IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 31, 2018, accepted September 10, 2018, date of publication October 12, 2018, date of current version November 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2875677

Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record

JINGHE ZHANG[1], (Member, IEEE), KAMRAN KOWSARI[1,2], (Member, IEEE), JAMES H. HARRISON, JR.,[3,4,5], JENNIFER M. LOBO[3,5], AND LAURA E. BARNES[1,2,5], (Member, IEEE)

[1]Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22904, USA
[2]Sensing Systems for Health Lab, University of Virginia, Charlottesville, VA 22904, USA
[3]Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22904, USA
[4]Division of Laboratory Medicine Department of Pathology, University of Virginia, Charlottesville, VA 22904, USA
[5]Data Science Institute, University of Virginia, Charlottesville, VA 22904, USA

Corresponding author: Laura E. Barnes (lb3dp@virginia.edu)

This work was supported by the Jeffress Trust Award in Interdisciplinary Science. Patient2Vec is shared as an open source tool at https://github.com/BarnesLab/Patient2Vec.

ABSTRACT The wide implementation of electronic health record (EHR) systems facilitates the collection of large-scale health data from real clinical settings. Despite the significant increase in adoption of EHR systems, these data remain largely unexplored, but present a rich data source for knowledge discovery from patient health histories in tasks, such as understanding disease correlations and predicting health outcomes. However, the heterogeneity, sparsity, noise, and bias in these data present many complex challenges. This complexity makes it difficult to translate potentially relevant information into machine learning algorithms. In this paper, we propose a computational framework, *Patient2Vec*, to learn an interpretable deep representation of longitudinal EHR data, which is personalized for each patient. To evaluate this approach, we apply it to the prediction of future hospitalizations using real EHR data and compare its predictive performance with baseline methods. *Patient2Vec* produces a vector space with meaningful structure, and it achieves an area under curve around 0.799, outperforming baseline methods. In the end, the learned feature importance can be visualized and interpreted at both the individual and population levels to bring clinical insights.

INDEX TERMS Attention mechanism, gated recurrent unit, hospitalization, longitudinal electronic health record, personalization, representation learning.

I. INTRODUCTION

Longitudinal EHR data resemble text documents from many perspectives. A text document consists of a sequence of sentences, and a sentence is a sequence of words. Similarly, the longitudinal health record of a patient consists of a sequence of visits, and there is a list of clinical events, including diagnoses, medications, and procedures, that occur during a visit. Considering these similarities, representation learning methods for text documents in Natural Language Processing (NLP) have great potential to be applied to longitudinal EHR data.

Deep neural networks have become very popular in the NLP field and have been very successful in many applications, such as machine translation, question answering, text classification, document summarization, language modeling, etc. [1]–[8]. These networks excel at complex language tasks because they are capable of identifying high-order relationships, the network structure can encode language structures, and they allow the learning of a hierarchical representation of the language, i.e., representations for tokens, phrases, and sentences, etc.

Among a variety of deep learning methods, Recurrent Neural Networks (RNNs) have shown their effectiveness in NLP tasks because they have the ability to capture sequential information [7]–[10] which is inherent in human language. Traditional neural networks assume that inputs are independent of each other, while an RNN computes the output based on the current input as well as the "memory" from the previous computation. Although vanilla RNNs are not good at capturing long-term dependencies, many variants have been proposed and validated that are effective in addressing this issue.

2169-3536 © 2018 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

In the medical domain, it is critical that analytical results are interpretable, so that they can be understood and validated by a human with expert knowledge and so that knowledge captured by analysis can be used for process improvement. Traditional deep neural networks have the disadvantage that they lack interpretability. A substantial amount of work is ongoing to make sense of the ''black box'', and the attention mechanism [11] is one of the more effective methods recently developed to make the output of these algorithms more interpretable.

Health care is undergoing unprecedented change, and there is a great potential and demand for personalized care strategies. Personalized medicine, also called precision medicine, has previously focused on optimizing therapy to better fit the genetic makeup of the patient or the disease (e.g., the genetic susceptibility of cancer to specific chemotherapy strategies). The availability of EHR data and advances in machine learning create the potential for another type of personalization of healthcare. This type of personalization has become ubiquitous in our daily life. For example, customers have come to expect personalized search on Google and personalized product recommendations on Amazon and Netflix, based on their charactersitics and previous experiences with the systems. Personalization of healthcare processes, based on a patient's phenotype (physical and medical characteristics) and healthcare experiences as documented in the health record, may also improve ''customer'' satisfaction and it has the additional potential to improve healthcare efficiency, lower costs, and yield better outcomes. We believe that representation learning methods can capture a personalized representation of the important heterogeneities in patients' phenotypes and medical histories at the population-level, and make these representations available to drive healthcare decisions and strategies.

This research is based on RNN models and the attention mechanism with the objective of learning a personalized, interpretable, and complete representation of patients' medical records. Our proposed framework is capable of learning a personalized representation for each patient from a sequence of clinical events. A hierarchical attention mechanism learns personalized weights of clinical events, including hospital visits and the procedures that they contain. These weights allow us to interpret the relative importance and roles of clinical events in the learned representations both at individual and population levels. The ultimate goal is more accurate prediction and better insight into the critical elements of healthcare processes that can be used to improve healthcare delivery.

The rest of this paper is organized as follows: Section II summarizes the variants of RNNs and the attention mechanism, as well as their application to EHR data. Section III presents an overview of the proposed *Patient2Vec* representation learning framework, and Section IV elaborates the details of the algorithms. In Section V, the proposed framework is evaluated for a prediction task and we compare its performance with other baseline methods. In addition to prediction performance, we further interpret the learned representations with visualizations on example patients and events. Finally, Section V provides a summary of this work.

II. RELATED WORK

In this section, we present an overview of a gated recurrent unit, a type of RNN, which is capable of capturing long-term dependencies. Then we briefly introduce attention mechanisms in neural networks that allow the network to attend to certain regions of data, which is inspired by the visual attention mechanism in humans. Additionally, we summarize the RNN networks and attention mechanisms previously used to mine EHR data.

A. RECURRENT NEURAL NETWORKS (RNN)

RNNs are expected to learn long-term dependencies by taking the previous state and the new input in the computation at the current time step t. However, vanilla RNNs are incapable of capturing the dependencies when the sequence is very long due to the vanishing gradient problem [12]. Many variants of the RNN network have been proposed to address this issue, and long short term memory (LSTM) is one of the most popular models used nowadays in NLP tasks [7], [8], [13]–[16].

1) GATED RECURRENT UNIT (GRU)

GRU is a simplified version of LSTM [7]. The basic idea of GRU is to combat the vanishing gradient problem with a gating mechanism. Hence the general recurrent structure in GRU is identical to vanilla RNNs except that a GRU unit is used in the computation at each time step rather than a traditional simple recurrent unit.

In general, a GRU cell has two gates, i.e., a reset gate r and an update gate z. The reset gate is used to determine how to integrate the previous state into the computation of the current state, while the update gate determines how much the unit updates its activation.

Given the input $\mathbf{x}_t$ at time step t, the reset gate r_t is computed as presented in Equation 1

$$r_t = \sigma(\mathbf{U}_r \mathbf{x}_t + \mathbf{W}_r \mathbf{s}_{t-1}) \tag{1}$$

where $\mathbf{U}_r$ and $\mathbf{W}_r$ are the weight matrices of the reset gate and $\mathbf{s}_{t-1}$ is the hidden activation at time step $t-1$. A similar computation is performed for the update gate z_t at time step t, shown in Equation 2

$$z_t = \sigma(\mathbf{U}_z \mathbf{x}_t + \mathbf{W}_z \mathbf{s}_{t-1}) \tag{2}$$

where $\mathbf{U}_z$ and $\mathbf{W}_z$ are the weight matrices of update gate. The current hidden activation $\mathbf{h}_t$ is computed by

$$\mathbf{h}_t = (1 - z_t)\mathbf{h}_{t-1} + z_t \tilde{\mathbf{h}}_t \tag{3}$$

where $\tilde{\mathbf{h}}_t$ is the candidate activation at time step t. The computation of $\tilde{\mathbf{h}}_t$ is presented in Equation 4

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(r_t \odot \mathbf{h}_{t-1})) \tag{4}$$

