# Multilevel Self-Attention Model and its Use on Medical Risk Prediction

Xianlong Zeng[1,2], Yunyi Feng[1,2], Soheil Moosavinasab[2], Deborah Lin[2], Simon Lin[2], Chang Liu[1]
*1. School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, USA*
*2. The Research Institute at Nationwide Children's Hospital, Columbus, OH, USA*
*Email: Liuc@ohio.edu*

Various deep learning models have been developed for different healthcare predictive tasks using Electronic Health Records and have shown promising performance. In these models, medical codes are often aggregated into visit representation without considering their heterogeneity, e.g., the same diagnosis might imply different healthcare concerns with different procedures or medications. Then the visits are often fed into deep learning models, such as recurrent neural networks, sequentially without considering the irregular temporal information and dependencies among visits. To address these limitations, we developed a Multilevel Self-Attention Model (MSAM) that can capture the underlying relationships between medical codes and between medical visits. We compared MSAM with various baseline models on two predictive tasks, i.e., future disease prediction and future medical cost prediction, with two large datasets, i.e., MIMIC-3 and PFK. In the experiments, MSAM consistently outperformed baseline models. Additionally, for future medical cost prediction, we used disease prediction as an auxiliary task, which not only guides the model to achieve a stronger and more stable financial prediction, but also allows managed care organizations to provide a better care coordination.

*Keywords:* Claims data; Cost prediction; Disease prediction; Self-attention; Deep learning.

## 1. Introduction

Being able to detect the early onset of diseases and identify risk factors make early intervention and risk management possible. Due to the complex nature of the problem and the diverse factors affecting future health conditions, effective data-driven solutions were not possible until the accumulation of a large amount of health data from Electronic Health Records (EHR) during the last two decades. EHR data contain rich health information, including medical codes (e.g., diagnoses, procedure, and medications), place of services, clinical notes, laboratory tests, and medical costs. With the increasing volume of EHR data, many deep learning models have been developed and applied to various healthcare tasks, such as disease predictions [1-4], phenotyping learning [5, 6], embedding learning [7, 8] and future cost prediction [9, 10]. The critical challenge for most of these tasks is to obtain a good representation of patients' historical medical records.

In claims data, a specific type of EHR, each patient can be viewed as a sequence of medical visits (facility and pharmacy visits) ordered by time, and each visit contains a set of unordered medical codes. One common way to model this structured data is to aggregate the medical codes within a visit to form a visit-embedding, then feed the longitudinal visit-embeddings through a recurrent neural network (RNN) to generate the representation of a patient. However, obtaining the patient's representation via such an approach has two limitations: (1) aggregating multiple types of

medical codes as a bag of codes will lose the complex relationships among them. (2) traditional RNN is not able to capture the dependencies among the visits nor handle the irregular time intervals.

To address these two limitations, we propose a multilevel self-attention model (MSAM) that utilizes the self-attention mechanism [11] at both medical code-level and visit-level, respectively. MSAM first embeds the discrete medical codes into a continuous distributed space, and then feeds the code embeddings through a code-level self-attention layer to form a visit representation. The code-level self-attention layer can relate different codes of a visit and embed relevant contextual information into each medical code. This self-attention mechanism can help the model better "understand" the usage and severity of each medical code. Next, MSAM combines each visit embedding and its corresponding time embedding, and then feeds them through a visit-level self-attention layer to generate the patient representation. The time embedding and visit-level self-attention layer enable the model to handle the irregular time intervals between visits and capture the progression of diseases. Finally, the learned patient representation is combined with demographic information (e.g., age, sex, and prior medical cost) to predict future events. As shown in Figure 1, MSAM is designed to capture the underlying relationships within the medical claims.

