Deep State-Space Generative Model For Correlated Time-to-Event Predictions

Yuan Xue
yuanxue@google.com
Google Inc.

Denny Zhou
dennyzhou@google.com
Google Inc.

Nan Du
dunan@google.com
Google Inc.

Andrew M. Dai
adai@google.com
Google Inc.

Zhen Xu
zhenxu@google.com
Google Inc.

Kun Zhang
kunzhang@google.com
Google Inc.

Claire Cui
claire@google.com
Google Inc.

ABSTRACT

Capturing the inter-dependencies among multiple types of clinically-critical events is critical not only to accurate future event prediction, but also to better treatment planning. In this work, we propose a deep latent state-space generative model to capture the interactions among different types of correlated clinical events (e.g., kidney failure, mortality) by explicitly modeling the temporal dynamics of patients' latent states. Based on these learned patient states, we further develop a new general discrete-time formulation of the hazard rate function to estimate the survival distribution of patients with significantly improved accuracy. Extensive evaluations over real EMR data show that our proposed model compares favorably to various state-of-the-art baselines. Furthermore, our method also uncovers meaningful insights about the latent correlations among mortality and different types of organ failures.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning; Artificial intelligence.**

KEYWORDS

State Space Model; Generative Model; Survival Analysis;

ACM Reference Format:

Yuan Xue, Denny Zhou, Nan Du, Andrew M. Dai, Zhen Xu, Kun Zhang, and Claire Cui. 2020. Deep State-Space Generative Model For Correlated Time-to-Event Predictions. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20)*, August 23–27, 2020, Virtual Event, CA, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3394486.3403206>

This work is licensed under a Creative Commons Attribution International 4.0 License.

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

© 2020 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-7998-4/20/08.

<https://doi.org/10.1145/3394486.3403206>

1 INTRODUCTION

Time-to-event prediction (also known as survival analysis) investigates the distribution of time duration until the event of interest happens in the presence of event censorship. In the healthcare domain, it is an essential tool for modeling the risks of critical medical events and capturing of the relationship between the co-variants and the risks [9].

Recently, machine learning methods have been applied to time-to-event predictions to provide flexible modeling of the time distribution [6, 19, 22], and capture the nonlinear relationship between co-variants and the risk of an event [30]. Most of the prior work [6, 19, 31] on time-to-event prediction are limited in modeling a single type of event, and lack the capability of analyzing the correlations among the risks of multiple types of events. In real world, most events are by nature related to or even caused by one another. In particular, within the medical domain, death may be caused by either a single organ failure or a combination of multiple simultaneous organ failures which could significantly increase the risk of death in a non-linear fashion. Furthermore, the dysfunction or failure of one organ will also trigger the dysfunction/failure of another (e.g., kidney failure may be caused by liver damage). Therefore, predicting the next-occurrence of events of interest (e.g., death) heavily depends on the joint risks of other associated types of events (e.g., different organ failures).

Understanding and capturing the inter-dependencies among multiple types of clinically-critical events are important not only to deriving more accurate future event timing predictions, but also critical for designing respective treatment plans that could simultaneously handle multiple correlated life-threatening failure events. For instance, when designing optimal treatment plans for patients with comorbidities, the decision on whether a diabetic patient (who also has a renal disease) should receive dialysis or a renal transplant must be based on a joint prognosis of diabetes-related complications and end-stage renal failure. Overlooking the diabetes-related risks may lead to misguided therapeutic decisions.

Recognizing the necessity of a multi-type event model, a sequence of recent work [2, 5, 22, 26, 35, 36] propose multi-type event analysis with a particular parametric form of event relation, *the competing risk* [11], that is, the occurrence of one event precludes

the occurrence of another. However, uncovering the general temporal correlations among multiple types of events still remains an open question.

Recent wide adoption of electronic medical records (EMR) leads to the collection of an enormous amount of patient measurements over time in the form of time-series data. These retrospective data contain valuable information that captures the intricate relations among patient conditions, clinical interventions and outcomes, and present a promising avenue for accurately capturing the temporal progression of clinically critical events.

To accurately predict the temporal progression and establish the temporal correlations of multiple types of a patient’s critical medical events, we need a powerful model that is able to capture the dynamics of the underlying patient states from rich time-varying patient measurements and infer the inter-dependency of the occurrences of clinically critical events from these hidden dynamic states.

Therefore, in this work, we present a deep state-space generative model, which provides joint time-to-event predictions of multiple clinical events based on the EMR time-series data. Our model is able to simultaneously predict mortality risk and capture organ failure risk trajectories by leveraging the temporal progressive correlations between the past measurements and clinical interventions. More specifically, we have made the following contributions:

Technical Significance. We present a deep state space generative model, augmented with intervention forecasting, to provide a principle framework to capture the interactions among observations, interventions, critical event occurrences, latent patient states and their uncertainty. Based on the temporal dynamics of patients’ states, we develop a new discrete-time hazard rate model that provides flexible fitting of general time-to-event distributions without restricted parametric assumptions.

Clinical Value. The ability to jointly forecast multiple clinical events and identify their temporal correlations provides clinicians with a full picture of a patient’s medical condition and better supports them with decision making. Moreover, by demonstrating the correlations between the risk of these organ failure events and mortality, we also provide physicians with evidence to understand our mortality risk predictions.

2 RELATED WORK

This section mainly summarizes the related studies in literature as follows:

Clinical predictions. Deep learning models are increasingly used to improve the predictions of clinical outcomes, such as mortality or diagnosis [7, 8, 23, 29, 32]. These studies can be roughly categorized based on the data, models and the prediction tasks used. Our work uses clinical time-series data, similar to [23–25, 27, 33] to make predictions. It is related to the switching discrete state space model[14] used to predict clinical interventions in ICU, and the continuous state space model[28], which learns a treatment policy using deep reinforcement learning. However, our task, providing a joint forecast for the hazard rate of multiple correlated event, has not been considered by these previous work.

