Deep Stable Representation Learning on Electronic Health Records

Yingtao Luo
Carnegie Mellon University
Pittsburgh, USA
yingtaoluo@cmu.edu

Zhaocheng Liu
Kuaishou Technology
Beijing, China
lio.h.zen@gmail.com

Qiang Liu
Chinese Academy of Sciences
Beijing, China
qiang.liu@nlpr.ia.ac.cn

***Abstract*—Deep learning models have achieved promising disease prediction performance of the Electronic Health Records (EHR) of patients. However, most models developed under the I.I.D. hypothesis fail to consider the agnostic distribution shifts, diminishing the generalization ability of deep learning models to Out-Of-Distribution (OOD) data. In this setting, spurious statistical correlations that may change in different environments will be exploited, which can cause sub-optimal performances of deep learning models. The unstable correlation between procedures and diagnoses existed in the training distribution can cause spurious correlation between historical EHR and future diagnosis. To address this problem, we propose to use a causal representation learning method called Causal Healthcare Embedding (CHE). CHE aims at eliminating the spurious statistical relationship by removing the dependencies between diagnoses and procedures. We introduce the Hilbert-Schmidt Independence Criterion (HSIC) to measure the degree of independence between the embedded diagnosis and procedure features. Based on causal view analyses, we perform the sample weighting technique to get rid of such spurious relationship for the stable learning of EHR across different environments. Moreover, our proposed CHE method can be used as a flexible plug-and-play module that can enhance existing deep learning models on EHR. Extensive experiments on two public datasets and five state-of-the-art baselines unequivocally show that CHE can improve the prediction accuracy of deep learning models on out-of-distribution data by a large margin. In addition, the interpretability study shows that CHE could successfully leverage causal structures to reflect a more reasonable contribution of historical records for predictions.**

***Index Terms*—Healthcare informatics, causal inference, electronic health records, out-of-distribution**

I. Introduction

Healthcare predictive model for healthcare disease diagnosis based on Electronic Health Records (EHR) is a key engine for improving the quality of clinical care. [1]. In the US, nearly 96% of hospitals had a digital electronic health records (EHR) in 2015 [1], which emphasizes the importance of learning EHR. The comprehensive patient information (such as demographics, diagnoses, and procedures) in EHR provides valuable assistance for personal health status tracking and monitoring [2]–[6]. To predict the future diagnoses based on a patient's historical EHR, many deep learning models [7]–[10] are proposed with promising accuracy to discover the statistical correlations in the training distribution for predictions.

Despite their great successes, the challenge of the out-of-distribution (OOD) problem has not yet been fully addressed in previous works, which may cause sub-optimal learning performance on EHR. The I.I.D. hypothesis that most models are built upon does not hold true for practical situations due to the inevitable distribution shifts such as data selection bias and confounding factors [11]–[14]. We present the spurious correlation between diagnoses and procedures as a representative example of out-of-distribution problems in electronic health records. The diagnoses and procedures are often correlated as clinicians select treatments according to the patients' current and historical diagnosis records based on medical experience and knowledge. However, the patients' demographics and insurance information may vary a lot in the training and test datasets, causing the subtle correlation between diagnoses and procedures to vary in different environments. We argue that the EHR prediction models may be misled by the subtle dependency between diagnoses and procedures, resulting in spurious correlations between historical EHR and future prediction that are unstable when confronting the OOD data in practice. As a result, the learned statistical correlations cannot guarantee to be as effective on inference as on the training dataset.

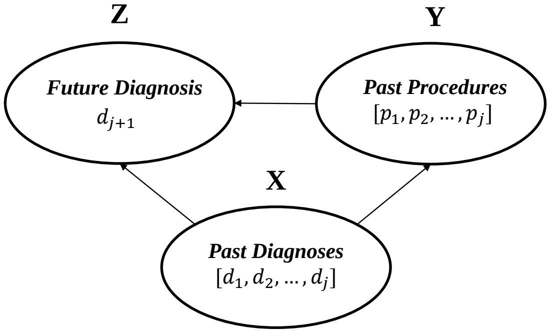

Fig. 1. The causal diagram of diagnosis prediction in EHR.