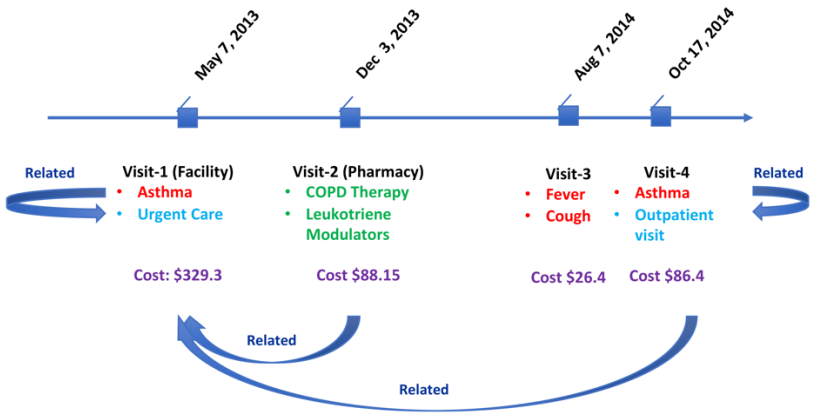


Fig. 1. An example sequence of the medical visits for a patient. There are many underlying relationships within medical claims data: 1) The asthma diagnoses in visit-1 and visit-4 might indicate different health concerns based on their nearby procedure codes. 2) The pharmacy claim (visit-2) is likely related to the first asthma visit, as it contains drugs related to asthma.

In the present study, we evaluated MSAM on two predictive tasks, future disease prediction and future medical cost prediction, with two large real-world datasets, MIMIC-3 and PFK. We compared our model to state-of-the-art approaches and our model demonstrated the best performance compared to these approaches. For cost prediction, we used disease prediction as an auxiliary task to achieve a stable and interpretable result. A case study was performed to demonstrate the reasonableness of the attention weight and interpretability of the predicted cost.

## 2. Related Work

### 2.1. *Future disease prediction*

One of the important prediction tasks in healthcare is future disease prediction. Accurate disease prediction results can help physicians to process patient's historical medical records and make automated clinical decision support possible. Riccardo Miotto et al. [12] first introduced a three-layer stacked autoencoder to model patient records and used the learned patient representation for future disease prediction. In order to capture the sequential information of medical visits, RNN-based models [13, 3] have been used to improve the prediction of future diseases. Later, many studies developed methods that could cope with the irregularity of the temporal gaps between visits. For example, Baytas et al. [5] proposed a time-aware RNN model that could handle the irregular time intervals between medical visits. Further, researchers have used TIMELINE [1], which uses a weight decay factor to learn the disease progression pattern and can distinguish chronic and acute diseases. Also, some deep learning models, such as GRAM [14] and KAME [15], have leveraged medical domain knowledge from medical ontology and achieved better prediction performance.

Although these models achieved a promising result for disease prediction, none of them take the complex relationships among medical codes into consideration. More recently, some studies [16, 17] started modeling the inherent relationships between different kinds of medical codes. MIME, proposed by Choi et al., can model the encounter structure of EHR and derive the visit embedding that is able to capture the dependencies among medical codes. Their study, however, heavily relied on the structure information within EHR data. This structure information might not exist in some EHR datasets such as claims data.

Compared with the aforementioned models, our MSAM not only can capture the underlying dependency between medical codes/visits automatically, but is also able to model the irregular visit time gaps. These two properties allow MSAM to effectively encode a patient's medical information, which improves prediction performance.

### 2.2. *Future medical cost prediction*

Medical cost is a proxy for resource use and has been operationalized in a variety of ways in health-related investigations (e.g., prices, charges, reimbursements, and indirect costs). In this study, we define medical cost as the actual paid amount to the accountable care organizations (ACOs) and narrow our analysis of previous research to studies with similar definitions and research goals.

Accurate forecasting of future medical cost is vital for healthcare organizations to coordinate care and resources and to evaluate the effectiveness of interventions. Three different kinds of approaches have been developed for predicting future medical cost: 1) rule-based models [18] [19], 2) machine learning models with cost-related predictors only [20, 21], and 3) machine learning models with cost-related features plus medical codes [22-24].

Many of the commercial solutions for cost prediction use rule-based models developed by medical experts. For example, the ACG system[a], developed by Johns Hopkins and the DxCG

[a] https://www.hopkinsacg.org/

model[b], implemented by Verisk (Jersey City, NJ) are two dominant models for predicting the future medical cost. Despite their outstanding performance, developing and maintaining these rule-based models consumes large amounts of resources.