Time-to-event predictions. Machine learning methods have been applied to time-to-event predictions. For example, recent

works have extended the classical Cox proportional hazards model with neural network-based co-variate encoding [15, 19] and with multi-task formulations [26, 35]. The work of [31] converts the time-to-event estimation to a discretized-time classification problem, while others use a continuous-time model based on Gaussian processes [2, 4, 10] or generative adversarial networks [6] to model the nonlinear relationship between co-variables and the time. Unlike existing works, our work combines a deep state space model with a discrete-time hazard model to support more flexible distribution model. The predicted hazard rates of correlated events are calibrated in time, which provide better interpretations on the influence among the events.

Deep state-space models. A few recent works [3, 12, 13, 18, 20, 21, 34] have extended the traditional linear Gaussian state space model to handle non-linear relations via neural networks. The goal of these works is to fit a generative state space model to a sequence of observations and actions, while ours is to capture the inherent state transition dynamics and use it for time-to-event prediction. In particular, we focus on the modeling and learning of hazard rate functions of different events which share the common underlying state. In addition, these existing works have presented different models, methods and neural network architectures to infer the latent states more accurately. From this perspective, our work is complementary to these existing works. The encoding network in our method can leverage any existing architecture. In the experiment, we adopted the architecture presented in [20].

3 CORRELATED TIME-TO-EVENT PREDICTIONS

We first describe our usage scenario and problem setting then provide the formal definition of our learning task in this section. Electronic medical record (EMR) provides a longitudinal database where each patient medical history including, lab and vital measurements, medication orders and administrations, conditions and diagnosis, medical procedures, is recorded. We are utilizing the measurements and interventions from EMR in the form of time-series data to predict the time of future critical medical events such as mortality and organ failures. To capture the uncertainty of time, we estimate the distribution of time duration between the time of prediction and the event occurrence as well as the risk of experiencing the events of interest at any given time.

Formally, consider a longitudinal EMR system with N patients. We discretize and calibrate patient i ’s longitudinal records to a time window $[1, T_i]$, where time 1 and T_i represent the time when the patient first and last interacts with the system¹. Note that T_i , also called censor time in survival analysis, can vary for different patients i . In this paper, we focus on personalized predictions. When the context is clear, we simplify notation T_i with T . We consider two types of time series data in EMR:

- **Observations $\mathbf{x}$,** a real-valued vector of O -dimension. Each dimension corresponds to one type of clinical measurement including vital signs and lab results (e.g., mean blood pressure, serum lactate). We use $\mathbf{x}_{1:T}$ to denote the sequence of measurements at discrete time points $t = 1, \dots, T$;

¹For inpatient prediction, this period refers to the start and end of an inpatient encounter, instead of the entire patient history.

Figure 1: Time-calibrated multiple event risk trajectory. Given the past blood pressure readings (red dots) and Dopamine dosage (green square) in the upper left, the model captures how the risk of different types of clinical outcomes (mortality, neural system failure, etc.) can change with time in the future at the bottom right.

- **Interventions $\mathbf{u}$,** a real-valued vector of I -dimension. Each dimension corresponds to one type of clinical intervention, and its value indicates the presence and the level of intervention such as the dosage of medication being administrated or the settings of a mechanical ventilator. Similarly, $\mathbf{u}_{1:T}$ denotes the sequence of interventions at $t = 1, \dots, T$.

At prediction time t^* , given the sequence of observations and interventions $\mathbf{x}_{1:t^*}, \mathbf{u}_{1:t^*}$, we estimate the distribution of time for a set of clinically significant events. We represent an event e with a tuple (c, t^e) , where t^e denotes the time to the event from t^* and c is the censorship indicator. If the event is observed, then $t^e \leq T$ and $c = 0$; Otherwise the event is censored and $t^e = T$ and $c = 1$.

The time-to-event distribution is well captured by two functions:

- **Survival function $S^e(t) = \Pr(t^e \geq t)$,** a monotonically decreasing function representing the probability of t^e not earlier than t ;
- **Hazard function $\lambda^e(t)$** representing the instantaneous rate of an event occurrence at time t given that no event occurred before time t .

As detailed in Sec. 4.2, $\lambda^e(t)$ determines $S^e(t)$ and captures the instantaneous risk of a patient experiencing event e at t . As thus, the time-to-event prediction can be achieved by estimating $\lambda^e(t^* + \tau)$ where $\tau \in [1, H]$, H being the maximum time length of the prediction horizon, for a set of events of interest $e \in E$.

Our learning task is to estimate the conditional probability distribution $\lambda^e(t^* + \tau | \bar{\mathbf{x}}, \bar{\mathbf{u}})$, where $\bar{\mathbf{x}}, \bar{\mathbf{u}}$ represents the historical value of the observation and intervention time-series up to the prediction time $\mathbf{x}_{1:t^*}, \mathbf{u}_{1:t^*}$. Fig. 1 illustrates an example of our prediction task for four types of organ failure events and mortality with two co-variants: non-invasive mean blood pressure (observation) and Dopamine drip rate (intervention). The prediction is made at time 8 marked as vertical dashed lines, with forecast horizon up to time 30. From the figure, we can see that the input features (i.e., co-variants) are represented as time-series data where values are measured at discrete time points². The predicted event hazard rate vary over the forecast horizon with neural system failure hazard rate and coagulation failure hazard rate showing similar time-varying behaviors.

²Note that the values from EMR are typically measured at irregular time intervals. We will explain in Sec. 7.1 how we handle such input data.

4 MODEL FORMULATION

In this section, we first describe our proposed deep state-space model, augmented with intervention forecasting, which provides a principled way to capture the interactions among observations, interventions, patient states and their uncertainty. Based on this model, we further present a novel latent-state-generated hazard rate formulation for correlated time-to-event predictions.

Figure 2: Graphical Model of State-based Hazard Rate.

