Time-to-Event Prediction for Correlated Clinical Events based on State Space Model

Yuan Xue, Denny Zhou, Nan Du, Andrew Dai, Zhen Xu, Kun Zhang, Claire Cui
yuanxue,dennyzhou,dunan,adai,zhenxu,kunzhang,claire@google.com
Google

ABSTRACT

Understanding and capturing the interactions among multiple events are important for not only deriving more accurate time progression of these events, also critical for designing treatment plans that simultaneously handles multiple events. In this work, we propose a deep state space model to capture the interactions among multiple clinically critical events (e.g., acute kidney failure, mortality) by explicitly modeling the dynamics of patients' latent physiological states. Based on the shared physiological states, we are able to not only provide more accurate estimation of distribution of survival time for each event but capture their relations. Comprehensive empirical evaluations show that our proposed model compares favorably to the state-of-the-art methods on real EMR data.

1 INTRODUCTION

Time-to-event prediction (also called survival analysis) investigates the distribution of time duration until the event of interests happen in presence of event censorship. It is an essential tool in healthcare domain.

Recently, machine learning methods have been applied to time-to-event predictions to provide flexible modeling of the time distribution [1], and modeling of the nonlinear relationships between co-variants and the risk of an event [2]. For example, existing works have extended the classical Cox proportional hazards model with neural network-based covariate encodings [3, 4] and with multi-task formulations [5, 6]. The work of [7] converts the time-to-event estimation to a discretized-time classification problem, while others use a continuous-time model based on Gaussian processes [8–10] or generative adversarial networks [11] to model the nonlinear relationship between covariates and the time.

Most of these prior works [3, 12] on survival analysis are only applicable to a single survival prediction task and lack the capability of analyzing the correlations among multiple survival predictions. In reality, most events are related to or even caused by one another. For example, in medical domain, death may be caused by a single organ failure or multiple simultaneous organ failures, with multi-organ failure significantly increases the risk of death in a non-linear fashion. Further, the dysfunction or failure of one organ will also cause the dysfunction/failure of another (e.g. kidney failure maybe caused by liver damage). Thus the prediction of events of interest will be influenced by their simultaneous risks of developing related diseases.

KDD '19, August 4 - 8, 2019, Anchorage, Alaska

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Understanding and capturing the interactions among multiple events are important for not only deriving more accurate time progression of these events, also critical for designing treatment plans that simultaneously handles multiple events. For examples. 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.

On the other hand, the wide adoption of electronic medical records (EMR) has resulted in 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 relationships among patient conditions and the onset of multiple clinically critical events.

In this paper, we present a deep state space generative model, which provides simultaneous time-to-event prediction of multiple clinical events. Specifically, our model includes a joint prediction of mortality risk, organ failure risk trajectories based on patterns in temporal progressions, and correlations between past measurements and clinical interventions. The contributions of this paper are as follows: 1) We present a deep state space generative model, augmented with intervention forecasting, which provides a principled way to capture the interactions among observations, interventions, critical event occurrences and true physiological states. 2) Based on the temporal dynamics of patients' physiological states, we develop a new discrete-time hazard rate model that provides flexible fitting of general survival time distributions. The ability to jointly forecast multiple clinical events provides clinicians with a full picture of a patient's medical condition and better supports them with decision making.

2 TIME-TO-EVENT PREDICTION FOR MULTIPLE CORRELATED EVENTS

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 system[1]. 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: 1) **observations 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

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

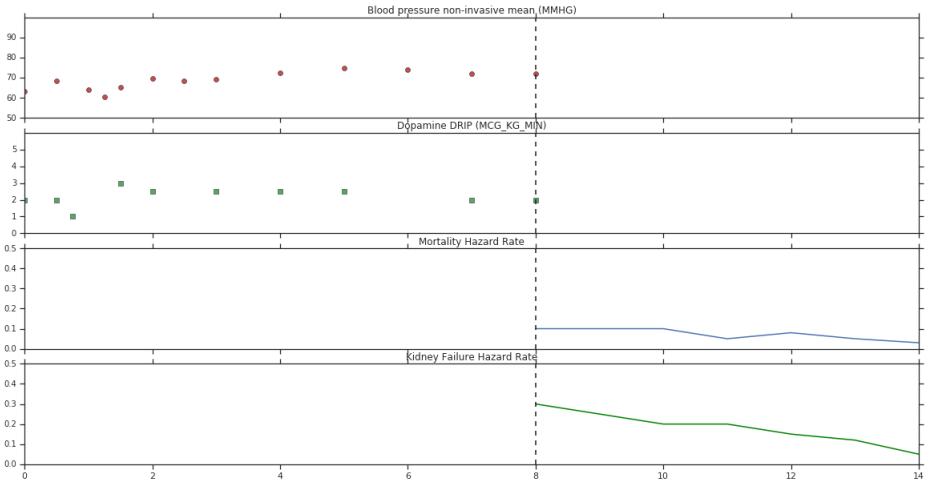

Figure 1: Time-calibrated multiple event risk trajectory.

pressure, serum lactate). We use $\mathbf{x}_{1:T}$ to denote the sequence of measurements at discrete time points $t = 1, ..., T$; 2) **interventions 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 $t = 1, \ldots, T$.