Data-driven approaches such as machine learning and deep learning, provide another strategy to predict future medical cost without relying on manually developed rules. Cowen et al. [22] applied regression models to aggregated medical codes and prior cost to model future medical cost. However, this method ignores the temporal information within medical data. Bertsimas et al. [24] developed a CART mode, which considered temporal patterns from the cost features. They found that adding aggregated medical features barely improved their model performance. Additionally, Morid et al. [10] captured the spike features (i.e., the fluctuation of prior medical cost) to model future cost. These two methods utilized temporal information about prior medical cost and largely improved the prediction performance. This improvement suggests that temporal information is vital for modeling future medical cost.

Compared to the above models, our MSAM for cost prediction can not only further leverage the irregular temporal information, but also take advantage of the underlying relationships within medical claims. In addition, to mitigate the training difficulties caused by the highly skewed cost data, we utilized disease prediction as an auxiliary task to achieve a stable prediction.

## 3. Methods

This section will introduce the terminology and notation we use to describe the dataset and model (section 3.1), followed by a general description of MASM (section 3.2), the self-attention encoder unit (section 3.3), and the loss function adopted for future diseases prediction and future medical cost prediction (section 3.4). The source code of this work is freely available on GitHub (https://github.com/1230pitchanqw/MSAM)

### 3.1. *Terminology and Notation*

Each patient in our datasets was represented as a sequence of medical visits $v_1, v_2, \dots, v_i$ ordered by service date $t$. The $i$-th visit $v_i$ is represented by a set of codes that include diagnoses, procedures and prescriptions $\{c_1, c_2, \dots, c_j\} \subseteq |C|$, where $C$ represents the entire set of medical codes.

### 3.2. *Model Architecture*

As shown in Figure 2, medical codes within each medical visit were first projected into a m-dimensional continuous embedding space via a trainable code embedding matrix $W_c$, Then the medical codes were passed through the code-level encoder and aggregated into a visit embedding $v_i$ via the following equation,

$$v_i = \sum_{k=1}^{|v_i|} S_c\left(c_k \middle| c_1, c_2, \dots, c_j\right), \quad (1)$$

[b] https://www.cotiviti.com/solutions/quality-and-performance/dxcg-intelligence

where $|v_i|$ denotes the number of codes within visit $|v_i|$ and $S_c$ denotes the code-level self-attention encoder (the detail of self-attention encoder is shown in Section 3.3).

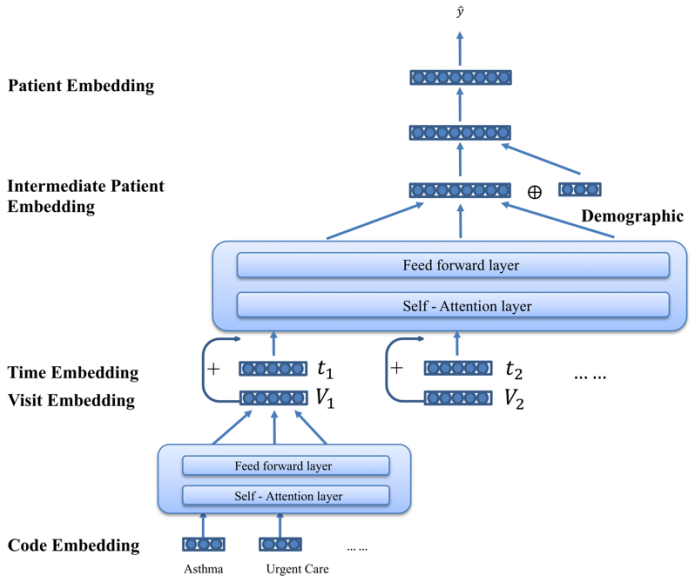


Fig. 2. The MSAM architecture.

Next, we added the time embedding, generated by the time encoding function $TE$, for each medical visit in order to capture the irregular service date information:

$$e_i = v_i + TE(\Delta t_i), \tag{2}$$

where $\Delta t_i$ represents the time interval between the visit $v_i$ and the time the model makes the prediction. And $e_i$ is the combination of the time embedding and the visit embedding. There are many possible functions for encoding time. In our experiments, we used the sinusoid encoding function [11].

After we obtaineds the vector representation for each medical visit, we aggregated the visit vectors via the visit-level self-attention encoder $S_v$ and formed the intermediate patient representation $u$ as follows:

$$u = \sum_{l=1}^{|u|} S_v\left(v_l | v_1, v_2, \ldots, v_i\right), \tag{3}$$

where $|u|$ denotes the number of visits and $S_v$ denotes the visit-level self-attention encoder.

