Multi-scale Temporal Memory for Clinical Event Time-Series Prediction

Jeong Min Lee and Milos Hauskrecht

University of Pittsburgh, Pittsburgh PA 15260 USA
jlee@cs.pitt.edu, milos@pitt.edu

Abstract. The objective of this work is to develop and study dynamic patient-state models and patient-state representations that are predictive of a wide range of future events in the electronic health records (EHRs). One challenge to overcome when building predictive EHRs representations is the complexity of multivariate clinical event time-series and their short and long-term dependencies. We address this challenge by proposing a new neural memory module called Multi-scale Temporal Memory (MTM) linking events in a distant past with the current prediction time. Through a novel mechanism implemented in MTM, information about previous events on different time-scales is compiled and read on-the-fly for prediction through memory contents. We demonstrate the efficacy of MTM by combining it with different patient state summarization methods that cover different temporal aspects of patient states. We show that the combined approach is 4.6% more accurate than the best result among the baseline approaches and it is 16% more accurate than prediction solely through hidden states of LSTM.

Keywords: Electronic Health Records (EHRs), Clinical Event Time-series Prediction, Neural Network, Sequence Prediction

1 Introduction

Electronic health records (EHRs) are longitudinal collections of clinical information that cover many aspects of patient care in hospitals. It consists of patient demographics, records of the administration of medication, past procedures, lab test results, various physiological signals, and other significant events related to patient care. The EHRs and events recorded therein can be used for a variety of purposes, such as prediction of adverse events [25] and mortality risk scores [29], detection of deviations in care [8, 9], automatic diagnosis [21, 24], lab value estimation [18–20], or intelligent retrieval of similar patient cases from the database of past patients [28].

The objective of this work is to develop and study dynamic patient-state models and patient-state representations that are predictive of a wide range of future events in the electronic health records. Such representations can characterize well the patient state for many different problems mentioned above. Defining good predictive representation of EHRs is a challenging problem due to the inherent complexities of the EHR-based multivariate event time-series. In general, EHRs

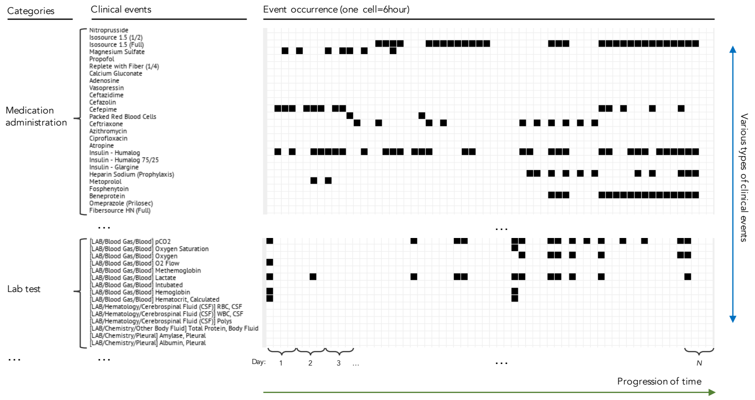

Fig. 1. A part of a patient's record in real-world EHRs (MIMIC-3 database) represented as a sequence of multi-hot vectors. Each vector indicates occurrence or non-occurrence of an event during a segmented time-window (e.g., 6 hours).

can consist of several thousands of clinical events corresponding to different types of medication, lab tests, medical procedures, physiological signals, etc. For example, MIMIC-3 [12], a widely used publicly available ICU Database, records more than 30,000 different types of clinical events. However, many clinical events are sparse and infrequent. Briefly, the average number of medication administration events per patient per admission is 10.1, lab test events 7.3, and procedures 1.5. To deal with the challenges of high-dimensionality and sparsity, deep learning based approaches have shown promising results in modeling EHRs-derived data and sequences. Two major deep learning approaches have been studied: latent-space embedding models [2, 4, 23] and neural temporal models based on RNNs and LSTMs [3, 5, 7, 15, 29].