In the following, we present the causal view analyses to discuss how the correlation between diagnoses and procedures as an example can cause the spurious relationship for model prediction. As shown in Fig. 1, the causal diagram of future diagnosis prediction in EHR consists of two sequences of features, i.e. "past diagnoses" (X) and "past procedures" (Y). By the ignorability assumption in causality [15], any other potential confounders are considered uncorrelated to both diagnoses and procedures. Therefore, these potential confounders will not be discussed for a better understanding of the scheme. Here, we discuss the case where the X and Y are correlated, thus the causal effect of each feature cannot be accurately estimated by deep learning models. In EHR, each diagnosis has an impact on the current and future procedures. At the same time, both diagnoses and procedures influence the "future diagnosis" (Z). Because doctors give treatments based on the same medical knowledge, diagnoses and procedures are strongly correlated. Due to the strong correlation of $X \rightarrow Y$, it is hard for deep learning models to learn a stable relationship of $X \rightarrow Z$ and/or $Y \rightarrow Z$. As an instance, patients diagnosed with diabetes can take insulin, and diabetes may cause puffiness. With the strong correlation between diabetes and insulin, a machine learning model has a great chance to learn that insulin causes puffiness. Moreover, the correlation between the two variables may be different in various data distributions, which causes difficulty for model generalization to Out-Of-Distribution (OOD) data. For example, procedures may vary among different insurance types and only some of them can cover insulin drugs, which may result in diet control treatment or non-insulin drug treatments such as Exenatide and Liraglutide for some patients with the same diagnoses. Therefore, models trained by one type of insurances may not always generalize to new insurances.

An intuitive approach to address this problem is to remove the dependency between X and Y, so that this selection bias in the training dataset will not affect the inference phase. In our case, if the model does not exploit the correlation between diabetes and insulin in a selected training dataset for predicting puffiness, it will find that $X \rightarrow Z$ is the truly stable relationship that reflects causation. To this end, we are interested in a method that discovers stable correlations reflecting causal effect of each feature across different environments, which is free of data biases introduced by the distribution shifts between training and inference. Such a method demonstrating the ability to find a more causal model can point out that potential of future studies and deployments on various machine learning healthcare systems. Sadly, many algorithms such as domain generalization [16], causal transfer learning [17] and invariant causal prediction [18] cannot deal with distribution shifts unobserved in the training data.

To obtain a stable correlation structure between each variable and the final prediction, a strand of variable decorrelation technique [19], [20] is proposed for linear models. Its basic notion is to remove the dependencies between variables through a sample weighting method and make the correlation structure between each variable and the prediction free of the confounding factors of other variables. In Fig. 1, the arrow from each diagnosis X to each procedure Y will be removed, which leaves the causal diagram with independent X and Y to accurately estimate their contributions. While the concept of variable decorrelation is tempting for healthcare systems, how to extend it to a high-dimensional nonlinear deep learning model with sequential data can be difficult. There are two challenges to tackle. First, with nonlinear neural layers, the nonlinear correlation in deep learning healthcare cannot be measured and eliminated by linear methods. Second, the sample weighting should be redefined to accommodate the sequential data that any past diagnoses can have an impact on a future procedure along the time. It is vital to efficiently remove the dependencies of all combinations of diagnoses and procedures without excessive computational complexity.

In this paper, we propose a causal representation learning method for sequential diagnosis prediction in EHR, called Causal Healthcare Embedding (**CHE**). To address the two challenges, first, we use Hilbert Schmidt Independence Criterion (HSIC) [21], [22] that measures the norm of cross-covariance from X to Y, the degree of dependence between X and Y for feature decorrelation [23], which can align with the nonlinear neural models. By minimizing HSIC(X,Y), i.e. the degree of dependence between diagnoses and procedures, we expect that the model will get rid of the spurious relationships that are hard to generalize to OOD data. Second, as pointed out by [24], treatments can be represented by latent factors as an alternative for estimating causality. While it is computationally expensive and inaccurate to calculate the binary sample weighting for all sequential combinations of treatments [25], we apply HSIC on the two low-dimensional latent representations generated by diagnoses and procedures. By minimizing the HSIC in the loss function, the dependency between X and Y can be minimized throughout the training. Without spurious relationships caused by unstable correlations, deep learning models can exploit the causation between each feature and the prediction. Moreover, the learning of causation can improve the model generalization to different environments without necessarily knowing the training data a priori. We perform the proposed CHE on multiple baseline models for predicting the future diagnosis and show by extensive experiments that it can improve these models by large margins.

