TrajSurv: Learning Continuous Latent Trajectories from Electronic Health Records for Trustworthy Survival Prediction

Sihang Zeng ZENGSH@UW.EDU
Department of Biomedical Informatics and Medical Education
University of Washington
Seattle, WA, USA

Lucas Jing Liu JLIU6@FREDHUTCH.ORG
Fred Hutch Cancer Center
Seattle, WA, USA

Jun Wen JUN_WEN@HMS.HARVARD.EDU
Department of Biomedical Informatics
Harvard University
Boston, MA, USA

Meliha Yetisgen MELIHAY@UW.EDU
Department of Biomedical Informatics and Medical Education
University of Washington
Seattle, WA, USA

Ruth Etzioni[*] RETZIONI@FREDHUTCH.ORG
Fred Hutch Cancer Center
Seattle, WA, USA

Gang Luo[*] LUOGANG@UW.EDU
Department of Biomedical Informatics and Medical Education
University of Washington
Seattle, WA, USA

Abstract

Trustworthy survival prediction is essential for clinical decision making. Longitudinal electronic health records (EHRs) provide a uniquely powerful opportunity for the prediction. However, it is challenging to accurately model the continuous clinical progression of patients underlying the irregularly sampled clinical features and to transparently link the progression to survival outcomes. To address these challenges, we develop TrajSurv, a model that learns continuous latent trajectories from longitudinal EHR data for trustworthy survival prediction. TrajSurv employs a neural controlled differential equation (NCDE) to extract continuous-time latent states from the irregularly sampled data, forming continuous latent trajectories. To ensure the latent trajectories reflect the clinical progression, TrajSurv aligns the latent state space with patient state space through a time-aware contrastive learning approach. To transparently link clinical progression to the survival outcome, TrajSurv uses latent trajectories in a two-step divide-and-conquer interpretation process. First, it explains how the changes in clinical features translate into the latent trajectory's evolution

[*] Co-senior authors.

using a learned vector field. Second, it clusters these latent trajectories to identify key clinical progression patterns associated with different survival outcomes. Evaluations on two real-world medical datasets, MIMIC-III and eICU, show TrajSurv's competitive accuracy and superior transparency over existing deep learning methods.

1. Introduction

Accurate and transparent survival prediction is crucial for trustworthy clinical decision making Alabdallah (2025). Longitudinal electronic health records (EHRs), which capture patients' evolving clinical status through irregularly sampled clinical features, provide rich temporal information for survival prediction Lee et al. (2019). Although deep learning models have been developed to leverage this information and enhance accuracy, two key challenges remain unresolved: accurately modeling continuous clinical progression and transparently linking that progression to survival outcomes Xie et al. (2022).

To accurately model the continuous clinical progression, it is important to aggregate irregularly sampled clinical features within a continuous-time framework in a clinically aligned way. Recurrent neural networks (RNNs) Nagpal et al. (2021); Lee et al. (2019) aggregate features at discrete times, potentially discarding the underlying continuous-time patterns. A recent approach, SurvLatentODE Moon et al. (2022) leverages neural ordinary differential equations (NODEs) to aggregate features into continuous-time latent states. However, without high-quality supervision signals to guide the latent states at earlier time points, it may learn latent states' trajectories that do not consistently align with the patient's actual clinical progression, even if the ultimate latent state is shown to be clinically relevant. This may lead to suboptimal modeling of clinical progression and suboptimal prediction accuracy.

To transparently link the clinical progression to the outcome, it is essential to provide an end-to-end interpretation of how changes in clinical features, i.e., feature velocities, lead to the outcome. However, existing models mainly interpret the contribution of clinical features' absolute values at observed time points Lee et al. (2019), but they do not fully explain how feature velocities relate to survival. This is a critical gap, as clinicians recognize the predictive value of feature velocities in longitudinal data besides their absolute values. For example, creatinine kinetics are used to define acute kidney injury Waikar and Bonventre (2009).

To address these challenges, we introduce TrajSurv, which learns continuous latent trajectories from longitudinal EHR for trustworthy survival prediction. To ensure the accurate modeling of continuous clinical progression, TrajSurv uses a neural controlled differential equation (NCDE) Kidger et al. (2020) to aggregate clinical feature changes over time into continuous latent trajectories, and designs a time-aware contrastive learning (TACL) objective to explicitly align the latent trajectories with actual clinical progression. TACL not only improves the prediction accuracy, but also learns clinically aligned latent trajectories that split the model into feature-to-trajectory and trajectory-to-outcome processes. This enables a divide-and-conquer approach for end-to-end transparency, leveraging a learned vector field and latent trajectory clustering. The implementation of TrajSurv is available at https://github.com/zengsihang/TrajSurv.

Generalizable Insights about Machine Learning in the Context of Healthcare

- **Accurate modeling of continuous clinical progression improves trustworthiness for longitudinal EHR analysis.** With NCDE and TACL, TrajSurv accurately models the continuous clinical progression, improving prediction accuracy and enabling transparency.
- **Divide-and-conquer approach improves end-to-end model transparency.** TrajSurv achieves its end-to-end transparency by splitting the model into feature-to-trajectory and trajectory-to-outcome steps for interpretation.

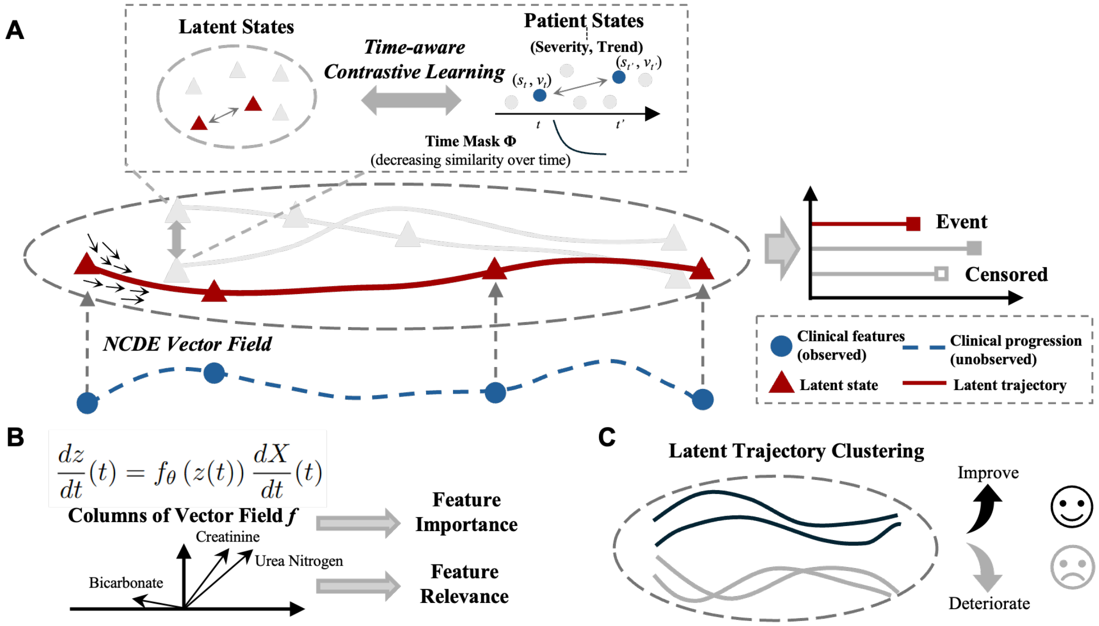


Figure 1: An illustration of TrajSurv and its two-step interpretation. (A) Model architecture. (B) TrajSurv's vector field interpretation. (C) TrajSurv's latent trajectory clustering.

2. Related Work

Deep learning has demonstrated efficacy in survival prediction from longitudinal EHRs, capturing complex temporal dependencies. Recurrent Deep Survival Machines (RDSM) Nagpal et al. (2021) and Dynamic-DeepHit Lee et al. (2019) leveraged RNNs and aggregated irregularly sampled clinical features at discrete times, which may not accurately model the continuous clinical progression. Although Dynamic-DeepHit offered interpretability through feature importance, attention weights, and predicted risks, neither method fully explained how the clinical features' changes over time relate to survival outcomes. Recent continuous-time models, including NODEs Chen et al. (2018), ODE-RNN Rubanova et al. (2019), and NCDEs Kidger et al. (2020), map irregularly sampled time series into the continuously evolved latent states in the latent space through differential equations. SurvLatentODE Moon et al. (2022) leveraged ODE-RNN to model time-varying clinical features. Still, the interpretation was limited to the last latent state, potentially lacking clinical relevance at earlier time points due to limited supervision, which may affect the model's accuracy and transparency. CoxSig Bleistein et al. (2024), employing NCDEs, focused on theoretical advancements and offered limited insight into NCDE's latent trajectory. In this study, we demonstrate how TrajSurv's continuous latent trajectories ensure accurate and transparent survival prediction.

3. Methods

3.1. Problem Formulation

3.1.1. Survival Prediction

We consider a right-censored survival dataset $\mathcal{D} = \{(\mathcal{X}^i, T^i, \delta^i)\}_{i=1}^N$, consisting of N patients with longitudinal EHR data $\mathcal{X}^i$, time-to-event T^i, and event indicator δ^i. For each patient i, the longitudinal EHR data $\mathcal{X}^i = \{(t_j^i, x_j^i)\}_{j=1}^{n_i}$ represents an irregularly sampled time series of clinical features, where $x_j^i \in \mathbb{R}^d$ is a d-dimensional clinical feature vector at time t_j^i, and $t_0^i = 0$ is defined the same for all patients as a starting time point (e.g., time of admission). $\mathcal{X}^i$ is irregularly sampled, which means that the total number of observations n_i and intervals between observations vary across patients, with clinical features containing missing values. The time-to-event T^i is defined by the time from the last observation $t_{n_i}^i$ to the event of interest (e.g., in-hospital death) or censoring (e.g., discharge). The event indicator $\delta^i = 1$ if the event occurs and $\delta^i = 0$ if censored.