One challenging issue related to predictive EHRs representations that have not been adequately addressed is how to properly model temporal dependencies among many different clinical events. More specifically, individual event-time-series in EHRs may have a different temporal dependency with respect to precursor events. Briefly, some events may strongly dependent on recently occurred events. For example, an administration of phenylephrine depends on the occurrence of hypotension (low blood pressure state) in connection with recent intubation. Lee and Hauskrecht [15] show that modeling such short-term dependency can improve the predictability of multivariate future events. However, other events may depend on more distant events. For example, valve replacement surgery in the distant past may impact the necessity of warfarin treatment. While neural temporal models (RNN or LSTM) can in principle model these long-range dependencies, the recurrent computations can easily dilute and attenuate such information in the hidden state [26]. In this work, we address the problem of modeling long-term dependencies in multivariate clinical event time-series by proposing a new type of information channel linking events in a distant past with the current prediction time. Through a novel mechanism called Multi-scale Temporal Memory (MTM), information about previous events on different time-scales is compiled and read on-the-fly for prediction through memory contents. The main benefit of this approach is that it is a modular and predictive signal from this module that can be combined with predictive signals from other patient state summarization modules.

We demonstrate the efficacy of MTM by combining it with different patient state summarization methods that cover different temporal aspects of patient states, including recent context module [15], recurrent temporal mechanism [16], and hidden states of LSTM [10]. We test the proposed approach on real-world clinical event time-series. We compare predictive performance (i.e., AUPRC) of the proposed combined approach with baseline models. We demonstrate that the combined approach is 4.6% more accurate than the best among the baselines and it is 16% more accurate than prediction solely through hidden states of LSTM.

2 Background

In this section, we introduce the EHR-based multivariate time-series and the prediction problem. Then, we review clinical event time-series based on neural temporal models.

2.1 Multivariate Clinical Event Time-Series

A patient's EHR is defined by a sequence of time-stamped clinical events $U = \{u_j\}_j$, where each event $u_j = (e_j, t_j)$ consists of a pair of type of the event $e_j \in E$ and timing of the event $t_j \in \mathbb{R}_{\geq 0}$. E is a set of all types of clinical events. As events in EHRs occur in continuous time, t_j is non-negative real value. One way to model the event time-series on real-valued continuous-time is by using point processes [27] such as a Poisson process or a Hawkes process [14]. However, point processes-based approaches are hard to optimize directly, and existing works for clinical event time-series [17, 22] explore multivariate event time-series with a relatively small number of events. Due to this limitation, multivariate event time-series are often converted to discrete-time event time-series. By sweeping the original time-series with a fixed-sized time window (e.g., 6 hours), the time-series is segmented to a sequence of non-overlapping bins, where each bin represents events occur during the time-window. Then, events occurred or non-occurred during a time window are represented as a binary multi-hot vector $\mathbf{y} \in \{0, 1\}^{|E|}$. With this discretization method, a patient's records in EHRs are represented as a sequence of the multi-hot vectors $\mathbf{y}_1, \cdots, \mathbf{y}_t$. Figure 1 shows an exemplar multi-hot vector representation of a patient's record.

The prediction problem we want to tackle can be then defined to predict the occurrence and non-occurrence of a wide range of EHR-related events in the future time step $\mathbf{y}_{t+1}$ given a sequence of patient history $\mathbf{y}_1, \cdots, \mathbf{y}_t$.

2.2 Neural Temporal Models for Clinical Event Time-Series

With the benefits of the flexible end-to-end training and combined feature representation learning capabilities, models based on neural architectures have been successfully adopted to various time-series modeling tasks. In the following, we summarize the approaches to clinical event time-series modeling and prediction.