To be noted, our proposed CHE method is a plug-and-play module for the diagnosis prediction task. It can be easily and adaptively incorporated with various diagnosis prediction models. To summarize, our contributions are listed as follows:

- For the first time, we propose a causal representation learning method for healthcare diagnosis prediction, which removes the dependencies between variables such as diagnoses and procedures by sample weighting on the latent representation.
- We show that the proposed method can learn a stable correlation between each causal feature to the prediction, which makes predictions stable across different data distributions without knowing the training data a priori.
- Our proposed method is a plug-and-play module that can work well with diagnosis prediction models in a variety of scenarios. We prove that its computational complexity is identical to model training, without the concerns of estimating a exponentially growing number of treatments caused by the combinatorial nature of sequential data.
- Extensive experiments on public datasets show that our method increases both NDCG and ACC of five state-of-the-art baselines by a significant margin, especially when

applying to data of different sources and/or distributions.

II. Related Work

In this section, we review the existing works for mining the EHR data, especially the state-of-the-art models on disease diagnosis prediction. Moreover, we introduce some related works on counterfactual prediction and variable decorrelation.

A. EHR Data Mining

The mining of EHR is essential for improving the healthcare management of patients. Many tasks that aim at improving healthcare quality can be identified as EHR data mining, such as risk prediction [2], [4], [5], [26], disease progression [27]–[29], phenotyping [30]–[32], diagnosis prediction [9], [33], [34]. Owing to the sequential pattern of EHR data, Recurrent Neural Networks (RNNs) are naturally suitable, and Long Short-Term Memory (LSTM) [35] has been successfully applied. RETAIN [7] presents a reverse time attention model that preserves interpretability. Dipole [3] incorporates bidirectional RNN for making a prediction based on EHR. Camp [9] uses demography information in co-attention model for diagnosis prediction. ConCare [10] proposes to incorporate multi-head self-attention to model the sequential data of EHR. StageNet [6] integrates time intervals between visits into LSTM to model the stages of health conditions. INPREM [36] applies Bayesian neural network in an attention-based prediction model for improving the model interpretability. HiTANet [37] proposes hierarchical time-aware attention networks for health risk prediction. LSAN [29] combines both long- and short-term information in EHR to make predictions. SETOR [38] utilizes ontological representation and neural ordinary equation for diagnosis prediction. Meanwhile, multi-sourced data is also considered in recent works [39], [40]. Besides, constructed on sequential prediction models, medical knowledge graphs are modeled to provide some prior knowledge for more accurate predictions [8], [33], [41], [42].

B. Counterfactual Prediction

Counterfactual learning [43] is an important direction of research in causal inference [15], [44]. Counterfactual learning can enable people to estimate the probability of counterfactual events and eventually identify the unbiased causal relationships between events. The existing counterfactual learning approaches usually reweight samples based on propensity scores [45]–[48], which indicate the probabilities of observation under different environments. Under the binary treatment setting, balancing the sample weights in the loss function can remove confounding bias to make causal prediction [25], [49]. Recent works further extend counterfactual learning to the multi-level treatment [50] and bundle treatment [24] settings. Meanwhile, Permutation Weighting (PW) [25] conducts permutation on observed features for calculating propensity scores. These methods directly control input variables, therefore, the accurate estimation of propensity score when the number of treatments in sequential data is growing exponentially can be a challenge.

