IEEE Open Journal of the Computer Society

Received 16 December 2020; accepted 12 January 2021. Date of publication 18 January 2021;
date of current version 12 February 2021. The review of this paper was arranged by Associate Editor Xuan Song.

Digital Object Identifier 10.1109/OJCS.2021.3052518

Multi-View Deep Learning Framework for Predicting Patient Expenditure in Healthcare

XIANLONG ZENG [1], SIMON LIN [2], AND CHANG LIU [3] (Member, IEEE)

[1]Electrical Engineering and Computer Science, Ohio University, Athens, OH 45701 USA
[2]RISI, Nationwide Children's Hospital, Columbus, OH 43205 USA
[3]School of EECS, Ohio University, Athens, OH 45701 USA

CORRESPONDING AUTHOR: XIANLONG ZENG (e-mail: xz926813@ohio.edu)

This work was supported in part by the U.S. Department of Commerce under Grant BS123456.

ABSTRACT Accurately predicting patient expenditure in healthcare is an important task with many applications such as provider profiling, accountable care management, and capitated medical payment adjustment. Existing approaches mainly rely on manually designed features and linear regression-based models, which require massive medical domain knowledge and show limited predictive performance. This paper proposes a multi-view deep learning framework to predict future healthcare expenditure at the individual level based on historical claims data. Our multi-view approach can effectively model the heterogeneous information, including patient demographic features, medical codes, drug usages, and facility utilization. We conducted expenditure forecasting tasks on a real-world pediatric dataset that contains more than 450,000 patients. The empirical results show that our proposed method outperforms all baselines for predicting medical expenditure. These findings help toward better preventive care and accountable care in the healthcare domain.

INDEX TERMS Administrative claims data, deep learning, electronic health record, expenditure prediction, machine learning.

I. INTRODUCTION

The increasing healthcare expenditures represent a significant challenge to healthcare providers and care organizations. As reported by the Centers for Medicare & Medicaid Services (CMS), the national health expenditure (NHE) for the United States grew 4.6% to $3.6 trillion in 2018 (i.e., $11,172 per person) and accounted for 17.7% of Gross Domestic Product (GDP). Specifically, Medicare spending grew 6.4% to $750.2 billion, and Medicaid grew by 3.0% to $597.4 billion.[1] The healthcare system is likely to become unsustainable unless medical cost growth is kept in check [1]. It is imperative to control the healthcare expenditure increase and reduce the medical cost for each individual.

Claims data, a special kind of Electronic Health Records (EHR) mainly for billing purposes, contains longitudinal patient health records including demographics, diagnoses, procedures, medications, facility usages, and expenditures. The claims data is one of the richest available sources for estimating patients' health conditions. The increasing amount of claims data provides a new, promising approach to tackle healthcare expenditure problems. Leveraging the historical claims, one can develop data-driven models to reveal important insights associated with the expenditure patterns. Specifically, an accurate medical cost predictive model at the individual level can help to identify patients with high medical risk and deliver a better quality of care.

Existing approaches for patient expenditure prediction usually rely on handcrafted features and linear regression-based models [2], [3]. For example, the Diagnostic Cost Groups (DCG) [4] model applies linear regression to predict healthcare expenditure based on the diagnostic categories manually designed by domain experts. Bertsimas *et al.* [5] developed a Classification And Regression Tree (CART) based on the aggregate medical codes and handcrafted cost features. These models help to predict healthcare expenditures. However, they are suffering from the following limitations: 1) They heavily rely on domain knowledge to group high-dimensional medical

[1]https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData

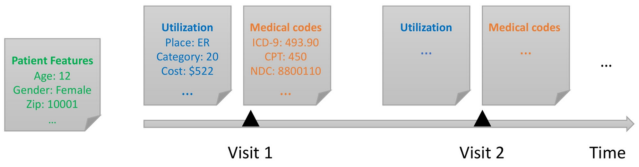

FIGURE 1. An illustrative example of claims data. Each patient can be represented as three data fields: demographic features, facility utilization, and medical codes.

codes into semantic-related categories. 2) they fail to utilize the rich information within claims data, such as facility usage, temporal information, and medical code correlation. 3). The linear regression-based model limits the predictive power and yields a sub-optimal model performance. This study aims to address these limitations in the literature.

Our work is motivated by several observations. First, the information in claims data is heterogeneous. As shown in Figure 1, different fields, such as medical codes, facility usages, and demographics, may have different characteristics and data structures. These fields are informative for expenditure prediction but hard to be modeled by general-purpose machine learning or deep learning models. To fully leverage these fields, a framework that can handle the multimodal inputs is needed. Second, there are more than 20,000 unique medical codes in the claims data. Representing medical code using the bag-of-words approach results in a sparse, high-dimensional vector, while aggregating them into categories requires expert knowledge and loses granular information. We need a method to capture the semantic meaning of medical codes and efficiently represent them in a denser vector. Third, different medical codes may have different importance for predicting expenditure. For example, *Malignant neoplasm (ICD-9 162)* could be more important than *impacted cerumen (ICD-9 380.4)* for estimating future medical expenditure. Thus, a mechanism to calculate the importance of different medical codes is needed.

This paper proposes a multi-view deep learning framework to capture the heterogeneous information within claims data. Our framework incorporates different data fields as different views. Specifically, the proposed model leverages a feedforward neural network to embed the non-sequential demographic features, an attention-based bidirectional recurrent neural network to capture the sequential facility usage, and a hierarchical attention network for learning medical code information. The attention mechanism calculates the importance of input variables and provides an interpretation of the predicted outcome. We demonstrate that the proposed multi-view deep learning framework achieves promising model performance for expenditure prediction compared to various baselines on a large pediatric claims data. The effectiveness of each view in the framework is evaluated via an ablation study. A case study is conducted to validate the learned personalized attention weights. Our model is also used to identify high utilizers and shows the potential to provide better population care management. In summary, the key contributions of our study are as follows:

1) We develop a multi-view deep learning framework to model the heterogeneous information within claims data. Our proposed method scales to input variables of thousands of dimensions and millions of patients without relying on expert domain knowledge.
2) The experimental results demonstrate that our approach can better predict healthcare expenditure at the individual level (R^2 >0.3). An ablation study is conducted to illustrate the benefits of exploiting different data fields in claims data.
3) Our framework shows better performance for selecting future high utilizers. This improvement implies that our model could enable care management entities to identify millions more in future costs from high utilizers compared to baseline approaches and better-allocate finite healthcare resources accordingly.

