Risk Prediction on Electronic Health Records with Prior Medical Knowledge

Fenglong Ma
SUNY Buffalo
Buffalo, NY, USA
fenglong@buffalo.edu

Jing Gao
SUNY Buffalo
Buffalo, NY, USA
jing@buffalo.edu

Qiuling Suo
SUNY Buffalo
Buffalo, NY, USA
qiulings@buffalo.edu

Quanzeng You
Microsoft AI & Research
Redmond, WA
quanzeng.you@microsoft.com

Jing Zhou
Ehealth Inc
Mountain View, CA
jing.zhou@ehealth.com

Aidong Zhang
SUNY Buffalo
Buffalo, NY
azhang@buffalo.edu

ABSTRACT

Predicting the risk of potential diseases from Electronic Health Records (EHR) has attracted considerable attention in recent years, especially with the development of deep learning techniques. Compared with traditional machine learning models, deep learning based approaches achieve superior performance on risk prediction task. However, none of existing work explicitly takes prior medical knowledge (such as the relationships between diseases and corresponding risk factors) into account. In medical domain, knowledge is usually represented by discrete and arbitrary rules. Thus, how to integrate such medical rules into existing risk prediction models to improve the performance is a challenge. To tackle this challenge, we propose a novel and general framework called PRIME for risk prediction task, which can successfully incorporate discrete prior medical knowledge into all of the state-of-the-art predictive models using posterior regularization technique. Different from traditional posterior regularization, we do not need to manually set a bound for each piece of prior medical knowledge when modeling desired distribution of the target disease on patients. Moreover, the proposed PRIME can automatically learn the importance of different prior knowledge with a log-linear model. Experimental results on three real medical datasets demonstrate the effectiveness of the proposed framework for the task of risk prediction[1].

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Applied computing** → **Health informatics**;

KEYWORDS

Healthcare Informatics; Prior Medical Knowledge; Posterior Regularization

ACM Reference Format:
Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. 2018. Risk Prediction on Electronic Health Records with Prior Medical Knowledge. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220020

1 INTRODUCTION

With the immense accumulation of *Electronic Health Records* (EHR) being available, the analysis of such data enables researchers and healthcare providers to get closer to the goal of personalized medicine. However, raw EHR data has its own issues, such as high dimensionality, temporality, sparsity, irregularity and bias [5]. These challenges dramatically increase the difficulty of directly applying traditional machine learning or statistical models [13, 14, 27, 28, 32] to predict patients' potential diseases, which is a core task in medical domain, named *risk prediction*. Therefore, it is crucial to develop more powerful models for solving the challenges introduced by the raw EHR data in risk prediction task.

Recently, deep learning models have shown the ability of directly extracting meaningful features from raw electronic health records in many domains, including computational phenotyping [1, 4], diagnosis prediction [7, 8, 20], risk prediction [2, 3, 5, 9, 25], and so on. Especially for risk prediction task, attention-based recurrent neural networks (RNN) are employed to predict the disease of Heart Failure in [9]. Convolutional neural networks (CNN) are also introduced to capture the local temporal characteristics of patients' visits and predict the risks of diseases [2, 3, 5], with improvement in performance.

Though the aforementioned deep learning based models have achieved good performance in the risk prediction task, they all ignore the importance of **prior medical knowledge**, such as the relationships between diseases and their corresponding risk factors. As we all know, prior medical knowledge plays an important role in healthcare domain. When a patient visits a doctor, the doctor first reviews the current symptoms, and then takes a careful review on medical history, such as *medications*, *smoking history*, *alcohol use*, and *diseases of family history*, which are risk factors of diseases.

[1]The PRIME source code is publicly available at http://www.acsu.buffalo.edu/~fenglong.

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3220020

With the current symptoms and patient's past medical history, the doctor may have an initial diagnosis for this patient. For example, the symptoms of a patient are *rapid irregular heartbeat associated with shortness of breath*, *increased need to urinate at night*, *chest pain* and *fainting*. He/She has been suffering *high blood pressure* and *coronary artery disease* more than *eight years*. According to the experience (or prior medical knowledge) and current symptoms, the doctor can quickly diagnose that the patient may have heart failure rather than other diseases. It is because *high blood pressure* and *coronary artery disease* are two key risk factors of heart failure. Therefore, considering prior medical knowledge is essential for risk prediction task.

However, it is extremely difficult to directly apply prior medical knowledge to EHR data. On the one hand, the medical knowledge is arbitrary or heterogeneous. Some diseases may be related to age (continuous value), while others are caused by the habits such as smoking or drinking (categorical value). On the other hand, almost all the medical knowledge is represented by rules. Thus, transforming the *discrete* arbitrary medical rules into the *continuous* real values is a thought-provoking problem. Even if we can obtain the real-valued representations of prior medical knowledge, how to reasonably combine the knowledge with the predictive models is still a challenge.

Posterior regularization [12] is an effective technique to convert the discrete knowledge into continuous real-valued features by modeling the posterior distribution as a constrained posterior feature set. However, the main drawback of directly applying posterior regularization technique is that it needs to manually set a bound for each constraint feature, which is impractical in medical domain. For example, when predicting the risk of heart failure disease for a patient, the doctor may consider the frequency of historical diseases (called underlying diseases in medical domain) and their durations. Here, the frequency and durations of underlying diseases can be modeled as constraint features. It is hard to set exact bound values for these two constraint features to determine whether the patient has heart failure or not. Obviously, the key challenge here is how to automatically learn the bound values of constraint features and guarantee the predictive performance meanwhile.

To tackle all the aforementioned challenges, in this paper, we propose a novel predictive framework PRIME, which can successfully integrate heterogeneous discrete PRIor MEdical knowledge into the predictive models to improve the performance. Specifically, the framework can employ all the existing deep learning based approaches as the basic predictive model, such as recurrent neural networks (RNN) and convolutional neural networks (CNN). To automatically learn the bounds of constraint features, we use a log-linear model in the proposed PRIME instead of modeling the posterior distribution as a constrained posterior set. It not only makes the training process of the proposed model more efficient, but also learns different weights for different constraint features. We conduct experiments on three medical datasets. The results show that the proposed framework PRIME is able to incorporate heterogeneous prior medical knowledge and outperforms existing risk prediction models.

It is worthwhile to highlight the contributions of the proposed framework as follows:

- To the best of our knowledge, this is the first attempt to take prior medical knowledge into account for risk prediction task.
- We propose a novel framework PRIME, which models prior medical knowledge as posterior regularization and learns the desired posterior distribution with a log-linear model.
- The proposed PRIME is a general model, which can be easily applied to any predictive models in healthcare. Moreover, it is able to distinguish the importance of different prior knowledge contributed to the risk prediction.
- Experimental results on three medical datasets demonstrate that the proposed PRIME is effective for the task of risk prediction.

In the following sections, we first review existing work in Section 2. In Section 3, we introduce the background information on deep learning based risk prediction models and posterior regularization technique. The details of the proposed PRIME are presented in Section 4. In Section 5, we conduct experiments on three real EHR datasets and demonstrate the effectiveness of the proposed PRIME. The limitation of the proposed framework is discussed in Section 6. Finally, we conclude this work in Section 7.

2 RELATED WORK

In this section, we briefly review existing studies which are closely related to our work, including deep learning based models for healthcare applications and posterior regularization techniques with deep learning models.