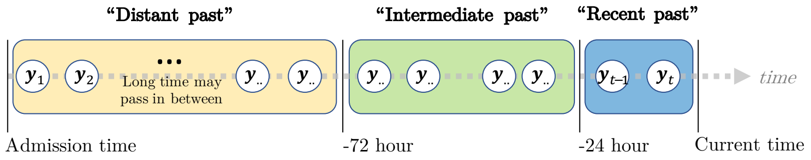

Fig. 2. MTM summarizes past history with multiple temporal scales

Word-to-Vector Models (Word2Vec). The Word2Vec (e.g., CBOW, Skip-gram) [23] learns low-dimensional embeddings of words and documents in NLP. Continuous Bag-of-Word (CBOW) [23] predicts the probability distributions of a center (target) word given the word's neighborhood (context) words. Skip-gram [23] is similar to Word2Vec, but the context and the target are switched around. For clinical tasks, Word2Vec models have been adopted to process a sequence of clinical events instead of words. More specifically, for the CBOW-based approach, recent events in a fixed-size recent history window (e.g., 48 hours) are set as the context and an event that occurs shortly after the history window is set as the target. Word2Vec models have been successfully applied to predict e.g. hospital visits [4]. One drawback of the Word2Vec models is that they cannot fully model the sequential information, as they treat the events in the past equally when pooling (summing or averaging) past event embeddings. Besides, the size of the neighborhood (context) window is limited to a certain number of events (e.g., 20 or 40). Hence, those events that occur outside of the window cannot be used for modeling.

RNN and LSTM based Approaches. The sequential models based on RNN and LSTM [10] resolve the problems by summarizing the information from each past step via hidden states. The hidden states correspond to a real-valued (latent) representation of patient states. RNN and LSTM have been successfully applied to many clinical event predictions such as medication prescriptions [1, 3], heart failure onset [6], and ICU mortality risk [29].

One advantage of RNNs is that it can model all events in the entire sequence without a length-span limit, unlike Word2Vec. However, RNN and LSTM models may encounter problems when modeling long sequences. Briefly, the loss (training objective function) is computed at the end of each sequence and the signal is passed to parameters at each time step via Back Propagation Through Time (BPTT). For RNN and LSTM, the length of the sequence is n. A long sequence (large n) can hinder the propagation of the loss signal to parameters, negatively affecting their training [11]. Our proposed work tackles this challenge by creating a direct path of length 1 connecting the current time step with a predictive event that occurred in the distant past.

3 Methodology

In this work, we propose Multi-scale Temporal Memory (MTM), a new neural temporal based model that summarizes a clinical event history and generates a predictive signal for occurrence and non-occurrence of future multivariate clinical events.

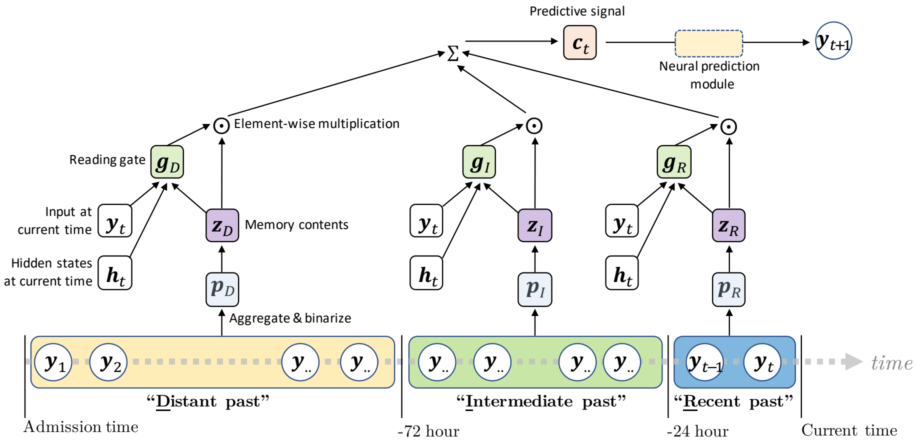

Fig. 3. Overview MTM's architecture: Given a sequence of multivariate patient state history $\mathbf{y}_1, \dots, \mathbf{y}_t$, we **(1)** aggregate and binarize past history by each time-scale $\mathbf{p}_*, * \in \{D, I, R\}$, **(2)** compose memory contents $\mathbf{z}_*$, **(3)** compute reading gate $\mathbf{g}_*$, **(4)** read memory contents referring reading gate and merge contents of multi-scale temporal memory, and **(5)** make a predictive signal $\mathbf{c}_t$ for neural prediction module.