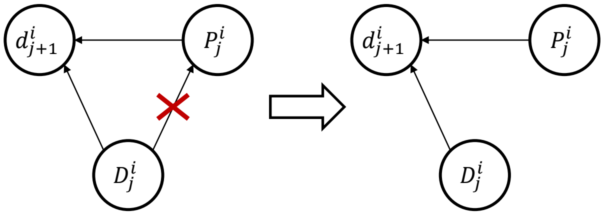

Fig. 2. Decorrelation between past diagnosis sequence D_j^i and past procedure sequence P_j^i until the j-th visit for patient v_i, for the more accurate and stable prediction of the diagnosis d_{j+1}^i of a future time $j+1$.

C. Variable Decorrelation

Recently, stable learning methods have been proposed to perform variable decorrelation for learning causal features in models from biased data. Stable learning can be viewed as another perspective of causal learning technique, in which there is no implicit treatments and the distribution of unobserved samples is unknown [19]. Existing stable learning methods are mostly investigated in linear models. Specifically, most methods conduct stable learning via decorrelation among features of samples, which tries to make the feature distribution closer to independently identically distribution [19], [20], [51]. Sample Reweighted Decorrelation Operator (SRDO) [52] generates some unobserved samples, and trains a binary classifier to get the probabilities of observation for weighting the observed samples. Recently, some works investigate to conduct stable learning on neural networks, such as on convolutional neural networks [53] and graph neural networks [54].

III. Preliminary

In this section, we first formulate the diagnosis prediction problem. Then, we discuss causality in EHR data.

A. Problem Formulation

In the EHR data, we have a set of patients $V = \{v_1, v_2, ..., v_{|V|}\}$, and patient v_i has t^i visits. Diagnoses and procedures are both represented in International Classification of Diseases, Ninth Revision (ICD-9)[1] medical codes, where we have M unique diagnosis medical codes and N unique procedure medical codes. For each patient v_i with j visits, there exists a historical diagnosis sequence $D_j^i = \left[d_1^i, d_2^i, ..., d_j^i\right]$ and a historical procedure sequence $P_j^i = \left[p_1^i, p_2^i, ..., p_j^i\right]$. Each diagnosis and procedure are M-dimensional multi-hot vector and N-dimensional multi-hot vector respectively, which means that $d_j^i \in \{0,1\}^M$ and $p_j^i \in \{0,1\}^N$, where $1 \leq j \leq t^i$. In this work, we would like to predict future diagnoses, i.e., predicting what diseases a patient will have in the future, based on historical EHR. Specifically, in this work, given D_j^i and P_j^i, we need to predict future diagnosis d_{j+1}^i. The descriptions of notations are shown in Table I.

[1] https://www.cdc.gov/nchs/icd/icd9.htm

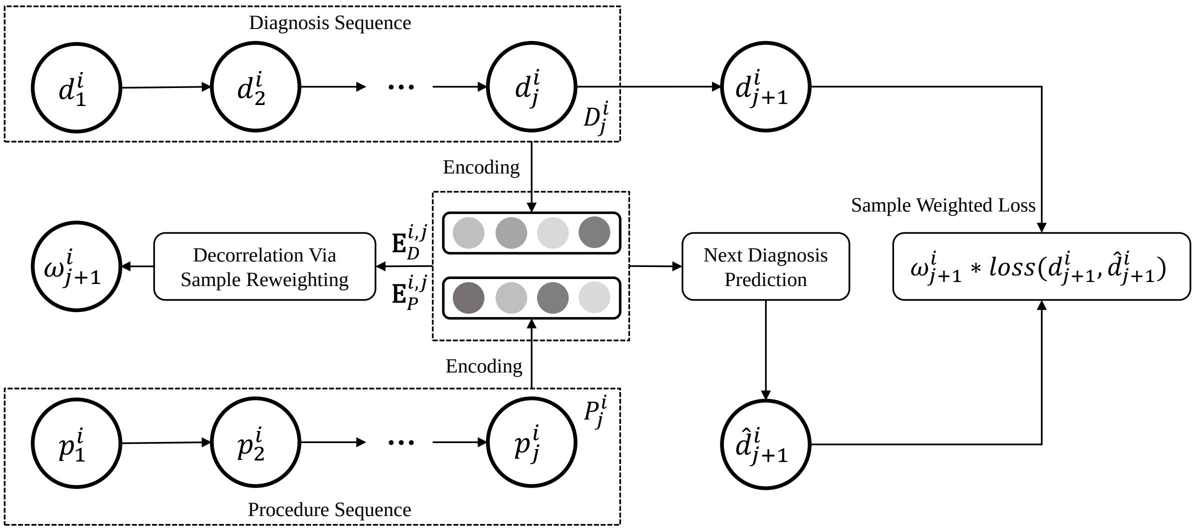

Fig. 3. The schematic diagram of decorrelation between past diagnoses D_j^i and past procedures P_j^i via sample weighting.

