Health-ATM: A Deep Architecture for Multifaceted Patient Health Record Representation and Risk Prediction

Tengfei Ma*† Cao Xiao*‡ Fei Wang§

Abstract

Leveraging massive electronic health records (EHR) brings tremendous promises to advance clinical and precision medicine informatics research. However, it is very challenging to directly work with multifaceted patient information encoded in their EHR data. Deriving effective representations of patient EHRs is a crucial step to bridge raw EHR information and the endpoint analytical tasks, such as risk prediction or disease subtyping. In this paper, we propose Health-ATM, a novel and integrated deep architecture to uncover patients' comprehensive health information from their noisy, longitudinal, heterogeneous and irregular EHR data. Health-ATM extracts comprehensive multifaceted patient information patterns with attentive and time-aware modulars (ATM) and a hybrid network structure composed of both Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). The learned features are finally fed into a prediction layer to conduct the risk prediction task. We evaluated the Health-ATM on both artificial and real world EHR corpus and demonstrated its promising utility and efficacy on representation learning and disease onset predictions.

1 Introduction

The broad adoption of electronic health record (EHR) systems has provided clinicians and researchers unprecedented resource and opportunity for conducting health informatics research. However, there are many challenges of working directly with raw EHR, such as sparsity, longitudinality, irregularity, etc. Therefore, deriving effective and robust representations for EHR is a critical step for bridging EHR and endpoint analytical task, such as risk prediction or disease subtyping [18].

From the data mining perspective, EHR representation learning is essentially the feature engineering process. Recently, there are quite a few different techniques being proposed for conducting such process. To name a few, Wang *et al.* [30] proposed a convolutional matrix factorization approach to detect temporal patterns from patient EHR corpus. Zhou *et al.* [35] aggregated raw EHR events into super features and used them to represent patients' health records, so that the dimensionality of feature space got reduced and the representations became much denser. More recently, researchers have also started to apply deep learning methods to extract more abstract EHR representations. For example, Cheng *et al.* [6] proposed to represent each patient's record as a temporal matrix with time on one dimension and clinical events on the other dimension, and then build a four layer convolutional neural network model for extracting temporal patterns as well as performing risk prediction. In [25], the authors presented a three-layer stack of denoising autoencoders to capture hierarchical regularities and dependencies in the aggregated EHRs and used it to facilitate predictive modeling. In addition, other deep models were also proposed with particular challenges being targeted, e.g. more attention on relevant clinical events and better interpretability [10] or including more focus on temporal challenges in modeling [4, 8].

Despite the initial success of these works, there are still many challenges that the existing research has not addressed. We list some of them as follows.

- **Irregularity and heterogeneity**: Due to the complexity arising from the heterogeneous manifestation and progression of many diseases, even for patients having the same disease, they might have heterogeneous health conditions or comorbidities (i.e. co-occurring conditions), which are encoded in their health records. Such complexity and heterogeneity is hard to represent and would affect learning and prediction.
- **Target-Awareness**: Neural processes of knowledge involve "attention" on relevant information[20, 13]. Although patients' EHR includes various types and huge amounts of clinical events, a large part of them could be irrelevant. Therefore, we follow the principle of neural processing to focus more on the most pertinent sets of clinical events, rather than using all available

*equal contributions
†IBM T.J. Watson Research Center
‡AI for Healthcare, IBM Research
§Weill Cornell Medical School, Cornell University

information. The attention has been introduced in learning health record representation, e.g. the RETAIN model [10]. However, the attention weights in [10] are generated only from the hidden states of the recurrent neural networks without considering the target information. Such strategy consequently weakens the interpretability and effectiveness of the model. In fact the experimental results in [10] also show that the RETAIN model did not gain much performance boost from adopting the attention mechanism.

- **Temporality**: Patients' health conditions evolve over time. The temporality encoded in time stamps of the clinical events reveals important information on impending patient health conditions [8]. In most related deep models, time stamps are either not accounted in the learned representation or not accounted in disease prediction[4] or only used to predict near-term subsequent events [8]. Ignoring modeling time stamps may compromise the prediction performance depending on the nature of events, disease mechanism and other factors.

To address those challenges, we propose an integrated deep architecture called Health-ATM (Attentive Time-aware Model) for patient representation learning and risk predictions. Building upon a novel time-aware convolutional neural network as well as a task-specific target-aware attention mechanism, the proposed model could achieve multifaceted characterization of clinical event patterns (e.g. contextual, temporal, time-aware, structural, and relevant) that occur across multiple hierarchies (single events of different categories, events during the same visit, temporal events scatter over several visits, etc). Consequently the Health-ATM could explore patient health information in a comprehensive way. It is worthwhile to highlight the following contributions of the proposed model.