2.1 Deep Learning for Healthcare

For most healthcare applications, the first step is to extract effective phenotypes from longitudinal EHR [1, 3–5, 7–9, 13, 14, 16, 20, 21, 24, 26–29, 32]. Traditional electronic phenotyping approaches are mainly based on matrix factorization [27, 28, 32] and tensor factorization [13, 14]. Recently, deep learning based models have shown their superior ability to learn complex patterns from high dimensional, noisy and temporal EHR data. Multi-layer perception (MLP) is used to learn the representations of phenotypes [4] and medical codes [7]. However, MLP based models do not consider the temporal nature of the EHR data. To model the temporal EHR data, recurrent neural networks (RNN) are applied to predict patients' health status [8, 9, 20, 24, 26] and patient subtyping [1]. Convolutional neural networks (CNN) focus on capturing local temporal dependency among EHR data and are used for predicting multiple diseases [25] and for other related task.

Risk prediction is an important yet challenging task in healthcare domain. Choi *et al.* [9] try to use attention-based recurrent neural networks to predict the risk of heart failure disease. Cheng *et al.* [5] apply the CNN model to analyze discrete patient EHR data. Che *et al.* [2] propose to use the pretrained embeddings of medical features in the CNN model to improve the prediction performance. In [3], the authors build a semi-supervised deep learning model with generative adversarial networks for the risk prediction task.

Compared with all the aforementioned predictive models, the proposed framework PRIME has the following advantages: (1) It takes prior medical knowledge into account, and (2) it is a general model that can include any state-of-the-art predictive model when

modeling patients' visits. The prior knowledge guides the predictive models to learn better sub-optimal parameters, which finally leads to good predictive performance.

2.2 Posterior Regularization in Deep Learning

The proposed framework PRIME is inspired by posterior regularization [12], which has been successfully introduced into deep learning models for sentiment classification [17] and machine translation [31] in natural language processing. Hu *et al.* [17] add the first-order logic rules into convolutional neural networks to further enhance the performance of sentiment classification task. However, this work still needs to manually set the bound values of constrained posterior features.

Different from the work [17], we use a log-linear model to represent the desired distribution. Employing log-linear models not only enables the proposed PRIME to incorporate prior medical knowledge as real-valued features, but also makes the proposed framework differentiable.

3 BACKGROUND

In this section, we first describe the EHR data used in this paper, then introduce the basic deep learning based risk prediction models, and finally present the posterior regularization technique.

3.1 EHR Data Description

The EHR data consists of patients' time-ordered visiting records. Let $\mathcal{P}$ denote the set of all the patients, where $|\mathcal{P}|$ is the number of patients in the EHR data. For each patient $p \in \mathcal{P}$, there are T_p time-ordered visits $V_1^{(p)}, V_2^{(p)}, \cdots, V_{T_p}^{(p)}$. We denote $C = \{c_1, c_2, \cdots, c_{|C|}\}$ as the set of all the diagnosis codes or medical events, and $|C|$ represents the number of unique diagnosis codes. Each visit $V_t^{(p)}$ includes a subset of diagnosis codes, which is denoted by a vector $\mathbf{x}_t^{(p)} \in \{0,1\}^{|C|}$. The i-th element in $\mathbf{x}_t^{(p)}$ is 1 if $V_t^{(p)}$ contains diagnosis code c_i. Demographical information of patients is also recorded for each visit, such as *gender*, *ethnicity* and *age*. For each patient, we use $\mathbf{g}^{(p)}$ to denote his/her demographical information at time T_p. For simplicity, we drop the superscript (p) when it is unambiguous in the following sections.

3.2 Basic Risk Prediction Models

In this paper, we separately apply two basic deep learning models used for risk prediction: One is a convolutional neural network (CNN) with a 1D convolutional layer over time-ordered visits and a max pooling layer, which has been used in previous work [2, 3, 5], and the other is a basic long-short term memory network (LSTM) [15].

The input of the predictive model is the EHR records of the p-th patient, denoted by $\mathbf{X}^{(p)} = \{\mathbf{x}_t^{(p)}\}_{t=1}^{T_p} \in \mathbb{R}^{T_p \times |C|}$. Since the input $\mathbf{X}^{(p)}$ is too sparse and with high dimensionality, it is natural to learn its low-dimension and meaningful embeddings. Thus, we first embed the input $\mathbf{x}_t$ into visit-level representations $\mathbf{v}_t \in \mathbb{R}^k$ as follows:

$$\mathbf{v}_t = \mathbf{W}_v \mathbf{x}_t + \mathbf{b}_v, \tag{1}$$

where $\mathbf{W}_v \in \mathbb{R}^{k \times |C|}$ and $\mathbf{b}_v \in \mathbb{R}^k$ are parameters to be learned, and k is the size of latent representations. Next, we provide the details of these two predictive models.

CNN Predictive Model. We first apply the convolutional operation only over the temporal dimension of $\mathbf{V}^{(p)} = \{\mathbf{v}_t^{(p)}\}_{t=1}^{T_p} \in \mathbb{R}^{T_p \times k}$. In order to capture the temporal dependencies among multiple visits, we use a combination of m filters with s different window sizes. Let l denote the size of a time window, and then $\mathbf{v}_{t:t+l-1}$ represents the concatenation of l visits from $\mathbf{v}_t$ to $\mathbf{v}_{t+l-1}$. A filter $\mathbf{W}_f \in \mathbb{R}^{l \times k}$ is applied on the window of l visits to produce a new feature $f_t \in \mathbb{R}$ with the ReLU activation function as follows:

$$f_t = \text{ReLU}(\mathbf{W}_f \mathbf{v}_{t:t+l-1} + b_f),$$

where $b_f \in \mathbb{R}$ is a bias term, and $\text{ReLU}(f) = \max(f, 0)$. This filter is applied to each possible window of visits in the whole description $\{\mathbf{v}_{1:l}, \mathbf{v}_{2:l+1}, \cdots, \mathbf{v}_{T_p-l+1:T_p}\}$ to generate a feature map $\mathbf{f} \in \mathbb{R}^{T_p-l+1}$ as follows:

$$\mathbf{f} = [f_1, f_2, \cdots, f_{T_p-l+1}].$$

To obtain the most important feature, max pooling technique [10] is used over the feature map $\mathbf{f}$, i.e., $\hat{f} = \max(\mathbf{f})$. We can see that each filter produces a feature. Since we have m filters with s different window sizes, the final vector representation of the p-th patient can be obtained by concatenating all the extracted features, i.e., $\mathbf{z}^{(p)} \in \mathbb{R}^{ms}$.

Finally, a fully connected softmax layer is applied to produce prediction probabilities as follows:

$$\hat{\mathbf{y}}_p = \text{softmax}(\mathbf{W}_y \mathbf{z}^{(p)} + \mathbf{b}_y), \tag{2}$$

where $\mathbf{W}_y \in \mathbb{R}^{N \times ms}$ and $\mathbf{b}_y \in \mathbb{R}^N$ are the learnable parameters, and N is the number of target diseases. In this work, we focus on the binary prediction task, i.e., $N = 2$.

LSTM Predictive Model. We use the basic LSTM unit [15] in the predictive model, whose behavior is controlled by a set of three gates: input, output and forget gates. The memory unit accumulates the useful information from the input $\mathbf{v}_t$ at time t based on the values of the gates, and stores the information in its internal state. The final output $z^{(p)}$ from LSTM is the vector representation of patient p. Finally, Eq. (2) is used for prediction.