TABLE I
NOTATIONS AND DESCRIPTIONS

Notation	Description
V	The set of patients in the EHR data
v_i	The i-th patient in the EHR data
t^i	The number of visits of the i-th patient
D_j^i	The diagnoses of the i-th patient until the j-th visit
d_j^i	The diagnosis of the i-th patient at the j-th visit
P_j^i	The procedures of the i-th patient until the j-th visit
p_j^i	The procedure of the i-th patient at the j-th visit
$\mathbf{E}_D^{i,j}$	The embedding of the historical diagnoses D_j^i
$\mathbf{E}_d^{i,j}$	The embedding of the diagnosis d_j^i
$\mathbf{E}_P^{i,j}$	The embedding of the historical procedures D_j^i
$\mathbf{E}_p^{i,j}$	The embedding of the procedure p_j^i
$\mathbf{W}$	The sample weights for all patients
ω_j^i	The sample weight of the i-th patient at the j-th visit
r	The dimensionality of the embedding space

B. Causal View Analysis

Several causal links in medical knowledge affect the causal relationships in EHR. First, historical diagnoses can affect future procedures, in which doctors select procedures according to the patients' historical data based on medical experience. Second, both historical diagnoses and procedures can have impacts on future diagnoses, since 1) diseases have development processes that are accompanied by complications and 2) procedures affect the development and rehabilitation of diseases. Consequently, these relationships lead to a strong correlation between the two sequences of historical diagnoses D_j^i and procedures P_j^i, which we denote as $D_j^i \rightarrow P_j^i$. Both diagnoses and procedures can have causal effects on future diagnosis, i.e. $D_j^i \rightarrow d_{j+1}^i$ and $P_j^i \rightarrow d_{j+1}^i$, as shown in Fig. 1. Causal inference theory [15], [44] requires that the conditional covariance between "Treatment" and "Covariates" to be zero. In our case, both diagnoses and procedures can be treatment(s) and/or covariates, so we ensure the conditional covariance of $D_j^i \rightarrow P_j^i$ to be zero.

The strong correlation between diagnoses and procedures hinders machine learning models from learning causal relationships for future diagnosis. Without the ability to remove $D_j^i \rightarrow P_j^i$, machine learning models may only rely on D_j^i or P_j^i to predict d_{j+1}^i depending on the training data distribution. For example, patients with diabetes usually take insulin to cure the disease. Diabetes may cause some complications, such as retinopathy and cataract, even when insulin is taken. If diabetes and insulin frequently occur together, it is hard for deep learning models to identify whether diabetes or insulin causes complications. If the model weighs on the occurrence of insulin for predicting complications, it cannot generalize to scenarios when diabetes and insulin do not occur together that often. On the other hand, if the correlation between D_j^i and P_j^i is eliminated, the model can estimate the importance of diabetes and insulin for future diagnoses independently, as shown in Fig. 2. As changing insulin to other prescriptions neither influence the distribution of diabetes (the correlation is eliminated) nor the distribution of complications (the correlation does not exist in medicine), the model will rely solely on diabetes to predict possible complications.