The task of survival prediction is to model the survival distribution given $\mathcal{X}^i$ up to the last observation, predicting the hazard function:

$$\lambda(t|\mathcal{X}^i) = \lim_{\Delta t \to 0} \frac{P(t \le T^i \le t + \Delta t | T^i \ge t, \mathcal{X}^i)}{\Delta t} \tag{1}$$

or equivalently, the survival function $S(t|\mathcal{X}^i) = P(T^i > t|\mathcal{X}^i)$.

In this study, we aggregate longitudinal EHR data $\mathcal{X}^i$ in a shared latent space $\mathcal{Z}$, where the patient status at time t is encoded as a latent state $z_t^i \in \mathcal{Z}$. The latent state z_t^i summarizes the patient's historical information up to t without knowing the clinical features after t, and the continuous latent trajectory $\tau^i = \{z_t^i, t \in [0, t_{n_i}^i]\}$ summarizes how latent states evolve in $\mathcal{Z}$. We seek to train a clinically aligned latent space $\mathcal{Z}$ for accurate and transparent survival prediction, where the continuous latent trajectory τ^i represents patients' clinical progression. For simplicity, we omit the superscript i in the following sections, except where differentiation between patients is required.

3.1.2. Clinical Alignment

To make the continuous latent trajectory clinically meaningful, we have to align the entire latent space with the clinical meaning of patient states, where a patient state captures the patient's current condition and the trend in the current condition. We define clinical alignment for latent states z_t at a given time point t as the property whereby close latent states correspond to similar patient conditions and their ongoing trends, while distant latent states correspond to distinct conditions or trends. This definition allows us to assess the local clinical relevance of the latent space at any specific time. However, directly comparing latent states at different time points does not provide a reliable measure of clinical alignment across these points. This is because the distance between z_t and $z_{t'}$ (where $t \neq t'$) could arise from genuine differences in patient states or simply from drift over time. Therefore, we consider entire latent trajectories τ, which inherently encode both time and the evolution of patient states. We define clinical alignment for latent trajectories as the property whereby similar trajectories represent similar evolutions of patient conditions, while dissimilar trajectories represent distinct evolutions. This definition focuses on the overall pattern of change over time, rather than point-to-point comparisons of individual latent states. As the trend information is inherently included in the evolution of patient conditions, it is not used as a separate metric to define clinical alignment for latent trajectories.

To practically capture clinical alignment, we utilize existing clinical assessments of severity as proxies for the patients condition. These assessments represent established clinical knowledge about patient conditions. Examples of such assessments include the Sequential Organ Failure Assessment (SOFA) score and the Acute Physiology And Chronic Health Evaluation (APACHE) score for intensive care unit (ICU) settings, and the Model for End-Stage Liver Disease (MELD) score for chronic liver disease. We denote the severity at time t as s_t and define the ongoing trend of severity as $v_t = (s_{t+\Delta t} - s_{t-\Delta t})/2\Delta t$, representing the changing rate of severity at t. The choice of Δt is tailored to the specific clinical scenario, ensuring that the severity trend captures meaningful changes in the patient's condition over time, rather than merely reflecting noise from small time intervals. Both severity and its trend can be high-dimensional if the assessment involves multiple components (e.g., different components in the SOFA score).

3.2. TrajSurv: Model Architecture

In this section, we introduce the architecture of TrajSurv. TrajSurv uses an NCDE to map irregularly sampled clinical features in longitudinal EHR into continuous latent trajectories, where the last latent states are linked to the survival outcomes. We further design a time-aware contrastive learning objective (TACL) to align latent trajectories with actual clinical progression. Figure 1A shows an illustration of TrajSurv.

3.2.1. Mapping Longitudinal EHR to Continuous Latent Trajectories

TrajSurv uses NCDE as an encoder to aggregate temporal information from longitudinal EHR data. NCDE is a latent state-based method that learns continuous-time latent states driven by an irregularly sampled input process Kidger et al. (2020). Similar to previous work Seedat et al. (2022), the use of NCDE in TrajSurv is motivated by its ability to capture the underlying continuous process from irregularly sampled clinical features, rather than discretely modeling the data (e.g., RNNs and transformers) or modeling through latent trajectories that only depend on initial values (e.g., NODEs).

Specifically, TrajSurv maps longitudinal EHR $\mathcal{X} = \{(t_j, x_j)\}_{j=0}^{n}$ into a continuous latent trajectory $\tau = \{z_t, t \in [0, t_n]\}$. Following previous practices Kidger et al. (2020); Morrill et al. (2021), the longitudinal EHR is interpolated into a continuous control signal $\{X_t, t \in [0, t_n]\}$ through the cubic Hermite splines with backward differences scheme such that $X_{t_j} = (x_j, t_j) \in \mathbb{R}^{d+1}$. The initial observation X_0 is mapped into a d_z-dimensional latent space using a feed forward network (FFN) $g_\phi : \mathbb{R}^{d+1} \to \mathbb{R}^{d_z}$. Subsequent continuous-time latent states are the solutions to an NCDE parameterized by a vector field $f_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z \times (d+1)}$:

$$z_t = z_0 + \int_0^t f_\theta(z_s) dX_s \tag{2}$$

where $z_0 = g_\phi(X_0)$, $t \in [0, t_n]$, and the integral is a Riemann-Stieltjes integral. Therefore, the latent trajectory τ is the solution to an NCDE controlled by the underlying process of irregularly sampled clinical features. This transformation is realized through the vector field f_θ, which maps the changes of clinical features dX_t into the evolution of latent states.

3.2.2. Linking Latent Trajectory to Survival Outcome

Because TrajSurv transforms irregularly sampled clinical features into evolved latent states, the last latent state z_{t_n} accumulates the temporal information up to t_n. Therefore, z_{t_n} is used as an aggregated predictor for the survival outcome. Similar to DeepSurv Katzman et al. (2018), we use a nonlinear Cox proportional hazard model leveraging an FFN for more accurate prediction. In particular, the hazard function $\lambda(t|\mathcal{X})$ is decomposed into a baseline hazard $\lambda_0(t)$ and the exponential of a risk score r, where r is derived from z_{t_n} through an FFN $G_\eta : \mathbb{R}^{d_z} \to \mathbb{R}$:

$$\lambda(t|\mathcal{X}) = \lambda(t|z_{t_n}) = \lambda_0(t) \cdot r = \lambda_0(t) \cdot \exp\left(G_\eta(z_{t_n})\right) \tag{3}$$

We optimize TrajSurv's survival prediction using two loss functions. The first is the negative partial likelihood loss, commonly applied in Cox proportional hazard models Cox (1972). For any $t \geq 0$, the risk set $R(t)$ is defined as the index set of patients still at risk (i.e., surviving equal or longer than t): $R(t) = \{i|T_i \geq t\}$. The negative log partial likelihood loss $\mathcal{L}_{PL}$ is then formulated as:

$$\mathcal{L}_{PL} = -\frac{1}{N_{\delta=1}} \sum_{i=1}^{N} \mathbb{I}(\delta^i = 1) \log \frac{\exp(r^i)}{\sum_{k \in R(T^i)} \exp(r^k)} \tag{4}$$

where $N_{\delta=1}$ is the number of patients who experienced the event and $\mathbb{I}(.)$ is the indicator function. Minimizing $\mathcal{L}_{PL}$ corresponds to maximizing the risk score of patients who had events, relative to those who survived longer. To further improve the concordance of predicted risk scores between patients, similar to previous work Wang et al. (2021), we introduce a smoothed pairwise ranking loss $\mathcal{L}_{PR}$ to encourage patients with shorter survival times to have higher risk scores through pairwise ranking, defined as:

$$\mathcal{L}_{PR} = \frac{1}{N_{\delta=1}} \sum_{i=1}^{N} \sum_{k \in R(T^i)} \mathbb{I}(\delta^i = 1)\sigma(r^k - r^i) \tag{5}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function, providing a smoothed and differentiable approximation of the indicator function. The final survival loss $\mathcal{L}_{SURV}$ is the sum of the two loss functions.

$$\mathcal{L}_{SURV} = \mathcal{L}_{PL} + \mathcal{L}_{PR} \tag{6}$$

3.2.3. Clinical Alignment of Latent Trajectory

Previous studies have shown that linking the last latent state z_{t_n} to the survival outcome could learn clinically meaningful last states, where closer states represent similar outcomes Moon et al. (2022). However, because there is no explicit supervision on prior latent states, TrajSurv's latent trajectory τ is not guaranteed to be clinically aligned, i.e., the proximity of latent trajectories may not correspond to similar clinical progression, though they ultimately arrive at a clinically meaningful distribution. This may affect trustworthiness because the clinical progression information is not accurately and transparently captured. To ensure TrajSurv's latent trajectories reflect clinical progression, we design a TACL objective.

While aligning two continuous-time processes, the latent trajectory and clinical progression, is challenging, the TACL objective simplifies this alignment by only aligning anchor latent states with high-quality labels while accounting for the time intervals between them. It is inspired by the anchor-based methods that select representative anchor points to enhance computational efficiency and reduce noise Yang et al. (2025), facilitating robust alignment. Although the alignment is performed at discrete time points for a latent trajectory, we expect it to propagate across the timeline due to the time-aware design and the continuous-time nature of the latent space, allowing for consistent alignment throughout the entire trajectory.