Finally, we concatenated the intermediate patient embedding $u$ and the one-hot encoded demographic embedding $d$, then stacked three fully connected feedforward layers to obtain the patient embedding $p$. We also used a skip-connection [25] between each layer to increase the representative power:

$$p = F([u, d]) + F, \tag{4}$$

Where $F(u)$ contains three feedforward blocks and " $+ F$" represent the skip-connection operation between each layer.

### 3.3. *Self-attention Encoder Unit*

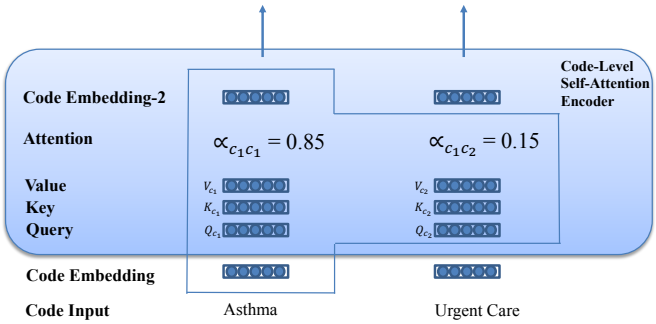


Fig. 3. The architecture of the self-attention layer (for code-level).

The implications of medical codes and visits vary depending on the context. To capture this contextual information, we applied the two self-attention units to both code-level and visit-level. A self-attention unit contains a self-attention layer, normalization layer and a feed forward residual connected layer. Figure 3 illustrates the architecture of the self-attention layer and the equations are shown below:

$$S_c\left(c_k|c_1, c_2, \ldots, c_j\right) = \sigma_v(f\left(c_k + \sum_{l=1}^{|v|} \propto_{c_k c_l} c_k\right)) \tag{5}$$

$$\propto_{c_k c_1}, \propto_{c_k c_2}, \ldots, \propto_{c_k c_{|v|}} = softmax\left(\frac{Q_{c_k} K_{c_1}}{\sqrt{m}}, \frac{Q_{c_k} K_{v_2}}{\sqrt{m}}, \ldots, \frac{Q_{v_k} K_{v_{|v|}}}{\sqrt{m}}\right) \tag{6}$$

$$Q_{c_k} = W_q c_k \tag{7}$$

$$K_{c_k} = W_k c_k \tag{8}$$

where $\sigma$ denotes the residual connection and layer normalization, $f$ denotes the feedforward block. $W_q, W_q \in R^{m \times m}$ are weight matrices for generating query and key vectors $Q_{c_k}$ and $K_{c_k}$. $\propto_{c_k c_1}$ denotes the attention score for a code $c_l$ when generating the vector representation of code $c_k$.

### 3.4. *Loss Function*

Disease prediction is a multiclass classification task, whereas the medical cost prediction is a regression task. Accordingly, we used negative log-likelihood loss function for disease prediction and mean-squared-error for cost prediction,

$$Loss_d = -(y\, log\hat{y} + (1 - y)\log(1 - \hat{y})), \tag{9}$$

$$Loss_c = -\frac{1}{2}(y - \hat{y})^2 \tag{10}$$

where $y$ is the target value and $\hat{y}$ is the predicted value.

Medical cost is highly skewed and can be affected by many personal or accidental factors such as transportation accident, sports damage, and even financial status. Thus, in order to mitigate the

uncertainty and stabilize the predicted result, we jointly performed disease prediction as an auxiliary prediction task with a ratio coefficient $\lambda$ for medical cost prediction.

$$Loss_{aux} = \lambda\, Loss_c - (u\, log\hat{u} + (1-u)\log(1-\hat{u})) \quad (11)$$

## 4. Experiments

### 4.1. *Source of Data*

Medical Information Mart for Intensive Care III (MIMIC-3) [26] is a freely accessible dataset that contains medical records for around 38,000 patients from an intensive care unit (ICU) over 11 years. MIMIC-3 does not contain medical cost information and is therefore only used for future disease prediction. Patients with less than two medical records were excluded from the experiments.