Let $\boldsymbol{\theta}$ be the set of all the parameters in the CNN/LSTM model, and the prediction probability vector $\hat{\mathbf{y}}_p$ can also be denoted by model *posterior distribution* $P(\mathbf{y}_p|\mathbf{X}^{(p)}; \boldsymbol{\theta})$, where $\mathbf{y}_p$ is the ground truth. The cross-entropy between the ground truth $\mathbf{y}_p$ and the prediction probabilities $\hat{\mathbf{y}}_p$ is used to calculate the loss. Thus, the objective function of risk prediction is the average of cross-entropy:

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{|\mathcal{P}|} \sum_{p=1}^{|\mathcal{P}|} \left(\mathbf{y}_p^\top \log(\hat{\mathbf{y}}_p) + (\mathbf{1} - \mathbf{y}_p)^\top \log(\mathbf{1} - \hat{\mathbf{y}}_p) \right). \tag{3}$$

Though the predictive models have shown their superior ability for the risk prediction task, they all ignore the importance of prior medical knowledge. For example, it is known that the heart works harder than it has to if the blood pressure is high. In other words, *high blood pressure* is an important factor to judge whether the patient will suffer the heart failure disease in the future. Therefore,

it is crucial to design a new framework for integrating priori medical knowledge into risk prediction model.

3.3 Posterior Regularization

Posterior regularization [12] is proposed to incorporate indirect supervision (i.e., priori medical knowledge) via structural constraints on posterior distributions of latent variables. The goal of posterior regularization is to restrict the space of model posteriors using priori knowledge to guide the model towards desired parameter distributions. Let $q(\mathbf{y}_p)$ denote the desired distribution of patient p. The posterior regularized loss function is defined as

$$\mathcal{F}(\boldsymbol{\theta}, \boldsymbol{q}) = \mathcal{L}(\boldsymbol{\theta}) + \alpha \frac{1}{|\mathcal{P}|} \sum_{p=1}^{|\mathcal{P}|} \min_{q \in Q} \mathrm{KL}\left(q(\mathbf{y}_p) || P(\mathbf{y}_p | \mathbf{X}^{(p)}; \boldsymbol{\theta})\right), \quad (4)$$

where α is a hyper-parameter to balance the preference between the loss of predictive model (Eq. (3)) and posterior regularization, and $\mathrm{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence to measure the difference between the desired distribution $q(\mathbf{y}_p)$ and the posterior distribution $P(\mathbf{y}_p|\mathbf{X}^{(p)}; \boldsymbol{\theta})$ of the predictive model. Q is a set of constraints for posterior information and defined as:

$$Q = \{q(\mathbf{y}_p) \ : \mathbb{E}_q[\boldsymbol{\phi}(\mathbf{X}^{(p)}, \mathbf{y}_p)] \leq \mathbf{b}\},$$

where $\boldsymbol{\phi}(\mathbf{X}^{(p)}, \mathbf{y}_p)$ is the set of constraint features and $\mathbf{b}$ is the (known) bound of constraint feature expectations. However, in risk prediction task, it is hard to specify the value of $\mathbf{b}$ to effectively bound the exceptions of constraint features. For example, the risk factors of heart failure include *high blood pressure*, *diabetes*, *heart attack*, and so on[2]. Even for the experienced doctors, they hardly provide the exact bounds of different risk factors. The other challenge is that the same risk factor may cause multiple diseases. Taking *diabetes* as an example, it causes not only heart failure, but also chronic kidney disease[3]. As the expectation of the same risk factor causing different diseases may be different, it is even more difficult to set different bound values for different diseases, and thus directly applying such posterior regularization techniques may not be practical.

4 RISK PREDICTION FRAMEWORK WITH PRIORI MEDICAL KNOWLEDGE

To tackle the aforementioned challenges in Section 3, in this work, we propose a novel framework PRIME that incorporates posterior regularization technique [12] into risk prediction. We first introduce the proposed framework and then present how to design constraint features with priori medical knowledge for the target disease.

4.1 The Proposed Framework PRIME

Figure 1 shows the overview of the proposed framework PRIME for the task of risk prediction. Given the input data $\mathbf{X}^{(p)}$, to predict its true label vector $\mathbf{y}_p$, we can use the predictive model to obtain the prediction probability vector $\hat{\mathbf{y}}_p = P(\mathbf{y}_p|\mathbf{X}^{(p)}; \boldsymbol{\theta})$. The main objective of the proposed PRIME is to integrate the prior medical

[2]https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142

[3]https://www.mayoclinic.org/diseases-conditions/chronic-kidney-disease/symptoms-causes/syc-20354521

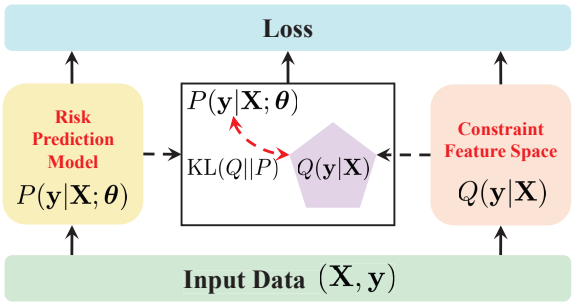


Figure 1: Overview of the Proposed Framework PRIME.

knowledge into the basic risk prediction model. To achieve this goal, a desired distribution $q(\mathbf{y}_p)$ is introduced along with posterior regularization technique. However, as we discussed in Section 3.3, we cannot directly optimize Eq. (4) to obtain the optimal parameters for risk prediction model. To solve the first challenge, that is how to specify the bound $\mathbf{b}$ for different constraint features, we use a log-linear model [22] to represent the desired distribution $q(\mathbf{y}_p)$. The objective function can be rewritten as follows:

$$\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \mathcal{W}) = \mathcal{L}(\boldsymbol{\theta}) + \alpha \frac{1}{|\mathcal{P}|} \sum_{p=1}^{|\mathcal{P}|} \mathrm{KL}\left(\tilde{\mathbf{y}}_p || P(\mathbf{y}_p|\mathbf{X}^{(p)}; \boldsymbol{\theta})\right) + \beta \mathcal{L}'(\boldsymbol{\Gamma}, \mathcal{W}), \quad (5)$$

where the desired distribution $\tilde{\mathbf{y}}_p = Q(\mathbf{y}_p|\mathbf{X}^{(p)}; \boldsymbol{\Gamma}, \mathcal{W})$ that encodes priori medical knowledge is defined as follows:

$$Q(\mathbf{y}_p|\mathbf{X}^{(p)}; \boldsymbol{\Gamma}, \mathcal{W}) = \frac{\exp\{\boldsymbol{\Gamma} \cdot \boldsymbol{\phi}(\mathbf{X}^{(p)}, \mathbf{y}_p; \mathcal{W})\}}{\sum_{\mathbf{y}'_p} \exp\{\boldsymbol{\Gamma} \cdot \boldsymbol{\phi}(\mathbf{X}^{(p)}, \mathbf{y}'_p; \mathcal{W})\}}, \quad (6)$$

where $\boldsymbol{\Gamma}$ is the learnable confidence matrix for different constraint feature categories according to prior medical knowledge, which will be illustrated in Section 4.2. Introducing the parameter set $\mathcal{W}$ into the constraint feature function makes the proposed model successfully distinguish the difference among multiple pieces of priori knowledge in the same category. In this way, we do not need to manually specify the bound vector $\mathbf{b}$. β is the hyper-parameter, and $\mathcal{L}'(\boldsymbol{\Gamma}, \mathcal{W})$ is the average cross entropy between the desired distribution $\tilde{\mathbf{y}}_p$ and the ground truth $\mathbf{y}_p$, which is defined as follows:

$$\mathcal{L}'(\boldsymbol{\Gamma}, \mathcal{W}) = -\frac{1}{|\mathcal{P}|} \sum_{p=1}^{|\mathcal{P}|} \left(\mathbf{y}_p^\top \log(\tilde{\mathbf{y}}_p) + (\mathbf{1} - \mathbf{y}_p)^\top \log(\mathbf{1} - \tilde{\mathbf{y}}_p)\right).$$

From Eq. (5), we can observe that the proposed approach is a general framework for incorporating knowledge into the predictive model, which can be applied to any prediction task in medical domain, including, but not limited to, risk prediction, diagnosis prediction and survivability prediction. Moreover, the flexibility of log-linear models makes the proposed framework easily represent the arbitrary priori knowledge as constraint features. Furthermore, it is easy to optimize the objective function (Eq. (5)) with standard stochastic gradient descent algorithms when we employ the differentiable log-linear models. Finally, the design of the desired distribution in Eq. (6) can successfully tackle the problem of manually setting bound values for constraint features. The proposed PRIME

can automatically assign different confidence levels for the same constraint feature when predicting the risk of different diseases.

