Temporal Label Smoothing for Early Event Prediction

Hugo Yèche [*1] **Alizée Pace** [*1 2 3] **Gunnar Rätsch** [+1 2 4] **Rita Kuznetsova** [+1]

Abstract

Models that can predict the occurrence of events ahead of time with low false-alarm rates are critical to the acceptance of decision support systems in the medical community. This challenging task is typically treated as a simple binary classification, ignoring temporal dependencies between samples, whereas we propose to exploit this structure. We first introduce a common theoretical framework unifying dynamic survival analysis and early event prediction. Following an analysis of objectives from both fields, we propose Temporal Label Smoothing (TLS), a simpler, yet best-performing method that preserves prediction monotonicity over time. By focusing the objective on areas with a stronger predictive signal, TLS improves performance over all baselines on two large-scale benchmark tasks. Gains are particularly notable along clinically relevant measures, such as event recall at low false-alarm rates. TLS reduces the number of missed events by up to a factor of two over previously used approaches in early event prediction.

1. Introduction

Early event prediction (EEP) is a time-series task concerned with determining whether an event will occur within a fixed time horizon. Key to safety-critical operations such as environmental monitoring [1], EEP is also highly relevant to clinical decision-making, where the deployment of in-patient risk stratification models can significantly improve patient outcomes and facilitate resource planning [2]. For instance, the National Early Warning Score (NEWS), a simple rule-based model predicting acute deterioration in critical care units, has been demonstrated to reduce in-patient mortality [3; 4]. Deteriorating patient signals are often identified by mining large quantities of existing medical data and associated patient outcomes, which has sparked a growing interest in machine learning and medical literature. Applications of such adverse event prediction models include alarm systems for delirium [5], septic shock [6], as well as circulatory or kidney failure in the intensive care unit (ICU) [7; 8].

Still, prediction systems often suffer from high false-alarm rates with limited usefulness in a practical context [2], despite the development of deep learning architectures addressing issues of high dimensionality, irregular sampling, or informative missingness in time-series [6; 8; 9; 10]. The typically rare occurrence and noisy definition of events of interest induce challenging, highly imbalanced datasets for model training [8], yet early event prediction remains largely considered as a simple binary classification task [7; 11; 9; 12].

In this work, we systematically study different choices of objective functions for this task and outline a novel, simple, yet best-performing approach to early event prediction. In particular, we argue that leveraging the temporal structure of early event prediction is critical to improving model performance. The dynamic survival analysis (DSA) framework [13], for instance, which aims to regress the time until a unique event of interest occurs, enforces structural properties across timepoints and studied horizons [14; 15]. As shown in Figure 1, inspired by this, we propose to induce monotonicity in model predictions over time with Temporal Label Smoothing (TLS). This novel regularization strategy also mirrors our expected confidence in the strength of prediction signals over time.

Contributions. The contributions of our work are threefold: (i) First, we adapt and benchmark existing approaches from the survival literature to early event prediction, highlighting theoretical similarities between these frameworks. We bridge the gap with prior work [8; 13; 16], showing that these enforce temporal structure properties in model predictions. (i) Next, we introduce a simple method to achieve this for our single-horizon prediction framework[1]. (iii) Finally, we explore real-world event prediction tasks and demon-

[*]Equal contribution [+] Co-supervised [1]Department of Computer Science, ETH Zürich, Switzerland [2]ETH AI Center, ETH Zürich, Switzerland [3]Max Planck Institute for Intelligent Systems, Tübingen, Germany [4]Swiss Institute for Bioinformatics, Zürich, Switzerland. Correspondence to: Hugo Yèche <hyeche@ethz.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]All code is made publicly available at `https://github.com/ratschlab/tls`.

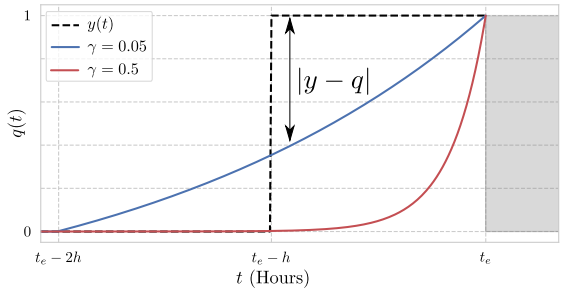


Figure 1: **Illustration of temporal label smoothing** for early event prediction. Predictions are carried out over a horizon h and t_e is the time of the next event, shaded in grey. True labels in black. γ controls the smoothing strength of surrogate labels q.

strate the performance gains of our method, particularly on clinically relevant metrics. Ablations show that this effectively focuses training on datapoints with a stronger predictive signal.

2. Problem Formalism and Related Work

We start by formalizing the early event prediction task and highlight its similarities and distinctions with survival analysis. After discussing its typical training objectives, we outline some temporal structure properties induced by label definition – which lead to novel optimization objectives.

2.1. Early event prediction (EEP)

We assume access to a dataset of irregular time series of covariates $\mathbf{x}_{i,t}$ and binary event labels $e_{i,t}$ encoding whether an event of interest is occurring at time t in series of index i. Each sample is a sequence $\{(\mathbf{x}_{i,0}, e_{i,0}), \ldots, (\mathbf{x}_{i,T_i}, e_{i,T_i})\}$ of length T_i. In the clinical setting, this could correspond to individual patient trajectories as time series of observations, with labeled events such as organ failure or death. For clarity, we drop index i unless explicitly needed.

For each point t along a time series, the covariates observed up to this point are denoted $\mathbf{X}_t = [\mathbf{x}_0, \ldots, \mathbf{x}_t]$ and the absolute time of the next event is given by $t_e = \arg\min_{\tau:\tau \geq t}\{e_\tau : e_\tau = 1\}$. Our task consists of modeling the probability of this event occurring within a fixed prediction horizon h: $y^h(t) = P(t > t_e - h|\mathbf{X}_t)$. In practice, we only access hard, binary labels $y^h(t) = \mathbb{1}[t > t_e - h]$. Estimates of this event probability, denoted $\hat{y}^h(t)$, are typically obtained by maximizing label likelihood through binary classification. As our task focuses specifically on early modeling, no prediction is carried out if the event is currently occurring.

Comparison to survival analysis. Both early event prediction and survival analysis are concerned with modeling the occurrence of an event of interest. These tasks differ in their variable of interest when applied to time series. Survival analysis is focused on studying event probability as a function of time-to-event h for a fixed timepoint t. It aims at modeling the survival function $S(h|\mathbf{X}_t) = P(t_e - t > h|\mathbf{X}_t)$. Early event prediction, in contrast, is concerned with event probability as a function of time t for a fixed horizon h. As a result, for a fixed $\{t, h\}$ and under the assumption of an event occurring only once, we have: $y^h(t) = 1 - S(h|\mathbf{X}_t)$. A dynamic survival analysis (DSA) model could therefore be used for EEP, fixing the horizon to that of interest. This leads to a first experimental question: can the survival objective, which considers all event horizons, improve performance on early event prediction at fixed h?

2.2. Optimization objectives for EEP

We compare relevant training objectives for early event prediction in Table 1, with further detail in Appendix A.4. Prior work on EEP typically focuses on addressing issues of class imbalance through loss reweighting techniques. Static class reweighting was used for sepsis or circulatory failure prediction [25; 7] through a balanced cross-entropy, which assigns a higher weight to samples from the minority class [17]. Still, performance improvements with this objective remain limited on highly imbalanced prediction tasks [26]. In contrast, dynamic reweighting methods such as focal loss and extensions [18; 27] induce a learning bias towards samples with high model uncertainty, typically harder to classify. This approach can improve the prediction of disease progression from imbalanced datasets [19; 20] but does not consider patterns of sample informativeness over time. Whereas class-imbalance techniques are not designed to account for any temporal structure between samples, these methods give higher importance to positive samples from the minority class, which are located closer to the event.

2.3. Preserving temporal structure

In this section, we highlight how different frameworks for early event prediction or dynamic survival analysis enforce some temporal structure properties induced by the task.

Temporal structure. Another important distinction must be made between early event prediction and typical classification tasks, in which data is independent and identically distributed (i.i.d.). Both in EEP and in survival analysis, labels are dependent over time. Within a patient stay, the design of our task induces the following temporal structure properties:

$$\text{Time monotonicity: } y^h(t) \leq y^h(t + \delta t) \quad (1)$$

$$\text{Horizon monotonicity: } y^h(t) \geq y^{h+\delta h}(t) \quad (2)$$

$$\text{Consistency: } y^h(t) = y^{h-\delta t}(t + \delta t) \quad (3)$$

Table 1: **Related work.** Comparison of different relevant training objectives. Early event labels and model predictions at time t are denoted $y_t^h = \mathbb{1}[t > t_e - h]$ and $\hat{y}_t^h \in [0, 1]$, dropping horizon h when fixed. Hazard function labels and predictions are denoted $\lambda_t^h = \mathbb{1}[t = t_e - h]$ and $\hat{\lambda}_t^h$. Binary cross-entropy is denoted by $H[l \parallel p] = -l \log p - (1 - l) \log(1 - p)$. Temporal structure properties are time monotonicity (Eq. 1), horizon monotonicity (Eq. 2), and consistency (Eq. 3). Additional details are provided in Appendix A.4.

Training objective	Previously used for event prediction	Temporal structure (1)	(2)	(3)	Loss function, summed over label values
Cross-entropy [11; 7]	✓	✗	✗	✗	$\sum_t H[y_t \parallel \hat{y}_t]$
Balanced cross-entropy [17]	✓	✗	✗	✗	$\sum_t \omega H[y_t \parallel \hat{y}_t]$
Focal loss [18; 19; 20]	✓	✗	✗	✗	$\sum_t \omega \lvert y_t - \hat{y}_t \rvert^{\zeta} H[y_t \parallel \hat{y}_t]$
Label smoothing [21]	✗	✗	✗	✗	$\sum_t H[q^{LS}(y_t) \parallel \hat{y}_t]$
Multi-horizon prediction [8; 15]	✓	✗	✓	✗	$\sum_t \sum_h H[y_t^h \parallel \hat{y}_t^h]$
Survival analysis likelihood [22; 23]	✗	✗	✓	✗	$\sum_h H[\lambda_0^h \parallel \hat{\lambda}_0^h]$
Landmarking [13; 16]	✗	✗	✓	✓	$\sum_t \sum_h H[\lambda_t^h] \parallel \hat{\lambda}_t^h]$
TCSR [24]	✗	✗	✓	✓	$\sum_t \sum_h H[\hat{\lambda}_{t+1}^{h-1} \parallel \hat{\lambda}_t^h]$
Temporal label smoothing	✓	✓	✗	✗	$\sum_t H[q^{TLS}(t) \parallel \hat{y}_t]$

for $\delta t, \delta h > 0$. Note that each property can be obtained from the other two.

Temporally structured objectives. Some early event prediction and survival analysis objectives induce the above structural properties in model predictions.

In multi-horizon prediction (MHP), the EEP framework is modified to output event predictions over multiple horizons [8; 28; 12]. Predictions are enforced to be monotonically decreasing over the horizon [8], such that if $h \leq h'$, then $\hat{y}^h(t) \geq \hat{y}^{h'}(t)$, as in Eq. 2. This has been shown to improve event prediction performance on the horizon of interest h.

Survival analysis also enforces horizon monotonicity if the survival function is modeled through the hazard function, defined as $\lambda(h|\mathbf{X}_t) = P(t_e - t = h | t_e - t \geq h, \mathbf{X}_t)$. The survival likelihood can then be maximized through binary cross-entropy on the hazard function [23; 29], recovering survival as follows: $S(h|\mathbf{X}_t) = \prod_{k=1}^{h} (1 - \lambda(k|\mathbf{X}_t))$. Equation 2 is enforced by the positivity of the hazard. Interestingly, recent works in DSA directly model the survival function [14; 15], and lose this temporal inductive bias.

Methods extending survival analysis to the dynamics setting [13], where t is allowed to vary, are designed to enforce temporal consistency across horizons (Eq. 3 can also be written in terms of the hazard function). For each timestep t in the training data, landmarking adjusts the prediction horizon to $h - t$, learning the hazard distribution for all horizons and timesteps jointly [13; 16]. A parallel can be drawn between multi-horizon prediction in EEP and landmarking in DSA, with a key difference in the likelihood considered: MHP maximizes event prediction probability, whereas landmarking deals with hazards.

Finally, whereas landmarking induces temporal consistency across labels, Maystre and Russo [24] directly enforces consistency across hazard predictions $\hat{\lambda}(h|\mathbf{X}_t)$. This can be achieved through dynamic programming, substituting ground truth labels with predictions from following time steps.