- **Hybrid CRNN to capture global structure and local features**: To resolve irregularity and heterogeneity requires better feature abstraction from raw EHR. Therefore we propose a hybrid convolutional recurrent neural network for joint feature extraction and temporal summarization. In the CRNN structure, a bidirectional GRU enables the networks to capture global structure via modeling the nonconsecutive event interaction and information decaying of patients' clinical events. Joint modeling of CNNs assumes these temporal interactions are in different abstraction levels and can be extracted by temporal convolution operators. A pooling strategy on these local correlations also extracts invariant regularities.

- **Time-aware convolution**: Time stamps are already demonstrated useful in predicting future events [14]. For example, Choi *et al.* [8] integrated time information into their RNN based models to predict the occurrence and timing of near-term subsequent events. Our work differentiates from previous works since task-wise we focus on prediction of far-future events. While model-wise, we design a new time-aware convolutional layer by integrating time stamps into the original convolutional layer and re-weight more abstract temporal interactions. The weights in convolution would be adjusted according to the event time-stamps, so that the assumption that temporally close events would be more relevant to the prediction target [34] could be captured.

- **Target-aware attention**: Doctors make diagnoses based on clinical evidences that are related to the target diseases or conditions. We mimic this heuristic using a target-aware attention design. This component allows the model to pay more attention on relevant tokens with respect to the disease of interest. Different from the use of attention mechanism in [10], we embed the target events into the same space as the observed events, and use not only hidden vector representation of the observed events but also the embedding vector of the target event to get the attention weights for each token, so that the learned model could benefit from information of target events and become more interpretable. As a result, our model can learn effective task-specific strategies for where to look on. The experimental results also suggest that a target-aware attention model could facilitate learning and improve prediction .

2 Related Work

In this section we briefly review the existing works that are closely related to patient representation learning. The patient representation learning (i.e. patient phenotyping) is a key step to resolve the data challenges in raw EHR and prepare them for downstream data-driven tasks, e.g. predictive modeling [19]. Existing works on EHR representation learning include traditional methods and deep learning approaches.

Traditional methods often take raw clinical events or event groups as features and obtain EHR representation as combinations of clinical events via some optimization procedure. They often focus more on demonstrating that some chronic diseases and their corresponding medication events are predictive, for example, the vector based approach in [32],tensor based

approach in [17], multilinear sparse logistic regression model in [31], and convolutional matrix factorization approach [30] for modeling the temporalities among clinical events.

Recent years, deep models become a preferred approach for EHR representation learning due to their ability to learn complex patterns from data to characterize higher-level correlations among clinical events. Various deep mechanism or embedding techniques have been adopted to generate better EHR representations. For example, word2vec embedding was used to learn low-dimensional representations of medical concepts [12] to generate dense and contextual-embedded features. [6] proposed to represent each patient's EHR as a temporal matrix with time on one dimension and clinical events on the other dimension, and then build a four layer convolutional neural network (CNN) for extracting phenotypes and perform prediction. In [25], the authors presented a three-layer stack of denoising autoencoders to capture hierarchical regularities and dependencies in the aggregated EHRs and used it to facilitate clinical predictive modeling.

To make EHR based disease prediction, sequential prediction approaches such as recurrent neural networks (RNN) or RNN-variants were often applied. For example, [11] developed an RNN based model to predict early onset of heart failure. Later in [10], it has been demonstrated reverse-order RNN performs better at some clinical prediction tasks. In addition, the attention mechanism has been explored to allow the model to focus on influential past visits or events, e.g. the RETAIN model [10]. However, their attention strategy is not target-aware since the attention weights in [10] are generated from only the hidden states of the RNN for the observed clinical events. This strategy weakens the interpretability and effectiveness of the model. Other than the attention mechanism, other deep models were also proposed with particular challenges being targeted, e.g. more focus on temporal challenges in modeling [4, 8], or incorporating hierarchical information inherent to medical ontologies [9]. However, there still lacks an integrated model that can simultaneously and fully represent multifaceted patient information to uncover patients' comprehensive health condition.

