Representation Learning of Longitudinal Electronic Health Record Data for Patient Characterization and Prediction of Health Outcomes

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

in partial fulfillment
of the requirements for the degree

Doctor of Philosophy

by

Jinghe Zhang

December 2017

APPROVAL SHEET

This Dissertation
is submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Author Signature: ______________________

This Dissertation has been read and approved by the examining committee:

Advisor: Laura E. Barnes

Committee Member: Donald E. Brown

Committee Member: Stephen D. Patek

Committee Member: James H. Harrison

Committee Member: Jennifer M. Lobo

Committee Member: Christopher C. Moore

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, School of Engineering and Applied Science

December 2017

Representation Learning of Longitudinal Electronic Health Record Data for Patient Characterization and Prediction of Health Outcomes

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Jinghe Zhang

December

2017

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, this data remains largely unexplored, but presents 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 this data present many complex challenges. This complexity makes it difficult to translate the potentially relevant information into machine learning algorithms.

To that end, this research provides contributions to the interpretable representation of complex, sparse, high-dimensional data comprised of various medical events, such as diagnoses, medications and procedures. The goal of this dissertation is to propose new computational frameworks for representing longitudinal EHR data for improved patient characterization and developing optimized predictive models. To illustrate the utility of the proposed frameworks, the designed algorithms will be applied to a variety of risk prediction problems including the early detection of diabetes, comorbid risk prediction of chronic diseases, and prediction of hospitalization. Furthermore, the designed algorithms are evaluated against other state-of-the-art representation approaches, and the learned representations are visualized and interpreted to deepen clinical sights. In addition to assisting clinical decision making, the methods proposed in this research could be applied to other complex temporal knowledge representation tasks within and outside the healthcare domain.

Acknowledgements

It is a great pleasure to thank my committee members for their time and support. First, I would like to express my sincere gratitude to my advisor, Dr. Laura Barnes, for her constant encouragement and guidance throughout my study at UVa. Without her support, I wouldn't be able to complete this dissertation. Also, I would like to thank my committee members Dr. Donald Brown and Dr. Stephen Patek for their advice and help on my graduate study. Their wisdom and attitude inspire me greatly in my research. In addition, I would like to thank Dr. James Harrison, Dr. Jennifer Lobo, and Dr. Christopher Moore for their guidance on my projects that have greatly deepened my understanding and interests in this area.

Most importantly, I would like to express my deepest appreciation to my parents and sisters for their persistent encouragement, support, and confidence on me. I appreciate everything they have done for me and I would not be where I am without them. It is their great love to me that have enlightened every moment of my life. I am so fortunate to have such a great family and they are my greatest motivation to pursue my dreams.

Finally, I would like to thank my dear friends. I will always remember our true friendship and the good times and laughs that we have shared. Special thanks to all the people that have helped me during my Ph.D. study.

To my beloved family

Table of Contents

List of Tables

List of Figures

Abbreviations

AFV	Aggregated frequency vector
AHRQ	Agency for healthcare research and quality
ATV	Aggregated transition vector
AUC	Area under curve
BiRNN-MGE	Bidirectional RNN with medical group embedding
BiRNN-MVE	Bidirectional RNN with medical vector embedding
BPS	Bag-of-pattern in sequences
CBOW	Continuous bag-of-words
CDC	Centers for disease control and prevention
CDR	Clinical database repository
CKD	Chronic kidney disease
CNN	Convolutional neural network
CPT	Current procedural terminology
DNN	Deep neural network
EHR	Electronic health record
ESRF	End stage renal failure
FRNN-MGE	Froward RNN with medical group embedding
FRNN-MVE	Forward RNN with medical vector embedding
GBT	Gradient boosting trees
GDP	Gross domestic product

GSP	Generalized sequential pattern
GRU	Gated recurrent unit
ICD-9	International classification of diseases, ninth version
ICD-10	International classification of diseases, tenth version
ICU	Intensive care unit
LR	Logistic regression
LSTM	Long short term memory
MLP	Multilayer perceptron
NLP	Natural language processing
OOB	Out-of-bag
RF	Random forest
RNN	Recurrent neural network
TEMR	Temporal event matrix representation
TF-IDF	Term frequency - inverted document frequency

Chapter 1

Introduction

Electronic health systems have been widely implemented in the United States and across the world [7]. The availability of tremendous health data provides promising opportunities for public health research. In particular, electronic health record (EHR) data have become very popular in clinical decision support systems, such as predictive modeling of health outcomes. However, many challenges exist when utilizing EHR data for such tasks. In this chapter, we introduce the utilization of EHR data in predictive modeling and its challenges. Then, we define the research problems and the data used in this work briefly.

1.1 Utilization of EHR Data in Clinical Decision Support

Rich EHR data can facilitate clinical research in understanding disease correlations and detecting health outcomes in advance, where machine learning plays an important role in this process. In particular, predictive models have been applied to help with decision-making in many medical domains. These include the prediction of breast cancer, type 2 diabetes, cardiovascular disease, and mortality for critically-ill hospitalized adults to name a few [8–11]. According to the model output, a high-risk patient will be referred to intensive interventions and monitoring (e.g., screening and counseling) to prevent adverse outcomes. These models have the potential to reduce the mortality rates and improve the quality of life of high-risk patients and control cost and complications for low-risk patients [12]. Notably, predictive models have become vital tools to assist with medical decision-making and can bring benefits to both healthcare providers and patients.

Predictive modeling of health outcomes can be treated as a classification problem in which many well-established classifiers are actively used in this arena, such as logistic regression and random forest. In [13], researchers studied the cardiovascular risk prediction problem using comprehensive EHR data, in which a few popular machine learning approaches are adopted, including logistic regression, generalized additive models, and gradient boosting trees. The experimental results demonstrate an improvement in predictive performance when compared to traditional clinical scoring systems. This and other similar research in applying machine learning methods to large healthcare datasets have demonstrated great potential in advancing clinical care.

1.2 Challenges

The wide implementation of electronic health systems has a great research potential for understanding unknown disease correlations and for better characterization of patients. However, it is generally very challenging to analyze EHR data for a variety of reasons, presented as follows [14, 15]:

- Complexity: EHR data are generally high-dimensional and sparse, and contain information collected from multiple sources, which makes it hard to integrate them into a universal feature space while preserving all useful information. In addition to single clinical events, there are multivariate and nested sequences stored in EHR systems, and it is difficult for machines to understand the patterns.

- Heterogeneity: EHR systems store huge amounts of health histories of many patients, whose characteristics and medical conditions could be very diverse from each other. The heterogeneity in patients' medical histories and characteristics brings another challenge to analytical tools for clinical decision support. Apart from the heterogeneity between patients, there is also a fairly high level of variation in clinicians' practices. For instance, if a diabetic patient visits a physician for an acute condition unrelated to the chronic disease, the physician might code diabetes in the electronic record for this patient or not. In addition, a patient with chronic diseases might schedule a regular doctor visit every 30 days while a similar patient might have a doctor visit every 45 days. These heterogeneities make it challenging to mine EHR data.

- Interpretability: Physicians make clinical decisions based on precise knowledge of the patients' health conditions, medical histories, and other characteristics. Patterns learned by machine learning methods from EHR databases provide supplemental or even unknown knowledge to support the decision making. Thus, in addition to improving clinical decision support, it requires the patterns discovered by machines to be interpretable to humans for further validation and knowledge discovery.

- Time-invariance: Patients' medical histories are not aligned in absolute time. Thus, representations that capture information in relative time perspective are needed for further modeling.

- Scalability: EHR systems store medical histories from heterogeneous data sources of large numbers of patients; patients with chronic conditions can have very complex and long histories. This aspect contributes to the large scale data contained in the EHR database, and ultimately requires efficient computational methods for analysis.

Hence, EHR data is complex and cannot be readily utilized in machine learning methods. This makes it a challenging task to extract useful information for further modeling and analytics.

1.3 Problem Definition

Although the large scale data from EHR systems has great potential to support clinical care and to discover unknown disease correlations, its complexity poses many challenges in utilizing this data for health analytics. Thus, the objective of this research is to develop representation learning frameworks for longitudinal EHR data to facilitate further analytics on patient characterization and prediction of health outcomes. Formally, representation learning is learning representations of the data that make it easier to extract useful information when building classifiers or other predictors [16].

In this research, we focus on addressing five problems in developing the representation learning frameworks.

- Temporal Information: A patient's medical history is expressed as multiple visits with distinct time stamps and clinical events, including the diagnoses, prescribed medications and procedures, etc. Overall, a patient's medical history in EHR data is a nested sequence of events with implicit temporal relationships between events. Thus, an effective representation of a patient's EHR data needs to capture the temporal information as well.

- High-dimensionality and Sparsity: Patients' clinical conditions and events that occur during visits are heterogeneous, resulting in a very high-dimensional feature space. For instance, there are $14,000$ ICD-9 diagnosis codes and $68,000$

ICD-10 diagnosis codes. In the meantime, each patient has a very small portion of events occurring in the medical history such that the feature space is very sparse.

- Predictive Modeling: Predictive modeling plays an increasingly important role in supporting clinical decision making, where early warning systems based on predictive modeling have been beneficial to both health care providers and patients. Thus, one of the aims of this research is to develop representation learning frameworks to better support predictive modeling of diseases or health outcomes of interest.

- Interpretation: Clinical decision making relies on precise knowledge of patient's medical condition and characteristics. Thus, a representation interpretable by a human is more desirable from two perspectives. First, it can deepen clinical insights that were previously unknown and could be used by clinicians to assist health and medical research. Second, the knowledge learned by mining EHR data can be evaluated by human experts to validate its utility.

- Personalization: Health care systems have experienced many changes over the past few decades. Access to a tremendous amount of health data and revolutionary discoveries in the medical world has promoted the demand and interest in personalized health care [17]. Although personalized medicine has been studied extensively in genetics research, little work has been conducted to help with personalized clinical decision making via mining EHR data [18, 19].

  In fact, in the prediction of certain health outcomes or in clinical decision making, it is very likely that a particular clinical event is of great importance for one group of patients, but not for another group. Furthermore, the relative importance of clinical events in one patient's medical history is most likely distinct from that of a different patient. Thus, one of the aims of this research is to learn a personalized representation for each patient to provide more accurate prediction of health outcomes and a better understanding of a patient's medical history, which can ultimately help with the delivery of personalized medicine.

In brief, this research aims to develop novel representation learning frameworks for longitudinal EHR data that can capture temporal relationships between clinical events, cope with the high-dimensionality and sparsity in the data, and help improve the performance of predictions on diseases and health outcomes. Finally, the learned representation is interpretable to allow further validation and to bring meaningful clinical insights.

1.4 Data

This research uses the EHR data from the University of Virginia (UVA) Clinical Database Repository (CDR), which is a de-identified data warehouse managed by the Division of Biomedical Informatics in the Department of Public Health Sciences [20].

This CDR contains information of all patient encounters in the UVA Health System, including limited patient demographics, such as age and sex, and inpatient and outpatient visit details with diagnoses, procedures, and medications. The diagnoses are coded primarily in the International Classification of Diseases, Ninth Version (ICD-9), and a small portion are in the Tenth Version (ICD-10) format [21,22]. The procedures are primarily coded in the Current Procedural Terminology (CPT) and some are in ICD-9 [21,23]. This research is based on 75 months of data beginning in September 2010, including $2,343,651$ encounters of $473,915$ distinct patients. On average, each patient has 4.95 visits during this time period.

1.5 Dissertation Overview

In this research, novel computational frameworks are proposed to learn effective and interpretable representations of patients' medical histories from a longitudinal EHR database. The rest of the dissertation is organized as follows.

Chapter 2 discusses related work on data representation and representation learning approaches. First, we review the commonly used and state-of-the-art representation in general, particularly in natural language processing (NLP), such as TF-IDF, and Word2Vec. Then, representation learning approaches for EHR data are summarized. We also provide an overview of popular classification models which are used for further predictive modeling to evaluate the proposed frameworks.

Chapters 3 to 5 present three computational representation learning frameworks to address the challenges in utilizing longitudinal EHR data. A general overview and the detailed algorithms are provided for each designed framework. Then, the proposed methods are evaluated with three prediction problems, respectively, using EHR data from real clinical settings. The performance of the models are compared with popular representation baselines. Additionally, the learned representations are visualized and interpreted to bring clinical insights.

Finally, this dissertation is concluded with a summary of the contributions of this research and a discussion on future directions in Chapter 6.

Chapter 2

Literature Review

In this chapter, we review the state-of-the-art data representations in general and for EHR data. First, we summarize the previous studies on representation learning, including feature vectors, graphs, and distributed representation, especially in natural language processing. Second, we focus on the data representation employed for EHR data. Finally, we describe popular classification models which are used for predictive modeling with the learned representations in the healthcare domain.

2.1 Data Representation in Natural Language Processing

The EHR data containing clinical events is similar to text documents with words in natural language processing (NLP) domain. Thus, the data representation approaches used for texts could be potentially used for EHR data representation. There are two types of word embeddings — frequency-based embedding and prediction-based embedding. In this section, we present a popular representation learning method for text data in each of the above categories, respectively.

2.1.1 Frequency-based Embedding: TF-IDF

TF-IDF is a classic representation approach for text data, and it is the most well-known document representation schema in information retrieval. In general, TF-IDF is a term-weighting scheme according to its importance in a document indicated by the statistics of occurrence of the term [24]. TF-IDF consists of two parts: TF is short for term frequency and IDF is inverted document frequency. Here, the frequency count of term t in document d is denoted as $tf(t,d)$. Usually, TF is normalized due to the variation in document lengths and repeated occurrences. A popular TF normalization method is sublinear TF scaling, which normalizes TF by the most frequent term in the document, as presented in Equation 2.1.

$$TF(t,d) = \begin{cases} \alpha + (1-\alpha)\frac{tf(t,d)}{max_t tf(t,d)} & tf(t,d) > 0 \\ 0 & otherwise \end{cases} \tag{2.1}$$

IDF assigns higher weights to rare terms since terms occurring in fewer documents are more discriminative. Equation 2.2 shows the formula to calculate IDF.

$$IDF(t) = 1 + log(\frac{N}{df(t)}) \tag{2.2}$$

where N is the total number of documents in the corpus and $df(t)$ is the counts of documents containing term t. Straightforwardly, TF-IDF weighting is the combination of TF and IDF, as shown in Equation 2.3 [25].

$$w(t,d) = TF(t,d) \times IDF(t) \tag{2.3}$$

Thus, a document d is represented by a vector $v(d)$ and each dimension corresponds to a distinct term t with value $w(t,d)$.

TF-IDF is based on the idea of bag-of-words and it is a very intuitive method, which has been demonstrated to be empirically effective and easy to implement. However, it has ignored the ordering between terms in a document and assumes term independence [24].

2.1.2 Prediction-based Embedding: Word2Vec

Bag-of-words representation ignores the ordering of words and fails to capture much of the semantics. Additionally, the vocabulary size is usually very large such that bag-of-words methods always result in a very high-dimensional and sparse vector representation of documents [26]. Word2Vec, a group of methods to learn fixed-length dense vector representations using a shallow neural network trained by prediction tasks, is proposed to address these issues [26].

There are two variants of the Word2Vec algorithm: CBOW (continuous bag-of-words) and skip-gram. The former one uses context to predict a target word, while the latter one uses a word to predict context. We describe the structure of skip-gram, presented in Figure 2.1 and its computation in the following.

Given a word w_t represented with a unique vector v_{w_t}, we have a sequence of words $\{w_1, w_2, \cdots, w_T\}$ and the objective of the prediction task is to maximize the average log probability of context words presented as follows [26, 27].

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \sum_{-k \leq j \leq k, j \neq 0} log(p(w_{t+j}|w_t)) \tag{2.4}$$

where k is the size of the context window. Usually, softmax function is used for classification problems, such that [28]

$$p(w_c|w_t;\theta) = \frac{exp(v_{w_t}^{\intercal} v'_{w_c})}{\sum_{j=1}^{V} exp(v_{w_t}^{\intercal} v'_{w_j})} \tag{2.5}$$

where V is the vocabulary size, and v_{w_t} and v'_{w_t} are the input and output vector of

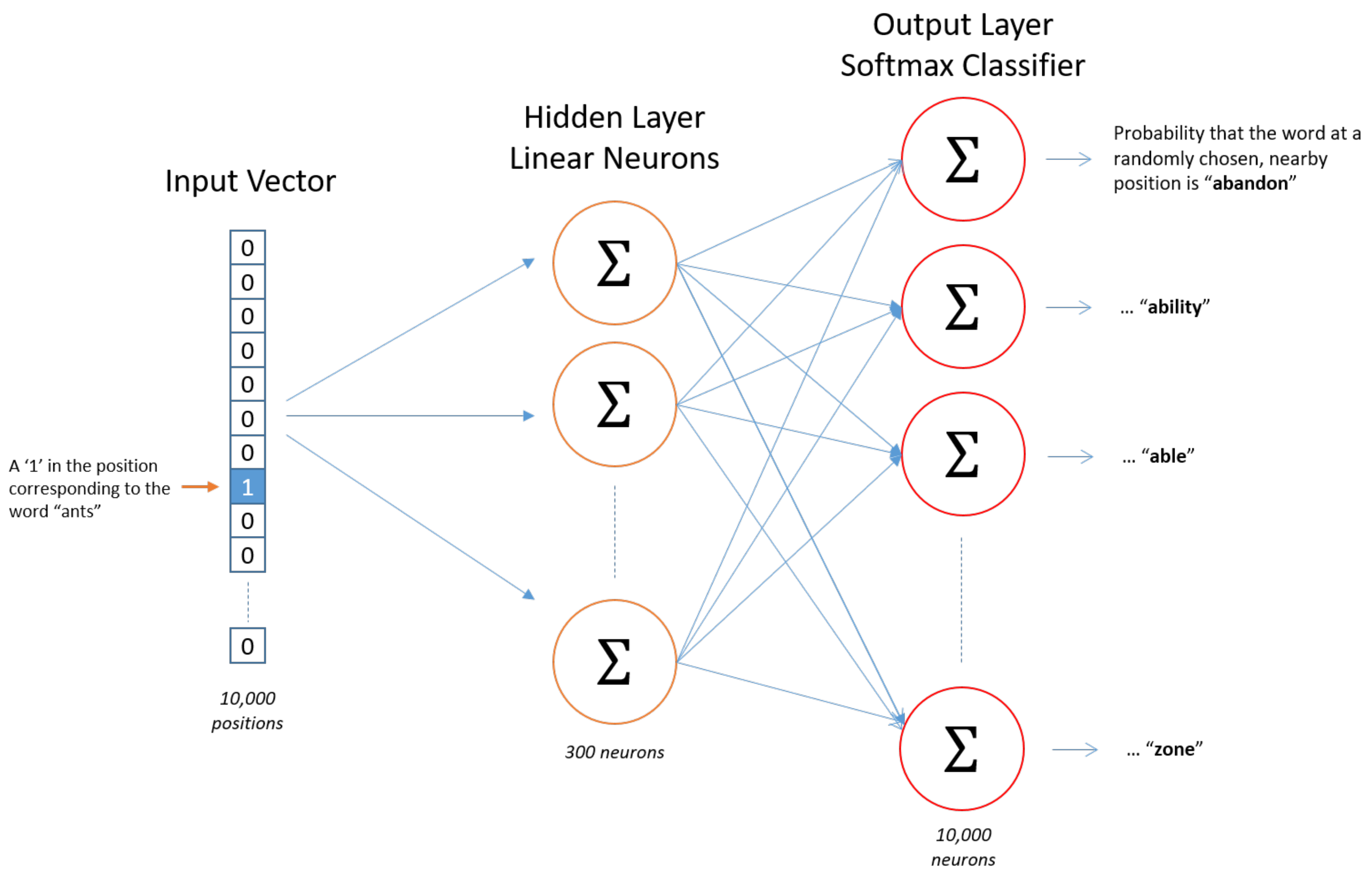


Figure 2.1: The skip-gram model for vector representation of medical codes [1]

word w_t. With the learned word embeddings, words sharing common contexts are located close to each other in the new vector space. Finally, a document is represented by the concatenation or sum of words in it.

With this effective representation of words and documents, it has achieved successes in many applications in the NLP field. Considering the similarity between text documents and longitudinal EHR data, the representation methods have great potential to be applied to the representation of clinical events in EHR data.

2.2 Data Representation of EHR Data

In this section, we discuss the popular representations of EHR data, which have been the basis of many predictive modeling tasks.

2.2.1 Aggregated Frequency Vector

Usually, frequency and presence (or absence) are used as the representations for categorical features, where presence or absence is coded as a binary variable [29, 30]. In [29], the diagnosis of diabetes is predicted based on the past diagnoses information, medication and procedure orders, and lab tests from patients' medical records. For each patient, a feature vector is constructed based on the longitudinal health data. When using frequencies, a feature value is generated by counting the occurrences of the associated clinical event in the pre-defined time window. Otherwise, it uses a binary variable as the feature value where 1 indicates that the associated event is present in the patient's medical history, and 0 otherwise [30].

Thus, a frequency vector or a binary vector indicating presence of clinical events is constructed to represent a patient's medical history in EHR data. However, the temporal information are omitted which might adversely affect the performance of predictive models.

2.2.2 Bag-of-Pattern in Sequences

Merely using frequencies omits the temporal orders of events, which might include vital information. Thus, frequent subsequences learned using sequence mining methods are used as features to represent patient's medical history [31–34].

Generalized Sequential Pattern (GSP) algorithm is a classical method for sequence mining, which is primarily based on the aprior algorithm [35]. GSP makes multiple passes over the database and starts with counting all single items. Using the frequent items discovered in the first pass, sequences with 2 items are generated and the supports are counted in the second pass. The GSP algorithm repeats this process until no more frequent sequences are found [32]. Many advanced algorithms have been proposed to improve the effectiveness and efficiency of sequence mining, such as SPAM, SPADE, and CM-SPADE [33, 34, 36].

Given the frequent bag-of-pattern sequences discovered by sequence mining algorithms, a vector representation is constructed with the counts or presence of these features. This representation learning approach is able to capture the temporal relationships between clinical events. However, it is very likely that important yet infrequent patterns are lost, while some frequent patterns which might not be useful are included in such a representation.

2.2.3 Other Advanced Representations

More advanced representation learning approaches have been proposed to preserve the temporal information in longitudinal EHR data. In [14, 37], longitudinal EHR data are represented in a temporal event matrix representation (TEMR) in which the columns correspond to time units and rows represent events. Further, a comprehensive patient signature is built by representing the EHR data in multiple time windows with different lengths.

Additionally, pairwise transitions have also been proposed to capture the short-term dependencies between clinical events [38]. Given each patient's visit sequence, pairs of events and their temporal orders A_{ij} in a visit sequence are extracted, where $A_{ij} \in \{A_{11}, A_{12}, \ldots, A_{nn}\}$ refers to the co-occurrences of diagnoses A_i and A_j in two distinct visits (i.e., the p^{th} and q^{th} visits, and $|p - q| = 1, 2, 3, \ldots$), and that A_i occurs prior to A_j. Then, f_{ij}^x is introduced to denote the transition from A_i to A_j in the sequence of patient x, which could be simple counts, binary feature indicating presence, or functions of time interval between events.

Table 2.1: The pairwise transition matrix of patient x

	A_1	A_2	A_3	A_4	A_5
A_0	f_{01}^x	f_{02}^x	f_{03}^x	f_{04}^x	f_{05}^x
A_1	f_{11}^x	f_{12}^x	f_{13}^x	f_{14}^x	f_{15}^x
A_2	f_{21}^x	f_{22}^x	f_{23}^x	f_{24}^x	f_{25}^x
A_3	f_{31}^x	f_{32}^x	f_{33}^x	f_{34}^x	f_{35}^x
A_4	f_{41}^x	f_{42}^x	f_{43}^x	f_{44}^x	f_{45}^x
A_5	f_{51}^x	f_{52}^x	f_{53}^x	f_{54}^x	f_{55}^x

Then, to learn the informative features from the temporal representation constructed above, various feature extraction approaches are proposed including latent factor model [14], non-negative matrix factorization [31], and simple χ^2 test [38].

Moreover, the graph-based representation approaches are proposed to capture temporal dependencies between clinical events, such as decomposing temporal graphs into bases [31], creating temporal clinical signatures and discovering temporal patterns using nonnegative matrix factorization based methods [14, 37, 39], and so forth [40]. In [31], the event sequence of each patient is represented with a temporal graph and the direction of the edges are based on the temporal orders between events. Further, the weight of each edge is defined by an exponential distribution of the time differences between two events. Then, the temporal graphs are decomposed into graph bases and each patient is represented by the graph bases and their associated weights. In general, these advanced approaches are capable of learning sequential patterns in the EHR data, and they achieve comparable predictive performance to the BPS representation method in terms of AUC according to [31].

Additionally, researchers have attempted to further learn the graph-based representation using deep learning approaches customized for longitudinal EHR data [41]. First, the medical record of each patient is converted to an attributed graph structure. The attributes of the vertices of the graph are the medical events (e.g., diagnosis codes) while the attributes of the edges are the temporal relationships between the vertices representing the correlation between events. Then, a convolutional neural network is adapted to handle irregular graphs. It initializes the distance between medical events by constructing a local tensor and then customizes the deep learning infrastructure to learn the correlation between events. Finally, the learned representations are used for further predictive modeling on the comorbid risk of chronic diseases.

Poor representation of data lacking vital information can adversely affect predictive models, while an appropriate data representation is the cornerstone for further advancements in analytics and modeling. Although the advanced approaches are capable of learning more sophisticated patterns of clinical events in patients' EHR data, they introduce more complex data structures, need prior clinical knowledge to guide the model construction, and require extra efforts to transform the findings to assist clinical decision support. Thus, it is imperative to develop effective and intuitive representation learning methods for the representation of EHR data. In this research, our work continues to explore other possibilities for advanced representation of EHR data.