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 to 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 $t^e <= T$ and $c = 0$; if the event is censored then $t^e = T$ and $c = 1$. The time-to-event distribution is well captured by: 1) *Survival function* $S^e(t) = \Pr(t^e \geq t)$, a monotonically decreasing function representing the probability of t^e not earlier than t; 2) *Hazard function* $\lambda^e(t)$ representing the rate of an event at time t given that no event occurred before time t. As detailed in Sec. 3.2, $\lambda^e(t)$ determines $S^e(t)$ and captures the risk of a patient experiencing event e at t. As thus, our prediction task is to estimate $\lambda^e(t^* + \tau)$ where $\tau \in [1, H]$ for a set of events of interest $e \in E$.

Fig. 1 illustrates the time-calibrated event risk predictions for two events: kidney failure and mortality along with two time-series covariants: non-invasive mean blood pressure and Dopamine drip rate. From the figure, we are able to see how the risk of these two events correlate and change over time.

3 MODEL

3.1 State Space Model

To provide a joint time-to-event prediction of multiple events, we need a powerful model that captures the temporal correlations among 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

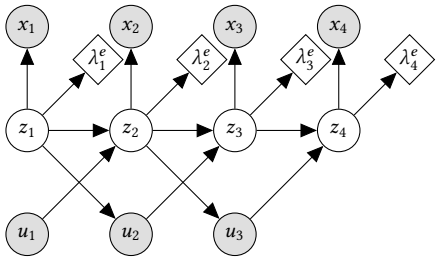

Figure 2: Graph Model For State-based Hazard Rate Generation.

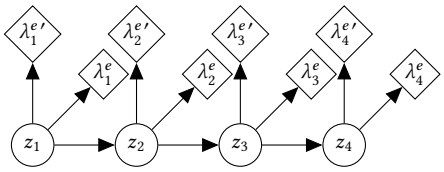

Figure 3: Graph Model For Multi-Event Hazard Rate Generation.

defined as:

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{u}_t) \sim \mathcal{N}(\mathcal{A}_t(\mathbf{z}_{t-1}) + \mathcal{B}_t(\mathbf{u}_t), \mathbf{Q}) \quad \text{Transition} \quad (1)$$

$$p(\mathbf{x}_t|\mathbf{z}_t) \sim \mathcal{N}(C(\mathbf{z}_t), \mathbf{R}) \quad \text{Emission} \quad (2)$$

where Eq. (1) defines the state transition: 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), 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 Gaussian state space models. For example, by setting $\mathcal{A}, \mathcal{B}, C$ to be matrices, we obtain linear

state space models. By parameterizing $\mathcal{A}, \mathcal{B}, C$ via deep neural networks, we have deep Gaussian state space models.

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.

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

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, which is more meaningful clinically.

3.2 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 [12]. 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

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

where $\mathcal{L}^e$ can be either a linear model or neural network to map the Gaussian variable $\mathbf{z}_t$ to a deterministic value. The discrete survival function at time t can be written as $S^e(t) = (1 - \lambda_t^e)S^e(t-1)$. Let $S^e(0) = 1$. for all events. The above recursion leads to

$$S^e(t) = \prod_{s=1}^{t}(1 - \lambda_s^e). \tag{5}$$

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

$$f^e(t) = \lambda_t^e \prod_{s=1}^{t-1}(1 - \lambda_s^e). \tag{6}$$

All event $e \in E$ are generated from the shared states but with its individual generation function $\mathcal{L}^e$. Fig. 3 shows a graph model for two events e, e'[2].

4 VARIATIONAL INFERENCE FOR TIME TO EVENT PREDICTION

Our state space model is fully specified by the generative parameter $\theta = (\mathcal{A}, \mathcal{B}, C, \mathcal{D}, \mathcal{L}^e, e \in E)$. In this section, we present three learning learning objectives and their associated variational lower bounds that support the clinical forecast tasks as described in Sec. 2. We also present the algorithm and the neural network models used for learning.

[2]Observation and intervention nodes are omitted in this figure for clear illustration.