3 Method

We are now ready to present the Health-ATM, a deep architecture that provides synergistic representation learning to predict the occurrence of a disease e (or multiple diseases) given initial input represented as sequences of clinical events of Length T: $\boldsymbol{x}_1$, $\boldsymbol{x}_2$,..., $\boldsymbol{x}_T$. Each clinical event $\boldsymbol{x}_i$ is represented by a medical code and associated with a visiting time stamp t_i;

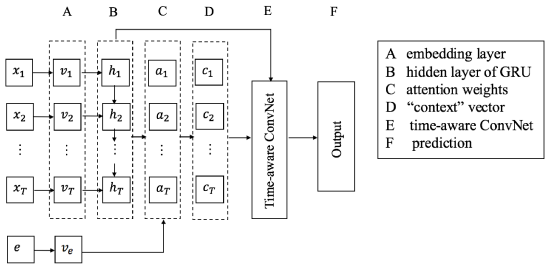

Figure 1: The proposed Health-ATM architecture.

and the target event is generally a diagnosis event which indicates a disease. A visit may contain multiple codes at the same time, but here we regard each code as a single clinical event, because we want to get finer-grained and interpretable attention weights between target and observed event in the following layers. Taking embedded clinical events as input, we extend an RNN model with adding a novel target-aware attention mechanism to take advantage of the target embedding information; and a CNN structure to a time-aware ConvNet to capture the time stamp information. The Health-ATM architecture is illustrated as in Figure. 1. In the following we describe the detailed design of the Health-ATM architecture.

3.1 Basic Layers: Contextual Embedding and RNN

Learning distributed representations or word embeddings has proved particularly useful in various natural language processing tasks, and has also gained initial success in medical concept embeddings [14, 12].

Similar to [14], we processed the EHR training dataset so that diagnosis codes, medication codes, procedure codes are laid out in a temporal order, and each code represents a clinical event. Then using the context window size of 15, and applying the CBOW model of word2vec [24], we were able to project all these medical codes into the same lower dimensional space, where similar or related codes are embedded close to one another.

The output of the contextual embedding layer $\mathbf{v}$ would be directly fed into the RNN layer to further generate temporal patient representations. RNN is a natural tool to model sequences and has received much attention in this field already [11, 4, 8, 10]. Following the procedures in [11], we adopt the GRU-based RNN structure [7] in our model. In addition, we extend the single-directional GRU to bi-directional GRU to capture both previous and future context information.

3.2 Target-Aware Attention Mechanism for Interpretable Patient Representation

Attentive neu-

ral networks have been successfully applied to a wide range of tasks, such as machine translation [3] and reading comprehension [16]. In healthcare, there has also been some works starting to study the benefit from the attention mechanism, e.g. the RETAIN model [10]. However, their strategy of only generating attention weights from the hidden states of the recurrent neural networks weakens the interpretability and effectiveness of the model, hence does not provide much performance gain.

Here for Health-ATM, we introduce an alternative attention mechanism to allow the model to pay more attention on target-relevant clinical events, as well as make the results more interpretable. Intuitively, not all clinical events are relevant to the target disease. When a doctor diagnoses a disease, the doctor may only examine the possibly relevant historical events based on their professional experiences. So "attention" should not be only decided by the historical events, but also impacted by the target. In our problem, the target disease is also represented by a medical code (a diagnosis) which occurs in the EHR training data. Hence after embedding all medical codes in Section 3.1, the target is in the same embedding space as the input events, and we can take advantage of the information from target events.

Then we calculate the attention weights for each position in the input sequence. Particular, given the vector representation of the i^{th} event $\boldsymbol{h}_i$ (i.e. the RNN hidden state at position i), the vector representation of the target disease $\boldsymbol{v}_e$ (i.e. the embedding vector of the target diagnosis event e), the attention weight between $\boldsymbol{v}_e$ and $\boldsymbol{h}_i$ is computed by:

$$a_{i,e} = \frac{\exp(g_{i,e})}{\sum_j \exp(g_{j,e})} \tag{3.1}$$

$$g_{i,e} = F(\boldsymbol{h}_i, \boldsymbol{v}_e) = \tanh(\boldsymbol{h}_i \cdot \boldsymbol{W} \cdot \boldsymbol{v}_e). \tag{3.2}$$

where W is the weight matrix parameter for $g_{i,e}$. The attention weights could indicate the importance of the corresponding clinical events when predicting a special target disease. Then we reformulate the hidden states in previous layers to get new context vectors and use them as input for the next layer: $\boldsymbol{c}_i = a_{i,e}\boldsymbol{h}_i$ for $i = 1, 2..., T$.

3.3 Time-Aware ConvNet for Temporal Information Integration In the EHR data, each clinical event is associated with a time stamp. We develop a time-aware convolutional neural network above the recurrent layers in order to utilize higher-level features as well as to integrate the time stamps to address the two following considerations.