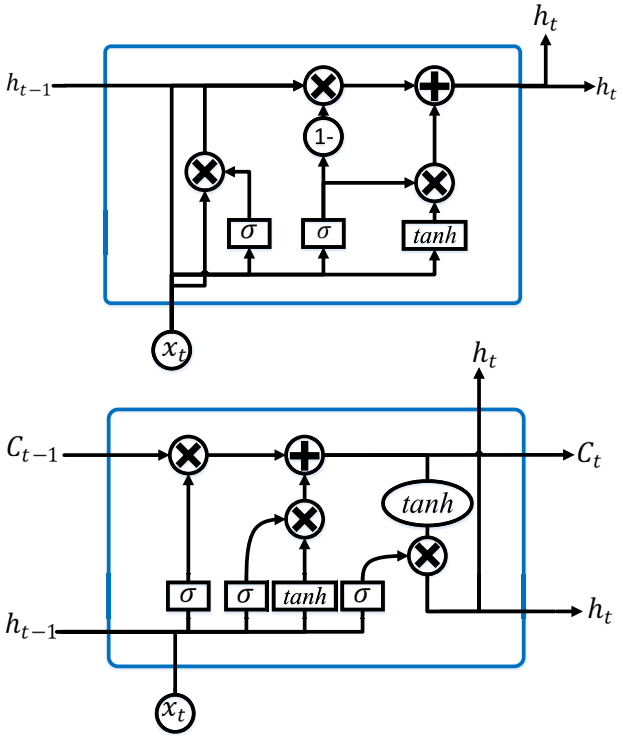

FIGURE 1. The top figure is a GRU gating unit and bottom figure shows an LSTM unit [7].

where $\mathbf{U}$ and $\mathbf{W}$ are weight matrices and $\odot$ represents element-wise multiplication. Figure 1 presents a graphical illustration of the GRU [7] and one unit of LSTM.

GRU is capable of learning long-term dependencies [17] due to the additive component of update from t to $t+1$ in the gating mechanism. Consequently, important features will be carried forward in the input stream while irrelevant information will be dropped. When the reset gate is 0, the network is forced to drop previous states and reset with current information. Moreover, the method provides shortcuts such that the error is easily backpropagated without vanishing too quickly [5], [18]. Hence, the GRU is well-suited to learn long-term dependencies in sequence data.

2) LONG SHORT-TERM MEMORY (LSTM)

An LSTM unit is similar to a GRU, but with one more gate in an LSTM unit (as shown in Figure 1). LSTM also preserves long term dependencies more effectively than basic RNN. This is particularly useful to overcome the vanishing gradient problem [19]. Although LSTM has a chain-like structure similar to RNN, LSTM uses multiple gates to carefully regulate the amount of information that will be allowed into each node state. Figure 1 shows the basic cell of an LSTM model. A step by step explanation of an LSTM cell is as following:
Input gate:

$$i_t = \sigma(\mathbf{W}_i[\mathbf{x}_t, \mathbf{h}_{t-1}] + b_i), \tag{5}$$

Candid memory cell value:

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{W}_c[\mathbf{x}_t, \mathbf{h}_{t-1}] + b_c), \tag{6}$$

Forget gate activation:

$$f_t = \sigma(\mathbf{W}_f[\mathbf{x}_t, \mathbf{h}_{t-1}] + b_f), \tag{7}$$

New memory cell value:

$$\mathbf{C}_t = i_t * \tilde{\mathbf{C}}_t + f_t \mathbf{C}_{t-1}, \tag{8}$$

Output gate value:

$$o_t = \sigma(\mathbf{W}_o[\mathbf{x}_t, \mathbf{h}_{t-1}] + b_o), \tag{9}$$

$$\mathbf{h}_t = o_t \tanh(\mathbf{C}_t), \tag{10}$$

In the above description all $\mathbf{b}$ represent bias vectors, all $\mathbf{W}$ represent weight matrices, and $\mathbf{x}_t$ is used as input to the memory cell at time t. Also,the i, c, f, o indices refer to input, cell memory, forget and output gates respectively. An RNN can be biased when later words are more influential than the earlier ones.

Empirically, LSTM and GRU achieve comparable performance in many tasks but there are fewer parameters in a GRU, which makes it a little faster to learn and able to generalize with fewer data [20].

B. ATTENTION MECHANISM

Attention mechanisms, inspired by the visual attention system found in humans, have become popular in deep learning. Attention allows the network to focus on certain regions of data, while perceiving other regions with "low resolution". In addition to higher accuracy, it also facilitates the interpretation of learned representations. We elaborate an attention mechanism on an RNN network, and Figure 2 presents a graphical illustration.

According to Figure 2, a variable-length weight vector α is learned based on hidden states [11]. Then a global context vector is computed based on weights α and all the hidden states to create the final output. Equation 11 presents the computation of the weight vector $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_T\}$, where T is the length of the sequence

$$\alpha_1, \alpha_2, \cdots, \alpha_T = f(\mathbf{W}_\alpha \mathbf{h} + b_\alpha) \tag{11}$$

and where f is a nonlinear activation function, usually *softmax* or tanh. Then, the context vector c is constructed as:

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t \tag{12}$$

Thus, the network puts more attention on the important features for the final prediction which can improve the model performance. An additional benefit is that the weights can be utilized to understand the importance of features such that the models are more interpretable. The attention mechanism has been introduced to both Convolutional Neural Networks (CNNs) and RNNs for various tasks and has achieved many successes in the fields of computer vision and NLP [11], [21], [22].

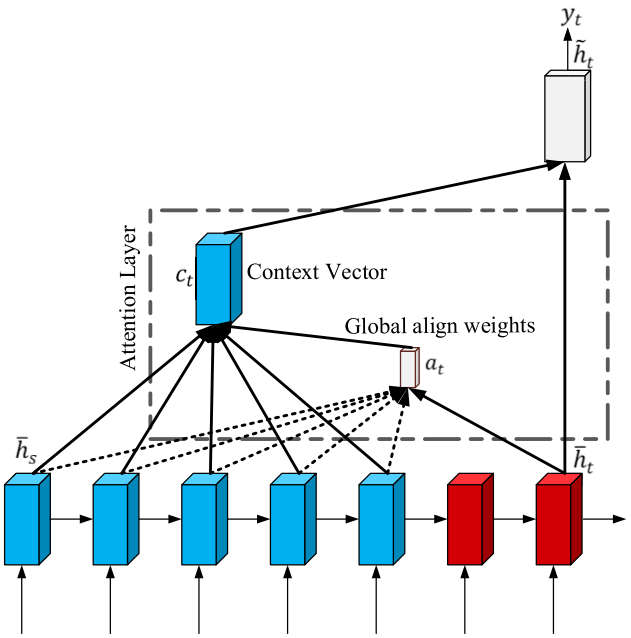

FIGURE 2. The global attention model.

C. DEEP LEARNING IN EHR DATA

Previous studies on EHR data mainly use statistical methods or traditional machine learning techniques. Recently researchers have started adapting deep learning approaches to this data [23], [24], including textual notes, temporal measurements of laboratory testing in the Intensive Care Unit (ICU), and longitudinal data in patient populations. Here, we summarize deep learning research in mining EHR data and focus on the studies using RNN-based models.

Hospitalized patients, especially patients in ICUs, are continuously monitored for cardiac, respiratory, and other physical functions, creating a large volume of sequential data in multiple dimensions. These measurements are utilized by physicians to make diagnostic and treatment decisions. The functions monitored may change over time and monitoring may be irregular, based on a patient's condition. It is very challenging for traditional machine learning methods to mine this multivariate time series data considering missing values, varying length, and irregular, non-simultaneous sampling. *Lipton et al.* [25] trained an LSTM with a replicated target to learn from these sequence data and used this model to make predictions of diagnoses. The data used in this research are time series of clinical measurements with continuous values, and the LSTM models outperformed logistic regression and MLP. *Che et al.* [26] developed a GRU-based model to address missing values in multivariate time series data, in which the missing patterns are incorporated for improved prediction performance. This work has been applied to the Medical Information Mart for Intensive Care III (MIMIC-III) clinical database to demonstrate its effectiveness in mining time series of clinical measurements with missing values [27]. Longitudinal EHR data including clinical events, such as diagnoses, medications, and procedures is also a potentially rich resource for predictive modeling. *Choi et al.* [28] analyze this data with a GRU network to forecast future clinical events, and it achieves a better prediction performance than comparison models such as logistic regression and MLP.

Difficulty in interpreting model behavior is one of the major drawbacks of using deep learning to mine EHR data. Some attempts have been made to address this issue. *Che et al.* [29] propose an interpretable mimic learning method which trains a mimic gradient boosting trees model to utilize predicted labels or features learned by deep learning models for final prediction [30]. Then the feature importances learned by the tree-based models are used for knowledge discovery. Attention mechanisms have been introduced recently to improve the interpretability of the prediction results of deep learning models in health analytics. *Choi et al.* [31] develop an interpretable model with two levels of attention weights learned from two reverse-time GRU models, respectively. The experimental results on EHR data indicate comparable prediction performance with conventional GRU models but more interpretable results. Our work continues the attempt to use attention mechanisms to improve the interpretability of RNN-based models.