The rest of the paper is organized as follows: Section 2 discusses the connection of the proposed approaches to relevant literature. Section 3 presents the technical details of our multi-view deep learning framework and experimental setup. Section 4 shows the results of our approach in comparison to baselines for expenditure prediction. Section 5 is the discussion, and Section 6 concludes the paper.

II. RELATED WORK

Medical expenditure is a proxy for health-related utilization, including prices, charges, and reimbursements. This study defines medical expenditure as the dollar amount paid to the care organizations and focuses on research with a similar definition. This section first reviews the existing methods for predicting medical expenditures in Section 2.1. Next, in Section 2.2, we present the deep learning models that have been used for mining electronic healthcare records (EHR).

A. MEDICAL EXPENDITURE MODELS

In the 1990s, a medical cost model, called the Chronic Disease Score (CDS) [6], [7], was developed to predict future medical costs and hospital visits based on pharmaceutical information. The authors collected 250,000 managed-care adult enrollees and utilized the medications within six months as the input variable to calculate the predicted cost using a linear regression model. Then, the Medicaid Rx model [8], developed for the Medicaid-insured population, utilized demographic and pharmacy data to adjust per-person payment toward healthcare plans. Two well-known models, Adjusted Clinical Groups (ACG) [9] and Diagnostic Cost Groups (DCG) [10], were developed to predict medical expenditures based on diagnostic data. The authors collected patients' medical claims from health maintenance organizations (HMO) and measured their morbidity burden. To apply a regression model, they hand-grouped thousands of diagnostic medical codes into hundreds of condition groups. Later, many other researchers have refined the ACG and DCG systems and have evolved into

different variants designed for different populations and purposes [11]. Despite the strong interpretability, these models 1) showed limited model performance in a real-world setting, 2) are biased toward a certain population, and 3) required a considerable amount of human effort to develop and maintain.

Machine learning and deep learning approaches provide another strategy to predict medical expenditures without relying on manually developed groupers [12]–[15]. Bertsimas *et al.* [5] developed a CART mode, which considered temporal patterns from the cost features. They have found that adding aggregated medical features barely improved their model performance. Additionally, Morid *et al.* [1] captured the spike features (i.e., the fluctuation of prior medical costs) to model future costs. These two models utilized temporal information of prior expenditures and largely improved the model performance. This improvement suggests that temporal information is vital for modeling future medical costs.

B. DEEP LEARNING IN HEALTHCARE

Deep learning models are widely used for mining electronic medical records (EHR), including patient phenotyping [16]–[18], representation learning [19] and disease progression [20]–[23]. Specifically, recurrent neural network-based models are extremely popular for their ability to model the sequential medical data. Dipole [24] utilized an attention-based bidirectional recurrent neural network (RNN) to model medical visits and predict future diagnosis codes based on learned representation. RETAIN [25] employed RNN with a reverse attention mechanism for heart failure prediction. KAME [26] and GRAM [27] integrated medical knowledge to increase the models' predictive power for rare disease prediction. GCT [21] used a Transformer-based model to learn the hidden EHR structure for predicting patient readmission and mortality rate. Despite using similar datasets, these studies aim at a different prediction task. Medical expenditure prediction is different from the aforementioned tasks, as the outcome has different types, distributions, and is sensitive to different factors.

III. METHODS

In this section, we first introduce the dataset and how we preprocess the data. Then we describe the objective and predictors. Next, we present the details of our multi-view deep learning framework. Finally, we present the baselines as well as the evaluation metrics.

A. DATA AND PREPROCESSING

Experiments were conducted on administrative claims data collected from the Medicaid program by Partner For Kids (PFK). PFK is one of the largest nonprofit health care providers in the United States and delivers coordinated services for children in south-central and southeastern Ohio. Our dataset contains more than 8,500,000 medical records of 450,000 patients from Jan 2013 to Dec 2014. To be included in the experiments, enrollees must maintain continued eligibility from Jan 2013 to Dec 2014. The continuous eligibility enforcement is applied to avoid including patients with short

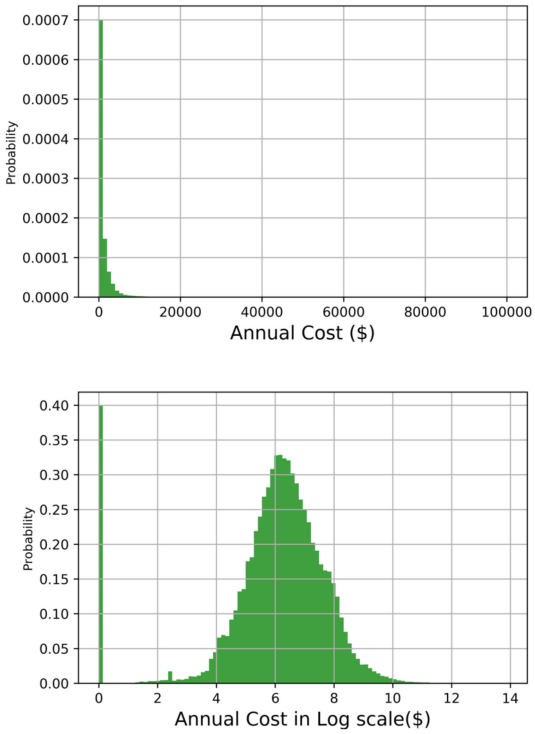

FIGURE 2. The distribution plot of per member per year dollar amount before log-transformation (top) and after log-transformation (bottomn). Due to the strong skewness (top), the model performance can be significantly affected and yield a sub-optimal predictive performance. The transformed outcome is normally distributed (bottom).

periods of enrollment but highly-variable profiles relative to the general population. Expenditure is defined here as the final paid amount, including professional, institutional, pharmaceutical, and dental costs. All negative paid amounts (about 4%) are converted to zero. 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.

B. OBJECTIVE AND PREDICTORS

Objective: Our objective was to predict healthcare expenditure at the individual level. In our predictive modeling, we used 2013's claims as the observation data and the medical expenditure in 2014 as the prediction target. Toward this goal, two predictive outcomes are conducted:

1) Per member per year dollar amount with the log scale (log-PMPY, log transformation of the total medical expenditure). The log-scale is applied to alleviate the skewness of the expenditure, as shown in Figure 2.
2) Rank percentiles of the per member per year dollar amount (pctl-PMPY, dividing the order rank of PMPY by the number of data points). Values of the rank percentiles range from 0 to 1.

Predictors: As shown in Table 1, there are three data fields in the claims data that are available as input variables:

TABLE 1. Summary of Predictors

Predictor
Non-sequential demographic features
Age, Gender, Zip, Prior medical/pharmacy cost.
Hierarchical medical codes sequences
Diagnoses codes (ICD-9-CM), Procedure codes (CPT & HCPCS), Medications (NDC).
Facility utilization sequences
Place of services, Category of services, Expenditure, Provider & Payer.