We use latent states at observation time points as anchor latent states. For simplicity, we denote the set of anchor latent states as $Z = \{z_k\} = \bigcup_{i=1}^{N} \{z^i_{t_j}\}_{j=0}^{n_i}$. To avoid confusion, in the remainder of this section, $|Z|$ will refer to the number of anchor latent states in a single batch, and the subscripts will refer to distinct anchor latent states and their properties.

Specifically, we align anchor latent states z_i across all patients in a batch with their corresponding severity s_i and severity trend v_i through contrastive learning Khosla et al. (2021). The contrastive learning guides anchor latent states with similar labels, which are severity and trend in our scenario, to be closer in the latent space. Because both severity and its trend are continuous variables, we employ a contrastive learning for regression framework called Rank-N-Contrast Zha et al. (2024). This framework learns the latent distance based on the rank of label distance. If we directly adopt Rank-N-Contrast, we have the negative log-likelihood loss:

$$\mathcal{L}_{RNC} = -\frac{1}{|Z|^2} \sum_{i=1}^{|Z|} \sum_{j=1}^{|Z|} \log \frac{\exp(\mathrm{sim}(z_i, z_j)/\kappa_1)}{\sum_{z_k \in \mathcal{S}_{i,j}} \exp(\mathrm{sim}(z_i, z_k)/\kappa_1)} \tag{7}$$

where $\mathcal{S}_{i,j} = \{z_k | k \neq i, |s_i - s_k| + \delta|v_i - v_k| > |s_i - s_j| + \delta|v_i - v_j|\}$ is the set of latent states with larger label distance with z_i than the label distance between z_i and z_j, δ is a hyperparameter that balances the severity and its trend, κ_1 is a hyperparameter that controls the sensitivity of contrast, and $\mathrm{sim}(\cdot, \cdot)$ is the similarity between two latent states.

However, this direct adoption does not account for the time intervals between latent states, which may inadequately capture the time-dependent nature of clinical progression. Therefore, we introduce a time mask $\Phi(t_i, t_j; \kappa_2) = \exp(-\frac{|t_i - t_j|}{\kappa_2})$ applied to the log-likelihood term in $\mathcal{L}_{RNC}$, forming our TACL objective. The time mask penalizes the contrasts between latent states separated by larger time intervals. This allows latent states with a large time interval to separate even if they share similar patient states. The hyperparameter κ_2 controls the strength of this penalty and can be adjusted for different clinical scenarios to reflect the time sensitivity of clinical progression. Formally, the TACL objective $\mathcal{L}_{TACL}$ is defined as:

$$\mathcal{L}_{TACL} = -\frac{1}{|Z|^2} \sum_{i=1}^{|Z|} \sum_{j=1}^{|Z|} \Phi(t_i, t_j; \kappa_2) \log \frac{\exp(\text{sim}(z_i, z_j)/\kappa_1)}{\sum_{z_k \in \mathcal{S}_{i,j}} \exp(\text{sim}(z_i, z_k)/\kappa_1)} \tag{8}$$

Overall, TrajSurv's loss function is:

$$\mathcal{L} = \mathcal{L}_{SURV} + \alpha \mathcal{L}_{TACL} \tag{9}$$

where α adjusts the relative strength of clinical alignment to survival prediction.

3.3. TrajSurv: Model Interpretation

While the relationship between changes in clinical features and survival outcomes is complex, TrajSurv's latent trajectory divides and conquers this process into two steps: mapping longitudinal EHR into continuous latent trajectories and linking latent trajectories to survival outcomes. The transparency can then be achieved by analyzing NCDE's vector field to understand how changes in clinical features lead to the evolution of latent trajectories, and by clustering latent trajectories to identify key clinical progression patterns linked to different survival outcomes. An illustration of model interpretation is shown in Figure 1B and C.

3.3.1. Step 1: Vector Field Interpretation

The vector field f_θ maps the changes of clinical features into continuous latent trajectories. To further understand this transformation, we first transform the integral form of the NCDE 2 to the derivative form:

$$\frac{dz}{dt}(t) = f_\theta(z(t)) \frac{dX}{dt}(t) \tag{10}$$

For a specific time point t, the vector of latent velocity $\frac{dz}{dt}(t)$ is computed as the product of the matrix $f_\theta(z(t))$ and the vector of clinical feature velocity $\frac{dX}{dt}(t)$. Expanding the matrix f_θ into column vectors, the latent velocity can be represented as the linear combination of the columns of f_θ weighed by the clinical feature velocity. Therefore, each column of f_θ represents the influence of a unit change in a specific clinical feature on the magnitude and direction of the latent state's moving velocity. We analyze the columns of f_θ as follows:

- **Feature Importance:** The magnitude of each column indicates the importance of the corresponding clinical feature in driving latent state evolution. Features with larger column magnitudes have a greater impact on the evolution of the latent state.
- **Feature Relevance:** The cosine similarity between two columns captures the relevance of the corresponding features in driving latent states evolution. If the cosine similarity is high, it suggests that similar changes in these features tend to move the latent state in similar directions. Conversely, if the cosine similarity is low, it indicates that similar changes in these features tend to move the latent state in opposing directions.

This approach provides an elegant and interpretable framework for understanding how changes in clinical features drive latent state evolution and produce the latent trajectory, through the feature importance and relevance. Practically, instead of analyzing the vector field at specific times for specific patients, we estimate the average vector field $\bar{f}_\theta$ by sampling latent states $z(t)$ across patients and time points and averaging their vector field $f_\theta(z(t))$. This average field provides an overall perspective on feature importance and relevance. We note that these average patterns may not generalize to individual cases and should be interpreted cautiously.

3.3.2. Step 2: Latent Trajectory Clustering

Trained with the TACL objective, TrajSurv's continuous latent trajectory is expected to be clinically aligned. This means that similar latent trajectories, reflecting the evolution of underlying patient states, can correlate with similar clinical progression and survival outcomes. To investigate this, we employ dynamic time warping (DTW) Müller (2007), a time series clustering method that finds optimal alignments between sequences of different lengths and outputs distances. Using DTW, we cluster latent trajectories ($\tau_i, i = 1, ..., N$) into C distinct groups. We visualize the cluster centroidsderived using the DTW barycenter averaging algorithmin the latent space. We compute the average severity trajectory over time for each cluster, aiming to reveal key clinical progression patterns. We further assess the association between the clusters and survival outcomes using Kaplan-Meier (KM) curves, hypothesizing that latent trajectory patterns may serve as prognostic indicators.

4. Experiments

In this section, we briefly introduce the experiment settings and refer readers to the Appendix for more details.

4.1. Dataset and Setup

We evaluated TrajSurv's survival prediction performance on two real-world EHR datasets in the ICU setting, MIMIC-III Johnson et al. (2016) and eICU Pollard et al. (2018). Similar to prior work Moon et al. (2022), the prediction task is to estimate time to in-hospital mortality based on longitudinal EHR data in the first 36 hours of admission. This prediction task is crucial for timely care and optimal resource allocation in ICU Moon et al. (2022).

In ICU, the SOFA score assesses the performance of six organ systems (respiration, coagulation, liver, cardiovascular, neurologic, and renal), with each component ranging from 0 to 4 and higher scores indicating more severe organ dysfunction. The overall SOFA score is the sum of these 6 components and higher scores indicate a higher risk of ICU mortality Shafigh et al. (2024). In the experiments, we used the SOFA score and its six components as s_t, representing patient conditions over time. SOFA scores s_t were processed hourly from forward-imputed data, and SOFA trends v_t were computed by the SOFA's average changing rate over 4-hour intervals ($\Delta t = 2$) to smooth fluctuations and capture meaningful trends.

4.2. MIMIC-III and eICU Cohort

The MIMIC-III cohort included 20,258 hospital admissions and 53 clinical features in 1-hour resolution. The eICU cohort included 116,503 hospital admissions and 53 clinical features in 1-hour resolution. Both cohorts featured irregularly sampled time series data. To ensure computational efficiency, the time series data was pre-processed to only include hourly time points with more than 20 observed features (See Appendix A.5 and B.1). For both cohorts, t_0 was defined as the time of hospital admission, and the time-to-event T was the time from the last record $(t_n, x_n); t_n \leq 36$ to in-hospital mortality. Patients who survived were right censored at discharge. Clinical features were standardized. Each dataset was randomly split into training (70%), validation (10%), and test sets (20%).

4.3. Comparison Methods

We compared the survival prediction performance of TrajSurv with three machine learning models, one clinical model, and four deep learning models. The machine learning models included the Cox proportional hazards model with elastic net penalty (CoxPH) Cox (1972), gradient-boosted Cox proportional hazards model (Boosting) Cox (1972); Friedman (2001), and random survival forest (RSF) Ishwaran et al. (2008). For the clinical model, we computed the longitudinal overall SOFA scores every 4 hour to create an aggregated 9-dimensional SOFA feature, and applied the Cox proportional hazards model (SOFA) for survival prediction. In the deep learning category, we evaluated models designed for survival prediction using longitudinal EHR data, including RDSM Nagpal et al. (2021), SurvLatent ODE (SLODE) Moon et al. (2022), and Dynamic-DeepHit (DDH) Lee et al. (2019).

4.4. Evaluation Metrics

We evaluated survival prediction performance using three commonly used time-dependent metrics: the concordance index (C-index), the time-dependent Brier score, and the dynamic area under the curve (AUC). The C-index and dynamic AUC measure the model's discrimination ability to distinguish patient risks. The Brier score assesses both the model's discrimination ability and its calibration. Specifically, we implemented the metrics using the *concordance_index_ipcw*, *brier_score*, and *cumulative_dynamic_auc* functions from scikit-survival module Pölsterl (2020). We computed the average across quartiles of follow-up times in the dataset for these metrics. We reported the average performance across 5 runs with different random seeds. Hyperparameters were tuned on the validation set.