Figure 3: Graphical Model of Multi-Event-Type Hazard Rate.

4.1 State Space Model

To provide a joint time-to-event prediction of multiple clinical events, we need a powerful model that captures the temporal correlations among clinical observations and interventions. To this end, we adopt a Gaussian state space model to explicitly model the latent patient physiological state as shown in Fig. 2. Let $\mathbf{z}_t$ be the latent variable vector that represents the physiological state at time t and $\mathbf{z}_{1:T}$ be the sequence of such latent variables. The system dynamics are defined via two equations:

$$\begin{aligned} p(\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{u}_t) &\sim \mathcal{N}(\mathcal{A}_t(\mathbf{z}_{t-1}) + \mathcal{B}_t(\mathbf{u}_t), \mathbf{Q}) && \text{Transition} \quad (1) \\ p(\mathbf{x}_t | \mathbf{z}_t) &\sim \mathcal{N}(\mathcal{C}(\mathbf{z}_t), \mathbf{R}) && \text{Emission} \quad (2) \end{aligned}$$

Eq. (1) defines the state transition. Specifically, function $\mathcal{A}$ defines the system transition without external influence, i.e., how patient state will evolve from $\mathbf{z}_{t-1}$ to $\mathbf{z}_t$ without intervention. $\mathcal{B}$ captures the effect of intervention $\mathbf{u}_t$ on patient state $\mathbf{z}_t$. In Eq. (2), $\mathcal{C}$ captures the relationship between internal state $\mathbf{z}_t$ and observable measurements $\mathbf{x}_t$. $\mathbf{Q}$ and $\mathbf{R}$ are process and measurement noise covariance matrices. We assume them to be time-invariant. Eq. (1) and (2) subsume a large family of linear and non-linear state space models. For example, by setting $\mathcal{A}, \mathcal{B}, \mathcal{C}$ to be matrices, we obtain linear state space models. By parameterizing $\mathcal{A}, \mathcal{B}, \mathcal{C}$ via deep neural networks, we have a deep state space model.

Intervention Forecast. Contrary to classical state space models, where interventions are usually considered as external factors, when inferring patient states from EMR data, interventions are an integral part of the system, as they are determined by clinicians based on their estimation of patient states and medical knowledge/clinical guidelines. To model this relationship, we augment the state space model with additional dependency from $\mathbf{z}_t$ to $\mathbf{u}_{t+1}$ as shown in Fig. 2.

$$p(\mathbf{u}_t | \mathbf{z}_{t-1}) \sim \mathcal{N}(\mathcal{D}(\mathbf{z}_{t-1}), \mathbf{U}) \quad (3)$$

Similarly, in Eq.(3) $\mathcal{D}$ can be either a matrix for a linear model or parameterized by a neural network for a nonlinear model. For clinical predictions, there are two different questions one may ask: 1) what will happen if *no* intervention is applied; 2) what will happen if the patient receives expected interventions. Our model allows us to answer the second question.

4.2 State-based Discrete-time Hazard Rate

Recall that the hazard rate function describes the instantaneous rate of event occurrence at time t . In classical survival analysis, this rate is usually assumed to be constant over time and statically determined by the co-variants at the time of prediction [9]. Based on the physiological state-space model, we propose a new time-to-event estimation model where the hazard rate function is discretized per time step and dependent on the dynamic latent patient physiological state at that time. Specifically, the hazard rate λ_t^e of event e at time step t is modelled as

$$\lambda_t^e = \mathcal{L}^e(\mathbf{z}_t) \quad (4)$$

where $\mathcal{L}^e$ can be either a linear model or neural network to map the hidden variable $\mathbf{z}_t$ to a deterministic value. The discrete survival function at time t can be written as

$$S^e(t) = (1 - \lambda_t^e) S^e(t-1). \quad (5)$$

Let $S^e(0) = 1$. The above recursion leads to

$$S^e(t) = \prod_{s=1}^t (1 - \lambda_s^e). \quad (6)$$

The incidence density function is defined as $f(t^e) = \Pr(t^e = t)$ and is connected with λ_t^e via

$$f(t^e) = \lambda_t^e \prod_{s=1}^{t-1} (1 - \lambda_s^e). \quad (7)$$

All event $e \in E$ are generated from the shared states $\mathbf{z}_t$ but with its individual generation function $\mathcal{L}^e$. Fig. 3 shows a graph model for two events e, e' . Note that observation and intervention nodes are omitted in this figure for clear illustration.

5 VARIATIONAL INFERENCE

Our state space model is fully specified by the generative parameter $\theta = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{L}^e, e \in E)$. In this section, we present the learning objective and the associated variational lower bound that supports the time-to-event prediction task as described in Sec. 3.

Recall that time to event prediction estimates the time distribution of t^e at t^* based on the historical values of $\bar{\mathbf{x}}, \bar{\mathbf{u}}$. We first consider the log likelihood of one event e represented by (c, t^e) . There are two cases for this event:

- if $c = 1$, which means the event is censored/survived at t^e , then the likelihood is captured by its survival function $\log S_\theta(t^e | \bar{\mathbf{x}}, \bar{\mathbf{u}})$;
- if $c = 0$, which means the event is observed at t^e , then the likelihood is captured by its incidence density function $f_\theta(t^e | \bar{\mathbf{x}}, \bar{\mathbf{u}})$.