III. *Patient2Vec* SYSTEM MODEL

In this section, we provide an overview of the proposed hierarchical representation learning framework. This framework uses deep recurrent neural networks to capture the complex relationships between clinical events in the patient's EHR data and employs the attention mechanism to learn a personalized representation and to obtain relative feature importance. The proposed representation learning framework contains four steps and is presented graphically in Figure 3.

A. LEARNING VECTOR REPRESENTATIONS OF MEDICAL CODES

EHR data consists primarily of records of outpatient and inpatient visits to healthcare providers. These visit records include multiple clinical codes for diagnoses, symptoms, procedures, therapies, and other observations and events that occurred during the visit. Here, we treat the set of medical codes associated with a visit as a sentence consisting of words, except that there is no ordering in the words. Thus, we adopt the word2vec approach to construct a vector to represent each medical code.

B. LEARNING WITHIN-SUBSEQUENCE SELF-ATTENTION

Clinical visits are represented as the set of vectors for the codes associated with the visit. Because closely-spaced visits are usually related clinically, we employ a time window to split the sequence of visits into multiple subsequences of equal length. A subsequence might contain multiple visits if they occurred within the same time window, or there might be no visits during a particular time window yielding an empty subsequence. Thus we transform the original sequence

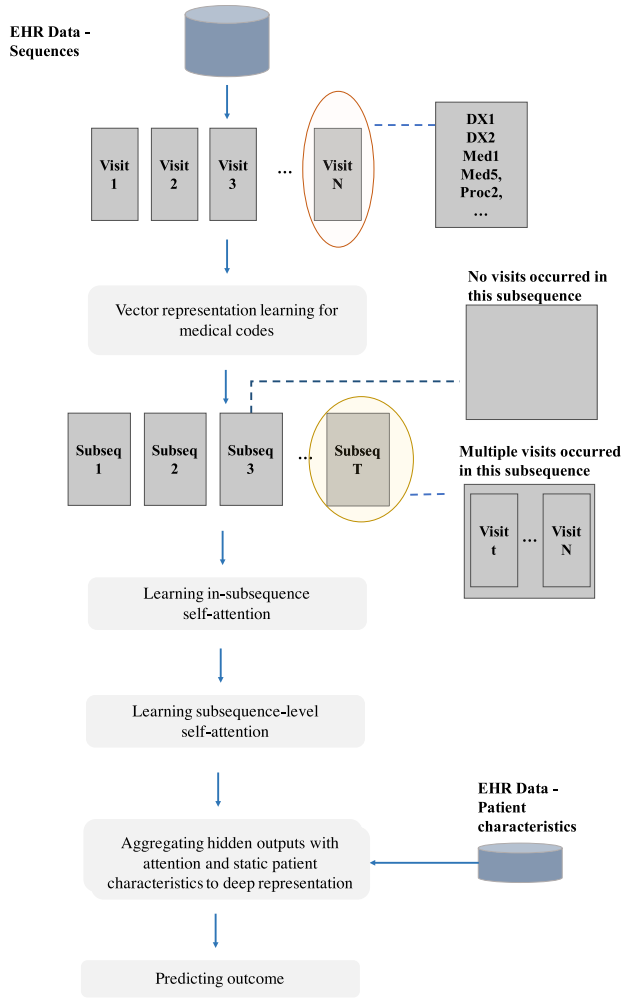

FIGURE 3. The *Patient2Vec* representation learning framework.

of irregularly-spaced visits into a sequence of subsequences with equal intervals, which is preferable for recurrent neural networks. The width of the subsequence window defines the time granularity of the method and its optimal width is related to the acuity (i.e., stability) of the clinical characteristics involved in the predication task. In future work it may be possible to define the relationship between clinical acuity and optimal subsequence width, or develop methods for learning an optimal width for a defined prediction task.

Because all medical events occurring within a subsequence are unlikely to contribute equally to the prediction of the target outcome, we cannot aggregate them with equal weights. Instead, we employ a self-attention mechanism which trains the network to learn the weights.

C. LEARNING SUBSEQUENCE-LEVEL SELF-ATTENTION

Given a sequence of subsequences with embedded medical codes, we are able to input it into a recurrent neural network to capture the temporal dependencies between events. However, the subsequences of visits are not contributing equally to the outcome. Hence, we employ another level of attention to learn the weights of the subsequences by the network itself for the outcome prediction.

D. CONSTRUCTING AGGREGATED DEEP REPRESENTATION

Given the learned weights and hidden outputs, we aggregate them into one universal vector for a comprehensive representation. In this step, the static information, such as age, gender, previous hospitalization history is added as extra features, to get a complete representation of a patient.

E. PREDICTING OUTCOME

Given the complete vector representation of a patient's EHR data, we add a logistic regression layer at the end for the prediction of outcome.

IV. PATIENT2VEC REPRESENTATION LEARNING ALGORITHM

In this section, we present the details of the proposed representation learning framework, which is based on a GRU network and a hierarchical attention mechanism. Figure 4 presents the structure of the proposed network with attention.

The proposed framework consists of five parts presented in the following: I) Learning vector representations of medical codes, II) Learning within-subsequence self-attention, III) Learning subsequence-level self-attention, IV) Constructing aggregated deep representation, V) Predicting outcome.

A. LEARNING VECTOR REPRESENTATIONS OF MEDICAL CODES

Given a patient's raw EHR data, a sequence of visits, we observe that a visit usually contains multiple medical codes. Hence, it is feasible to learn a vector to represent the medical code by capturing the relationships between the codes. In this work, we employ the classical word2vec algorithm, skip-gram. The basic idea of skip-gram is to learn a vector to represent each word such that the probability of the context to predict based on the target word is maximized. Hence, the vectors of similar words are close to each other in the learned feature space. In the skip-gram model, the vectors are learned by training a shallow neural network to predict the context words given an input word. Similarly, in our problem, the input is a medical code and the target to predict are the medical codes occurred in the same visit.

Hence, each subsequence is a matrix consisting of the vectors of medical codes occurred during this associated time window.

B. LEARNING WITHIN-SUBSEQUENCE SELF-ATTENTION

Given a sequence of subsequences encoded by vectors of medical codes, this step employs the within-subsequence attention which allows the network itself to learn the weights of vectors in the subsequence according to its contribution to the prediction target.

Here, we denote the sequence of patient i as $\mathbf{s}^{(i)}$, and $\mathbf{v}_t^{(i)}$ denotes the tth subsequence in sequence $\mathbf{s}^{(i)}$,