Recall that time to event prediction estimates the time distribution of t^e at t^* based on the historical values of $\bar{\mathbf{x}}, \bar{\mathbf{u}}$. Thus we first consider the log likelihood of one event e happening at time t^e:

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

where $\hat{\mathbf{z}} = \mathbf{z}_{1:t^e}$ and recall that $\lambda_t = \mathcal{L}(\mathbf{z}_t)$ and $S(t^e) = \prod_{s=1}^{t^e}[1-\lambda_s]$, $f(t^e) = \prod_{s=1}^{t^e-1}[1-\lambda_s]\cdot\lambda_{t^e}$.

The ELBO of the log event time likelihood is given as:

$$\begin{aligned}&(1-c)\cdot \mathop{\mathbb{E}}_{q_\phi(\hat{\mathbf{z}}|\bar{\mathbf{x}}, \bar{\mathbf{u}})}\left[\log f_\theta(t^e|\hat{\mathbf{z}})\right] + c \cdot \mathop{\mathbb{E}}_{q_\phi(\hat{\mathbf{z}}|\bar{\mathbf{x}}, \bar{\mathbf{u}})}\left[\log S_\theta(t^e|\hat{\mathbf{z}})\right] \qquad (7) \\ &- \mathbb{KL}(\, q_\phi(\hat{\mathbf{z}}|\bar{\mathbf{x}}, \bar{\mathbf{u}})||p_\theta(\hat{\mathbf{z}}|\bar{\mathbf{x}}, \bar{\mathbf{u}})) \\ &= (1-c)\cdot \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}_t|\bar{\mathbf{x}}, \bar{\mathbf{u}})}\left[\sum_{s=1}^{t^e-1}\log(1-p_\theta(\lambda_t|\mathbf{z}_t)) + p_\theta(\lambda_t|\mathbf{z}_t)\right] \\ &+ c \cdot \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}_t|\bar{\mathbf{x}}, \bar{\mathbf{u}})}\left[\sum_{s=1}^{t^e}\log(1-p_\theta(\lambda_t|\mathbf{z}_t))\right] \\ &- \sum_{t=1}^{t^e}\mathbb{KL}(\, q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1}, \bar{\mathbf{x}}, \bar{\mathbf{u}})||p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1}, \bar{\mathbf{u}})\,) \qquad (8)\end{aligned}$$

For a set of events $e \in E$, the loss function is a weighted sum of all the negative log event time likelihood: $-\sum_{e\in E}\log p_\theta(t^e|\bar{\mathbf{x}}, \bar{\mathbf{u}})$, which are based on the same latent state estimation and naturally lead to a multi-task training framework.

Give the ELBOs of the above tasks, our learning algorithm proceeds the following steps: 1) inference of $\mathbf{z}$ from $\mathbf{x}$, and $\mathbf{u}$ by an encoder network q_ϕ; 2) sampling based on the current estimate of the posterior $\mathbf{z}$ to estimate the weighted sum of the likelihood of all the events based on the generative model p_θ; 3) estimating gradients of the loss (negative ELBO) with respect to θ and ϕ depending on the task and updating parameters of the model. Gradients are averaged across stochastically sampled mini-batches of the training set. We follow the same model architecture as in [13] and use a LSTM as the encoder network, MLP for the state transition, observation emission, and hazard rate generation functions.

5 EXPERIMENTS

5.1 Dataset And Data Preprocessing

We use Medical Information Mart for Intensive Care (MIMIC) data [14] in our empirical study. MIMIC-III is a large open dataset comprising 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, hospital length of stay, survival data, and more.

We select inpatients from MIMIC-III who are still alive 48 hours after admission as our study cohort and predict their risk of in-hospital death at 48 hours after admission along with organ failure risks. For organ failures, we adopt the definition of sequential organ failure assessment score (SOFA score) [15] and take the experts' inputs to determined the threshold: 1) *Liver Failure* is defined as

bilirubin ≥ 6; 2) *Kidney Failure* as creatinine ≥ 2; 3) *Coagulation System Failure* as platelet < 100; 4) *Nervous system failure* as Glasgow score ≤ 10 We consider the following parameters that indicate the presence and the level of life interventions: 1) the dosage and drip rate of vasopressors and antibiotics 2) mechanical ventilation and dialysis machine settings. There are 42026 in-patient encounters included in the study with 3175 observed in hospital death. We select the 88 most frequently used observational data features and 17 types of vasopressors and antibiotics, 5 most recorded ventilation and dialysis machine settings as intervention features. All observation and intervention values are normalized using z-score, where the mean and standard deviation of each feature are computed based on training data.

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 [16]. 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. We need to differentiate the case where a missing value represents that the intervention is not performed or completed vs. the case where a missing value means the same setting is continued at this time step. We first derive the distribution of inter-medication-administration time and the inter-intervention-setting time. Then we 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 for its missing value. If it falls outside of this range, then a missing value is considered as no action.