Firstly, to capture those higher-level temporal interactions beyond what RNN can model, we design a hybrid CRNN structure for joint feature extraction and temporal summarization. In the hybrid structure, the previous bidirectional GRU layers enable the networks to capture global structure via modeling the interactions between nonconsecutive events as well as the information decays of patients' earlier clinical events. Joint modeling of CNNs assumes these temporal interactions occur in different abstraction levels and can be extracted by temporal convolution operators.

Secondly, for the convolutional layer, we modify the original CNN by integrating time stamps into the convolution operation and re-weight more abstract temporal interactions. Here the weights are adjusted according to the event time-stamps, since we assume temporally close events would be more likely to be relevant and have more influence on each other.

Structurally speaking, the time-aware ConvNet contains two layers: a multi-channel **time-aware convolutional layer** and a max-pooling layer. As for inputs, the time-aware ConvNet takes multiple channels, including both the hidden states of forward/backward GRUs and the two corresponding context vectors. This allows the model to benefit from different features. For each input channel $\boldsymbol{c}$, different from a general convolutional layer, the weights of the input variables are re-weighted by their time stamps in each convolution computation: $f_i(\boldsymbol{W}) = f(\boldsymbol{W} * (\boldsymbol{c}_{i:i+s-1} \odot \boldsymbol{d}^i_{i:i+s-1}) + \boldsymbol{b})$ where $\boldsymbol{c}_{i:i+s-1}$ is the input; s is the window size; $\boldsymbol{W}$ is the parameter; $*$ is the convolution operation, and $d^i_j = g(t_j)$ is a function of the time stamp t_j, where t_j indicate the days to the end of the sequence. We use a softmax function to normalize the time-aware weights $g(t_j) = \text{softmax}(\lambda \cdot t_j)$, where λ is a parameter. For the activation function f, we use rectification (ReLU). We use multiple filters to generate n filter maps, and then apply a max-pooling layer subsequently to aggregate the representation.

3.4 Prediction Layer with Cross-Entropy Loss The prediction layer takes the output of time-aware ConvNet as input, and outputs a binary label based on particular target disease specification. For the binary classification task here, we adopt a fully connected layer with a sigmoid function over the hidden vectors to generate the outputs. We choose the cross-entropy loss function as the objective function: $C_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i * \ln(\hat{y_i}) + (1 - y_i) * (1 - \ln(\hat{y_i})))$, where $\hat{y}$ is the prediction and y is the real label. We use the Adam algorithm [22] for optimization.

3.5 Regularization To generalize well and avoid overfitting, we employ the L2 regularization over all parameters in the Cross-entropy loss function, including

W_s, U_s, b_s in the GRU layers, W and b in the attention layer and convolution layer. In addition, we use dropout [27] for the hidden layers.

4 Experiments

4.1 Dataset Description We evaluate the quality of representation learning based on the task of disease prediction using two datasets, one is a congestive heart failure (CHF) cohort extracted from a proprietary real world EHR data (SNOW dataset), and the other is a publicly available artificial dataset (EMRbots).

The real-world CHF data (SNOW) is extracted from a warehouse including the records from 319,650 patients over 4 years. The CHF cohort includes patients and matching controls as defined by clinical experts. The criteria for being patients include 1) ICD-9 diagnosis of heart failure appeared in the EHR for two outpatient encounters, indicating consistency in clinical assessment, and 2) At least one medication was prescribed with an associated ICD-9 diagnosis of heart failure. The diagnosis date was defined as its first appearance in the record. These criteria have also been previously validated as part of Geisinger Clinical involvement in a Centers for Medicare and Medicaid Services (CMS) pay-for-performance pilot[26]. For matching controls, a primary care patient was eligible as a control patient if they are not in the case list, and had the same PCP as the case patient. More details could be found in [33]. The artificial EMRbots data is downloaded from online [1]. The data simulation criteria and procedure could be found in [21].

Table 1: Basic statistics of datasets

Dataset	SNOW (12)	SNOW (6)	EMR
# patients	2268	2191	1443
# controls	14526	13335	5287
# visits per patient	19.7	11.4	4.6
# codes per patient	41.0	23.9	4.6
# unique codes	1865	1865	529
# codes per visit	2.08	2.09	1

For the SNOW dataset, we evaluate the model performance in the task of predicting whether a patient could develop CHF within 1) 12-month observation window and 6-month prediction window, and 2) 6-month observation window and 6-month prediction window. For the first scenario, we select patients and controls who have at least 4 events in the observation window. For the second scenario, we select subjects who have at least 3 events in the observation window. For artificial data, we focus on the prediction of diabetes mellitus (target event code: E08). For each patient, we randomly select about 4 controls and we require all patients and controls include as least 4 medical visits. The basic statistics for the datasets are summarized in Table.1.