2.3 Predictive Models

2.3.1 Logistic Regression

Logistic regression, a generalized linear regression model, is one of the most commonly used statistical methods to model the relationships between the response variable and explanatory features. The logistic function is very similar to linear regression except that the response variable is binomial such that the odds ratio is:

$$\frac{p(x)}{1 - p(x)} \tag{2.6}$$

where $p(x)$ is the probability of the target event given explanatory variables x. Logistic regression models the logarithm of the ratio, shown in Equation 2.7 [42].

$$log\frac{p(x)}{1 - p(x)} = \beta_0 + \beta x \tag{2.7}$$

Let $Pr(Y = 1|X = x) = p(x; \beta)$ and we model p

$$p(x_i; \beta) = \frac{1}{1 + exp^{-(\beta_0 + \beta x_i)}} \tag{2.8}$$

The maximum likelihood function is used for parameter estimation and the log likelihood transforms products to sums. The log likelihood function is presented in Equation 2.9 [42].

$$\frac{1}{N}\sum_{i=1}^{N}(y_i log(p(x_i; \beta)) + (1 - y_i)log(1 - p(x_i; \beta))) \tag{2.9}$$

where N is the total number of observations and $y_i \in \{0, 1\}$ is the true label of observation x_i. In addition, lasso, which is originally the shrinkage in linear regression, can also be adapted to logistic regression which shrinks the coefficients towards zero [43]. Thus, the objective function of lasso logistic regression is the likelihood function of a vanilla logistic regression with a $l1$ penalty term on coefficients β.

Overall, logistic regression and its variants can be used as predictive models when there are two or more classes.

2.3.2 Random Forest

Random forest is one of the most popular bagging ensemble learning methods, also called bootstrap aggregating, which can be used for classification, regression and other tasks. The basic idea of ensemble methods is that a group of "weak learners" are combined to obtain a "strong learner", where each individual classifier is a "weak learner". Bagging is a powerful ensemble method that can reduce the variance of a single classifier with high variance, such as decision trees. Bagging of decision trees works as follows [44].

1. Create bootstrap samples by re-sampling with replacement from the dataset
2. Train a decision tree on each bootstrap sample
3. Make a prediction on a new observation by combining the outputs from each tree, where majority voting is a commonly used approach to get the final output

Bagging of a high-variance base algorithm can result in a more stable classifier. However, decision trees are greedy and the base trees are likely to be correlated which limits the improvement that can be achieved by bagging.

Random forest is an improvement from bagging of decision trees, since the former one takes a randomly sampled subset of features when growing a tree. In this way, the models are not correlated or are weakly correlated which works better with the bagging approach. In addition to better classification performance, random forest computes feature importance by measuring the normalized average decrease in out-of-bag (oob) error after permuting a certain feature in the oob cases. The feature importance can also be computed using Gini impurity criterion in node splitting, i.e., by summing up the Gini decreases of the feature across all trees. The feature importance calculated in these two ways are usually consistent with each other [45]. Random forest grows an ensemble of trees, and the construction and training of each tree in the forest is as follows [45].

1. Randomly sample with replacement from the original data and use the bootstrapped sample as the training set to grow a decision tree.

2. Randomly sample $m \ll M$ features at each node, where M is the total number of features in the original data. Then, the best split on the subset of features is used to split the node.

3. Grow each tree without pruning.

Then, the final prediction on a new observation is obtained by majority voting on the classification decisions by trees in the forest.

2.3.3 Multilayer Perceptron

Multilayer perceptron (MLP) is a group of feedforward artificial neural networks with one or more hidden layers. Figure 2.2 presents an example of the structure of MLP with two hidden layers.

Except the input layer, each node in the hidden layers or output layer has a nonlinear activation function. In a one-hidden-layer MLP,

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x)). \tag{2.10}$$

where $b^{(1)}$ and $b^{(2)}$ are bias terms in the hidden and output layer. and $W^{(1)}$ and $W^{(2)}$ are weight matrices, and G and s are activation functions. The output of the hidden layer in a MLP is $h(x) = s(b^{(1)} + W^{(1)}x)$, and two popular functions for s are $tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ and $sigmoid(z) = \frac{1}{1+e^{-z}}$. The final output of a MLP

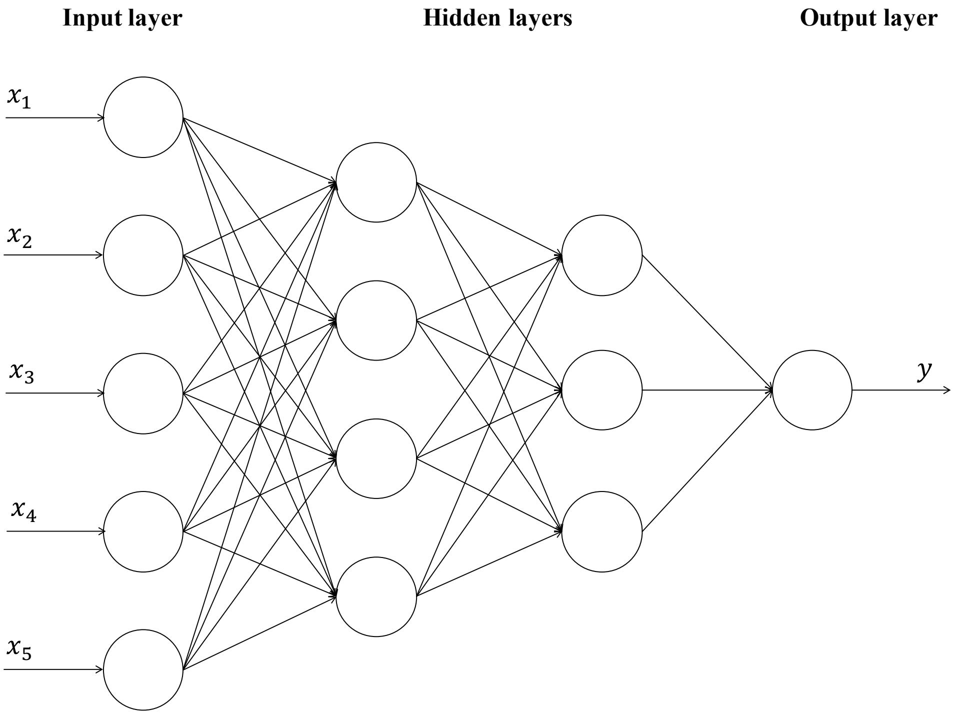


Figure 2.2: The structure of a multilayer perceptron (MLP) with two hidden layers

is $o(x) = G(b^{(2)} + W^{(2)}h(x))$, and the $softmax$ function is usually used for G in classification problems [46].

The MLP is trained by optimizing the weights that can minimize the loss function, where mean square error is often used to compute the loss for regression problems and cross entropy loss is popular for classification problems [47]. The set of weights to optimize is $\theta = \{W^{(2)}, b^{(2)}, W^{(1)}, b^{(1)}\}$ and the backpropagation method is used to compute the gradients [48]. Stochastic gradient descent with mini-batches is usually employed to minimize the loss.

There are two steps in the training process [46, 49]:

1. Forward: the input signal is propagated through the network until the output layer in the forward direction.

2. Backward: the error (loss) is calculated using pre-defined loss function, which is then propagated back through the network in the backward direction. In this process, weights in the network are adjusted.

The two steps are iterated until the loss stops decreasing, i.e., the network is converged. A MLP is able to learn the nonlinear relationships between the input and output vectors, and it demonstrates practical benefits when full theoretical models are not available. However, the disadvantage lies in the difficulties in training and interpretation [49].

2.3.4 Deep Learning

Deep learning has become a very popular machine learning method lately. It has achieved great success in many areas, such as computer vision and NLP. The deep structure and nonlinear activation functions allow deep learning models to learn complex patterns from the data where full theoretical models are not available. Moreover, the availability of big data makes deep learning models a feasible approach for many machine learning problems.

A deep neural network (DNN) is an artificial neural network with many hidden layers, differing from MLPs which usually have no more than two hidden layers. The most intuitive DNN is a type of feedforward neural network with fully connected hidden nodes in adjacent layers, and an example of the structure of DNN is presented in Figure 2.3.

Intuitively, we expect that a deeper MLP is more powerful than a shallow one. However, a deeper MLP introduces too many parameters to optimize which makes it hard for the network to train and would need much more data and time to converge. Thus, a MLP with many layers is not a mainstream method nowadays. Instead, more advanced deep networks have been developed to learn complex representations for classification and other tasks. Here, we present an overview of two popular deep learning models – convolutional neural networks and recurrent neural networks.

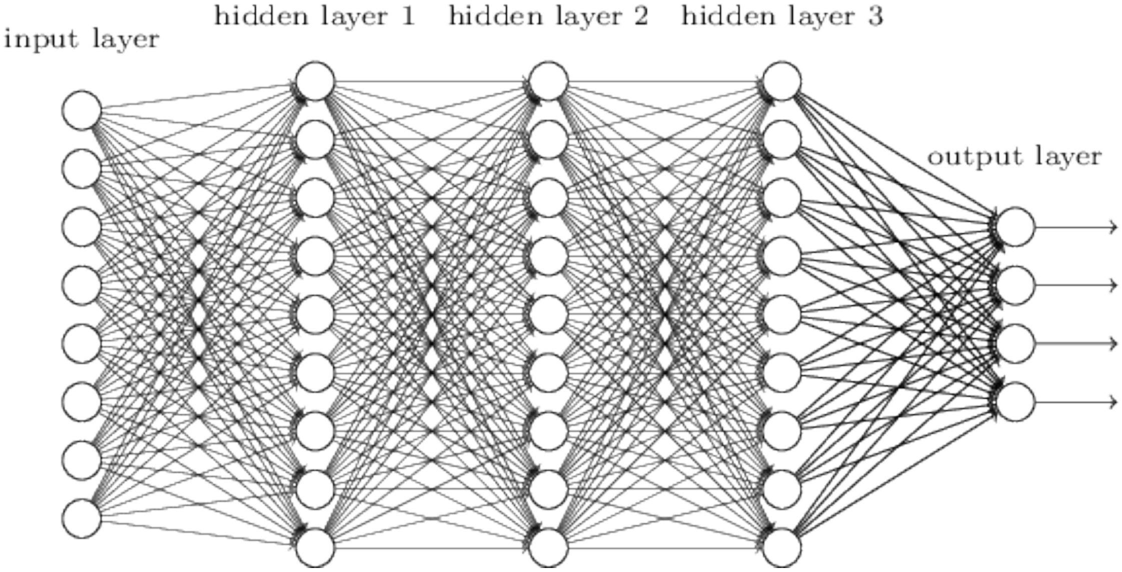


Figure 2.3: The structure of a deep neural network (DNN) [2]

Convolutional Neural Network (CNN)

CNNs are a group of feedforward artificial neural networks with one or more convolutional layers followed by fully connected layers [50]. Figure 2.4 presents the structure of a CNN and there are three major components in a CNN described in the following.

- Local connectivity: Instead of connecting the entire input signal to every neuron in the hidden layer, a CNN only makes connections in small, localized regions of the input signal, which is called a local receptive field of a hidden neuron [51]. Therefore, a neuron learns patterns from small local regions rather than from the entire input signal. Then, we slide the receptive field on the input signal to generate the first layer of feature maps, which is the convolutional layer. In this way, local patterns of the input signal are captured by the network, and this is inspired by how the early visual system works in biology [50].

- Shared weights: A neuron connects to a local receptive field using weights and biases in order to generate feature maps from the input signal. Here, the weights and biases are called filters and usually multiple filters with different sizes are

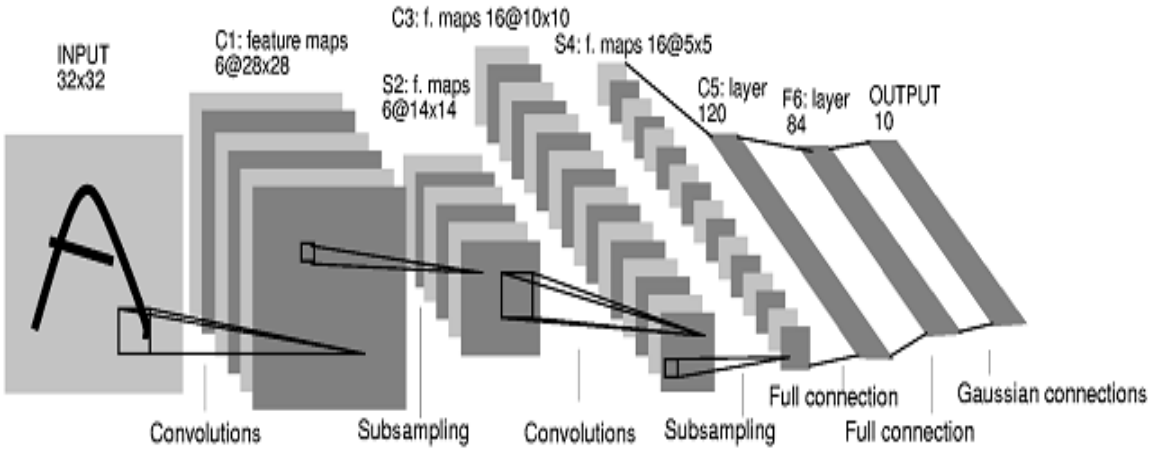


Figure 2.4: The structure of a convolutional neural network (CNN) [3]

employed. All neurons in the same layer use the same set of filters such that the weights and biases are shared among all neurons. The benefit of this strategy is that the number of parameters in the network is much smaller [2].

- Pooling: In CNNs, a convolutional layer is usually followed by a pooling layer such that the learned feature maps are simplified. There are two popular pooling strategies, i.e., max-pooling and $l2$-pooling. Max-pooling selects the most salient information in the feature maps, and $l2$-pooling takes the squared sum to condense the information in the feature maps [2]. Thus, the patterns learned become position-invariant and fewer parameters are needed in the following layers.

Multiple convolutional and pooling layers could be used in the network to learn a hierarchy of abstractions. At the end, fully connected layers are added to gather information learned from the previous layers and to make the final output. In brief, CNNs have a few nice properties, such as learning from local regions, subsampling salient features, learning a hierarchy of abstractions, and reduced number of features. Therefore, CNNs have been very successful in machine learning tasks on images and in many other fields.

Recurrent Neural Network (RNN)

Unlike CNNs which emphasize local connectivity, RNNs are capable of learning sequential information. As previously elaborated, traditional neural networks assume that inputs are independent of each other, while in many cases it is not realistic, such as in NLP fields. For example, it is clear that there is a dependency between the words in a sentence. In a RNN, in addition to taking the current element as the input, the computation on this element is also dependent on that of the previous elements. This resembles the dependencies in natural languages such that RNNs have been very successful in the NLP fields. To describe how a RNN works, we start with the mathematical definition of the computation of the RNN cell, as presented in Equation 2.11.

$$s_t = f(Ux_t + Ws_{t-1}) \tag{2.11}$$

where x_t and s_t are the input and hidden state at time step t. According to Equation 2.11, the hidden state at time step t is based on the current input x_t and the hidden state from the previous time step $t - 1$. W and U are weight matrices and f is an activation function, in which tanh and $ReLU$ are two popular nonlinear choices [52]. The output at time step t is:

$$o_t = g(Vs_t) \tag{2.12}$$

where V is a weight matrix and g is usually a nonlinear function such as $softmax$.

Figure 2.5 presents a graphical illustration of a RNN and its unfolding in time [4]. The RNN unfolded in time is equivalent to a deep feedforward neural network in which the weights are shared at all the hidden layers.

The RNNs are expected to store long-term dependencies, however, it has been demonstrated that it is not capable of learning from very long sequences [4]. Some

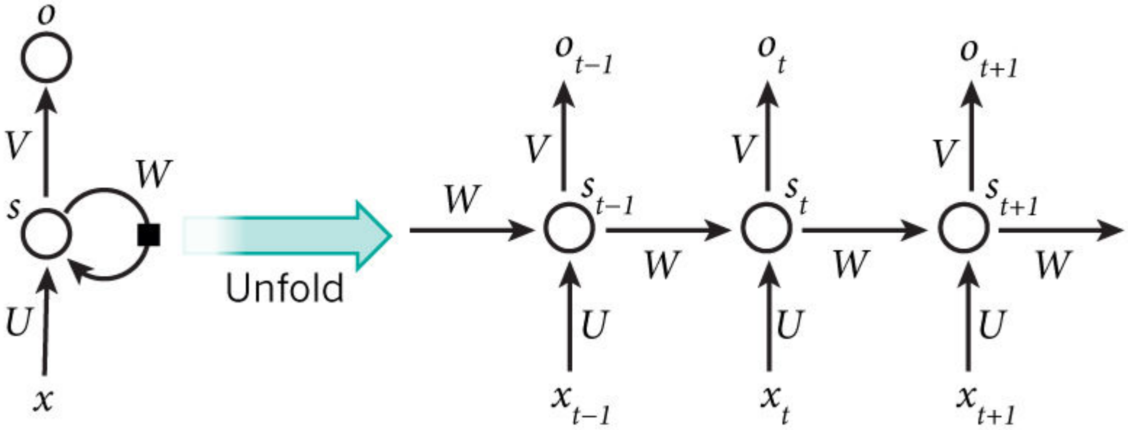


Figure 2.5: A recurrent neural network (RNN) and its unfolding in time [4]

variants have been proposed to address this issue, such as long short term memory and gated recurrent unit, which have been successfully applied to many NLP tasks [53,54].

2.4 Summary

Electronic health data is heterogeneous and cannot be readily expressed in a unified vector space, which makes it hard to be utilized for further modeling and analytics. In this chapter, we present an overview of the popular and state-of-the-art representation learning methods for text documents in the NLP field as well as for EHR data. Finally, we briefly introduce four types of classifiers, including logistic regression, bagging-based ensemble methods, multi-layer perceptron, and deep learning models, which are used in the later chapters for representation learning and predictive modeling. More literature reviews on deep learning approaches are presented in Chapter 5, including gated recurrent unit and the attention mechanism.

Chapter 3

SLR: A Sparse Longitudinal Representation of Electronic Health Record Data

In this chapter, we propose a novel framework to learn a sparse longitudinal representation of patients' medical histories in an EHR database. The proposed method is evaluated with the early detection of diabetes and the predictive performance is compared with widely used baselines. The learned representations are interpreted and visualized to bring clinical insights.

3.1 *SLR* System Model

Given (1) the training set, a patient's health record, consists of a sequence of visits and each visit containing a set of medical codes (shown in Fig. 3.1), namely, s_i, for patient i, and (2) each patient's label (i.e., 1 or 0) representing the outcome (e.g., $y_i = 1$ indicating that patient i has a diagnosis of disease of interest or the targeted outcome in their medical history), our research problem is to find a method to transform each s_i to a unified representation x_i, where $x_i \in R^d$, so as to maximize

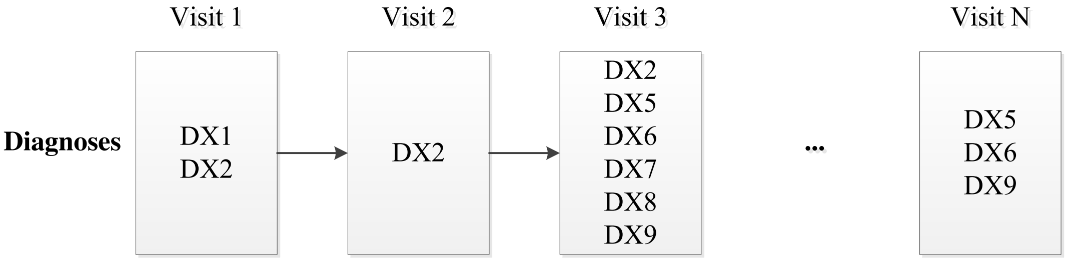


Figure 3.1: An example of a patient's profile in EHR systems

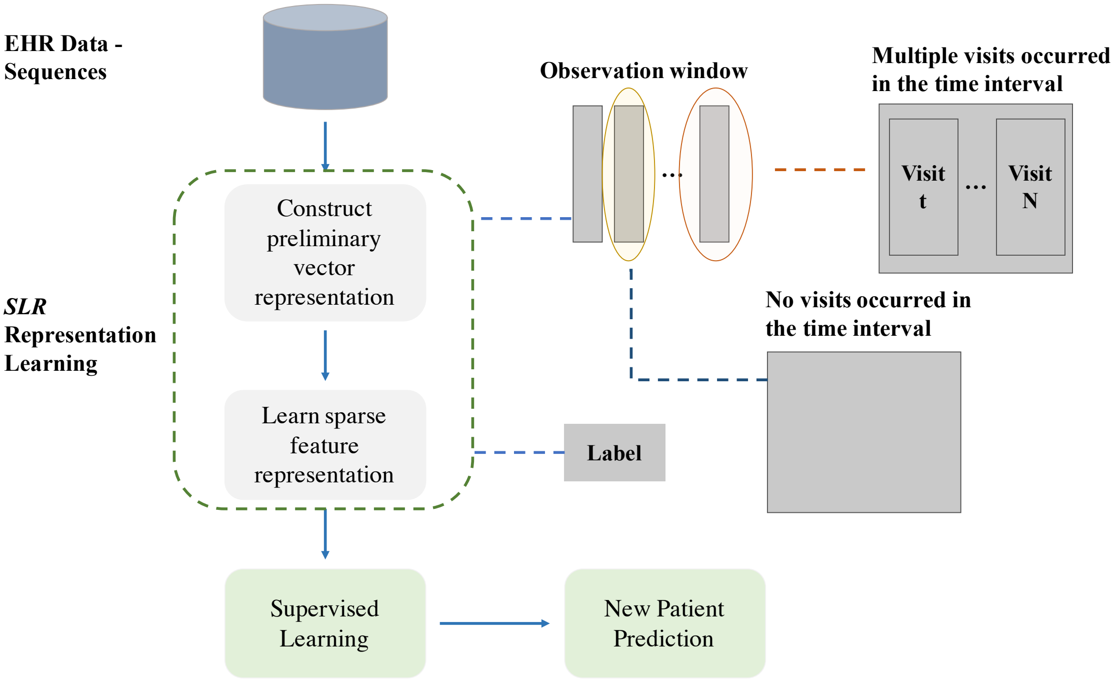

EHR Data - Sequences
Construct preliminary vector representation
Learn sparse feature representation
SLR Representation Learning
Observation window
...
Multiple visits occurred in the time interval
Visit t
...
Visit N
No visits occurred in the time interval
Label
Supervised Learning
New Patient Prediction

Figure 3.2: The *SLR* representation learning framework

the accuracy of any arbitrary classifier. That is:

$$\min \sum_{i=1}^{N} |y_i - \mathit{classifier}(x_i)|, \tag{3.1}$$

where N is the number of patients and the function $\mathit{classifier} : R^d \rightarrow \{0, 1\}$ refers to an arbitrary binary classifier on top of feature space R^d.

The *SLR* framework contains the following three components and Figure 3.2 presents a graphical illustration of the *SLR* framework.

1. **Representation Learning** — Given each patient's medical history and the corresponding label, this step transforms each patient's record into a unified vector space. Further, a sparse group lasso based algorithm [55] is employed to learn a sparse longitudinal representation of the patient's medical history.

2. **Supervised Learning** — Given the learned data representation from Step 1 and the label of each patient, this step trains a predictive model using supervised learning algorithms, such as logistic regression and random forest.

3. **New Patient Prediction** — Given a feature representation from Step 1, this step uses the predictive model (from Step 2) to predict the new patient's label. The outcome of this step is 1 or 0, which refers to whether the patient will have the targeted health outcome.

3.2 *SLR* Representation Learning Algorithm

Preliminary Vector Representation

As presented in Figure 3.1, the medical history of patient i is a sequence s_i with multiple visits and there are one or more clinical events during each visit. Commonly, the time intervals between visits are different intra and across patients. One straightforward way of representing this data is to use the counts or binary indicators of the clinical events in the sequence, however, temporal information is omitted. To cope with this issue, we develop an itemset representation which truncates the sequence s_i into T time windows with even intervals and an itemset of clinical events is constructed for each time window t_j. Then, the counts of clinical events in each time window is computed and concatenated as a preliminary representation x_i of patient i's medical history, i.e., $x_i = \{v_{i1}, v_{i2}, \cdots, v_{ij}, \cdots, v_{iT}\}$ and the associated time windows are $\{t_1, t_2, \cdots, t_j, \cdots, t_T\}$. Here, $v_{ij} = \{c_{j1}^{(i)}, c_{j2}^{(i)}, \cdots, c_{jp}^{(i)}, \cdots, c_{jP}^{(i)}\}$ is the count vector of clinical events $A_1, A_2, \cdots, A_p, \cdots, A_P$ at time window t_j, where P is the total number of distinct clinical events. Thus, all elements in a patient's preliminary representation are presented in Table 3.1. The final preliminary data representation of patient i is $x^{(i)} = \{c_{11}^{(i)}, c_{12}^{(i)}, \cdots, c_{TP}^{(i)}\}$.

Table 3.1: The preliminary vector representation x_i of Patient i

	A_1	A_2	A_3	$\cdots$	A_p	$\cdots$	A_P
t_1	$c_{11}^{(i)}$	$c_{12}^{(i)}$	$c_{13}^{(i)}$	$\cdots$	$c_{1p}^{(i)}$	$\cdots$	$c_{1P}^{(i)}$
t_2	$c_{21}^{(i)}$	$c_{22}^{(i)}$	$c_{23}^{(i)}$	$\cdots$	$c_{2p}^{(i)}$	$\cdots$	$c_{2P}^{(i)}$
t_3	$c_{31}^{(i)}$	$c_{32}^{(i)}$	$c_{33}^{(i)}$	$\cdots$	$c_{3p}^{(i)}$	$\cdots$	$c_{3P}^{(i)}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
t_j	$c_{j1}^{(i)}$	$c_{j2}^{(i)}$	$c_{j3}^{(i)}$	$\cdots$	$c_{jp}^{(i)}$	$\cdots$	$c_{jP}^{(i)}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
t_T	$c_{T1}^{(i)}$	$c_{T1}^{(i)}$	$c_{T1}^{(i)}$	$\cdots$	$c_{Tp}^{(i)}$	$\cdots$	$c_{TP}^{(i)}$

Figure 3.3 illustrates the medical history of an example patient and the construction of itemsets with associated time windows of this patient. In this example, the size of each time window is 4 months and an itemset is constructed for each time window. Then, the counts of clinical events in the itemsets are computed to represent the medical record at each window, while the events not present in the time window are assigned a value 0. Finally, the count vectors of clinical events in each of the time windows are concatenated to construct a preliminary representation of the patient's medical history.