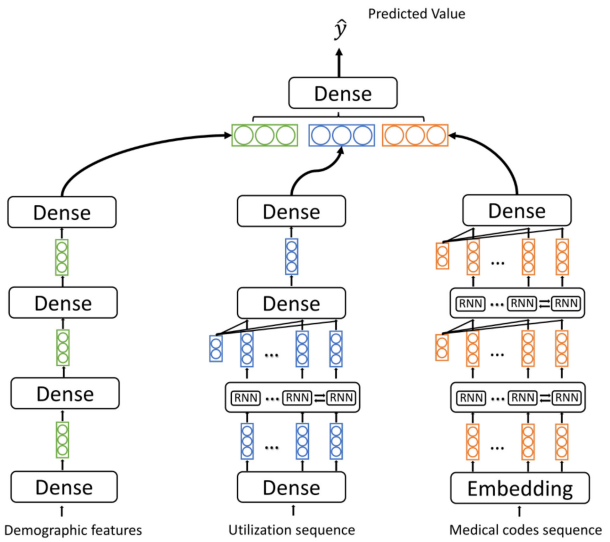

FIGURE 3. Multi-view deep learning framework.

non-sequential demographic features, hierarchical medical codes information, and facility utilization sequences. Among these features, diagnosis codes are encoded by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), procedure codes are encoded by the Current Procedure Terminology (CPT) and the Healthcare Common Procedure Coding System (HCPCS), and medication codes are encoded by the National Drug Codes (NDC). 5,567 unique diagnosis codes, 4,087 unique procedure codes, and 1,279 unique medication codes were identified during the study period.

C. OUR APPROACH: MULTI-VIEW DEEP LEARNING FRAMEWORK

Multi-view learning [39], [40] is a popular framework where heterogenous data are represented by multiple distinct fields, and each field is encoded through a unique learning module, as referred to as a view. The core of a multi-view learning framework is to learn efficient patient representation by incorporating the three data fields (i.e., demographics, medical codes, facility utilizations) as three different views. The architecture of our model is shown in Figure 3, where different color refers to a different learning module.

As shown in Figure 3, the first view is the demographic encoder, where demographic features are embedded into a vector representation via a feedforward neural network. The second view is the utilization encoder, where the utilization sequence is fed through an attention-based bidirectional recurrent neural network to generate the utilization vector. The third view is the medical code encoder, where the medical code representation is generated via a hierarchical attention neural network. Finally, the three vectors from the three views are concatenated to predict medical expenditure. The proposed model can be trained end-to-end.

First view, the demographic encoder: The raw demographic features are represented as the concatenation of binary encoded variables (e.g., age, gender, and zip code) and continuous variables (e.g., prior expenditures). Given the raw demographic feature vector d, a three-layer fully-connected neural network FNN_3 with $ReLU$ activation function is applied to learn the demographic representation, i.e., $r_d = FNN_3(d)$.

Second view, the utilization encoder: The utilization encoder is used to learn the patient representation from facility usage. The facility usage sequence $[u_1, u_2, \ldots, u_T]$ can provide important complementary information for predicting expenditure. For example, a patient who experiences a series of intense emergency room (ED) visits might indicate a high probability of future high utilization. To capture the longitudinal dependency between utilization sequences, an attention-based bidirectional recurrent neural network (BRNN) is used to learn the representation of utilization information. The BRNN consists of a forward and backward Long-Short Term Memory (LSTM) cell. Specifically,

$$e_i = W_u u_i + b_u \tag{1}$$

$$h_1, h_2, \ldots, h_T = BRNN_u (e_1, e_2, \ldots, e_T) \tag{2}$$

$$\alpha_i = \frac{\exp(\tanh(W_\alpha h_i + b_\alpha))}{\sum_{j=1}^{T} \exp\left(\tanh\left(W_\alpha h_j + b_\alpha\right)\right)} \tag{3}$$

$$r_u = \sum_{i=1}^{T} \alpha_i \odot h_i, \tag{4}$$

where α_i is the attention weight, e_i is the intermediate hidden representation of u_i, T is the length of the sequence, W_α, b_α are the attention parameters. The representation of utilization sequence r_u is the summation of the utilization hidden states $[h_1, h_2, \ldots, h_T]$ weighted by the corresponding attention weights $[\alpha_1, \alpha_2, \ldots, \alpha_T]$.

Third view, the medical codes encoder. In claims data, medical visits are represented as a set of medical codes. Considering that medical visits are usually clinically related when they are temporally close to each other, we employ a time window to split the sequence of visits into multiple subsequences with equal length. This processing step is commonly used in many healthcare predictive studies [28], [29]. Within each subsequence, there might be several medical visits or no visits at all. The width of the subsequence window size is a hyperparameter. In this study, we set the window size to one month. Given a patient represented as a sequence of subsequences, medical code $c_i \in \{0, 1\}^{|C|}$ within each subsequence was first

projected into a m-dimensional continuous space as follows,

$$d_i = ReLU\,(W_c\ c_i + b_c), \tag{5}$$

where $W_c \in R^{m\times|C|}$ is the weight matrix of medical codes, m is the size of embedding dimension, $|C|$ is the total number of medical codes. We observe that the medical codes within the subsequence are usually semantically related. Hence, skip-gram algorithm is employed to pretrain the medical code embedding and used the train vectors as the initialization of W_c. The skip-gram algorithm is first applied to learn the efficient representation of words in the natural language processing domain [30]. The basic idea is to learn a distributed vector representation of words by using each word in the sentence to predict its nearby words. Similarly, in our study, the input is a medical code, and the target is to predict medical codes within the same subsequence.

Next, the bidirectional recurrent neural network is applied to encode the medical code embeddings within the same subsequence. It takes the embedded code sequence $[d_1,\ d_2,\ \ldots, d_i]$ as input, and outputs the hidden representation $[g_1,\ g_2,\ \ldots, g_i]$. Because medical codes occurring within the same subsequence may have different informativeness, we cannot directly aggregate them with equal weights. Therefore, we leverage the attention mechanism to calculate the weights based on the hidden state. The calculation is shown as follows,

$$g_1,\ g_2,\ \ldots, g_i\ =\ BRNN_c\,(d_1,\ d_2, \ldots,\ d_i) \tag{6}$$

$$\beta_i = \frac{exp\left(tanh\left(W_\beta\ g_i + \ b_\beta\right)\right)}{\sum_{j=1}^{|v|} exp\left(\tanh\left(W_\beta\ g_j + b_\beta\right)\right)} \tag{7}$$

$$v = \sum_{i=1}^{|v|} \beta_i \odot\ g_i \tag{8}$$

where g_i is the intermediate hidden representation of d_i , β_i is the attention weight of medical code c_i, W_α, $b_\alpha$are the attention parameters, $|v|$ is the number of medical codes within the subsequence. The representation of subsequence is the summation of the medical code hidden states weighted by the corresponding attention weights.