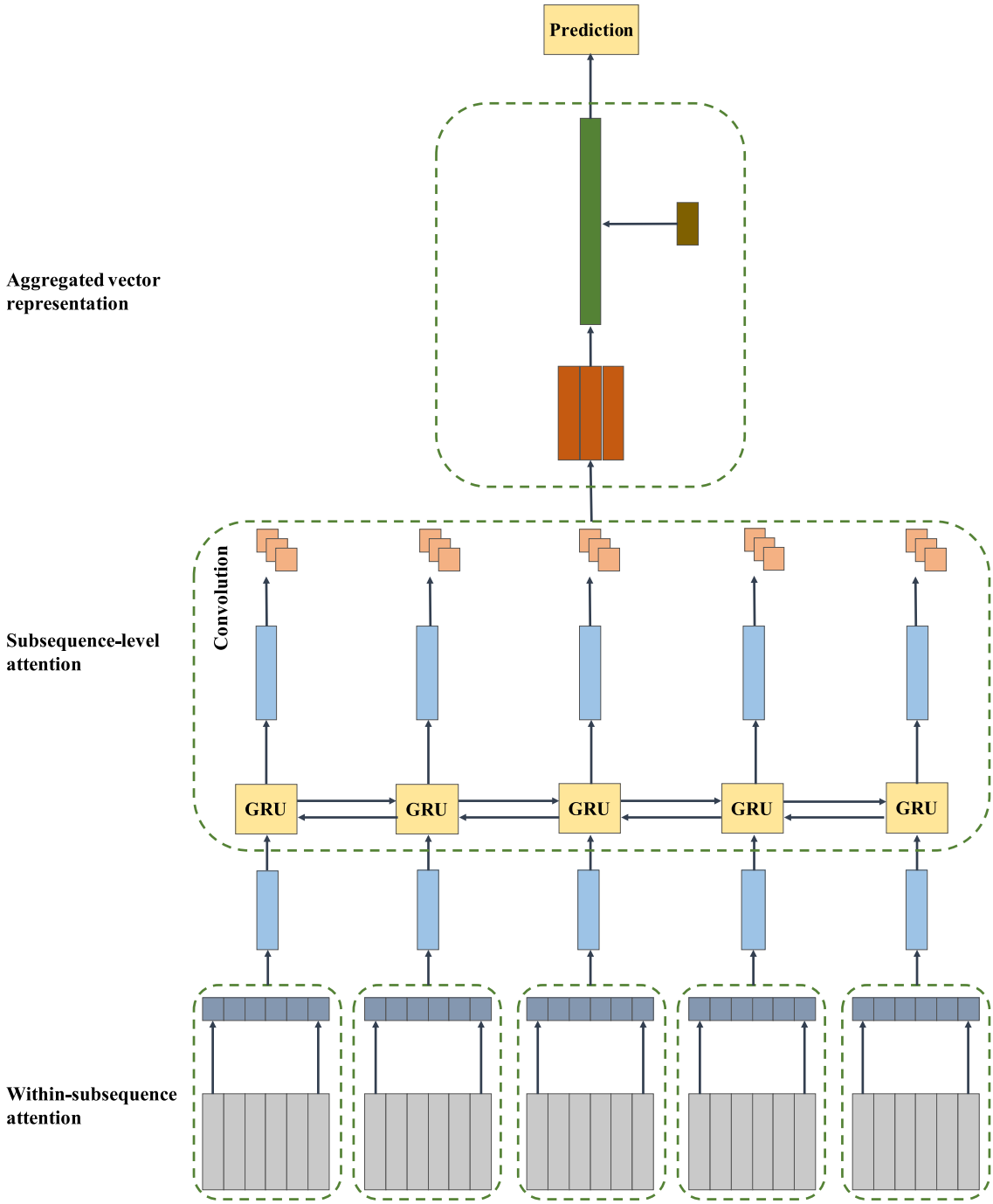

FIGURE 4. A graphical illustration of the network in the *Patient2Vec* representation learning framework.

where $t \in \{1, 2, \cdots, T\}$. Thus, $\mathbf{s}^{(i)} = \{\mathbf{v}_1^{(i)}, \cdots, \mathbf{v}_t^{(i)}, \cdots, \mathbf{v}_T^{(i)}\}$. To simplify the notation, we omit i in the following explanation. Subsequence $\mathbf{v}_t \in \mathbb{R}^{n \times d}$ is a matrix of medical codes such that $\mathbf{v}_t = \{v_{t_1}, v_{t_2}, \cdots, v_{t_j}, \cdots, v_{t_n}\}$, where $v_{t_j} \in \mathbb{R}^d$ is the vector representation of the jth medical code in the tth subsequence $\mathbf{v}_t$ and there are n medical codes in a subsequence. In real EHR data, it is very likely that the numbers of medical codes in each visit or time window are different, thus, we utilize the padding approach to obtain a consistent matrix dimensionality in the network.

To assign attention weights, we utilize the one-side convolution operation with a filter $\omega^\alpha \in \mathbb{R}^d$ and a nonlinear activation function. Thus, the weight vector α_t is generated for medical codes in the subsequence $\mathbf{v}_t$, presented in Equation 13.

$$\alpha_t = \tanh(Conv(\omega^\alpha, \mathbf{v}_t)) \tag{13}$$

where $\alpha_t = \{\alpha_{t_1}, \alpha_{t_2}, \cdots, \alpha_{t_n}\}$, and $\omega^\alpha \in \mathbb{R}^d$ is the weight vector of the filter. The convolution operation *Conv* is

presented in Equation 14.

$$\tilde{\alpha}_{t_j} = (\omega^{\alpha})^{\mathsf{T}} \mathbf{v}_{t_j} + b^{\alpha} \tag{14}$$

where b^{α} is a bias term. Then, given the original matrix $\mathbf{v}_t$ and the learned weights α_t, an aggregated vector $\mathbf{x}_t \in \mathbb{R}^d$ is constructed to represent the tth subsequence, presented in 15.

$$\mathbf{x}_t = \sum_{j=1}^{n} \alpha_{t_j} v_{t_j} \tag{15}$$

Given Equation 15, we obtain a sequence of vectors, $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T\}$, to represent a patient's medical history.

C. LEARNING SUBSEQUENCE-LEVEL SELF-ATTENTION

Given a sequence of embedded subsequences, this step employs the subsequence-level attention which allows the network itself to learn the weights of subsequences according to their contribution to the prediction target.

To capture the longitudinal dependencies, we utilize a bidirectional GRU-based RNN, presented in Equations 16.

$$\mathbf{h}_1, \cdots, \mathbf{h}_t, \cdots, \mathbf{h}_T = GRU(\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T) \tag{16}$$

where $\mathbf{h}_t \in \mathbb{R}^k$ represents the output by the GRU unit at the tth subsequence. Then, we introduce a set of linear and softmax layers to generate M hops of weights $\beta \in \mathbb{R}^{M \times T}$ for subsequences. Then, for the hop m

$$\gamma_{mt} = (\mathbf{w}_m^{\beta})^{\mathsf{T}} \mathbf{h}_t + b^{\beta} \tag{17}$$

$$\beta_{m_1}, \cdots, \beta_{m_T} = soft\,max(\gamma_{m_1}, \cdots, \gamma_{m_t}, \cdots, \gamma_{m_T}) \tag{18}$$

where $\mathbf{w}_m^{\beta} \in \mathbb{R}^k$. Thus, with the subsequence-level weights and hidden outputs, we construct a vector $\mathbf{c}_m \in \mathbb{R}^k$ to represent a patient's medical visit history with one hop of subsequence weights, presented in the following Equation 19.

$$\mathbf{c}_m = \sum_{t=1}^{T} \beta_{mt} \mathbf{h}_t \tag{19}$$

Then, a context vector $\mathbf{c} \in \mathbb{R}^{M \times k}$ is constructed by concatenating $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_M$.

D. CONSTRUCTING AGGREGATED DEEP REPRESENTATION

Given the context vector $\mathbf{c}$, this step integrates the patients characteristics $\mathbf{a} \in \mathbb{R}^q$ into the context vector for a complete vector representation of the patient's EHR data. In this research, the patient characteristics include demographic information and some static medical conditions, such as age, gender, and previous hospitalization. Thus, an aggregated vector is constructed, $\mathbf{c}' \in \mathbb{R}^{M \times k+q}$, by adding $\mathbf{a}$ as additional dimensions to the context vector $\mathbf{c}$.

E. PREDICTING OUTCOME

Given the vector representation of the complete medical history and characteristics of patients, $\mathbf{c}'$, we add a linear and a softmax layer for the final outcome prediction, as presented in Equation 20.

$$\hat{y} = softmax(\mathbf{w}^{c\mathsf{T}} \mathbf{c}' + b^{c}) \tag{20}$$

To train the network, we use cross-entropy as the loss function, presented in Equation 21.

$$L = -\frac{1}{N} \sum_{n=1}^{N} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) + \frac{1}{N} \sum_{n=1}^{N} ||\beta\beta^{\mathsf{T}} - \mathbf{I}||_F^2 \tag{21}$$

where N is the total number of observations. Here, y_i is a binary variable in classification problems, while model output $\hat{y}_i$ is real-valued. The second term in Equation 21 is to penalize redundancy if the attention mechanism provides similar subsequence weights for different hops of attention, which is derived from [32]. This penalty term encourages the multiple hops to focus on diverse areas and each hop focuses on a small area.

Thus, we obtain a final output for the prediction of outcomes and a complete personalized vector representation of the patient's longitudinal EHR data.

V. EVALUATION

A. BACKGROUND

Although health care spending has been a relatively stable share of the Gross Domestic Product (GDP) in the United States since 2009, the costs of hospitalization, the largest single component of health care expenditures, increased by 4.1% in 2014 [33]. Unplanned hospitalization is also distressing and can increase the risk of related adverse events, such as hospital-acquired infections and falls [34], [35]. Approximately 40% hospitalizations in the United Kingdom are unplanned and are potentially avoidable [36]. One important form of unplanned hospitalization is hospital re-admissions within 30 days of discharge, which is financially penalized in the United States. Early interventions targeted to patients at risk of hospitalization could help avoid unplanned admissions, reduce inpatient health care cost and financial penalties for providers, and reduce emergency department congestion [37].