SLR Representation Learning Algorithm

With the preliminary data representation, $x_i = \{v_{i1}, v_{i2}, \cdots, v_{ij}, \cdots, v_{iT}\}$, the proposed method based on the sparse group lasso algorithm [55] aims at learning a sparse set of events in the longitudinal order, as presented in Equation 3.2.

$$min_{\omega} L(X\omega, y) + (1-\alpha)\lambda \sum_{j=1}^{T} \sqrt{P} ||\omega_j||_2 + \alpha\lambda ||\omega||_1 \tag{3.2}$$

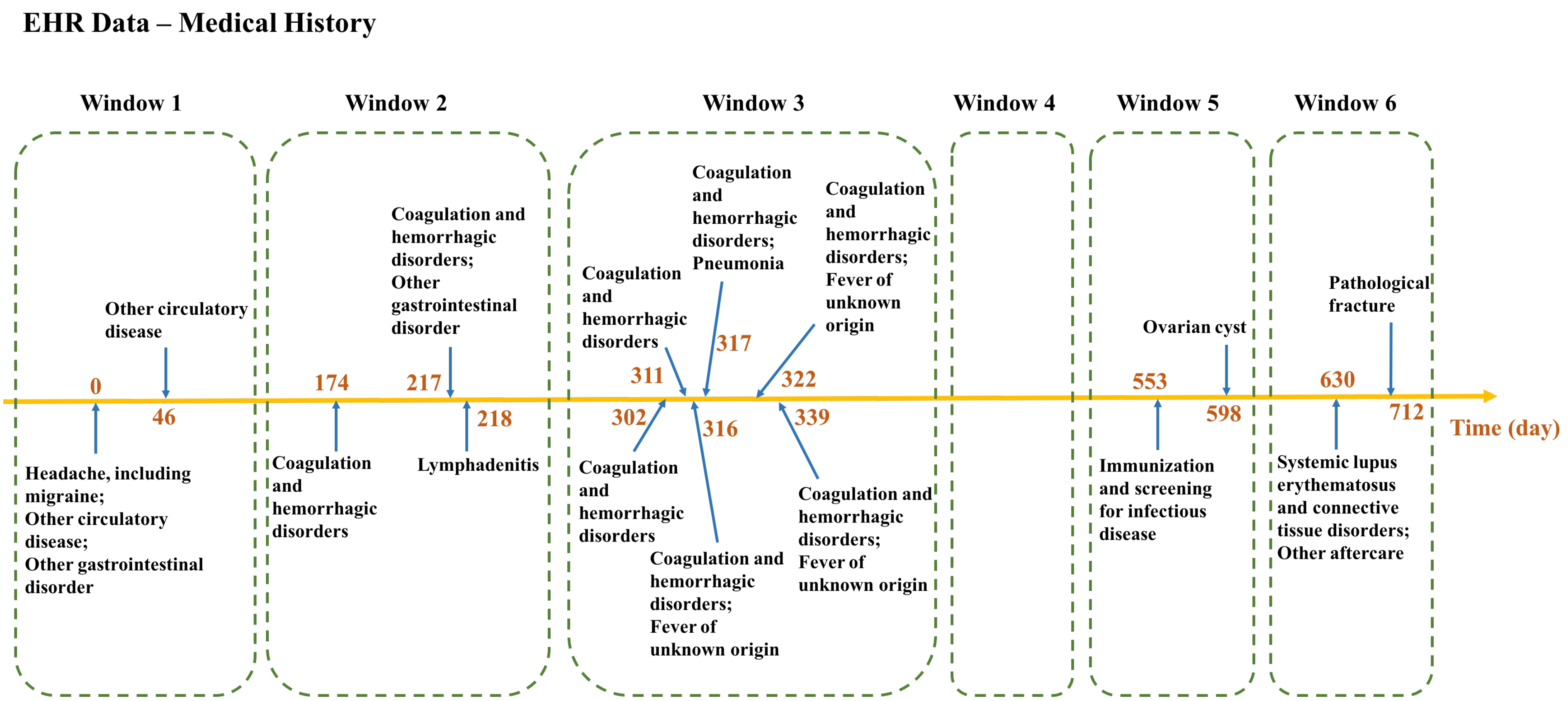

EHR Data – Medical History
Window 1
Window 2
Window 3
Window 4
Window 5
Window 6
Other circulatory disease
0
46
Headache, including migraine; Other circulatory disease; Other gastrointestinal disorder
174
Coagulation and hemorrhagic disorders
217
Coagulation and hemorrhagic disorders; Other gastrointestinal disorder
218
Lymphadenitis
Coagulation and hemorrhagic disorders
311
302
Coagulation and hemorrhagic disorders
Coagulation and hemorrhagic disorders; Pneumonia
317
316
Coagulation and hemorrhagic disorders; Fever of unknown origin
Coagulation and hemorrhagic disorders; Fever of unknown origin
322
339
Coagulation and hemorrhagic disorders; Fever of unknown origin
553
Immunization and screening for infectious disease
Ovarian cyst
598
630
Systemic lupus erythematosus and connective tissue disorders; Other aftercare
Pathological fracture
712
Time (day)

Figure 3.3: The EHR data of an example patient and the construction of time windows

where P is the number of distinct clinical events. j is the jth time window ($j \in \{1, 2, \cdots, T\}$) and $\omega_j = \{\omega_{j1}, \omega_{j2}, \cdots, \omega_{jP}\}$. $L(X\omega, y)$ is the loss function shown in Equation 3.3.

$$L(X\omega, y) = \frac{1}{N} \sum_{i=1}^{N} log(1 + exp(-y_i x_i^{\intercal} \omega)) \tag{3.3}$$

where N is the total number of patients. The objective function in *SLR* learning consists of three parts: error minimization, $l2$ penalty on each time window of features, and a $l1$ sparsity term on each feature throughout the entire medical history. Thus, *SLR* learns a sparse longitudinal representation of the medical history by minimizing the error in Equation 3.2. We propose Algorithm 1 to learn a sparse longitudinal representation from the EHR data.

This problem is the sum of a convex function and a penalty term, so that the coordinate descent algorithm is employed to learn the global optimum [55]. When $\alpha = 1$, Equation 3.2 becomes lasso, which learns a sparse representation of the clinical events, while it becomes group lasso when $\alpha = 0$, which is equivalent to ridge regression on groups of features. Here, α and λ are two tuning parameters to control the degree of sparsity in the *SLR* representation.

3.3 Evaluation

3.3.1 Background

Diabetes mellitus is a group of metabolic diseases characterized by hyperglycaemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion and/or insulin action [56, 57]. According to the American Diabetes Association, there are generally two etiopathogenetic categories of diabetes, type 1 and type 2 diabetes. The cause of the former one is absolute insulin deficiency. The latter one is more prevalent in the population and the cause is a combination

Algorithm 1: *SLR* Representation Learning

```
Input  : Set of sequences $S$, set of events $A$, number of time intervals $T$,
         number of patients $N$, number of distinct events $P$
Output: Set of selected events $A_s$
begin
    Construct Preliminary Data Representation $X$:
    Initialization:
    $X \leftarrow \phi$;
    for  $i = 1$ to $N$ do
        $x_i \leftarrow \phi$
        for  $j = 1$ to $T$ do
            $c_j^{(i)} \leftarrow \{c_{j1}^{(i)}, \cdots, c_{jp}^{(i)}, \cdots, c_{jP}^{(i)}\}$
            $x_i \leftarrow \{x_i, c_j^{(i)}\}$
        $X_i \leftarrow x_i$
    Learn SLR Representation $A_s$:
    Initialization:
    $A_s \leftarrow \phi$;
    Block coordinate descent to optimize (3.2): $\hat{\omega}$
    for $j = 1$ to $T$ do
        $\omega_j \leftarrow \{\omega_{j1}, \cdots, \omega_{jp}, \cdots, \omega_{jP}\}$
        if $\omega_j \neq \mathbf{0}$ then
            for $p = 1$ to $P$ do
                if $\omega_{jp} \neq 0$ then
                    $A_s \leftarrow A_s \cup \{A_{jp}\}$
    return $A_s$
```

of resistance to insulin action and an inadequate compensatory insulin secretory response [56]. According to the National Diabetes Statistics Report by the Centers for Disease Control and Prevention (CDC), 30.3 million Americans, i.e., 9.4% of the U.S. population, have diabetes in 2017 [58, 59].

Diabetes is now the third leading cause of death in the United States, which was ranked the seventh in 2015. The number of people diagnosed with diabetes has more than tripled over the last couple of decades [60]. Diabetes may cause other complications, such as kidney disease, vision loss, heart disease, high blood pressure, obesity, and depression [14, 60]. According to [61], almost 40% of patients with type 2 diabetes are suffering from chronic kidney disease and that diabetes is the leading cause of it in the United States. Diabetes also brings a huge economic burden. According to [62], the total estimated economic costs of diagnosed diabetes in the United States is $245 billion with $176 billion of medical costs and $69 billion due to reduced productivity. The medical expenditure of diabetic patients is approximately 2.3 times higher than the individuals without diabetes. Moreover, the economic costs are still increasing, and there has been a 41% increase from costs in 2007. In brief, diabetes imposes a substantial burden on the society in addition to the suffering of patients.

However, diabetes is often undiagnosed or misdiagnosed. Statistics from the American Diabetes Association shows that approximately 24% are undiagnosed among all diabetic patients [59]. Uncontrolled or untreated diabetes can lead to serious complications, such as blindness, heart disease, infections, neuropathy, or even death [63]. The early detection of diabetes can help patients take actions early and make their diabetes under control to reduce the risk of complications.

Hence, the *SLR* representation learning framework is applied to recognizing the misdiagnosed or undiagnosed diabetes using patients' EHR data. One of the objectives is to improve the performance of classifiers with the proposed representation, and the other objective is to bring clinical insights to deepen the understanding of disease correlations. In the rest of this chapter, we use the terminology "prediction" to indicate the early recognition of undiagnosed or misdiagnosed diabetes.

3.3.2 Experimental Design

In this research, we use the de-identified EHR data of 75 months beginning in September 2010 from the University of Virginia Health System. This dataset contains $2,343,651$ inpatient and outpatient visits of $473,915$ distinct patients. In this experiment, we focus on the diagnosis codes in the EHR data. Here, the diagnoses are coded primarily in ICD-9 and a small portion are ICD-10 codes. The raw diagnoses codes are further clustered into 283 groups according to the Clinical Classification Software from the Agency for Healthcare Research and Quality (AHRQ) [64].

In the experiment, we define the observation window and prediction period to validate the proposed method. For patients with diabetes diagnoses, we first extract the visits between 1.5 years to half a year prior to the first observed diagnoses of diabetes. The visits within 6 months of the first observed diabetic diagnosis are excluded for the purpose of early detection. Although the type 1 and type 2 diabetes are not distinguished in the AHRQ clinical classification, the majority are patients with type 2 diabetes. Additionally, Figure 3.4 illustrates the age distribution of diabetic patients with histogram and density plot, demonstrating that this population are primarily consist of elder patients.

For patients without observed diabetes diagnoses, we exclude the individuals when their visit histories are shorter than four years. Visits occurred in the first 12 months are used as observations for modeling, while the following 3 years are monitored to ensure the patients are not diagnosed with diabetes in the near future. Thus, a 12-month observation window is constructed for patients in each group, and a prediction window is defined to allow observation of the outcome. To better illustrate

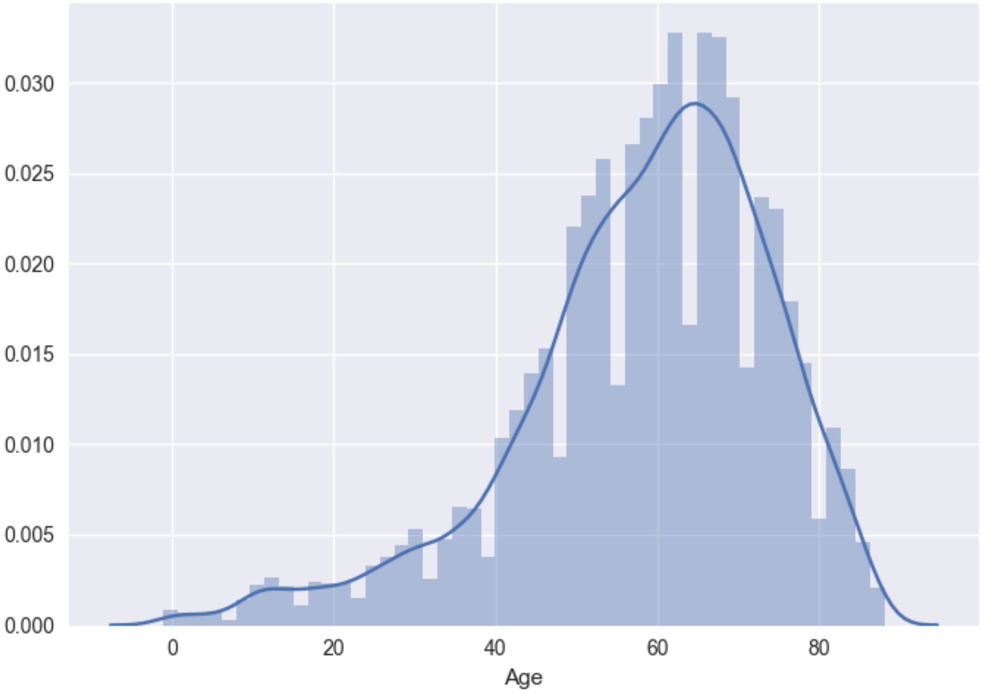


Figure 3.4: A graphical illustration of the age distribution of diabetic patients

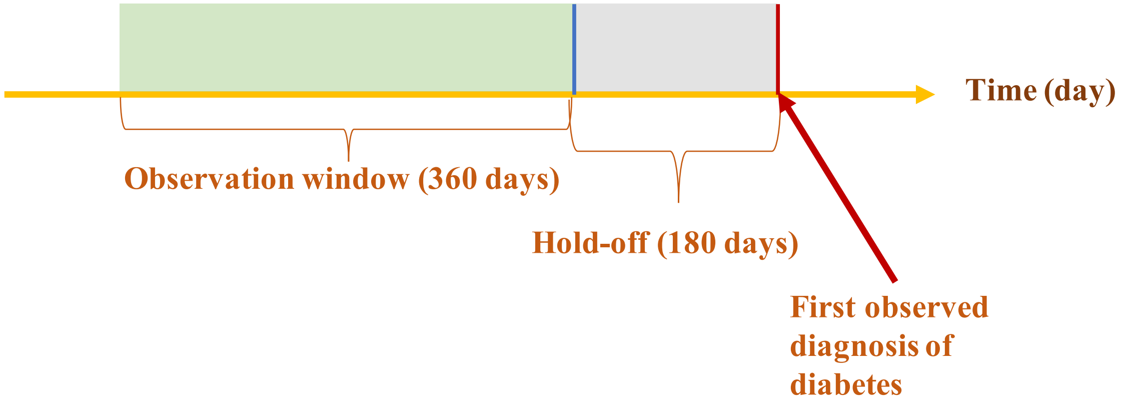

Time (day)
Observation window (360 days)
Hold-off (180 days)
First observed diagnosis of diabetes

Figure 3.5: A graphical illustration of the experimental setting for the early detection of diabetes

the experimental setting, we present the observation window, hold-off and onset of outcome event in Figure 3.5

Accordingly, there are $3,112$ and $27,555$ patients in the target and control groups, respectively. Table 3.2 presents the preliminarily selected diagnoses, which includes the diagnoses appear in at least 5% of patients' records in both populations. These preliminary diagnoses are used as the basis of the *SLR* framework. To validate the proposed representation learning framework, we compare the prediction performance of the proposed model with baseline approaches as follows.

- Aggregated Frequency Vector (AFV): A patient's record is represented as a count vector of medical events in the observation window. Each dimension is associated with a distinct medical event in patients' EHR data.

- Bag-of-pattern in Sequences (BPS): This method runs a widely used sequence pattern mining algorithm, CM-SPADE [34], to discover all frequent patterns in patients' EHR data. Thus, a patient's record is represented with a frequency vector of the discovered frequent patterns.

- Aggregated Transition Vector (ATV): The pairwise transitions between medical event pairs are counted and a patient's record is represented with a frequency

Table 3.2: The preliminarily selected diagnoses for the early detection of diabetes

Index	Diagnoses
1	Immunizations and screening for infectious disease
2	Coronary atherosclerosis and other heart disease
3	Nonspecific chest pain
4	Cardiac dysrhythmias
5	Other circulatory disease
6	Other lower respiratory disease
7	Esophageal disorders
8	Other liver diseases
9	Other gastrointestinal disorders
10	Chronic kidney disease
11	Genitourinary symptoms and ill-defined conditions
12	Nonmalignant breast conditions
13	Osteoarthritis
14	Other non-traumatic joint disorders
15	Spondylosis; intervertebral disc disorders; other back problems
16	Other connective tissue disease
17	Other bone disease and musculoskeletal deformities
18	Other injuries and conditions due to external causes
19	Abdominal pain
20	Malaise and fatigue
21	Medical examination/evaluation
22	Other aftercare
23	Other screening for suspected conditions (not mental disorders or infectious disease)
24	Other and unspecified benign neoplasm
25	Thyroid disorders
26	Nutritional deficiencies
27	Disorders of lipid metabolism
28	Other nutritional; endocrine; and metabolic disorders
29	Deficiency and other anemia
30	Mood disorders
31	Screening and history of mental health and substance abuse codes
32	Other nervous system disorders
33	Essential hypertension

vector of pairwise transitions.

In the experiment, 80% of patients are randomly selected from each group as training data, while the rest are used for testing. The final predicted labels are assigned by comparing the model output with the classification threshold tuned on the training set. The time window used to group clinical events is 90 days such that 4 subsets of features are constructed for the 1-year observation window. The size of time window is determined according to the U.S. Census Bureau stating that adults have an average of 3.9 visits to doctors annually [65]. However, this time window can be further tuned on specific populations for an optimal output. With the *SLR* and baseline representations of patients' medical records, we model the risk of diabetes using three classifiers, logistic regression with $l1$ penalty (LR), random forest (RF), and gradient boosting trees (GBT). Here, the RF and GBT consist of 100 trees, respectively, and the γ value in the $l1$ penalty term of the LR is tuned with 10-fold cross validation. The prediction performance of baselines and the proposed framework on the early detection task is evaluated with AUC (area under curve), sensitivity, specificity, and F2 score. Each experiment is repeated 50 times and we calculate the averages and standard deviations of the above metrics, respectively.

3.3.3 Experimental Results

The predictive performance of the classifiers based on the baselines and proposed *SLR* framework are presented in Table 3.3. Here, the results shown are based on the *SLR* learned with $\alpha = 0.8$ and $\lambda = 0.00016$.

According to the results table, the predictive performance of classifiers based on *SLR* outperform the baselines with a significant improvement in terms of sensitivity, AUC, and F2 score. The three classifiers based on *SLR* have a relatively high sensitivity greater than 80%, and the specificity are around 62%. The performance between the baselines are not significantly different, while the GBT classifiers achieve

Table 3.3: The predictive performance of baselines and the proposed *SLR* framework

		Sensitivity	Specificity	AUC	F2 score
AFV	LR	0.827 ± 0.012	0.332 ± 0.006	0.640 ± 0.011	0.283 ± 0.015
	RF	0.302 ± 0.016	0.830 ± 0.005	0.599 ± 0.010	0.444 ± 0.009
	GBT	0.694 ± 0.014	0.541 ± 0.005	0.670 ± 0.009	0.451 ± 0.009
BPS	LR	0.829 ± 0.012	0.333 ± 0.006	0.637 ± 0.009	0.290 ± 0.016
	RF	0.683 ± 0.011	$\mathbf{0.950 \pm 0.002}$	0.611 ± 0.011	0.446 ± 0.009
	GBT	0.699 ± 0.013	0.539 ± 0.006	0.674 ± 0.009	0.449 ± 0.009
ATV	LR	0.830 ± 0.014	0.338 ± 0.006	0.643 ± 0.010	0.447 ± 0.009
	RF	0.266 ± 0.011	0.902 ± 0.004	0.638 ± 0.008	0.268 ± 0.015
	GBT	0.506 ± 0.017	0.725 ± 0.005	0.656 ± 0.013	0.383 ± 0.016
SLR	LR	0.859 ± 0.010	0.620 ± 0.005	0.802 ± 0.007	0.563 ± 0.012
	RF	0.818 ± 0.011	0.630 ± 0.006	0.807 ± 0.006	0.579 ± 0.011
	GBT	$\mathbf{0.880 \pm 0.009}$	0.600 ± 0.006	$\mathbf{0.822 \pm 0.007}$	$\mathbf{0.583 \pm 0.011}$

slightly better results than the other two classifiers in general. Additionally, the LR classifiers based on baseline representations emphasize more on the positive class such that they achieve high sensitivity but very low specificity. However, the RF classifiers tend to have a high specificity but low sensitivity. The GBT models are capable of achieving a more balanced prediction in the two classes. Overall, the proposed *SLR* framework is capable of improving the prediction performance.

To illustrate the impact of the tuning parameters in the two penalty terms, Figure 3.6 presents the numbers of features selected using the proposed framework with different α and λ, respectively. Accordingly, we found that using a larger λ with fixed α obtains a sparser set of features. This trend holds true for all α values from 0 to 1. In the *SLR* framework, α balances the degree of sparsity over all features and the penalization in groups of features, where the groups are constructed corresponding to the time window in which the event occurs. A larger α indicates that the algorithm is more toward sparsity on events occurred throughout the entire observation window, while a smaller α tends to penalize more at the group level. According to Figure 3.6, a larger α demonstrates a smoother change in terms of the number of selected features, while a smaller α focuses more on the group-level penalty such that the changes in

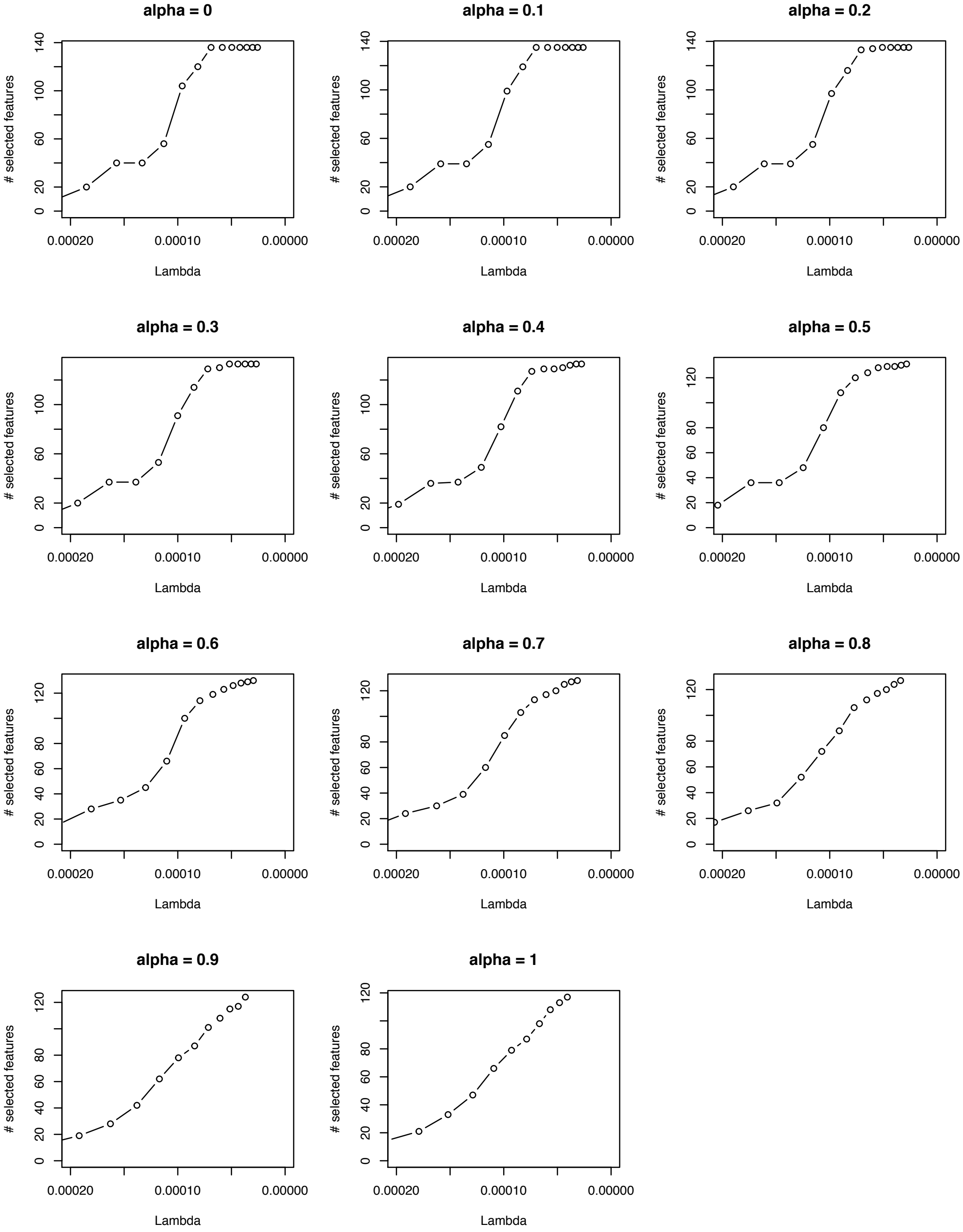


Figure 3.6: Numbers of features selected by *SLR* with different regularization strengths

the number of selected features are less smooth.

In addition, we also compare the model performance when α changes from 0 to 1, as shown in Figure 3.7. Accordingly, the predictive performance of the model increases when λ decreases with fixed α. When $\alpha = 0$, Equation 3.2 only focuses on the group-level penalty and a larger λ forces many groups of features to be discarded which results in a low AUC. Similarly, the increase of AUC when λ decreases is not smooth in such cases.

In brief, Figures 3.6 and 3.7 further demonstrate that we can balance the group-level penalty and feature-level sparsity by α, and λ controls the penalty on both levels. Hence, increasing λ gives more sparse representation in general, while a larger α emphasizes more on the feature-level sparsity and a smaller α shows more penalty on the group-level.