4.2 Experiment Setting Since the prediction tasks are binary classification problems, we choose the area under the receiver operating characteristic curve (AUC) and the negative log-likelihood as two measures.

We evaluate the models with 5-fold cross-validation strategy and report the average performance. For each iteration we split 10% of the training set into a development set (which is used to determine the hyperparameters) and keep the remaining 90% as the real training set.

The hyperparameters of our Health-ATM model are finally set as follows: 1) we use window size of 15 for word2vec and train medical code vectors of 100 dimensions on each training data. 2) the hidden layer size of GRU is 50. 3) number of filters for CNN is 64 and each filter window size is 5[1]. We use max-pooling layer following the convolution with pooling size (3,1). 4) dropout rate is 0.5. Training is done through Adam [22] at learning rate 0.0002 with shuffled mini-batches (batch size 100).

4.3 Results of Model Performance We choose the following models as our baselines. All models were implemented using Theano 0.8.0[28]. The implementation details are described as follows.

- **RNN**: We implement RNN as a bi-directional GRU. The word embedding sequences are used as input, and a prediction layer via logistic regression is applied over the hidden layer.
- **CRNN**: Before the last layer of logistic regression, we use a conventional CNN over the hidden layer of RNN (bi-directional GRU).
- **Attentive CRNN** A degenerated version of Health-ATM in terms of replacing the time-aware convolution with a standard convolution layer.
- **Time-Aware CRNN** Another degenerated version of Health-ATM in terms of removing the attention mechanism. The input to the ConvNet only consists of two channels: the forward GRU outputs and the backward GRU outputs.

In addition, we also compared our model with state-of-the-art sequence prediction models for healthcare.

[1] i.e. filter shape is (5,50). For artificial data we used a different filter window: 3, and the other hyperparameters are the same as the SNOW data.

- **Logistic Regression** We use the same input as our model, and add L2-regularization with a coefficient 0.01 for logistic regression.

- **RETAIN**: To make fair comparison, we keep the hyperparamters of RETAIN similar to our model: 100-dimension word embedding, 100-dimension GRU layer, minibatch size of 100, dropout for the hidden layer at rate 0.5, L2-regularization for all the parameters. The original RETAIN did not utilize the pre-trained word vectors, but used the one-hot vectors as input and insert an embedding layer instead. In this work, we use the same pre-trained word vectors for RETAIN as the ones used in our own model.

- **CNN-SDM**: Following the description in Cheng et al 2016 [6], we split each patient sequence into 5 sub-frames, and use the Slow Fusion CNN model (SF-CNN) with the following setting for the SNOW data: the filter window is 3 with 105 feature maps; the activation function of the convolutional layer is rectify and the pooling function is mean instead of max; the dropout rate is 0.5; the L2-regularization coefficient is 0.001; and the mini-batch size is 50. For EMRbots data, considering the sequences are generally very short, we set the number of sub-frames as 2 and only use the Late Fusion CNN model (Slow Fusion cannot be applied to 2 sub-frames); the filter window is also changed into 2.

- **CNN-NIPS** This model differentiates from a basic CNN model by using a set of filters with different widths. For comparison, we use our pre-trained word vectors as input and do not fine tune the embedding matrix. The filter shapes used for the convolution layer are set as {2, 3, 4}; other hyper-parameters are same as our Health-ATM model. (The code is from https://github.com/ych133/ConvNet_EHR_Risk_Prediction).

Table 2 and 3 compare the prediction performance of Health-ATM against its reduced models and all baseline methods.

From the results, Health-ATM outperforms all its reduced models and state-of-the-art baselines in terms of the generalization power of prediction tasks (e.g. testing AUC). Among the reduced models of Health-ATM, the RNN is the same as the model in [11]. CRNN performs better than the RNN due to a convolution layer that could capture the higher-level patterns. The attentive CRNN further gains performance advantage over the CRNN thanks to the additional attention mechanism that allows the model to focus more on target-related events. With time-aware design, the model gains more performance improvement compared with the CRNN since the integration of time-stamp emphasizes the temporally close events to be more relevant. The proposed architecture, Health-ATM, since having all the benefits mentioned above, achieves the best testing AUC on all prediction tasks.