Given the embedded subsequence vectors v, a bidirectional RNN is again used to encode the subsequence, then apply the attention mechanism to measure the importance of the subsequences. This yield,

$$l_1,\ l_2,\ \ldots, l_T\ =\ BRNN_v\,(v_1,\ v_2,\ \ldots, v_T) \tag{9}$$

$$\gamma_i = \frac{exp\left(tanh\left(W_\gamma l_i\ +\ b_\gamma\right)\right)}{\sum_{j=1}^{T} exp\left(tanh\left(W_\gamma l_j + b_\gamma\right)\right)} \tag{10}$$

where γ_i is the attention weight of subsequence v_i, W_γ and b_γ are the attention parameters. The representation of the hierarchical medical code information is formulated as: $r_c = \sum_{i=1}^{T} \gamma_i \odot\ l_i$. The final patient representation r is the concatenation of the vectors learned from different views, and the predicted expenditure is calculated as follows,

$$r = [\ r_d,\ r_u, r_c] \tag{11}$$

$$\widehat{y_i} = ReLU\ (W_o r\ + b_o) \tag{12}$$

where W_o and b_o are the output parameters. *ReLU* is used as the output activation function as the expenditure should not be negative in the real-world setting.

Objective function: Our target, medical expenditure, is a continuous value, where $y_i \in R$. Therefore, for model training, the mean square error is used as the objective function,

$$L = \frac{1}{N}\sum_{i=1}^{N} (y_i - \widehat{y_i})^2, \tag{13}$$

where y_i and $\widehat{y_i}$ are the true and predicted expenditure of the *i-th* patient, N is the total number of training data.

Implementation details. We set the maximum training epoch to 100 to guarantee convergence. Early stopping and dropout layer (with 0.5 dropout rate) are applied to avoid overfitting. The embedding size and the LSTM unit size are set to 200 to ensure sufficient predictive power. Model is implemented with TensorFlow, and the code is publicly available at our codebase.[2]

D. BASELINE AND EVALUATION METRIC

We evaluate our model's performance with various baseline methods that have shown promising performance in the previous study, including (1) Lasso [3]: Linear regression with L1 regularizer. (2) Ridge [3]: Linear regression with L2 regularizer. In the experiment, the L1 and L2 coefficients are set to 1. (3) DT [31]: Decision Tree. (4) RF [5]: random forest with 100 estimators (5) GBM [32]: gradient boosting machine, we used the implementation by [33] with default hyperparameters. (6) FNN [34]: feedforward neural network with three hidden layers. *ReLU* is used as the activation function of each hidden layer, and a dropout layer with a 0.5 dropout rate is applied after each hidden layer.

For all baseline models, the demographic features are directly used as part of the input variables. Facility usages and medical codes are first aggregated due to the sequential nature and high-dimensionality. Specifically, diagnostic codes were grouped into 251 categories using Clinical Classifications Software (CCS). Medication codes were grouped into 307 drug classes using the NDC Directory from the Food and Drug Administration (FDA). Approximately 600 input variables are used for baselines. More details can be found in our codebase.

To evaluate the performance of predicting future patient expenditure for each method, we use four measures:

1) Coefficient of determination (R^2). R^2 is often used to evaluate the predictive performance of expenditure forecasting algorithms. It compares the accuracy of the predictive value with respect to the mean target value. A higher R^2 indicates a better fitting model.

$$R^2\,(x, y) = 1 - \frac{\sum_{i=1}^{N} (y_i - x_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{14}$$

[2]https://github.com/1230pitchanqw/multi-view

TABLE 2. The Results for Predicting *Log-Pmpy*

	R^2	MAE	RMSE	PCC
Ridge [3]	0.229 (0.005)	1.111 (0.011)	1.676 (0.018)	0.479 (0.005)
Lasso [3]	0.231 (0.006)	1.113 (0.011)	1.676 (0.018)	0.480 (0.006)
DT [31]	-0.351 (0.018)	1.483 (0.009)	2.220 (0.013)	0.314 (0.003)
RF [5]	0.183 (0.007)	1.164 (0.007)	1.726 (0.016)	0.466 (0.006)
GBM [32]	0.269 (0.017)	1.068 (0.009)	1.632 (0.017)	0.520 (0.005)
FNN [34]	0.207 (0.023)	1.219 (0.048)	1.700 (0.037)	0.529 (0.007)
Multi-view$_{\text{plain}}$	0.300 (0.007)	1.055 (0.021)	1.598 (0.014)	0.551 (0.007)
Multi-view	0.301 (0.008)	1.053 (0.018)	1.596 (0.016)	0.552 (0.006)

TABLE 3. The Results for Predicting *Pctl-Pmpy*

	R^2	MAE	RMSE	PCC
Ridge [3]	0.324 (0.007)	0.195 (0.001)	0.238 (0.002)	0.570 (0.007)
Lasso [3]	0.325 (0.007)	0.195 (0.001)	0.238 (0.002)	0.570 (0.006)
DT [31]	-0.133 (0.010)	0.238 (0.001)	0.309 (0.002)	0.423 (0.005)
RF [5]	0.326 (0.001)	0.189 (0.001)	0.238 (0.001)	0.582 (0.002)
GBM [32]	0.375 (0.001)	0.185 (0.001)	0.229 (0.001)	0.613 (0.005)
FNN [34]	0.203 (0.046)	0.222 (0.007)	0.258 (0.008)	0.634 (0.009)
Multi-view$_{\text{plain}}$	0.360 (0.027)	0.197 (0.006)	0.233 (0.005)	0.633 (0.005)
Multi-view	0.400 (0.002)	0.183 (0.001)	0.224 (0.001)	0.637 (0.002)

2) Mean absolute error (MAE) and Root mean square error (RMSE). MAE and RMSE are two straight forward evaluation metrics for regression models. MAE and RMSE are calculated as:

$$MAE\ (x, y) = \frac{1}{N}\sum_{i=1}^{N}|y_i - x_i| \tag{15}$$

$$RMSE\ (x, y) = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}(y_i - x_i)^2} \tag{16}$$

3) Pearson Correlation Coefficient (PCC). PCC measures the linear relationship between two lists of values. The result of PCC varies between -1 to 1, with 0 indicating no correlation. In our expenditure prediction task, higher PCC indicates better predictive outcomes. The correlation coefficient is calculated as follows:

$$PCC\ (x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{17}$$

IV. RESULT

A. FUTURE EXPENDITURE PREDICTION

In this subsection, we examined the model performance on the expenditure prediction task. Table 2 and Table 3 showed the predictive results of our proposed model and baselines with the two objective outcomes, i.e., log-PMPY and pctl-PMPY, respectively. According to the tables, our multi-view approach can significantly outperform all baseline methods ($p<0.01$).