3.3.4 Visualization & Interpretation

To get a better understanding of the learned *SLR* representation, we illustrate the correlations between the selected features in the observation window and the outcome, i.e., diabetes mellitus, in Figure 3.8. We observe that the diagnoses temporally closer to the observation of the outcome are more positively correlated while the selected features that are negatively correlated are mostly in the first time window, i.e., relatively much earlier than the outcome onset. In the very early time window, the events are either learned as uncorrelated or have negative correlations with the outcome. Such negatively correlated events are *other screening for suspected conditions*, *medical examination/evaluation*, etc.

Disorders of lipid metabolism and *chronic kidney disease* (CKD) are the most positively correlated diagnoses with the outcome, and the correlations are stronger when they are closer to the onset of the outcome. Here, the *disorders of lipid metabolism* includes hypercholesterolemia, hyperglyceridemia, hyperlipidemia, lipoprotein defi-

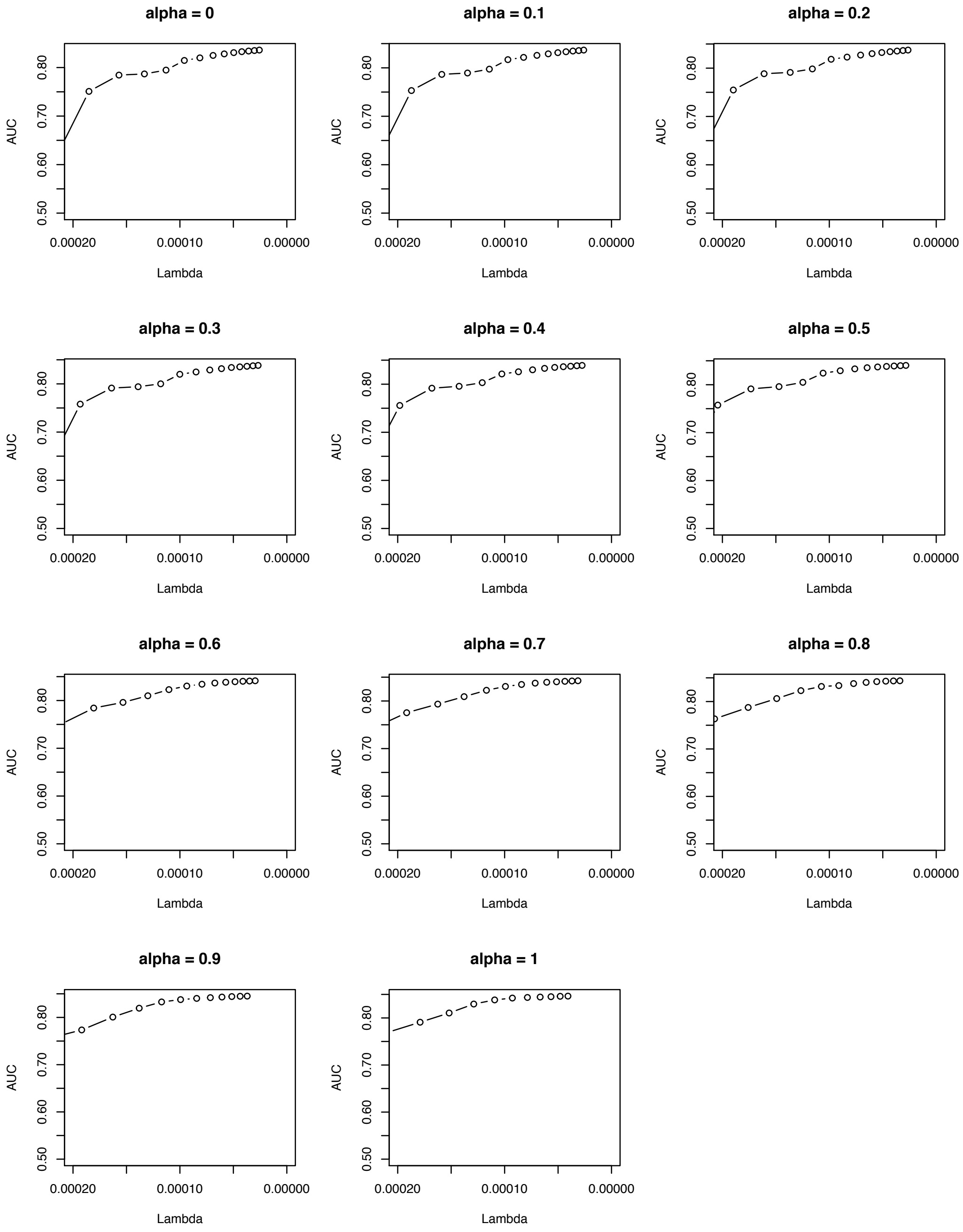


Figure 3.7: The predictive performance with features selected by *SLR* with different regularization strengths

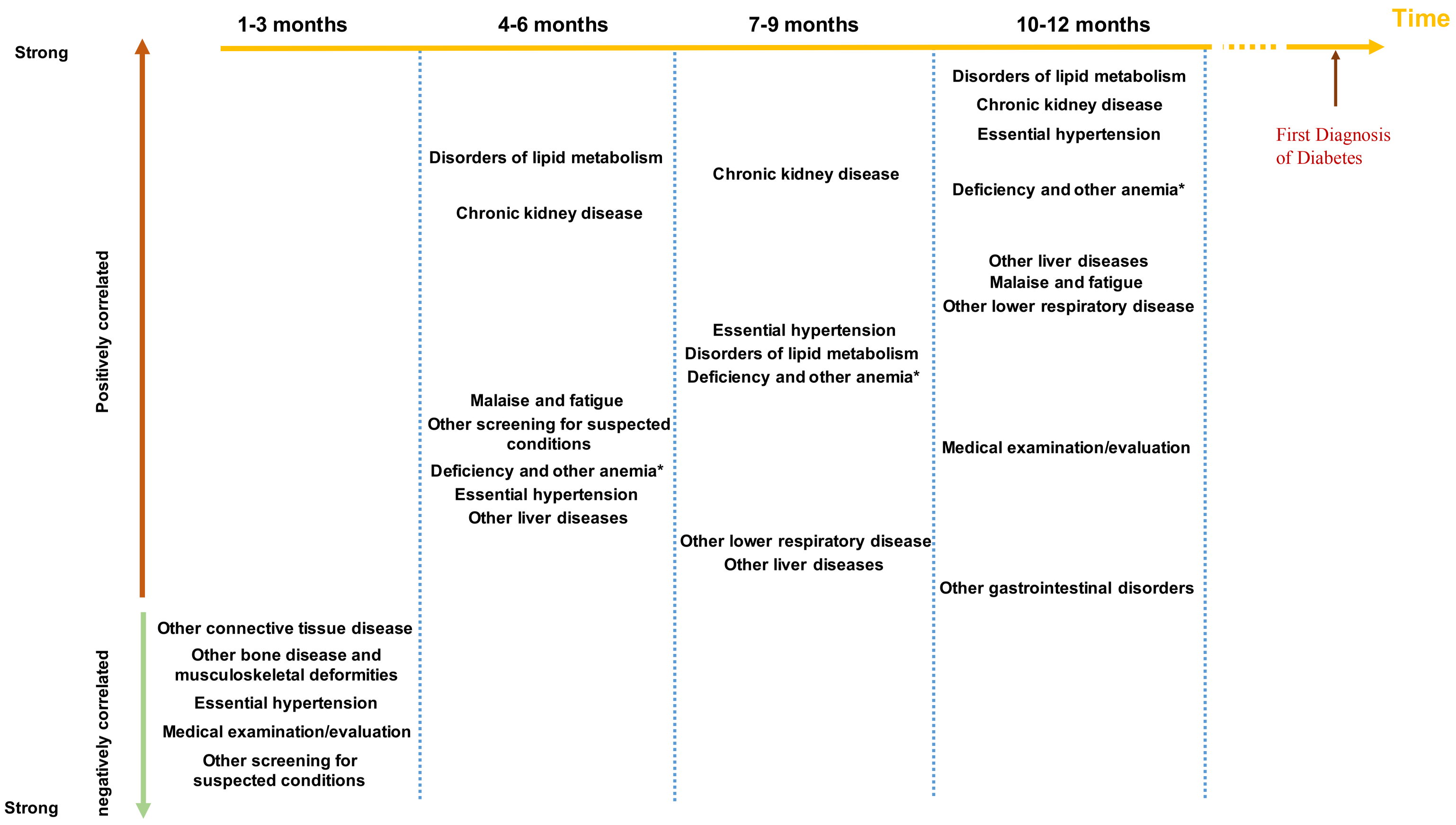

1-3 months
4-6 months
7-9 months
10-12 months
Time
Strong
Positively correlated
negatively correlated
Strong
First Diagnosis of Diabetes
Other connective tissue disease
Other bone disease and musculoskeletal deformities
Essential hypertension
Medical examination/evaluation
Other screening for suspected conditions
Disorders of lipid metabolism
Chronic kidney disease
Malaise and fatigue
Other screening for suspected conditions
Deficiency and other anemia*
Essential hypertension
Other liver diseases
Chronic kidney disease
Essential hypertension
Disorders of lipid metabolism
Deficiency and other anemia*
Other lower respiratory disease
Other liver diseases
Disorders of lipid metabolism
Chronic kidney disease
Essential hypertension
Deficiency and other anemia*
Other liver diseases
Malaise and fatigue
Other lower respiratory disease
Medical examination/evaluation
Other gastrointestinal disorders

Figure 3.8: Selected features by *SLR* and their correlations to the first diagnosis of diabetes

ciencies, and so forth [64]. According to [66, 67], hypertriglyceridemia and hyperlipidemia are common lipid abnormality in persons with type 2 diabetes. Research has also shown that the presence of hypertriglyceridemia may be used to predict the presence of coexistent diabetes since abnormalities in triglyceride metabolism may be a fundamental factor in the pathogenesis of diabetes [68]. Additionally, it is not rare to see lipoprotein abnormalities in untreated, hyperglycemic diabetic patients [69]. As to *chronic kidney disease*, previous literature has shown that the prevalence of CKD is high among people with diagnosed diabetes (39.6%), undiagnosed diabetes (41.7%) and prediabetes (17.7%), compared to the general population without diabetes (10.6%) [70].

Moreover, *essential hypertension*, *deficiency and other anemia*, *malaise and fatigue*, and *other liver disease* are also positively correlated with the outcome. Here, *deficiency and other anemia* refers to iron deficiency anemias, other deficiency anemias, and/or hereditary anemias according to [64]. Numerous medical research have shown that these are prevalent symptoms in prediabetic and diabetic patients [71–75]. According to Mayo Clinic, *essential hypertension* and diabetes are linked by hyperinsulinemia, which is often caused by insulin resistance and might eventually lead to type 2 diabetes [71, 76, 77]. Obesity is the most common cause of insulin resistance and compensatory hyperinsulinemia and it is a strong risk factor of elevated blood pressure [78]. Moreover, diabetes and *essential hypertension* are the leading causes of *chronic kidney disease*, where we also observe the correlations between these diseases in Figure 3.8. However, *essential hypertension* and *other liver disease* occurred much earlier show a weak correlation with the outcome than those diagnoses occurred later and temporally closer to the outcome onset. In addition, the diagnoses of *other gastrointestinal disorders* occurred closer to the first observed diabetes is weakly correlated with it.

In general, the learned *SLR* representation is consistent with the previous medical

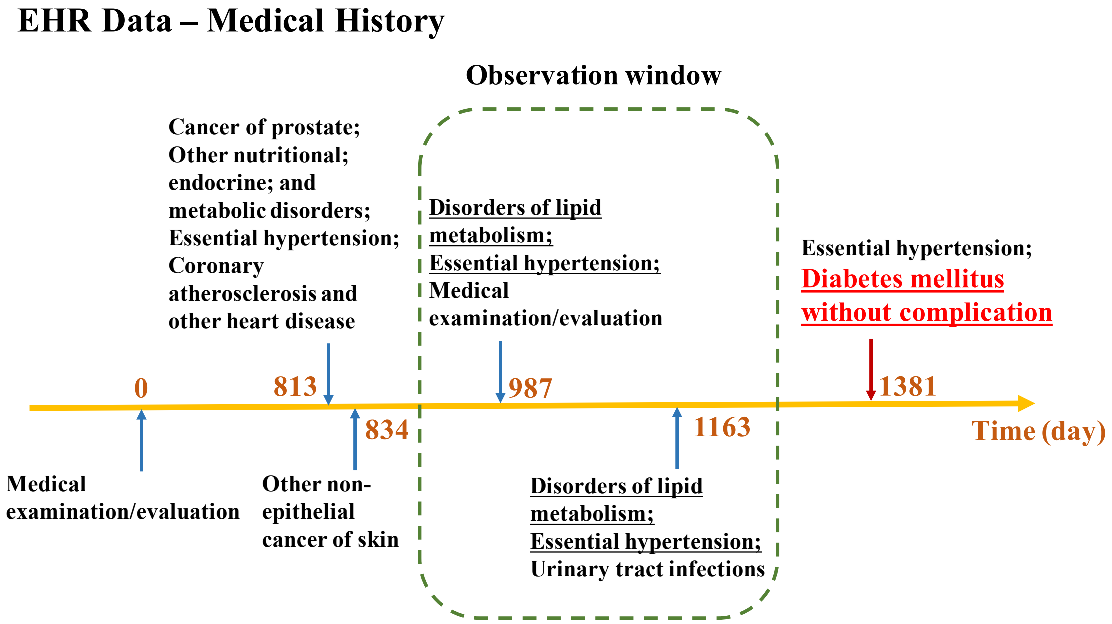

EHR Data – Medical History
Observation window
Cancer of prostate; Other nutritional; endocrine; and metabolic disorders; Essential hypertension; Coronary atherosclerosis and other heart disease
Disorders of lipid metabolism; Essential hypertension; Medical examination/evaluation
Essential hypertension; Diabetes mellitus without complication
0
813
987
1381
834
1163
Time (day)
Medical examination/evaluation
Other non-epithelial cancer of skin
Disorders of lipid metabolism; Essential hypertension; Urinary tract infections

Figure 3.9: The medical history of an example diabetic patient

research. In addition to improving the accuracy of early detection of diabetes, this research could also be used to deepen clinical understanding of disease correlations.

To provide a more straightforward understanding of the learned representation and its benefit to the early detection of diabetes, we present two example patients who are diagnosed with diabetes at least 18 months after their first visits recorded in the EHR system in Figures 3.9 and 3.10. We observe that some initial diagnoses in the patients' medical records are unrelated to diabetes, while later in the observation window some strong risk factors appear, such as *disorders of lipid metabolism* and *essential hypertension*, highlighted with underscore in the figures. Both example patients are predicted as positive by the classifiers based on the *SLR* representation.

3.4 Summary

In this chapter, we propose a representation learning framework, *SLR*, to learn a sparse longitudinal representation of EHR data. This work improves the performance of predictive models as well as deepens the understanding of disease correlations. We

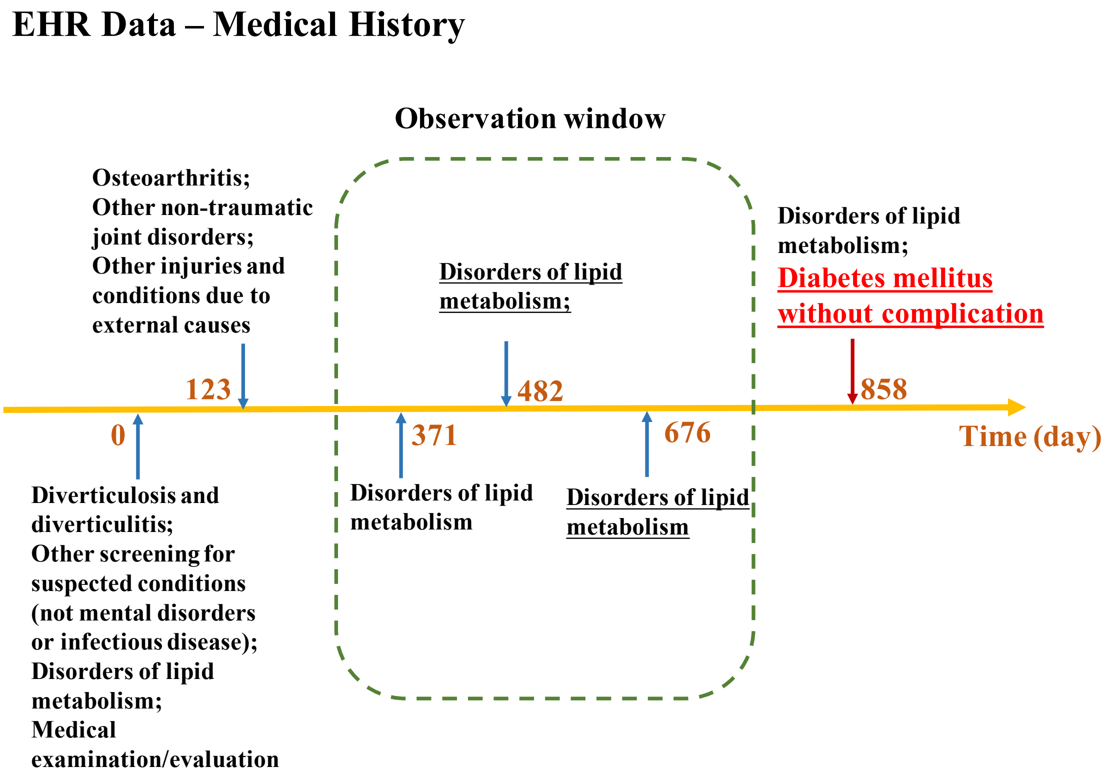

EHR Data – Medical History
Observation window
Osteoarthritis;
Other non-traumatic joint disorders;
Other injuries and conditions due to external causes
123
Disorders of lipid metabolism;
482
Disorders of lipid metabolism;
Diabetes mellitus without complication
858
0
Diverticulosis and diverticulitis;
Other screening for suspected conditions (not mental disorders or infectious disease);
Disorders of lipid metabolism;
Medical examination/evaluation
371
Disorders of lipid metabolism
676
Disorders of lipid metabolism
Time (day)

Figure 3.10: The medical history of a second example diabetic patient

apply this framework to the early detection of diabetes using patients' longitudinal EHR data. The experimental results demonstrate that the proposed *SLR* representation is capable of achieving a more accurate prediction and that it also uncovers the correlations between diseases which are found to be consistent with previous medical research.

One limitation of this framework is multicollinearity with clinical events occurring in multiple time windows. However, penalty terms are used to reduce overfitting of data in such models. Other possibilities to avoid overfitting can be explored in future work. Another limitation is that we use the general diagnosis categories of ICD-9 and ICD-10 codes from the AHRQ classification scheme [64]. The current coarse categories might introduce information loss, so future work will consider more specific diagnosis codes and other grouping strategies. Additionally, the clinical events can be expanded to include procedures and medications for a more comprehensive representation of patients' medical histories.

Chapter 4

WB-SLR: Weighted Bagging of Sparse Longitudinal Representation of Electronic Health Record Data

In this chapter, we propose a novel framework to learn a sparse longitudinal representation of EHR data based on the bagging strategy. The proposed method is evaluated with the early detection of chronic kidney disease (CKD) in diabetic patients and the predictive performance is compared with widely used baselines. The learned representations are interpreted and visualized to bring clinical insights.

4.1 *WB-SLR* System Model

Given patients' longitudinal EHR data, the research problem remains the same as in Chapter 3, i.e., to find a method to transform each sequence s_i into a unified representation for an accurate prediction. In this section, we propose *WB-SLR* which learns a weighted bagging of *SLR*s to provide a more stable and comprehensive representation.

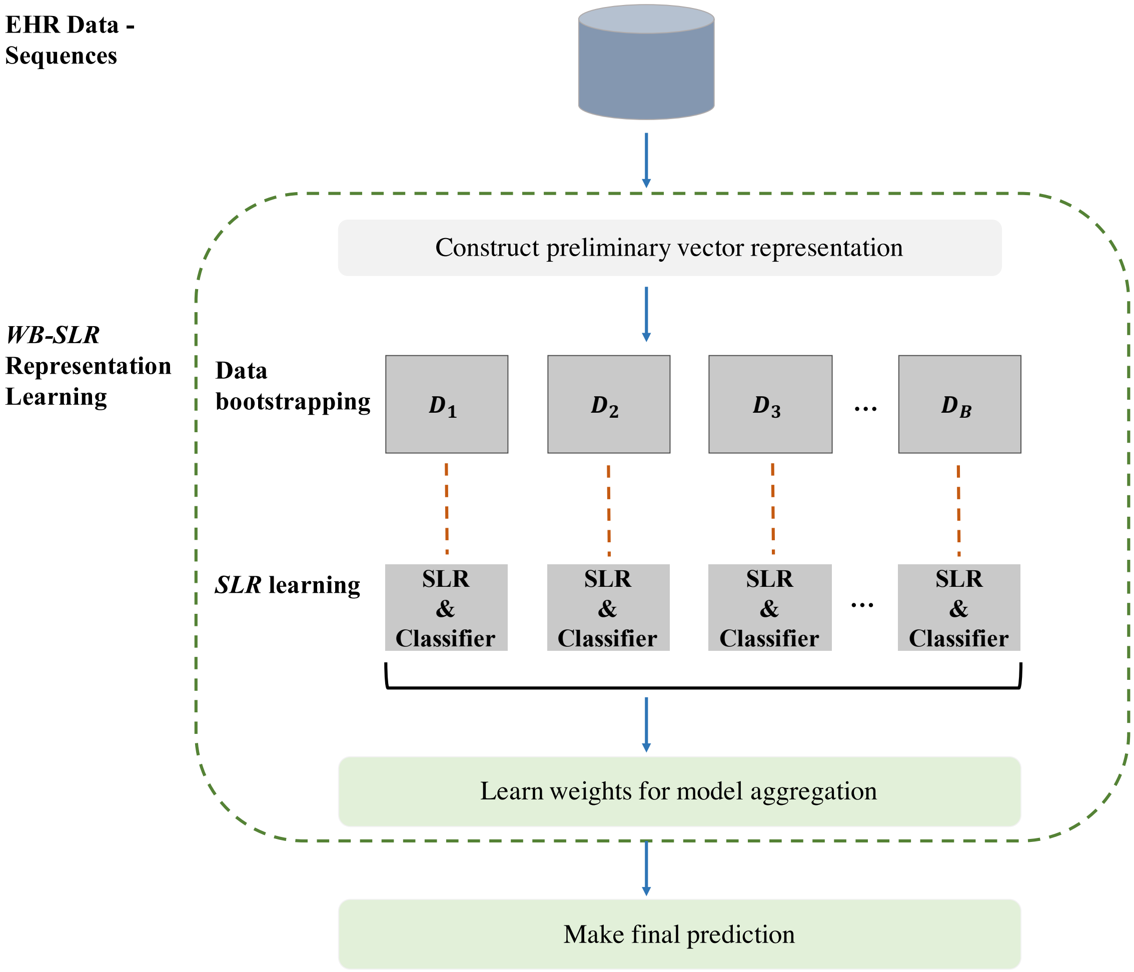


Figure 4.1: The *WB-SLR* representation learning framework

The *WB-SLR* framework consists of four steps and Figure 4.1 presents a graphical illustration of the designed framework.

1. **Constructing Preliminary Representation** — Given each patient's medical record, this step transforms a sequence into the preliminary vector representation of clinical events at multiple time windows, as elaborated in 3.2.

2. **Learning *SLR* Ensemble** — Given the training set of preliminary vector representations, this step first performs bootstrapping on the training set. Then, a *SLR* representation and an associated logistic regression classifier are learned on each bootstrapped sample.

3. **Learning Weights for Model Aggregation** — Given the ensemble of classifiers and associated *SLR*s, the classifiers are aggregated linearly with weights such that the oob error is minimized.

4. **Making Final Prediction** — This step makes the final prediction for each patient by weighted aggregation of the output from each single classifier in the ensemble.

4.2 *WB-SLR* Representation Learning Algorithm

In this section, we elaborate the details of the *WB-SLR* representation learning framework consisting of three parts – *SLR* ensemble learning, weighted model selection, and new patient prediction, as described in the following.

SLR Ensemble Learning

There are two steps to obtain the ensemble of *SLR*s and the associated classifiers:

1. **Constructing data bootstraps** — Given the training set of preliminary vector representation D, we sample B bootstraps, denoted as $\{D_1, \cdots, D_b, \cdots, D_B\}$. Each bootstrap D_b is created by randomly sampling with replacement from D.

2. **Learning *SLR*s** — Given bootstrap D_b, a sparse representation l_b is learned using the *SLR* framework proposed in Chapter 3. Additionally, an associated classifier $C(l_b, D_b)$ is trained to make prediction according to Equations 3.2.

Accordingly, a classifier and the associated *SLR* representation is learned on each bootstrap, which are the basis for weighted model selection.

Weighted Model Aggregation

In this part, we provide the details of model aggregation using oob weighting, which is derived from the work by Rao and Tibshirani [79]. The steps are elaborated as follows:

1. **Prediction on oob samples** — Given the bootstraps $\{D_1, \cdots, D_b, \cdots, D_B\}$ and the training set D, we use K_i to denote the indices of the bootstrapped samples that do not contain patient i. Then, each classifier trained based on the bootstrapped samples in K_i makes a prediction on the label of observation i. The final prediction on the oob sample is aggregated with the following Equation 4.1:

$$\hat{y}_i(W) = \frac{1}{|K_i|} \sum_{b \in K_i} \omega_b C(l_b, D_b) \tag{4.1}$$

   where $W = \{\omega_1, \omega_2, \cdots, \omega_b, \cdots, \omega_B\}$.