3.1 Multi-Scale Temporal Memory

MTM summarizes patient history using multiple information channels where each channel covers the history in different temporal scales. We hypothesize that information on past event occurrences on different time range may have different importance for predicting future event occurrence. To process patient history on multiple time scales, MTM segments the patient history into three folds as shown in Figure 2: distant past (e.g., from the beginning of admission to 72 hours before current time), recent past (e.g., within 24 hours from current time), and "intermediate past" (time range between boundaries of distant past and recent past). Contents of the memory are composed based on the types of events that occurred in each segmented window. Then, considering factors about current patient states, the model reads contents of the multi-scale memory and generates a predictive signal that will be combined with other neural temporal mechanisms that cover different aspects of clinical event time-series to generate a final prediction for next multivariate events. In the following, we describe MTM in detail and the neural framework for the next multivariate events prediction.

Composing memory contents. Given a segmented patient history (depicted in Figure 2) on multiple time-scales, we compose memory contents for each time-scale with the following steps: (1) We aggregate patient states vectors $\{\mathbf{y}_i\}_i$ of each temporal segment $* \in \{D, I, R\}$ into a single multivariate vector $\mathbf{p}_*$ through binarization. $\{D, I, R\}$ denote distant, intermediate, and recent pasts respectively. (2) We compose contents of the memory $\mathbf{z}_* \in \mathbb{R}^{|E|}$ through linear projection followed by non-linear activation:

$$\mathbf{z}_* = \tanh\left(\mathbf{W}_* \mathbf{p}_* + \mathbf{b}_*\right) \tag{1}$$

where $\mathbf{W}_* \in \mathbb{R}^{|E|\times|E|}$ and $\mathbf{b}_* \in \mathbb{R}^{|E|}$ are trainable parameters for each time-scale. Through linear projection with $\mathbf{W}_*$, we extract information about the events that occurred in a specific temporal segment.

Reading memory contents. To comprehensively determine the amount of memory contents to be read for each prediction task (multivariate target events), MTM computes reading gates $\mathbf{g}_* \in \mathbb{R}^{|E|}$ considering three factors: (1) current patient state reflected on input $\mathbf{y}_t$, (2) recent dynamics of patient state reflected on hidden states $\mathbf{h}_t$ from LSTM, and the contents of the memory itself $\mathbf{z}_*$.

$$\mathbf{g}_* = \sigma(\mathbf{W}_h\mathbf{h}_t + \mathbf{W}_y\mathbf{y}_t + \tilde{\mathbf{W}}_*\mathbf{z}_*) \tag{2}$$

where σ denotes logistic sigmoid activation function and $\mathbf{W}_h \in \mathbb{R}^{|E|\times r}, \mathbf{W}_y \in \mathbb{R}^{|E|\times|E|}, \tilde{\mathbf{W}}_* \in \mathbb{R}^{|E|\times r}$ are parameters to learn and r is dimension of hidden state. The predictive signal $\mathbf{c}_t \in \mathbb{R}^{|E|}$ is computed as a linear combination of reading gates and memory contents for each temporal scale:

$$\mathbf{c}_t = \mathbf{g}_D \odot \mathbf{z}_D + \mathbf{g}_I \odot \mathbf{z}_I + \mathbf{g}_R \odot \mathbf{z}_R \tag{3}$$

where $\odot$ is element-wise multiplication.