Next, we will introduce how to design constraint features to integrate priori medical knowledge into the desired distribution for risk prediction in detail.

4.2 Constraint Feature Design

Since different diseases have different risk factors, we cannot use the same constraint feature with prior medical knowledge to predict these diseases. Fortunately, in medical domain, doctors have classified risk factors into five main categories: patient characteristics, underlying diseases, disease duration, genetics and family history. In the following, we formally provide the design of these constraint features.

Patient Characteristics

In healthcare, it is natural to consider the characteristics of patients such as gender, age and ethnicity, when predicting the risk of diseases. For example, people of certain races, including *Blacks*, *Hispanics*, *American Indians* and *Asian-Americans*, are at higher risk of suffering type 2 diabetes[4]. Since COPD develops slowly over years, most people are *at least 40 years old* when symptoms begin[5]. Thus, it is important to design constraint features for patient characteristics. In this paper, we mainly focus on two characteristics of patients: ethnicity and age.

Given the demographical information $\mathbf{g}^{(p)} = [g_e^{(p)}, g_a^{(p)}]$ of patient p and the corresponding label $\mathbf{y}_p$, the feature on ethnicity can be defined as follows:

$$\phi_e(\mathbf{X}^{(p)}, \mathbf{y}_p) = \begin{cases} 1 & \text{if } g_e^{(p)} \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases},$$

where $\mathcal{E}$ denotes the set of races related to the prediction. Since the value of ϕ_e is either 1 or 0, thus the ethnicity vector $\boldsymbol{\phi_e} = [1, 1]$ or $[0, 0]$. To model the different importance on cases and controls, the confidence vector $\boldsymbol{\gamma_e}$ is introduced for the constraint feature ethnicity.

For most of diseases, the risk increases as the patients get older. Thus, the commonly used logistic function is introduced to model the effect of age as follows:

$$\phi_a(\mathbf{X}^{(p)}, \mathbf{y}_p; w_y^{(a)}) = \left\{1 + \exp\{-w_y^{(a)}(g_a^{(p)} - \psi)\}\right\}^{-1},$$

where $w_y^{(a)} \in \mathbb{R}$ is the disease specific parameter to model the influence of age for risk prediction. If the disease is not sensitive to age, then $w_y^{(a)} \to +\infty$. ψ is a predefined scalar. In this paper, we use age groups instead of real ages of patients and set $\psi = 9$ (i.e., the age from 40 to 45). Thus, the age feature vector $\boldsymbol{\phi_a} = [\phi_a(w_0^{(a)}), \phi_a(w_1^{(a)})]$, and $\boldsymbol{\gamma_a}$ is its corresponding confidence vector.

Underlying Diseases

Underlying diseases of patients are the key risk factors for the prediction. Different underlying diseases may have different contributions for the target disease prediction. For example, the underlying diseases of heart failure include high blood pressure, coronary artery disease, diabetes, and so on. If the diagnosis codes about high blood pressure always appear in a patient's visiting records compared with other diseases' codes, then the probability of high blood pressure causing heart failure is higher than that of other underlying diseases.

To fully make use of all the underlying diseases, we first obtain theses diseases for each risk prediction task denoted as $\mathcal{U}$, and then calculate the frequency of those underlying diseases in patient p's visits, which is represented by $\mathbf{u}_p$. The reason is that the greater the frequency, the higher the risk. Additionally, the effect of different underlying diseases is different for the final disease prediction. Therefore, the constraint features of underlying diseases are designed as follows:

$$\phi_u(\mathbf{X}^{(p)}, \mathbf{y}_p; \mathbf{w}_y^{(u)}) = \begin{cases} \left\{1 + \exp(-\mathbf{w}_y^{(u)} \cdot \mathbf{u}_p)\right\}^{-1} & \text{if } sum(\mathbf{u}_p) > 0 \\ 0 & \text{if } sum(\mathbf{u}_p) = 0 \end{cases},$$

where $\mathbf{w}_y^{(u)} \in \mathbb{R}^{|\mathcal{U}|}$ is the leaned parameter to represent the different effect of different underlying diseases, $|\mathcal{U}|$ is the number of underlying diseases, and $sum(\mathbf{u}_p)$ is the sum of $\mathbf{u}_p$. The underlying disease vector is $\boldsymbol{\phi_u} = [\phi_u(\mathbf{w}_0^{(u)}), \phi_u(\mathbf{w}_1^{(u)})]$, and its importance vector is $\boldsymbol{\gamma_u}$.

Disease Duration

Similar to the frequency of underlying diseases, the duration of underlying diseases is another import factor for risk prediction. If a patient p has been diagnosed *high blood pressure* for five years, and the other patient p' has the disease for only one month, then the risk of suffering the disease *heart failure* on patient p is much higher than that on patient p'. In order to obtain the duration of underlying diseases, we first find the start time $t_d^{(p)}$ of a certain underlying disease d from patients' visiting records, and then calculate the duration using $T_p - t_d^{(p)}$. Finally, the duration of diseases is denoted as $\mathbf{d}_p$. Based on $\mathbf{d}_p$, the constraint features on disease duration is defined as follows:

$$\phi_d(\mathbf{X}^{(p)}, \mathbf{y}_p; \mathbf{w}_y^{(d)}) = \begin{cases} \left\{1 + \exp(-\mathbf{w}_y^{(d)} \cdot \mathbf{d}_p)\right\}^{-1} & \text{if } sum(\mathbf{d}_p) > 0 \\ 0 & \text{if } sum(\mathbf{d}_p) = 0 \end{cases},$$

where $\mathbf{w}_y^{(d)} \in \mathbb{R}^{|\mathcal{U}|}$ is similar to $\mathbf{w}_y^{(u)}$ to model the difference among underlying diseases, and $\boldsymbol{\phi_d} = [\phi_d(\mathbf{w}_0^{(d)}), \phi_d(\mathbf{w}_1^{(d)})]$ with confidence vector $\boldsymbol{\gamma_d}$.