All baseline models, except the decision tree (DT), achieve promising predictive performance as indicated by the R^2. More than 20% of the variation can be explained by these baselines with the log-PYPM predictive outcome. The PCC value is more than 0.5, and MAE/RMSE reaches 1.1 and 1.6, respectively. GBM achieves the best model performance for all measures among all baselines. The superior performance of GBM is likely because of its strong predictive power: it combines many decision trees in series, and each is focusing on errors from the previous one. On the other hand, we observe that the decision tree method performs the worst among all models. This result is likely because the decision tree memorizes some noise from the data and fails to generalize.

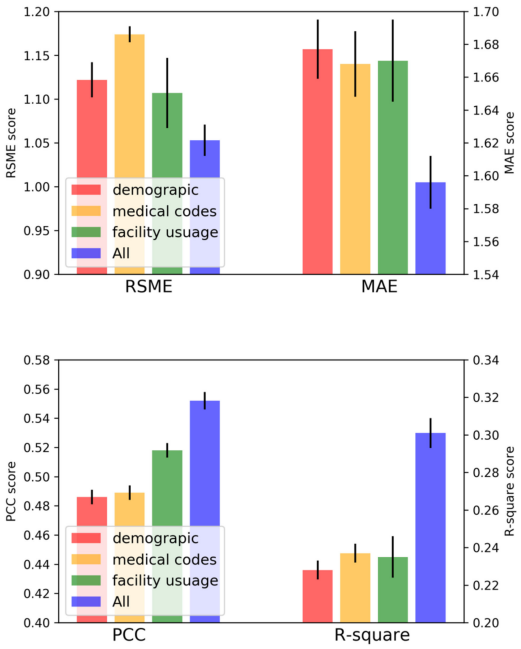

FIGURE 4. Effectiveness of the multi-view framework with log-PMPY predictive outcome. The lower RSME and MAE indicate better model performance (top). The higher PCC and R^2 indicate better model performance (bottom).

The attention mechanism is widely used in the sequential problem for boosting model performance and providing interpretability to the model. We validate the attention mechanism's effectiveness by removing the attention layer, i.e., Multi-view$_{\text{plain}}$. Comparing Multi-view and Multi-view$_{\text{plain}}$, we can observe that the attention mechanism slightly improves the predictive power.

The superior predictive performance of our model can be attributed to the efficient representation that learns from the claims data. The information in claims data is heterogeneous and has different structures. The multi-view framework incorporates different information as different views via a multi-view deep learning framework. By employing three different encoders to encode the three different data fields, our proposed method is able to accumulate relevant historical healthcare information and learn the succinct feature presentation of patients. Our framework also leverages the embedding technique to capture medical codes' semantic meaning and utilize attention mechanisms to focus on important information.

B. EFFECTIVENESS OF MULTI-VIEW FRAMEWORK

To illustrate the benefit of the multi-view approach, we analyzed the model performance using a single-view approach. Figures 4 and 5 show the experimental results under four evaluation metrics for predicting log-PMPY and pctl-PMPY, respectively. We can observe that the multi-view approach is

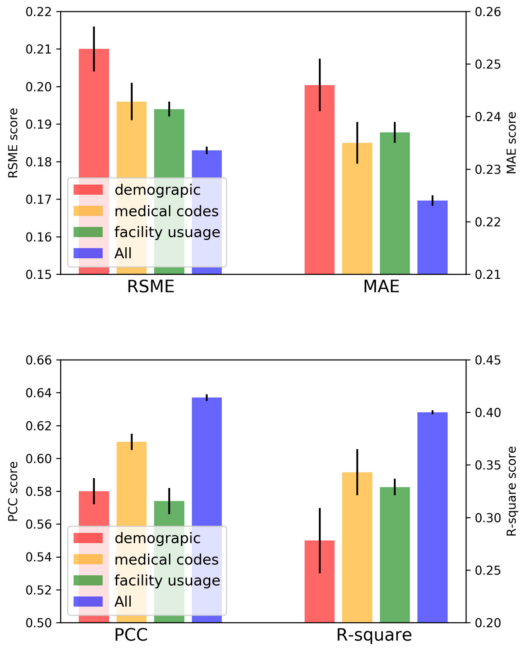

FIGURE 5. **Effectiveness of the multi-view framework with pctl-PMPY predictive outcome.**

better than the three single-view approaches under all evaluation measures. This improvement demonstrates the effectiveness of the multi-view framework. The model with demographic features yields reasonably-well outcomes. This observation indicates that age, gender, and prior total healthcare expenditures are good predictors for healthcare expenditure prediction. This finding was also reported by previous studies [12]. In addition, the model with medical codes information yields superior performance compared to the model with demographic features. This may be because the medical codes contain richer health status information than demographic features. Overall, combining three different information can significantly improve the predictive power (yield a 30% improvement on R^2). This performance improvement validates the effectiveness of our proposed multi-view framework.

C. EVALUATION OF MEDICAL CODE EMBEDDING

The quality of the medical code embedding is important for capturing the semantic information of medical codes. We evaluated the learned medical code embedding process using the mapping table between medical codes and medical code vectors. Specifically, we performed nearest neighbor queries to determine whether the embeddings have successfully captured medical codes' semantic similarity. For each medical code query, we sorted all other codes by the cosine distance to the query and displayed the top 5 nearest medical codes. Table 4 shows the medical codes selected as queries and their nearest neighbors.

The results in Table 4 are consistent with common intuition and clinical ontology. The diseases that share similar semantic meanings are close to each other. For example, *asthma (ICD9 493.90)* is close to other *asthma diseases (ICD9 493.92, 493.02, 493.91, 493.81)*, while *wheezing (ICD9 786.07)* is a common disease that often presents in asthma [35]. Therefore, it makes clinical sense that the top 5 most similar medical codes to *asthma (ICD 9 493.90)* are asthma-related diseases and *wheezing*. The top 5 most similar medical codes to *dental caries (ICD9 521.00)* are all dental procedures *(CPT D2330, D2335, D1510, D2331, D3220)*. This result is expected, and it indicates that our vector representations can also capture the semantic relationships between diagnosis codes and procedure codes.