Partner for Kids (PFK) is one of the largest pediatric ACOs for Medicaid enrollees in central and southeastern Ohio. Our PFK dataset contains 146,287 enrollees' medical claims from 2013 to 2014 with two years of continuous eligibility. In accordance with the Common Rule (45 CFR 46.102[f]) and the policies of Nationwide Children's Institutional Review Board, this study used a limited dataset and was not considered human subjects research and thus not subject to institutional review board approval.

### 4.2. *Dataset preprocessing*

For future disease prediction using the MIMIC-3 dataset, the dataset was constructed using previous medical records to predict the disease of the next visit. Diagnosis codes and procedure codes were extracted from the records. In order to improve model performance and outcome stability, we aimed to predict the grouped diagnosis categories instead of the specific diagnosis. We used Clinical Classification Software[c] (CCS) to group the diagnosis codes into around 280 categories.

Table 1. Statistical information of the MIMIC-3 and the PFK datasets.

| Dataset | MIMIC-3 | PFK |
|---|---|---|
| # of patients | 7,537 | 146,287 |
| Age avg. | -- | 8.5 |
| Male pct. | -- | 51% |
| Avg. # of visits per patient | 1.6 | 8.9 |
| Avg. # of codes per visit | 15.9 | 5.0 |
| # of Diagnosis, Procedure, Drug | (4894, 1442, --) | (7497, 4499, 338) |
| # of CCS categories | 282 | 291 |
| Avg. cost per patient | -- | $1282.1 |
| Median cost per patient | -- | $514.8 |

For prediction experiments using the PFK dataset, we used the prior year's medical information to predict the disease or medical cost of the next year. Diagnoses, procedures, and medications (drug

[c] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

class) were extracted from claims for disease prediction. Diagnoses, procedures, medications, age, gender, and medical cost (i.e., the actual paid amount to the ACO) were extracted for future cost prediction. The facility and pharmacy visits were grouped by month. We converted all negative paid amounts into 0 and removed all claims with empty service date (less than 0.1% of such claims). Table 1 lists the detailed statistics about the two datasets after preprocessing.

### 4.3. *Implementation details*

All models including baselines were implemented using TensorFlow. The dimensionality of code embedding and visit embedding were chosen from {100, 200}. The number of self-attention head and feedforward block were chosen from {1, 2, 3}. The auxiliary coefficient $\lambda$ was chosen from $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. Hyper-parameters that yield the best model performance on the validation dataset were then used for testing. More detail information is shown in the GitHub repository.

## 5. Results

### 5.1. *Future disease prediction*

**Objective**: Predict all diagnosis codes in the next visit for the MIMIC-3 dataset and predict all diagnosis codes in the next year for the PFK dataset. **Evaluation Metric:** *Recall@k* is defined as the number of successfully recalled medical codes from the k recalled codes divided by the number of true positive diagnosis codes. *Recall@k* reflects the accuracy of clinical diagnostic decision and is widely used in many disease prediction tasks [2, 3].

Table 2. The *recall@k* of disease prediction task. The values within parentheses indicate the standard deviation from 5 different random data splits.

| Dataset | Model | *Recall@10* | *Recall@20* | *Recall@30* |
|---|---|---|---|---|
| MIMIC-3 | Most Frequent | 0.2255 | 0.3646 | 0.4716 |
| | MLP [12] | 0.2132(0.014) | 0.3526(0.018) | 0.4687(0.022) |
| | RNN [2] | 0.3451(0.009) | 0.5093(0.009) | 0.6210(0.009) |
| | B-RNN | 0.3603(0.011) | 0.5247(0.010) | 0.6389(0.009) |
| | AB-RNN [3] | 0.3671(0.011) | 0.5466(0.011) | 0.6474(0.010) |
| | TIMELINE [1] | -- | -- | -- |
| | MSAM | **0.4027(0.012)** | **0.5783(0.012)** | **0.6830(0.011)** |
| PFK | Most Frequent | 0.4412 | 0.6123 | 0.7185 |
| | MLP | 0.4612(0.003) | 0.6589(0.005) | 0.7761(0.001) |
| | RNN | 0.5014(0.003) | 0.6810(0.002) | 0.7887(0.002) |
| | B-RNN | 0.5193(0.007) | 0.6987(0.005) | 0.7880(0.003) |
| | AB-RNN | 0.5392(0.001) | 0.6995(0.002) | 0.7908(0.002) |
| | TIMELINE | 0.5397(0.001) | 0.7079(0.002) | 0.7929(0.001) |
| | MSAM | **0.5514(0.004)** | **0.7187(0.002)** | **0.7948(0.002)** |

Table 2 shows the experimental results of our MSAM and baseline models on both MIMIC-3 and PFK dataset. The Most Frequent model uses the top k most frequently occurred diagnosis code as the prediction. The best model performance in each column is marked in bold.