For TrajSurv, we further evaluated the clinical alignment between latent states and patient states by computing Spearman's correlation between latent distance and SOFA or SOFA trend distance at each time point.

5. Results on Real Data

5.1. Survival Prediction Performance

Table 1 presents the survival prediction performance across comparison models and TrajSurv. Compared to existing models, TrajSurv achieved a higher C-index and dynamic AUC

Table 1: Survival Prediction Performance Comparison, mean ± std. **Bold** values indicate the best performance among all models or deep learning models for a given metric.

		MIMIC-III			eICU		
	Model	C-index	Brier	AUC	C-index	Brier	AUC
ML	CoxPH	0.706 ± 0.012	0.058 ± 0.002	0.698 ± 0.012	0.751 ± 0.007	0.044 ± 0.001	0.733 ± 0.007
	Boosting	0.766 ± 0.014	0.054 ± 0.002	0.754 ± 0.014	0.781 ± 0.005	0.043 ± 0.001	0.761 ± 0.005
	RSF	0.784 ± 0.013	**0.053 ± 0.002**	0.774 ± 0.014	0.804 ± 0.004	**0.042 ± 0.001**	0.783 ± 0.004
Clinical	SOFA	0.754 ± 0.012	0.053 ± 0.003	0.729 ± 0.013	0.755 ± 0.006	0.044 ± 0.001	0.726 ± 0.006
DL	RDSM	0.772 ± 0.012	0.061 ± 0.002	0.757 ± 0.012	0.796 ± 0.004	0.048 ± 0.001	0.779 ± 0.004
	SLODE	0.762 ± 0.018	0.064 ± 0.004	0.745 ± 0.020	0.780 ± 0.007	0.045 ± 0.000	0.756 ± 0.007
	DDH	0.768 ± 0.014	0.072 ± 0.005	0.748 ± 0.014	0.818 ± 0.005	0.051 ± 0.002	0.796 ± 0.005
	TrajSurv	**0.803 ± 0.011**	**0.056 ± 0.002**	**0.790 ± 0.013**	**0.823 ± 0.006**	**0.044 ± 0.001**	**0.803 ± 0.006**

and comparable Brier score, indicating improved discrimination and competitive calibration performance. (For more details, see Appendix)

Ablation Study To examine how our additional objectives improve the performance over vanilla NCDE, we conducted ablation studies on MIMIC-III by removing the time mask in $\mathcal{L}_{TACL}$ (A1), removing the entire $\mathcal{L}_{TACL}$ (A2), and removing both $\mathcal{L}_{TACL}$ and $\mathcal{L}_{PR}$ (A3). As shown in Figure 2A and B, comparing A2 and A3, we found that adding $\mathcal{L}_{PR}$ improves the performance over vanilla NCDE. Comparing A1 and A2, we found that contrastive learning slightly improves the discrimination ability, potentially due to additional supervision for the latent states. TrajSurv outperforms both A1 and A2, implying the validity of our model design involving the time-dependent nature of clinical progression.

Cross-Cohort Generalization To assess out-of-distribution generalization, we performed a cross-cohort evaluation by training TrajSurv on the full MIMIC-III training set and testing it on a random sample of 4,000 patients from the eICU dataset. The model achieved a C-index of 0.760 and a Brier score of 0.038, demonstrating robust performance on unseen data from different hospital systems.

5.2. Clinical Alignment Performance

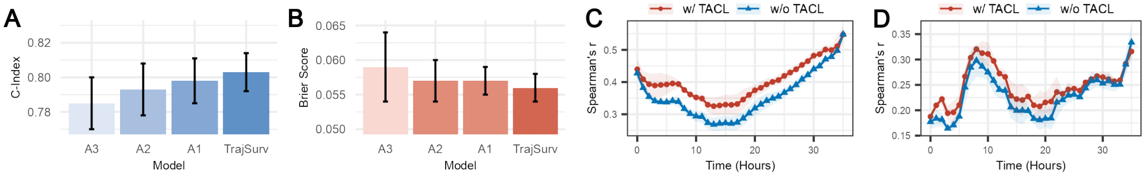


Figure 2: Ablation study showing improved (A) C-index and (B) Brier score. Spearman's correlation between latent distance and label distance, (C) SOFA distances, and (D) SOFA trend distances, with or without the TACL objective.

Figure 2C and D shows how the correlations change over time with and without TACL on MIMIC-III data. Without TACL, the model has relatively high correlations with SOFA and SOFA trend at the end of the 36-hour period, while they drop significantly in the middle. This suggests that despite the clinical alignment of the last latent state, the evolution of latent states does not align with the clinical progression. Trained with the TACL objective, TrajSurv retains the correlations at the end, and has significantly higher correlations with SOFA and slightly higher correlations with SOFA trend in the middle of the 36 hours. This suggests that latent states across the time horizon are more clinically aligned with patient states. Therefore, the evolution of latent states better reflects the clinical progression.

5.3. Interpretation

In this section, we validate TrajSurv's two-step interpretation process with known clinical knowledge on the test set of MIMIC-III data at the population level.

5.3.1. Feature Importance and Relevance from TrajSurv's Vector Field

From the columns of $\bar{f}_\theta$ (Figure 3A), we obtained feature importance and feature relevance following the procedure in section 3.3.1. As shown in Figure 3B, the top 15 influential features are ranked by the magnitude of the columns in the average vector field. Among the most important are blood urea nitrogen, total bilirubin, white blood cell count, creatinine, and hematocrit. These findings are consistent with previous ICU studies identifying these as significant indicators of patient states Chia et al. (2021); Iwase et al. (2022). For instance, blood urea nitrogen, creatinine, hematocrit, lactate, and PH were identified among the top 6 significant clinical features in a previous study Chia et al. (2021). The feature importance derived from the vector field provides a dynamical perspective, suggesting that changes in these features drive latent state shifts more rapidly than others.

Additionally, we plotted a heatmap of cosine similarities between columns of the vector field to reveal specific relationships among features (Figure 3C), providing an understanding of feature relevance in driving latent state transitions. For instance, creatinine and urea nitrogen display high positive cosine similarity, suggesting that their similar change patterns result in latent state shifts in nearly identical directions. This alignment is clinically consistent, as both are indicators of kidney function, where elevated levels may indicate renal impairment Hosten (1990). In contrast, urea nitrogen and bicarbonate show a high negative cosine similarity, meaning their parallel increases or decreases drive latent states in opposing directions. This finding aligns with evidence that urea nitrogen and bicarbonate are often inversely related in certain conditions Papadoyannakis et al. (1984); Balakrishnan et al. (2011). For example, bicarbonate supplementation is associated with reduced blood urea nitrogen levels in chronic renal failure patients Papadoyannakis et al. (1984).

These relationships derived from TrajSurv's vector field confirm known clinical associations while illustrating how TrajSurv captures the dynamical interdependencies between features that influence latent state evolution. We found these interpretations to be robust across different random data splits. For instance, blood urea nitrogen, total bilirubin, and white blood cell count were consistently ranked as top-5 important features, and the cosine similarity between creatinine and urea nitrogen remained stable at 0.60 ± 0.06 across three different random seeds. Note that the feature importance and relevance were data-driven and should be interpreted with caution.

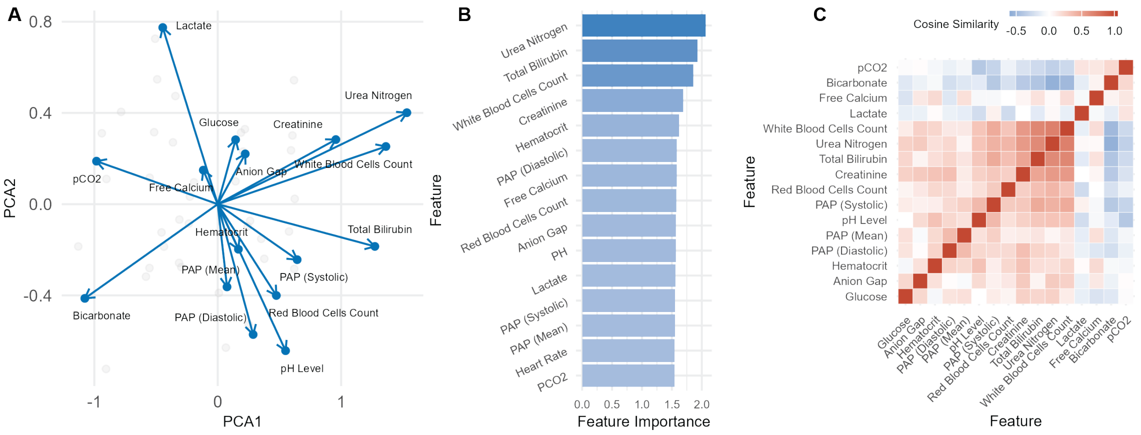


Figure 3: Feature importance and relevance derived from TrajSurv's vector field. (A) Columns of the average vector field; (B) Features with top 15 importance; (C) Cosine similarities between selected features.