TABLE 4. **Example of the Nearest Medical Code Query**

Query	493.90 (asthma)	250.00 (diabetes)	780.39 (convulsions)
Top 5 most similar	493.92 786.07 (wheezing) 493.02 493.91 493.81	250.13 250.02 250.03 K0552 (pump supplies) 250.11	345.90 (epilepsy) 781.0 (abnormal move) 95819 (EEG test) 780.02 780.31
Query	487.1 (influenza)	521.00 (dental caries)	311 (depressive disorder)
Top 5 most similar	487.8 780.60 (fever) 079.99 (viral infection) 04-0800 (tamiflu) 465.9 (respiratory infection)	D2330 (dental proc) D2335 D1510 D2331 D3220	300.9 (mental disorder) V62.84 (suicide ideation) 296.20 (depress) 296.23 296.34

* The description of the medical code is shown in the parenthesis. Diagnosis code with the same first three digits belongs to the same disease category (e.g., 493.90 and 493.92 both represent asthma-related diseases). Redundant descriptions are ignored for display purposes.

D. INTERPRETATION VIA ATTENTION MECHANISM: CASE STUDY

Interpretation is vital in the healthcare domain. In addition to predictive performance analysis, a case study with two high expenditure patients is conducted to interpret the model prediction. We first present the profiles of the selected patients, then we display the relative importance of the medical codes using a heatmap. As shown in Figure 6 (top), Patient A was an 8-year-old female who spent $8,589 in 2013. The predicted next year expenditure of this patient is $35,596, while the actual cost is $43,914. The cost-driving disease of patient A was *phenylketonuria (ICD9 270.1)* and *personality disorder (ICD9 301)*.

Figure 6 (bottom) shows the attention weights of medical codes at each time window. The attention weights represent the relative importance for predicting medical expenditures, where the darker color indicates a stronger correlation between the medical code and the predictive outcome. Among the displayed medical codes in Figure 6 (bottom), we observed that *phenylketonuria (ICD9 270.1)* has the highest attention weights across the different time periods. This finding is consistent with the medical literature as *phenylketonuria* is considered a severe disease that can lead to intellectual disability, seizures, and mental disorders [36]. It is also a chronic

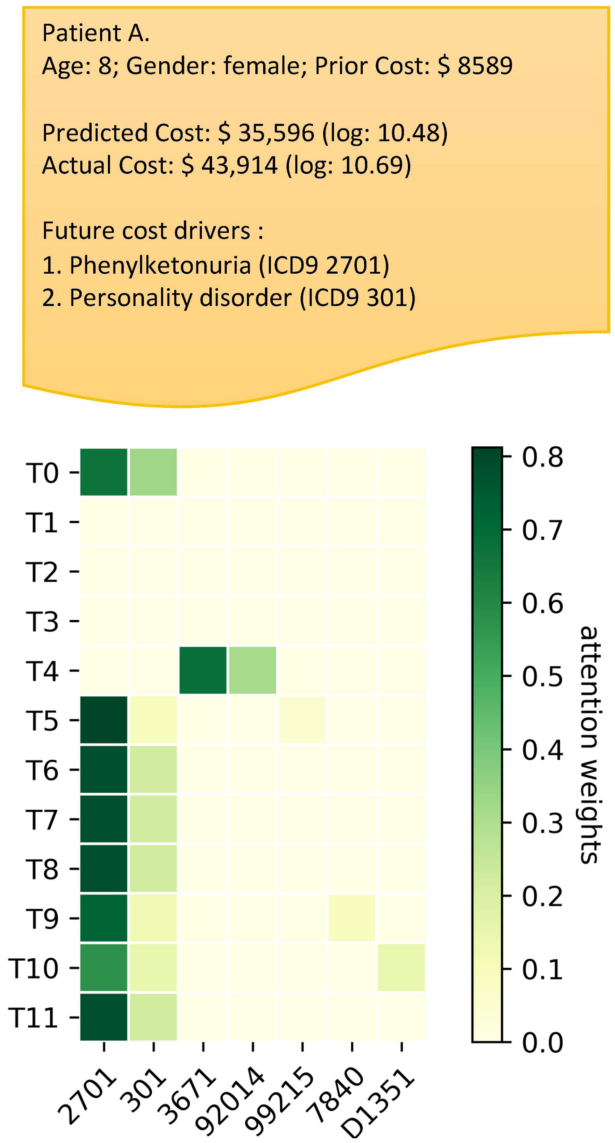

FIGURE 6. The profile (top) and attention weight heatmap (bottom) for Patient A.

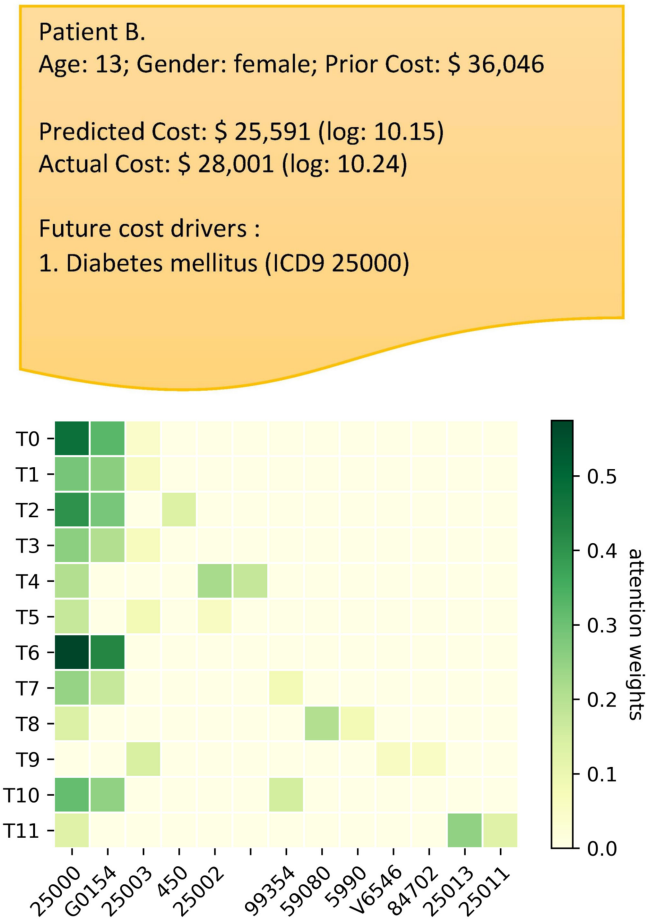

FIGURE 7. The profile (top) and attention weight heatmap (bottom) for Patient B.

condition that often lasts for years or even lifelong. Compared to *phenylketonuria (ICD9 270.1)* and *personality disorder (ICD9 301)*, the attention weights on *headache (ICD9 784.0)* and *dental sealant procedure (D1351)* are relatively small. These two codes are both related to acute diseases and thus less likely to influence future medical expenditure.