2. **Learning of oob weighting** — Given $(\hat{y}_i(W), y_i)$ for observation i in the training set, we optimize the weights W of all classifiers in the ensemble. Considering that this is a classification problem, we employ the negative log-likelihood as the objective function with the constraints that $\omega_b \geq 0$ for all $b \in \{1, 2, \cdots, B\}$, as presented in Equation 4.2.

$$\begin{aligned} \underset{W}{\text{minimize}} \quad & -\sum_{i=1}^{N} y_i log \hat{y}_i(W) + (1 - y_i) log(1 - \hat{y}_i(W)) \\ \text{subject to} \quad & W \geq \mathbf{0}. \end{aligned} \tag{4.2}$$

   Then, the truncated Newton method, Newton conjugate gradient algorithm [80], is utilized to solve the optimization problem in Equation 4.2, and the optimal weights are denoted as $\hat{W} = \{\hat{\omega}_1, \cdots, \hat{\omega}_b, \cdots, \hat{\omega}_B\}$.

New Patient Prediction

Given the learned representations and classifiers in the ensemble with associated weights, we make final prediction for a new patient i as presented in Equation 4.3.

$$\hat{y}_i(\hat{W}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\omega}_b C(l_b, D_b) \tag{4.3}$$

Thus, the *WB-SLR* framework introduces a weighted bagging of the *SLR*s to obtain the final output in order to achieve more stable and more accurate prediction.

4.3 Evaluation

4.3.1 Background

Chronic kidney disease (CKD) is a general term of heterogeneous disorders characterizing the gradual loss of renal function over time [81–83]. There are five stages of CKD: Stage 1 - kidney damage with normal kidney function; Stage 2 - kidney damage with mild loss of kidney function; Stage 3 - mild to severe loss of kidney function; Stage 4 - severe loss of kidney function; and Stage 5 - also known as end stage renal failure (ESRF), which indicates kidney failure requiring dialysis or transplant for survival [84]. CKD could eventually progress to kidney failure, which is fatal without dialysis or a kidney transplant [83]. It increases the risk of mortality, decreased quality of life, as well as serious complications, such as cardiovascular disease, anemia, mineral and bone disorders, fractures, and cognitive decline. [85]

The worldwide prevalence of CKD in the general population is 13.4%, and it imposes a huge economic burden globally [86]. According to CDC, 30 million Americans, i.e., 15% of adults population, are estimated to have CKD. The prevalence of CKD has been increasing since the last couple of decades, which is partly due to the increased prevalence of diabetes and hypertension [87]. Medicare spending exceeds

$50 billion in 2013 for CKD patients ages 65 or older, which is 20% of all medical spending for this age group [84].

Diabetes and high blood pressure are the main risk factors of CKD and close to half of the CKD patients also have diabetes and/or self-reported cardiovascular disease [84]. This study by Bailey et al. confirms the high prevalence of CKD, 43.5%, in patients with type 2 diabetes [88]. However, CKD starts with impaired renal function and is usually asymptomatic until the later stages [83, 86]. According to CDC, almost half of the patients with severely reduced kidney function but not on dialysis are not aware of having CKD, and it is also unaware to approximately 96% of people with kidney damage or mildly reduced renal function [89].

Early detection and treatment of CKD can slow the progression of kidney damage by controlling the underlying cause [83, 85]. Early medical interventions can also prevent the risk of complications, especially cardiovascular disease, which is the leading cause of morbidity and mortality in dialysis patients. It has been widely demonstrated that interventions in the conservative phases of CKD is more effective and should be performed as early as possible [90]. Considering the prevalence of CKD in diabetic patients, the difficulty in recognizing it in the early stage, and its extreme importance of early interventions, we apply the proposed representation learning framework, *WB-SLR*, to this clinically meaningful problem, i.e., recognizing misdiagnosed and undiagnosed CKD in diabetic patients. In the rest of this chapter, we use the the terminology "prediction" to indicate the early recognition of undiagnosed or misdiagnosed CKD in diabetic patients.

4.3.2 Experimental Design

In this research, we use the de-identified EHR data of 75 months beginning in September 2010 from the University of Virginia Health System. This dataset contains $2,343,651$ inpatient and outpatient visits of $473,915$ distinct patients. In this experiment, we focus on the diagnosis codes in the EHR data. Here, the diagnoses are coded primarily in ICD-9 and a small portion are ICD-10 codes. The raw diagnoses codes are further clustered into 283 groups according to the Clinical Classification Software from AHRQ [64].

In the experiment, we define the observation window and prediction period to validate the proposed method. We first extract all patients with diabetes diagnoses in their medical record. For diabetic patients with CKD diagnoses to be included in the cohort, their diabetes are diagnosed at least 1.5 years prior to the first observed diagnoses of CKD. Thus, it allows adequate time to observe the health conditions to construct features and allows the hold-off window to be used for the purpose of early detection.

For positive patients, their visits during 1.5 to 0.5 year prior to the first observed CKD are used to construct features for prediction. The visits within 6 months of the first observed CKD diagnosis are excluded. For diabetic patients without observed CKD diagnoses, we exclude the individuals when their visit histories are shorter than 2.5 years after the first observed diagnosis of diabetes. Visits occurred in the first 12 months are used as observations for modeling, and the next 0.5 year is the hold-off window same as that for the positive patients. The following 1 year is monitored to ensure the patients are not diagnosed with CKD in the near future. Thus, a 12-month observation window is constructed for patients in each group, and a prediction window is defined to allow observation of the outcome. To better illustrate the experimental setting, we present the observation window, hold-off and onset of outcome event in Figure 4.2.

Diagnoses appear in less than 5% of the patients' records in both populations are excluded as rare events. Table 4.1 presents the preliminarily selected diagnoses, which are used as the basis of the *WB-SLR* framework. Overall, there are 395 and $6,259$ patients in the target and control group, respectively. The target group is randomly

Table 4.1: The preliminarily selected diagnoses for the early detection of chronic kidney disease (CKD) in diabetic patients

Index	Diagnoses
1	Coronary atherosclerosis and other heart disease
2	Nonspecific chest pain
3	Other and ill-defined heart disease
4	Cardiac dysrhythmias
5	Congestive heart failure; nonhypertensive
6	Other circulatory disease
7	Pleurisy; pneumothorax; pulmonary collapse
8	Other lower respiratory disease
9	Esophageal disorders
10	Other liver diseases
11	Other gastrointestinal disorders
12	Other diseases of kidney and ureters
13	Genitourinary symptoms and ill-defined conditions
14	Osteoarthritis
15	Other non-traumatic joint disorders
16	Spondylosis; intervertebral disc disorders; other back problems
17	Other connective tissue disease
18	Other bone disease and musculoskeletal deformities
19	Other injuries and conditions due to external causes
20	Abdominal pain
21	Malaise and fatigue
22	Medical examination/evaluation
23	Other aftercare
24	Other screening for suspected conditions (not mental disorders or infectious disease)
25	Other and unspecified benign neoplasm
26	Thyroid disorders
27	Diabetes mellitus without complication
28	Diabetes mellitus with complications
29	Nutritional deficiencies
30	Disorders of lipid metabolism
31	Fluid and electrolyte disorders
32	Other nutritional; endocrine; and metabolic disorders
33	Deficiency and other anemia
34	Mood disorders
35	Screening and history of mental health and substance abuse codes
36	Other nervous system disorders
37	Essential hypertension

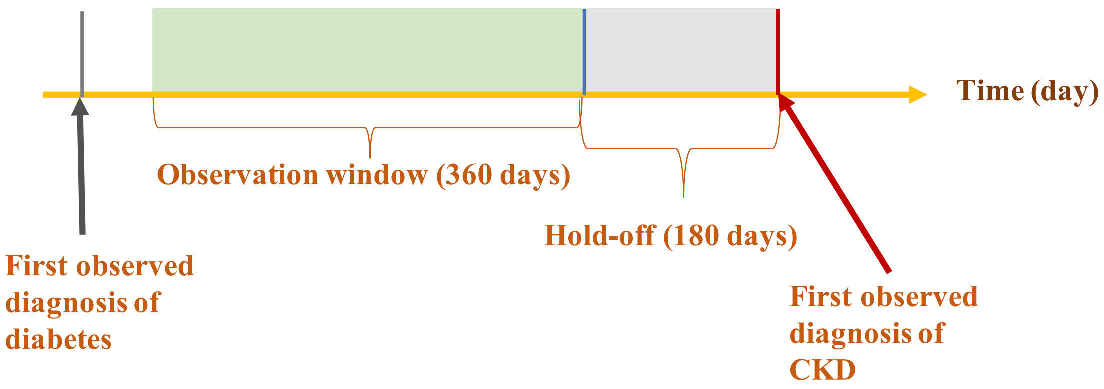

Time (day)
Observation window (360 days)
Hold-off (180 days)
First observed diagnosis of diabetes
First observed diagnosis of CKD

Figure 4.2: A graphical illustration of the experimental setting for the comorbid risk prediction of CKD in diabetic patients

split into training, validation and testing sets with a 2:1:2 ratio. In other words, 40% are used for training, another 40% are used for testing, and the rest 20% are used for parameter tuning. As to the negative group, we randomly extract 593 patients since the prevalence of CKD in diabetic patients is approximately 40%. Then, this negative group is randomly split into training, validation and testing with the same ratio as used for the positive class.

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

- Aggregated Frequency Vector (AFV): A patient's record is represented as a count vector of medical events in the observation window. Each dimension is associated with a distinct medical event in patients' EHR data.

- Bag-of-pattern in Sequences (BPS): This method runs a widely used sequence pattern mining algorithm, CM-SPADE [34], to discover all frequent patterns in patients' EHR data. Thus, a patient's record is represented with a frequency vector of the discovered frequent patterns.

- Aggregated Transition Vector (ATV): The pairwise transitions between medical event pairs are counted and a patient's record is represented with a frequency vector of pairwise transitions.

- *SLR*: A sparse longitudinal representation proposed in Chapter 3.
- Bagged *SLR*: It first learns a *SLR* representation and an associated classifier on each bootstrap sample. Then, the final prediction is obtained by majority voting over all the classifiers.

The time window used to group clinical events is 120 days such that 3 subsets of features are constructed for the 1-year observation window. According to the U.S. Census Bureau, adults have an average of 3.9 visits to doctors annually [65]. Here, we use a slightly larger time window in this experiment considering that the dataset is small. However, this time window can be further tuned on specific populations for an optimal output. With *WB-SLR* and baseline representations of patients, we model the comorbid risk of CKD using three classifiers, logistic regression with $l1$ penalty (LR), random forest (RF), and gradient boosting trees (GBT) with 20 trees, respectively. The hyperparameters of classifiers are tuned on the validation set. The performance of baselines and the proposed framework on the comorbid risk prediction task is evaluated with AUC (area under curve), sensitivity, specificity, and F2 score. Each experiment is repeated 50 times and we calculate the averages and standard deviations of the above metrics, respectively.

4.3.3 Experimental Results

The predictive performance of the classifiers based on the baseline representations and proposed *WB-SLR* framework are presented in Table 4.2. Here, the results shown are based on the *WB-SLR* learned with $\alpha = 0.7$ and $\lambda = 0.0005$.

According to the results table, the predictive performance of classifiers based on *WB-SLR* outperform the baselines in terms of AUC. The classifier based on *WB-SLR* achieves a relatively higher AUC by approximately 5% than the bagged *SLR*. The *WB-SLR* achieves a highly balanced prediction result on the target and control groups by comparing the sensitivity and specificity. With the baseline representations, the

Table 4.2: The predictive performance of baselines and the proposed *WB-SLR* framework

		Sensitivity	Specificity	AUC	F2 score
AFV	LR	0.649 ± 0.030	0.734 ± 0.025	0.712 ± 0.025	0.660 ± 0.037
	RF	0.557 ± 0.033	0.890 ± 0.017	0.774 ± 0.020	0.671 ± 0.026
	GBT	0.623 ± 0.029	0.891 ± 0.028	0.796 ± 0.021	0.667 ± 0.023
BPS	LR	0.592 ± 0.042	0.728 ± 0.026	0.757 ± 0.022	0.592 ± 0.031
	RF	$\mathbf{0.975 \pm 0.011}$	0.287 ± 0.026	0.732 ± 0.027	0.803 ± 0.020
	GBT	0.831 ± 0.026	0.684 ± 0.026	0.807 ± 0.018	0.749 ± 0.024
ATV	LR	0.500 ± 0.033	0.888 ± 0.027	0.697 ± 0.026	0.546 ± 0.035
	RF	0.598 ± 0.032	0.842 ± 0.022	0.791 ± 0.021	0.712 ± 0.028
	GBT	0.558 ± 0.031	0.885 ± 0.017	0.796 ± 0.025	0.722 ± 0.024
SLR	LR	0.747 ± 0.035	0.899 ± 0.018	0.835 ± 0.015	0.772 ± 0.031
	RF	0.753 ± 0.033	$\mathbf{0.903 \pm 0.019}$	0.847 ± 0.026	0.799 ± 0.030
	GBT	0.857 ± 0.024	0.775 ± 0.025	0.877 ± 0.017	0.810 ± 0.016
Bagged *SLR*		0.829 ± 0.023	0.865 ± 0.020	0.842 ± 0.020	0.818 ± 0.028
WB-SLR		0.835 ± 0.025	0.852 ± 0.012	$\mathbf{0.891 \pm 0.018}$	$\mathbf{0.820 \pm 0.027}$

GBT and RF classifiers generally outperform LR classifiers in terms of AUC and F2 score. In general, the GBT models demonstrates a slightly better performance than RFs. The performance between the first four baselines are not significantly different, while the *SLR* based representations are capable of improving the prediction performance significantly. In terms of AUC, the bagged *SLR* achieves a more accurate prediction than the LR classifier based on *SLR*, while the RF and GBT models based on *SLR* outperform the bagged *SLR*. This finding is consistent with the properties of bagging and random forest described in Chapter 2. In general, the proposed *WB-SLR* is able to achieve a more accurate prediction compared to baselines and the improvement could be more significant if there are more observations to allow better weight learning for model aggregation.

In addition to the predictive modeling on CKD in diabetic patients, we apply *WB-SLR* to the risk prediction of diabetes, in which the details of this task are elaborated in Chapter 3. The predictive performance are presented in Table 4.3.

Comparing the predictive performance in Tables 4.3 and 3.3, we observe that

Table 4.3: The predictive performance of *WB-SLR* framework on the early detection of diabetes

	Sensitivity	Specificity	AUC	F2 score
WB-SLR	0.885 ± 0.009	0.610 ± 0.006	0.846 ± 0.006	0.623 ± 0.005

WB-SLR achieves a more accurate prediction, especially in terms of AUC and F2 score. Thus, it indicates that the weighted bagging of SLRs are capable of improving the predictive performance.

4.3.4 Visualization & Interpretation

In addition to predictive performance, we elaborate the selected features by the *SLR* with highest weight in *WB-SLR* in the early detection of CKD among diabetic patients. We illustrate the correlations between the selected features in the observation window and the outcome in Figure 4.3. In general, we observe that the diagnoses temporally closer to the observation of outcome are more positively correlated with it while the selected features that are negatively correlated are mostly in the first time window, i.e., occurred much earlier than the outcome onset. Negatively correlated events in the very early time window include *non-specific chest pain*, *osteoarthritis*, *thyroid disorders*, *other connective tissue disease*, *essential hypertension*, *diabetes with/without complications*, and so forth.

However, the diabetic diagnoses become a very important positive risk factor when it is later in the time window, especially the *diabetes mellitus with complications.* Other positively correlated risk factors throughout the observation window are *deficiency and other anemia*, *other aftercare*, *other injuries and conditions due to external causes*, and heart and cardiovascular diseases, including *congestive heart failure*, *cardiac dysrhythmias*, *other and ill-defined heart disease*, etc. Here, *deficiency and other anemia* refers to iron deficiency anemias, other deficiency anemias, and/or hereditary anemias according to [64], which are found to be correlated to CKD ac-

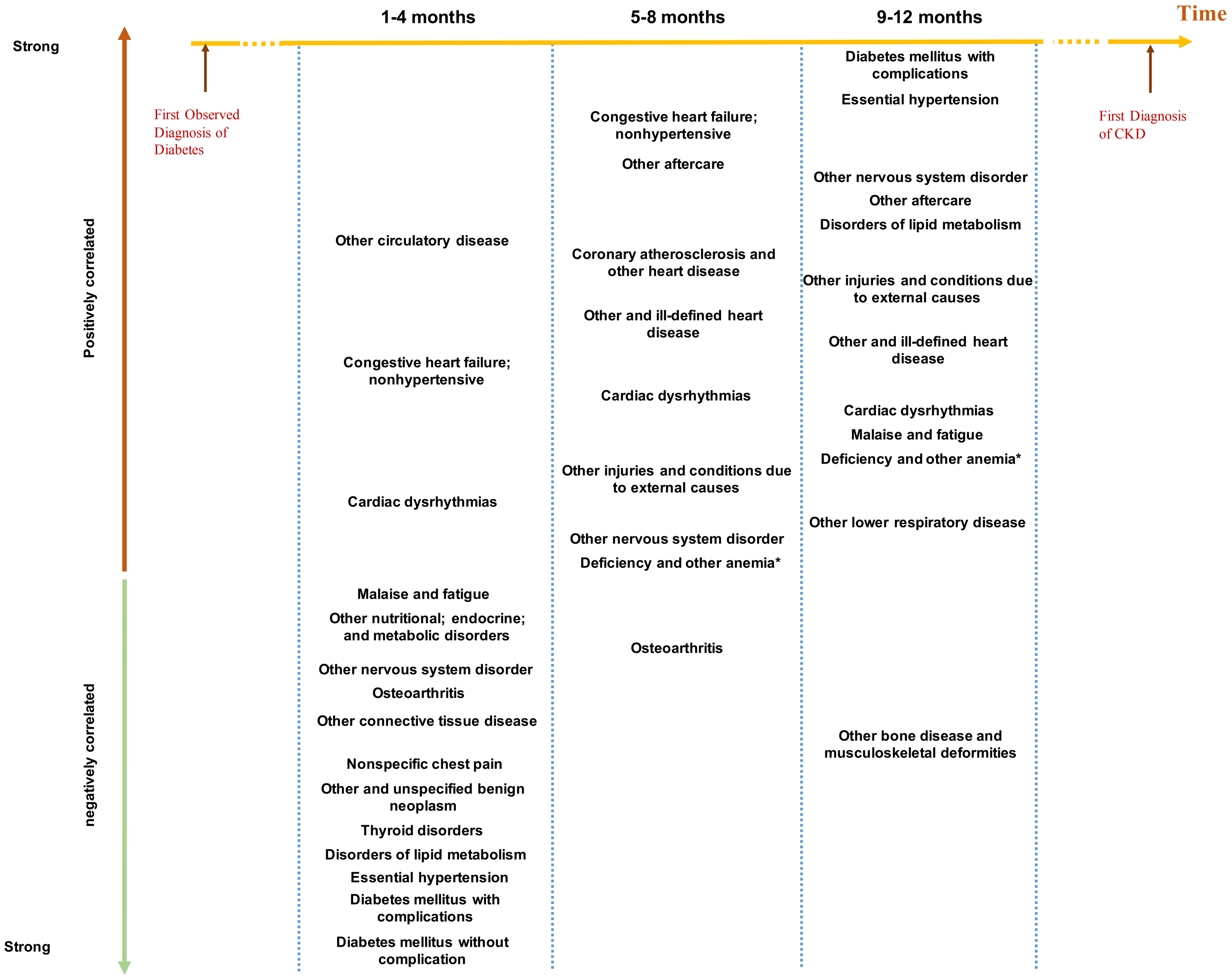

1-4 months
5-8 months
9-12 months
Time
Strong
Positively correlated
negatively correlated
Strong
First Observed Diagnosis of Diabetes
First Diagnosis of CKD
Other circulatory disease
Congestive heart failure; nonhypertensive
Cardiac dysrhythmias
Malaise and fatigue
Other nutritional; endocrine; and metabolic disorders
Other nervous system disorder
Osteoarthritis
Other connective tissue disease
Nonspecific chest pain
Other and unspecified benign neoplasm
Thyroid disorders
Disorders of lipid metabolism
Essential hypertension
Diabetes mellitus with complications
Diabetes mellitus without complication
Congestive heart failure; nonhypertensive
Other aftercare
Coronary atherosclerosis and other heart disease
Other and ill-defined heart disease
Cardiac dysrhythmias
Other injuries and conditions due to external causes
Other nervous system disorder
Deficiency and other anemia*
Osteoarthritis
Diabetes mellitus with complications
Essential hypertension
Other nervous system disorder
Other aftercare
Disorders of lipid metabolism
Other injuries and conditions due to external causes
Other and ill-defined heart disease
Cardiac dysrhythmias
Malaise and fatigue
Deficiency and other anemia*
Other lower respiratory disease
Other bone disease and musculoskeletal deformities

Figure 4.3: The set of selected features with highest weight in the *WB-SLR* framework and their correlations to CKD in diabetic patients

cording to [91–93]. In fact, previous literature have shown that the prevalence of anemia increases as kidney function decreases [93–95].

Additionally, we observe that *essential hypertension* and a group of cardiovascular and heart diseases are strongly correlated with the onset of CKD. Previous research have demonstrated that *essential hypertension* is one of the leading causes of CKD together with diabetes [96]. According to [97], there is a close interrelation between cardiovascular disease and kidney disease, and the disease of one organ may cause dysfunction of the other, which could ultimately lead to the failure of both organs [97]. Additionally, both cardiovascular and heart diseases and kidney disease are complications of diabetes. It is also likely that the diabetic patients developed cardiovascular and heart problems because of uncontrolled diabetes such that we observe the diagnoses of CKD later as another complication of diabetes, apart from the damages to renal function by hypertension and cardiovascular diseases.

In Figure 4.3, *other nervous system disorder*, *other aftercare* and *other injuries and conditions due to external causes* are also shown to be positively correlated to CKD. According to [98,99], neurological complications are prevalent in CKD patients and occur in almost all patients with severe CKD, which might affect both the central and peripheral nervous systems. Here, the *other aftercare* diagnosis group includes aftercare following surgeries and long-term (current) use of drugs, such as insulin. In the diagnosis group *other injuries and conditions due to external causes*, there are diagnoses of injuries due to accidents as well as other nonspecific abnormal toxicological findings which includes abnormal levels of drugs or heavy metals in blood, urine, or other tissue [64]. Medical research has shown that exposure to heavy metals and chronic use of drugs known to be potentially nephrotoxic can lead to CKD [100,101].

It is clinically counterfactal to observe *disorders of lipid metabolism*, *essential hypertension*, and *diabetes with/without complications* being negatively correlated factors earlier in the observation window, while they become strong positive risk factors

Table 4.4: Diagnoses positively correlated with prediction outcome in at least one third of *SLR* models

	Months in Observation Window		
	1-4	5-8	9-12
Coronary atherosclerosis and other heart disease		Y	Y
Nonspecific chest pain		Y	
Other and ill-defined heart disease		Y	Y
Cardiac dysrhythmias	Y	Y	Y
Congestive heart failure; nonhypertensive	Y	Y	Y
Other circulatory disease	Y	Y	Y
Pleurisy; pneumothorax; pulmonary collapse			Y
Other lower respiratory disease	Y	Y	Y
Other diseases of kidney and ureters			Y
Diabetes mellitus without complication		Y	Y
Diabetes mellitus with complications		Y	Y
Nutritional deficiencies			Y
Deficiency and other anemia		Y	Y
Disorders of lipid metabolism			Y
Other nervous system disorders			Y
Essential hypertension		Y	Y
Spondylosis; intervertebral disc disorders; other back problems	Y		
Other connective tissue disease			Y
Other injuries and conditions due to external causes			Y
Other and unspecified benign neoplasm			Y
Thyroid disorders		Y	Y
Malaise and fatigue		Y	Y
Other aftercare		Y	Y

later. The potential rationale is that these diagnoses are prevalent in the early observation window in both the negative and positive patient cohorts, however, the factors making real differences are the occurrences of them later in the observation period or constant occurrences of these diagnoses. Thus, the early observation of these factors are learned as being negatively correlated to the onset of CKD.

To get a more comprehensive understanding of the learned *WB-SLR* representation, we further interpret the learned *WB-SLR* representation in Table 4.4. It shows the diagnoses at the three distinct time windows during the observation period which are selected as being positively correlated to the outcome by at least one third of the *SLR*s in *WB-SLR*. Accordingly, the *diabetes mellitus with/without complications*, *other diseases of kidney and ureters*, *other circulatory disease*, *deficiency and other anemia*, as well as heart diseases, including *congestive heart failure*, *cardiac dysrhythmias*, etc., appear to be important factors. These findings are consistent with the medical literature.

Interestingly, we observe that some diagnoses are positively correlated throughout the entire observation window, such as *cardiac dysrhythmias* and *congestive heart failure*, while some are more correlated when occurred early in the observation window, such as *spondylosis, intervertebral disc disorders, other back problems.* According to the National Kidney Foundation, there are three early warnings signs of kidney disease, including lower back pain in which the pain from kidney might be recognized as lower back pain [102]. It is very straightforward to observe that the diagnoses of *other diseases of kidney and ureters* as a positive risk factor of CKD, since CKD starts with impaired renal function and may not be apparent until the kidney function is significantly impaired [83]. Moreover, some diagnoses occur later in the observation window, i.e., closer to the onset of CKD, are more positively correlated to it, including *diabetes with complications*, *essential hypertension*, *other and ill-defined heart disease*, *other nervous system disorders*, *deficiency and other anemia*, and *nutritional deficiencies.* According to medical research, deficiencies of nutritions including Vitamin D is also positively correlated to CKD [95, 103–105].