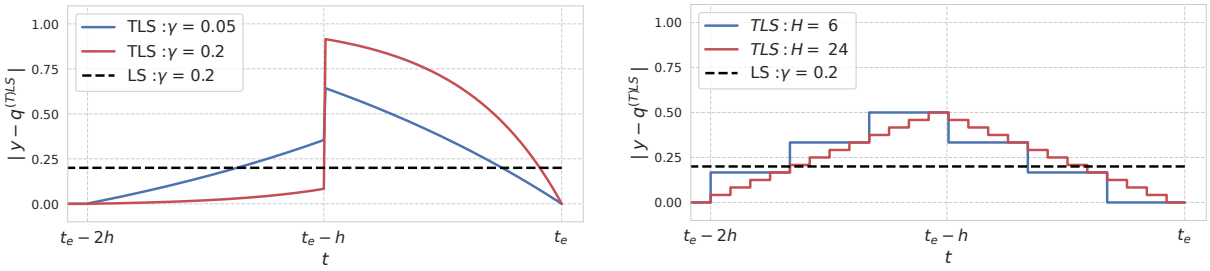
Overall, all methods discussed enforce forms of temporal monotonicity or consistency over *horizons* (Eqs. 2 and 3). In contrast, Equation 1 is most relevant to early event prediction, where h is fixed: we propose a training objective explicitly designed to preserve this form of temporal structure.

3. Temporal Label Smoothing

We introduce temporal label smoothing, our approach to enforce the structural property most relevant to our task (Eq. 1). Thanks to prior knowledge of the labels' structure, our approach focuses training on relevant timesteps and overcomes issues with noisy label boundaries.

Temporal label smoothing substitutes the original label distribution y in the cross-entropy objective with a time-dependent distribution $q(t)$. We constrain this surrogate target to be *monotonically increasing with time*. In practice, as illustrated in Figure 2, this increases smoothing strength around the label boundary $t_e - h$, reducing prediction certainty in this region, which we show to be prone to high error rates in Section 5.1.

Recent work in dynamic survival analysis also proposes to replace labels in the training objective [24], this time



(a) Parametrization q^{exp}. (b) Parametrization q^{step}.

Figure 2: **Label smoothing strength over time** under different parametrizations, with $(h_{min}, h_{max}) = (0, 2h)$. Note that $|y - q^{(T)LS}|$ corresponds to the difference in optimum y^* between the smoothed objective and cross-entropy. The black dashed line represents this difference for regular label smoothing. Smoothing function q^{step} is equivalent to multi-horizon prediction with a unique output.

with predictions at different time points to enforce temporal consistency (Eq. 3). In practice, as demonstrated experimentally in Section 5, we find this approach to be unstable and to converge poorly on real datasets with long time series and large event horizons. In contrast, we propose to replace labels with a prediction-independent distribution fixed *a priori*, and thus less prone to optimization challenges.

Smoothing parametrizations. We propose various temporal smoothing parametrizations for q_t in Appendix A.2. Experimental results suggest that an exponential parametrization, defined as follows, performs best on considered tasks.

$$q^{exp}(t) = \begin{cases} 0 & \text{if } t \leq t_e - h_{max} \\ e^{-\gamma(t_e - t - d)} + A & \text{if } t_e - h_{max} < t < t_e - h_{min} \\ 1 & \text{if } t \geq t_e - h_{min} \end{cases}$$

Parameters h_{min} and h_{max} define the time range over which we apply smoothing, namely $[t_e - h_{max}, t_e - h_{min}]$. Under this constraint, parameters $\{d, A\}$ are defined to enforce q_t to be continuous at boundary points (see Appendix A.2). Finally, γ controls the smoothing strength at a given time.

3.1. Link with label smoothing

A comparison must be drawn with label smoothing [21] which replaces binary cross-entropy labels y with a smooth version q between 0 and 1. By shifting the optimum from y to q, label smoothing prevents models overconfidence, which could improve robustness against the noisy nature of event prediction [30; 31]. Still, despite recent extensions [32; 33; 34], label smoothing remains designed for i.i.d. classification problems. Based on prior knowledge of the temporal structure in our task, our approach also modulates smoothing as a function of time. To the best of our knowledge, we are the first work to introduce a temporal dependence to label smoothing.

3.2. Link with multi-horizon prediction

Temporal label smoothing effectively adapts the contribution of each sample to reflect prior knowledge about the structure of event prediction labels. Under simplifying assumptions justified empirically in Section 5.2, we show that MHP can be seen as a special case of temporal label smoothing. Unlike this method, TLS does not require any architectural change.

Proposition 1. *Under the assumption that model outputs are* ***equal*** *for all horizons $\{h_1, \dots h_H\}$ (rather than monotonically increasing), MHP is equivalent to temporal label smoothing parameterized with q^{step}:*

$$q^{step}(t) = \begin{cases} 0 & \text{if } \quad t < t_e - h_H \\ 1 - \frac{k}{H} & \text{if } \quad t_e - h_{k+1} \leq t < t_e - h_k \quad \forall k \leq H - 1 \\ 1 & \text{if } \quad t \geq t_e - h_1 \end{cases}$$

Proof. See Appendix A.1. □

Proposition 1 frames MHP as a special case of TLS with parametrization q^{step}. This function is defined as a sequence of step functions in time and is illustrated in Figure 2b.

4. Experimental Setup

4.1. Early prediction tasks

We demonstrate the effectiveness of our method on different clinical early prediction tasks to understand its added value. These tasks are established in existing literature and published benchmarks and deal with electronic health records from the ICU, where early prediction of organ failure or acute deterioration is critical to patient management [2]. Clinical events are labeled following internationally accepted criteria [35; 26].

Our work is first evaluated on the prediction of acute circulatory failure within the next $h = 12$ hours, as defined in the HiRID-ICU-Benchmark (HiB) [26]. This task is based on

the publicly available HiRID dataset [7], containing high-resolution observations of over 33,000 ICU admissions. We also investigate early prediction of patient mortality, or *decompensation*, within a horizon of $h = 24$ hours – a widely studied task in the machine learning literature [36]. We use the framework defined in the MIMIC-III Benchmark (M3B) [35] for the MIMIC-III dataset [37], counting approximately 40,000 patient stays. Positive label prevalence is 4.3% and 2.1% of time points for circulatory failure and decompensation prediction respectively. Further details on task definition and data pre-processing are provided in Appendix B.

Alternative tasks. To investigate a third clinical event prediction task, we also considered predicting respiratory failure in intensive care patients [26]. Unfortunately, ambiguous labeling led to close to random performance for all considered methods. Instead, we benchmarked TLS and baselines on a subtask with better defined labels, prediction of the onset of mechanical ventilation and reached similar conclusions to other tasks in Section 5. Experimental details are included in Appendix E.

4.2. Benchmarking strategy

Baselines. We quantify the added value of our method by comparing its performance to alternative learning approaches used for early event prediction (EEP) and dynamic survival analysis (DSA), discussed in Section 2. Our first baselines consist of balanced cross-entropy [17] and focal loss [18], popular sample reweighting methods for imbalanced tasks. We also implement multi-horizon prediction as a multi-output model trained to predict event occurrence over different horizons between 0 and $2h$. Note that for a fair comparison, we set $(h_{min}, h_{max}) = (0, 2h)$ in TLS. As in Tomašev et al. [8], a cumulative distribution function layer on logits enforces the monotonicity of predictions (Eq. 2). We also compare to DSA objectives, with landmarking [13] and the recently proposed TCSR [24] and DDRSA [38]. Finally, we also compare our method to conventional label smoothing [21] to confirm that our method's performance can be attributed to its temporal dependency.

Architecture choice. As our method and baselines are model-agnostic and only vary in terms of optimization objective, a unique model architecture is used for each task, selected through a random search on cross-entropy validation performance. Following a published benchmark on the HiRID dataset [26], we use a GRU [39] architecture for the circulatory failure task. For decompensation prediction, transformers [40] outperform the LSTM-based models [41] originally proposed in the M3B benchmark [35], and are thus used in our work. As recommended by Tomašev et al. [8], we apply l_1-regularization to input embedding layers, which improves performance on both tasks.

Hyperparameters introduced by baselines or by our method, such as strength term γ in smoothing parametrization q^{exp}, are optimized through grid searches on the validation set. Further implementation details are provided in Appendix C.

4.3. Evaluation metrics

To account for the highly imbalanced nature of clinical early prediction tasks, the area under the precision-recall curve (AUPRC) provides more insight than the area under the receiver operating characteristic curve (AUROC): under a low prevalence of positive samples, precision is more sensitive to false alarms than specificity [42]. Still, "area under the curve" metrics can be poorly representative of clinical usefulness, as improvements in low precision regions can dominate such global metrics but remain incompatible with the low false alarm rates required for clinical deployment. Thus, to better assess model performance in this context, we also measure performance at a clinically motivated operating point through recall at 50% precision [28]. To ensure that conclusions made for this operating point also hold at higher precision constraints, we also plot full precision-recall curves.

In addition to *timestep-level* metrics, which measure prediction performance at each data point, we also evaluate models in an event-based approach [7; 8]. Following Tomašev et al. [8]'s definition, an event prediction is positive if the model outputs a positive prediction at any time over the h hours before the event. The threshold defining a positive prediction is chosen based on a precision lower bound. We also use a stepwise criterion with a 50% precision. This allows us to measure the event recall of our approach in comparison to published baselines. Unless stated otherwise, we always report mean performance with 95% confidence intervals on the mean computed over ten training runs.

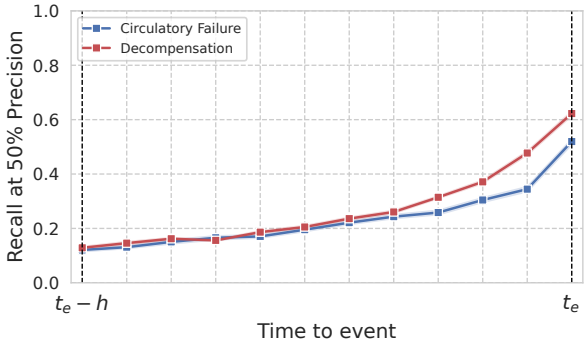


Figure 3: **Comparison of naive performance as a function of time** on both tasks using cross-entropy. Events should be predicted at a horizon h from an event at time t_e. Performance is reported at time increments $\frac{h}{12}$.

Table 2: **Performance of different training objectives for early prediction.** Recall is reported at a 50% timestep-level precision. In **bold**, we highlight best-performing methods with statistically significant p-values (< 0.05) under paired Student's t-tests [43] compared with the next-best method marked italic (last row). Note that cross-entropy is a special case of weighted cross-entropy and focal loss, which performs best in this setting. Hence, the first three lines are identical.

Task	Circulatory Failure (HiRID)			Decompensation (MIMIC-III)		
Training objective	AUPRC	Timestep Recall	Event Recall	AUPRC	Timestep Recall	Event Recall
Cross-entropy [11; 7]	39.1 ± 0.4	29.3 ± 0.9	82.8 ± 1.3	34.5 ± 0.4	28.2 ± 0.5	69.7 ± 1.0
Weighted CE [17]	39.1 ± 0.4	29.3 ± 0.9	82.8 ± 1.3	34.5 ± 0.4	28.2 ± 0.5	69.7 ± 1.0
Focal loss [18; 19; 20]	39.1 ± 0.4	29.3 ± 0.9	82.8 ± 1.3	34.5 ± 0.4	28.2 ± 0.5	69.7 ± 1.0
Label smoothing [21]	39.3 ± 0.4	29.9 ± 0.8	83.8 ± 1.3	33.9 ± 0.3	27.7 ± 0.5	68.8 ± 1.0
Multi-horizon [8; 15]	*39.6* ± 0.5	*30.3* ± 1.0	85.2 ± 1.7	*34.9* ± 0.3	*28.6* ± 0.5	*70.3* ± 0.6
Landmarking [13; 16]	*39.6* ± 0.3	30.1 ± 0.6	*89.1* ± 0.8	34.0 ± 0.5	27.2 ± 0.6	68.8 ± 1.1
TCSR [24]	36.0 ± 0.4	26.5 ± 0.8	89.0 ± 2.1	28.6 ± 1.2	19.9 ± 1.4	68.4 ± 1.0
DDRSA [38]	39.4 ± 0.3	29.3 ± 0.7	87.7 ± 1.5	32.2 ± 1.3	24.0 ± 2.3	65.7 ± 1.0
Temporal label smoothing	$\mathbf{40.6} \pm 0.3$	$\mathbf{32.3} \pm 0.7$	$\mathbf{92.5} \pm 0.5$	$\mathbf{35.5} \pm 0.3$	$\mathbf{29.3} \pm 0.4$	$\mathbf{71.8} \pm 0.8$
p-value	**0.002**	**0.004**	**<0.001**	**0.004**	**0.02**	**0.002**

5. Results

In this section, we validate the following claims: (1) temporal label smoothing yields practical performance improvement along clinically-motivated metrics, and (2) achieves this by leveraging temporal structure and modulating prediction confidence as a function of event proximity.

5.1. Prediction performance

Overall, our results highlight that TLS improves performance over other approaches proposed to address the challenges of early clinical prediction. We occasionally focus on circulatory failure prediction for brevity; see Appendix D for similar conclusions on decompensation.

Necessity of temporal inductive biases. As visualized in Figure 3, training EEP as a simple binary classification with a cross-entropy objective shows a reduction in recall between event time t_e and prediction horizon $t_e - h$. This suggests a weakening in the discriminative signal associated with events and an increase in noise close to the label boundary, where performance is the poorest. In fact, we argue that correct predictions in this region, close to $t_e - h$, are not as critical as ones near t_e: missing an imminent event is more severe. Mirroring the decrease in both signal strength and clinical importance of predictions as the time-to-event increases, model confidence should also decrease, focusing instead on more critical time windows.