Figure 7 presents the profile and medical code attention weights of Patient B. As shown in Figure 7 (top), this patient was a 13-year-old female with a $36,046 annual expenditure in 2013. The predicted cost for Patient B is $25,591 and the actual cost is $ 28,001. Looking through medical records of patient B, we identify one driver disease: *Diabetes mellitus (ICD9 250.00)*. The relative importance of medical codes for predicting Patient B's future expenditure is depicted in Figure 7 (bottom). We can easily observe that *diabetes mellitus (ICD9 250.00)* and *Direct skilled nursing services (HCPCS G0154)* are the two significant medical codes. *Diabetes mellitus* is a chronic condition that requires consistent care intervention [37]. *G0154* is a valid HCPCS code for *Direct skilled nursing services of a licensed nurse in the home health or hospice setting.* We also observe many diabetes-related medical codes such as *ICD9 250.13*, *ICD9 250.11, ICD9 250.03*, and *ICD9 250.02.* All these medical codes are under the *Diabetes mellitus genre (ICD 250).*

E. HIGH UTILIZER SELECTION

One application of healthcare expenditure predictive models is patient stratification. Using the predicted value as the risk score, we can risk-stratify patients and provide care coordination to the high-risk individuals. In this subsection, we compared the stratification performance of our multi-view model with LASSO and GBM, the two baselines that show competitive performance on expenditure prediction. As shown in Table 5, patients selected by our approach are more likely to incur a higher medical expenditure in the following year. Specifically, the top 1% population (218 individuals in the testing data) selected by our model had a combined healthcare cost of approximately $3 million, which is $1 million higher

TABLE 5. The Result for Selecting Future High Utilizers

	Top 5% (n=1,094)			Top 3% (n=656)			Top 1% (n=218)		
	Lasso	GBM	Multi-view	Lasso	GBM	Multi-view	Lasso	GBM	Multi-view
Mean ($)	6,061	6,587	6,807	7,075	8,119	8,683	10,294	11,108	15,432
Total ($)	6,630,818	7,206,625	7,447,195	4,641,250	5,326,719	5,696,399	2,244,173	2,421,603	3,364,390
Diff ($)	816,377	240,570	--	1,055,149	369,680	--	1,120,217	942,787	--

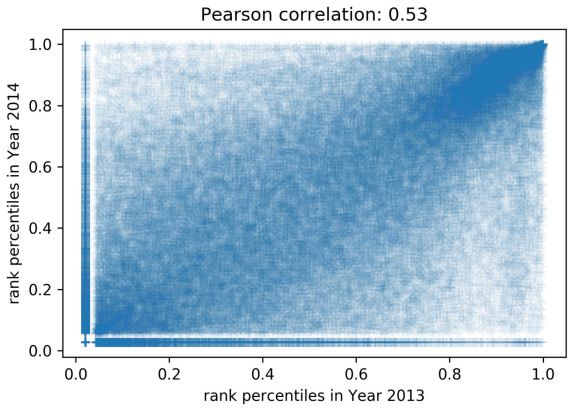

FIGURE 8. Scatter plot of expenditure percentiles between two time periods. The correlation score is 0.53, indicating that patient expenditures are positively correlated between year 2013 and year 2014. The upper right corner is denser, implying the high-cost patients are more temporally consistent.

than that of the patients identified by the LASSO model. This result implies that the ACO could potentially influence an additional $1 million in cost compared to LASSO.

V. DISCUSSION

The goal of this study was to develop an accurate model for expenditure prediction based on historical claims. Some prior studies [38] have suggested that healthcare expenditures may be episodic and not consistent, while some literature [12] shows that healthcare expenditures are significantly temporally correlated. The benefit of modeling prior claims for expenditure prediction depends on the degree of randomness in the healthcare utilization pattern. Thus, we examined the temporal correlation of annual expenditure at the individual level on our pediatric dataset.

Figure 8 presents the ranking percentile scatter plot of the patient expenditure in two time periods (i.e., year 2013 and year 2014). In Figure 8, we can observe a large portion of points spreading along the diagonal line. This spreading pattern indicates that patients are more likely to have a consistent healthcare expenditure. This observation is also confirmed by the Pearson Correlation (>0.5). Pearson Correlations range from -1 to 1 with 0 implying no correlation. The Pearson Correlation between two ranking percentiles is 0.53, indicating a strong positive correlation between prior expenditure and future expenditure (i.e., if prior expenditure increases, future expenditure will likely increase).

Our study developed a deep learning model that can incorporate the heterogeneous information within claims data as different views. Utilizing the embedding learning processes, our model eliminates the reliance on domain knowledge to handcraft medical codes into semantic similar categories. Through extensive analysis, we found that our model outperformed baselines on the expenditure prediction task and the high utilizer selection task. These findings highlight the potential of our model to provide better population healthcare care management.

This study has several limitations. First, the findings and conclusions were made based on the experimental results on a state Medicaid pediatric claims dataset. The results and observations may vary by state, insurer plan, or payer type. Second, we conducted the study under continuous eligibility across 2 years. This eligibility enforcement may have distorted the underlying population. Third, although the attention mechanism provides a certain level of interpretation for the predictive outcome, few actionable factors are provided for efficient intervention.

Our future work will try to address these limitations. We plan to extend our study scope by gathering additional data from different types of healthcare programs. We will also collaborate with physicians and domain experts to deploy the model into a real-world setting to provide better guidance for care management.

VI. CONCLUSION

This study proposes a multi-view deep learning framework to learn efficient and interpretable patient representation for medical expenditure prediction. Our approach leverages a feedforward neural network, an attention-based bidirectional recurrent neural network, and a hierarchical attention network to exploit heterogeneous information in claims data from different views. Experimental results show our approach outperforms various baselines on a real-world pediatric dataset for predicting medical expenditure in the following calendar year. We conducted a case study to interpret the learned feature importance, which can help understand the model's decisions. Moreover, we applied the learned model on a high utilizer selection task. The promising result implies that our model could enable care management entities to better allocate healthcare resources.

REFERENCES

[1] M. A. Morid, O. R. L. Sheng, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, "Healthcare cost prediction: Leveraging fine-grain temporal patterns," *J. Biomed. Inform.*, vol. 91, 2019, Art. no. 103113.

[2] A. S. Ash *et al.*, "Using diagnoses to describe populations and predict costs," *Health Care Financing Rev.*, vol. 21, pp. 7–28, 2000.