Genetics & Family History

Many diseases are caused by abnormalities in an individual's genome[6]. For example, the uncommon genetic disorder *alpha-1-antitrypsin deficiency* is the cause of some cases of COPD. To design the constraint feature for genetics, we first collect a set of genetic disorders $\mathcal{G}$ which are related to the target disease. Let $C^{(p)}$ denote all the diagnosis codes in patient p's visits $\mathbf{X}^{(p)}$. The value of constraint feature is 1 as long as the intersection of $C^{(p)}$ and $\mathcal{G}$ is not empty. The formal mathematical formulation is given as follows:

$$\phi_g(\mathbf{X}^{(p)}, \mathbf{y}_p) = \begin{cases} 1 & \text{if } C^{(p)} \cap \mathcal{G} \neq \emptyset \\ 0 & \text{otherwise} \end{cases}.$$

[4] https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

[5] https://www.mayoclinic.org/diseases-conditions/copd/symptoms-causes/syc-20353679

[6] https://www.genome.gov/10001204/specific-genetic-disorders/

Similar to the constraint feature ethnicity, the value of ϕ_g is 1 or 0. Thus, $\boldsymbol{\phi_g} = [0, 0]$ or $[1, 1]$, and $\boldsymbol{\gamma_g}$ is the confidence vector.

Some diseases are related to the disease history of the whole family, such as chronic kidney disease. We collect the set of family history disorders $\mathcal{H}$, and then provide the constraint feature function as follows:

$$\phi_h(\mathbf{X}^{(p)}, \mathbf{y}_p) = \begin{cases} 1 & \text{if } C^{(p)} \cap \mathcal{H} \neq \emptyset \\ 0 & \text{otherwise} \end{cases},$$

and $\boldsymbol{\phi_h} = [0, 0]$ or $[1, 1,]$ with the confidence vector $\boldsymbol{\gamma_h}$.

Note that (1) in the proposed PRIME framework, the confidence matrix $\mathbf{\Gamma}$ and weights $w \in \mathcal{W}$ of risk factors belonging to different categories can be learned automatically. (2) We use the weighted combination of all the risk factors to predict the risk of diseases, i.e., $\mathbf{\Gamma} \cdot \boldsymbol{\phi}(\mathbf{X}^{(p)}, \mathbf{y}_p; \mathcal{W}) = \boldsymbol{\gamma_e} \odot \boldsymbol{\phi_e} + \boldsymbol{\gamma_a} \odot \boldsymbol{\phi_a} + \boldsymbol{\gamma_u} \odot \boldsymbol{\phi_u} + \boldsymbol{\gamma_d} \odot \boldsymbol{\phi_d} + \boldsymbol{\gamma_g} \odot \boldsymbol{\phi_g} + \boldsymbol{\gamma_h} \odot \boldsymbol{\phi_h}$ in Eq. (6), where $\odot$ is the element-wise multiplication. (3) If the patient p does not have any underlying diseases, i.e., $sum(\mathbf{u}_p) = 0$, the desired distribution $Q(\mathbf{y}_p|\mathbf{X}^{(p)}; \mathbf{\Gamma}, \mathcal{W})$ will be close to $[0.5, 0.5]$, which cannot correctly represent the real distribution. To avoid this phenomenon, we force $Q(\mathbf{y}_p|\mathbf{X}^{(p)}; \mathbf{\Gamma}, \mathcal{W}) = P(\mathbf{y}_p|\mathbf{X}^{(p)}; \boldsymbol{\theta})$ when $sum(\mathbf{u}_p) = 0$.

4.3 Prediction

In the training process, our goal is to learn a set of parameters by minimizing the objective function Eq. (5), i.e.,

$$\hat{\boldsymbol{\theta}}, \hat{\mathbf{\Gamma}}, \hat{\mathcal{W}} = \underset{\boldsymbol{\theta}, \mathbf{\Gamma}, \mathcal{W}}{\operatorname{argmin}} \left\{ \mathcal{J}(\boldsymbol{\theta}, \mathbf{\Gamma}, \mathcal{W}) \right\}.$$

Given the learned parameters, we can predict the risk for an unseen patient $\mathbf{X}^{(p)}$ according to

$$\hat{y}_p = \operatorname{argmax} \left\{ P(\mathbf{y}_p|\mathbf{X}^{(p)}; \hat{\boldsymbol{\theta}}) \right\}. \tag{7}$$

Though Eq. (7) can make predictions for given patients, it ignores the effect of prior medical knowledge. Thus, we use the following formulation to predict the risk of patients:

$$\hat{y}_p = \operatorname{argmax} \left\{ P(\mathbf{y}_p|\mathbf{X}^{(p)}; \hat{\boldsymbol{\theta}}) + Q(\mathbf{y}_p|\mathbf{X}^{(p)}; \hat{\mathbf{\Gamma}}, \hat{\mathcal{W}}) \right\}. \tag{8}$$

5 EXPERIMENTS

To fairly evaluate the effectiveness of the proposed framework PRIME, three real EHR datasets are used, including heart failure, COPD and chronic kidney disease cohorts. The experimental results show that integrating prior medical knowledge indeed improves the performance of onset prediction. Moreover, the proposed PRIME framework is able to learn the importance of different risk factors for the final prediction. Next, we start this section by introducing the datasets and experimental settings, and then provide detailed performance comparison between the proposed PRIME and state-of-the-art approaches.

5.1 Datasets

The datasets are extracted from a real EHR database, and three cohorts are identified: heart failure, COPD and chronic kidney disease. The statistics of these three datasets are listed in Table 1. The goal of this work is to predict whether a patient is from the case or control group as a binary classification task. For each dataset, we first identify a set of optional case patients according to the medical diagnosis guidelines, and then domain experts help us confirm whether the patients suffer these diseases. Finally, a set of group matched controls is collected according to patient demographics and clinical characteristics. For each case patient, we denote the date of disease confirmation, i.e., the operation criterion date, then track back from this date, hold off the visits within the prediction window (270 days), and finally use the remaining visits before the prediction window as the patient's input data. For each control patient, we hold off the last one year's visits and use the remaining visits as the input data. We remove the ICD-9 codes which appear less than 5 times in the datasets, and exclude patients who made less than 5 visits.

Table 1: Statistics of Datasets.

Dataset	Heart Failure	COPD	Kidney Disease
# of cases	2,403	4,807	3,201
# of controls	5,168	11,487	7,020
# of visits	247,792	518,996	345,676
Avg. # of visits per patient	32.73	31.85	33.82
# of unique ICD-9 codes	4,130	5,132	4,714
Avg. # of codes per visit	2.83	2.71	2.81

5.2 Experimental Setup

In this subsection, we first describe the traditional and state-of-the-art approaches for risk prediction which are used as baselines, and then introduce the implementation details. Finally, we outline the measures used for evaluation.

Baseline Approaches

To validate the performance of the proposed framework for risk prediction task, we implement the following methods:

• *Traditional classification approaches.* We compare the proposed PRIME with logistic regression (LR), support vector machine (SVM) and random forest (RF). The input data is the frequency of all the diagnosis codes appeared in all the visits. When implementing these approaches with *scikit-learn*[7], we follow the same setting as mentioned in previous work [2, 3].

• *Deep learning approaches.* We use four deep learning models as baselines, including two recurrent neural networks (GRU [6] and LSTM [15]), and two prediction approaches: RETAIN [9] and CNN [2, 3, 5]. For GRU, LSTM and RETAIN, we set $k = 256$ in Eq. (1) and the hidden size as 256. For CNN, we set the size of filter windows (l) from 2 to 5 with $s = 100$ filter maps. We also use regularization (l_2 norm with the coefficient 0.001) and drop-out strategies (the drop-out rate is 0.5) for all the approaches.