C. The Effect of Correlation in Embedding Space

The correlation is easy to understand in linear models, considering that multicollinearity problem has been widely studied in linear models [55]. Diagnosis prediction models are mostly based on deep learning approaches, thus the discussion should be done in the embedding space. We first encode diagnoses and procedures into an embedding space by any

EHR encoders

$$\mathbf{E}_D^{i,j} = Encoder\left(D_j^i\right), \tag{1}$$

$$\mathbf{E}_P^{i,j} = Encoder\left(P_j^i\right), \tag{2}$$

For diagnosis prediction, we can learn a deep learning model $f(\cdot)$ that satisfies

$$d_{j+1}^i = f\left(\mathbf{E}_D^{i,j}, \mathbf{E}_P^{i,j}\right). \tag{3}$$

Considering the strong correlation between D_j^i and P_j^i, we can predict one of their embeddings based on another as

$$\mathbf{E}_D^{i,j} = g_{P\to D}\left(\mathbf{E}_P^{i,j}\right) + \varepsilon_{P\to D}, \tag{4}$$

$$\mathbf{E}_P^{i,j} = g_{D\to P}\left(\mathbf{E}_D^{i,j}\right) + \varepsilon_{D\to P}, \tag{5}$$

where $g_{P\to D}(\cdot)$ and $g_{D\to P}(\cdot)$ are deep learning models, $\varepsilon_{P\to D}$ and $\varepsilon_{D\to P}$ are residual errors. $g_{P\to D}(\cdot)$ and $g_{D\to P}(\cdot)$ are both deep learning models, so their parameters can be merged with $f(\cdot)$. Therefore, Eq. (3) can rewritten as

$$d_{j+1}^i = f\left(\varepsilon_{P\to D}, \mathbf{E}_P^{i,j}\right), \tag{6}$$

or

$$d_{j+1}^i = f\left(\mathbf{E}_D^{i,j}, \varepsilon_{D\to P}\right). \tag{7}$$

Due to the strong correlation between D_j^i and P_j^i, the model can rely on either $\mathbf{E}_D^{i,j}$ or $\mathbf{E}_P^{i,j}$ to make diagnosis prediction, which is unstable when the correlation changes. Thus, we need to decorrelate past diagnoses and past procedures to obtain $g_{P\to D}\left(\mathbf{E}_P^{i,j}\right) = 0$ and $g_{D\to P}\left(\mathbf{E}_D^{i,j}\right) = 0$. Accordingly, as illustrated in Fig. 2, we plan to remove $\mathbf{E}_D^{i,j} \to \mathbf{E}_P^{i,j}$ to learn more accurate, stable and causal diagnosis prediction models. Then, the contribution of D_j^i and P_j^i to predicting d_{j+1}^i can be free of the interference of unstable $D_j^i \to P_j^i$.

IV. Methodology

In this section, we introduce the sample weighting method with independence testing statistics to conduct causal disease diagnosis prediction, which is a plug-and-play module that aligns well with deep learning models.

A. Hilbert Schmidt Independence Criterion

The removal of dependencies between $\mathbf{E}_D^{i,j}$ and $\mathbf{E}_P^{i,j}$ is at the core of sample weighting. To measure the dependency for optimization, we introduce HSIC [21], [22], an independence testing statistics as the Hilbert-Schmidt norm of the cross-covariance operator between the distributions in Reproducing Kernel Hilbert Space (RKHS). Consider the measurable, positive definite kernel k of variables and the corresponding RKHS H [56]. For all $h_D \in H_D, h_P \in H_P$, the cross-covariance operator Σ_{DP} from H_D to H_P is:

$$\begin{aligned}\langle h_D, \Sigma_{DP} h_P\rangle = \mathbb{E}_{DP}[h_D(\mathbf{E}_D^{i,j}) h_P(\mathbf{E}_P^{i,j})] \\ -\mathbb{E}_D[h_D(\mathbf{E}_D^{i,j})]\mathbb{E}_P[h_P(\mathbf{E}_P^{i,j})].\end{aligned} \tag{8}$$

As proved by [53], if the product of k_D and k_P is characteristic, $\mathbb{E}[k_D(\mathbf{E}_D^{i,j}, \mathbf{E}_D^{i,j})] < \infty$ and $\mathbb{E}[k_P(\mathbf{E}_P^{i,j}, \mathbf{E}_P^{i,j})] < \infty$, we have

$$\Sigma_{DP} = 0 \Longleftrightarrow D \perp P, \tag{9}$$

which means that if $\mathbf{E}_D^{i,j}$ cannot be transformed into $\mathbf{E}_P^{i,j}$ via a nonlinear operator, the two variables are independent.