What's worth noting is, among the state-of-the-art baselines, RETAIN also employs attention mechanism but does not show obvious improvement over the basic RNN model (sometimes even performs worse). This could also be verified that RETAIN has similar performance as a basic RNN in [10], demonstrating that their strategy could not leverage the power of attention possibly due to the loss of target information. For a fair comparison, our reduced attention model (Attentive CRNN) without time stamps not only performs better than RETAIN, but also consistently outperforms our own degenerated model without attentions. This shows effectiveness of our new target-aware attention strategy.

5 Model Analysis for Congestive Heart Failure Prediction

In this section, we demonstrate the utility of Health-ATM with predicting the diagnosis of congestive heart failure (CHF). The CHF is a disease that is hard to detect or has heterogeneous manifestation. With CHF, the heart is weakened and would cause heart pump blood at an abnormal rate and thus induces low cardiac output and causes problems of blood congestion backing up into the lungs and tissues. Depending on the progression stage that CHF is detected, on average 50% of patients will have an average life expectancy of five years. For those diagnosed at their advanced stages, up to 90% will pass away within one year. Therefore, an early and accurate risk warning would improve the quality of care and prevention greatly.

Paying attention to patient-specific CHF-relevant events is critical for accurate detection. In addition, taking the inter-event time duration into consideration could also improve the prediction. From the results discussed in Table 2, we already see the performance gain in prediction due to introduce the target-aware attention and the time-sensitive design. In the following we will demonstrate 1) the interpretability of attention and 2) the importance of time-awareness in the discrimination of CHF patients from controls.

5.1 Interpretability of Target-aware Attention

Fig. 2 is the visualization of the attention contributions from various clinical events in a high-CHF-risk subject's record (risk = 0.5279, we use the sigmoid output of the last layer as the risk value instead of a binary label). We only highlight variables with top attention weights.

	Model	AUC (SNOW 12-mo)	AUC (SNOW 6-mo)	AUC (EMRbots)
Reduced	RNN (Choi et al 2016 [11])	0.6702±0.0077	0.66636±0.0078	0.7769±0.0133
	CRNN	0.6967± 0.0136	0.6642 ± 0.0035	0.7909 ± 0.0144
	Attentive CRNN	0.7096± 0.0134	0.6700 ± 0.0073	0.7913± 0.0063
	Time-Aware CRNN	0.7243± 0.0261	0.6810± 0.0172	0.9363±0.9363
Proposed	**Health-ATM**	**0.7286± 0.0181**	**0.6900± 0.0042**	**0.9434±0.0149**
Baselines	Logistic Regression	0.6496±0.0150	0.6077±0.0047	0.7897±0.0097
	RETAIN (Choi et al 2016 [10])	0.6683±0.0084	0.6572 ± 0.0150	0.7671 ± 0.0125
	CNN NIPS (Che et al 2016 [5])	0.6722±0.0135	0.6511 ± 0.0130	0.7794±0.0092
	CNN SDM (Cheng et al 2016 [6])	0.6902 ±0.0117	0.6576 ±0.0069	0.7884 ± 0.0113

Table 2: Comparison with baseline models in terms of AUC.

	Model	neglog (SNOW 12-mo)	neglog (SNOW 6-mo)	neglog (EMRbots)
Reduced	RNN [11]	0.3799±0.0153	0.3950 ±0.0144	0.3613 ±0.4281
	CRNN	0.3595± 0.0161	0.3885±0.0115	0.3537 ± 0.0172
	Attentive CRNN	0.3546±0.0114	0.3846±0.0121	0.3465± 0.0223
	Time-aware CRNN	0.3474± 0.0147	**0.3762±0.0076**	0.2092±0.0159
Proposed	**Health-ATM**	**0.3464± 0.0127**	0.3768±0.0115	**0.1979± 0.0271**
Baselines	Logistic Regression	0.4247±0.0133	0.4381±0.0142	0.3976±0.0194
	RETAIN [10]	0.3779±0.0106	0.3858±0.0165	0.3703 ± 0.0259
	CNN NIPS [5]	0.3715±0.1659	0.4023±0.0086	0.3576±0.0124
	CNN SDM [6]	0.3661±0.0097	0.3924 ± 0.0135	0.4025 ± 0.0129

Table 3: Comparison with baseline models in terms of negative log-likelihood.