In this research, we apply our proposed representation learning framework to the risk prediction of future hospitalization. Many studies have been conducted by researchers to predict the risk of 30-day readmission, or the admission risk of a particular population, such as patients with Ambulatory Care Sensitive Conditions (ACSCs), patients with heart failure, etc. [38]–[41]. Here, we focus on the general population and the objective is to predict the risk of all-cause hospitalization using longitudinal EHR data.

B. EXPERIMENTAL DESIGN

In this research, we use de-identified EHR data from the University of Virginia Health System covering 75 months beginning in September 2010. This dataset contains 2,343,651

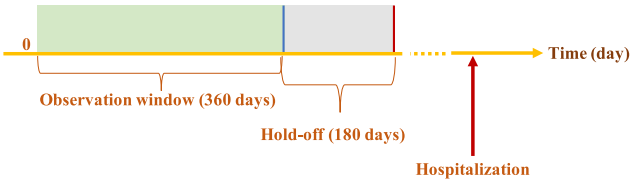

FIGURE 5. A graphical illustration of the experimental setting for the risk prediction of hospitalization.

inpatient and outpatient visits of 473,915 distinct patients. We extracted visit data with diagnosis, medication, and procedure codes.

We defined the observation window and prediction period to validate the proposed method. We first extract all patients with a medical record of at least 1.5 years, where the first year is the observation window and the medical records in this time window are used for feature construction. The following 6 months is the hold-off period for the purpose of early detection. For the positive class, we take all patients who have hospitalization after the first 1.5 years in their medical history, while the negative class consists of patients who have no hospitalization after 1.5 years. To better illustrate the experimental setting, we present the observation window, hold-off and onset of outcome event in Figure 5. Here, the medical codes include diagnosis, medication, and procedure codes, and a vector representation is learned for each code. In this dataset, diagnoses are primarily coded in ICD-9 and a small portion is ICD-10 codes, while procedures are mainly using CPT codes with a few ICD-9 procedure codes. The codes of medications are using the pharmaceutical categories. Overall, there are 94 distinct medication categories, 34,419 distinct diagnoses codes, and 7,895 distinct procedure codes in the EHR data. The dimension of the learned vectors of medical codes is set to 100. Medical codes that appear in less than 50 patients medical records are excluded as rare events.

To construct the subsequences of medical codes, we use l days as the time window. Figure 6 presents the cumulative histogram and density plot of the numbers of visits in the observation window, and we observe that the majority of patients have a small number of visits during the observation window (less than 25% of patients have more than 4 visits). Thus, we set l to 90 days, which split the observation window into 4 subsequences.

Within each subsequence, the number of distinct medical codes were computed and patients with more medical codes in a subsequence than the 95% quantile were excluded from the dataset. Overall, there are 8,841 and 89,101 patients in the target and control groups, respectively. Each group is randomly split into training, validation and testing sets with a 7:1:2 ratio. Thus, 70% are used for training, another 20% is used for testing, and the rest 10% are used for parameter tuning and early stopping. The stochastic gradient descent algorithm is used in training to minimize the cross-entropy loss function, shown in Equation 21.

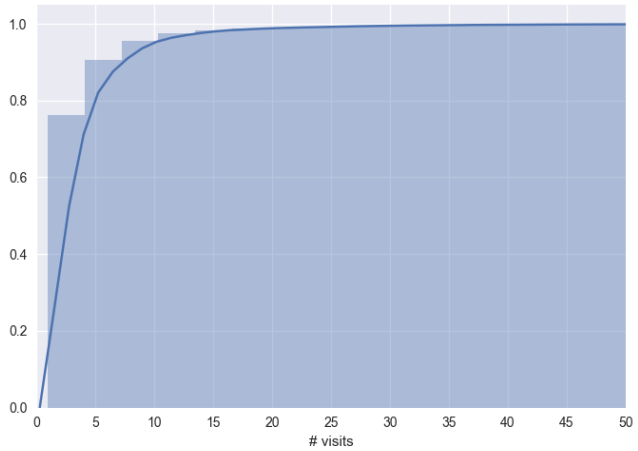

FIGURE 6. The cumulative histogram and density plot of patients' numbers of visits.

To evaluate the proposed representation learning framework, we compare the prediction performance of the proposed model with baseline approaches as follows.

1) LOGISTIC REGRESSION (LR)

The inputs are the aggregated counts of grouped medical codes over the entire observation window. Since the dimensionality of raw medical codes is huge, AHRQ clinical classifications of diagnoses and procedures are used to achieve a more general clustering of medical codes [42]. The medication codes are the pharmaceutical classes. Furthermore, patient characteristics and previous inpatient visit are also considered, where age and gender are demographic information, and a binary indicator is utilized to represent the presence of the previous hospitalization. Hence, the input is a 436-dimensional vector representing a patient's medical history and characteristics.

2) MULTI-LAYER PERCEPTRON (MLP)

A multi-layer perceptron is trained to predict hospitalization using the same inputs for logistic regression. Here, we use a one hidden layer MLP with 256 hidden nodes.

3) FORWARD RNN WITH MEDICAL GROUP EMBEDDING (FRNN-MGE)

We split the sequence into subsequences with equal interval l. The input at each step is the counts of medical groups within the associated time interval, and the patient characteristics are appended as additional features in the final logistic regression step. Here, the RNN is a forward GRU (or LSTM [18]) with one hidden layer and the size of the hidden layer is 256.

4) BIDIRECTIONAL RNN WITH MEDICAL GROUP EMBEDDING (BIRNN-MGE)

The inputs used for this baseline is the same as the one for the FRNN-MGE [15]. The RNN used here is a bidirectional GRU with one hidden layer and the size of the hidden layer is 256.

TABLE 1. The predictive performance of baselines and the proposed *Patient2Vec* framework.

Methods	Sensitivity	Specificity	AUC	F2 score
LR	0.637 ± 0.010	0.728 ± 0.003	0.721 ± 0.006	0.434 ± 0.006
MLP	0.727 ± 0.013	0.617 ± 0.004	0.713 ± 0.007	0.423 ± 0.007
RETAIN	0.553 ± 0.012	0.710 ± 0.003	0.663 ± 0.007	0.370 ± 0.008
FRNN-MGE	0.636 ± 0.012	0.739 ± 0.004	0.759 ± 0.006	0.438 ± 0.009
BiRNN-MGE	0.600 ± 0.012	$\mathbf{0.777 \pm 0.003}$	0.768 ± 0.007	0.439 ± 0.009
FRNN-MVE	0.753 ± 0.011	0.676 ± 0.004	0.785 ± 0.006	0.470 ± 0.008
BiRNN-MVE	0.724 ± 0.010	0.707 ± 0.003	0.788 ± 0.005	0.473 ± 0.008
Patient2Vec	$\mathbf{0.769 \pm 0.010}$	0.694 ± 0.004	$\mathbf{0.799 \pm 0.005}$	$\mathbf{0.492 \pm 0.007}$

5) FORWARD RNN WITH MEDICAL VECTOR EMBEDDING (FRNN-MVE)

We split the sequence into subsequences with equal interval l. The input at each step is the vector representation of the medical codes within the associated time interval, and the patient characteristics are appended as additional features in the final logistic regression step. Here, the RNN is a forward GRU (or LSTM [28]) with one hidden layer and the size of the hidden layer is 256.

6) BIDIRECTIONAL RNN WITH MEDICAL VECTOR EMBEDDING (BIRNN-MVE)

The inputs used for this baseline is the same as the one for the FRNN-MVE [25]. The RNN used here is a bidirectional GRU or LSTM [15] with one hidden layer and the size of the hidden layer is 256.

7) RETAIN

This model uses reverse time attention mechanism on RNNs for an interpretable representation of patient's EHR data [31]. The inputs are the same as the one for FRNN-MGE, which takes the counts of medical grouping within each time interval to construct features. Similarly, the two RNNs used for generating weights are GRU-based and the size of the hidden layers are 256.

8) *Patient2Vec*

The inputs are the same as that for FRNN-MVE. One filter is used when generating weights for within-subsequence attention, and three filters are used for subsequence-level attention. Similarly, the RNN used here is GRU-based and there is one hidden layer and the size of the hidden layer is 256.

The inputs of all baselines and *Patient2Vec* are normalized to have zero mean and unit variance. We model the risk of hospitalization based on *Patient2Vec* and baseline representations of patients' medical histories, and the model performance is evaluated with Area Under Curve(AUC), sensitivity, specificity, and F2-score. The validation set is used for parameter tuning and early stopping in the training process. Each experiment is repeated 20 times and we calculate the averages and standard deviations of the above metrics, respectively.