3.2 Neural-based Prediction Framework

We combine the predictive signal from MTM with additional patient history summarization methods that cover different temporal aspects of patient states. We use recent-context module [15], recurrent temporal mechanism [16] and hidden states of LSTM. Briefly the recent-context module projects current time-step input $\mathbf{y}_t$ to a target event space with a learnable parameters $\mathbf{W}_r \in \mathbb{R}^{|E|\times|E|}$ and $\mathbf{b}_r$ to get the "recent bias" term $\mathbf{b}_\kappa$:

$$\mathbf{b}_\kappa = \mathbf{W}_r\mathbf{y}_t + \mathbf{b}_r \tag{4}$$

The recurrent temporal mechanism captures information about periodic (repeated) events using a special recurrent mechanism based on probability distributions of inter-event gaps. It outputs two target event-specific periodicity-based predictive signals that use different sources of periodic information: $\boldsymbol{\alpha}^e \in \mathbb{R}$ signal is based on an interval of current patient's event time-series and $\boldsymbol{\beta}^e \in \mathbb{R}$ signal is compiled from a pool of past patient data in training set. Details of the signal generation processes can be found in [16]. We also use LSTM to derive dynamics of patient state through hidden state. To compute hidden state, we first project input $\mathbf{y}_t$ to low-dimensional space with embedding matrix: $\mathbf{W}_{emb} \in \mathbb{R}^{d\times|E|}$: $\mathbf{x}_t = \mathbf{W}_{emb}\mathbf{y}_t$. Based on previous time step's hidden state $\mathbf{h}_{t-1}$ and $\mathbf{x}_t$, we compute new hidden state $\mathbf{h}_t \in \mathbb{R}^r$:

$$\mathbf{h}_t = \mathrm{LSTM}(\mathbf{h}_{t-1}, \mathbf{x}_t) \tag{5}$$

Given predictive signals $\{\boldsymbol{\alpha}^e, \boldsymbol{\beta}^e, \mathbf{h}_t, \mathbf{c}_t, \mathbf{b}_\kappa\}$, we first combine periodicity-based signals for each target event type with hidden state through concatenation:

$$\boldsymbol{\gamma}^e = [\mathbf{h}_t; \boldsymbol{\alpha}^e; \boldsymbol{\beta}^e] \tag{6}$$

Then, we project $\boldsymbol{\gamma}^e$ to a scalar $\lambda^e \in \mathbb{R}$ through $\mathbf{w}_e \in \mathbb{R}^{1\times r+2}$ and $b_e \in \mathbb{R}$. We apply the same procedure to all events $e \in E$ and concatenate all λ^e:

$$\lambda^e = \mathbf{w}_e\boldsymbol{\gamma}^e + b_e \quad \boldsymbol{\lambda} = [\lambda^1; \dots ; \lambda^{|E|}] \tag{7}$$

Final prediction for next multivariate event is computed as follows:

$$\hat{\mathbf{y}}_{t+1} = \sigma(\boldsymbol{\lambda} + \mathbf{b}_\kappa + \mathbf{c}_t) \tag{8}$$

We use the binary cross-entropy to compute loss $\mathcal{L}$ and parameters of the model are learned through a stochastic gradient descent optimization algorithm (Adam) [13].

$$\mathcal{L} = \sum_t -[\mathbf{y}_t \cdot \log \hat{\mathbf{y}}_t + (\mathbf{1} - \mathbf{y}_t) \cdot \log(\mathbf{1} - \hat{\mathbf{y}}_t)] \tag{9}$$

4 Experiments

In this section, we evaluate our approach on MIMIC-3, an ICU EHRs dataset.

4.1 Experiment Setup

Clinical Data. We extract 5137 EHRs of patients from MIMIC-3 database using the following criteria: (1) adult patient, (2) length of stay is between 48 and 480 hours, (3) data are recorded in Meta Vision system, one of the two systems used to create MIMIC-3 database. We randomly split 5137 patients into train and test sets using 8:2 ratio.

Then, multivariate event time-series are generated by segmenting all sequences with a time-window (W=1). As mentioned in Section 2.1, at each i-th window we obtain multi-hot vector $\mathbf{y}_i \in \{0, 1\}^{|E|}$ by aggregating and making binary all events occur within the time range of the window. For the types of clinical events (E), we use events in the categories of medication administration, lab results, procedure, and physiological results. Among all types of events in the first three categories, we filter out those events observed in less than 500 different patients. For physiological events, we select 16 important event types with the help of a critical care physician. To this end, we get 63 medication events, 41 procedure events, 155 lab test events, and 16 physiological signal events ($|E|$=275).