Further recall that $\lambda_t^e = \mathcal{L}^e(\mathbf{z}_t)$ and $S^e(t^e) = \prod_{s=1}^{t^e} [1 - \lambda_s^e]$, $f^e(t^e) = \prod_{s=1}^{t^e-1} [1 - \lambda_s^e] \cdot \lambda_{t^e}^e$. Thus $f_\theta^e(t^e)$ and $\log S_\theta^e(t^e)$ are independent from observations and interventions conditioned on hidden state $\hat{\mathbf{z}}$, where $\hat{\mathbf{z}} = \mathbf{z}_{1:t^e}$. Putting both cases together and based on the graph model in Fig. 2, we have the log likelihood of e as follows:

$$\begin{aligned} \log p_\theta(t^e | \bar{\mathbf{x}}, \bar{\mathbf{u}}) &= \underbrace{(1 - c) \cdot \log f_\theta^e(t^e | \bar{\mathbf{x}}, \bar{\mathbf{u}})}_{\text{event is observed at } t^e} \\ &+ \underbrace{c \cdot \log S_\theta^e(t^e | \bar{\mathbf{x}}, \bar{\mathbf{u}})}_{e \text{ is censored/survived at } t^e} \\ &= (1 - c) \cdot \log \int_{\hat{\mathbf{z}}} p_\theta(\hat{\mathbf{z}} | \bar{\mathbf{x}}, \bar{\mathbf{u}}) f_\theta^e(t^e | \hat{\mathbf{z}}) \\ &+ c \cdot \log \int_{\hat{\mathbf{z}}} p_\theta(\hat{\mathbf{z}} | \bar{\mathbf{x}}, \bar{\mathbf{u}}) S_\theta^e(t^e | \hat{\mathbf{z}}) \end{aligned}$$

This log likelihood is intractable when inferring the posterior $p_\theta(\hat{\mathbf{z}} | \bar{\mathbf{x}}, \bar{\mathbf{u}})$. We adopt the variational inference method by introducing a variational distribution q_ϕ that approximates this posterior. The evidence lower bound (ELBO) of the log event time likelihood is thus given as:

$$\begin{aligned} &\underbrace{(1 - c) \cdot \mathbb{E}_{q_\phi(\hat{\mathbf{z}} | \bar{\mathbf{x}}, \bar{\mathbf{u}})} [\log f_\theta^e(t^e | \hat{\mathbf{z}})] + c \cdot \mathbb{E}_{q_\phi(\hat{\mathbf{z}} | \bar{\mathbf{x}}, \bar{\mathbf{u}})} [\log S_\theta^e(t^e | \hat{\mathbf{z}})]}_{\text{time-to-event prediction loss}} \\ &- \underbrace{\text{KL}(q_\phi(\hat{\mathbf{z}} | \bar{\mathbf{x}}, \bar{\mathbf{u}}) || p_\theta(\hat{\mathbf{z}} | \bar{\mathbf{x}}, \bar{\mathbf{u}}))}_{\text{regularization loss}} \end{aligned} \quad (8)$$

The lower bound in Eq.(8) has two components: 1) the log likelihood loss for time-to-event prediction; 2) the regularization loss which measures the difference between the encoder and the simple prior distribution of the latent state $\mathbf{z}$ given the transition model between $\mathbf{z}_{t-1}$ and $\mathbf{z}_t$ as defined in the state space model (Eq.(1)). Similar to [20], this ELBO can be factorized along time as:

$$\begin{aligned} & (1-c) \cdot \mathbb{E}_{q_\phi(\mathbf{z}_t|\bar{\mathbf{x}},\bar{\mathbf{u}})} \left[\sum_{s=1}^{t^e-1} \log(1 - \mathcal{L}^e(\mathbf{z}_t)) + \mathcal{L}^e(\mathbf{z}_t) \right] \\ & + c \cdot \mathbb{E}_{q_\phi(\mathbf{z}_t|\bar{\mathbf{x}},\bar{\mathbf{u}})} \left[\sum_{s=1}^{t^e} \log(1 - \mathcal{L}^e(\mathbf{z}_t)) \right] \\ & - \sum_{t=1}^{t^e} \text{KL}(q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1}, \bar{\mathbf{x}}, \bar{\mathbf{u}}) || p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1}, \bar{\mathbf{u}})) \end{aligned} \quad (9)$$

For a set of events e , the loss function is a sum of all the negative log event time likelihood: $-\sum_{e \in E} \log p_\theta(t^e|\bar{\mathbf{x}}, \bar{\mathbf{u}})$, each of which is based on the same latent state estimation and the hazard rate generation function associated with its event type.

6 MODEL ARCHITECTURE

We describe our learning algorithm and the neural network models used for learning in this section. As shown in Fig. 4, give the ELBO, our learning algorithm proceeds the following steps:

- Inference of $\hat{\mathbf{z}}$ from $\bar{\mathbf{x}}$ and $\bar{\mathbf{u}}$ by an encoder network q_ϕ . We follow the same model architecture as in [20] and use a bi-directional LSTM as the encoder network.
- Sampling based on the current estimate of the posterior $\hat{\mathbf{z}}$.
- Estimate the next step latent state $\mathbf{z}_{2:t}$ via the generative model p_θ and compute the regularization loss. We use two multi-layer perceptrons (MLP) for the state transition module – one for state transition without external influence; the other for the effect of intervention on state transition.
- Estimate the hazard rate for each type of event of interest via the generative model p_θ . Each event has a separate hazard rate emission module which is a MLP.
- Survival function and incidence density function are computed based on the estimated hazard rate, from which the negative log event likelihood loss is computed for each types of event.
- The likelihood loss of all events and the KL-divergence regularization loss are aggregated as the training loss (negative ELBO).
- Estimate the gradients of the loss with respect to θ and ϕ and updating parameters of the model. Gradients are averaged stochastically across mini-batches of the training set.

Our model is implemented in TensorFlow [1], and will be open-sourced³.

7 EXPERIMENTS

We extensively evaluate our proposed deep state-space model (DSSM) over real temporal event data showing that it has better predictive performance for time-to-event prediction, and is able to uncover

	Positive	Negative	Excluded
Mortality	4,277	41,843	4,557
Kidney Failure	2,056	33,655	14,966
Liver Failure	1,474	39,469	9,734
Coagulation Failure	2,496	35,496	12,685
Nervous system Failure	3,475	28,340	18,862

Table 1: Statistics of different event prediction tasks.

meaningful insights about the latent correlation among different types of events.