Timestep-level performance. In Table 2, we find TLS to outperform baselines across all metrics for circulatory failure and decompensation[1]. The full precision-recall curve of models trained with the best objectives is shown in Figure 5b: TLS improves recall for all precision thresholds beyond 50%, a low false-alarm region of particular clinical relevance [2].

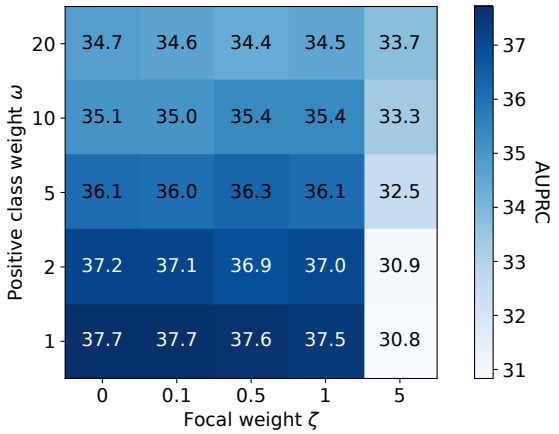
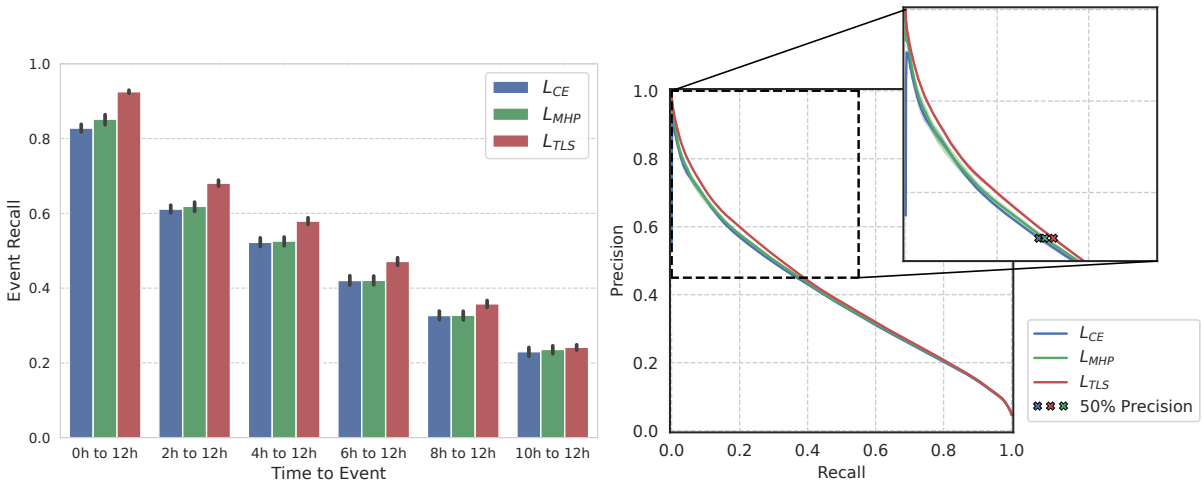


Figure 4: **Performance loss with class reweighting methods** on circulatory failure prediction (validation). Balanced cross-entropy corresponds to $\zeta = 0$.

In contrast, loss reweighting methods designed to tackle class imbalance were found to reduce performance on all tasks over traditional cross-entropy, as shown in Figure 4. For weighted cross-entropy, we attribute it to the increase in

[1]Despite overlapping confidence intervals between multi-horizon and TLS on decompensation due to individual training run variability, we can reject the null hypothesis that MHP has a higher performance than our method (p-values < 0.05)



(a) Event-level performance for a 50% timestep-level precision threshold.

(b) Precision-recall curve. Inset shows the clinically-applicable region with precision $> 50\%$.

Figure 5: **Clinically-oriented performance analysis** of different training objectives on circulatory failure prediction (CE: cross-entropy, MHP: multi-horizon prediction, TLS: temporal label smoothing).

false alarms resulting from the drive to improve recall. This further reduces the low precision of all models, thus negatively affecting the AUPRC. On the other hand, focal loss down-weighs confident samples in training, constraining the model to focus on samples with uncertain predictions. In the context of noisy labeling, as is the case close to our class boundary, data points with ambiguous signals cannot be correctly predicted and thus dominate the loss, impeding improvements in other regions of input space. We analyze model performance over time in Section 5.2 to further support this hypothesis.

Empirical comparison to dynamic survival analysis. Despite the similarities between the tasks of early event prediction and dynamic survival analysis, survival objectives were not found to markedly improve performance on the former, as shown in the second block of Table 2. A likely explanation for this is that the survival likelihood is trained to predict events potentially occurring at horizons much greater than that of interest in EEP. As signal strength decreases with the time-to-event, errors from distant events dominate the loss – leading to poor performance on long time series. This finding goes in the direction of recent works [15; 14] in dynamic survival analysis, which train a fixed (multi-)horizon model as in EEP.

Finally, the prediction-dependent label smoothing in TCSR [24], designed to improve survival performance on short-sequence survival tasks, did not improve performance our EEP tasks either. Training was found to be unstable due to error propagation over long sequences.

Clinically relevant performance. As highlighted in Figure 5a, TLS improves performance over other training objectives in predicting overall adverse event episodes throughout a stay. For circulatory failure, temporal label smoothing is able to predict 7.4% more events than the closest baseline designed for EEP (multi-horizon prediction): this corresponds to reducing the number of missed events by a factor of 2, from 303 to 152 out of 2045 events in the test set on average. Within the events captured by TLS but not by MHP, models trained with our objective predict them on average 104 minutes before their occurrence, giving clinicians sufficient time to take action and avoid patient degradation. We also note here the benefit of adapting dynamic survival analysis to the EEP setting, with landmarking and TCSR performing best in circulatory failure event recall, after TLS. As these methods also enforce temporal structure, this result further motivates our approach, which achieves even greater performance gains and suggests promise in using survival likelihood objectives for early event prediction.

5.2. Illustrative insights

We propose ablations to build intuition around our proposed method. In particular, we aim to understand how temporal smoothing works and why it outperforms other training approaches for early prediction tasks.

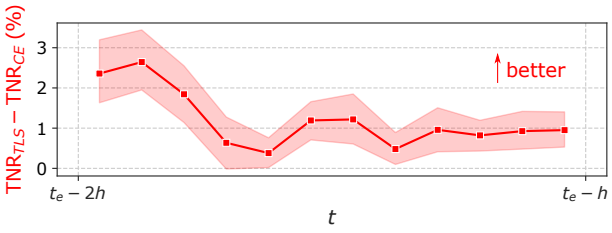
Empirical comparison to multi-horizon prediction. In our theoretical discussion in Section 3.2, we demonstrated how MHP is a restriction of label smoothing with a step function $q^{step}(t)$. This claim relies on the constraint to produce a unique prediction across all considered horizons,

Table 3: **Do MHP's multiple outputs improve performance over TLS with q^{step}?** We provide p-values for the paired Student-t test [43] on the null hypothesis H_0:MHP $\leq$ TLS. With no statistically significant improvements ($p < 0.05$), we justify our assumption in Proposition 1.

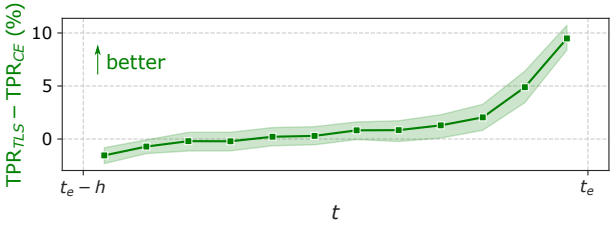
Task	Circulatory Failure (HiRID)			Decompensation (MIMIC-III)		
Training objective	AUPRC	Timestep Recall	Event Recall	AUPRC	Timestep Recall	Event Recall
MHP	39.6 ± 0.5	30.3 ± 1.0	85.2 ± 1.7	34.9 ± 0.3	28.6 ± 0.5	70.3 ± 0.6
TLS (q^{step})	39.3 ± 0.2	29.4 ± 0.8	83.4 ± 1.2	35.2 ± 0.3	29.2 ± 0.4	70.4 ± 0.7
p-value (H_0)	0.11	0.10	**0.03**	0.95	0.97	0.40

reflecting the design of our method. We verify the impact of this assumption by measuring performance gains afforded by learning distinct predictions per horizon. As shown in Table 3, we only find statistical evidence for slight performance gain over using q^{step} on event recall for circulatory failure. Thus, models do not appear to leverage this additional flexibility offered by MHP. With superior results on all event- and timestep-based experiments, and a simpler implementation, we find temporal label smoothing to be a superior training objective to MHP in early prediction tasks.

Performance over time. To better understand the mechanism of action of TLS, we study the difference in performance between TLS and the cross-entropy objective over time in Figure 6. TLS results in a significant increase in true positive and negative rates when prediction time is far from the label boundary ($t = t_e - 2h$ or $t = t_e$). In particular, the performance gains close to the event time t_e explains the better recall of imminent events in Figure 5a.



(a) True negative rate (TNR).



(b) True positive rate (TPR).

Figure 6: **Performance improvement over time** for TLS over cross-entropy on circulatory failure. Timestep-level metrics computed for a precision of 0.5 over two-hour bins.

In contrast, the prediction model trained with TLS is less competitive where smoothing is strongest, near $t_e - h$, but, as expected, this performance loss remains minor. This result validates our hypothesis that the signal is too noisy in the boundary region for any model to recover the original label distribution. From a clinical perspective, errors made in the boundary region are less critical, as they result in the latest false positives or earliest false negatives. Overall, TLS not only improves global event prediction performance but allows these gains to occur at more critical times for clinicians.

6. Limitations and Further Work

Our method is built upon the idea that signal strength decays as the distance to the event increases. This is a valid assumption in the clinical context, as shown in Figure 3. However, it might not hold for other fields which require early prediction of events. In such cases, TLS may not improve performance around the boundary region as it enforces lower confidence there.

Also, our method introduces an additional hyperparameter over cross-entropy and a label parametrization to optimize, as mentioned in Sections 3 and 4. We propose one function which performs best across considered tasks (exponential, again inspired by survival analysis), but further work could be carried out on investigating alternative parametrizations, potentially motivated by theoretical analysis.

Finally, the temporal structure properties presented in our work may not hold in more complex variations of our task, such as in a competing risk setting, where our soft label reparametrization may not be as useful. Further experimental studies are needed to establish the benefit of TLS in this context.

Based on these limitations, promising avenues of further work include combining TLS with survival objectives and using temporal label smoothing for survival regression tasks to explore the relative benefits of different temporal inductive biases. Doing so would also pave the way to explore TLS in the context of competing risks by building upon the rich literature existing in the survival analysis field [14].

7. Conclusion

Early event prediction is paramount to the development of clinical decision support systems, with a demonstrated potential to improve patient outcomes [3]. Still, this task remains relatively poorly studied in the machine learning literature, with few training solutions tailored to address its challenges or to exploit its intrinsic temporal structure. We demonstrate that this can be achieved by adapting and significantly improving approaches from the survival analysis literature [13; 16]. This also motivates us to design a simple, yet top-performing training framework that leverages the structure of event signals over time. We show that multi-horizon prediction, a heuristic used to improve early prediction, can be formalized as a realization of our framework.

Temporal label smoothing empirically outperforms all considered baselines on various tasks and datasets, *with significant improvements* in clinically-relevant evaluation metrics. Our ablation studies show that it effectively focuses training on data points with a stronger predictive signal.

Looking ahead, we expect that temporal label smoothing will be leveraged to develop more clinically reliable systems for risk prediction of infrequent adverse events. Further research on tailored machine learning solutions to improve real-world decision support holds promise for better clinical care and operations management.

Acknowledgements

We thank the anonymous reviewers and area chair for their careful review and proofreading which helped improve the quality of this work. We would also like to thank Alexander Immer and Luca Maystre for the insightful discussion around dynamic survival analysis.

This project was supported by grant #2022-278 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology) and by ETH core funding (to G.R). This publication was made possible by an ETH AI Center doctoral fellowship to A.P.

References

[1] 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. doi: 10.1175/JAMC-D-15-0297.1. URL `https://journals.ametsoc.org/view/journals/apme/55/11/jamc-d-15-0297.1.xml`.

[2] 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):1–10, 2020.

[3] Gary B Smith, David R Prytherch, Paul Meredith, Paul E Schmidt, and Peter I Featherstone. The ability of the national early warning score (news) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, 2013.

[4] Anne Pullyblank, Alison Tavaré, Hannah Little, Emma Redfern, Hein le Roux, Matthew Inada-Kim, Kate Cheema, and Adam Cook. Implementation of the national early warning score in patients with suspicion of sepsis: evaluation of a system-wide quality improvement project. *British Journal of General Practice*, 70 (695):e381–e388, 2020.