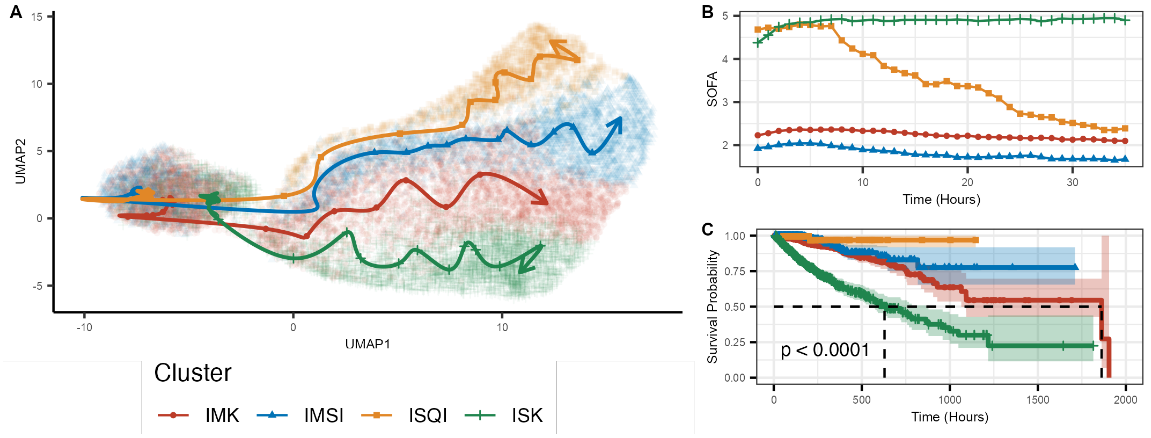


Figure 4: Clustering of latent trajectories with (A) the centroids of clusters; (B) the average SOFA trajectories of clusters; and (C) survival curves of clusters.

5.3.2. Latent Trajectory Clustering

Clustering latent trajectories using DTW, we obtained four clusters. Figure 4A shows the centroids of four clusters in the latent space, dividing the space into subdomains. Examining the average SOFA score trajectories (Figure 4B) of these clusters reveals four distinct disease progression patterns. ISQI cluster starts with high SOFA scores that decrease rapidly (initially severe, quickly improving - ISQI). ISK cluster maintains a consistently high SOFA score with a slight increase early on, indicating an initially severe condition that gradually worsens (initially severe, keep - ISK). IMK cluster and IMSI cluster exhibit lower SOFA scores, with scores remaining stable over time for the former (initially mild, keep - IMK) and slowly improving for the latter (initially mild, slowly improving - IMSI). These clusters suggest that TrajSurvs continuous latent trajectories represent different clinical progression patterns in the ICU.

The KM survival curve for each cluster reveals distinct survival patterns aligned with the clinical progression identified in the latent trajectories, hence stratifying risks. The ISQI subgroup, characterized by severe initial conditions that rapidly improve, shows the highest survival probability over time, suggesting that patients with early improvement after critical onset are less likely to experience adverse outcomes. In contrast, the ISK subgroup, with consistently high severity and a gradually worsening trend, displays the lowest survival probability, indicating a sustained high risk. The IMSI and IMK subgroups, representing milder initial conditions with slow improvement or stability, show intermediate survival outcomes. The statistically significant separation among the KM curves highlights the strength of TrajSurvs continuous latent trajectories in capturing meaningful clinical progression patterns, effectively stratifying risk based on the evolved patient conditions.

5.4. Case Study

A case study of two patients in MIMIC-III demonstrates TrajSurvs patient-level interpretation of how latent trajectories lead to predictions consistent with actual clinical progression.

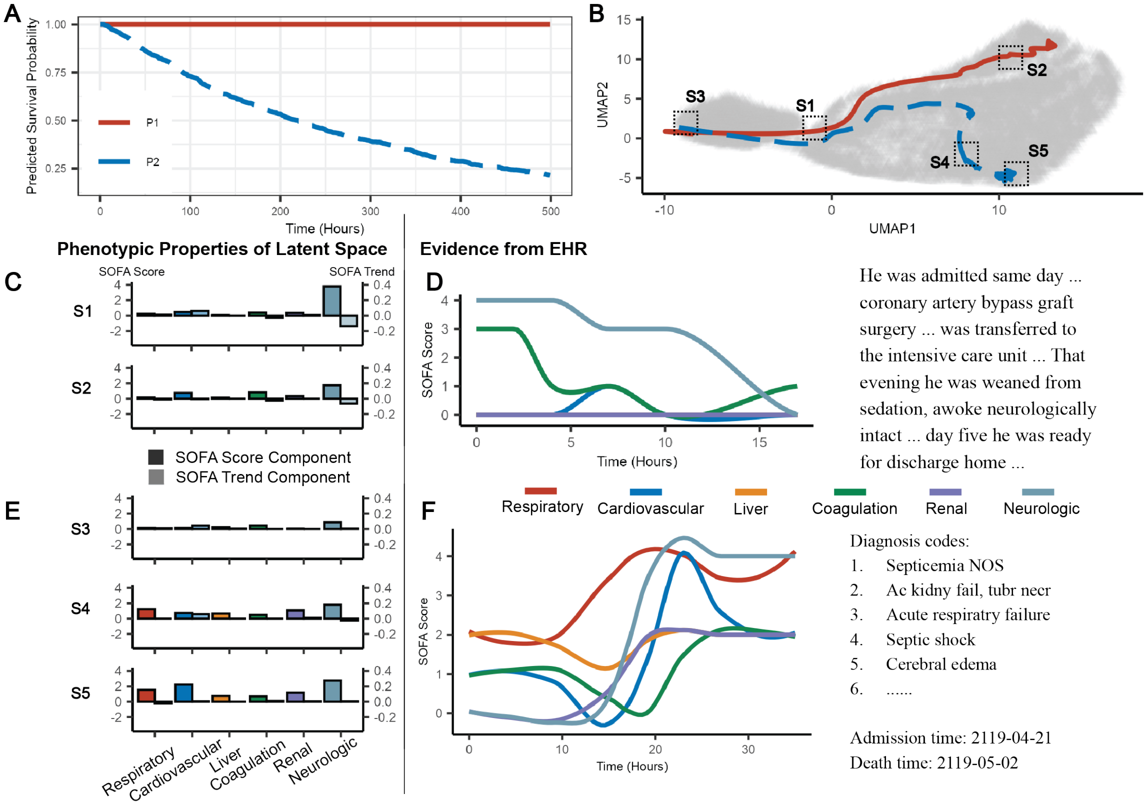
For patient P1, TrajSurv predicts a favorable survival outcome after the last observation (Figure 5A). This prediction is supported by several aspects of P1s latent trajectory. First, as shown in Figure 5B, P1s latent trajectory falls within the ISQI subgroup, indicating a generally positive prognosis. Additionally, P1's latent trajectory sequentially goes through regions S1 and S2 on the latent space, where the phenotypic properties of latent space (see Appendix for the calculation) show severe but improving neurologic symptoms at S1 and mild symptoms at S2 (Figure 5C). This is consistent with P1's actual SOFA trajectory and discharge note, confirming neurologic improvement post-coronary artery bypass surgery on the first admission day (Figure 5D).

For patient P2, TrajSurv predicts a poor survival outcome following the last observation (Figure 5A). This result is supported by P2s latent trajectory (Figure 5B). P2s trajectory shows a sudden shift from the IMSI subdomain to ISK subdomain, suggesting an abrupt deterioration mid-visit. P2's trajectory passes through S3, S4, and S5 regions, which indicate initial presentation with mild symptoms (S3) that progress to multiple organ dysfunction (S4 and S5). In the absence of discharge notes, we compared this interpretation to P2s actual SOFA trajectory, diagnosis codes, and actual survival time, showing consistency with a sepsis diagnosis and rapid decline leading to death within 300 hours.

This case study suggests TrajSurvs potential to provide clinicians with interpretable insights consistent with patient outcomes and clinical progression.

6. Discussion

In this study, we propose TrajSurv, a method for survival prediction that learns continuous latent trajectories from longitudinal EHR. TrajSurv leverages NCDE to capture the continuous process of evolving clinical features, mapping this evolution to continuous latent trajectories. We further incorporate a TACL objective to ensure the clinical alignment of these trajectories, improving both accuracy and transparency. Paired with a two-step



He was admitted same day ... coronary artery bypass graft surgery ... was transferred to the intensive care unit ... That evening he was weaned from sedation, awoke neurologically intact ... day five he was ready for discharge home ...

Diagnosis codes:

1. Septicemia NOS
2. Ac kidny fail, tubr necr
3. Acute respiratry failure
4. Septic shock
5. Cerebral edema
6. ......

Admission time: 2119-04-21
Death time: 2119-05-02

Figure 5: Case studies of two patients (P1 and P2). (A) Predicted survival probability after the last observation; (B) Continuous latent trajectories until the last observation; (C) Average phenotypic properties of S1 and S2 of the latent space; (D) Evidence from EHR for P1, including the ground truth SOFA trajectory and the discharge note; (E) Average phenotypic properties of S3, S4, and S5; (F) Evidence from EHR for P2, including the ground truth SOFA trajectory and the Diagnosis codes.

divide-and-conquer interpretation process, TrajSurv enables trustworthy survival prediction, achieving competitive prediction performance and superior transparency.

With increasing emphasis on the need for reliable and transparent AI systems in healthcare, TrajSurv provides a viable direction for the transparent modeling of longitudinal EHR by learning clinically aligned continuous latent trajectories and interpreting the model through a divide-and-conquer approach. The divide-and-conquer interpretation process provides a dynamical perspective on how the temporal changes of clinical features link to the outcome, enriching existing interpretation methods and fostering trust in the model's outputs. By providing a clear and visual representation of clinical progression, TrajSurv contributes to the development of trustworthy AI tools for clinical use.