It is not surprising to observe that the diabetic diagnoses are found to have a positive correlation with CKD. Both *diabetes mellitus without complications* and *diabetes mellitus with complications* occur from the 5 to 8 months window to the end of the observation period. It implicitly mimics the progression of CKD due to diabetes. We also observe that *lower respiratory disease* and *pleurisy, pneumothorax, pulmonary collapse* are positive factors, especially the former one being positively correlated

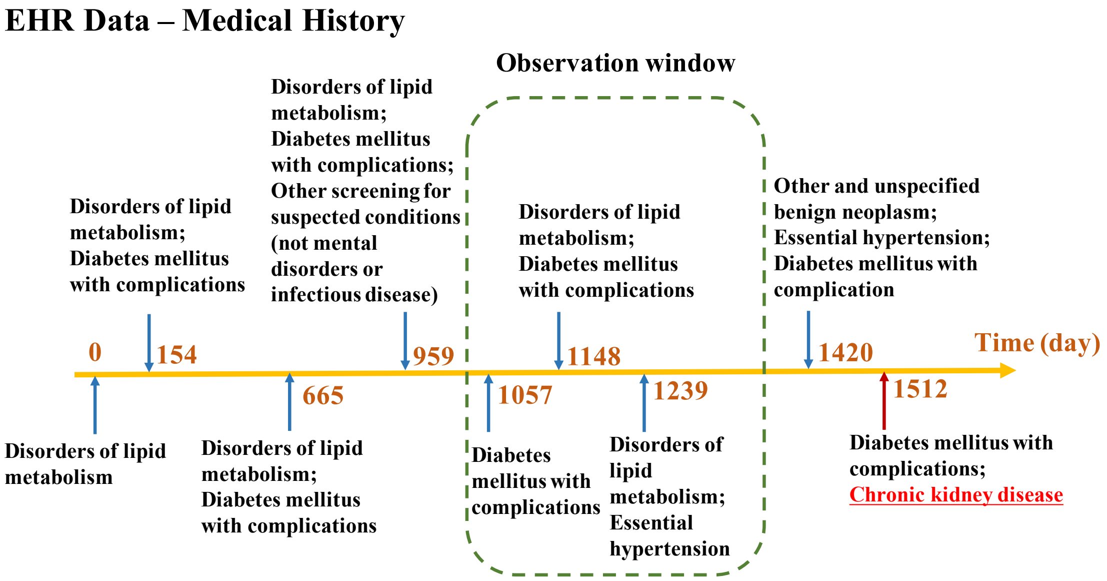

EHR Data – Medical History
Observation window
Disorders of lipid metabolism; Diabetes mellitus with complications; Other screening for suspected conditions (not mental disorders or infectious disease)
Disorders of lipid metabolism; Diabetes mellitus with complications
Disorders of lipid metabolism; Diabetes mellitus with complications
Other and unspecified benign neoplasm; Essential hypertension; Diabetes mellitus with complication
0
154
959
1148
1420
Time (day)
665
1057
1239
1512
Disorders of lipid metabolism
Disorders of lipid metabolism; Diabetes mellitus with complications
Diabetes mellitus with complications
Disorders of lipid metabolism; Essential hypertension
Diabetes mellitus with complications; Chronic kidney disease

Figure 4.4: The medical history of an example diabetic patient who developed CKD later

throughout the entire observation window. In fact, lung and kidney function are intimately related and the respiratory complications of CKD include pulmonary edema, fibrinous pleuritis, pulmonary calcification, and so forth [106]. Thus, the two groups of diagnoses related to lung function are recognized as positively correlated factors to CKD.

To provide a more straightforward understanding of the learned representation and its benefit to the comorbid risk prediction of CKD in diabetic patients, we present the medical histories of two example diabetic patients in Figures 4.4 and 4.5. Figure 4.4 shows the EHR data of a diabetic patient who developed CKD later, while Figure 4.5 illustrates the medical history of a diabetic patient with no observed CKD in the record. We observe that *disorders of lipid metabolism* and *diabetes mellitus with complications* occur repeatedly in the first patient's medical record, and those are found to be positively correlated with comorbid CKD. However, the diagnoses in the second patient's EHR data are less correlated to comorbid CKD according to the learned representation by *WB-SLR*. Both example patients are predicted correctly

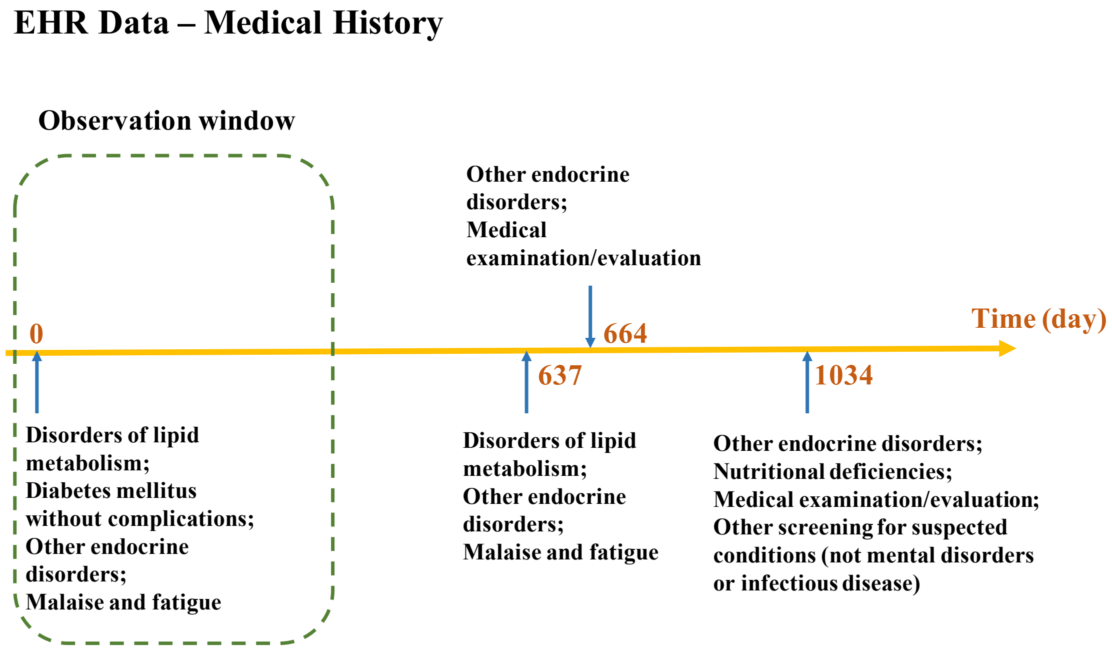

EHR Data – Medical History
Observation window
Other endocrine disorders; Medical examination/evaluation
664
Time (day)
0
637
1034
Disorders of lipid metabolism; Diabetes mellitus without complications; Other endocrine disorders; Malaise and fatigue
Disorders of lipid metabolism; Other endocrine disorders; Malaise and fatigue
Other endocrine disorders; Nutritional deficiencies; Medical examination/evaluation; Other screening for suspected conditions (not mental disorders or infectious disease)

Figure 4.5: The medical history of an example diabetic patient without CKD in the record

using the proposed *WB-SLR* representation learning framework.

In general, the learned *WB-SLR* representation is consistent with the previous medical research. In addition to improving the accuracy of the early detection of CKD in diabetic patients, this research could also be used to deepen clinical understanding of disease correlations.

4.4 Summary

In this chapter, we propose a novel representation learning framework, *WB-SLR*, to learn a comprehensive and stable representation of patients' EHR data. This method utilizes the *SLR* framework, the bagging strategy, and model aggregation based on oob weighting. We apply this framework to the early detection of CKD in diabetic patients using longitudinal EHR data. The experimental results demonstrate that the proposed *WB-SLR* representation framework is capable of achieving a more accurate prediction and it also uncovers the correlations between diseases which are found to be consistent with previous medical research.

Alternative model selection and combination methods will be explored to aggregate the ensemble of classifiers and representations learned from bootstrapped samples. Moreover, future improvement on the proposed model could employ the strategy of random forest which samples a subset of features when growing a tree. This could potentially reduce the correlation between the single models in the ensemble and further improve the prediction accuracy. Again, future work will consider more specific diagnosis codes and other grouping strategies to reduce the information loss introduced by using the current AHRQ clinical classification scheme. Similarly, other types of clinical events, such as procedures and medications, could be added to learn a more comprehensive representation of patients' medical histories. Regularization is introduced to reduce multicollinearity and other potential approaches to address this issue will be explored as future directions.

In the experiment, we model the comorbid risk of CKD in diabetic patients with a small population under the current experimental setting, i.e., 395 individuals are identified under the setting of having their first CKD diagnoses at least 18 months after the first observed diabetes diagnoses. To address this sample size issue, we could potentially employ the medical records of CKD patients without diabetes or with concurrent diabetes. This is because that CKD in patients with or without diabetes are similar and most likely share some common risk factors or symptoms. Hence, there is great potential to improve the prediction performance by including those patients into the originally identified positive patients. We will also explore transfer learning approaches to transfer the knowledge learned from a similar population to address the target problem.

Chapter 5

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

In this chapter, we propose a computational framework to learn an interpretable deep representation of the longitudinal EHR data which is personalized for each patient. We first introduce the background and related works of this research. Then, we elaborate the framework and the representation learning algorithm in detail. Next, we apply this proposed framework to the risk prediction of hospitalization and discuss the predictive performance of *Patient2Vec* compared to baselines. In the end, the learned feature importance are visualized and interpreted at both the individual and population levels.

5.1 Introduction

Longitudinal EHR data are resembling text documents in 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 is consisting of a sequence of visits, and there are a list of clinical events, including diagnoses, medications, and procedures, occurred during a visit. Unlike text documents, there is usually no ordering between the events in a visit. Considering these similarities, the representation learning methods for text document in NLP has great potential to be applied to longitudinal EHR data.

Deep neural networks have become very popular in the NLP field and they have been very successful in many applications, such as machine translation, question answering, text classification, document summarization, language modeling, etc. [5,107, 108] It is very beneficial to use deep neural networks for such tasks, since these networks are capable of identifying high-order relationships. Additionally, the network structure can encode the language structures, and it allows 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, RNNs have shown their effectiveness in NLP tasks due to its capability of capturing sequential information which is natural in human language. Traditional neural networks assume that inputs are independent of each other, while a RNN computes the output based on the current input as well as the "memory" from the previous computation. Although the vanilla RNNs are not good at capturing long-term dependencies, many variants have been proposed and validated to be effective in addressing this issue.

In medical fields, it is critical that the analytical results are interpretable, such that it can be understood and validated by human with precise knowledge. However, a salient disadvantage of deep neural networks is the lack of interpretability. In order to make sense of the “black box”, many attempts have been made and the attention mechanism is one of the effective methods to make the results more interpretable.

Health care has experienced unprecedented changes over the past century, and there is a great potential and demand in personalized health care. Personalized medicine, also called precision medicine, is not a new term, yet most research are in the genetics field. However, the availability of EHR data and advances in machine learning have great potential to help make personalized health care accessible to patients. In fact, personalization has been ubiquitous in our daily life and we are experiencing it all the time. For example, there is personalized search on Google and personalized product recommendations on Amazon and Netflix. In addition to better customer experience, personalization might bring more benefits when applied in healthcare systems, such as better health outcomes and reduced psychological distress and costs. In representation learning, most methods capture the important features at population-level which might be distinctive between patients considering the heterogeneities in their medical histories and characteristics. Thus, it is important to learn a personalized representation of a patient’s medical history for personalized medicine and ultimately to achieve better healthcare outcomes.

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. The ultimate goal is to help achieve more accurate prediction, to bring clinical insights, and to facilitate the delivery of personalized medicine with such representations of EHR data. The rest of this chapter is organized as follows: Section 5.2 summarizes the variants of RNNs and the attention mechanism, as well as the applications of them on EHR data. Section 5.3 presents an overview of the proposed *Patient2Vec* representation learning framework, and Section 5.4 elaborates the details of the algorithms. In Section 5.6, the proposed framework is evaluated with 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 5.6 provides a summary of this work.

5.2 Related Work

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

Gated Recurrent Unit (GRU)

RNNs are expected to learn long-term dependencies by taking previous state and the new input in the computation at current time step t. However, the vanilla RNNs are incapable of capturing the dependencies when the sequence is very long due to the vanishing gradient problem [4]. Thus, 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. GRU is a simplified version of LSTM, and 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 x_t at time step t, the reset gate r_t is computed as presented in Equation 5.1.

$$r_t = \sigma(U_r x_t + W_r s_{t-1}) \tag{5.1}$$

where U_r and W_r are the weight matrices of reset gate, and 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 5.2.

$$z_t = \sigma(U_z x_t + W_z s_{t-1}) \tag{5.2}$$

where U_z and W_z are the weight matrices of update gate. The current hidden activation h_t is computed by

$$h_t = (1 - z_t)h_{t-1} + z_t \tilde{h}_t \tag{5.3}$$

where $\tilde{h}_t$ is the candidate activation at time step t and the computation of it is presented in Equation 5.4.

$$\tilde{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1})) \tag{5.4}$$

where U and W are weight matrices and $\odot$ represents element-wise multiplication. Figure 5.1 presents a graphical illustration of the GRU.

The GRU is capable of learning long-term dependencies 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 and maintained as it is, 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, it provides shortcuts such that the error is easily backpropagated without vanishing too

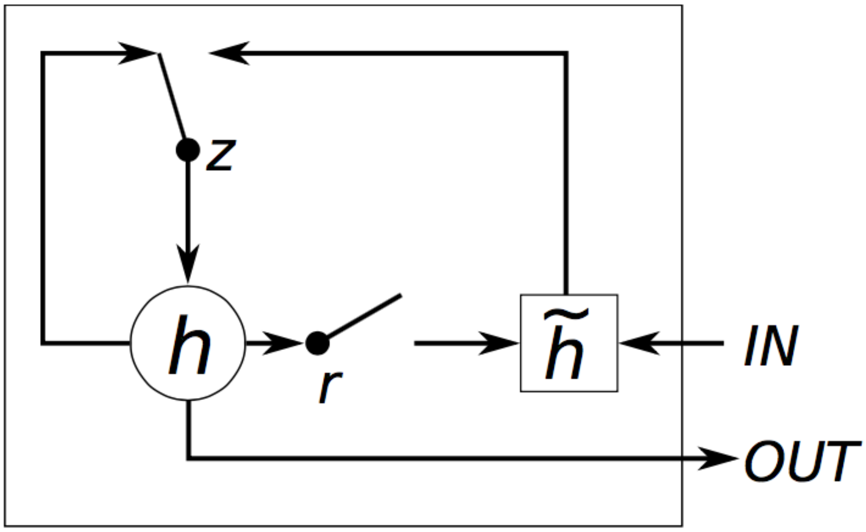


Figure 5.1: The GRU gating [5]

quickly [5, 109]. Hence, the GRU is well-suited to learn long-term dependencies in sequence data. The LSTM is similar to GRU, but with one more gate in a LSTM unit. 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 need fewer data to generalize [110].

Attention Mechanism

Attention mechanisms have become more and more popular in deep learning, which is inspired by the visual attention system found in human. Attention allows the network to focus on certain regions, while perceiving the other regions with "low resolution". In addition to higher accuracy, it also facilitates the interpretation of learned representations. We elaborate an attention mechanism on a RNN network, and Figure 5.2 presents a graphical illustration.

According to Figure 5.2, a variable-length weight vector α is learned based on the hidden states. Then a global context vector is computed based on weights α and all

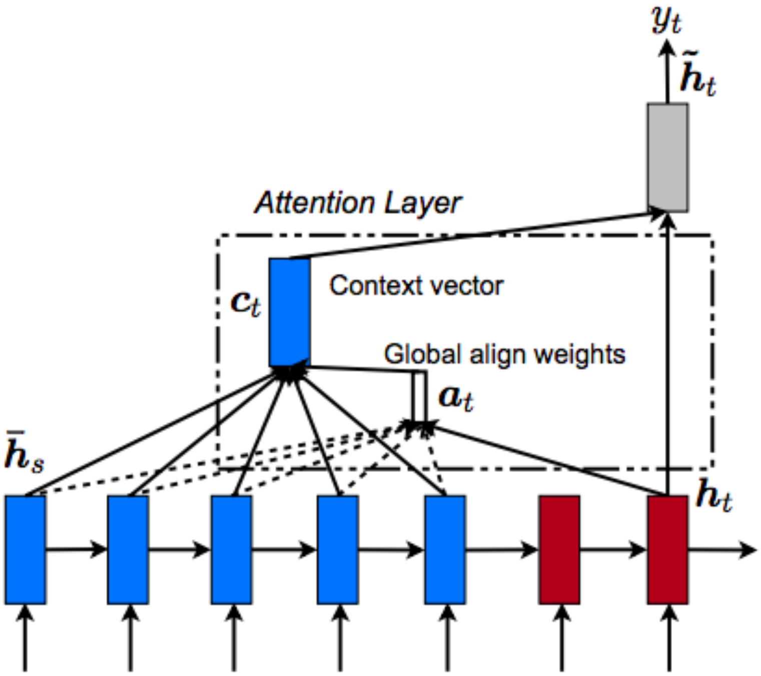


Figure 5.2: The global attention model [6]

the hidden states to make final output. Equation 5.5 presents the computation of weight vector $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_T\}$, where T is the length of the sequence.

$$\alpha_1, \alpha_2, \cdots, \alpha_T = f(W_\alpha h + b_\alpha) \tag{5.5}$$

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

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{5.6}$$

Thus, the network puts more attention to the important features in 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 CNNs and RNNs for various tasks and have achieved many successes in computer vision and NLP fields [6, 111, 112].

Deep Learning in EHR Data

Previous studies on health analytics of EHR data are mainly using statistical methods or traditional machine learning models. Nowadays, researchers have started adapting deep learning approaches to this area, including textual notes, temporal measurements of lab tests in the intensive care unit (ICU), and longitudinal EHRs in the health systems. Here, we summarize the deep learning research in mining EHR data and focus on the studies using RNN-based models.

Hospitalized patients, especially patients in ICUs are monitored on their clinical conditions, such that large amounts of clinical measurements are generated. These measurements are utilized by physicians to make diagnostic and treatment decisions. However, it is very challenging for traditional machine learning methods to mine this multivariate time series data considering the missing values, varying length, and irregular sampling. Lipton et al. train a LSTM with replicated target to learn from these sequence data and use this model to make predictions of diagnoses [113]. The data used in this research are the time series clinical measurements with continuous values and the LSTM models outperform logistic regression and MLP. Furthermore, Che et al. develop a GRU-based model to address the missing values in multivariate time series data, in which the missing patterns are incorporated for improved prediction performance [114]. This work has been applied to the MIMIC-III clinical database to demonstrate its effectiveness in mining time series of clinical measurements with missing values [115]. In addition to time series clinical measurements, longitudinal EHR data with clinical events, such as diagnoses, medications, and procedures is also a rich resource to explore. Choi et al. leverage this data with a GRU network to forecast future clinical events, and it achieves a better prediction performance than baselines such as logistic regression and MLP [116].

However, the difficulty in interpretation is one of the major drawbacks of using deep learning to mine EHR data. Some attempts have been made to address this issue. In [117], it proposes an interpretable mimic learning method which trains a mimic gradient boosting trees model to utilize the predicted labels or features learned by the deep learning models for final prediction [118]. Then, the feature importance learned by the tree-based models are used for knowledge discovery. Recently, attention mechanisms have been introduced to improve the interpretability of the prediction results of deep learning models in health analytics. In [119], it develops 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 improve the interpretability of RNN-based models with attention mechanisms for the representation learning of longitudinal EHR data.

5.3 *Patient2Vec* System Model

In this section, we provide an overview of the proposed hierarchical representation learning framework which utilizes deep recurrent neural networks to capture the complex relationships between clinical events in patient's EHR data. Additionally, it employs the attention mechanism to learn a personalized representation and to obtain the relative feature importance.

The proposed representation learning framework contains four steps and Figure 5.3 presents a graphical illustration of this framework.

1. **Learning vector representations of medical codes** — In the EHR data, there are often times multiple medical codes within a visit. Here, we treat the set of medical codes in a visit as a sentence consisting of words, except that

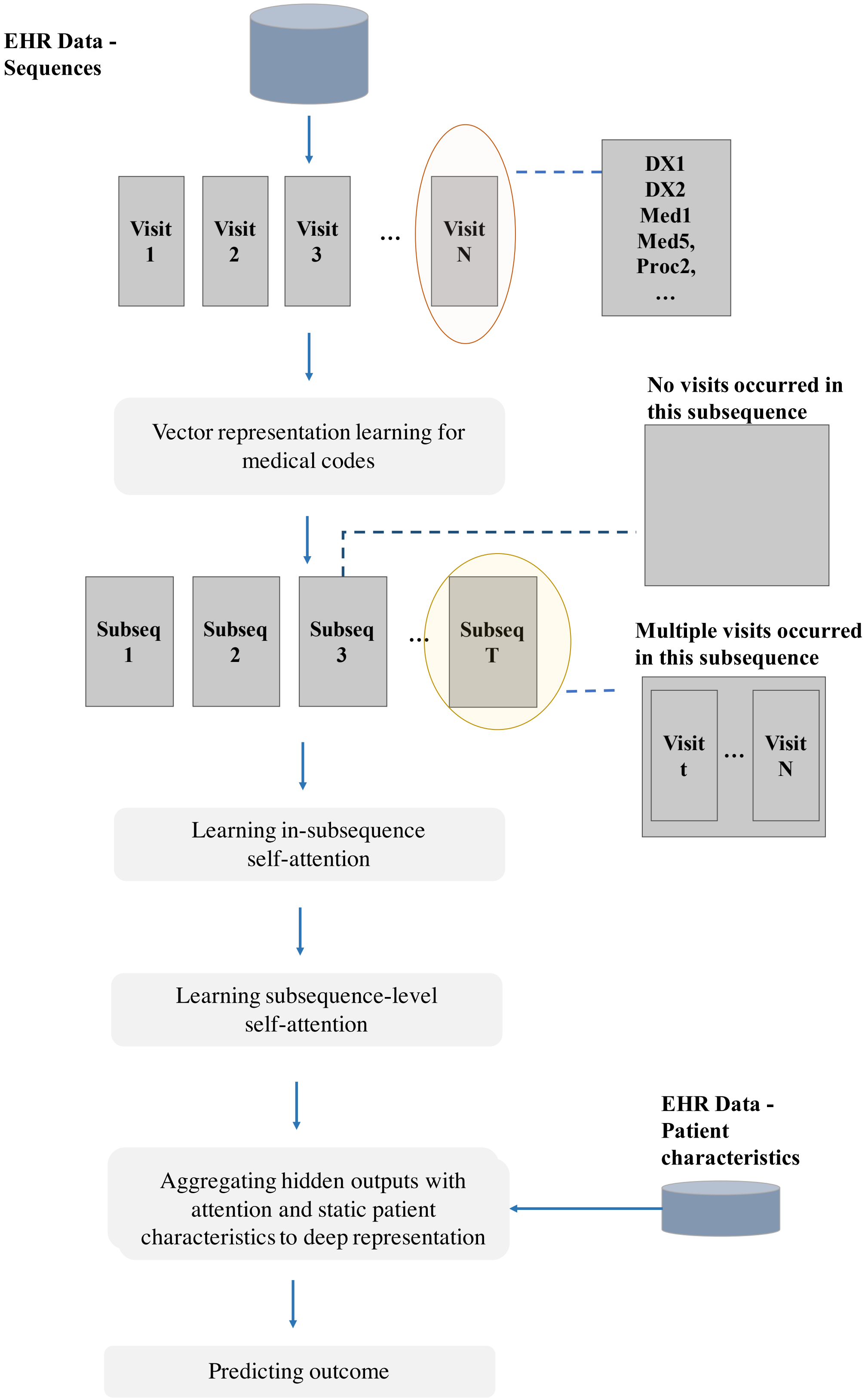

EHR Data -
Sequences
Visit 1
Visit 2
Visit 3
...
Visit N
DX1
DX2
Med1
Med5,
Proc2,
...
Vector representation learning for medical codes
No visits occurred in this subsequence
Subseq 1
Subseq 2
Subseq 3
...
Subseq T
Multiple visits occurred in this subsequence
Visit t
...
Visit N
Learning in-subsequence self-attention
Learning subsequence-level self-attention
EHR Data -
Patient characteristics
Aggregating hidden outputs with attention and static patient characteristics to deep representation
Predicting outcome

Figure 5.3: The *Patient2Vec* representation learning framework

there is no ordering in the words. Thus, we adopt the word2vec approach to learn a vector to represent each medical code.

2. **Learning within-subsequence self-attention** — Given the vector representation of medical codes, we are able to represent a visit with the codes occurred during this visit. However, some visits might be highly correlated with each other such that it is not clinically meaningful to separate them. Thus, we employ a time window to split the sequence of visits into multiple subsequences. Consequently, a subsequence might contain multiple original visits if they occurred within the same time window, or there might be no visits during a particular time window, which makes the subsequence empty. In this way, we are also able to transform this sequence into a sequence of subsequences with equal interval, which is preferable for recurrent neural networks.

   However, the medical events occurred within a subsequence are not contributing equally to the prediction of target outcome. Thus, we cannot aggregate them with equal weights, but instead we employ a self-attention mechanism which trains the network to learn the weights by itself.

3. **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.

4. **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.

5. **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.

5.4 *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 5.4 presents the structure of the proposed network with attention. The proposed framework consists of five parts presented in the following.

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, as described in Chapter 2.1 for medical code embedding. 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.