[3] 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," *Med. Care*, vol. 36, pp. 1108–1113, 1998.

[4] A. K. Rosen, S. A. Loveland, J. J. Anderson, C. S. Hankin, J. N. Breckenridge, and D. R. Berlowitz, "Diagnostic cost groups (DCGs) and concurrent utilization among patients with substance abuse disorders," *Health Serv. Res.*, vol. 37, pp. 1079–1103, 2002.

[5] D. Bertsimas *et al.*, "Algorithmic prediction of health-care costs," *Oper. Res.*, vol. 56, pp. 1382–1392, 2008.

[6] D. O. Clark, M. Von Korff, K. Saunders, W. M. Balugh, and G. E. Simon, "A chronic disease score with empirically derived weights," *Med. Care*, pp. 783–795, 1995, doi: 10.1097/00005650-199508000-00004

[7] M. Von Korff, E. H. Wagner, and K. Saunders, "A chronic disease score from automated pharmacy data," *J. Clin. Epidemiol.*, vol. 45, pp. 197–203, 1992.

[8] P. A. Fishman, M. J. Goodman, M. C. Hornbrook, R. T. Meenan, D. J. Bachman, M. C. O'Keeffe Rosetti, "Risk adjustment using automated ambulatory pharmacy data: The RxRisk model," *Med. Care*, vol. 41, pp. 84–99, 2003.

[9] J. P. Weiner, B. H. Starfield, D. M. Steinwachs, and L. M. Mumford, "Development and application of a population-oriented measure of ambulatory care case-mix," *Med. Care*, pp. 452–472, 1991, doi: 10.1097/00005650-199105000-00006

[10] Y. Zhao *et al.*, "Measuring population health risks using inpatient diagnoses and outpatient pharmacy data," *Health Serv. Res.*, vol. 36, pp. 180–193, 2001.

[11] Y. Zhao *et al.*, "Predicting pharmacy costs and other medical costs using diagnoses and drug claims," *Med. Care*, vol. 43, pp. 34–43, 2005.

[12] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, "Machine learning approaches for predicting high cost high need patient expenditures in health care," *Biomed. Eng. Online*, vol. 17, no. 1, pp. 1–20, 2018, doi: 10.1186/s12938-018-0568-3

[13] C.-Y. Kuo, L.-C. Yu, H.-C. Chen, and C.-L. Chan, "Comparison of models for the prediction of medical costs of spinal fusion in Taiwan diagnosis-related groups by machine learning algorithms," *Healthcare Inform. Res.*, vol. 24, no. 1, 2018, Art. no. 29.

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

[15] C. Wu, F. Wu, Y. Huang, and X. Xie, "NICE: Neural in-hospital cost estimation from medical records," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, doi: 10.1145/3357384.3358130

[16] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards heterogeneous temporal clinical event pattern discovery," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2012, doi: 10.1145/2339530.2339605

[17] J. Zhou, F. Wang, J. Hu, and J. Ye, "From micro to macro," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2014, doi: 10.1145/2623330.2623711

[18] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "`Deepr`: A convolutional net for medical records," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 22–30, Jan. 2017.

[19] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, and J. Sun, "Multi-layer representation learning for medical concepts," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1495–1504.

[20] C. Che, C. Xiao, J. Liang, B. Jin, J. Zho, and F. Wang, "An RNN architecture with dynamic temporal matching for personalized predictions of parkinson's disease," in *Proc. 2017 SIAM Int. Conf. Data Mining*, 2017, pp. 198–206.

[21] E. Choi *et al.*, "Learning the graphical structure of electronic health records with graph convolutional transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 606–613.

[22] B. L. P. Cheung and D. Dahl, "Deep learning from electronic medical records using attention-based cross-modal convolutional neural networks," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Inform. (BHI)*, Las Vegas, NV, USA, 2018, pp. 222–225.

[23] X. Zeng, Y. Feng, S. Moosavinasab, D. Lin, S. Lin, and C. Liu, "Multi-level self-attention model and its use on medical risk prediction," *Pacific Symp. Biocomput.*, vol. 25, pp. 115–126, 2020.

[24] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and G. J. Dipole, "Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 1903–1911.

[25] 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 *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3512–3520. 2016.

[26] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "KAME: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, New York, NY, USA: ACM, 2018, pp. 743–752.

[27] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 787–795.

[28] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2Vec: A personalized interpretable deep representation of the longitudinal electronic health record," *IEEE Access*, vol. 6, pp. 65333–65346, 2018, doi: 10.1109/access.2018.2875677

[29] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits Transl. Sci. Proc.*, vol. 2016, pp. 41–50, 2016.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Adv. Neural Inf. Process. Syst.*, vol. 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2013, pp. 3111–3119.

[31] S. Sushmita *et al.*, "Population cost prediction on public healthcare datasets," in *Proc. 5th Int. Conf. Digit. Health 2015 - DH '15*. 2015, doi: 10.1145/2750511.2750521

[32] B. Hartman, R. Owen, and Z. Gibbs, "Predicting high-cost health insurance members through boosted trees and oversampling: An application using the HCCI database," *North Amer. Actuarial J.*, pp. 1–9, 2020.

[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, New York, NY, USA: ACM, 2016, pp. 785–794.

[34] B. Lahiri and N. Agarwal, "Predicting healthcare expenditure increase for an individual from medicare data," in *Proc. ACM SIGKDD Workshop Health Inform.*, academia.edu., 2014. [Online]. Available: http://www.academia.edu/download/34334926/healthcare.pdf

[35] R. K. Modlin, *Clinical Methods: The History, Physical, and Laboratory Examinations*, vol. 2, Military Medicine, 1977, pp. 761–761, doi: 10.1093/milmed/142.10.761a

[36] C. C. Scriver, "A simple phenylalanine method for detecting phenylketonuria in large populations of newborn infants by R. Guthrie and A. Susi, *Pediatrics*, vol. 32, pp. 318–343, 1963," *Pediatrics*, vol. 102, pp. 236–237, 1998.

[37] M. I. Harris, *Diabetes in America*, 2nd ed. Diabetes Research and Clinical Practice, 1995, Art. no. 75.

[38] T. L. Johnson *et al.*, "For many patients who use large amounts of health care services, the need is intense yet temporary," *Health Affairs*, vol. 34, pp. 1312–1319, 2015.

[39] H. Cao, S. Bernard, L. Heutte, and R. Sabourin, "Dissimilarity-based representation for radiomics applications," Mar. 2018, *arXiv:1803.04460*.

[40] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 1863–1883, Oct. 2019.