7.1 Dataset and Data Preprocessing

We use Medical Information Mart for Intensive Care (MIMIC) data [16] in our empirical study. MIMIC-III is a large open EMR dataset containing information relating to patients admitted to critical care units at a large tertiary care hospital. Data includes vital signs, medications, laboratory measurements, procedure codes, diagnostic codes, and more.

We consider inpatients from MIMIC-III who are still alive 48 hours after admission and predict their risk of in-hospital death at 48 hours after admission along with the risk of 4 multiple organ failures based on the SOFA score [17] definition:

- **Kidney Failure:** creatinine $\geq 2\text{mg/dl}$
- **Liver Failure:** bilirubin $\geq 2\text{mg/dl}$
- **Coagulation Failure:** platelet $< 100 \times 10^3/\mu\text{l}$
- **Nervous system Failure:** Glasgow coma scale ≤ 12

As organ failures can be recurrent, we only consider the first occurrence of each type of organ failure in an encounter as the event of interest. The statistics of these events from MIMIC-III are provided in Table 1. The positive (negative) columns are the number of encounters where the corresponding events are observed (unobserved) after the prediction time. The excluded column counts the encounters where the corresponding event occurs before the prediction time. We only include the encounters free of any organ failures in our study. There are 38485 adult in-patient encounters included in the study. The mean length of stay of the encounters is around 10 days. As a side note, the high number of excluded encounters is related to the strong association between organ failures and ICU admission. And the organ failures we included in our study are usually newly developed after ICU admission.

We select 96 most frequently used lab measurements and vital signs as observation features, 8 types of vasopressors and antibiotics and 6 most recorded ventilation and dialysis machine settings as intervention features. For intervention features, we consider the following parameters to indicate the presence and the level of medical interventions: 1) the dosage and drip rate of medication administrations, 2) mechanical ventilation and dialysis machine settings.

As different coding systems are used in MIMIC-III, we harmonize the medical codes corresponding to the same lab/vital measurement as a single feature. In addition, we standardize the unit when a medical code is used with multiple units or without a unit. All observation and intervention values are normalized using z-score,

³<https://github.com/Google-Health/records-research/state-space-model>

Figure 4: Loss Computation Process. The state-space computation task (leftmost) captures the changing dynamics of patients’ latent states $z_{1:t}$ based on past observations and interventions. Each following task of event prediction (mortality, kidney failure, etc.) has its own survival and density function that depend on these latent states $z_{1:t}$.

where the mean and standard deviation of each feature are computed based on the MIMIC-III dataset. The outlier measurements are removed from training. The details of feature selection and data preprocessing are reported in the supplemental material.

Observational data is recorded at irregular intervals in EMR, resulting in a large number of missing values when sampled at regular time steps. Handling missing value in observation data has been investigated in recent works [7]. For lab measurements and vital signs, we adopt a simple method where the most recent value is used to impute the missing ones. For interventions, the situation is more complex and not handled in any existing works. Specifically, we need to differentiate the case where a missing value represents that the intervention is not performed or has been completed vs. the case where a missing value means the same setting is continued at this time step. To address this issue, we follow the observation that most (continuous) medication are administrated at regular intervals and organ support machine settings are also regularly adjusted. We first derive the distribution of inter-medication-administration and inter-intervention-setting time, then pick the 90-percentile time as the cut-off threshold. If two consecutive interventions are within the time range of their corresponding thresholds, then we consider the missing value as an indication of a continuous action and use the last setting as its missing value. If it falls outside of this range, then a missing value is considered as no action.

7.2 Prediction Performance

We first evaluate the overall performance of time-to-event predictions for each event. The following metrics are used:

- **C-index** (i.e., concordance index) measures the extent to which the ordering of actual event times of pairs agrees with the ordering of their predicted risk. It is a widely used discriminative metric for evaluating the performance of survival models.
- **AUC-ROC and AP** (a.k.a. AUPRC) within two fixed prediction windows $[0, 24]$ hr and $[0, 48]$ hr. This metric evaluates the short-term prediction performance, while C-index evaluates the overall model prediction power. To accurately compute AUC-ROC and AP for an event prediction within a fixed time window in the presence of censorship, we only consider 1) the events which are observed within this window as positives, and 2) the events which are either observed or censored outside of the window as negatives. These are the cases where we can be sure that the event does not occur in the window. If an event is censored within the window (e.g., the patient is discharged within the window without the specified event being observed), it is not included in the computation.

We compare our methods with two state-of-art methods:

- **DeepSurv** [19] is a Cox proportional hazards deep neural network that models the interactions between a patient’s covariates and the outcome. DeepSurv model uses a MLP to

Figure 5: Event hazard rate and true occurrence time for both observed (in solid dot) and censored (in light cross) events. For each prediction task, the fitted hazard rate is able to capture the true event time accurately, while for the censored events, the respective hazard rates are low as expected.

Figure 6: Different correlated trajectories of Multi-Organ Failure and Mortality. In Case 1 and 2, mortality is highly correlated with other organ failures as indicated by the learned hazard rates. In contrast, for Case 3, the neural system failure and coagulation have little influence on mortality, which is also reflected in the learned hazard rate curve.

first encode the co-variants then use them as the coefficients in the Cox model. In our experiment, we use a 3-layer MLP with sigmoid activation as the encoder and the most recent observation values of each feature as the co-variants.

- **DRSA** [31] refers to Deep Recurrent Survival Analysis, which is a model based on the recurrent neural network. In this model, the l -th RNN cell predicts the instantaneous hazard rate at time step l . This method can be considered as the

deterministic counterpart of our deep state space encoding. The key difference between our work is that there is no regularization loss as in Eq(8) in DRSA. In addition, DRSA considers a static co-variant x , while our model considers the time series of observations and interventions as features from which the intrinsic dynamics of the hidden states are learned and used to forecast the hazard rate. For fair comparison, in our experiment, we use an improved version of

DRSA, where the time series of observation and intervention values are encoded using LSTM as co-variant $\mathbf{x}$.