A key innovation in TrajSurv lies in its TACL objective. This objective works as a soft regularization to the latent space by explicitly aligning the latent trajectory with actual clinical progression while respecting its time-dependent nature. By contrasting latent representations of patient states at different time points, TACL encourages the model to learn trajectories that are not only predictive of survival but also clinically meaningful. This enables clinically meaningful vector field analysis and cluster patterns in our two-step interpretation process. Through our ablation, we found that TACL improved both performance and clinical alignment.

The patient's clinical progression is a complex and dynamic process, and TrajSurvs TACL offers insights into how latent trajectories align with this process. Our analysis reveals that TACL learns latent states that correspond more closely to severity levels than to severity trends (Figure 2C and D). This finding is consistent with prior research showing that mean and peak SOFA scores are stronger ICU prognosticators than changes in SOFA (Δ-SOFA) Ferreira et al. (2001), suggesting that TACL may prioritize the most salient dimensions of multidimensional patient states. Future research may further refine this alignment by incorporating more comprehensive patient state definitions or advancing alignment techniques. As one of the early efforts, TrajSurv establishes a framework for transparent longitudinal EHR modeling by quantifying patient states, aligning latent trajectories with clinical progression, and interpreting with a divide-and-conquer approach.

Our work builds upon and extends previous research utilizing NCDEs. While studies like TE-CDE Seedat et al. (2022) and CoxSig Bleistein et al. (2024) have employed NCDEs for tasks like continuous-time counterfactual prediction and dynamical survival prediction, they have largely underexplored the rich information contained within the latent trajectory and vector field. TrajSurv specifically leverages NCDE's capacity to represent the continuous nature of clinical progression transparently and quantitatively, linking clinical feature evolution to the latent trajectory. Crucially, the feature importance and relevance derived from the NCDE vector field provide a novel and elegant way to interpret how changes in different clinical features jointly influence changes in the latent space. For example, similar changes in creatinine and urea nitrogen lead to the similar transition of latent states.

Finally, TrajSurv holds considerable potential for clinical translation. Its accurate survival prediction and risk stratification can facilitate clinical resource allocation and decision-making. Critically, as shown in the case study, the interpretable outputs provided by TrajSurv can empower clinicians by providing insights into the prediction process, fostering trust and acceptance in clinical settings. While comprehensive clinical validation across diverse cohorts remains necessary, we believe TrajSurv represents a significant step toward trustworthy AI tools in clinical practice and can inspire the development of future explainable AI models for longitudinal EHR data.

Limitations Our model design relies on an existing clinical assessment like SOFA as the supervision signal. While these assessments offer a practical way to quantify patient states, they do not fully represent patient states and reduce the model flexibility in various clinical contexts, as not all clinical contexts provide such assessments. Future research may explore purely data-driven methods to ensure clinical alignment or define patient states more comprehensively using more adaptable labels, such as knowledge graphs or key clinical variables, rather than pre-existing assessments. Furthermore, future research is needed to understand the clinical alignment between latent trajectories and clinical progression from a more holistic view. Finally, population-level properties may not fit individual cases, and we leave individual-level vector field analysis for future studies.

References

Abdallah Alabdallah. *Towards Trustworthy Survival Analysis with Machine Learning Models*. PhD thesis, Halmstad University Press, 2025. URL https://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-55202.

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

M Balakrishnan, R Tucker, BE Stephens, and JM Bliss. Blood urea nitrogen and serum bicarbonate in extremely low birth weight infants receiving higher protein intake in the first week after birth. *Journal of Perinatology*, 31(8):535–539, 2011.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The journal of machine learning research*, 13(1):281–305, 2012.

Linus Bleistein, Van-Tuan Nguyen, Adeline Fermanian, and Agathe Guilloux. Dynamical survival analysis with controlled latent states. *arXiv preprint arXiv:2401.17077*, 2024.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Alvin Har Teck Chia, May Sze Khoo, Andy Zhengyi Lim, Kian Eng Ong, Yixuan Sun, Binh P Nguyen, Matthew Chin Heng Chua, and Junxiong Pang. Explainable machine learning prediction of icu mortality. *Informatics in Medicine Unlocked*, 25:100674, 2021.

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

Flavio Lopes Ferreira, Daliana Peres Bota, Annette Bross, Christian Mélot, and Jean-Louis Vincent. Serial evaluation of the sofa score to predict outcome in critically ill patients. *Jama*, 286(14):1754–1758, 2001.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Zhichen Gong and Huanhuan Chen. Dynamic state warping, 2017. URL https://arxiv.org/abs/1703.01141.

Adrian O Hosten. Bun and creatinine. *Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition*, 1990.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. 2008.

Shinya Iwase, Taka-aki Nakada, Tadanaga Shimada, Takehiko Oami, Takashi Shimazui, Nozomi Takahashi, Jun Yamabe, Yasuo Yamao, and Eiryo Kawakami. Prediction algorithm for icu mortality and length of stay using machine learning. *Scientific reports*, 12 (1):12912, 2022.

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.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18:1–12, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. URL https://arxiv.org/abs/2004.11362.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems*, 33:6696–6707, 2020.

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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Intae Moon, Stefan Groha, and Alexander Gusev. Survlatent ode: A neural ode based time-to-event model with competing risks for longitudinal data improves cancer-associated venous thromboembolism (vte) prediction. In *Machine Learning for Healthcare Conference*, pages 800–827. PMLR, 2022.

James Morrill, Patrick Kidger, Lingyi Yang, and Terry Lyons. Neural controlled differential equations for online prediction tasks. *arXiv preprint arXiv:2106.11028*, 2021.

Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

Chirag Nagpal, Vincent Jeanselme, and Artur Dubrawski. Deep parametric time-to-event regression with time-varying covariates. In *Survival prediction-algorithms, challenges and applications*, pages 184–193. PMLR, 2021.

Chirag Nagpal, Willa Potosnak, and Artur Dubrawski. auton-survival: an open-source package for regression, counterfactual estimation, evaluation and phenotyping with censored time-to-event data. *arXiv preprint arXiv:2204.07276*, 2022.

NJ Papadoyannakis, CJ Stefanidis, and M McGeown. The effect of the correction of metabolic acidosis on nitrogen and potassium balance of patients with chronic renal failure. *The American journal of clinical nutrition*, 40(3):623–627, 1984.

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

Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL http://jmlr.org/papers/v21/20-729.html.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. *arXiv preprint arXiv:2206.08311*, 2022.

Navid Shafigh, Morteza Hasheminik, Elnaz Shafigh, Haleh Alipour, Shahram Sayyadi, Neda Kazeminia, Batoul Khoundabi, and Sara Salarian. Prediction of mortality in icu patients: A comparison between the sofa score and other indicators. *Nursing in Critical Care*, 29 (6):1619–1622, 2024.

Sushrut S. Waikar and Joseph V. Bonventre. Creatinine kinetics and the definition of acute kidney injury. *Journal of the American Society of Nephrology : JASN*, 20 3:672–9, 2009. URL https://api.semanticscholar.org/CorpusID:902086.

Lu Wang, Yan Li, and Mark Chignell. Combining ranking and point-wise losses for training deep survival analysis models. In *2021 IEEE international conference on data mining (ICDM)*, pages 689–698. IEEE, 2021.

Feng Xie, Han Yuan, Yilin Ning, Marcus Eng Hock Ong, Mengling Feng, Wynne Hsu, Bibhas Chakraborty, and Nan Liu. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of biomedical informatics*, 126:103980, 2022.

Beihua Yang, Peng Song, Yuanbo Cheng, Shixuan Zhou, and Zhaowei Liu. Enhanced tensor based embedding anchor learning for multi-view clustering. *Information Sciences*, 690: 121532, 2025.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36, 2024.

Appendix A. Additional Results

A.1. Model Calibration

To assess TrajSurv's calibration, we generated calibration plots across quartiles of follow-up times on MIMIC-III data. Figure S1 illustrates that TrajSurvs predicted risk scores align well with observed outcomes at the median follow-up time but show slight underestimation at the 25% quartile and slight overestimation at the 75% quartile. Overall, TrajSurvs calibration is competitive with other methods.

A.2. Phenotypic Properties of Latent Space

To visually examine the phenotypic properties of the clinically aligned latent space, we plotted the dominant phenotypes and corresponding severity across regions in the latent space on MIMIC-III data. For each region, we averaged the SOFA component scores of the 50 nearest latent states, effectively representing the typical phenotype within that area. The dominant phenotype is then obtained from the highest average SOFA component in that region. As shown in Figure S2, distinct regions in this space represent different clinical phenotypes, while close states reflect consistent dominant phenotypes and severity. For instance, the lower corner of the latent space corresponds to severe cardiovascular conditions. This demonstrates that the proximity in TrajSurv's latent space indicates similar patient phenotypic patterns.

A.3. Feature Importance and Relevance of All Features

Figure S3 and S4 show the full feature importance and relevance figures from the vector field interpretation on MIMIC III. We note that these average patterns obtained from data-driven analysis should be carefully interpreted and should not be directly used in clinical practice.

A.4. Interpretation Method Comparison

TrajSurv's interpretation is a two-step process: (1) vector field interpretation for the feature-to-trajectory mapping, and (2) latent trajectory clustering for the trajectory-to-outcome linkage. The vector field interpretation offers insights distinct from, yet complementary to, methods like SHAP Lundberg and Lee (2017) or permutation importance Altmann et al. (2010):

- It explains how changes in clinical features dynamically influence the magnitude and direction of the patient's latent trajectory over continuous time.
- It allows assessment of feature relevance by measuring how similarly features drive this trajectory evolution via cosine similarity of vector field columns.

Table S1 compares different interpretation methods. We highlight that TrajSurv combines vector field interpretation with latent trajectory clustering, hence enabling end-to-end interpretation from a dynamical perspective.