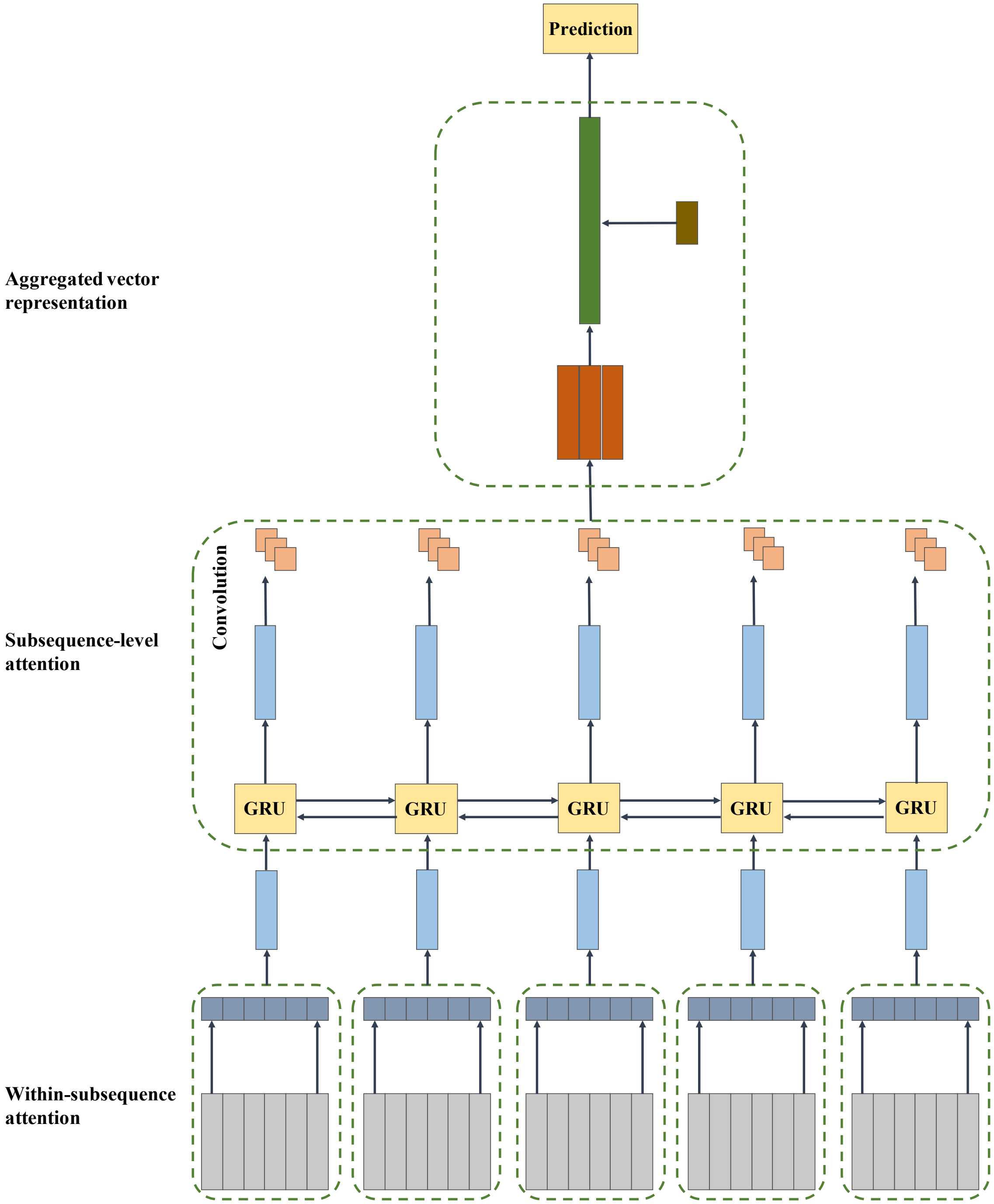

Prediction
Aggregated vector representation
Convolution
Subsequence-level attention
GRU
GRU
GRU
GRU
GRU
Within-subsequence attention

Figure 5.4: A graphical illustration of the network in the *Patient2Vec* representation learning framework

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 $s^{(i)}$, and $v_t^{(i)}$ denotes the tth subsequence in sequence $s^{(i)}$, where $t \in \{1, 2, \cdots, T\}$. Thus, $s^{(i)} = \{v_1^{(i)}, \cdots, v_t^{(i)}, \cdots, v_T^{(i)}\}$. To simplify the notation, we omit i in the following explanation. Subsequence $v_t \in \mathbb{R}^{n \times d}$ is a matrix of medical codes such that $v_t = \{v_{t_1}, v_{t_2}, \cdots, v_{t_j}, \cdots, v_{t_n}\}$, where $v_{t_j} \in \mathbb{R}^d$ is the vector representation of the jth medical code in the tth subsequence 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 v_t, presented in Equation 5.7.

$$\alpha_t = \tanh(Conv(\omega^\alpha, v_t)) \tag{5.7}$$

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 5.8.

$$\tilde{\alpha}_{t_j} = (\omega^\alpha)^\intercal v_{t_j} + b^\alpha \tag{5.8}$$

where b^α is a bias term. Then, given the original matrix v_t and the learned weights α_t, an aggregated vector $x_t \in \mathbb{R}^d$ is constructed to represent the tth subsequence, presented in 5.9.

$$x_t = \sum_{j=1}^{n} \alpha_{t_j} v_{t_j} \tag{5.9}$$

Given Equation 5.9, we obtain a sequence of vectors, $x = \{x_1, x_2, \cdots, x_t, \cdots, x_T\}$, to represent a patient's medical history.

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 5.10.

$$h_1, h_2, \cdots, h_t, \cdots, h_T = GRU(x_1, x_2, \cdots, x_t, \cdots, x_T) \tag{5.10}$$

where $h_t \in \mathbb{R}^k$ 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

$$\gamma_{mt} = (w_m^{\beta})^{\intercal} h_t + b^{\beta} \tag{5.11}$$

$$\beta_{m1}, \beta_{m2}, \cdots, \beta_{mt}, \cdots, \beta_{mT} = softmax(\gamma_{m1}, \gamma_{m2}, \cdots, \gamma_{mt}, \cdots, \gamma_{mT}) \tag{5.12}$$

where $w_m^{\beta} \in \mathbb{R}^k$. Thus, with the subsequence-level weights and hidden outputs, we construct a vector $c_m \in \mathbb{R}^k$ to represent a patient's medical visit history with one hop of subsequence weights, presented in the following Equation 5.13.

$$c_m = \sum_{t=1}^{T} \beta_{mt} h_t \tag{5.13}$$

Then, a context vector $c \in \mathbb{R}^{M \times k}$ is constructed by concatenating c_1, c_2, $\cdots$, c_M.

Constructing aggregated deep representation

Given the context vector c, this step integrate the patients characteristics $a \in \mathbb{R}^q$ 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, $c' \in \mathbb{R}^{M \times k + q}$, by adding a as additional dimensions to the context vector c.

Predicting outcome

Given the vector representation of the complete medical history and characteristics of patients, c', we add a linear and a softmax layer for the final outcome prediction, as presented in Equation 5.14.

$$\hat{y} = softmax(w^{c\intercal} c' + b^c) \tag{5.14}$$

To train the network, we use cross entropy as the loss function, presented in Equation 5.15.

$$L = -\frac{1}{N} \sum_{n=1}^{N} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) + \frac{1}{N} \sum_{n=1}^{N} ||\beta \beta^\intercal - \mathbf{I}||_F^2 \tag{5.15}$$

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 5.15 is to penalize redundancy if the attention mechanism provides similar subsequence weights for different hops of attention, which is derived from [120]. 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.

5.5 Evaluation

5.5.1 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, increase by 4.1% in 2014 [121]. More importantly, unplanned hospitalization can be distressing and can increase the risk of related adverse events, such as hospital-acquired infections and falls [122, 123]. In fact, hospitalization, one of the most expensive types of health care treatment, are sometimes avoidable. Approximately 40% hospitalizations in the United Kingdom are unplanned and are potentially avoidable [124]. Early interventions targeted to patients at-risk of hospitalization could help avoid unplanned admissions, reduce inpatient health care cost, reduce emergency department congestion, and so forth [125]. Thus, it is imperative to predict the risk of hospitalization such that interventions could be designed to prevent unnecessary admissions.

In this research, we apply our proposed representation learning framework on 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. [10, 126–128]. Here, we focus on the general population and the objective is to predict the risk of all-cause hospitalization using longitudinal EHR data.

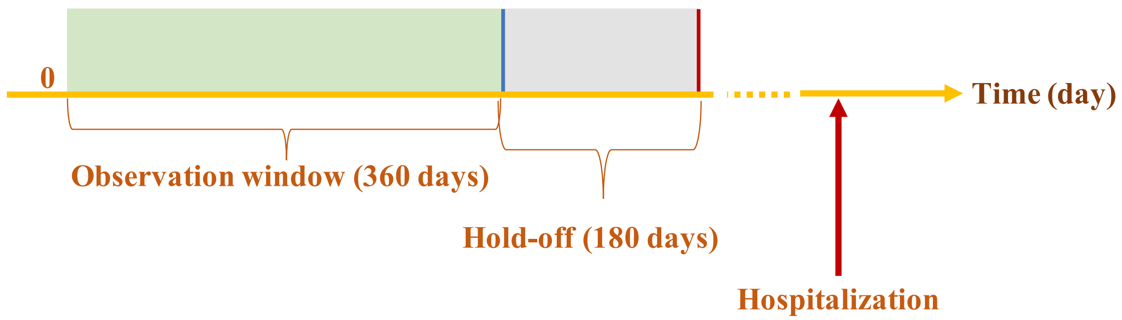


Figure 5.5: A graphical illustration of the experimental setting for the risk prediction of hospitalization

5.5.2 Experimental Design

In this research, we use the de-identified EHR data of 75 months beginning in September 2010 from the University of Virginia Health System. This dataset contains $2,343,651$ inpatient and outpatient visits of $473,915$ distinct patients. In this experiment, we focus on the diagnosis, medication, and procedure codes in the EHR data.

In the experiment, we define 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 is 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 5.5.

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 are ICD-10 codes, while procedures are mainly using CPT codes with a few ICD-9 procedure codes. The codes of medications are

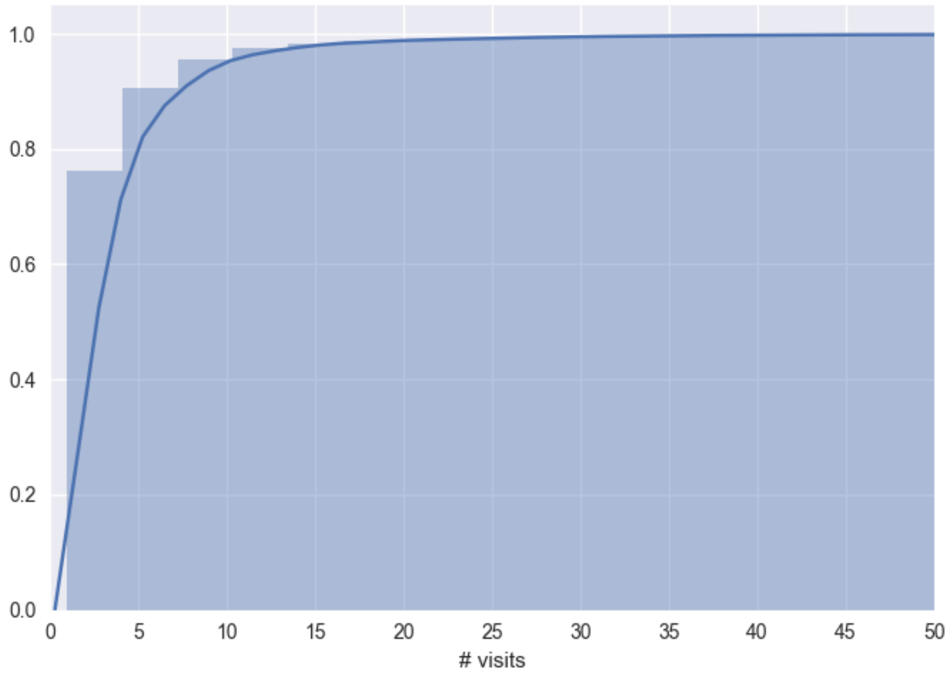


Figure 5.6: The cumulative histogram and density plot of patients' numbers of visits

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 are set to 100. The raw medical codes appear in less than 50 patients' medical records are considered as rare events such that these codes are excluded.

To construct the subsequences of medical codes, we use l days as the time window to split the sequence. Figure 5.6 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. In fact, there are only less than 25% of patients with more than 4 visits. Thus, we set l to 90 days in which a sequence is split into 4 subsequences.

Within each subsequence, the number of distinct medical codes are computed and patients with more medical codes in a subsequence than the 95% quantile are 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% are 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 5.15.

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

- **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 are huge, AHRQ clinical classifications of diagnoses and procedures are used to achieve a more general clustering of medical codes [64]. 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 previous hospitalization. Hence, the input is a 436-dimensional vector representing a patient's medical history and characteristics.

- **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.

- **Froward 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 with one hidden layer and the size of the hidden layer is 256.

- **Bidirectional RNN with medical group embedding (BiRNN-MGE)** — The inputs used for this baseline is the same as the one for the FRNN-MGE. The RNN used here is a bidirectional GRU with one hidden layer and the size of the hidden layer is 256.

- **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 with one hidden layer and the size of the hidden layer is 256.

- **Bidirectional RNN with medical vector embedding (BiRNN-MVE)** — The inputs used for this baseline is the same as the one for the FRNN-MVE. The RNN used here is a bidirectional GRU with one hidden layer and the size of the hidden layer is 256.

- **RETAIN** — This model uses reverse time attention mechanism on RNNs for an interpretable representation of patient's EHR data [119]. 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.

- ***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 are evaluated with AUC, sensitivity, specificity, and F2 score. 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.

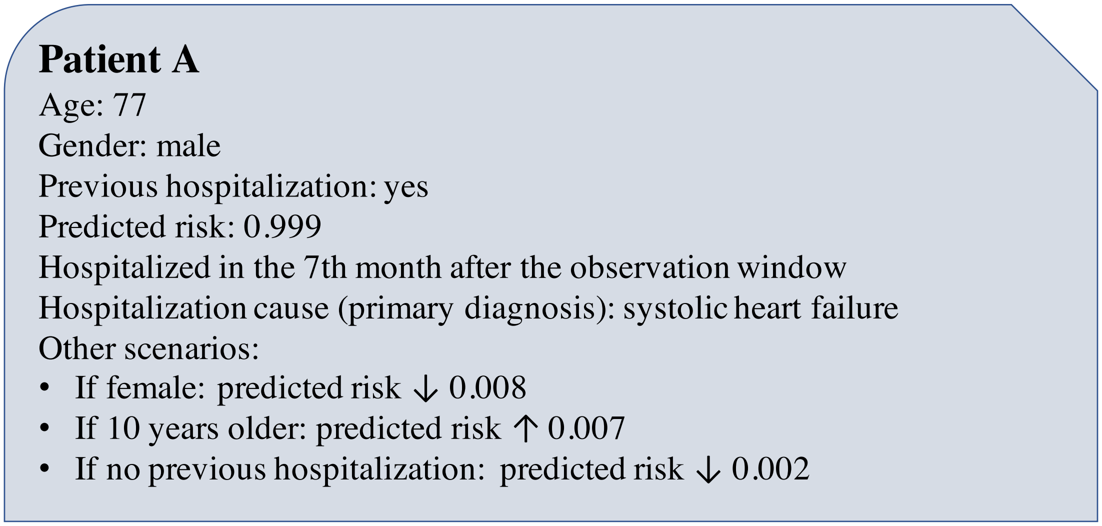
5.5.3 Experimental Results

The predictive performance of *Patient2Vec* and baselines are presented in Table 5.1. The results shown here for the RNN-based models are based on time interval $l = 90$ days to construct subsequences.

Table 5.1: The predictive performance of baselines and the proposed *Patient2Vec* framework

	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	$\mathbf{0.799 \pm 0.005}$	$\mathbf{0.492 \pm 0.007}$

According to Table 5.1, 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



Patient A

Age: 77
Gender: male
Previous hospitalization: yes
Predicted risk: 0.999
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

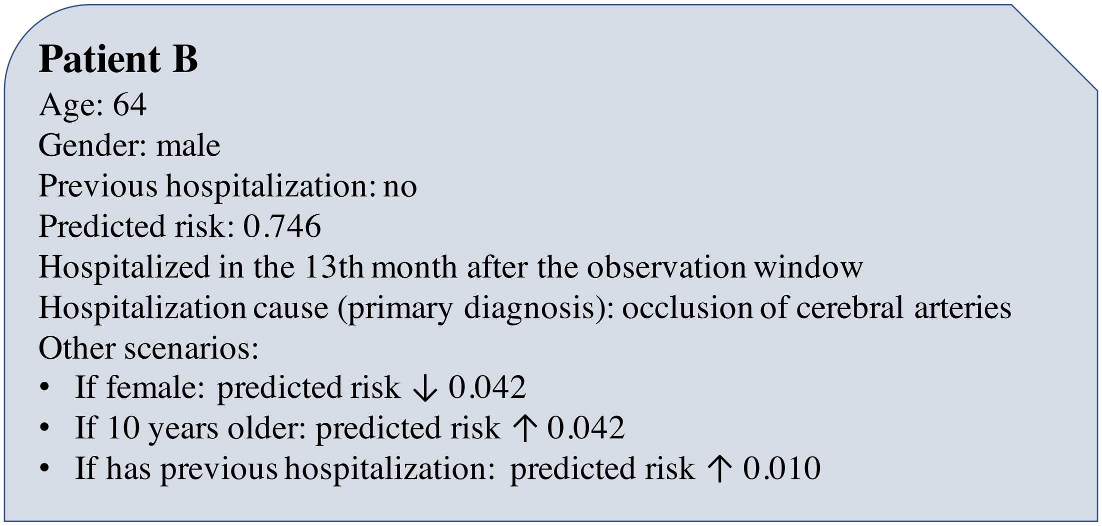
Figure 5.7: The profile of Patient A

F2 score.

5.5.4 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 examples in the following. Figures 5.7 and 5.8 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 5.7, Patient A is a male patient who has hospitalization history in the observation window and is admitted to the hospital later in 7 months for congestive heart failure. The predicted risk is 99.9%, 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. A heat map in Figure 5.9 shows the relative importance of the medical events in this patient's medical record at each time



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 5.8: The profile of Patient B

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 outcome. Accordingly, we observe that the last subsequence is the most important with respect to hospitalization risk, and the first two subsequences have relatively smaller weights, while the second last subsequence is the least important one.

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*, *antihypertensive*, *anticonvulsant*, *anticoagulant*, 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*

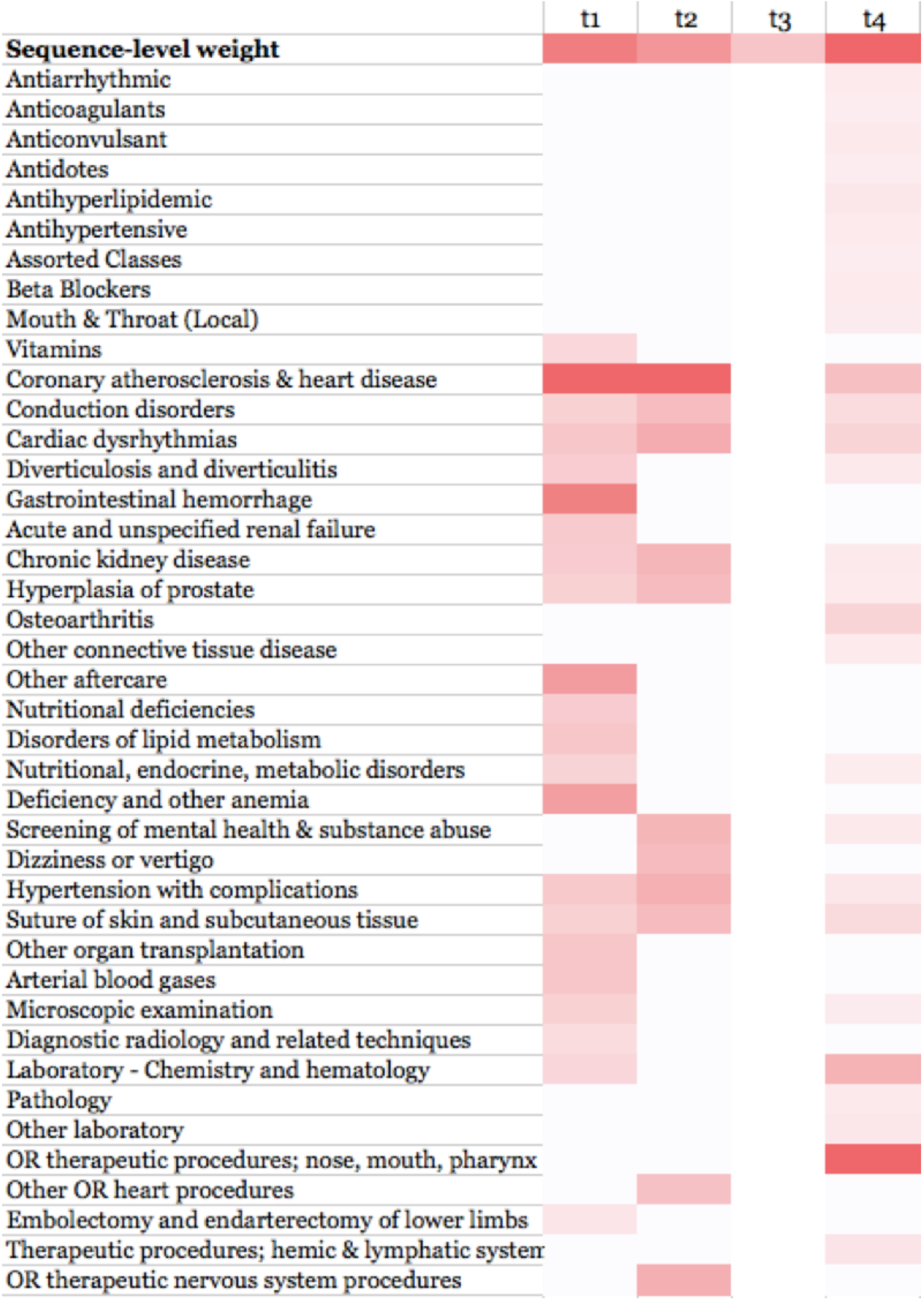


Figure 5.9: The heat map showing feature importance for Patient A

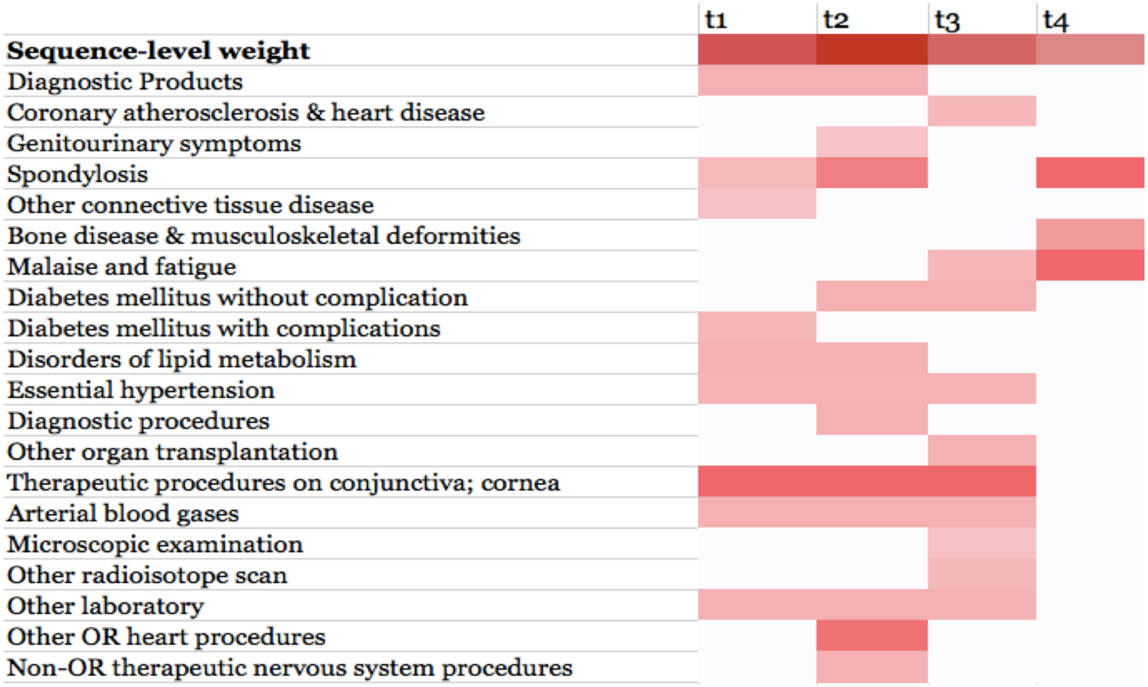


Figure 5.10: The heat map showing feature importance for Patient B

procedures, and *other OR therapeutic nervous system procedures*. Additionally, we observe that the kidney disease related diagnoses and procedures appear to be important features as well. 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 5.8 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, a heat map in Figure 5.10 presents the relative importance of medical groups in the subsequences, as well as the subsequence-level weights to the risk of hospitalization. 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 blood or urine testing, or used as radiopaque contrast agents for x-rays or CT scans for diagnostic purposes.

Additionally, we attempt to interpret the learned representation and feature importance at the population-level. In Table 5.2, we present the top 20 clinical groups with high weights among hospitalized patients in the test set.

According to Table 5.2, 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*

Table 5.2: 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

back problems. As to procedures, the most important ones 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 5.3 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 are consistent with medical literature.

5.6 Summary

In this chapter, we propose a representation learning framework, *Patient2Vec*, to learn a personalized interpretable deep representation of EHR data based on recurrent neural networks and attention mechanism. This work improves the performance

Table 5.3: 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

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 personalized medicine and to bring clinical insights.

In this experiment, the proposed *Patient2Vec* framework is evaluated with the risk prediction of all-cause hospitalization, and it can be applied to predict hospitalization in more specific populations or for certain causes.

Chapter 6

Conclusions & Future Directions

In this chapter, we summarize the proposed frameworks for representation learning of longitudinal EHR data, as well as the contributions. Then, we conclude this chapter with discussions on the future directions of this research.

6.1 Conclusions

First, this dissertation describes the wide implementation of EHR data which makes it a feasible resource to support clinical decision making. Then, we review the characteristics of longitudinal EHR data as well as the difficulties in utilizing it due to a variety of challenges including its complexity, heterogeneity, interpretation requirements and so forth. Hence, we introduce our research objective to address these issues by developing effective and efficient representation learning frameworks for EHR data and to facilitate future analytics. This dissertation also briefly describes the dataset used for this research, which is longitudinal EHR data from real clinical settings.

Secondly, we review the previous studies on popular and state-of-the-art representation learning methods in NLP, considering the longitudinal EHR data is alike text documents in many perspectives. Additionally, this dissertation summarizes the representation methods for EHR data, such as aggregated counts and other advanced approaches. This dissertation also presents an overview of a variety of classification models. Although these models are not directly learning representations, they are used for further predictive modeling based on the learned representations, such as for patient characterization and prediction of health outcomes.

Then, this dissertation proposes three representation learning frameworks, which are *SLR*, *WB-SLR*, and *Patient2Vec*.

- *SLR* focuses on learning a sparse longitudinal representation of patients' EHR data and it is evaluated with the early detection of diabetes based on patients' medical histories of diagnoses codes. The experimental results show an improved prediction performance compared with baseline representations. This dissertation further interprets the learned representation with an analysis on the selected features and the medical histories of example patients, and the findings are generally consistent with medical literature.