The Proposed Approaches

PRIME is the proposed framework for risk prediction, which integrates prior medical knowledge with posterior regularization technique. PRIME$_r$ and PRIME$_c$ are two implementations of PRIME, which use LSTM and CNN as the basic predictive model respectively. The settings of PRIME$_r$ and PRIME$_c$ are the same as those of LSTM and CNN. Besides, we set $\alpha = \beta = 0.01$ for PRIME$_r$, and $\alpha = 0.01$ and $\beta = 0.1$ for PRIME$_c$. PRIME$_r-$ and PRIME$_c-$ have

[7] http://scikit-learn.org/stable/

the same settings with PRIME$_r$ and PRIME$_c$ except for the final prediction step. PRIME$_r-$ and PRIME$_c-$ use Eq. (7), but PRIME$_r$ and PRIME$_c$ apply Eq. (8) for the risk prediction.

Details of Designing Constraint Features
To clearly show the details of designing constraint features for each prediction task, we first list all the underlying diseases used for the three prediction tasks in Table 2. Next, we introduce how to calculate the constraint features: underlying disease and disease duration. For each kind of underlying diseases, if one of the diagnosis codes appears in the patients' visits, then the counter of this disease adds 1. The duration of each underlying disease is calculated from the first appeared date to the end and measured by months. If the frequency of underlying diseases is smaller than 3, then we set it as well as its duration as 0 in our experiments.

Table 2: Diagnosis Codes (ICD9) of Underlying Diseases. "$*$" means that all the codes in this diagnosis group are included.

Disease	ICD-9 Codes
High Blood Pressure	401, 401.0, 401.1, 401.9, 402.0, 402.00, 402.1, 402.10, 402.9, 402.90
Coronary Artery Disease	414.00, 414.01, 414.0
Diabetes	250.$*$
Congenital Heart Defects	V13.65
Valvular Heart Disease	424.0
Alcohol Use	305.0, 305.00, 305.01, 305.02, 305.03
Smoking	305.1, V15.82, E869.4
Obesity	278, 278.0, 278.00, 278.01, 278.02, 278.03
Asthma	493.$*$
Abnormal Kidney Structure	794.4
Exposure to dusts & Chemicals	V87.2

The constraint features used in **heart failure** prediction task include *age*, *underlying diseases* and *their durations*. The underlying disease set $\mathcal{U}$ consists of *high blood pressure*, *coronary artery disease*, *diabetes*, *congenital heart defects*, *valvular heart disease*, *alcohol use*, *smoking* and *obesity*. The constraint features for predicting the risk of **COPD** are *age*, *genetics* (the diagnosis code 273.4, i.e., $\mathcal{G} = \{273.4\}$), *underlying diseases* and *durations*. The underlying diseases include *smoking*, *asthma* and *exposure to dusts and chemicals*. For the task of **kidney disease** prediction, we use *age*, *ethnicity*, *diseases of family history*, *underlying diseases* and *their durations*. Specifically, ethnicity set $\mathcal{E}$ includes *African-American*, *Native American* and *Asian-American*. The diagnosis codes about family history (i.e., $\mathcal{H}$) are V18.6, V18.61, V18.69. The underlying diseases are *high blood pressure*, *diabetes*, *smoking*, *obesity* and *abnormal kidney structure*.

Implementation Details & Evaluation Strategies
We implement all the deep learning baselines and the proposed framework PRIME with PyTorch 0.2.0. For training models, we use Adadelta [30] with a mini-batch size of 50. We randomly divide the datasets into the training, validation and testing set in a 0.75:0.10:0.15 ratio. The validation set is used to select the best values of parameters. We repeat all the approaches 10 times and report the average performance.

We use *F1 Score*, *Accuracy*, and the area under the receiver operating characteristic curve (*AUROC*) as measures for comparing the performance of all the methods in three risk prediction tasks.

5.3 Performance Evaluation

Table 3 shows the performance of all the approaches on all the three real world medical datasets. We can observe that the proposed approaches achieve the best performance compared with all the baselines in terms of the values of all the measures.

On the Heart Failure dataset, the overall performance of traditional approaches LR, RF and SVM is worse than that of the deep learning based approaches. This illustrates that employing deep learning techniques to model the high dimensional and sparse EHR data is effective for risk prediction task. In the four deep learning based baselines, GRU and LSTM perform better than RETAIN and CNN. Since RETAIN applies attention mechanisms, training RETAIN needs abundant EHR data. The size of the Heart Failure dataset is relatively small, and thus the performance of RETAIN is worse than that of GRU and LSTM. The advantage of CNN is to capture the local temporal important features. However, heart failure is a chronic disease, which needs to capture the longtime characteristics of disease evolution. RNN based models can correctly recognize these features on the Heart Failure dataset, which leads to better performance compared with CNN.

For the proposed four approaches, PRIME$_r$ achieves the best performance. We can observe that the performance of both PRIME$_r$ and PRIME$_r-$ is better than that of the basic predictive model LSTM. Similarly, the values of all the measures on both PRIME$_c$ and PRIME$_c-$ are higher than those on CNN. These observations strongly confirm that prior medical knowledge can help the predictive models to improve the performance.

On the COPD dataset, the performance of RETAIN is better than that of GRU and LSTM, which shows that the attention mechanism starts to work. Among all the baselines, the performance of CNN is the best. Even for the proposed PRIME$_r$ and PRIME$_r-$, the values on all the measures are smaller than those of CNN. The reason is that unlike some diseases, COPD has a clear cause, which is directly related to cigarette smoking. CNN has superior ability to capture these local important features, i.e., the diagnosis codes about smoking in visits. Thus, it achieves better performance compared with other approaches. However, after integrating prior medical knowledge using posterior regularization, i.e., the proposed approach PRIME$_c$ significantly improves over CNN. This again confirms that taking prior medical knowledge into account is effective for risk prediction task.

Since the characteristics of patients suffering kidney disease are very clear, the traditional classification approach RF can achieve comparable performance with deep learning based ones. Even on the simple dataset, incorporating prior medical knowledge can still improve the predictive performance. On the Kidney Disease dataset, we also observe that the performance of the basic model LSTM is comparable with that of the proposed PRIME$_r$. This is because we do not tune the best values of the hyper-parameters α and β. These two parameters are sightly sensitive to the dataset. Nevertheless,

Table 3: Performance on the Three Real World Medical Datasets.

Model		Heart Failure			COPD			Kidney Disease		
		AUROC	F1 Score	Accuracy	AUROC	F1 Score	Accuracy	AUROC	F1 Score	Accuracy
Traditional Classification	LR	0.8810	0.8383	0.9048	0.8940	0.8559	0.9206	0.9147	0.8922	0.9335
	RF	0.8755	0.8444	0.9137	0.8801	0.8478	0.9202	0.9235	0.9145	0.9491
	SVM	0.8424	0.7734	0.8590	0.8400	0.7711	0.8715	0.8940	0.8545	0.9067
Deep Learning	GRU	0.9047	0.8854	0.9357	0.9014	0.8772	0.9349	0.9263	0.9146	0.9485
	RETAIN	0.8913	0.8661	0.9251	0.9110	0.8925	0.9431	0.9225	0.9133	0.9485
	LSTM	0.9034	0.8827	0.9339	0.9041	0.8812	0.9370	0.9267	0.9164	0.9498
	CNN	0.8994	0.8712	0.9260	0.9181	0.8968	0.9444	0.9284	0.9161	0.9491
This Work	PRIME$_r-$	0.9059	0.8881	0.9374	0.9048	0.8859	0.9399	0.9258	0.9107	0.9455
	PRIME$_c-$	0.8944	0.8709	0.9278	0.9204	0.9005	0.9464	0.9331	0.9201	0.9511
	PRIME$_r$	**0.9126**	**0.8955**	**0.9410**	0.9052	0.8868	0.9403	0.9276	0.9118	0.9459
	PRIME$_c$	0.9070	0.8788	0.9295	**0.9211**	**0.9014**	**0.9468**	**0.9362**	**0.9236**	**0.9530**

the proposed PRIME$_c$ outperforms other approaches on the Kidney Disease dataset.