A.5. Additional Experiments on Model Generalization

We conducted additional experiments on TrajSurv's generalization using the MIMIC-III dataset. The results demonstrate TrajSurv's robustness across different settings.

Full vs. Reduced Data We compared the full data (hourly data from the raw database) and the reduced data (only including hours with more than 20 observed features). The C-index is 0.806 on the full data and 0.803 on the reduced data, with about 3 times faster training on the reduced data. Therefore, we used reduced data during our experiments, which achieves computational efficiency while preserving fair comparison.

Short Trajectory We performed an experiment using only the data between 18 and 36 hours after admissions, rather than the full 36-hour data in the main paper. TrajSurv achieves a C-index of 0.788, compared to 0.803 with 36-hour data, indicating robustness despite information loss.

Limited Training Data To assess performance with limited training data, we experimented with a 20% training, 40% validation, and 40% test split on MIMIC-III data. TrajSurv achieves a C-index of 0.752, compared to 0.746 of RSF, demonstrating generalizability even with reduced training data.

A.6. Interpolation Method

TrajSurv employs cubic Hermite splines with backward differences for input path interpolation. This is recommended by prior NCDE literature Kidger et al. (2020); Morrill et al. (2021) for its smoothness, online processing capability, and efficient integration. To explore alternatives, we experimented with rectilinear interpolation, which is also an online method, yielding a C-index of 0.789, slightly below TrajSurv with cubic splines (C-index 0.803).

Appendix B. Experimental Details

B.1. MIMIC-III Cohort

The MIMIC-III cohort includes 20,258 hospital admissions and 53 clinical features in 1-hour resolution. The dataset has 1,743 events. The 53 clinical features were the lab and chart events with top occurrence and demographics. The clinical features include arterial blood pressure diastolic, arterial blood pressure mean, arterial blood pressure systolic, alanine aminotransferase, alkaline phosphatase, anion gap, aspartate aminotransferase, base excess, basophils, bicarbonate, bilirubin total, calcium total, calculated total carbon dioxide, chloride, creatinine, central venous pressure, eosinophils, free calcium, glucose, heart rate, hematocrit, hemoglobin, international normalized ratio, lactate, lymphocytes, magnesium, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, monocytes, oxygen saturation, pulmonary artery pressure diastolic, pulmonary artery pressure mean, pulmonary artery pressure systolic, partial pressure of carbon dioxide, pH, phosphate, platelet count, partial pressure of oxygen, potassium, potassium whole blood, prothrombin time, partial thromboplastin time, red cell distribution width, red blood cell count, respiratory rate, sodium, temperature, urea nitrogen, white blood cell count, gender, age, and body mass index.

Table S1: Comparison of Interpretation Methods.

	SHAP / Permutation Importance	**Attention Mechanisms**	**TrajSurv's Vector Field Interpretation**
Primary Insight	Contribution of each feature to a specific outcome	Which input features/time steps are most influential for an outcome	How feature changes drive the **evolution & direction** of latent state trajectories
Temporal Aspect	Typically static for a given prediction; can be applied at different times but doesn't inherently model evolution	Highlights salient inputs/time steps for discrete predictions	**Models continuous-time dynamics**; captures how features influence on **trajectory evolution**
Focus of Importance	Importance for the magnitude of the final prediction	Importance for the final output	Importance for latent trajectory evolution (**magnitude and direction**)
Feature Interaction	Can compute SHAP interaction values	Does not directly provide feature interactions	**Relevance in co-directing trajectory evolution**
Granularity	Insight into outcome prediction	Insight into outcome prediction	Insight into the feature-to-trajectory **dynamical process**

The MIMIC-III cohort is inherently irregularly sampled. For computational feasibility and a fair comparison with baselines, we only included hourly time points where more than 20 features were observed. This cohort has an average of 4.02 hourly clinical observations per patient within the first 36 hours and an overall feature missingness rate of 57.7% across all time steps.

B.2. eICU Cohort

The eICU cohort includes 116,503 hospital admissions and 53 clinical features in 1-hour resolution. The dataset has 7,958 events. The clinical features include basophils, eosinophils, lymphocytes, monocytes, polymorphonuclear leukocytes, alanine aminotransferase, aspartate aminotransferase, blood urea nitrogen, fraction of inspired oxygen, bicarbonate, hematocrit, hemoglobin, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, mean platelet volume, prothrombin time, international normalized ratio, red blood cell count, red cell distribution width, white blood cell count (x1000), albumin, alkaline phosphatase, anion gap, bedside glucose, calcium, chloride, creatinine, glucose, magnesium, pH, partial pressure of carbon dioxide, partial pressure of oxygen, phosphate, platelet count (x1000), potassium, sodium, total bilirubin, total protein, temperature, arterial oxygen saturation, heart rate, respiration rate, central venous pressure, end-tidal carbon dioxide, systemic arterial systolic pressure, systemic arterial diastolic pressure, systemic arterial mean pressure, pulmonary artery systolic pressure, pulmonary artery diastolic pressure, pulmonary artery mean pressure, age, and body mass index.

The eICU cohort is also inherently irregularly sampled, and similarly preprocessed. This cohort has an average of 3.15 hourly clinical observations per patient within the first 36 hours and an overall feature missingness rate of 55.0% across all time steps.

B.3. Hyperparameter Tuning

Hyperparameters were tuned on the validation set. Specifically, for TrajSurv, we conducted a random search Bergstra and Bengio (2012) to tune the hyperparameters. The searching space of key hyperparameters is shown in Table S2. The search ranges were informed by common practice, prior work (e.g., κ_1 based on Rank-N-Contrast Zha et al. (2024)), and empirical observations. Hyperparameters were tuned for the highest C-index on the validation set, with early stopping (patience 5) to prevent overfitting.

B.4. Optimization

We used AdamW Loshchilov and Hutter (2019) to optimize TrajSurv. For the NCDE module, we used the *torchcde* package with *torchdiffeq* backend. We trained TrajSurv for 100 epochs, with early stopping based on the C-index on the validation set, with patience 5. TrajSurv's training and evaluation were performed on a single NVIDIA Tesla T4 or RTX 2080 Ti GPU.

B.5. Comparison Methods Implementation

All baseline models underwent hyperparameter tuning using grid search on the validation set. For CoxPH, RSF, and Boosting, we used the implementations in the scikit-survival library Pölsterl (2020). For RDSM, we used the auton-survival package Nagpal et al. (2022). DDH and SLODE were implemented using the authors' original publicly available code.

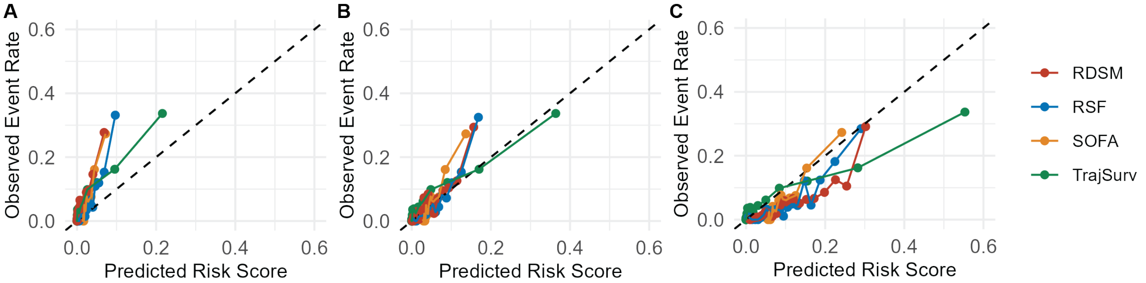


Figure S1: Calibration plot comparing TrajSurv with RSF, SOFA, and RDSM at (A) 25% quartile, (B) median, and (C) 75% quartile follow-up time.

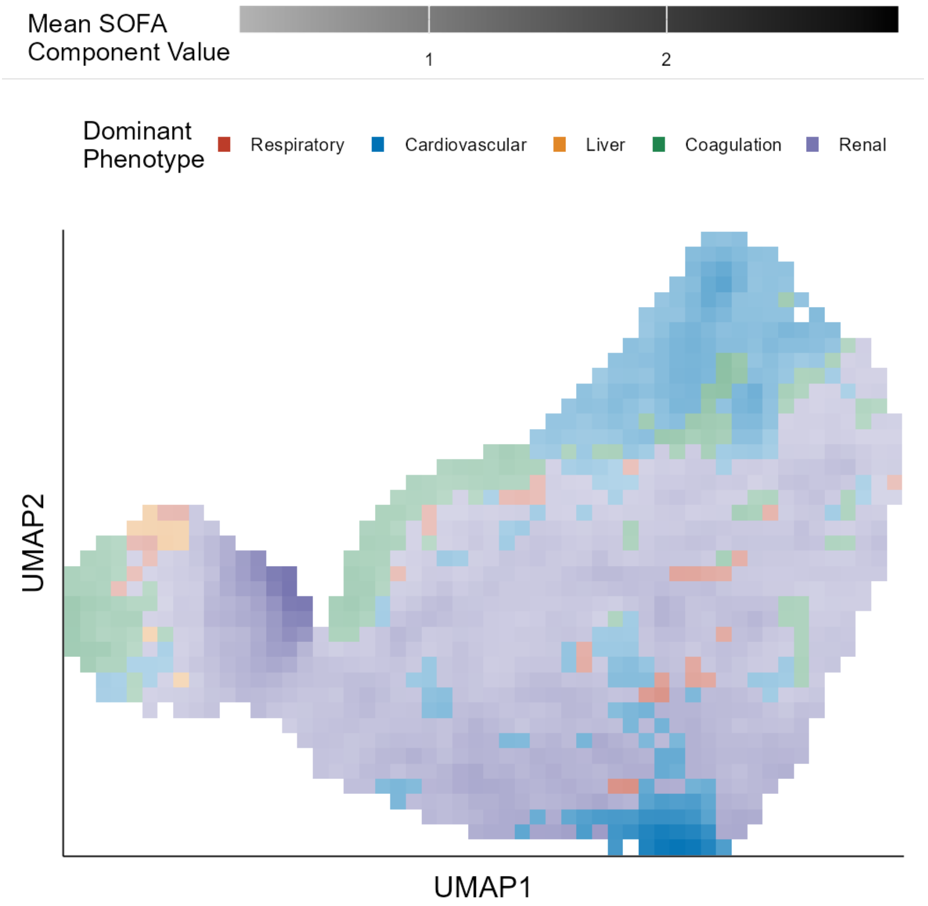


Figure S2: Dominant phenotypes in different regions of latent space. Different colors represent different dominant phenotypes, and the transparency of the color corresponds to the average severity (SOFA component) of the dominant phenotypes.