- *WB-SLR* employs the bagging approach to build an ensemble of classifiers and *SLR* representations on bootstrapped samples. Then, the learned models are combined with a weighting strategy, in which the weights are optimized to minimize the oob error. The final prediction is computed as the weighted output from the classifiers in the ensemble. This dissertation elaborates the evaluation of *WB-SLR* with the early detection of CKD using longitudinal EHR data of diagnoses codes, and the *WB-SLR* achieves an improvement on the prediction performance. Similarly, we visualize the learned representation of the *SLR* with the highest weight in the ensemble and the medical histories of example patients. To provide a more comprehensive understanding of the learned representation using *WB-SLR*, we further analyze the diagnoses groups learned as being positively correlated with the outcome event in at least one third of the classifiers in the ensemble. The findings are also supported by medical literature.

- *Patient2Vec* learns a personalized interpretable representation based on the longitudinal EHR data of mainly medical events, including diagnoses, procedures, and medications. Instead of using the clinical groupings of medical codes, this work utilizes the word2vec algorithm to learn a vector representation of each medical code. The dissertation elaborates the structure and algorithm of *Patient2Vec* representation learning framework in detail. In brief, the framework proposes a GRU based model with two-level attentions of multiple hops. Thus, *Patient2Vec* is capable of capturing the complex relationships between clinical events in patients' medical histories such that the prediction performance is improved compared with baseline methods.

  Meanwhile, feature importance are available in the learned representation, which indicates improved interpretability in *Patient2Vec*. A distinguishing characteristics of *Patient2Vec* is that it learns a personalized representation for each individual based on the medical history. Due to the heterogeneity of patients in terms of their EHR data, the feature importance are likely to be distinctive between patients. The *Patient2Vec* is designed to capture the relative importance of a clinical event among all events in a patient's medical record according to its correlation with the target outcome. Thus, the personalization in the learned representation is an additional benefit of *Patient2Vec*. This dissertation then illustrates the importance of clinical events in the medical histories of two example patients to demonstrate the interpretability and personalization achieved by *Patient2Vec*. Additionally, we attempt to interpret the feature importance at the population level by summarizing the top diagnoses, procedures, and medications in terms of the attention weights aggregated in the hospitalized population, as well as among hospitalized patients for specific causes.

In summary, this dissertation proposes three representation learning frameworks for longitudinal EHR data, and then describes the evaluation of the frameworks with prediction tasks using EHR data collected from real clinical setting. Additionally, the learned representations are analyzed and interpreted extensively to bring clinical insights.

In the experiments, we use clinical codes to learn a representation of a patient's medical history. One of the limitations is that the codes in the EHR systems are primarily used for billing purposes rather than prognosis. Additionally, the diagnoses codes in EHR systems are incapable of indicating the severity of illness. Moreover, the diagnoses of chronic diseases, such as diabetes and CKD, are subjective to some degree such that the diagnosis codes of a particular chronic disease could be initially coded anytime in a time range of multiple years. Hence, the diagnoses and/or procedure codes in EHR data might not be perfectly accurate or appropriate for risk prediction of diseases or health outcomes. However, this rich data source is valuable with tremendous patient information and large populations for research in assisting clinical decision making and understanding disease correlations. Ultimately, it provides promising opportunities to help improve health outcomes.

6.2 Contributions

As part of the summary, we elaborate the contributions of this research in the following.

- This research proposes three computational frameworks for representation learning of longitudinal EHR data such that the learned representations can be used for further analytics, such as predictive modeling.

- The learned representations using proposed frameworks are validated with three clinically meaningful and important prediction tasks, respectively, using longitudinal EHR data from real clinical setting. The results demonstrate that the learned representations are capable of improving the prediction performance compared with baseline representations.

- In addition to better prediction performance, this research emphasizes on the interpretability of the learned representations in order to bring clinical insights, deepen the understanding of disease correlations, and discover relationships between medical events.
- One of the designed frameworks attempts to learn a personalized representation for each patient to provide more precise characterization, more accurate prediction of health outcomes, and a better understanding of a patient's medical history for personalized medicine.
- In addition to EHR data, the methods proposed in this research could be applied to complex temporal knowledge representation tasks on sequence data in other domains.

6.3 Future Directions

The work presented in this dissertation can be extended in many directions. For example, other potential methods can be explored to avoid overfitting of data in the *SLR* and *WB-SLR* frameworks, as well as extending the attention mechanism in *Patient2Vec* to allow the network to learn weights of patient characteristics in a more automatic manner. In the following, we summarize other open areas for future exploration.

Random Feature Subset in WB-SLR: The bagging strategy is utilized in *WB-SLR* to improve prediction performance and to reduce variance. Future improvement on the proposed model could employ the strategy of random forest which randomly samples a subset of features when growing each tree. This could potentially reduce the correlation between the base models in the ensemble and further improve the prediction accuracy.

Transfer Learning: It is very likely that there is a small patient cohort for certain representation learning and predictive modeling tasks in clinical domain. In the experiment of early detection of CKD in diabetic patients, the sample size of the positive class is small which limits the performance of the proposed methods and further predictive modeling. To address this issue, we could potentially employ the medical records of CKD patients without diabetes or with concurrent diabetes. This is because that CKD in patients with or without diabetes are similar to each other and most likely share some common risk factors or symptoms. Hence, there is a potential to improve the prediction performance by including those CKD patients into the originally identified positive class. Thus, transfer learning approaches will be explored to transfer the knowledge learned from a similar population to address the target problem.

Clinical Data Coding Mechanisms: In the evaluation of *SLR* and *WB-SLR* frameworks, we use the general diagnosis and procedure categories of ICD-9 and ICD-10 codes from the AHRQ clinical classification scheme [64]. This current coarse categories might introduce information loss, and future work will explore more specific medical codes and other grouping strategies.

Other Data Sources: In addition to discrete clinical events, the proposed framework can be extended to incorporate lab results in the future for more comprehensive representation and higher prediction performance.

Other Applications: The proposed representation learning frameworks could be further validated on more EHR datasets and could be utilized for more applications. In the evaluation of *Patient2Vec*, the experiment in this current work is on the prediction of all-cause hospitalization, which could be refined to the risk prediction of hospitalization in a certain patient population or for a certain cause. Ultimately, the proposed methods could be applied to other complex temporal knowledge representation and prediction tasks within and outside the health care domain.

References

[1] C. McCormick, "Word2vec tutorial - the skip-gram model." http://www.mccormickml.com, 2016, [Accessed on October 2, 2017].

[2] M. Nielsen, "Neural networks and deep learning," http://neuralnetworksanddeeplearning.com/, 2017, [Accessed on October 4, 2017].

[3] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[6] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[7] A. K. Jha, C. M. DesRoches, E. G. Campbell, K. Donelan, S. R. Rao, T. G. Ferris, A. Shields, S. Rosenbaum, and D. Blumenthal, "Use of electronic health records in US hospitals," *New England Journal of Medicine*, vol. 360, no. 16, pp. 1628–1638, 2009.

[8] J. Lindstrom and J. Tuomilehto, "The diabetes risk score: A practical tool to predict type 2 diabetes risk," *Diabetes Care*, vol. 26, no. 3, pp. 725–731, 2003.

[9] G. C. M. Siontis, I. Tzoulaki, K. C. Siontis, and J. P. A. Ioannidis, "Comparisons of established risk prediction models for cardiovascular disease: systematic review," *BMJ*, vol. 344, 2012.

[10] 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 Systems with Applications*, vol. 42, no. 20, pp. 7110–7120, Nov. 2015.

[11] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, "Data mining in healthcare and biomedicine: A survey of the literature," *Journal of Medical Systems*, vol. 36, no. 4, pp. 2431–2448, May 2011.

[12] S. M. Domchek, A. Eisen, K. Calzone, J. Stopfer, A. Blackwood, and B. L. Weber, "Application of breast cancer risk prediction models in clinical practice," *Journal of Clinical Oncology*, vol. 21, no. 4, pp. 593–601, 2003.

[13] E. H. Kennedy, W. L. Wiitala, R. A. Hayward, and J. B. Sussman, "Improved cardiovascular risk prediction using nonparametric regression and electronic health record data," *Medical care*, vol. 51, no. 3, p. 251, 2013.

[14] N. Lee, A. F. Laine, J. Hu, F. Wang, J. Sun, and S. Ebadollahi, "Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients," in *Healthcare Informatics, Imaging and Systems Biology (HISB), 2011 First IEEE International Conference on.* IEEE, 2011, pp. 250–257.

[15] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature reviews. Genetics*, vol. 13, no. 6, p. 395, 2012.

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[17] A. Snowdon, C. Alessi, and K. Schnarr, *" It's All about Me": The Personalization of Health Systems.* Ivey Business School, International Centre for Health Innovation, Western University, 2014.

[18] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: a patient-centered framework," *Journal of general internal medicine*, vol. 28, no. 3, pp. 660–665, 2013.

[19] W.-L. Liao and F.-J. Tsai, "Personalized medicine: a paradigm shift in healthcare," *BioMedicine*, vol. 3, no. 2, pp. 66–72, 2013.

[20] University of Virginia School of Medicine, "UVa clinical data repository," https://med.virginia.edu/phs/office-of-the-chair/administrative-divisions/division-of-biomedical-informatics/uva-clinical-data-repository/, [Accessed on September 25, 2017].

[21] V. N. Slee, "The international classification of diseases: ninth revision (ICD-9)," *Annals of internal medicine*, vol. 88, no. 3, pp. 424–426, 1978.

[22] W. H. Organization *et al.*, "ICD-10: international statistical classification of diseases and related health problems: tenth revision," 2004.

[23] M. Beebe, J. Dalton, M. Espronceda, D. Evans, R. Glenn, and G. Green, "CPT: Standard: Current procedural terminology," *Amer Medical Assn*, 2007.

[24] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[25] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1986.

[26] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.

[27] D. Meyer, "How exactly does word2vec work?" 2016.

[28] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[29] K. Ng, J. Sun, J. Hu, and F. Wang, "Personalized predictive modeling and risk factor identification using patient similarity," *AMIA Summits on Translational Science Proceedings*, vol. 2015, p. 132, 2015.

[30] S. H. Huang, P. LePendu, S. V. Iyer, M. Tai-Seale, D. Carrell, and N. H. Shah, "Toward personalizing treatment for depression: predicting diagnosis and severity," *Journal of the American Medical Informatics Association: JAMIA*, vol. 21, no. 6, pp. 1069–1075, Dec. 2014.

[31] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. Sydney, Australia: ACM, 2015, pp. 705–714.

[32] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Data Engineering, 1995. Proceedings of the Eleventh International Conference on.* IEEE, 1995, pp. 3–14.

[33] M. J. Zaki, "Spade: An efficient algorithm for mining frequent sequences," *Machine learning*, vol. 42, no. 1, pp. 31–60, 2001.

[34] P. Fournier-Viger, A. Gomariz, M. Campos, and R. Thomas, "Fast vertical mining of sequential patterns using co-occurrence information," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining.* Springer, 2014, pp. 40–52.

[35] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.

[36] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2002, pp. 429–435.

[37] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, and A. Laine, "A framework for mining signatures from event sequences and its applications in healthcare data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 2, pp. 272–285, Feb 2013.

[38] J. Zhang, H. Xiong, Y. Huang, H. Wu, K. Leach, and L. E. Barnes, "M-SEQ: Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data," in *Big Data (Big Data), 2015 IEEE International Conference on.* IEEE, 2015, pp. 2569–2577.

[39] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2012, pp. 453–461.

[40] D. Gotz, F. Wang, and A. Perer, "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data," *Journal of Biomedical Informatics*, vol. 48, pp. 148–159, Apr. 2014.

[41] J. Zhang, J. Gong, and L. Barnes, "HCNN: Heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records," in *Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017 IEEE/ACM International Conference on.* IEEE, 2017, pp. 214–221.

[42] C. R. Shalizi, *Advanced Data Analysis from an Elementary Point of View.* Cambridge University Press, 2015.

[43] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[44] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[45] ——, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[46] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural networks and learning machines.* Pearson Upper Saddle River, NJ, USA:, 2009, vol. 3.

[47] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010.* Springer, 2010, pp. 177–186.

[48] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[49] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14, pp. 2627–2636, 1998.

[50] A. Ng, "Convolutional neural network," http://ufldl.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork/, 2017, [Accessed on October 4, 2017].

[51] F.-F. Li, J. Johnson, and S. Yeung, "Convolutional neural networks for visual recognition," http://cs231n.github.io/convolutional-networks/, 2017, [Accessed on October 4, 2017].

[52] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[54] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[55] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[56] A. D. Association *et al.*, "Diagnosis and classification of diabetes mellitus," *Diabetes care*, vol. 37, no. Supplement 1, pp. S81–S90, 2014.

[57] K. G. M. M. Alberti and P. f. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. part 1: diagnosis and classification of diabetes mellitus. provisional report of a who consultation," *Diabetic medicine*, vol. 15, no. 7, pp. 539–553, 1998.

[58] Centers for Disease Control and Prevention (CDC), "National diabetes statistics report, 2017," https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf, 2017, [Accessed on September 27, 2017].

[59] American Diabetes Association, "Statistics about diabetes," http://www.diabetes.org/diabetes-basics/statistics/, 2017, [Accessed on September 27, 2017].

[60] The Advisory Board, "Diabetes is the 3rd-not 7th-leading cause of death, study suggests," https://www.advisory.com/daily-briefing/2017/01/30/diabetes-2, 2017, [Accessed on September 27, 2017].

[61] R. A. Bailey, Y. Wang, V. Zhu, and M. F. Rupnow, "Chronic kidney disease in us adults with type 2 diabetes: an updated national estimate of prevalence based on kidney disease: Improving global outcomes (KDIGO) staging," *BMC research notes*, vol. 7, no. 1, p. 415, 2014.

[62] American Diabetes Association *et al.*, “Economic costs of diabetes in the US in 2012,” *Diabetes care*, vol. 36, no. 4, pp. 1033–1046, 2013.

[63] M. O’Malley, “What its like to have uncontrolled type 2 diabetes,” https://www.everydayhealth.com/type-2-diabetes/living-with/what-its-like-to-have-uncontrolled-type-2-diabetes/, 2016, [Accessed on September 27, 2017].

[64] Agency for Healthcare Research and Quality (AHRQ), “Clinical classifications software (CCS) for ICD-9-CM,” https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp, 2015, [Accessed on September 10, 2016].

[65] United States Census Bureau, “Americans are visiting the doctor less frequently, Census Bureau Reports,” https://www.census.gov/newsroom/releases/archives/health_care_insurance/cb12-185.html, 2017, [Accessed on September 27, 2017].

[66] S. Subramanian and A. Chait, “Hypertriglyceridemia secondary to obesity and diabetes,” *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, vol. 1821, no. 5, pp. 819–825, 2012.

[67] F. L. Dunn, “Hyperlipidemia and diabetes,” *Medical Clinics of North America*, vol. 66, no. 6, pp. 1347–1360, 1982.

[68] D. Berkowitz, “Gout, hyperlipidemia, and diabetes interrelationships,” *Jama*, vol. 197, no. 2, pp. 77–80, 1966.

[69] S. L. Abbate and J. D. Brunzell, “Pathophysiology of hyperlipidemia in diabetes mellitus.” *Journal of Cardiovascular pharmacology*, vol. 16, pp. S1–S7, 1990.

[70] L. C. Plantinga, D. C. Crews, J. Coresh, E. R. Miller, R. Saran, J. Yee, E. Hedgeman, M. Pavkov, M. S. Eberhardt, D. E. Williams *et al.*, “Prevalence of chronic kidney disease in us adults with undiagnosed diabetes or prediabetes,” *Clinical Journal of the American Society of Nephrology*, pp. CJN–07 891 109, 2010.

[71] E. Ferrannini, D. Santoro, and V. Manicardi, “The association of essential hypertension and diabetes.” *Comprehensive therapy*, vol. 15, no. 11, pp. 51–58, 1989.

[72] C. Fritschi and L. Quinn, “Fatigue in patients with diabetes: a review,” *Journal of psychosomatic research*, vol. 69, no. 1, pp. 33–41, 2010.

[73] M. C. Thomas, “The high prevalence of anemia in diabetes is linked to functional erythropoietin deficiency,” in *Seminars in nephrology*, vol. 26, no. 4. Elsevier, 2006, pp. 275–282.

[74] D. R. Bosman, A. S. Winkler, J. T. Marsden, I. C. Macdougall, and P. J. Watkins, "Anemia with erythropoietin deficiency occurs early in diabetic nephropathy," *Diabetes care*, vol. 24, no. 3, pp. 495–499, 2001.

[75] Ö. Tarim, A. Küçükerdogan, Ü. Günay, Ö. Eralp, and İ. Ercan, "Effects of iron deficiency anemia on hemoglobin a1c in type 1 diabetes mellitus," *Pediatrics international*, vol. 41, no. 4, pp. 357–362, 1999.

[76] G. Bönner, "Hyperinsulinemia, insulin resistance, and hypertension." *Journal of cardiovascular pharmacology*, vol. 24, pp. S39–49, 1994.

[77] M. R. Castro, "Is hyperinsulinemia a form of diabetes?" http://www.mayoclinic.org/diseases-conditions/type-2-diabetes/expert-answers/hyperinsulinemia/faq-20058488, 2017, [Accessed on September 27, 2017].

[78] J. M. Olefsky, "The insulin receptor: its role in insulin resistance of obesity and diabetes," *Diabetes*, vol. 25, no. 12, pp. 1154–1161, 1976.

[79] J. S. Rao and R. Tibshirani, "The out-of-bootstrap method for model averaging and selection," *University of Toronto*, 1997.

[80] J. R. Shewchuk *et al.*, "An introduction to the conjugate gradient method without the agonizing pain," 1994.

[81] A. S. Levey and J. Coresh, "Chronic kidney disease," *The lancet*, vol. 379, no. 9811, pp. 165–180, 2012.

[82] A. S. Levey, J. Coresh, E. Balk, A. T. Kausz, A. Levin, M. W. Steffes, R. J. Hogg, R. D. Perrone, J. Lau, and G. Eknoyan, "National kidney foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification," *Annals of internal medicine*, vol. 139, no. 2, pp. 137–147, 2003.

[83] Mayo Clinic, "Chronic kidney disease," http://www.mayoclinic.org/diseases-conditions/chronic-kidney-disease/home/ovc-20207456, 2017, [Accessed on September 27, 2017].

[84] National Institute of Diabetes and Digestive and Kidney Diseases, "Kidney disease statistics for the United States," https://www.niddk.nih.gov/health-information/health-statistics/kidney-disease, 2016, [Accessed on September 28, 2017].

[85] V. Jha, G. Garcia-Garcia, K. Iseki, Z. Li, S. Naicker, B. Plattner, R. Saran, A. Y.-M. Wang, and C.-W. Yang, "Chronic kidney disease: global dimension and perspectives," *The Lancet*, vol. 382, no. 9888, pp. 260–272, 2013.

[86] N. R. Hill, S. T. Fatoba, J. L. Oke, J. A. Hirst, C. A. OCallaghan, D. S. Lasserson, and F. R. Hobbs, "Global prevalence of chronic kidney disease–a systematic review and meta-analysis," *PLoS One*, vol. 11, no. 7, p. e0158765, 2016.

[87] J. Coresh, E. Selvin, L. A. Stevens, J. Manzi, J. W. Kusek, P. Eggers, F. Van Lente, and A. S. Levey, "Prevalence of chronic kidney disease in the United States," *JAMA*, vol. 298, no. 17, pp. 2038–2047, 2007.

[88] D. M. Naranjo, L. Fisher, P. A. Areán, D. Hessler, and J. Mullan, "Patients with type 2 diabetes at risk for major depressive disorder over time," *The Annals of Family Medicine*, vol. 9, no. 2, pp. 115–120, 2011.

[89] Centers for Disease Control and Prevention (CDC), "National chronic kidney disease fact sheet, 2017," https://www.cdc.gov/diabetes/pubs/pdf/kidney_factsheet.pdf, 2017, [Accessed on September 29, 2017].

[90] F. Locatelli, L. Del Vecchio, and P. Pozzoni, "The importance of early detection of chronic kidney disease," *Nephrology Dialysis Transplantation*, vol. 17, no. suppl 11, pp. 2–7, 2002.

[91] J. L. Babitt and H. Y. Lin, "Mechanisms of anemia in ckd," *Journal of the American Society of Nephrology*, pp. ASN–2 011 111 078, 2012.

[92] L. Gotloib, D. Silverberg, R. Fudin, and A. Shostak, "Iron deficiency is a common cause of anemia in chronic kidney disease and can often be corrected with intravenous iron." *Journal of nephrology*, vol. 19, no. 2, pp. 161–167, 2006.

[93] W. McClellan, S. L. Aronoff, W. K. Bolton, S. Hood, D. L. Lorber, K. L. Tang, T. F. Tse, B. Wasserman, and M. Leiserowitz, "The prevalence of anemia in patients with chronic kidney disease," *Current medical research and opinion*, vol. 20, no. 9, pp. 1501–1510, 2004.

[94] C.-y. Hsu, C. E. McCulloch, and G. C. Curhan, "Epidemiology of anemia associated with chronic renal insufficiency among adults in the United States: results from the third national health and nutrition examination survey," *Journal of the American Society of Nephrology*, vol. 13, no. 2, pp. 504–510, 2002.

[95] N. M. Patel, O. M. Gutiérrez, D. L. Andress, D. W. Coyne, A. Levin, and M. Wolf, "Vitamin D deficiency and anemia in early chronic kidney disease," *Kidney international*, vol. 77, no. 8, pp. 715–720, 2010.

[96] R. C. Atkins, "The epidemiology of chronic kidney disease," *Kidney international*, vol. 67, pp. S14–S18, 2005.

[97] M. Liu, X. Li, L. Lu, Y. Cao, R. Sun, S. Chen, and P. Zhang, "Cardiovascular disease and its relationship with chronic kidney disease," *Eur Rev Med Pharmacol Sci*, vol. 18, no. 19, pp. 2918–26, 2014.

[98] A. V. Krishnan and M. C. Kiernan, "Neurological complications of chronic kidney disease," *Nature Reviews Neurology*, vol. 5, no. 10, pp. 542–551, 2009.

[99] R. Arnold, T. Issar, A. V. Krishnan, and B. A. Pussell, "Neurological complications in chronic kidney disease," *JRSM cardiovascular disease*, vol. 5, p. 2048004016677687, 2016.

[100] N. C. C. for Chronic Conditions (Great Britain), "Chronic kidney disease: national clinical guideline for early identification and management in adults in primary and secondary care." Royal College of Physicians, 2008.

[101] R. Kazancioğlu, "Risk factors for chronic kidney disease: an update," *Kidney international supplements*, vol. 3, no. 4, p. 368, 2013.

[102] National Kidney Foundation, "3 early warning signs of kidney disease," https://www.kidney.org/blog/kidney-cars/3-early-warning-signs-kidney-disease, 2017, [Accessed on September 27, 2017].

[103] A. Levin, G. Bakris, M. Molitch, M. Smulders, J. Tian, L. Williams, and D. Andress, "Prevalence of abnormal serum vitamin D, PTH, calcium, and phosphorus in patients with chronic kidney disease: results of the study to evaluate early kidney disease," *Kidney international*, vol. 71, no. 1, pp. 31–38, 2007.

[104] E. A. González, A. Sachdeva, D. A. Oliver, and K. J. Martin, "Vitamin D insufficiency and deficiency in chronic kidney disease," *American journal of nephrology*, vol. 24, no. 5, pp. 503–510, 2004.

[105] R. E. LaClair, R. N. Hellman, S. L. Karp, M. Kraus, S. Ofner, Q. Li, K. L. Graves, and S. M. Moe, "Prevalence of calcidiol deficiency in CKD: a cross-sectional study across latitudes in the United States," *American Journal of Kidney Diseases*, vol. 45, no. 6, pp. 1026–1033, 2005.

[106] D. J. Pierson, "Respiratory considerations in the patient with renal failure," *Respiratory care*, vol. 51, no. 4, pp. 413–422, 2006.

[107] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[108] M. M. Lopez and J. Kalita, "Deep learning applied to nlp," *arXiv preprint arXiv:1703.03091*, 2017.

[109] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[110] D. Britz, "Recurrent neural network tutorial," http://www.wildml.com/2015/10/, 2015, [Accessed on October 5, 2017].

[111] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[112] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.

[113] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[114] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *arXiv preprint arXiv:1606.01865*, 2016.

[115] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, 2016.

[116] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, 2016, pp. 301–318.

[117] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Interpretable deep models for icu outcome prediction," in *AMIA Annual Symposium Proceedings*, vol. 2016. American Medical Informatics Association, 2016, p. 371.

[118] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[119] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504–3512.

[120] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[121] C. M. Torio and B. J. Moore, "National inpatient hospital costs: The most expensive conditions by payer, 2013," https://www.hcup-us.ahrq.gov/reports/statbriefs/sb204-Most-Expensive-Hospital-Conditions.jsp, 2016, [Accessed on October 2, 2017].

[122] 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," *Medical care*, vol. 52, no. 8, p. 751, 2014.

[123] 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," *Quality and safety in health care*, vol. 17, no. 3, pp. 216–223, 2008.

[124] S. Purdey and A. Huntley, “Predicting and preventing avoidable hospital admissions: a review.” *The journal of the Royal College of Physicians of Edinburgh*, vol. 43, no. 4, pp. 340–344, 2012.

[125] Canadian Institute for Health Information, “Early identification of people at-risk of hospitalization,” https://secure.cihi.ca/free_products/HARP_reportv_En.pdf, 2013, [Accessed on October 2, 2017].

[126] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani, “Risk prediction models for hospital readmission: a systematic review,” *Jama*, vol. 306, no. 15, pp. 1688–1698, 2011.

[127] G. Giamouzis, A. Kalogeropoulos, V. Georgiopoulou, S. Laskar, A. L. Smith, S. Dunbar, F. Triposkiadis, and J. Butler, “Hospitalization epidemic in patients with heart failure: risk factors, risk prediction, knowledge gaps, and future directions,” *Journal of cardiac failure*, vol. 17, no. 1, pp. 54–75, 2011.

[128] 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,” *European Respiratory Journal*, vol. 10, no. 4, pp. 822–827, 1997.