Baseline Methods. We compare our method (HS-RC-PP-MTM) with the following set of baseline models predicting a wide range of future events $\mathbf{y}_{t+1}$:

- **Logistic Regression with Recent Context (RC)** uses the current events. It amounts to use the recent bias term in Equation (4): $\hat{\mathbf{y}}_{t+1} = \sigma(\mathbf{W}_p\mathbf{y}_t + \mathbf{b}_p)$
- **Hidden States from LSTM (HS)** uses hidden states of LSTM in Equation (5) with linear projection and sigmoid activation: $\hat{\mathbf{y}}_{t+1} = \sigma(\mathbf{W}_q\mathbf{h}_t + \mathbf{b}_q)$
- **HS + Recent Context (HS-RC)** [15] uses hidden states of LSTM with the recent bias term $\mathbf{b}_\kappa$ in Equation (4): $\hat{\mathbf{y}}_{t+1} = \sigma(\mathbf{W}_r\mathbf{y}_t + \mathbf{b}_r + \mathbf{b}_\kappa)$
- **HS + RC + Periodicity Predictor (HS-RC-PP)** [16] uses combination of hidden states of LSTM, the recent bias term $\mathbf{b}_\kappa$ and periodicity signal $\boldsymbol{\alpha}, \boldsymbol{\beta}$ from [16]. It computes prediction with $\boldsymbol{\lambda}$ in Equation (7): $\hat{\mathbf{y}}_{t+1} = \sigma(\boldsymbol{\lambda} + \mathbf{b}_\kappa)$

Models	All-events	Medication	Lab test	Physio signal	Procedure
RC	18.26	35.52	4.11	45.23	34.67
HS	26.30	41.16	12.03	81.49	35.90
HS-RC	26.50	41.66	12.07	81.61	36.25
HS-RC-PP	26.76	42.82	12.04	81.70	36.29
HS-RC-PP-MTM	28.00	43.84	13.80	81.84	36.35

Table 1. Prediction results (AUPRC) for all events and each event category

Evaluation Metrics. We evaluate the quality of predictions by calculating the area under the precision-recall curve (AUPRC). The reported AUPRC values (for the different methods) are averaged over all target events.

Implementation Details. For the experiments, we use embedding size $d = 64$, a fixed learning rate=0.005 and minibatch size=128. The size of the hidden state r is determined by the internal cross-validation from $(128, 256, 512)$. To prevent over-fitting, L_2 weight decay regularization is applied to all models and the weight is determined by the internal cross-validation.

4.2 Results

The second column of Table 1 shows the overall experiment results for predicting all types of events. The proposed model (HS-RC-PP-MTM) outperforms all baselines. Particularly, it outperforms HS-RC-PP by 4+%. With this, we can observe the benefit of multi-scale memory capturing dependencies that are not covered by other patient history summarization methods, including LSTM.

We further analyze the experiment results by dividing them into 4 event categories. As shown in Table 1 (column 3-6), we observe the performance gain of MTM is higher for medication and lab test events. Notably, lab tests are the hardest events to predict compared to other categories, 14+% performance gain from MTM for lab test prediction clearly shows its effectiveness.

We also experiment with two additional window sizes (W=6,12). As Table 2 shows, larger segmentation window increases overall predictability. This is expected as larger window size results in the multivariate vector $\mathbf{y}_i$ with more event occurrences and it increases prior probability which directly affects the AUPRC score. Especially, we observe a pattern that the gap between HS-RC-PP-MTM and HS-RC-PP is decreasing as the window size is increased. An implication of this observation is that, for longer sequences (event time-series generated from W=1 based window-segmentation), MTM brings more value than it does for shorter sequences (e.g., W=6,12).