C. EXPERIMENTAL RESULTS

The predictive performance of *Patient2Vec* and baselines are presented in Table 1. The results shown here for the RNN-based models are based on time interval $l = 90$ days to construct subsequences.

According to Table 1, the RNN-based models are generally capable of achieving higher prediction performance in terms of sensitivity, AUC and F2 score, except for the RNN models based on medical group embedding which have lower sensitivity. Among all RNN-based approaches, the ones based on vector embedding outperform those based on medical group embedding in terms of sensitivity, AUC, and F2 score. The bidirectional RNN models generally have higher specificity but lower sensitivity than the forward RNN models, while the bidirectional ones have comparable AUC and F2 score with the forward ones, respectively. Generally, the proposed *Patient2Vec* framework outperforms the baseline methods, especially in terms of sensitivity and F2 score.

D. VISUALIZATION & INTERPRETATION

In addition to predictive performance, we interpret the learned representation by understanding the relative importance of clinical events in a patient's EHR data. Considering the feature importance learned by *Patient2Vec* are personalized for an individual patient, we illustrate it with two example patients. Figures 8 and 9 present the profiles of two individuals, Patient A and Patient B, respectively. To facilitate the interpretation, instead of using raw medical codes, we present the clinical groups from the AHRQ clinical classification software on diagnoses and procedure codes, as well as pharmaceutical groups for medications.

According to Figure 8, Patient A is a male patient who has hospitalization history in the observation window and is admitted to the hospital seven months after the end of the observation window for congestive heart failure. The predicted risk is 96.4%, while the risk decreases for female patients or patients without hospitalization history. It is also not surprising to observe an increased risk for older patients. The heat map in Figure 7 shows the relative importance of the medical events in this patient's medical record at each time window and the first row of the heat map presents the subsequence-level attention. The darker color indicates a stronger correlation between the clinical events and the

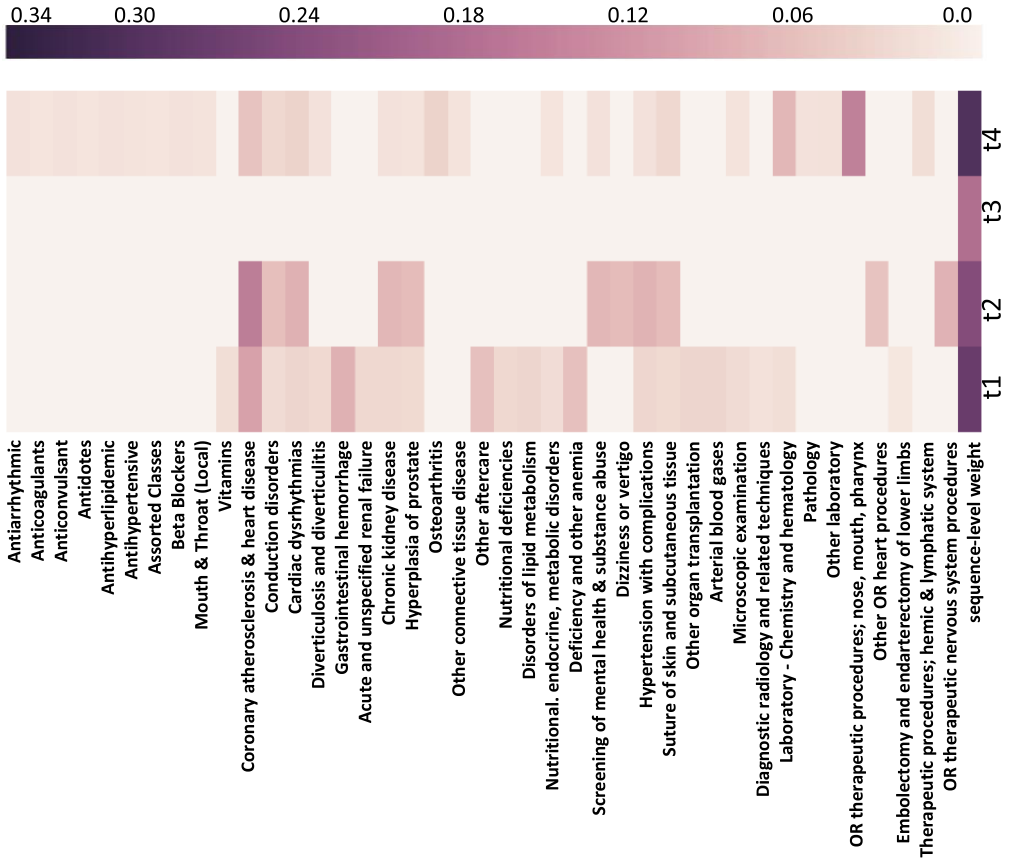

FIGURE 7. The heat map showing feature importance for Patient A.

Patient A
Age: 77
Gender: male
Previous hospitalization: yes
Predicted risk: 0.964
Hospitalized in the 7th month after the observation window
Hospitalization cause (primary diagnosis): systolic heart failure
Other scenarios:
- If female: predicted risk ↓ 0.008
- If 10 years older: predicted risk ↑ 0.007
- If no previous hospitalization: predicted risk ↓ 0.002

FIGURE 8. The profile of Patient A.

outcome. Accordingly, we observe that the last subsequence, t4, is the most important with respect to hospitalization risk, followed by $t1$, $t2$, and $t3$ in order of importance.

Among all the clinical events in the subsequence $t4$, we observe that the *OR therapeutic procedures (nose, mouth, and pharynx)*, *laboratory (chemistry and hematology)*, *coronary atherosclerosis & other heart disease*, *cardiac dysrhythmias*, and *conduction disorders* are the ones with the highest weights, while other events such as *other connected tissue disease* are less important in terms of future hospitalization risk. Additionally, some medications appear to be informative as well, including *beta blockers*, *antihypertensives*, *anticonvulsants*, *anticoagulants*, etc. In the first-time window, the medical events with high weights are *coronary atherosclerosis & other heart disease*, *gastrointestinal hemorrhage*, *deficiency and anemia*, and *other aftercare*. In the next subsequence, the most important medical events are heart diseases and related procedures such as *coronary atherosclerosis & other heart disease*, *cardiac dysrhythmias*, *conduction disorders*, *hypertension with complications*, *other OR heart procedures*, and *other OR therapeutic nervous system procedures*. We also observe that the kidney disease related diagnoses and procedures appear to be important features. Throughout the observation window, the *coronary atherosclerosis & other heart disease*, *cardiac dysrhythmias*, and *conduction disorders* constantly show high weights with respect to hospitalization risk, and the findings are consistent with medical literature.

Figure 9 presents the profile of Patient B, which is a male patient without hospitalization in the observation window. This patient is hospitalized for occlusion of cerebral arteries approximately one year after the observation window, and the predicted risk is 74.6%. For a similar patient who is 10 years older or with previous hospitalization history, the risk increases by 4.2% and 1%, respectively, while there is a smaller risk of hospitalization for a female patient. To illustrate the medical events of Patient B, the heat map in Figure 10 depicts the relative importance of medical groups

Patient B
Age: 64
Gender: male
Previous hospitalization: no
Predicted risk: 0.746
Hospitalized in the 13th month after the observation window
Hospitalization cause (primary diagnosis): occlusion of cerebral arteries
Other scenarios:
- If female: predicted risk ↓ 0.042
- If 10 years older: predicted risk ↑ 0.042
- If has previous hospitalization: predicted risk ↑ 0.010

FIGURE 9. The profile of Patient B.