- **DSSM** is our proposed deep state space model. This model encodes the temporal correlations among observations, interventions and hidden states, which are rolled-out to the prediction horizon step by step to generate the hazard rate for each event.

The hyperparameters including the learning rate, the hidden state size for LSTM, the number of units and layers for MLP, the size of time step are tuned. The experiment uses a hidden state size of 50 for LSTM and 32 hidden units with 3 layers for all MLPs, including the state transition MLP, the intervention effect MLP and observation emission MLP. A learning rate of 0.0001 is used for **DSSM** and **DRSA** and learning rate of 0.0005 is used for **DeepSurv**. In the experiment, each time step takes 12 hours. The prediction rolls out to 240 time steps, beyond which a constant hazard rate which is the same as the last time step is used.

To demonstrate the performance variance in our evaluation, we use 10-fold cross validation. For each fold, we split the dataset into train/eval/test according to 80%/10%/10% based on the hash value of the patient ID. We estimate the standard error of the mean based on the sample standard deviation of these 10 folds. Table 3 reports the mean and the standard error for the three models. We can see that our proposed method **DSSM** outperforms **DeepSurv** and **DRSA** on all the metrics. In addition, our model brings significant improvement on the short-term predictions in terms of both AUC-ROC and AP, as the state roll-outs tend to be more accurate at the closer forecast horizon. Though **DRSA** also rolls out the state of a LSTM for hazard rate prediction, its performance is limited by the LSTM’s tendency to use recent history. While our approach incorporates the regularization loss that minimizes the KL divergence between the encoded state and the prior distribution of the latent state under the state transition model, which encourages the true state dynamics to be learnt.

7.3 Hazard Rate Trajectory

We now zoom in to individual patients and study the hazard rate dynamics. Fig. 5 plots the hazard rate trajectory within the first 240hr after the prediction time for mortality and organ failures separately on each row. All plots share the same y-axis range $[0, 0.2]$. The first three figures on each row plot the cases where the corresponding event happens to the patient. The true event time is plotted in the figure as a dot. The last three figures plot the cases where the event is not observed and the censor time (when the patient is discharged) is plotted as a cross.

Comparing these figures, we can clearly see that the hazard rate is significantly higher for the observed cases than the cases where events are not happened in the observation window. In addition, the flexible discretized-time model provides more accurate instantaneous hazard rate estimation, as reflected in the fact that a large portion of the events happen around the time where the hazard rate is peaked. This is in contrast with the constant hazard assumption which is usually made in the conventional survival analysis[9]. Lastly we observe that all the trajectories tend to converge to a base value after around 150 hrs due to the fact that no new data comes in after the prediction point. This baseline values vary for different patients and different events.

	Neural.	Liver	Kidney	Coagulation
Case 1	0.596	0.745	0.741	0.577
Case 2	0.810	0.838	0.914	0.854
Case 3	0.477	0.565	0.784	0.5270

Table 2: Correlations between Mortality and Organ Failures.

7.4 Hazard Rate Correlations

We further show that our model can reveal the correlations among these events which can in turn facilitate the understanding of the progression of patient conditions. Fig. 6 plots the hazard rate trajectories along with their true event time of three patient encounters – Case 1 and 2 corresponding to mortality cases and Case 3 corresponding to a survived case. For presentation clarity, only the organ failures that happened within the window of 240hr are plotted. For both mortality case, multiple organ failures are predicted with high hazard rate around the prediction of the increased mortality hazard rate. This prediction is validated by the sequence of true organ failure and mortality. For Case 3, the neural system failure and coagulation failure are predicted at high risk at the beginning of the prediction horizon. The true event time for both failures are 1hr after prediction time. They increase the mortality risk slightly at the beginning. All three risks – neural system failure, coagulation failure and mortality drop significantly and the patient is discharged at 266hr.

From the figure, we can also see that the mortality hazard rate trajectory is strongly correlated with the trajectory of organ failures. In order to quantify the correlation, we cross-correlate organ failure trajectories with mortality trajectory and show the correlation coefficients in Table. 2. Though mortality has positive correlations with all organ failures, the level of correlation varies by individual patient and different organs, which reveals valuable insights. In Case 1, mortality is predicted to be highly correlated with the kidney and liver failure, and both are also predicted with high predicted risks. Similarly, in Case 2, mortality is highly correlated with liver, kidney, coagulation failures, and all are predicted with high risks. This provides an explanation for the mortality events in addition to the mortality risk prediction. In contrast, for Case 3, though the neural system failure and coagulation failure are predicted to have high risk, the correlation between mortality and neural system and coagulation is relatively low as shown in Table. 2, indicating mortality is less influenced by these two organ failures in this case.

Comparing with the single mortality event prediction, our correlated predictions provide insights into why mortality may happen. This offers clinicians with a full picture of a patient’s medical condition and better supports them with better decision making.

8 CONCLUSIONS

We proposed a deep latent state-space generative model to capture the relations between patients’ mortality risk and the associated organ failure risks. Based on the learned patients’ states, we further develop a new formulation of the hazard rate function to fit general discrete-time survival distribution of observed events.