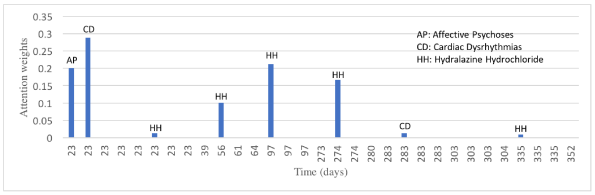

Figure 2: Visualizing attention weights for a patient with heart failure risk 0.5279

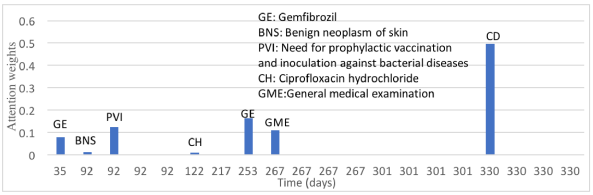

Figure 3: Visualizing attention weights for a patient with heart failure risk 0.1451

From the figure, we observe that the attention was initially on the subject's another condition: affective psychoses, then more attention is shifted to cardiac dysrhythmia, a CHF-related symptom of irregular heart rhythm or abnormal heart pacing. For the rest of this record, Hydralazine Hydrochloride, a CHF medication, contributes most to the prediction.

As a contrast, Fig. 3 helps visualize the attention weights for a lower-CHF-risk subject (risk = 0.1451). At the beginning, the patient took Gemfibrozil, a cholesterol medication. Then the patient's condition on skin problems, benign neoplasm, and bacterial infection got attention. Next, Ciprofloxacin hydrochloride, an antibiotic medication was used to treat bacterial infections. And the patient refilled Gemfibrozil and did some general medical exam. For some time later the patient experienced symptoms of cardiac dysrhythmia. However, since cardiac dysrhythmia alone is not a discriminating indicator of CHF, the patient is not predicted as high CHF risk.

5.2 Analysis of Time-awareness Even with the sophisticated attention mechanism, without considering time stamps of clinical events, we could still miss predicting some patients' impending CHF onset. The big performance gain with time-aware component also indicates the importance of temporal information. To better understand these scenarios, we analyze 22 cases whose impending CHF onset risk is predicted high but would be predicted as low risk when time information is removed from the full model. These patients were actually diagnosed CHF within the 6 months beyond

the observation period. This is to say, without modeling time stamps, these patients' disease would not be detected timely, and could miss their best timing for treatment.

Examining the 22 cases indicates that the visit time duration and the frequency change of some clinical events may explain the performance difference in prediction tasks. The discoveries are also backed by literature evidences. Here are two selected cases.

Case I: Increased frequency of hospital revisits during the year-long consumption of Simvastatin Simvastatin is a statin that treats high cholesterol to reduce risks of heart attack, stroke, and blood vessel problems [29]. One time-sensitive example in the dataset is: a patient has been taking Simvastatin for nearly a year. Over the year, the patient visited the clinics more and more often and was diagnosed with all kinds of complications including joint disorder, disorder of breast, disorders of lipid metabolism, and hypertension.

Without modeling time-stamps, the sequence of events is labeled with low CHF-risk probably due to all the complications the patient suffered are considered normal reactions due to Simvastatin inhibit the mevalonate pathway involved in CoQ-10 synthesis and cause such disorders [23]. However, given the increasing frequency in hospital revisits and diagnoses, it may suggest patient has been suffering from a highly suspected side effect of statins: they could in fact cause cholesterol elevation, not reduction, thus contribute to the severity of CHF [15], and eventually leads to CHF onsets. The proposed model correctly catches this pattern and makes correct prediction.

Case II: Long-term and frequent recurrence of cardiac dysrhythmias Cardiac dysrhythmia is a condition that the patient either has irregular heart rhythm or abnormal heart rate or pacing. Without considering time information, cardiac dysrhythmias itself is not a discriminating marker in detecting CHF. Some dysrhythmias (e.g. sinus arrhythmia) are considered normal conditions [2].

However, with considering time information, for example, constant recurrence of cardiac dysrhythmias may indicate impending or onset of CHF. From the results, we observes some patient has simple patterns of recurring and frequent diagnosis of cardiac dysrhythmias together with some occasional conditions such as hypertension over a year. Imagine if these events occur sparsely over a decade, we cannot tell whether they are actually associated with CHF. However, with their frequent occurrence over only a year, these patient would have higher risk of CHF onsets. The proposed model catches this pattern and makes correct prediction.