As shown in Table 2, our proposed MSAM consistently outperformed all baseline models on MIMIC-3 and PFK. Among all baselines, The Most Frequent and MLP models underperformed compared to advanced deep learning models. By contrast, RNN-based models showed promising model performance. Compared to vanilla RNN, bidirectional RNN (B-RNN) is able to remember medical information from both far and recent medical visits. With the help of the attention mechanism, the attention-bidirectional RNN (AB-RNN) can better measure the relationships between medical visits. TIMELINE is also an RNN-based model, it introduces a time factor that can capture the irregular time gaps between medical visits. As a result, as the RNN variant can capture more medical information, the model gains more predictive power and thus shown better performance.

MSAM utilized the time embedding to capture the irregular temporal information and employed two self-attention units to capture the underlaying relationships within medical codes and medical visits. Compared to all baseline models, the *recall@k* of MSAM was higher for both the MIMIC-3 and PFK datasets. This improvement shows that modeling irregular temporal information and underlaying relationships can increase the prediction performance.

### 5.2. *Future cost prediction*

**Objective**: Predict the medical cost (i.e. the actual paid amount to the ACOs) in the next year for the PFK dataset. **Evaluation Metric:** Mean absolute error (MAE) is defined as the absolute difference between the predicted cost and the true cost. MAE is used to measure how close each prediction outcome was to the target value.

Table 3. The evaluation results for cost prediction task, the values within parentheses indicate the standard deviation form 5 different random data splits.

| Model | Medical Features | $ MAE (SD) |
|---|---|---|
| Most Recent | Not Included | 1120.7 (39.6) |
| CART [24] | | 1318.7 (27.5) |
| LASSO | | 1033.1 (34.3) |
| XGBOOST | | 953.7 (28.2) |
| CART [24] | Aggregated | 1316.2 (22.4) |
| LASSO | | 1030.0 (25.9) |
| XGBOOST | | 991.5 (26.4) |
| RNN | Sequential | 944.5 (31.9) |
| B-RNN | | 942.6 (34.8) |
| AB-RNN | | 981.2 (39.3) |
| TIMELINE | Irregular Temporal | 937.3 (31.5) |
| MSAM | | 860.8 (35.6) |
| $MSAM_{AUX}$ | | **847.7 (27.7)** |

As shown in Table 3, the two MSAM variants outperformed baselines under the MAE evaluation metric and the MSAM with auxiliary task achieved the best performance across all models as well as the lowest variance across all advanced deep learning models. The "Most Recent" model used the last year's medical cost as the prediction. The "Medical Features" column indicates the medical

information that is utilized by the corresponding model. The best model performance is marked in bold.

Table 3 shows that adding aggregated medical features to traditional machine learning algorithms barely improved the prediction performance. This observation was also reported by Bertsimas et al. [24] and indicates that the aggregated medical information largely overlaps with the medical cost information. On the contrary, deep learning models can fully utilize sequential or temporal medical information. This advantage enables the model to capture the progression of a patient's health condition and thus grants the model more predictive power, leading to higher model performance compared to traditional machine learning models.

From the transition between aggregated data to sequential data, and then from sequential data to irregular temporal data, the model gains more medical information. This increase in medical information helps to improve the performance of the deep learning model. Consequently, the MSAM achieved the best model performance compared to all baseline models. This result confirms that being able to capture the underlying relationships can further increase the model's predictive power. In addition, the implemented auxiliary task (disease prediction) mitigated the random nature of incurring the medical cost and stabilized the prediction. Notably, $\text{MSAM}_{\text{AUX}}$ achieved the lowest MAE score across all models and also the lowest standard deviation across all deep learning models.