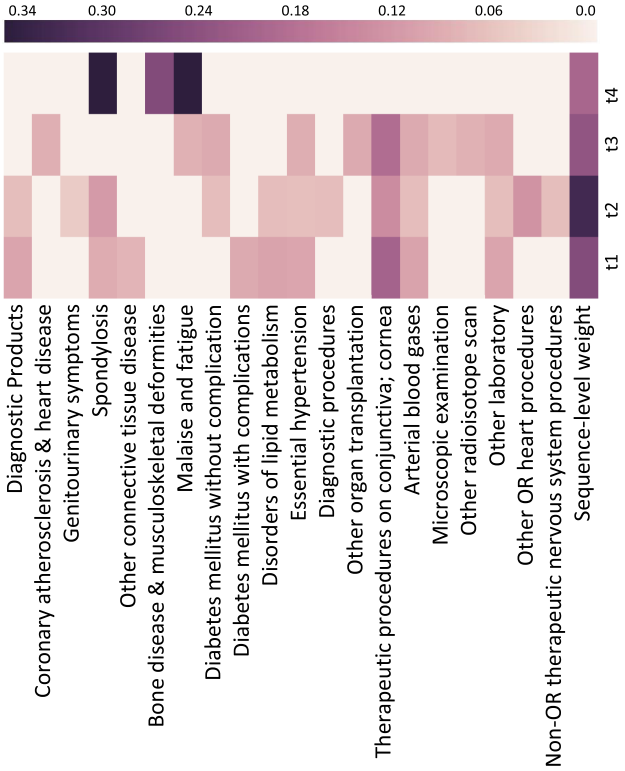

FIGURE 10. The heat map showing feature importance for Patient B.

in the subsequences, as well as the subsequence-level weights for hospitalization risk. Similarly, the darker color indicates a stronger correlation between the clinical events and the outcome. Accordingly, we observe that the second subsequence appears to be the most important, while the last one is less predictive of future hospitalization. In fact, the medical events in the last time window are *spondylosis, intervertebral disc disorders, other back problems* and *other bone disease & musculoskeletal deformities*, and *malaise and fatigue*, which are not highly related to the cause of hospitalization of Patient B.

In the most predictive subsequence, $t2$, we observe that *other OR heart procedures*, *genitourinary symptoms*, *spondylosis, intervertebral disc disorders, other back problems*, *therapeutic procedures on eyelid, conjunctiva, and cornea*, and *arterial blood gases* have high attention weights. In the earliest time window, the most important medical events also include *therapeutic procedures on eyelid, conjunctiva, and cornea*, *arterial blood gases*, while *diabetes*, *hypertension* as well as diagnostic products show their relatively high importance. Throughout the observation window, medical events *spondylosis, intervertebral disc disorders, other back problems*, *therapeutic procedures on eyelid, conjunctiva, and cornea* are constantly with high attention weights. Here, diagnostic products is a medication class, which include barium sulfate, iohexol, gadopentetate dimeglumine, iodixanol, tuberculin purified protein derivative, iodixanol, regadenoson, acetone (urine), and so forth. These medications are primarily for blood or urine testing, or used as radiopaque contrast agents for x-rays or CT scans for diagnostic purposes.

TABLE 2. The top clinical groups with high weights in hospitalized patients.

Index	Clinical Groups
Diagnoses	
1	Essential hypertension
2	Other connective tissue disease
3	Spondylosis; intervertebral disc disorders; other back problems
4	Other lower respiratory disease
5	Disorders of lipid metabolism
6	Other aftercare
7	Diabetes mellitus without complication
8	Screening and history of mental health and substance abuse codes
9	Other nervous system disorders
10	Other screening for suspected conditions (not mental disorders or infectious disease)
Procedures	
1	Other OR therapeutic procedures on nose; mouth and pharynx
2	Suture of skin and subcutaneous tissue
3	Other therapeutic procedures on eyelids; conjunctiva; cornea
4	Laboratory - Chemistry and hematology
5	Other laboratory
6	Other OR therapeutic procedures of urinary tract
7	Other OR procedures on vessels other than head and neck
8	Therapeutic radiology for cancer treatment
Medications	
1	Diagnostic Products
2	Analgesics-Narcotic

Additionally, we attempt to interpret the learned representation and feature importance at the population-level. In Table 2, we present the top 20 clinical groups with high weights among hospitalized patients in the test set.

According to Table 2, the most predictive diagnosis groups for future hospitalization are chronic diseases, including *essential hypertension*, *diabetes*, *lower respiratory disease*, *disorders of lipid metabolism*, and musculoskeletal diseases such as *other connective tissue disease* and *spondylosis, intervertebral disc disorders, other back problems*. The most

TABLE 3. The top diagnosis groups with high weights in patients hospitalized for *osteoarthritis*, *septicemia*, *acute myocardial infarction*, *congestive heart failure*, and *diabetes mellitus with complications*, respectively.

Index	Diagnosis Groups
In patients admitted for *osteoarthritis*	
1	Osteoarthritis
2	Other connective tissue disease
3	Other non-traumatic joint disorders
4	Spondylosis; intervertebral disc disorders; other back problems
5	Other aftercare
In patients admitted for *septicemia*	
1	Essential hypertension
2	Diabetes mellitus without complication
3	Disorders of lipid metabolism
4	Other lower respiratory disease
5	Other aftercare
In patients admitted for *acute myocardial infarction*	
1	Coronary atherosclerosis and other heart disease
2	Medical examination/evaluation
3	Other screening for suspected conditions (not mental disorders or infectious disease)
4	Other lower respiratory disease
5	Disorders of lipid metabolism
In patients admitted for *congestive heart failure*	
1	Congestive heart failure (nonhypertensive)
2	Coronary atherosclerosis and other heart disease
3	Cardiac dysrhythmias
4	Diabetes mellitus without complication
5	Other lower respiratory disease
In patients admitted for *diabetes mellitus with complications*	
1	Diabetes mellitus with complications
2	Diabetes mellitus without complication
3	Other aftercare
4	Other nutritional; endocrine; and metabolic disorders
5	Fluid and electrolyte disorders

important procedures are some OR therapeutic procedures and laboratory tests, such as the OR procedures on nose, mouth, and pharynx, vessels, urinary tract, eyelid, conjunctiva, cornea, etc. It is not surprising to see that diagnostic products are showing with high weights, considering these medications are used in testing or examinations for diagnostic purposes.

Moreover, we present the top diagnoses groups with high weights in patients hospitalized for different primary causes. Table 3 shows the top 5 diagnosis groups with high weights in patients admitted for *osteoarthritis*, *septicemia (except in labor)*, *acute myocardial infarction*, *congestive heart failure (nonhypertensive)*, and *diabetes mellitus with complications*, respectively. Accordingly, we observe that the most important diagnoses for hospitalization risk prediction in population admitted for osteoarthritis are musculoskeletal diseases such as connective tissue disease, joint disorders, and spondylosis. However, the diagnoses with highest weights in the patients admitted for septicemia are chronic diseases including essential hypertension, diabetes, disorders of lipid metabolism, and respiratory disease. The top diagnoses have many overlaps between the populations admitted for acute myocardial infarction and for congestive heart failure, considering both populations are admitted for heart diseases. Here, the overlapped diagnosis groups include coronary atherosclerosis and other heart diseases and lower respiratory diseases. As for patients admitted for diabetes with complications, the top diagnoses are diabetes with or without complications, nutritional, endocrine, metabolic disorders, and fluid and electrolyte disorders. In general, the learned feature importance is consistent with medical literature.

VI. DISCUSSION

Our proposed framework is applied to the prediction of hospitalization using real EHR data that demonstrates its prediction accuracy and interpretability. This work could be further enhanced by incorporating the follow-up information on the negative patient population and investigate if it indeed shows an improved health outcome or the patient is hospitalized elsewhere. *Patient2Vec* employs a hierarchical attention mechanism, allowing us to directly interpret the weights of clinical events. In future work, we will extend the attention to incorporate demographic information for a more comprehensive and automatic interpretation.

Although we apply *Patient2Vec* to the early detection of long-term hospitalization, i.e., at least 6 months after the previous hospitalization, it could be used to predict the risk of 30-day readmission to help prevent unnecessary rehospitalizations.

VII. CONCLUSION

In this paper, we propose a representation learning framework, *Patient2Vec*, to learn a personalized interpretable deep representation of EHR data based on recurrent neural networks and the attention mechanism. This work improves the performance of predictive models as well as deepens the understanding of disease correlations. We apply this framework to the risk prediction of hospitalization using patients' longitudinal EHR data. The experimental results demonstrate that the proposed *Patient2Vec* representation is capable of achieving a more accurate prediction than baselines approaches. Moreover, the learned feature importance in the representations are interpreted both at the individual and population levels to facilitate clinical insights.

In this work, the proposed *Patient2Vec* framework is evaluated with the risk prediction of all-cause hospitalization, but in the future could be applied to predict hospitalization in more specific populations, other health related prediction problems, or domains outside of health.

REFERENCES

[1] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.

[2] Y. Kim. (2014). "Convolutional neural networks for sentence classification." [Online]. Available: https://arxiv.org/abs/1408.5882

[3] J. Howard and S. Ruder. (2018). "Universal language model fine-tuning for text classification." [Online]. Available: https://arxiv.org/abs/1801.06146

[4] M. M. Lopez and J. Kalita. (2017). "Deep learning applied to NLP." [Online]. Available: https://arxiv.org/abs/1703.03091

[5] K. Cho *et al.* (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation." [Online]. Available: https://arxiv.org/abs/1406.1078

[6] A. L. Nobles, J. J. Glenn, K. Kowsari, B. A. Teachman, and L. E. Barnes, "Identification of imminent suicide risk among young adults using text messages," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, p. 413.

