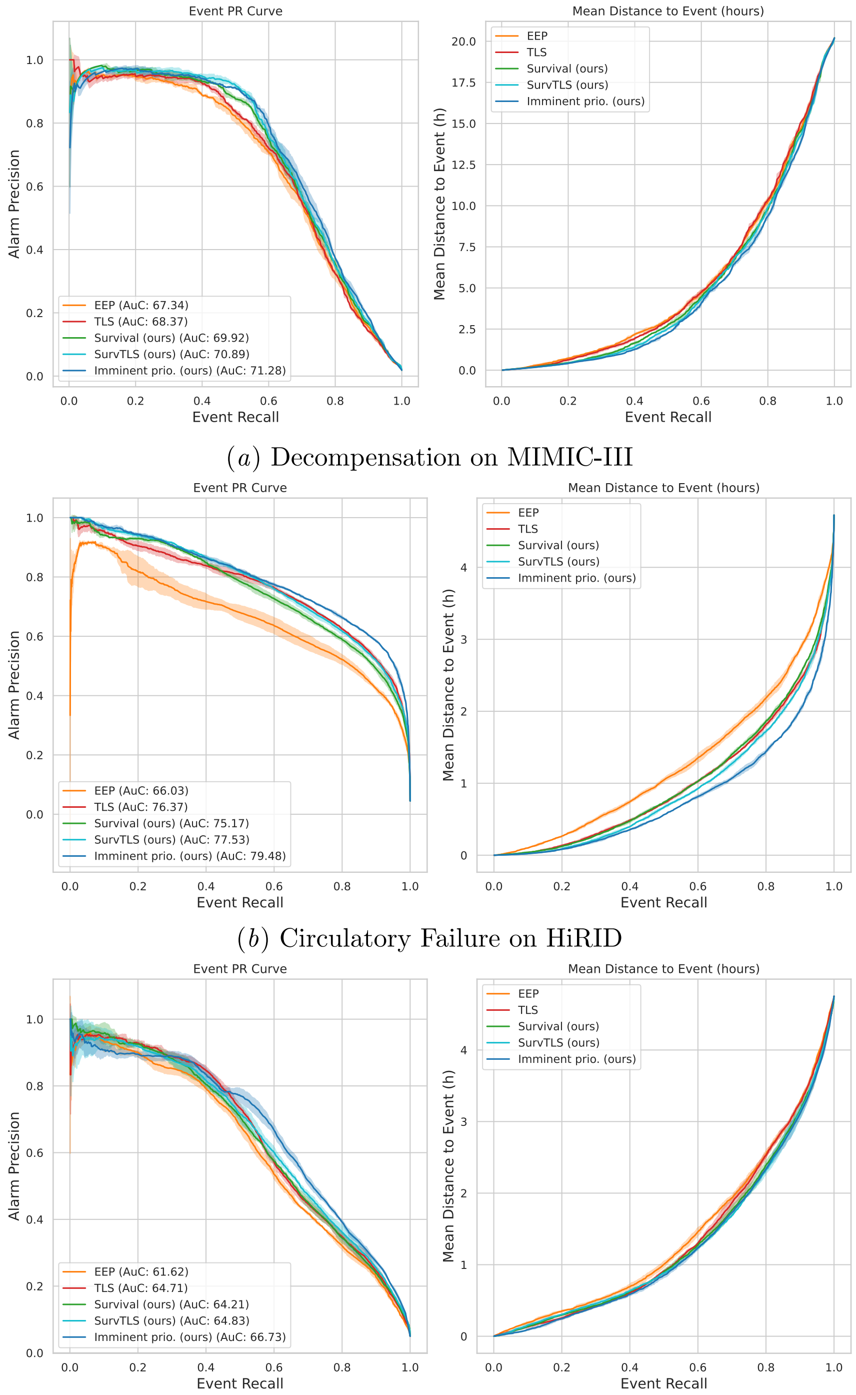
Table 7: Hyperparameter search range for *mechanical ventilation*, In **bold** are parameters selected by grid search.

Hyperparameter	Values
Drop-out	(**0.0**, 0.1, 0.2, 0.3)
Depth	(1, 2, **3**, 4)
Hidden Dimension	(128, 256, **512**,1024)
L1 Regularization	(1e-1, 1, **10**, 100)

Table 8: Hyperparameter search range for *decompensation*, In **bold** are parameters selected by grid search.

Hyperparameter	Values
Drop-out	(0.0, 0.1, **0.2**, 0.3)
Depth	(2, 3, **4**, 5)
Hidden Dimension	(128, **256**, 512 ,1024)
L1 Regularization	(1e-1, **1**, 10, 100)

Event Performance Curves In Figure 10 we show alarm precision and mean distance to event curves plotted over levels of event recall (sensitivity of the alarm policy conditioned on a risk predictor).



(a) Decompensation on MIMIC-III

(b) Circulatory Failure on HiRID

(c) Ventilation on HiRID

Figure 10: Event-based performances of the alarm policy under different models. We show the event-based alarm precision (Hyland et al., 2020) and also plot the mean distance of the first alarm to the event horizon at each event detection sensitivity level.

Table 9: Event Performance on HiRID Circulatory Failure at fixed alarm precisions.

Metric	Event Recall			Mean Dist. (h)		
	@ 60% P.	*@ 70% P.*	*@ 80% P.*	*@ 60% P.*	*@ 70% P.*	*@ 80% P.*
EEP	67.1±0.0	46.6±0.0	24.5±0.1	1.61±0.04	0.93±0.02	0.36±0.02
+ TLS	83.3±0.0	70.9±0.0	52.8±0.0	**1.99±0.04**	**1.40±0.01**	**0.81±0.02**
Survival	79.0±0.0	65.0±0.0	48.2±0.0	1.81±0.02	1.19±0.02	0.69±0.01
+ survTLS	82.5±0.0	69.5±0.0	54.4±0.0	1.85±0.02	1.26±0.02	0.78±0.02
+ Imminent prio.	**88.7±0.0**	**75.0±0.0**	**55.3±0.0**	1.91±0.03	1.24±0.04	0.70±0.00

Table 10: Event Performance on HiRID Ventilation at fixed alarm precisions.

Metric	Event Recall			Mean Dist. (h)		
	@ 60% P.	*@ 70% P.*	*@ 80% P.*	*@ 60% P.*	*@ 70% P.*	*@ 80% P.*
EEP	55.7±0.0	49.0±0.0	39.7±0.0	1.25±0.04	0.96±0.04	0.69±0.03
+ TLS	58.2±0.1	52.9±0.0	44.7±0.0	1.23±0.01	1.01±0.00	**0.73±0.01**
Survival	58.2±0.0	50.8±0.0	41.2±0.0	1.20±0.03	0.93±0.03	0.66±0.01
+ survTLS	60.3±0.0	52.3±0.0	42.7±0.0	1.24±0.02	0.94±0.01	0.69±0.01
+ Imminent prio.	**64.5±0.1**	**58.0±0.1**	**43.8±0.0**	**1.43±0.02**	**1.14±0.03**	0.66±0.03