### 5.3. *Case study for the self-attention mechanism*

To explore how the self-attention mechanism works on claims data, we limited the number of the attention heads to one and analyzed the code-level attention weights via a case study. We selected four visits for each of the following diseases: *diabetes (ICD9-250.00)*, *asthma (ICD9-493.90)*, and *convulsion (ICD9-780.39)*. Figure 4 shows the code-level attention scores for encoding these three codes. The x-axis represents the visit-id and each visit contains 3 to 4 medical codes, while the y-axis represents the attention score. From Figure 4, we can observe that the attention score was different when the medical code co-occurs with different contextual medical codes. These differences indicate the self-attention mechanism enables each medical code to express different health concern given different neighboring codes.

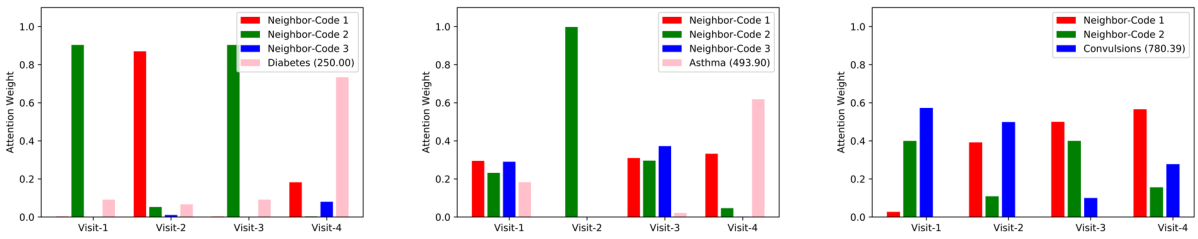


Fig. 4. Attention score analysis. Left: the attention scores on contextual medical codes when embedding *Diabetes*; Middle: the attention scores on contextual medical codes when embedding *Asthma*; Right: the attention scores on contextual medical codes when embedding *Convulsions*;

Next, to better illustrate how the self-attention mechanism works, we listed the detail information of one of the asthma visits (the visit-1 of asthma visits in Fig 4) in Table 4 and analyzed the attention weight from the clinical perspective. As shown in the table, attention scores [0.18, 0.24, 0.29, 0.29] were obtained for encoding the medical code *asthma (ICD-493.90)*. The 0.24 attention

score shows that a significant amount of the attention was put on medical codes *outpatient visit (CPT-99214)*. This attention score allows the model to carry certain pieces of information: the severity of the asthma is not high since it is an outpatient visit instead of an emergency visit. Similarly, the two 0.29 attention scores allowed the model to embed information: the asthma disease is well addressed by proper medications.

Table 4. Diagnosis codes in visit-1 (i.e. "CodeSet-1") and their corresponding attention scores.

| Date | Diagnosis Codes | Attention |
|---|---|---|
| Sep/03 | **Asthma, unspecified type (ICD-493.90)** | 0.18 |
| Sep/03 | Outpatient visit (CPT-99214) | 0.24 |
| Sep/05 | Asthma/COPD Therapy - Beta Adrenergic Agents (Drug) | 0.29 |
| Sep/05 | Medical Supplies & DME - Respiratory Therapy (Drug) | 0.29 |

## 6. Conclusion

In this work, we developed a multilevel self-attention model (MSAM) that can model the complex claims data and predict future disease and future medical cost. By utilizing the self-attention units, time embedding and the auxiliary task, MSAM is able to capture the underlying relationships among medical claims, handle the irregularity time gaps between medical visits and stabilized the prediction result. We examined the predictive performance of MSAM on two real-world healthcare datasets, MIMIC-3 and PFK. Our proposed MSAM outperforms all baseline models on the two predictive tasks evaluated. We also provide a case study to illustrate the effectiveness of the self-attention mechanism.

In the future, we plan to test MSAM on more health-related tasks such as high-risk patient selection and preventable cost prediction.

## 7. Bibliography

[1] T. Bai, S. Zhang, B. L. Egleston and S. Vucetic., "Interpretable representation learning for healthcare via capturing disease progression through time.," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[2] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, 2016.

[3] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun and J. Gao., "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017.

[4] A. Rajkomar, E. Oren, K. Chen, H. N, D. AM and P. J. Liu., "Scalable and accurate deep learning for electronic health records," *npj Digit Med,* pp. 1-10, 2018.