To validate learned weight matrices for multi-scale memory contents ($W_*, * \in \{D, I, R\}$ in Equation (1)), we extract the top 3 events for an exemplar target event **extubation** in Table 3. We can see that MTM properly learns and gives higher weights to the intubation event, PEEP (setting of the mechanical ventilation), and fentanyl, analgesics used during mechanical ventilation. Additional examples are compiled in Table 4.

Models	W=1	W=6	W=12
HS-RC-PP	26.76	36.68	40.07
HS-RC-PP-MTM	28.00 (4.6 +%)	37.28 (1.6 +%)	40.34 (0.6 +%)

Table 2. Prediction results by varying time-series segmentation window settings

Distant past ($*$=D)	Intermediate past ($*$=I)	Recent past ($*$=R)
(Med) Potassium Chloride	(Proc) PEEP	(Proc) PEEP
(Med) KCL	(Med) Fentanyl	(Physio) Inspired O2 Fraction
(Proc) Intubation	(Proc) Intubation	(Med) Fentanyl

Table 3. Top 3 past events predictive of **extubation**, based on the value from learned memory content parameter W_* for each temporal range in Equation (1).

5 Conclusion

We proposed a novel mechanism called Multi-scale Temporal Memory (MTM) to model long-term dependencies in EHR-derived clinical event time-series. With MTM, information about past events on different time-scales is compiled and read on-the-fly for prediction through memory contents. We demonstrate the efficacy of MTM by combining it with different patient state summarization methods that cover different temporal aspects of patient states. We show that the combined approach is 4.6% more accurate than the baseline approaches and it is 16% more accurate than the prediction based on the popular LSTM summarization approach.

In the future we plan to study ways of relaxing hard segmentations of past history. That is, we plan to automatically identify the memory content and the timing information for past events that are important for predicting the next events. One possible direction is to design an attention mechanism capable of aggregating event history via a specialized kernel that considers both (a) the type of target and context events and (b) timing of events.

Acknowledgement. The work in this paper was supported by NIH grant R01GM088224. The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

1. Bajor, J.M., Lasko, T.A.: Predicting medications from diagnostic codes with recurrent neural networks. In: ICLR (2017)
2. Bengio, Y., et al.: A neural probabilistic language model. Journal of machine learning research (Feb), 1137–1155 (2003)
3. Choi, E., et al.: Doctor ai: Predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference. pp. 301–318 (2016)
4. Choi, E., et al.: Multi-layer representation learning for medical concepts. In: The 22nd International Conference on Knowledge Discovery and Data Mining (2016)