The squared Hilbert-Schmidt norm of the cross-covariance operator Σ_{DP} can be approximated by the unbiased calculation in the embedding space as

$$\text{HSIC}(\mathbf{E}_D, \mathbf{E}_P) = \frac{1}{|V|(t^i - 1)} \sum_{i=1}^{|V|} \sum_{j=1}^{t^i - 1} \text{HSIC}_{local}(\mathbf{E}_D^{i,j}, \mathbf{E}_P^{i,j}). \tag{10}$$

Specifically, if r denotes the hidden dimensionality, we can calculate the HSIC of $\mathbf{E}_D^{i,j} \in \mathbb{R}^r$ and $\mathbf{E}_P^{i,j} \in \mathbb{R}^r$ by

$$\text{HSIC}_{local}(\mathbf{E}_D^{i,j}, \mathbf{E}_P^{i,j}) = \frac{1}{(r-1)^2} Tr(K_d J K_p J), \tag{11}$$

where Tr is the trace of a matrix, $J = I - 1/r$ with I as an r-order identity matrix. K_D and K_P are any kernel matrices. We can consider RBF kernel to calculate

$$K_d(x_1, x_2) = \exp(-\frac{\|x_1 - x_2\|_2^2}{\sigma^2}), \tag{12}$$

where $x_1, x_2 \in \mathbf{E}_D^{i,j}$ represent the values in different dimensions of the latent representation. Therefore, $x_q \in \mathbb{R}^1$ for $\forall q \in [1, ..., r]$. Similarly, there is $K_p(x_1, x_2)$ where $x_1, x_2 \in \mathbf{E}_P^{i,j}$ represent different dimensions of the latent representation. The kernel tricks of $K_d, K_p \in \mathbb{R}^{r\times r}$ can approximately calculate HSIC rapidly. In this way, for each time's visit by each patient, the cross-covariance in the embedding space from diagnoses to procedures can be measured by HSIC.

B. Loss Functions

Inspired by feature decorrelation techniques [51], [52], we propose to minimize HSIC by the sample weighting technique to mitigate the dependency between diagnoses and procedures in the embedded space. We use $\mathbf{W}$ to denote sample weights, where the weight for patient i at the j-th visit is denoted as ω_j^i. We denote the weighted samples as $\mathbf{WE}_D$ and $\mathbf{WE}_P$, where the weighted samples of patient i at the j-th visit are denoted as $\omega_j^i \mathbf{E}_D^{i,j}$ and $\omega_j^i \mathbf{E}_P^{i,j}$.

To minimize the correlation between diagnoses and procedures, we propose to optimize $\mathbf{W}$ with HSIC as follows

$$\mathbf{W}^* = \operatorname*{argmin}_{\mathbf{W}} \text{HSIC}(\mathbf{WE}_D, \mathbf{WE}_P). \tag{13}$$

Meanwhile, we define the cross-entropy loss for the classification task of the diagnosis prediction as

$$Enc, Prd = \operatorname*{argmin}_{Enc, Prd} \sum_{i=1}^{|V|} \sum_{j=1}^{t^i - 1} \omega_j^i(n) \mathbf{L}_j^i, \tag{14}$$

where

$$\mathbf{L}_j^i = L(Prd(Enc(D_j^i), Enc(P_j^i)), d_{j+1}^i), \tag{15}$$

where L denotes the cross-entropy loss function. Enc represents the encoder that maps diagnoses and procedures into the embedding space. Prd represents the final prediction layer that maps the latent representation into the one-hot probability vector. The architectures of Enc and Prd depend on the base model our method is used upon. L is also based on the specific loss function used in the base model.