[5] Andrew Wong, Albert T Young, April S Liang, Ralph Gonzales, Vanja C Douglas, and Dexter Hadley. Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA network open*, 1(4):e181018–e181018, 2018.

[6] Josef Fagerström, Magnus Bång, Daniel Wilhelms, and Michelle S Chew. Lisep lstm: a machine learning algorithm for early detection of septic shock. *Scientific reports*, 9(1):1–8, 2019.

[7] 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.

[8] 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.

[9] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten M. Borgwardt. Set functions for time series. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4353–4363. PMLR, 2020. URL http://proceedings.mlr.press/v119/horn20a.html.

[10] Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=4c0J6lwQ4_.

[11] Simon Meyer Lauritsen, Mads Ellersgaard Kalør, Emil Lund Kongsgaard, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artif. Intell. Medicine*, 104:101820, 2020. doi: 10.1016/j.artmed.2020.101820. URL https://doi.org/10.1016/j.artmed.2020.101820.

[12] Subhrajit Roy, Diana Mincu, Eric Loreaux, Anne Mottram, Ivan Protsyuk, Natalie Harris, Yuan Xue, Jessica Schrouff, Hugh Montgomery, Alistair Connell, et al. Multitask prediction of organ dysfunction in the intensive care unit using sequential subnetwork routing. *Journal of the American Medical Informatics Association*, 28(9):1936–1946, 2021.

[13] Hans C Van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007.

[14] 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.

[15] 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.

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

[17] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[19] Zhe Wang, Yiwen Zhu, Dongdong Li, Yichao Yin, and Jing Zhang. Feature rearrangement based deep learning system for predicting heart failure mortality. *Computer methods and programs in biomedicine*, 191: 105383, 2020.

[20] Subhrajit Roy, Diana Mincu, Lev Proleev, Negar Rostamzadeh, Chintan Ghate, Natalie Harris, Christina Chen, Jessica Schrouff, Nenad Tomašev, Fletcher Lee Hartsell, et al. Disability prediction in multiple sclerosis using performance outcome measures and demographic data. In *Conference on Health, Inference, and Learning*, pages 375–396. PMLR, 2022.

[21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

[22] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 00359246. URL http://www.jstor.org/stable/2985181.

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

[24] Lucas Maystre and Daniel Russo. Temporally-consistent survival analysis. In *Advances in Neural Information Processing Systems 36*, 2022.

[25] Joseph Futoma, Sanjay Hariharan, and Katherine A. Heller. Learning to detect sepsis with a multitask gaussian process RNN classifier. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1174–1182. PMLR, 2017. URL http://proceedings.mlr.press/v70/futoma17a.html.

[26] 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.

[27] Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Xiaojie Shi, Shuyang Cheng, and Dragomir Anguelov. Polyloss: A polynomial expansion perspective of classification loss functions. *CoRR*, abs/2204.12511, 2022. doi: 10.48550/arXiv.2204.12511. URL https://doi.org/10.48550/arXiv.2204.12511.

[28] Nenad Tomašev, Natalie Harris, Sebastien Baur, Anne Mottram, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Valerio Magliulo, et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nature Protocols*, 16(6):2765–2787, 2021.

[29] Erin Craig, Chenyang Zhong, and Robert Tibshirani. Survival stacking: casting survival analysis as a classification problem. *arXiv preprint arXiv:2107.13480*, 2021.

[30] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6448–6458. PMLR, 2020. URL http://proceedings.mlr.press/v119/lukasik20a.html.

[31] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4696–4705, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.

[32] Weizhi Li, Gautam Dasarathy, and Visar Berisha. Regularization via structural label smoothing. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 1453–1463. PMLR, 2020. URL http://proceedings.mlr.press/v108/li20e.html.

[33] Clara Meister, Elizabeth Salesky, and Ryan Cotterell. Generalized entropy regularization or: There's nothing special about label smoothing. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6870–6886. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.615. URL https://doi.org/10.18653/v1/2020.acl-main.615.

[34] Julian Lienen and Eyke Hüllermeier. From label smoothing to label relaxation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 8583–8591. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17041.

[35] 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.

[36] David Bellamy, Leo Celi, and Andrew L Beam. Evaluating progress on machine learning for longitudinal electronic healthcare data. *arXiv preprint arXiv:2010.01149*, 2020.

[37] Alistair EW 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):1–9, 2016.

[38] Niranjan Damera Venkata and Chiranjib Bhattacharyya. When to intervene: Learning optimal intervention policies for critical events. In *Advances in Neural Information Processing Systems*.

[39] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014. URL http://arxiv.org/abs/1412.3555.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von

Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

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

[42] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

[43] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[44] FJ Richards. A flexible growth function for empirical use. *Journal of experimental Botany*, 10(2):290–301, 1959.

[45] David Collett. *Modelling survival data in medical research*. CRC press, 2015.

[46] Zhengzheng Xing, Jian Pei, and Philip S. Yu. Early prediction on time series: A nearest neighbor approach. In Craig Boutilier, editor, *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*, pages 1297–1302, 2009. URL http://ijcai.org/Proceedings/09/Papers/218.pdf.

[47] Guoliang He, Yong Duan, Tieyun Qian, and Xu Chen. Early prediction on imbalanced multivariate time series. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 1889–1892, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450322638. doi: 10.1145/2505515.2507888. URL https://doi.org/10.1145/2505515.2507888.

[48] Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11842–11851. PMLR, 2021. URL http://proceedings.mlr.press/v139/yang21m.html.

[49] 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.

[50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[52] NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. Cuda, release: 10.2.89, 2020. URL https://developer.nvidia.com/cuda-toolkit.

[53] Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, and Evan Shelhamer. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014.

[54] The gin-config Team. gin-config python packaged. https://github.com/google/gin-config, 2019.

A. Theoretical Details

A.1. Multi-Horizon prediction: proof of Proposition 1

Equivalency between MHP and TLS objectives. Recalling the formalism of multi-horizon prediction outlined in Section 3.2, true labels and model predictions at time t can be rewritten as $\mathbf{y}_t = [y_t^{h_1}, \dots, y_t^{h}, \dots, y_t^{h_H}]$ and $\hat{\mathbf{y}}_t = [\hat{y}_t^{h_1}, \dots, \hat{y}_t^{h}, \dots, \hat{y}_t^{h_H}]$, where H is the number of horizons considered. The training objective for this datapoint becomes:

$$L^{MHP}(\mathbf{y}_t, \hat{\mathbf{y}}_t) = -\frac{1}{H}\sum_{k=1}^{H} y_t^{h_k}\log(\hat{y}_t^{h_k}) + (1-y_t^{h_k})\log(1-\hat{y}_t^{h_k})$$

The assumption that $\{\hat{y}_t^{h_k}\}_k$ is **equal** for all k allows to rewrite the objective as follows:

$$L^{MHP}(\mathbf{y}_t, \hat{\mathbf{y}}_t) = -\left[\log(\hat{y}_t)\frac{1}{H}\sum_{k=1}^{H} y_t^{h_k} + \log(1-\hat{y}_t)\frac{1}{H}\sum_{k=1}^{H}(1-y_t^{h_k})\right]$$

with $\hat{y}_t$ being the common prediction shared across all horizons. This equation can now be viewed as a temporal label smoothing objective with smoothed labels $q^{step}(t) = \frac{1}{H}\sum_{k=1}^{H} y_t^{h_k}$:

$$L^{MHP}(\mathbf{y}_t, \hat{\mathbf{y}}_t) = -\left[\log(\hat{y}_t)\cdot q^{step}(t) + \log(1-\hat{y}_t)\cdot\left(1-q^{step}(t)\right)\right]$$

Smoothing parametrization. Next, we aim to recover the explicit form of $q^{step}(t)$. Without loss of generality, we assume that horizons $\{h_k\}_k$ are in ascending order. The temporal dependency between samples, formalized in Equation 1), results in the following relationship between predictions at horizons h_u and h_v :

$$v \leq u \quad \text{and} \quad y_t^{h_v} = 1 \implies y_t^{h_u} = 1 \tag{4}$$

$$v \geq u \quad \text{and} \quad y_t^{h_v} = 0 \implies y_t^{h_u} = 0 \tag{5}$$

Thanks to the above property, we can determine $q^{step}(t)$ by studying three cases of multi-horizon labels, illustrated in Figure 7. For notational simplicity, we define the time-to-event as $d_e(t) = t_e - t$.

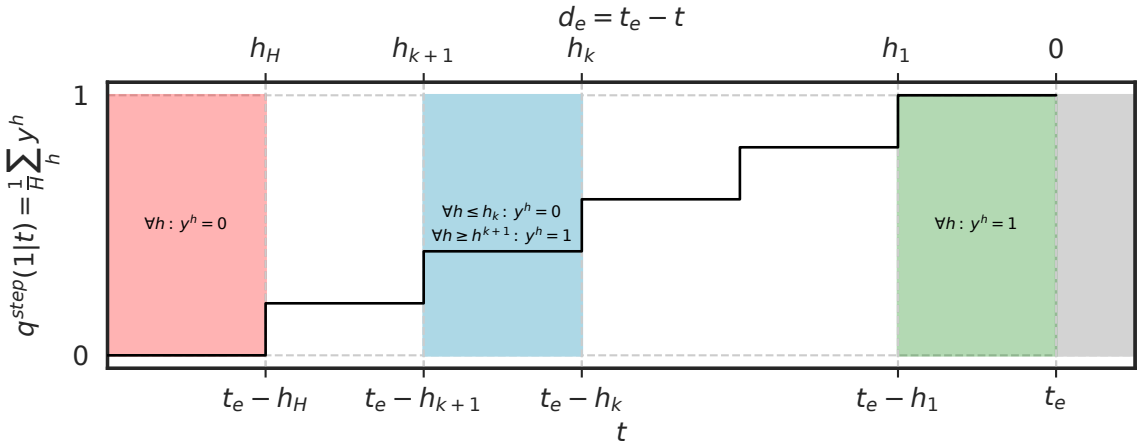


Figure 7: **Label values for multi-horizon prediction**, and conversion to smoothed labels $q^{step}(t)$.

Case 1: $d_e(t) \leq h_1$.
From label definition. we have that $y_t^{h_1} = 1$ if $d_e(t) \leq h_1$. As h_1 is the smallest horizon, following Equation 4, we have $y_t^{h_c} = 1, \forall c \in [\![1, H]\!]$. We can rewrite the objective as:

$$\begin{aligned} L^{MHP}(\mathbf{y}_t, \hat{\mathbf{y}}_t) &= -\log(\hat{y}_t) \\ &= -[q^{step}(t)\log(\hat{y}_t) + (1-q^{step}(t))\log(1-\hat{y}_t)] \end{aligned}$$

where $q^{step}(t) = 1$.

Case 2: $d_e(t) > h_H$.
Similarly, if $d_e(t) > h_H$, then $y_t^{h_H} = 0$ which implies $y_t^{h_c} = 0, \forall c \in [\![1, H]\!]$ from Equation 5. The objective can be rewritten as:

$$\begin{aligned} L^{MHP}(\mathbf{y}_t, \hat{\mathbf{y}}_t) &= -\log(1 - \hat{y}_t) \\ &= -[q^{step}(t)\log(\hat{y}_t) + (1 - q^{step}(t))\log(1 - \hat{y}_t)] \end{aligned}$$

where $q^{step}(t) = 0$.