From Table 3, we can safely conclude that integrating prior medical knowledge into existing risk prediction model can help it improve the predictive performance. Moreover, utilizing posterior regularization technique to model the prior medical knowledge with risk prediction approach is effective and reasonable.

Table 4: Statistics of Constraint Features on Three Datasets.

Group	Heart Failure		COPD		Kidney Disease	
	$= 0$	> 0	$= 0$	> 0	$= 0$	> 0
Case	424	1,979	3,033	1,774	588	2,613
Control	3,649	1,519	11,023	464	5,076	1,944
Sum	4,073	3,498	14,056	2,238	5,664	4,557

5.4 Importance of Constraint Features

The main contribution of this work is to introduce prior medical knowledge into the predictive model. To model the prior knowledge, posterior regularization technique is applied. The challenge of posterior regularization is how to design constraint features, which is introduced in Section 4.2. Next, we conduct experiments on the constraint features to illustrate the reasonableness of the proposed framework. We count four numbers in Table 4: The number of patients in case/control group with $sum(\mathbf{u}_p) = 0$ and $sum(\mathbf{u}_p) > 0$. We can observe that more than 50% patients have no constraint features on underlying diseases and durations. Especially on the COPD dataset, there are 86.3% patients without underlying diseases in the constraint feature set (i.e., $sum(\mathbf{u}_p) = 0$). Thus, we cannot directly use constraint features to predict the labels of patients. However, even only using a small part of patients with constraint features (i.e., $sum(\mathbf{u}_p) > 0$), the proposed framework can learn better model parameters and achieve better performance compared with the basic predictive models. This also can be observed from Table 3. Thus, designing constraint features for risk prediction task is necessary.

5.5 Constraint Feature Analysis

The advantage of the proposed PRIME is to automatically learn the weights for different risk factors and constraint feature categories. Next, we quantitatively show the weights learned by the proposed framework and qualitatively illustrate the reasonableness of the learned weights.

Confidence of Feature Categories. Figure 2 shows the normalized confidence scores $\mathbf{\Gamma}$ learned by PRIME$_r$ on the Heart Failure dataset, where the normalizer is Softmax function. We can observe that the six weights are different, and the weights on the risk prediction are higher than those on the non-risk prediction. From Eq. (6), we can observe that only according to the confidence matrix $\mathbf{\Gamma}$, the proposed model PRIME$_r$ cannot determine the labels of patients. This is because they are also related to the weights on the constraint features.

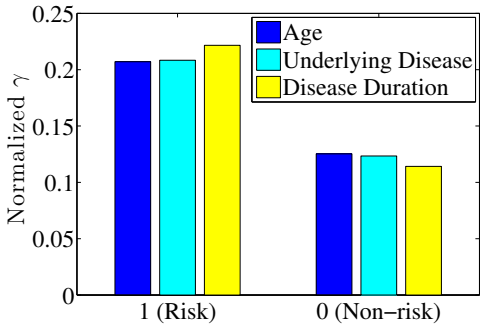
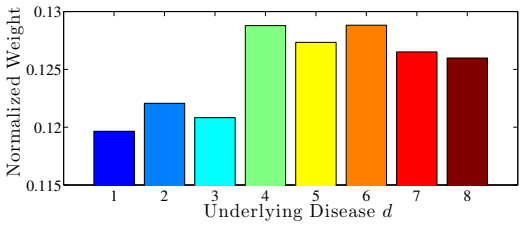
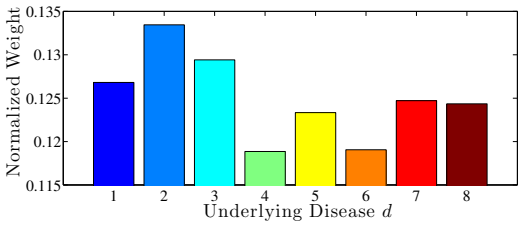


Figure 2: Confidence Matrix Learned by PRIME$_r$ on the Heart Failure Dataset.

Weights of Constraint Features. Figures 3 and 4 show the weights learned by the proposed framework PRIME$_r$ on the Heart Failure dataset for the constraint features: underlying diseases and disease duration respectively. From Figure 3(a), we can observe that for the prediction of case patients, *congenital heart defects*, *valvular heart disease*, *alcohol use* play important roles for the case patients'

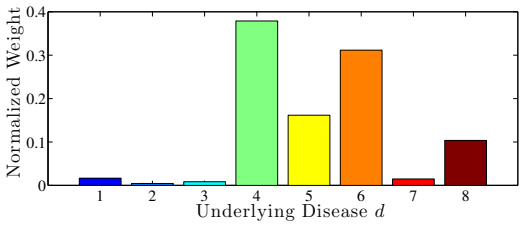


(a) $\mathbf{w}_1^{(u)}$ (Risk)

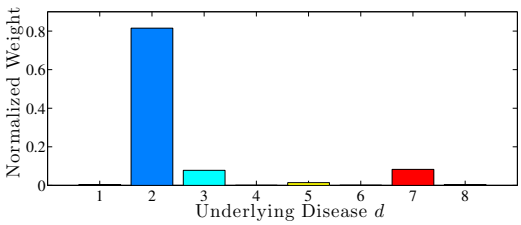


(b) $\mathbf{w}_0^{(u)}$ (Non-risk)

Figure 3: Learned Weights by PRIME$_r$ for Underlying Diseases on the Heart Failure Dataset. X-axis represents different underlying diseases, which are in the order of 1-*high blood pressure*, 2-*coronary artery disease*, 3-*diabetes*, 4-*congenital heart defects*, 5-*valvular heart disease*, 6-*alcohol use*, 7-*smoking* and 8-*obesity*. Since the values of the learned weights may be negative, we use softmax function to normalize the weight vector. Y-axis represents the normalized weights.



(a) $\mathbf{w}_1^{(d)}$ (Risk)



(b) $\mathbf{w}_0^{(d)}$ (Non-risk)

Figure 4: Learned Weights by PRIME$_r$ for Disease Duration on the Heart Failure Dataset.

prediction. Congenital heart defect[8] is one or more abnormalities in the heart's structure that the patients are born with. One of complications of congenital heart defects is heart failure. Valvular heart disease[9] may cause heart failure when one or more of the valves do not open or close properly. Some studies [11] have been shown that heavy drinking increases the risk of heart failure.

Figure 3(b) shows the weights of underlying diseases on the control patients. The weight of *high blood pressure*, *coronary artery disease* and *diabetes* is much higher than that of other risk factors. It dose not mean that these three factors are not the risk factors for the prediction of heart failure disease. The reason is that when constructing the control patients for cases, we consider patients' underlying diseases. Since these three diseases are common ones, they all frequently appear in the visits of both case and control patients.

For the learned weights for disease duration shown in Figure 4, the overall trends are similar with those estimated for underlying diseases. These two figures demonstrate that the proposed framework PRIME can learn different weights for different risk factors according to the characteristics of input data. In this way, the proposed framework PRIME successfully tackles the drawback of existing posterior regularization models [12, 17]. Due to the limitation of space, we do not show the weights on the COPD and Kidney dataset as the patients are similar with that in the weights exhibited on the Heart Failure dataset.