Table S2: Hyperparameter Tuning Ranges. **Bold** values indicate the final selected hyperparameters.

Hyperparameter	Range
α	$[0.75, \mathbf{1.0}, 1.25]$
κ_1	$[1, \mathbf{2}]$
κ_2	$[10, \mathbf{30}, 50]$
δ	$[10, \mathbf{20}]$
d_z	$[32, \mathbf{64}]$

Appendix C. More Background Information

C.1. Neural Controlled Differential Equations

Neural controlled differential equations (NCDEs) extend neural ordinary differential equations (NODEs) to handle incoming data in an irregularly sampled time series setting.

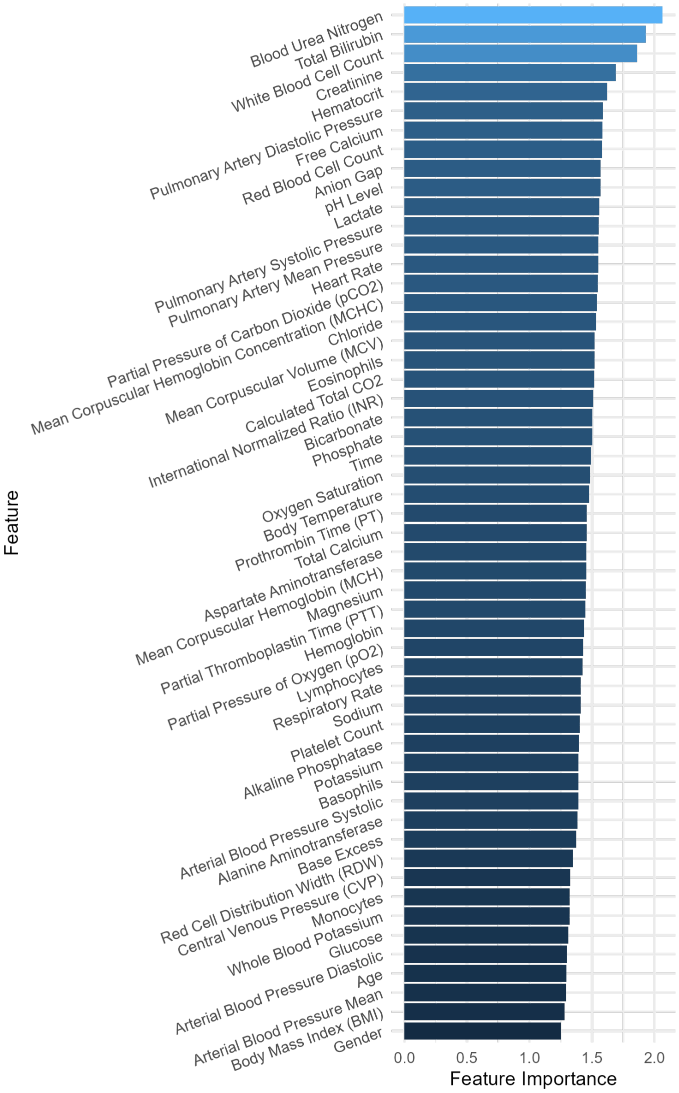


Figure S3: Feature importance of all clinical features.

NCDEs leverage the mathematical framework of controlled differential equations (CDEs) to provide continuous-time modeling that dynamically adapts to data observations. A key advantage of NCDEs is their ability to utilize adjoint backpropagation for efficient training. The training of NCDEs has an overall memory footprint of $\mathcal{O}(L + H)$, where $L = t_n - t_0$ and H is the memory footprint of the vector field. This contrasts previous work on NODEs for time series, which requires $\mathcal{O}(LH)$ memory. Kidger et al. (2020)

C.2. Dynamic Time Warping

Dynamic Time Warping (DTW) is a technique used to measure the similarity between two time series of different lengths or varying speed. Previous work has also extended DTW to series of latent states, called dynamic state warping Gong and Chen (2017), which is similar

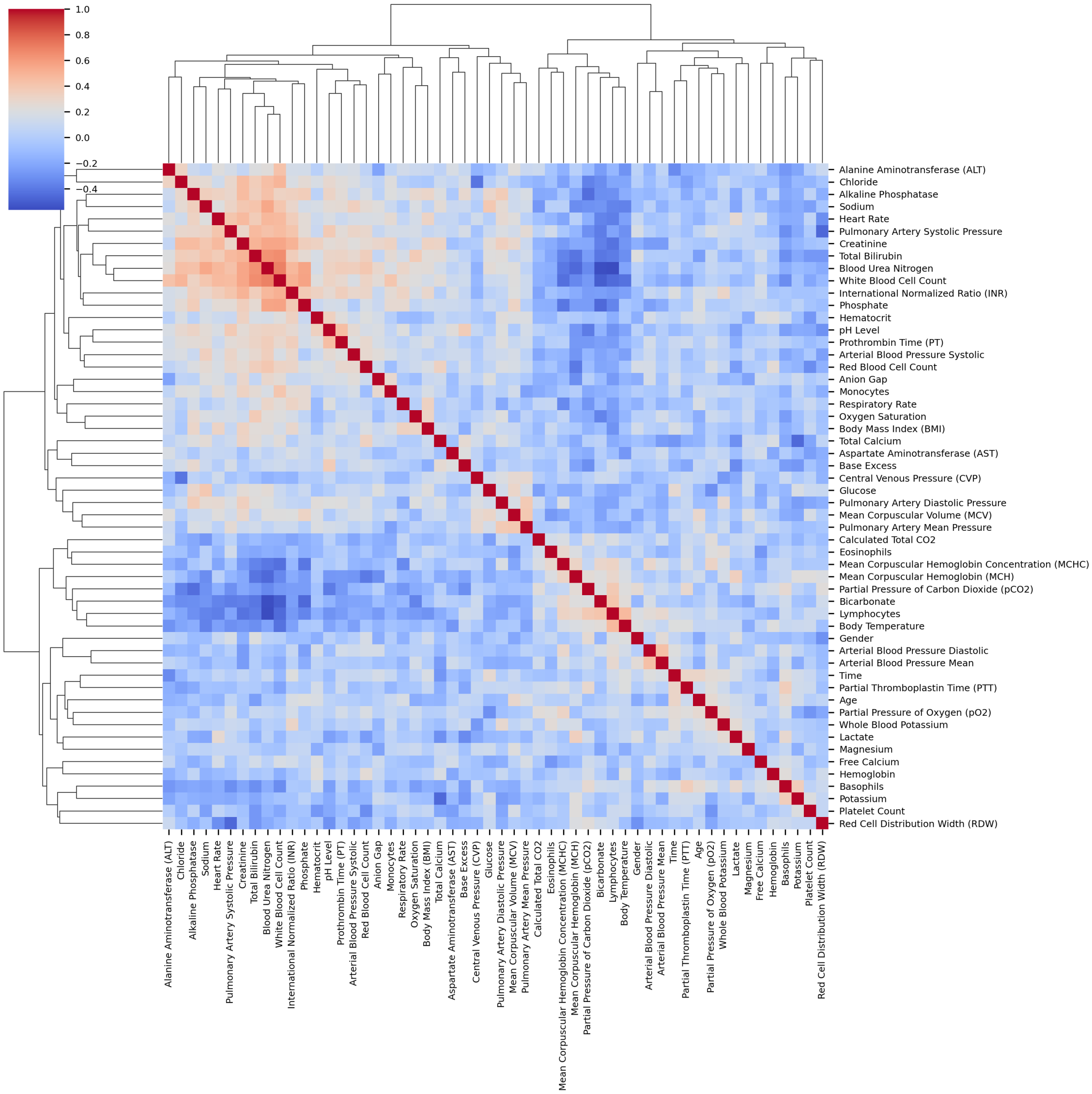


Figure S4: Heatmap of the cosine similarities between clinical features from the average vector field.

to our latent trajectory clustering. The core of DTW involves constructing a cost matrix where each cell (i, j) represents the distance between points in the two time series, and then finding the optimal warping path through this matrix. The accumulated cost matrix D is built using the following recursive formula:

$$D(i, j) = \text{cost}(i, j) + \min[D(i-1, j), D(i, j-1), D(i-1, j-1)],$$

where $\text{cost}(i, j)$ is the distance between the i-th and j-th points of the two respective time series. The final DTW distance is the value of $D(n, m)$, where n and m are the lengths of the two time series.

In our latent trajectory clustering, the time series in DTW are the hourly extracted latent states from the continuous latent trajectories and the cost is the L2 distance between latent states.