Case 3: $\exists k \in [\![1, H-1]\!] \quad \text{s.t} \quad h_k < d_e(t) \leq h_{k+1}$.
Following the same reasoning as in the first two cases, we now have a specific index k which separates positive and negative labels. We have $y_t^{h_c} = 0, \forall c \in [\![1, k]\!]$ and $y_t^{h_c} = 1, \forall c \in [\![k+1, H]\!]$. This allows to rewrite the objective as follows:

$$\begin{aligned} L^{MHP}(\mathbf{y}_t, \hat{\mathbf{y}}_t) &= -[\frac{H-k}{H}\log(\hat{y}_t) + \frac{k}{H}\log(1 - \hat{y}_t)] \\ &= -[q^{step}(t)\log(\hat{y}_t) + (1 - q^{step}(t))\log(1 - \hat{y}_t)] \end{aligned}$$

where

$$q^{step}(t) = \frac{H-k}{H}.$$

This defines a new smoothing parametrisation q^{step}:

$$q^{step}(t) = \begin{cases} 1 - \frac{k}{H} & \text{if} \quad h_k \leq d_e(t) < h_{k+1} \quad \forall k \leq H-1 \\ 1 & \text{if} \quad d_e(t) \leq h_1 \\ 0 & \text{if} \quad d_e(t) > h_H \end{cases}$$

Thus, $\forall d_e(t) > 0$, we find that $L^{MHP} = L^{TLS}$ when smoothed labels are defined as q^{step}. This concludes our proof. □

A.2. Temporal label smoothing functions

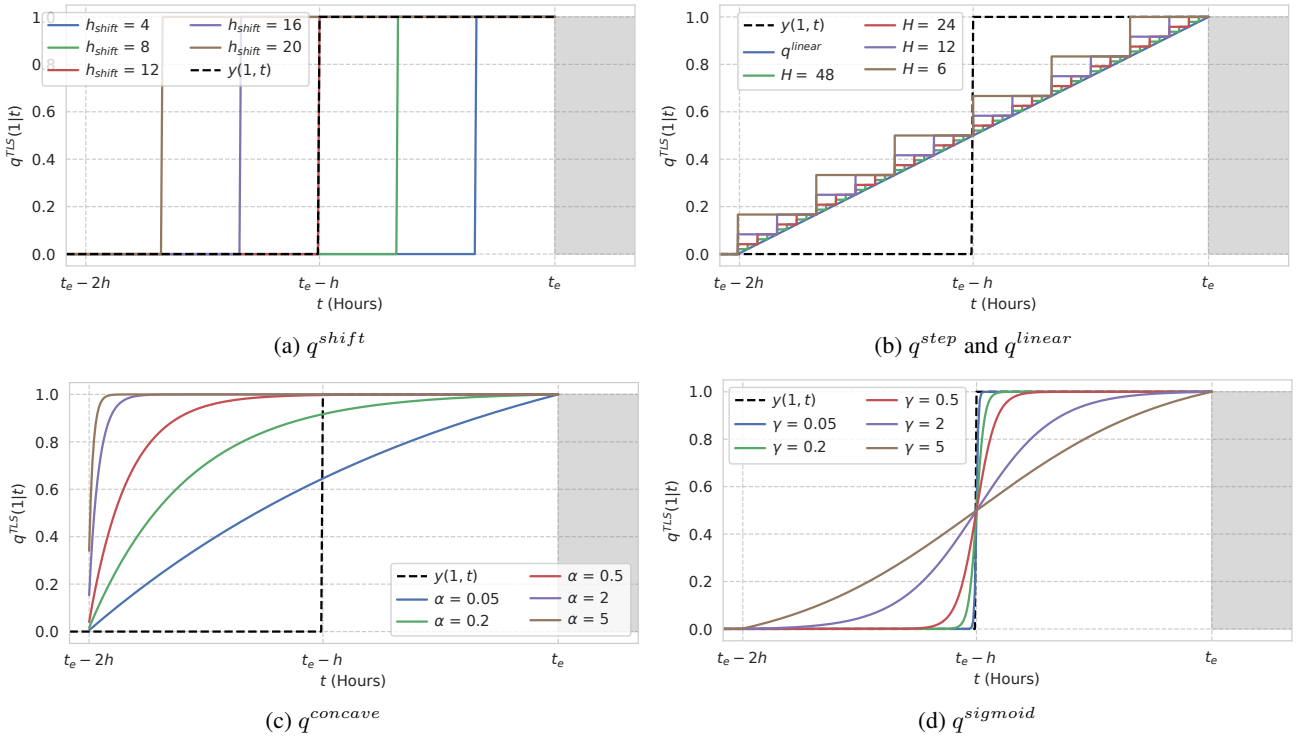
Motivated by prior work [8; 22], we compare the performance of various smoothing functions $q(t)$. All proposed parametrizations are continuous and monotonous increasing functions that satisfy boundary conditions $q(t_e - 2h) = 0$ and $q(t_e) = 1$. As evidenced in Table 4, we find exponential label smoothing to perform best or as well as others across all tasks and metrics. Performance as a function of hyperparameter setting can be visualized in Figure 9. All model and hyperparameter selection were carried out on the validation set, including the final choice of parametrization function.

Table 4: **Performance of different smoothing functions on early prediction tasks.** Timestep-level recall is reported at a 50% precision.

Task	Circulatory Failure		Decompensation	
Method	AUPRC	Recall	AUPRC	Recall
q^{step}	39.3 ± 0.2	29.4 ± 0.8	35.2 ± 0.3	29.2 ± 0.4
q^{shift}	40.1 ± 0.3	31.8 ± 0.6	34.5 ± 0.4	28.2 ± 0.5
q^{linear}	39.4 ± 0.3	29.7 ± 0.8	35.1 ± 0.4	29.2 ± 0.6
$q^{sigmoid}$	39.4 ± 0.3	29.7 ± 0.8	34.9 ± 0.4	28.8 ± 0.5
$q^{concave}$	39.4 ± 0.3	29.7 ± 0.8	35.1 ± 0.4	29.2 ± 0.6
q^{exp}	$\mathbf{40.6} \pm 0.3$	$\mathbf{32.3} \pm 0.7$	$\mathbf{35.5} \pm 0.3$	$\mathbf{29.3} \pm 0.4$

Shifted boundary labels. Shifting the prediction horizon or label boundary in training can be viewed as a form of temporal label smoothing, in which class labels are inverted within a prediction window of interest. This defines the following smoothing parametrization $q^{shift}(t)$:

$$q^{shift}(t) = \mathbb{1}\left[t \geq t_e - h_{shift}\right] \tag{6}$$



(a) q^{shift} (b) q^{step} and q^{linear}

(c) $q^{concave}$ (d) $q^{sigmoid}$

Figure 8: **Illustration of temporal label smoothing** with alternative smoothing parametrizations.

where h_{shift} is a hyperparameter controlling the horizon of the smoothed labels ($h_{shift} = h$ corresponds to cross-entropy training). The strength of this smoothing function is illustrated in Figure 8a.

Figure 9 outlines the performance of this alternative smoothing parametrization as a function of h_{shift}. For decompensation, shifting the label boundary closer to the event time decreases performance. On circulatory failure, performance does improve over traditional cross-entropy training as the label horizon is brought closer to the event of interest, which can be interpreted as an inductive bias similar to that induced by the exponential smoothing function.

Linear label smoothing. The most straightforward extension to the step function q^{step} described in Section 3.2 is a linear label smoothing corresponding to the case $H \to +\infty$.
Our parametrization $q^{linear}(t)$ is thus defined as follows:

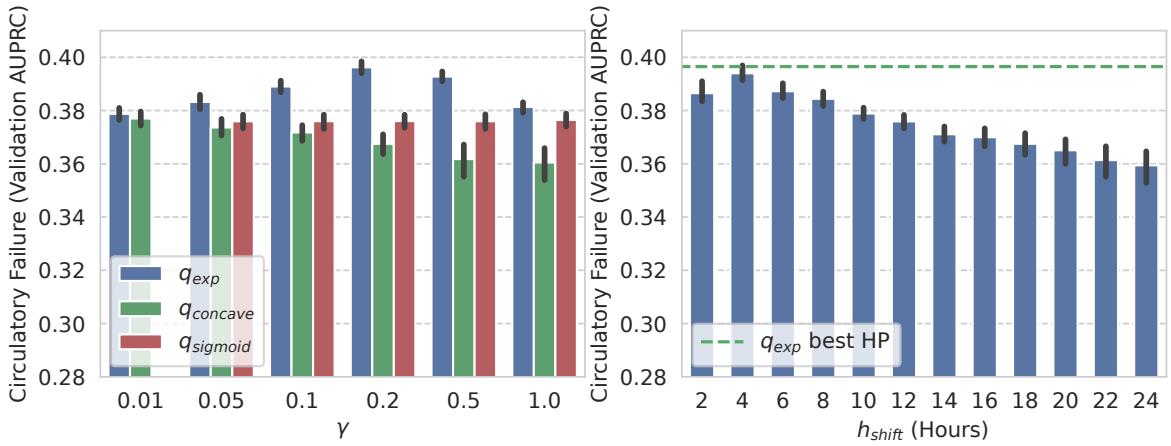
$$q^{linear}(t) = \begin{cases} 0 & \text{if} \quad t \leq t_e - 2h \\ 1 - \frac{t_e - t}{2h} & \text{if} \quad t > t_e - 2h \end{cases} \tag{7}$$

We illustrate the impact of the number of steps H in Figure 8b.

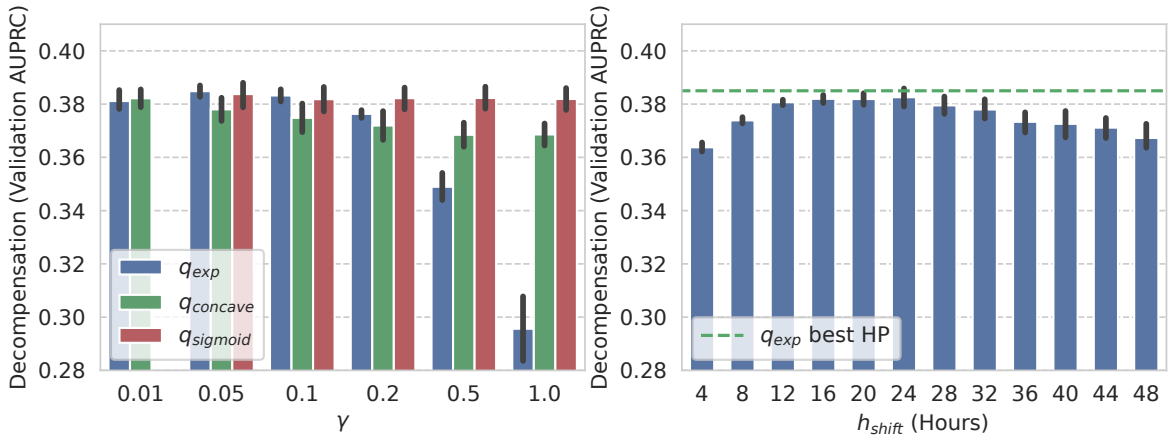
Sigmoidal label smoothing. Another natural direction to explore is to smooth labels starting from the true distribution, a unique step function at $t = t_e - h$. This can be achieved by defining q_t as a generalized logistic function [44]:

$$q^{sigmoid}(t) = \begin{cases} 0 & \text{if} \quad t \leq t_e - 2h \\ \frac{K - A}{1 + e^{\frac{t_e - t - d}{\gamma}}} + A & \text{if} \quad t > t_e - 2h \end{cases} \tag{8}$$

where K, A and d are three constants fixed by imposing the boundary conditions at $t = t_e - 2h$ and $t = t_e$, as well as



(a) Circulatory failure.



(b) Decompensation.

Figure 9: **Validation AUPRC performance of temporal label smoothing as a function of smoothing hyperparameters**, with different smoothing parameterizations. (Left) Performance for different smoothing strengths γ with $q^{exp}, q^{concave}, q^{sigmoid}$; (Right) Performance for different prediction horizons h_{shift} with q^{shift} smoothing.

$q(t_e - 2h) = \frac{1}{2}$. This yields:

$$K = -Ae^{\frac{2h-d}{\gamma}}$$
$$A = \frac{e^{\frac{-d}{\gamma}} + 1}{e^{\frac{-d}{\gamma}} - e^{\frac{2h-d}{\gamma}}}$$
$$d = h$$

As shown in Figure 8d, γ controls the smoothing strength, interpolating between the true distribution $\delta_{y=1}$ as $\gamma \to 0$ and q^{linear} when $\gamma \to +\infty$.

Exponential label smoothing. The smoothing function we find to perform best is the exponential decay one. This idea is motivated by survival analysis, where patient survival probability can be modeled as the exponential decay of a cumulative hazard function [22; 45]. In practice, as defined in Section 3, our exponential smoothing function $q^{exp}(t)$ is defined as follows:

$$q^{exp}(t) = \begin{cases} 0 & \text{if} \quad t \leq t_e - 2h \\ e^{-\gamma(t_e - t - d)} + A & \text{if} \quad t > t_e - 2h \end{cases} \tag{9}$$

where parameters $\{d, A\}$ are set to satisfy boundary conditions:

$$A = -e^{-\gamma(2h-d)}$$
$$d = -\frac{1}{\gamma}\ln\left(1 - e^{-\gamma 2h}\right)$$

Here, γ also controls the smoothing strength between q^{linear} when $\gamma \to 0$ and $q(t) = 0\,\forall t < t_e$ when $\gamma \to +\infty$.

Overall, despite q^{shift} achieving good results on circulatory failure, q^{exp} statistically outperforms this smoothing parameterization for both tasks on validation metrics. An interesting avenue for further work would be to combine exponential smoothing with the boundary shift approach, or effectively change (h_{min}, h_{max}), which was fixed to $(0, 2h)$ in our work for a fair comparison to multi-horizon prediction.

Concave exponential label smoothing. Finally, to mirror the behavior of the exponential smoothing function away from linear interpolation and investigate its effect on performance, we designed the following concave smoothing function $q^{concave}$:

$$q^{concave}(t) = \begin{cases} 0 & \text{if} \quad t \leq t_e - 2h \\ 1 - e^{-\gamma(d - t_e + t)} + A & \text{if} \quad t > t_e - 2h \end{cases} \tag{10}$$

Parameters $\{d, A\}$ are identical to the convex smoothing function parameters, set to satisfy boundary conditions. The strength of this concave smoothing function is illustrated Figure 8c.

No performance gains were obtained through temporal label smoothing with a concave function, as shown in Figure 9. This smoothing function effectively penalizes false positives harder than false negatives, which is less adapted to our tasks of interest (in contrast to the convex q^{exp}). As a result, the best-performing concave parametrization is consistently obtained with the lowest value of γ, closer to a linear function choice.

A.3. Related time-series tasks

Comparison to early time-series classification. A distinction must be drawn between our task of early event prediction and that of early time-series classification. The latter has been more extensively explored in the literature [46; 47; 48], but addresses a distinct problem.