[8]https://www.mayoclinic.org/diseases-conditions/adult-congenital-heart-disease/symptoms-causes/syc-20355456

[9]https://www.mayoclinic.org/diseases-conditions/heart-valve-disease/symptoms-causes/syc-20353727

6 DISCUSSIONS

This paper presents PRIME, a deep learning based framework for risk prediction task. The proposed PRIME automatically incorporates discrete medical knowledge or rules into deep prediction models using posterior regularization. With such a design, the proposed framework achieves more accurate prediction results than the state-of-the-art baselines.

It is worth mentioning that we do not explicitly perform any missing value imputation for the input EHR data. Imputing EHR data is challenging as EHR data are not missing at random (NMAR) [18, 19, 23]. The proposed PRIME does not explicitly solve the problem of missing values. However, it does implicitly reduce the impact brought by missing values by employing the dropout technique, which is essentially equivalent to the random remove of some visits or codes. Thus, the proposed framework is more robust to missing visits.

The limitation of this work is that the proposed PRIME is only effective for common diseases. For rare and emerging diseases, since there is little medical knowledge about them, it is hard to incorporate any prior knowledge into deep learning predictive models. Thus, the proposed PRIME may achieve similar performance to the state-of-the-art baselines. In our future work, we will focus on how to improve predictive performance of risk prediction for rare diseases.

7 CONCLUSIONS

In this paper, we propose a general risk prediction framework PRIME, which can integrate prior medical knowledge into all the existing predictive models to improve the predictive performance. Specifically, we employ two state-of-the-art deep learning architectures–recurrent neural networks (RNN) and convolutional neural networks (CNN)–as the basic predictive models. To model the discrete and heterogeneous prior medical knowledge, posterior regularization technique is used. However, different from existing posterior regularization, we use a log-linear model to estimate the desired distributions of diseases. The benefit of the proposed approach is that it can automatically learn the weights for different prior medical knowledge. We validate the proposed framework on three real medical datasets. Experimental results show that the proposed PRIME outperforms existing risk prediction models. Finally, we qualitatively analyze the reasonableness of the weights learned by the proposed PRIME.

ACKNOWLEDGMENTS

The authors gratefully thank **Ran Huo** who is an MD candidate from Southern Medical University for helpful discussions. The authors would like to thank the anonymous referees for their valuable comments and suggestions, and NVIDIA Corporation with the donation of the Titan Xp GPU. This work is supported in part by the US National Science Foundation under grants IIS-1553411, IIS-1747614, IIS-1218393 and IIS-1514204. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient Subtyping via Time-Aware LSTM Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. 65–74.

[2] Zhengping Che, Yu Cheng, Zhaonan Sun, and Yan Liu. 2016. Exploiting Convolutional Neural Network for Risk Prediction with Medical Feature Embedding. In *Proceedings of NIPS Workshop on Machine Learning for Health (NIPS-ML4HC'16)*.

[3] Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. 2017. Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'17)*. 787–792.

[4] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep Computational Phenotyping. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. 507–516.

[5] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM'16)*. 432–440.

[6] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-decoder Approaches. *arXiv preprint arXiv:1409.1259* (2014).

[7] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. 1495–1504.

[8] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: Graph-based Attention Model for Healthcare Representation Learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. 787–795.

[9] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An Interpretable Predictive model for Healthcare Using Reverse Time Attention Mechanism. In *Proceedings of Advances in Neural Information Processing Systems (NIPS'16)*. 3504–3512.

[10] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) From Scratch. *Journal of Machine Learning Research (JMLR)* 12, Aug (2011), 2493–2537.

[11] Luc Djoussé and J Michael Gaziano. 2008. Alcohol consumption and heart failure: a systematic review. *Current atherosclerosis reports* 10, 2 (2008), 117–120.

[12] Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. 2010. Posterior Regularization for Structured Latent Variable Models. *Journal of Machine Learning Research (JMLR)* 11, Jul (2010), 2001–2049.

[13] Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. 2014. Limestone: High-throughput Candidate Phenotype Generation via Tensor Factorization. *Journal of Biomedical Informatics* 52 (2014), 199–211.

[14] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. 2014. Marble: High-throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Dining (KDD'14)*. 115–124.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9, 8 (1997), 1735–1780.

[16] George Hripcsak and David J Albers. 2012. Next-generation Phenotyping of Electronic Health Records. *Journal of the American Medical Informatics Association (JAMIA)* 20, 1 (2012), 117–121.

[17] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing Deep Neural Networks with Logic Rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*. 2410–2420.

[18] Jau-Huei Lin and Peter J Haug. 2008. Exploiting Missing Clinical Data in Bayesian Network Modeling for Predicting Medical Problems. *Journal of Biomedical Informatics* 41, 1 (2008), 1–14.

[19] Roderick JA Little and Donald B Rubin. 2014. *Statistical Analysis with Missing Data*. Vol. 333. John Wiley & Sons.

[20] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. 1903–1911.

[21] Fenglong Ma, Chuishi Meng, Houping Xiao, Qi Li, Jing Gao, Lu Su, and Aidong Zhang. 2017. Unsupervised Discovery of Drug Side-effects from Heterogeneous Data Sources. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. ACM, 967–976.

[22] Franz Josef Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*. 295–302.

[23] Rimma Pivovarov, David J Albers, Jorge L Sepulveda, and Noémie Elhadad. 2014. Identifying and Mitigating Biases in EHR Laboratory Tests. *Journal of Biomedical Informatics* 51 (2014), 24–34.

[24] Qiuling Suo, Fenglong Ma, Giovanni Canino, Jing Gao, Aidong Zhang, Pierangelo Veltri, and Agostino Gnasso. 2017. A Multi-task Framework for Monitoring Health Conditions via Attention-based Recurrent Neural Networks. In *Proceedings of the AMIA 2017 Annual Symposium (AMIA'17)*.

[25] Qiuling Suo, Fenglong Ma, Ye Yuan, Mengdi Huai, Weida Zhong, Jing Gao, and Aidong Zhang. 2017. Personalized Disease Prediction Using A CNN-Based Similarity Learning Method. In *Proceedings of The IEEE International Conference on Bioinformatics and Biomedicine (BIBM'17)*. 811–816.

[26] Qiuling Suo, Fenglong Ma, Ye Yuan, Mengdi Huai, Weida Zhong, Jing Gao, and Aidong Zhang. 2018. Deep Patient Similarity Learning for Personalized Healthcare. *IEEE Transactions on NanoBioscience* (2018).

[27] Fei Wang, Noah Lee, Jianying Hu, Jimeng Sun, and Shahram Ebadollahi. 2012. 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 (KDD'12)*. 453–461.

[28] Xiang Wang, Fei Wang, Jianying Hu, and Robert Sorrentino. 2014. Exploring Joint Disease Risk Prediction. In *AMIA Annual Symposium Proceedings (AMIA'14)*. 1180–1187.

[29] Ye Yuan, Guangxu Xun, Fenglong Ma, Qiuling Suo, Hongfei Xue, Kebin Jia, and Aidong Zhang. 2018. A Novel Channel-aware Attention Framework for Multi-Channel EEG Seizure Detection via Multi-view Deep Learning. In *Proceedings of the 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI'18)*. IEEE, 206–209.

[30] Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012).

[31] Jiacheng Zhang, Yang Liu, Huanbo Luan, Jingfang Xu, and Maosong Sun. 2017. Prior Knowledge Integration for Neural Machine Translation Using Posterior Regularization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, Vol. 1. 1514–1523.

[32] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. 2014. From Micro to Macro: Data Driven Phenotyping by Densification of Longitudinal Electronic Medical Records. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. 135–144.