	C-index	AUC@24	AP@24	AUC@48	AP@48
Mortality					
DeepSurv	0.769 (0.009)	0.911 (0.015)	0.25 (0.056)	0.861 (0.014)	0.228 (0.043)
DRSA	0.743 (0.016)	0.906 (0.025)	0.28 (0.0092)	0.837 (0.015)	0.186 (0.031)
DSSM	0.7769 (0.007)	0.949 (0.021)	0.375 (0.091)	0.873 (0.016)	0.258 (0.036)
Kidney Failure					
DeepSurv	0.826 (0.007)	0.957 (0.007)	0.315 (0.037)	0.901 (0.009)	0.319 (0.033)
DRSA	0.810 (0.007)	0.944 (0.011)	0.221 (0.039)	0.876 (0.011)	0.221 (0.035)
DSSM	0.829 (0.003)	0.981 (0.005)	0.485 (0.057)	0.914 (0.005)	0.365 (0.033)
Liver Failure					
DeepSurv	0.709 (0.015)	0.752 (0.018)	0.062 (0.017)	0.733 (0.015)	0.071 (0.020)
DRSA	0.702 (0.017)	0.778 (0.025)	0.056 (0.021)	0.745 (0.026)	0.056 (0.017)
DSSM	0.709 (0.012)	0.843 (0.019)	0.132 (0.032)	0.784 (0.019)	0.101 (0.002)
Coagulation Failure					
DeepSurv	0.831 (0.012)	0.875 (0.020)	0.239 (0.041)	0.863 (0.019)	0.265 (0.031)
DRSA	0.803 (0.007)	0.928 (0.011)	0.196 (0.029)	0.861 (0.009)	0.216 (0.015)
DSSM	0.835 (0.007)	0.942 (0.007)	0.292 (0.039)	0.890 (0.009)	0.272 (0.019)
Neural Sys. Failure					
DeepSurv	0.852 (0.003)	0.907 (0.005)	0.587 (0.013)	0.876 (0.004)	0.552 (0.006)
DRSA	0.849 (0.003)	0.948 (0.01)	0.68 (0.026)	0.870 (0.007)	0.526 (0.012)
DSSM	0.863 (0.004)	0.968 (0.006)	0.751 (0.02)	0.889 (0.005)	0.586 (0.008)

Table 3: Time-to-Mortality Prediction Performance. Parentheses denote standard error.

Extensive experiments over MIMIC datasets show that our proposed model not only outperforms several state-of-art baselines in terms of prediction accuracy, but also provides meaningful insights into the temporal relations among the multiple types of events. By demonstrating the correlations between different organ failures and mortality risk, we provide physicians with more evidence to have better decision-making.

REFERENCES

- [1] Martin Abadi and et al. 2015. TensorFlow: A System for Large-Scale Machine Learning.
- [2] Ahmed M. Alaa and Mihaela van der Schaar. 2017. Deep Multi-task Gaussian Processes for Survival Analysis with Competing Risks. In *NIPS*. 2329–2337.
- [3] Ahmed M. Alaa and Mihaela van der Schaar. 2019. Attentive State-Space Modeling of Disease Progression. In *NeurIPS*.
- [4] James E. Barretta and Anthony C. C. Coolena. 2010. Gaussian process regression for survival data with competing risks.
- [5] Alexis Bellot and Mihaela van der Schaar. 2018. Multitask Boosting for Survival Analysis with Competing Risks. In *NeurIPS*.
- [6] Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin, and Ricardo Henao. 2018. Adversarial Time-to-Event Modeling. In *ICML*.
- [7] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci. Rep.* 8, 1 (2018).
- [8] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2015. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. (Nov. 2015). arXiv:1511.05942
- [9] D. R. Cox. 1992. Regression models and life-tables. In *Breakthroughs in statistics*. 527–541.
- [10] Tamara Fernandez, Nicolas Rivera, and Yee Whye Teh. 2016. Gaussian Processes for Survival Analysis. In *NIPS*. 5021–5029.
- [11] J. P. Fine and R. J. Gray. 1999. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American statistical association* 94, 446 (1999), 496–509.
- [12] Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. 2017. A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning. In *NIPS*.
- [13] Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. 2016. Sequential Neural Models with Stochastic Layers. In *NIPS*.
- [14] Marzyeh Ghassemi, Mike Wu, Michael C. Hughes, Peter Szolovits, and Finale Doshi-Velez. 2017. Predicting intervention onset in the ICU with switching state space models. *American Medical Informatics Association (AMIA)*.
- [15] E. Giunchiglia, A. Nemchenko, and M. van der Schaar. 2018. RNN-SURV: A Deep Recurrent Model for Survival Analysis. In *International Conference on Artificial Neural Networks (ICANN)*.
- [16] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, A Freely Accessible Critical Care Database. *Scientific Data* 3 (2016). Article number: 160035.
- [17] Alan E. Jones, Stephen Trzeciak, and MD Jeffrey A. Kline. 2010. The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation.
- [18] Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. 2017. Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data. In *ICLR*.
- [19] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology* 18, 1 (2018), 24.
- [20] Rahul G. Krishnan, Uri Shalit, and David Sontag. 2015. Deep Kalman Filters. *CoRR abs/1511.05121* (2015).
- [21] Rahul G Krishnan, Uri Shalit, and David Sontag. 2017. Structured Inference Networks for Nonlinear State Space Models. In *AAAI*.
- [22] Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. 2018. DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks. In *AAAI*.
- [23] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. 2016. Learning to Diagnose with LSTM Recurrent Neural Networks. In *International Conference on Learning Representations (ICLR)*.
- [24] Zitao Liu and Milos Hauskrecht. 2013. Clinical Time Series Prediction with a Hierarchical Dynamical System. *Artificial Intelligence in Medicine* (2013), 227–237.
- [25] Zitao Liu and Milos Hauskrecht. 2016. Learning Adaptive Forecasting Models from Irregularly Sampled Multivariate Clinical Data. In *AAAI*.
- [26] Jiayu Zhou Dongxiao Zhu Lu Wang, Yan Li and Jieping Ye. 2017. Multi-task Survival Analysis. In *IEEE International Conference on Data Mining*.
- [27] Wu M, Ghassemi M, Feng M, Celi LA, Szolovits P, and Doshi-Velez F. 2017. Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database. *J Am Med Inform Assoc* (2017), 488–495.
- [28] Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Continuous State-Space Models for Optimal Sepsis Treatment: a Deep Reinforcement Learning Approach. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*. 147–163.
- [29] Alvin Rajkomar and et al. 2018. Scalable and Accurate Deep Learning with Electronic Health Records. *Digital Medicine* 1 (2018). Article number: 18.
- [30] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. 2016. Deep Survival Analysis. In *Proceedings of Machine Learning Research*. Vol. 56. 101–114.
- [31] Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. 2019. Deep Recurrent Survival Analysis. In *AAAI*.
- [32] Ying Sha and May D Wang. 2017. Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*.
- [33] Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. 2018. Attend and Diagnose: Clinical Time Series Analysis Using Attention Models. In *AAAI*.
- [34] Yuan Xue, Denny Zhou, Nan Du, Andrew M. Dai, Zhen Xu, Kun Zhang, and Claire Cui. 2019. Deep Physiological State Space Model for Clinical Forecasting. In *NeurIPS Workshop on Machine Learning for Health*.
- [35] Jieping Ye Yan Li, Jie Wang and Chandan K. Reddy. 2016. A Multi-Task Learning Formulation for Survival Analysis. In *22nd ACM SIGKDD*.
- [36] Quan Zhang and Mingyuan Zhou. 2018. Nonparametric Bayesian Lomax delegate racing for survival analysis with competing risks. In *NeurIPS*.