Considering a time series up to timestep t, early event prediction is concerned with classifying *whether* a particular event will occur between t and $t + h$, for a fixed horizon h. Predictions are made at each timepoint over the entire time series: as multiple samples arise from the same time series and therefore depend on one another over time, these should not be considered as i.i.d.

In contrast, early classification of time series aims to regress *the first timepoint* t at which a label for the entire time series can be predicted with a desired accuracy [46]. A single prediction is made, as soon as possible, for the entire series – which can be considered an independent sample from the dataset of time series. This latter task can be framed as an early prediction of the event "prediction is possible", where $h = \infty$, given a separate time-series classifier. As a result, an interesting avenue of further work would be to apply temporal label smoothing to the latter task.

On the other hand, early event prediction cannot be translated into a simple early classification problem. As a result, methods designed for early time-series classification are therefore not applicable to this problem setting.

A.4. EEP Objective Functions

In this section, we clarify the mathematical formalism behind our EEP baselines to facilitate comparison to temporal label smoothing. Most baselines explored effectively propose a modification of the cross-entropy objective often used for binary classification tasks, $L^{CE}(y, \hat{y}) = -y\log(\hat{y}) - (1 - y)\log(1 - \hat{y})$.

Weighted cross-entropy. To facilitate learning from highly imbalanced datasets, a common adjustment to the training objective consists of reweighting terms in the cross-entropy objective:

where hyperparameter ω determines the contribution of each class to the loss. Balanced cross-entropy is a special case of this objective, where weights are based on the prevalence of each class (ω is set as the inverse of the proportion of positive labels). Regular cross-entropy corresponds to the case where $\omega = 1/2$.

Focal loss. Denoting our output prediction as $\hat{y} = p_\theta(y = 1)$, the focal loss objective for binary classification of target y is a variant on the balanced cross-entropy loss:

$$L^{focal}(y, \hat{y}) = -\omega(1 - \hat{y})^\zeta y \log(\hat{y}) - (1 - \omega)\hat{y}^\zeta(1 - y)\log(1 - \hat{y})$$

where ω_y is a balancing weight for class y and ζ is the focal loss weight.

Multi-horizon prediction. As highlighted in Section 3.2, multi-horizon training can be formalized as the following objective:

$$L^{MHP}(\mathbf{y}_t, \hat{\mathbf{y}}_t) = -\frac{1}{H}\sum_{k=1}^{H} y_t^{h_k}\log(\hat{y}_t^{h_k}) + (1 - y_t^{h_k})\log(1 - \hat{y}_t^{h_k})$$

where true labels and model predictions are given by $\mathbf{y}_t = [y_t^{h_1}, \ldots, y_t^h, \ldots, y_t^{h_H}]$ and $\hat{\mathbf{y}}_t = [\hat{y}_t^{h_1}, \ldots, \hat{y}_t^h, \ldots, \hat{y}_t^{h_H}]$, for H distinct horizons.

Label smoothing. As introduced by Szegedy et al. [21], label smoothing consists of substituting the original label distribution $\delta_{y=c}$ in the cross-entropy objective $L^{CE}(y, \hat{y})$ by a smoothed version $q^{LS}(c|y)$. This surrogate distribution over classes c is defined as follows :

$$q^{LS}(c|y) = \delta_{y=c}(1 - \alpha) + u(c)\alpha \tag{11}$$

In the original approach, u is uniform and $\alpha \in [0, 1]$ controls the smoothing strength. By shifting the minimum of the objective function away from $\hat{y} = 1$, labels smoothing prevents the model from becoming overconfident during training. Alternative designs for u have been proposed [32; 33; 34] but are incompatible with the binary nature of adverse event prediction. In binary tasks, labeling is defined according to the positive class such that $y \in \{0, 1\}$ and $\hat{y} = p_\theta(y = 1)$. Label smoothing therefore becomes a linear interpolation with parameter α such that $q^{LS} = p(y = 1)$:

$$q^{LS} = (1 - \alpha)y + \alpha(1 - y) \tag{12}$$

As suggested by Lukasik et al. [30], label smoothing can be used to regularize early prediction models due to the inherently noisy nature of the task. It does not, however, account for the time dependency between samples of a given stay – highlighted in our problem formalism (Section 2.1). In contrast, temporal label smoothing modulates smoothing based on time t to infuse this prior knowledge into the training objective.

A.5. DSA Objective Functions

In this section, we detail how despite existing differences between EEP and DSA, we can train a model with a DSA objective while using it for EEP tasks at inference time. We then describe in detail the two baselines we consider from DSA: landmarking and TCSR.

From Survival Analysis to Early Event Prediction. Survival analysis is a statistical framework to model the time T until an event of interest occurs. This event is considered to be terminal, thus, it is **unique** and no observation is carried after it. In survival analysis, we assume access only to an initial observation of a patient state $\mathbf{X}_{i,0} = [\mathbf{x}_{i,0}]$, a survival time T_i and a censoring indicator c_i. If a patient was (right-)*censored*, thus did not experience an event before the last know survival time at T_i, then $c_i = 1$. Otherwise, we have that $c_i = 0$, which means the patient reached a terminal state at T_i. Given these, we can define three probability functions:

$$\begin{aligned}
\text{probability mass function:} \quad & f(k|\mathbf{X}) = P(T = k|\mathbf{X}_{i,0}) \\
\text{survival function:} \quad & S(k|\mathbf{X}) = P(T > k|\mathbf{X}) \\
\text{hazard function:} \quad & \lambda(k|\mathbf{X}) = P(T = k|T \geq k, \mathbf{X}) = P(T = k|T > k - 1, \mathbf{X})
\end{aligned}$$

Then, if we consider only non-censored and right-censored patients, the survival likelihood can be defined as follows:

$$\mathcal{L}_{surv} = \prod_i P(T = T_i|\mathbf{X}_{i,0})^{1-c_i} P(T > T_i|\mathbf{X}_{i,0})^{c_i}$$

Thus, when maximizing $\mathcal{L}_{surv}$, we aim to maximize the probability of failure at time T_i if the event occurred, or the probability of survival until at least T_i if the patient is censored. It has been shown [23; 29] that MLE on $\mathcal{L}_{surv}$ is equivalent to minimizing the binary cross-entropy between hazard function estimates and labels of the form $\lambda_i^h = \mathbb{1}\left[T_i = h \wedge c_i = 0\right]$. Thus, in practice when training a model with a survival likelihood, we minimize $\sum_{i=1}^{N}\sum_{h=1}^{T_i} -\lambda_i^h \log(\hat{\lambda}(h|\mathbf{X}_{0,i}))$. As mentioned in Section 2, using existing relation between f, S and λ, such as $f(k|\mathbf{X}) = h(k|\mathbf{X})S(k-1|\mathbf{X})$ and $S(k|\mathbf{X}) = 1 - \sum_{p=1}^{k} f(p|\mathbf{X}) = \prod_{p=1}^{k}(1 - h(p|\mathbf{X}))$, we can recover the model's probability estimate for an event to occur within a fixed horizon h as $\hat{y}^h = 1 - \hat{S}(h|\mathbf{X})$.

Landmarking. When multiple observations are available for a given patient, thus $\mathbf{X}_{i,t} = [\mathbf{x}_{i,0}, ..., \mathbf{x}_{i,t}]$, as in EEP, existing works [13; 16] have extended survival analysis to this dynamic context. This field is referred to as "dynamic survival analysis". As mentioned in Section 2, the most prominent technique to leverage these additional observations is landmarking, where the model is fitted with new triplets of the form $(\mathbf{X}_{i,t}, T_i - t, c_i)$. As in regular survival analysis, when using landmarking, we minimize binary cross-entropy on the hazard function of the form $\sum_{i=1}^{N}\sum_{i=t}^{T_i-1}\sum_{h=1}^{T_i-t} -\lambda_{t,i}^h \log(\hat{\lambda}(h|\mathbf{X}_{t,i}))$ with $\lambda_{t,i}^h = \mathbb{1}\left[T_i - t = h \wedge c_i = 0\right]$. As in regular survival analysis, we can recover the model's probability estimate for an event to occur within a fixed horizon h from a given timepoint t as $\hat{y}_t^h = 1 - \hat{S}(h|\mathbf{X}_t)$, which is the probability of interest in EEP tasks.

Temporally consistent survival regression (TCSR). Concurrently to our work, Maystre and Russo [24] proposed TCSR, a method based on a temporally consistent dynamic sample reweighting and label softening. Indeed, to enforce models estimate to match constraints from Equation 3, TCSR proposes to replace landmarking labels $[\lambda_{t,i}^1, \lambda_{t,i}^2, ..., \lambda_{t,i}^{T_i-t}]$ by $[\lambda_{t,i}^1, \hat{\lambda}(1|\mathbf{X}_{t-1,i}), ..., \hat{\lambda}(T_i - t - 1|\mathbf{X}_{t+1,i})]$. In addition, they also apply a reweighting according to the model estimate of the survival function, such that $w_{t,i}^1 = 1, w_{t,i}^2 = 1 - \hat{f}(1|\mathbf{X}_{t+1,i})$ and $w_{t,i}^h = \hat{S}(h-2|\mathbf{X}_{t+1,i}) \quad \forall h \geq 2$.

Dynamic Deep Recurrent Survival Analysis (DDRSA). In previous work, Ren et al. [49] propose an extension to survival analysis by modeling hazard distribution with a recurrent neural network. This model is trained to maximize a modified likelihood:

$$\mathcal{L}_{DRSA} = \prod_i P(T = T_i|\mathbf{X}_{i,0})^{(1-c_i)\alpha_{DRSA}} P(T \leq T_i|\mathbf{X}_{i,0})^{(1-c_i)(1-\alpha_{DRSA})} P(T > T_i|\mathbf{X}_{i,0})^{c_i(1-\alpha_{DRSA})}$$

Decreasing α_{DRSA} enforces the model to focus on learning censoring patterns over exact time to event prediction and vice-versa. As with landmarking for survival analysis, Venkata and Bhattacharyya [38] extends DDRSA to the dynamic case by considering all triplets of the form $(\mathbf{X}_{i,t}, T_i - t, c_i)$ in the above likelihood.

Handling of non-terminal events. Certain tasks in EEP tackle events that are terminal such as decompensation. There, the underlying assumption made in survival analysis regarding the terminality of states holds allowing to rely on a DSA approach for EEP as described above. However, in practice, most events from EEP, such as circulatory failure, are not terminal. This means that observations are carried out during and after an event. It also means other events of the same type can occur. To still use a survival analysis method for these tasks, we further split patient stays into episodes. Using EEP notations, for a patient indexed by i experiencing v events at times $t_{e_1}, ..., t_{e_v}$, respectively ending at times $s_{e_1}, ..., s_{e_v}$, we consider as distinct samples the episodes $[\mathbf{X}_{i,0}, ..., \mathbf{X}_{i,t_{e_1}-1}], [\mathbf{X}_{i,s_{e_1}}, ..., \mathbf{X}_{i,t_{e_2}-1}], ..., [\mathbf{X}_{i,s_{e_v}}, ..., \mathbf{X}_{i,T_i}]$. Note that this approach is consistent with EEP, where no prediction is carried out during an event.

B. Dataset Details

B.1. Task definition

In this section, we provide more details on the definition of our early prediction tasks for circulatory failure from HiB [26] and decompensation from M3B [35]. A breakdown of event prevalence for each clinical endpoint is given in Table 5.

Table 5: **Event prevalence analysis**, highlighting class imbalance. Positive timesteps are counted for 12-hour and 24-hour horizons for circulatory failure and decompensation respectively. Statistics are computed on the training set.

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

Circulatory failure is a failure of the cardiovascular system, detected in practice through elevated arterial lactate (> 2 mmol/l) and either low mean arterial pressure (< 65 mmHg) or administration of a vasopressor drug. Yèche et al. [26] defines a patient to be experiencing a circulatory failure event at a given time if those conditions are met for $2/3$ of time points in a surrounding two-hour window. Early prediction labels are then derived from these event labels as outlined in Section 2.1.

Decompensation refers to the death of a patient. Event labels are directly extracted from the MIMIC-III [37] metadata about the time of death of a patient. Early prediction labels are also extracted following Section 2.1. Note that decompensation can occur outside of the ICU stay if a patient is sent to a palliative unit, for instance, which can result in patient stays with fewer than 24 positive samples.

B.2. Pre-processing

We describe the pre-processing steps we applied to both datasets, HiRID and MIMIC-III.

Imputation. Diverse imputation methods exist for ICU time series. For simplicity, we follow the approach of original benchmarks [35; 26] by using forward imputation when a previous measure existed. The remaining missing values are zero-imputed after scaling, corresponding to a mean imputation.

Scaling. Whereas prior work explored clipping the data to remove potential outliers [8], we do not adopt this approach as we found it to reduce performance on early prediction tasks. A possible explanation is that, due to the rareness of events, clipping extreme quantiles may remove parts of the signal rather than noise. Instead, we simply standard-scale data based on the training sets statistics.