[5] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain and J. Zhou., "Patient subtyping via time-aware LSTM networks.," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017.

[6] D. Y. Ding, C. Simpson, S. Pfohl, D. C. Kale, K. Jung and N. H. Shah., "The Effectiveness of Multitask Learning for Phenotyping with Electronic Health Records Data," in *PSB*, 2019.

[7] Y. Choi, C. Y.-I. Chiu and D. Sontag, "Learning low-dimensional representations of medical concepts.," in *AMIA Summits on Translational Science Proceedings*, 2016.

[8] Beaulieu-Jones, B. K, I. S. Kohane and A. L. Beam., "Learning Contextual Hierarchical Structure of Medical Concepts with Poincairé Embeddings to Clarify Phenotypes.," in *PSB*, 2019.

[9] C. Yang, C. Delcher, E. Shenkman and S. Ranka., "Machine learning approaches for predicting high cost high need patient expenditures in health care," *biomedical engineering online,* vol. 17, p. 131, 2018.

[10] M. A. Morid, O. R, L. Sheng, K. Kawamoto, T. Ault, J. Dorius and S. Abdelrahman., "Healthcare cost prediction: Leveraging fine-grain temporal patterns.," *Journal of biomedical informatics,* vol. 91, pp. 130-113, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin., "Attention is all you need," in *Advances in neural information processing systems,*, 2017.

[12] R. Miotto, L. Li, B. A. Kidd and J. T. Dudley., "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *scientific reports,* vol. 6, p. 26094, 2016.

[13] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz and W. Stewart., "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism.," in *Advances in Neural Information Processing Systems*, 20146.

[14] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart and J. Sun., "GRAM: graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[15] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou and J. Gao., "Kame: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.

[16] P. Nguyen, T. Tran and S. Venkatesh., "Resset: A recurrent model for sequence of sets with applications to electronic medical records.," in *International Joint Conference on Neural Networks*, 2018.

[17] E. Choi, C. Xiao, W. Stewart and J. Sun., "Mime: Multilevel medical embedding of electronic health records for predictive healthcare," in *Advances in Neural Information Processing Systems*, 2018.

[18] G. C. Pope, R. P. Ellis, A. S. Ash, J. Z. Ayanian, D. W. Bates, H. Burstin, L. I. Iezzoni, E. Marcantonio and B. Wu., "Diagnostic cost group hierarchical condition category models for Medicare risk adjustment," Health Economics Research, Waltham, MA, 2000.

[19] A. S. Ash, R. P. Ellis, G. C. Pope, J. Z. Ayanian, D. W. Bates, H. Burstin, L. I. Iezzoni, E. MacKay and W. Yu., "Using diagnoses to describe populations and predict costs.," *Health care financing review,* vol. 21, no. 3, p. 7, 2000.

[20] E. W. Frees, X. Jin and X. Lin., "Actuarial applications of multivariate two-part regression models," *Annals of Actuarial Science,* vol. 7, no. 2, pp. 258-287, 2013.

[21] M. A. Morid, O. R. L. Sheng, K. Kawamoto, T. Ault, J. Dorius and S. Abdelrahman., "Healthcare cost prediction: Leveraging fine-grain temporal patterns," *Journal of biomedical informatics,* vol. 91, pp. 103-113, 2019.

[22] M. E. Cowen, D. J. Dusseau, B. G. Toth, C. Guisinger, M. W. Zodet and Y. Shyr., "Casemix adjustment of managed care claims data using the clinical classification for health policy research method," *Medical care,* pp. 1108-1113, 1998.

[23] I. Duncan, M. Loginov and M. Ludkovski., "Testing alternative regression frameworks for predictive modeling of health care costs.," *North American Actuarial Journal,* vol. 20, no. 1, pp. 65-87, 2016.

[24] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala and G. Wang., "Algorithmic prediction of health-care costs.," *Operations Research ,* vol. 56, no. 6, pp. 1382-1392, 2008.

[25] K. He, X. Zhang, S. Ren and J. Sun., "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[26] A. E. Johnson, T. J. Pollard, L. Shen, H. Lehman, Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi and R. G. Mark., "MIMIC-III, a freely accessible critical care database," in *Scientific data*, 2016.