Feature	unit
cardiac index	UNKNOWN UOM
cardiac output rate	L PER MIN
chloride	MEQ PER L
co2	MEQ PER L
creatinine kinase	IU PER L
creatinine	MG PER DL
creatinine kinase mb	NG PER ML
cvp	MMHG
eosinophils	PERCENT
exhaled minute ventilation low	
exhaled minute ventilation low	L PER MIN
expiratory ratio	RATIO
fio2 analyzed	TORR
glucose	MG PER DL
glucose poc	MG PER DL
heart rate	BPM
hematocrit	PERCENT
hemoglobin	G PER DL
inr	RATIO
inr	
insp pressure	CM H2O
insp time	S
inspiratory ratio	RATIO
ionized calcium	MEQ PER L
ketones urine	MG PER DL
lactate	MMOL PER L
lymphocytes diff	PERCENT
magnesium	MG PER DL
mean ch	PG
mean chc	PERCENT
mean cv	FL
minute ventilation obs	L PER MIN
monocytes	PERCENT
neutrophils urine	PERCENT
o2 flow	L PER MIN
o2 saturation	PERCENT
o2 saturation p	PERCENT
p co2	MMHG
p o2	MMHG
paw	CM H2O
peep observed	CM H2O
ph	PH
ph urine	PH
phosphorous	MEQ PER L
plateau pressure	CM H2O
platelet	
platelet	K PER UL
potassium	MEQ PER L
potassium urine	MEQ PER L
protein urine	MG PER DL
psv	UNKNOWN UOM
pt	S
ptt	S
rbc	PER UL
rdw	PERCENT
replacement rate	ML PER H
respiratory rate	BREATHS PER MIN
respiratory rate spont	BREATHS PER MIN
respiratory rate total	BPM
rrt output	ML
sodium	MEQ PER L
specific gravity urine	
svo2	PERCENT
tbili	MG PER DL
tco2	MEQ PER L
temperature	CEL
troponin t	NG PER ML
urine output	ML
urine output foley	ML
vt obs	ML PER BREATH
vt spont	ML PER BREATH
wbc count	K PER UL
weight	KG

Table 5: Observation Feature and Unit (Part 2)

Feature	unit
30042	MCG KG MIN
30043	MCG KG MIN
30044	MCG MIN
30047	MCG MIN
30120	MCG KG MIN
30127	MCG MIN
30306	MCG KG MIN
30307	MCG KG MIN
dialysate rate	ML PER H
fi o2	PERCENT
peep	CM H2O
pip	CM H2O
respiratory rate setting	BPM
vt set	ML PER BREATH

Table 6: Intervention Feature and Unit

9 SUPPLEMENTARY MATERIALS

9.1 Data Preprocessing Details

We preprocess the MIMIC-III dataset in the following steps.

- (1) **Code harmonization.** This is a manual process based on the inputs from clinical experts. In this step, the medical codes corresponding to the same measurements from different coding systems, including LONIC and MIMIC specific coding, are harmonized into the same entity. For example, serum creatinine is associated the following MIMIC-III specific codes: 220615, 50912, 1525, 3750, 791, and LONIC code 2160-0.
- (2) **Unit conversion.** This is an automated process with manual review. In MIMIC-III, a medical code may be used in multiple units and sometimes miss a unit. To determine whether these two entries correspond to the same measurement concept, we derive its value range and mean under different units and test whether they are similar to each other. The final results are reviewed manually.
- (3) **Outlier removal.** We derive the distribution of each measurement code after harmonization and unit conversion. We remove the outliers, defined as below $0.1 \times$ the value at 1 percentile or above $10 \times$ the value at 99 percentile.
- (4) **Value normalization.** We collect the mean and standard deviation over the cleaned dataset for each harmonized code and compute its z-score as feature value.

9.2 Features

We record the features and their units used in the experiment in Table 4-6.

Feature	unit
access pressure	MMHG
albumin	G PER DL
alt	IU PER L
anion gap	MEQ PER L
ap	IU PER L
arterial base excess	MEQ PER L
arterial bicarbonate	MEQ PER L
arterial pco2	MMHG
arterialph	PH
arterial po2	MMHG
ast	IU PER L
base excess	MEQ PER L
basophils	PERCENT
blood flow	ML PER MIN
bp diastolic invasive	MMHG
bp diastolic non invasive	MMHG
bp map invasive	MMHG
bp mean non invasive	MMHG
bp systolic invasive	MMHG
bp systolic non invasive	MMHG
bun	MG PER DL
calcium	
calcium	MEQ PER L

Table 4: Observation Feature and Unit (Part 1)