C. Implementation Details

Training details. For all models, we set the batch size according to the available hardware capacity. Because transformers are memory-consuming, we train the decompensation models with a batch size of 8 stays. On the other hand, we train the GRU model for circulatory failure with a batch size of 64. We early stopped each model training according to their validation loss when no improvement was made after 10 epochs.

Libraries. A full list of libraries and the version we used is provided in the `environment.yml` file. The main libraries on which we build our experiments are the following: pytorch 1.11.0 [50], scikit-learn 0.24.1[51], ignite 0.4.4, CUDA 10.2.89[52], cudNN 7.6.5[53], gin-config 0.5.0 [54].

Infrastructure. We follow all guidelines provided by `pytorch` documentation to ensure the reproducibility of our results. However, reproducibility across devices is not ensured. Thus we provide here the characteristics of our infrastructure. We trained all models on a single `NVIDIA RTX2080Ti` with a `Xeon E5-2630v4` core. Training took between 3 and 10 hours for a single run.

Uncertainty estimation. We compute uncertainty estimates over a population of 10 training instances with different seeds. This widely-used approach has the advantage to account for the stochasticity of the training procedure, which we found to be predominant in early prediction tasks. This approach differs from other work [12; 20; 8; 28] which computes uncertainty estimate by bootstrapping the test population. We found that using a pivot bootstrap estimator decreases confidence intervals by effectively increasing the population size. To be conservative with our results, we retained the former approach to

compute statistics across 10 training instances. We report the 95% confidence interval over the population means in all experiments.

Architecture choices We used the same architecture and hyperparameters reported giving the best performance on circulatory failure in Yèche et al. [26] and only optimized embedding regularization parameters [8]. Exact parameters are reported in Table 6. For decompensation, as we found a transformer architecture to perform better than originally proposed models [35], we carried out our own random search on validation AUPRC performance. The exact parameters for this task are reported in Table 7.

Table 6: **Hyperparameter search range** for circulatory failure with GRU [39] backbone. In **bold** are parameters selected by random search.

Hyperparameter	Values
Learning Rate	(1e-5, 3e-5, 1e-4, **3e-4**)
Drop-out	(**0.0**, 0.1, 0.2, 0.3, 0.4)
Depth	(1, **2**, 3)
Hidden Dimension	(32, 64, 128, **256**)
L1 Regularization	(1e-2, 1e-1, 1, **10**, 100)

Table 7: **Hyperparameter search range** for decompensation with Transformer [40] backbone. In **bold** are parameters selected by random search.

Hyperparameter	Values
Learning Rate	(1e-5, 3e-5, **1e-4**, 3e-4)
Drop-out	(0.0, 0.1, 0.2, **0.3**, 0.4)
Attention Drop-out	(0.0, **0.1**, 0.2, 0.3, 0.4)
Depth	(1, **2**, 3)
Heads	(**1**, 2, 4)
Hidden Dimension	(32, **64**, 128, 256)
L1 Regularization	(1e-2, **1e-1**, 1, 10)

C.1. Baseline implementation

Balanced cross-entropy. In the binary setting, the only hyperparameter of balanced cross-entropy is the relative contribution of the minority class to the loss, ω. As discussed in Section 5.2, no value of ω was found to improve validation performance over the non-balanced case $\omega = 1$.

Focal loss. A grid search over focal loss hyperparameters was also carried out. Similarly to balanced cross-entropy, on all tasks, no values of focal loss weight ζ or balancing weight ω were found to outperform regular cross-entropy corresponding to $\zeta = 0$ and $\omega = 1$.

Multi-horizon prediction. Following Tomašev et al. [8], we consider H horizons on both side of the true horizon h between 0 and $2h$. As we didn't find $H \longrightarrow +\infty$, to increase performance, we selected $H = 11$ (including true horizon h) compared to $H = 8$ in Tomašev et al. [8], which we found to perform slightly worse. This means we made a prediction every 2 hours for circulatory failure and every 4 hours for decompensation.

Label smoothing. Label smoothing [21], as defined in Section 3, is normally used in multi-class setting. We still compared our method to it for two reasons. First, to explore if it can help when dealing with a noisy signal as we claim is the case for

early event detection. Second, to ablate the impact of adding a temporal dependency to the method. Again, we select the hyperparameter α through a grid search. Interestingly, we found label smoothing to slightly improve performance over the validation set for all tasks as opposed to the results reported for the test set in Table 2. We found $\alpha = 0.05$ to perform best for both circulatory failure and decompensation.

Landmarking. For all tasks, landmarking was trained with the same architecture and parameters with the exception that our model return hazard estimates. In theory, we should make predictions until $h_{\max} = \max_i(T_i)$ corresponding to 2016 and 2805, for respectively circulatory failure and decompensation. Due to computing limation, as is common in practice, we truncated this horizon to 1000 for circulatory failure.

TCSR. As for landmarking, we considered $h_{\max}$ to be 1000 and 2805 for respectively circulatory failure and decompensation. In practice, we found that the dynamic nature of the label and weight assignment lead to great instability. To be able to train correctly models with this objective, we had to reduce learning rates to 5e-5 and 3e-5. More importantly, for circulatory failure, we used stop-gradient operation for predictions such that $\frac{d\mathcal{L}_{i,t}}{d\hat{\lambda}^h_{i,t+1}} = 0$. A similar approach for decompensation resulted in worse results, thus we did not use it for this task.

DDRSA. As for previous survival baselines, we considered $h_{\max}$ to be 1000 and 2805 for respectively circulatory failure and decompensation. The additional trade-off hyperparameter α_{DRSA} was selected through a grid search on the validation performance between 0 and 1 with increments of 0.1. We found optimal values to be 0.1 for circulatory failure et 0.2 for decompensation.

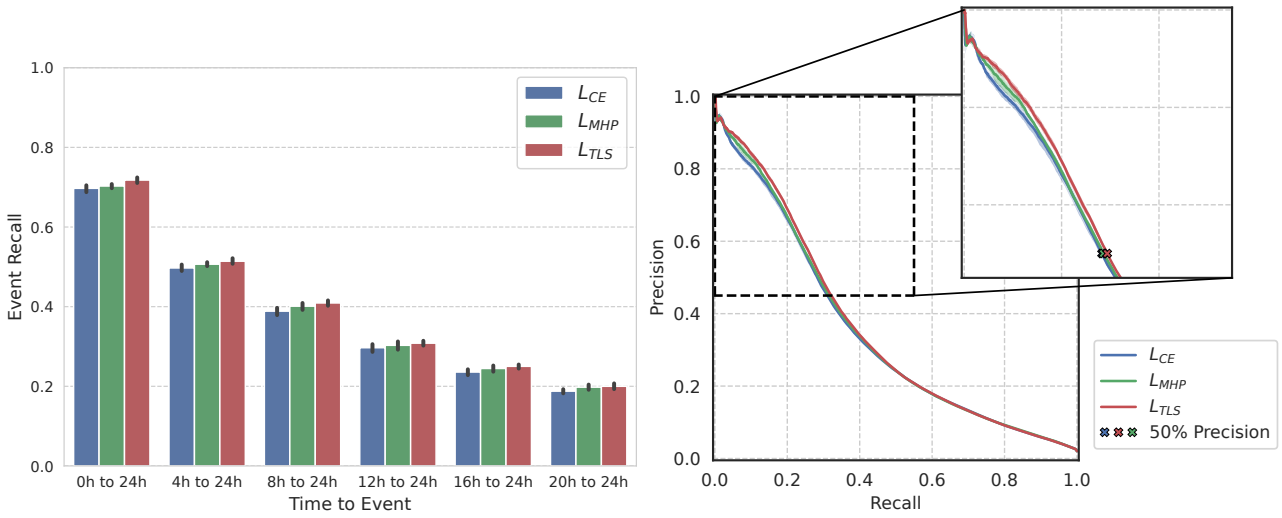
C.2. TLS implementation

```
def get_smoothed_labels(event_label_patient, smoothing_fn, h_true, h_min,
                  h_max, **kwargs):

   # Find when event label changes
   diffs = np.concatenate([np.zeros(1),
            event_label_patient[1:] - event_label_patient[:-1]], axis=-1)
   pos_event_change = np.where((diffs == 1) & (event_label_patient == 1))[0]

   # Handle patients with no events
   if len(pos_event_change) == 0:
      pos_event_change = np.array([np.inf])

   # Compute distance to closest event for each time point
   time_array = np.arange(len(event_label_patient))
   dist_all_event = pos_event_change.reshape(-1, 1) - time_array
   dist_to_closest = np.where(dist_all_event > 0,
                         dist_all_event, np.inf).min(axis=0)

   return smoothing_fn(dist_to_closest, h_true=h_true, h_min=h_min, h_max=h_max,
                                                    **kwargs)
```

Figure 10: **Temporal label smoothing algorithm.** Python-style code to obtain smooth early prediction labels from event labels.

TLS depends on two components, the temporal range over which we smooth labels, defined by h_{min} and h_{max}, and the smoothing function $q(t)$. Concerning the temporal range, for a fair comparison, we fix it to match MHP, thus for all experiments, we set $h_{min} = 0$ and $h_{max} = 2h$. For the smoothing function, we perform a grid search over the type of function discussed in Appendix A.2 and the smoothing strength parameter γ. For all experiments, we found q^{exp} to outperform other considered functions. Given validation performance, we used $\gamma = 0.2$ for circulatory failure and $\gamma = 0.05$ for decompensation.

As discussed in Section 3, contrary to MHP, TLS does not require any change to the architecture leading to a computational overhead. The smoothing of the labels can be easily integrated into the data loader, as shown in Figure 10.



(a) *Event recall* at 50% timestep-level precision.

(b) *Precision-recall curve.*

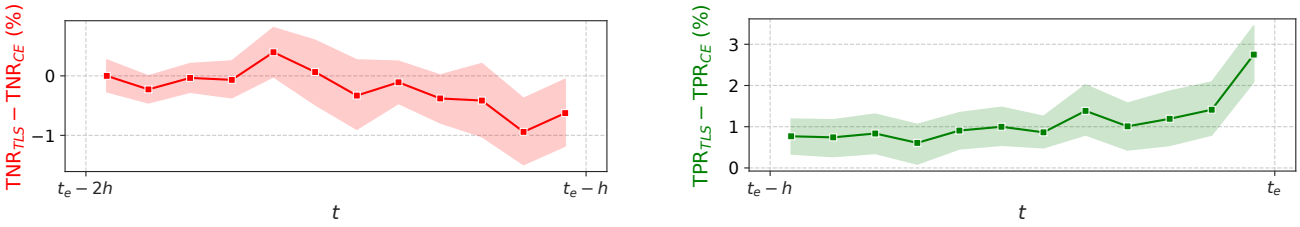
Figure 11: **Clinically relevant performance** on decompensation. Inset in (b) shows the clinically-applicable region with precision greater than 50%.

D. Additional Experiments and Ablation Studies

This section provides additional results and experiments to complete our findings from the main manuscript. Unless otherwise stated, mean results are shown with a 95% confidence interval on the mean shaded or in error bars.

D.1. Performance analysis for decompensation prediction

Event-level performance for decompensation prediction is given in Figure 11a. Results are similar to those on circulatory failure discussed in Section 5.1: temporal label smoothing improves recall of adverse event episodes over cross-entropy and MHP. Note that the improvements observed over the baselines in terms of event-recall between 0 and h are smaller than for circulatory failure, but are statistically significant as shown in Table2



(a) True negative rate (TNR).

(b) True positive rate (TPR).

Figure 12: **Performance improvement over time** for TLS over traditional cross-entropy on decompensation prediction. Timestep-level metrics computed for precision of 0.5 over two-hour bins.

The precision-recall curve obtained for timestep-level event prediction on this task is also given in Figure 11b. As for circulatory failure prediction, recall gains are concentrated in regions of low false-alarm rates (>50% precision) which are most clinically relevant.

Likewise, whereas recall near the label boundary $t_e - h$ is slightly negatively affected by temporal label smoothing in Figure 12, true positive rates are significantly improved leading up to the event time t_e. This mirrors the temporal smoothing pattern which favors higher model confidence away from the label boundary. As discussed in Section 5.2, this is aligned with clinical priorities in terms of model performance, as it ensures imminent events are better predicted.

D.2. Sub-group analysis

Populations in the intensive care unit are often heterogeneous. This has motivated recent works to focus on the fairness of deep learning across these sub-populations. In this analysis, we ensure that temporal label smoothing does not negatively affect performance in specific subgroups, compared to the objectives commonly used in the literature [8; 7; 11]. To achieve this, we measured event prediction performance across genders and age groups (below 50, between 50 and 70, and over 70 years old). As shown in Table 8, TLS matches or outperforms baseline performance across all studied subgroups, suggesting that the overall population-wide improvements are not achieved by disproportionally favouring specific cohorts. While some algorithmic bias can be observed across all methods, for instance in poorer decompensation performance amongst female patients, TLS does not appear to be amplifying this issue. In further work, we look forward to extending this analysis to more specific subgroups and studying the fairness of early event prediction methods for clinical applications.