5.2 Performance Evaluation

We use the following metrics for evaluating the time-to-event prediction performance: 1) **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 the most widely used metric for evaluating the performance of survival models. 2) **AUC-ROC and AP** (also called AUPRC) within two fixed time 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. We compare the following models in this study:

Table 1: Time-to-Mortality Prediction Results.

	C-Index	AUC-ROC@24	AP@24	AUC-ROC@48	AP@48
Deep Cox					
mortality	0.541384	0.604488	0.174972	0.533015	0.272601
liver failure	0.629617	0.664716	0.22288	0.670643	0.382857
kidney failure	0.625078	0.663327	0.231871	0.666792	0.390601
coagulation system failure	0.619296	0.647936	0.235104	0.651341	0.398273
nervous system failure	0.629617	0.664716	0.22288	0.670643	0.382857
DSSM					
mortality	0.724085	0.700137	0.248567	0.700137	0.211497
liver failure	0.718561	0.818175	0.413299	0.818175	0.555788
kidney failure	0.713212	0.809073	0.410789	0.789161	0.553899
coagulation system failure	0.711887	0.788607	0.406456	0.772243	0.55162
nervous system failure	0.718561	0.818175	0.413299	0.800478	0.555788

Deep Cox [3] is a baseline model. Cox Proportional Hazard model [12] is a widely used survival model. Deep Cox model uses a MLP to first encode the co-variants then use them as the co-efficients in the Cox model. In the experiment, we use a 3-layer MLP with sigmoid activation as the encoder and the most recent observation values as the patient co-variants.

DSSM is our proposed method, where the intervention-augmented deep state space model is roll-out to the predictionn horizon step by step to generate the hazard rate at that time for each event.

The hyperparameters including the learning rate and the hidden state size for LSTM are tuned. The experiment uses a hidden state size of 30 and learning rate of 0.0003. In the experiment, each timestep take 1 hour, and the prediction rolls out to 120 hours. From Table 1, we can see that our proposed method **DSSM** outperforms **Deep Cox** on all the metrics. In addition, the discrete-time state-based time-to-event estimation 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.

6 CONCLUSION

In this work, we present a joint prediction of organ failure risks and mortality risk. Our prediction model is built upon on the deep state space model of the hidden patient physiological state, which provides a principled way to capture the interactions among observations, interventions and true physiological state. Extensive experiment study over MIMIC datasets shows that our proposed outperforms the state-of-art methods.

REFERENCES

[1] Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Association for the Advancement of Artificial Intelligence*. 2018.

[2] Noémie Elhadad David Blei Rajesh Ranganath, Adler Perotte. Deep survival analysis. In *Proceedings of Machine Learning Research*, volume 56, pages 101–114.

[3] 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):24, 2018.

[4] E. Giunchiglia, A. Nemchenko, and M. van der Schaar. Rnn-surv: A deep recurrent model for survival analysis. In *International Conference on Artificial Neural Networks (ICANN)*, 2018.

[5] Jiayu Zhou Dongxiao Zhu Lu Wang, Yan Li and Jieping Ye. Multi-task survival analysis. In *IEEE International Conference on Data Mining*. 2017.

[6] Jieping Ye Yan Li, Jie Wang and Chandan K. Reddy. A multi-task learning formulation for survival analysis. In *22nd ACM SIGKDD*. 2016.

[7] Lei Zheng Zhengyu Yang Weinan Zhang Lin Qiu Yong Yu Kan Ren, Jiarui Qin. Deep recurrent survival analysis. In *AAAI*, 2019.

[8] Tamara Fernandez, Nicolas Rivera, and Yee Whye Teh. Gaussian processes for survival analysis. In *Advances in Neural Information Processing Systems 29*, pages 5021–5029. 2016.

[9] Ahmed M. Alaa and Mihaela van der Schaar. Deep multi-task gaussian processes for survival analysis with competing risks. In *Advances in Neural Information Processing Systems 30*, pages 2329–2337. 2017.

[10] James E. Barretta and Anthony C. C. Coolena. Gaussian process regression for survival data with competing risks. 2010.

[11] Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin, and Ricardo Henao. Adversarial time-to-event modeling. In *ICML*, 2018.

[12] D. R. Cox. Regression models and life-tables. In *Breakthroughs in statistics*, page 527–541. 1992.

[13] Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *CoRR*, abs/1511.05121, 2015.

[14] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 2016. Article number: 160035.

[15] Sequential organ failure assessment score. URL https://en.wikipedia.org/wiki/SOFA_score.

[16] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.*, 8(1), 2018.