6 Conclusion

In this work, we propose the Health-ATM, a novel deep architecture that uncovers patients' comprehensive health conditions underlying their noisy, longitudinal, heterogeneous and irregular sampled EHR data using an integrated deep architecture that extracts multifaceted patient patterns (e.g. contextual, temporal, time-aware, structural, and relevant) with attentive and time-aware modulars (ATM). We compared the performance of Health-ATM with its reduced models and a few state-of-the-art methods. Results of disease prediction from both real world and artificial data demonstrate the superior performance of Health-ATM against all baselines. In addition, we analyzed the interpretability of target-aware attentions, and importance of time stamps. Results and case analysis demonstrate the promising utility and efficacy for Health-ATM in EHR representation learning and disease onset predictions. Future directions include extending the architecture to be able to take heterogeneous data types as input, or consider general progression modeling as tasks.

Acknowledgements

The work of Fei Wang is supported by NSF IIS-1650723 and IIS-1716432.

References

[1] A simulated electronic medical record dataset., 2016.

[2] S. Atwood, C. Stanton, and J. Storey-Davenport. *Introduction To Basic Cardiac Dysrhythmias 4th Edition*. Jones and Bartlett Learning, 2013.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014.

[4] C. Che, C. Xiao, J. Liang, B. Jin, JY. Zhou, and F. Wang. An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease. In *SIAM International Conference on Data Mining*, 2017.

[5] Z. Che, Y. Cheng, Z. Sun, and Y. Liu. Exploiting convolutional neural network for risk prediction with medical feature embedding. In *NIPS 2016 Workshop on Machine Learning for Health (NIPS ML4HC)*, 2016.

[6] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 2016.

[7] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, 2014.

[8] Edward Choi, Mohammad Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Pre-

dicting clinical events via recurrent neural networks. In *Proceedings of Machine Learning for Healthcare*, 2016.

[9] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: graph-based attention model for healthcare representation learning. *CoRR*, abs/1611.07012, 2016.

[10] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

[11] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 2016.

[12] Y. Choi, CY. Chiu, and D. Sontag. Learning low-dimensional representations of medical concepts. *AMIA CRI.*, 2016.

[13] Robert Desimone and Duncan John. Neural mechanisms of selective visual attention. *Annual review of neuroscience.*, 18(1):193–222, 1995.

[14] Wael Farhan, Zhimu Wang, Yingxiang Huang, Shuang Wang, Fei Wang, and Xiaoqian Jiang. A predictive model for medical events based on contextual embedding of temporal sequences. *JMIR Medical Informatics*, 4(4), 2016.

[15] Duane Graveline and Malcolm Kendrick. The statin damage crisis 3rd ed. edition, 2012.

[16] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[17] Joyce C. Ho, Joydeep Ghosh, and Jimeng Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 115–124, New York, NY, USA, 2014. ACM.

[18] G. Hripcsak and DJ. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

[19] G. Hripcsak and DJ. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association.*, 20:117–121, 2013.

[20] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, November 1998.

[21] Uri Kartoun. A methodology to generate virtual patient repositories. *CoRR*, abs/1608.00570, 2016.

[22] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Leo Marcoff and Paul D. Thompson. The role of coenzyme {Q10} in statin-associated myopathy: A systematic review. *Journal of the American College of Cardiology*, 49(23):2231 – 2237, 2007.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA, 2013.

[25] R. Miotto, L. Li, BA. Kidd, and JT. Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports.*, 6, 2016.

[26] M. Pfisterer, P. Buser, H. Rickli, M. Gutmann, P. Erne, and P. Rickenbacher. Bnp-guided vs symptom-guided heart failure therapy. *JAMA: the journal of the American Medical Association.*, 301:383–392, 2009.

[27] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[28] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[29] HP. Van der, AA. Voors, WH. van Gilst, Bohm M., and DJ. van Veldhuisen. Statins in the treatment of chronic heart failure: A systematic review. *PLoS Medicine.*, 3, 2006.

[30] Fei Wang, Noah Lee, Jianying Hu, Jimeng Sun, and Shahram Ebadollahi. Towards heterogeneous temporal clinical event pattern discovery: A convolutional approach. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 453–461. ACM, 2012.

[31] Fei Wang, Ping Zhang, Buyue Qian, Xiang Wang, and Ian Davidson. Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 145–154, 2014.

[32] X. Wang, F. Wang, J. Hu, and R. Sorrentino. Exploring joint disease risk prediction. In *AMIA Annual Symposium Proceedings.*, pages 1180–1187, 2014.

[33] J. Wu, J. Roy, and WF. Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical Care.*, 48:S106–113, 2010.

[34] J. Zhao. Temporal weighting of clinical events in electronic health records for pharmacovigilance. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 375–381, Nov 2015.

[35] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 135–144, 2014.