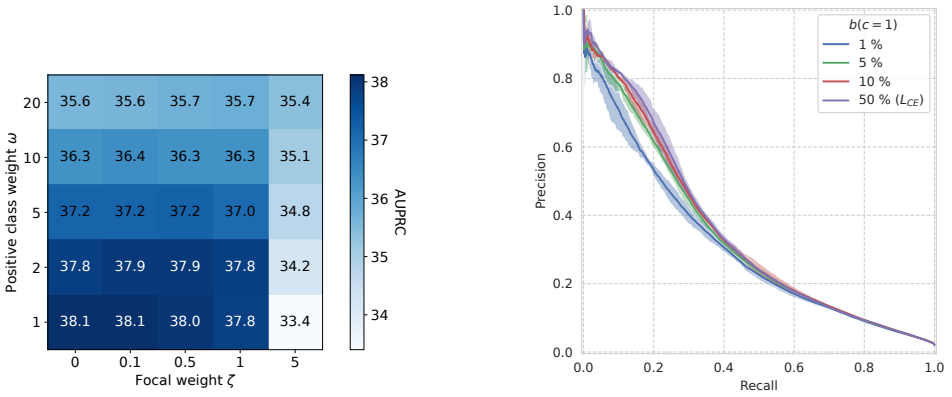
Table 8: **Sub-group performance analysis.** We color improvement above the 95% confidence interval in green.

Circulatory Failure	Age ≤ 50		$50 <$ Age ≤ 70		Age > 70		Female		Male	
Method	AUPRC	Recall	AUPRC	Recall	AUPRC	Recall	AUPRC	Recall	AUPRC	Recall
CE	40.4 ± 0.5	29.4 ± 0.6	38.8 ± 0.6	29.6 ± 1.1	39.2 ± 0.3	29.0 ± 1.0	39.3 ± 0.6	30.0 ± 0.7	39.1 ± 0.4	29.0 ± 1.0
TLS	40.4 ± 0.5	$\mathbf{32.7} \pm 1.0$	$\mathbf{41.1} \pm 0.4$	$\mathbf{32.6} \pm 0.7$	$\mathbf{40.0} \pm 0.3$	$\mathbf{31.7} \pm 0.7$	$\mathbf{41.2} \pm 0.3$	$\mathbf{32.8} \pm 0.6$	$\mathbf{40.4} \pm 0.3$	$\mathbf{32.0} \pm 0.8$
Δ(TLS-CE)	0.0	+ 3.3	+ 2.3	+ 3.0	+ 0.9	+ 2.7	+ 1.8	+ 2.9	+ 1.3	+ 3.0
Decompensation	Age ≤ 50		$50 <$ Age ≤ 70		Age > 70		Female		Male	
Method	AUPRC	Recall	AUPRC	Recall	AUPRC	Recall	AUPRC	Recall	AUPRC	Recall
CE	29.2 ± 0.8	25.3 ± 1.2	34.9 ± 0.9	27.4 ± 0.6	35.8 ± 0.2	29.4 ± 0.6	30.9 ± 0.4	24.8 ± 0.6	38.3 ± 0.6	31.4 ± 0.5
TLS	$\mathbf{30.5} \pm 0.5$	26.2 ± 1.1	$\mathbf{36.7} \pm 0.5$	$\mathbf{29.1} \pm 0.5$	$\mathbf{36.3} \pm 0.3$	$\mathbf{30.3} \pm 0.4$	$\mathbf{31.6} \pm 0.3$	$\mathbf{25.7} \pm 0.5$	$\mathbf{39.6} \pm 0.5$	$\mathbf{32.8} \pm 0.6$
Δ(TLS-CE)	+ 1.3	+ 1.0	+ 1.8	+ 1.7	+ 0.5	+ 0.9	+ 0.7	+ 0.9	+ 1.3	+ 1.4

D.3. Loss reweighting methods

Hyperparameter grid search results on decompensation prediction for different loss reweighting methods are shown in Figure 13a. Weighted cross-entropy and focal loss were also found to negatively affect performance in comparison to traditional cross-entropy. Likely explanations for these results are provided in Section 5.2: focal loss focuses training on noisily labeled samples, and weighted cross-entropy largely reduces precision.

We validate the latter hypothesis by visualizing precision-recall curves of models trained with this objective in Figure 13. With a relative weight for the positive class $\omega > 1$, weighted cross-entropy encourages a greater number of true positives to improve recall. Doing so also increases the of false positives, impairing precision. In Figure 13, as the starting precision of all cross-entropy models is poor, no discernible improvements in the recall can be observed as class weights are increased, whereas precision is markedly reduced in low-recall regions. This explains the overall reduction in AUPRC with this method.



(a) *Reduction in AUPRC (validation).* (b) *Reduction in precision.*

Figure 13: **Performance loss with class reweighting methods**, on decompensation prediction. (a) Balanced cross-entropy corresponds to $\zeta = 0$, focal loss to $\zeta \geq 0$. (b) Loss reweighting does not improve AUPRC because it significantly reduces precision. Balance weights correspond to $b(c)$. Similar results for circulatory failure prediction.

E. Alternative Early Prediction Tasks

As a third task to benchmark our method, we studied early prediction of respiratory failure, defined in Yèche et al. [26]. Unfortunately, this task has vague labels which result in all methods performing close to random. For transparency, we first provide results on this task and motivate our belief that this label ambiguity is caused by a very noisy estimate of a certain clinical variable (FIO_2). See Section E.1 for details.

As a related, alternative dataset, we define a related sub-task that does not rely on FIO_2: prediction of the onset of mechanical ventilation. For this task, we show that: (1) models do perform much better than random, which confirms our hypothesis on respiratory failure labeling, and (2) TLS improves again significantly over EEP baselines, with similar results to in Section 5.1.

Implementation details. For respiratory failure prediction, we used the transformer architecture and hyperparameters for respiratory failure reported in Yèche et al. [26]. For ventilation onset, we used a GRU model and selected hyperparameters based on a grid search over the validation AUPRC. This resulted in a 2-layer GRU with a hidden space dimensionality of 128 and no dropout. In both cases, we chose 10.0 as the l_1 regularization strength for the embedding module and used a batch size of 8 stays. For label smoothing, we found $\alpha = 0.1$ to give the best validation performance. We used $\gamma = 0.05$ (respiratory failure) and $\gamma = 0.1$ (ventilation onset) for temporal label smoothing with exponential parametrization.

E.1. Labeling issues for respiratory failure

Respiratory failure is defined as a P/F ratio (arterial pO_2 over FIO_2) below 300 mmHg [26]. This includes mild failure events, which results in high event prevalence in the HIRID dataset [7]: 38.6% of timepoints have a positive label, and 83% of patients undergo at least one event, with on average 1.8 events per positive patient. Despite this high prevalence, all EEP methods have a performance close to 60% AUPRC, as shown in Table 9 and as in Yèche et al. [26]. This corresponds to an enrichment factor (ratio of AUPRC of predictor vs. random classifier) with respect to a random classifier ($\approx 40\%$) of 1.5 for this task, compared to factors of 10 and 15 for circulatory failure and decompensation, respectively. Such a low performance suggests an inherent issue with labeling. Our hypothesis is that the estimation of FIO_2 is highly error-prone, which challenges the quality of respiratory failure labels and causes the low performance of all machine learning models considered. For completeness, we nevertheless show the results for respiratory failure (in addition to ventilation onset in this section and circulatory failure as well as decompensation in the main part).

E.2. Ablation study: onset of mechanical ventilation

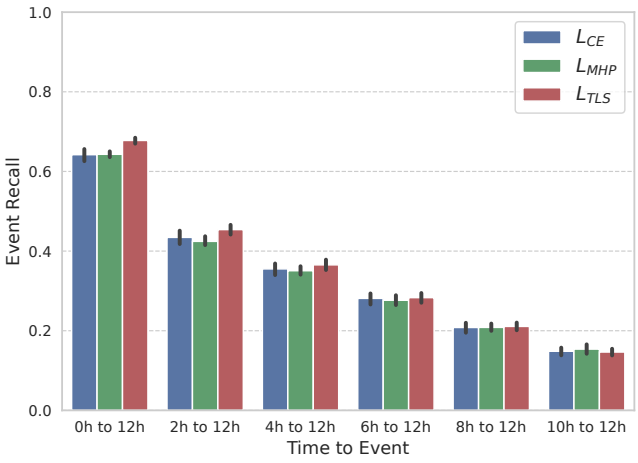
To verify the above hypothesis, we define a similar task independent of FIO_2 estimates and verify we can recover a better baseline performance. We focus on predicting whether a patient will be mechanically ventilated within the next 12 hours. Ventilation is a good proxy for severe respiratory distress but is not labeled based on a P/F ratio estimate. With a 5.6% timestep-level prevalence, baseline performance at 34% AUPRC in Table 9 is roughly 6.2 times better than a random classifier. This confirms that poor FIO_2 estimation underlies poor performance on respiratory failure prediction across all methods.

Table 9: **Performance of different training objectives for early prediction of respiratory failure and ventilation onset.** Recall is reported at a 50% timestep-level precision. In **bold**, we highlight best-performing methods with statistically significant p-values (< 0.05) under paired Student's t-tests [43] compared with the next-best method marked in italic.

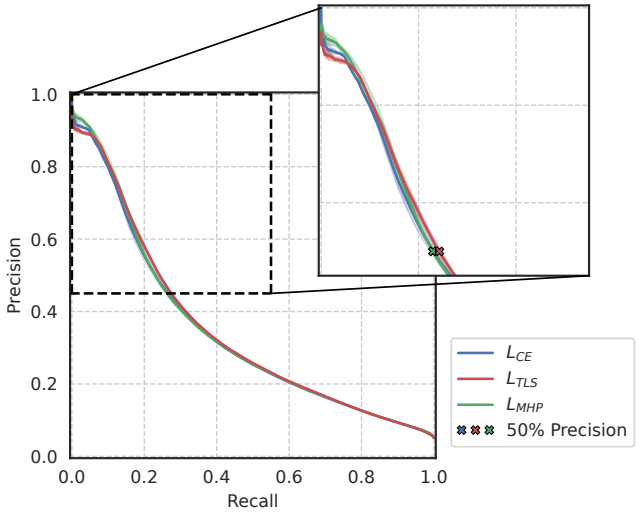
Task	Respiratory Failure (HiRID)			Ventilation Onset (HiRID)		
Training objective	AUPRC	Timestep Recall	Event Recall	AUPRC	Timestep Recall	Event Recall
Cross-entropy [11; 7]	60.5 ± 0.2	77.3 ± 0.5	94.9 ± 0.2	34.1 ± 0.4	*23.0* ± 1.1	64.2 ± 1.8
Multi-horizon [8; 15]	60.3 ± 0.1	76.6 ± 0.5	*95.0* ± 0.1	*34.4* ± 0.5	*23.0* ± 0.6	*64.3* ± 0.9
Temporal Label Smoothing	*60.4* ± 0.2	*77.0* ± 0.3	**95.3** ± 0.1	34.7 ± 0.4	**24.2** ± 0.7	**67.8** ± 0.9
p-value	0.15	0.14	**0.04**	0.25	**0.008**	**<0.001**
Enrichment Factor		**1.5**			**6.2**	

E.3. Temporal label smoothing performance for onset of mechanical ventilation

In this final section, we verify the benefits of TLS in predicting the onset of mechanical ventilation – a feasible task relative to the respiratory system. In Table 9 and Figure 14a, we find that TLS again improves performance in both timestep and event recall over multi-horizon prediction, and performs on par in terms of AUPRC. This is likely due to its lower performance at very low recall in Figure 14b. Finally, TLS again improves the true negative and positive rates away from the label boundary $t_e - h$ in Figure 15, which corresponds to more clinically relevant regions. All conclusions agree with our analysis on other tasks in Section 5.2.

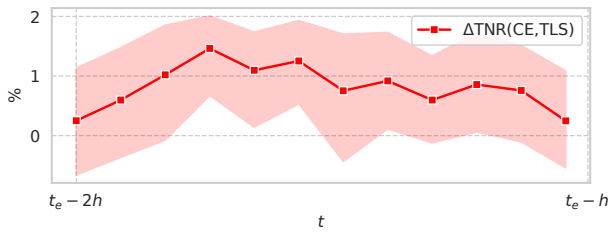


(a) *Event recall* at 50% timestep-level precision.

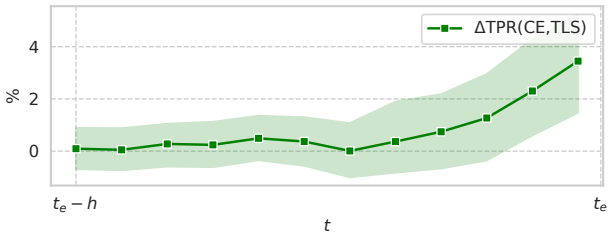


(b) *Precision-recall curve.* Inset shows the clinically-applicable region with precision greater than 0.5

Figure 14: **Clinically relevant performance** on ventilation onset.



(a) True negative rate (TNR)



(b) True positive rate (TPR)

Figure 15: **Performance improvement over time** for TLS over traditional cross-entropy on onset ventilation prediction. Timestep-level metrics computed for precision of 0.5 over two-hour bins.