5. Choi, E., et al.: Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In: Neural Information Processing Systems (2016)
6. Choi, E., et al.: Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association (2017)
7. Esteban, C., et al.: Predicting clinical events by combining static and dynamic information using RNN. In: Intl. Conf. on Healthcare Informatics (ICHI) (2016)
8. Hauskrecht, M., et al.: Outlier detection for patient monitoring and alerting. Journal of biomedical informatics **46**(1), 47–55 (2013)
9. Hauskrecht, M., et al.: Outlier-based detection of unusual patient-management actions: an icu study. Journal of biomedical informatics **64**, 211–221 (2016)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comp. (1997)
11. Hochreiter, S., et al.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies (2001)
12. Johnson, A.E., et al.: MIMIC-III, a freely accessible critical care database. Scientific data **3**, 160035 (2016)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
14. Laub, P.J., et al.: Hawkes processes. arXiv preprint arXiv:1507.02822 (2015)
15. Lee, J.M., Hauskrecht, M.: Recent-context-aware LSTM-based Clinical Time-Series Prediction. In: In Proceedings of AI in Medicine Europe (AIME) (2019)
16. Lee, J.M., Hauskrecht, M.: Clinical Event Time-series Modeling with Periodic Events. In: The Thirty-Third International Flairs Conference. AAAI (2020)
17. Liu, S., Hauskrecht, M.: Nonparametric regressive point processes based on conditional gaussian processes. In: Neural Information Processing Systems (2019)
18. Liu, Z., Hauskrecht, M.: Clinical time series prediction: Toward a hierarchical dynamical system framework. Artificial Intelligence in Medicine **65**(1), 5–18 (2015)
19. Liu, Z., Hauskrecht, M.: A regularized linear dynamical system framework for multivariate time series analysis. In: Twenty-Ninth AAAI Conference on Artificial Intelligence. pp. 1798–1804 (2015)
20. Liu, Z., Wu, L., Hauskrecht, M.: Modeling clinical time series using gaussian process sequences. In: Proceedings of the 2013 SIAM International Conference on Data Mining. pp. 623–631. SIAM (2013)
21. Malakouti, S., Hauskrecht, M.: Hierarchical adaptive multi-task learning framework for patient diagnoses and diagnostic category classification. In: International Conference on Bioinformatics and Biomedicine (BIBM) (2019)
22. Mei, H., Eisner, J.M.: The neural hawkes process: A neurally self-modulating multivariate point process. In: Neural Information Processing Systems (2017)
23. Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. In: neural information processing systems (2013)
24. Miotto, R., et al.: Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports (2016)
25. Nemati, S., et al.: An interpretable machine learning model for accurate prediction of sepsis in the ICU. Critical care medicine **46**(4), 547–553 (2018)
26. Pascanu, R., et al.: On the difficulty of training recurrent neural networks. In: International conference on machine learning. pp. 1310–1318 (2013)
27. Rasmussen, J.G.: Lecture notes: Temporal point processes and the conditional intensity function. arXiv preprint arXiv:1806.00221 (2018)
28. Wang, F., et al.: Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. Statistical Analysis and Data Mining: The ASA Data Science Journal **5**(1), 54–69 (2012)
29. Yu, K., et al.: Monitoring icu mortality risk with a long short-term memory recurrent neural network. In: Pac Symp Biocomput. World Scientific (2020)

A Examples of Top Past Events Predictive of Target Events

Table 4 shows top past events predictive of target events for the different temporal ranges (Distant, Intermediate, and Recent past) as identified by our methods. For example, the top predictive events for amiodarone (treats irregular heartbeat such as tachycardia) include metoprolol and diltiazem. Both of these are used to treat high blood pressure and heart issues. Similarly, past events predictive of diltiazem and labetalol (medications treating high blood pressure) include clinical events that are related to high blood pressure and heart function: digoxin, metoprolol, hydralazine, and nicardipine. Finally, the top past events predicting vasopressin (a medication treating a low blood pressure) include norepinephrine and phenylephrine that are also used to treat low blood pressure.

Distant past ($*$=D)	**Intermediate past**($*$=I)	**Recent past**($*$=R)
Target: (Med) Amiodarone		
(Med) Amiodarone	(Med) Amiodarone	(Med) Amiodarone
(Med) Diltiazem	(Med) Diltiazem	(Med) Metoprolol
(Lab) Urea Nitrogen, Urine	(Lab) Thyroid Stimulating Hormone	(Med) Diltiazem
Target: (Med) Diltiazem		
(Med) Diltiazem	(Med) Diltiazem	(Med) Diltiazem
(Lab) Digoxin	(Med) Metoprolol	(Med) Metoprolol
(Physio) Inspired O2 Fraction	(Med) Amiodarone	(Proc) EKG
Target: (Med) Labetalol		
(Med) Labetalol	(Med) Labetalol	(Med) Labetalol
(Med) Hydralazine	(Med) Hydralazine	(Med) Hydralazine
(Med) Nicardipine	(Med) Metoprolol	(Med) Haloperidol
Target: (Med) Vasopressin		
(Med) Vasopressin	(Med) Vasopressin	(Med) Vasopressin
(Proc) Ultrasound	(Med) Norepinephrine	(Med) Norepinephrine
(Med) Packed Red Blood Cells	(Med) Phenylephrine	(Med) Phenylephrine

Table 4. Top 3 preceding events for example target events based on the value from learned memory content parameter W_* for each temporal range in Equation (1).