[7] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "HDLTex: Hierarchical deep learning for text classification," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 364–371.

[8] K. Kowsari, M. Heidarysafa, D. E. Brown, K. J. Meimandi, and L. E. Barnes. (2018). "RMDL: Random multimodel deep learning for classification." [Online]. Available: https://arxiv.org/abs/1805.01890

[9] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush, "LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 667–676, Jan. 2018.

[10] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 6085.

[11] M.-T. Luong, H. Pham, and C. D. Manning. (2015). "Effective approaches to attention-based neural machine translation." [Online]. Available: https://arxiv.org/abs/1508.04025

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[13] D. Yogatama, C. Dyer, W. Ling, and P. Blunsom. (2017). "Generative and discriminative text classification with recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1703.01898

[14] T. Young, D. Hazarika, S. Poria, and E. Cambria. (2017). "Recent trends in deep learning based natural language processing." [Online]. Available: https://arxiv.org/abs/1708.02709

[15] M. Basaldella, E. Antolli, G. Serra, and C. Tasso, "Bidirectional LSTM recurrent neural network for keyphrase extraction," in *Proc. Italian Res. Conf. Digit. Libraries*. Springer, 2018, pp. 180–187.

[16] S. Ghosh, O. Vinyals, B. Strope, S. Roy, T. Dean, and L. Heck. (2016). "Contextual LSTM (CLSTM) models for large scale NLP tasks." [Online]. Available: https://arxiv.org/abs/1602.06291

[17] B. Yue, J. Fu, and J. Liang, "Residual recurrent neural networks for learning sequential representations," *Information*, vol. 9, no. 3, p. 56, 2018.

[18] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling." [Online]. Available: https://arxiv.org/abs/1412.3555

[19] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. ICML*, vol. 28, 2013, pp. 1310–1318.

[20] D. Britz. (2015). *Recurrent Neural Network Tutorial*. Accessed: Oct. 5, 2017. [Online]. Available: http://www.wildml.com/2015/10/

[21] V. Mnih *et al.*, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[22] A. M. Rush, S. Chopra, and J. Weston. (2015). "A neural attention model for abstractive sentence summarization." [Online]. Available: https://arxiv.org/abs/1509.00685

[23] T. Ma, C. Xiao, and F. Wang, "Health-Atm: A deep architecture for multifaceted patient health record representation and risk prediction," in *Proc. SIAM Int. Conf. Data Mining (SIAM)*, 2018, pp. 261–269.

[24] A. Rajkomar, E. Oren, and J. Dean, "Scalable and accurate deep learning with electronic health records," *Digit. Med.*, vol. 1, no. 1, 2018, Art. no. 18.

[25] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel. (2015). "Learning to diagnose with LSTM recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1511.03677

[26] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. (2016). "Recurrent neural networks for multivariate time series with missing values." [Online]. Available: https://arxiv.org/abs/1606.01865

[27] A. E. W. Johnson *et al.*, "Mimic-III, a freely accessible critical care database," *Sci. Data*, vol. 3, May 2016, Art. no. 160035.

[28] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. Mach. Learn. Healthcare Conf.*, 2016, pp. 301–318.

[29] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Interpretable deep models for ICU outcome prediction," in *Proc. AMIA Annu. Symp. Amer. Med. Inform. Assoc.*, 2016, pp. 371–380.

[30] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[31] 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. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3504–3512.

[32] Z. Lin *et al.* (2017). "A structured self-attentive sentence embedding." [Online]. Available: https://arxiv.org/abs/1703.03130

[33] C. M. Torio and B. J. Moore. (2016). *National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013*. [Online]. Available: https://www.hcup-us.ahrq.gov/reports/statbriefs/sb204-Most-Expensive-Hospital-Conditions.jsp

[34] E. Wallace, E. Stuart, N. Vaughan, K. Bennett, T. Fahey, and S. M. Smith, "Risk prediction models to predict emergency hospital admission in community-dwelling adults: A systematic review," *Med. Care*, vol. 52, no. 8, pp. 751–765, 2014.

[35] E. N. de Vries, M. A. Ramrattan, S. M. Smorenburg, D. J. Gouma, and M. A. Boermeester, "The incidence and nature of in-hospital adverse events: A systematic review," *Qual. Saf. Health Care*, vol. 17, no. 3, pp. 216–223, 2008.

[36] S. Purdey and A. Huntley, "Predicting and preventing avoidable hospital admissions: A review," *J. Roy. College Phys. Edinburgh*, vol. 43, no. 4, pp. 340–344, 2012.

[37] H. Ontario, "Early identification of people at risk of hospitalization: Hospital admission risk prediction (HARP)—A new tool for supporting providers and patients," 2013.

[38] B. Zheng, J. Zhang, S. W. Yoon, S. S. Lam, M. Khasawneh, and S. Poranki, "Predictive modeling of hospital readmissions using metaheuristics and data mining," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7110–7120, 2015.

[39] D. Kansagara *et al.*, "Risk prediction models for hospital readmission: A systematic review," *J. Amer. Med. Assoc.*, vol. 306, no. 15, pp. 1688–1698, 2011.

[40] G. Giamouzis *et al.*, "Hospitalization epidemic in patients with heart failure: Risk factors, risk prediction, knowledge gaps, and future directions," *J. Cardiac Failure*, vol. 17, no. 1, pp. 54–75, 2011.

[41] E. Prescott, A. M. Bjerg, P. K. Andersen, P. Lange, and J. Vestbo, "Gender difference in smoking effects on lung function and risk of hospitalization for COPD: Results from a danish longitudinal population study," *Eur. Respiratory J.*, vol. 10, no. 4, pp. 822–827, 1997.

[42] Agency for Healthcare Research and Quality (AHRQ). (2015). *Clinical Classifications Software (CCS) for ICD-9-CM*. [Online]. Available: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

JINGHE ZHANG received the M.Sc. degree in industrial and systems engineering from the State University of New York at Binghamton and the Ph.D. degree in systems engineering from the University of Virginia. She is currently a Lead Data Scientist at Target Corporation. Her research interests include natural language processing, machine learning, recommender systems, and health informatics.

KAMRAN KOWSARI received the M.Sc. degree from the Department of Computer Science, The George Washington University, Washington, DC, USA. He is currently pursuing the Ph.D. degree with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA, where he is also a member of the Sensing Systems for Health Lab. He has more than 10 years of experience in machine learning and software development. His experience includes numerous industrial and academic projects. His research interests include natural language processing, machine learning, deep learning, artificial intelligence, text mining, and unsupervised learning.

JAMES H. HARRISON, JR., received the M.D. and Ph.D. degrees in pharmacology from the Medical University of South Carolina, Charleston, SC, USA. He completed residencies in the Anatomic Pathology and Laboratory Medicine, Yale-New Haven Hospital, New Haven, CT, USA, and a Post-Doctoral Fellowship in environmental toxicology at Yale University, New Haven, CT, USA. He is currently an Associate Professor of pathology and the Director of the Laboratory Information Systems, University of Virginia Medical Center, and has also appointments at the Department of Public Health Sciences, UVA School of Medicine, and the Department of Systems and Information Engineering, UVA School of Engineering and Applied Sciences. He has over 25 years of experience in the field of medical informatics, including work in clinical laboratory information systems, electronic health records, clinical data analysis, and clinical data standards development.

JENNIFER M. LOBO received the Ph.D. degree in industrial engineering from North Carolina State University, Raleigh, NC, USA. She is currently an Assistant Professor of biomedical informatics with the Department of Public Health Sciences, University of Virginia. Her research interests involve using mathematical modeling and stochastic optimization methods to build models that simulate the natural course of disease. These models allow for an estimation of outcomes under different screening and treatment policies in the absence of randomized controlled trials and can be used to optimize screening and treatment decisions for patients with chronic diseases. Her projects include optimizing treatment for patients with type 2 diabetes, generating individualized decision analysis models for prostate cancer patients, and developing optimal imaging surveillance guidelines for recurrent kidney cancer.

LAURA E. BARNES received the Ph.D. degree in computer science from the University of South Florida, Tampa, FL, USA. She is currently an Associate Professor in systems and information engineering with the Data Science Institute, University of Virginia, where she directs the Sensing Systems for Health Lab which focuses on understanding the dynamics and personalization of health and well-being through mobile sensing and analytics.