C. Model Optimization

We iteratively optimize the weighted loss and the HSIC by

$$Enc_{n+1}, Prd_{n+1} = \operatorname*{argmin}_{Enc,Prd} \sum_{i=1}^{|V|} \sum_{j=1}^{t^i-1} \omega_j^i(n)\mathbf{L}_j^i, \tag{16}$$

and

$$\mathbf{W}(n+1) = \operatorname*{argmin}_{\mathbf{W}} \epsilon \cdot \mathrm{HSIC}(\mathrm{Diag}, \mathrm{Proc}), \tag{17}$$

where

$$\mathrm{Diag} = \mathbf{W}(n)Enc_{n+1}(D_j^i), \mathrm{Proc} = \mathbf{W}(n)Enc_{n+1}(P_j^i). \tag{18}$$

Enc_n, Prd_n and $\mathbf{W}(n)$ indicates encoder, final prediction layer and sample weights at the n-th iteration, and $\mathbf{W}(0)$ is initially set as ones. ϵ is a coefficient that balances the learning rates for updating the neural network and sample weights.

Eq. (16) and Eq. (17) are optimized iteratively, meaning that we first optimize the neural network and then optimize the HSIC for each iteration. Every two subsequences $\mathbf{E}_D^{i,j}$ and $\mathbf{E}_P^{i,j}$ of length j are fed into a neural network to calculate the cross-entropy loss, and sample weights are multiplied to the loss to update the model parameters. Then, we use the updated model to calculate the HSIC of $\mathbf{E}_D^{i,j}$ and $\mathbf{E}_P^{i,j}$ obtained by the encoder part of the model. The sample weighting reassigns the importance of each sample when calculating the loss function to remove the dependency between features.

To make the overall method presented in Section IV clearer, we show the pseudo-code of training the CHE method in Algorithm 1.

D. Complexity Analysis

The time complexity of calculating HSIC only grows with the hidden dimensionality r. By naive algorithms, the multiplication of K_d and K_p is $O(r^3)$, and the calculation of trace is also $O(r^3)$. For deep learning models, r is a hyperparameter and is thus trivial. The calculation of overall HSIC is $O(|V|t)$ if we denote $t = max(t^i)$ for $\forall i$, which is linearly proportional to the number of visits in the data. This is acceptably efficient. On the other hand, the number of treatments for each timestamp is M, the number of unique ICD-9 codes for diagnosis. Considering the combination of ICD-9 codes in a sequence, the total number of treatments can be as many as M^t, which makes the traditional counterfactual weighting to estimate propensity scores very expensive.

V. Experiments

In this section, we conduct extensive experiments to verify the effectiveness of our proposed CHE method.

Algorithm 1 The training process of CHE.

```
Require: Set of patients $V = \{v_1, v_2, ..., v_{|V|}\}$, each $v_i$'s
    diagnosis sequence $D_{j^i}^i = \left[d_1^i, d_2^i, ..., d_{j^i}^i\right]$ and procedure
    sequence $P_{j^i}^i = \left[p_1^i, p_2^i, ..., p_{j^i}^i\right]$, maximum epoch $Epoch$,
    and any BaseModel.
Ensure: Model parameters $Enc()$ and $Prd()$ in the Base-
    Model.
Initialize epoch indicator $n \leftarrow 0$.
Initialize best epoch indicator $n_{best} \leftarrow 0$.
Initialize $Enc_0()$ and $Prd_0()$ randomly.
Initialize sample weights $\mathbf{W}$. For every $\omega_j^i \in \mathbf{W}$, $\omega_j^i(0) \leftarrow$
    $1$ for $1 \le i \le |V|$ and $1 \le j \le t^i - 1$.
while early-stopping not reached and $n < Epoch$. do
    Update model parameters $Enc_{n+1}()$ and $Prd_{n+1}()$
        according to Eq. (18), while keeping $\mathbf{W}(n)$ fixed.
    Update sample weights $\mathbf{W}(n+1)$ according to Eq. (20),
        while keeping $Enc_{n+1}()$ and $Prd_{n+1}()$ fixed.
    $n \leftarrow n + 1$.
    Update $n_{best} \leftarrow n$, if better result is achieved on the
        validation set.
end while
return $Enc_{n_{best}}()$ and $Prd_{n_{best}}()$.
```

A. Datasets

We evaluate our proposed sequential stable learning method on two real-world datasets: MIMIC-III and MIMIC-IV.

- **MIMIC-III Dataset** We use diagnoses and procedures data from the Medical Information Mart for Intensive Care (MIMIC-III) database[2] [57], which contains patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Patients who had less than three admission records are excluded. After this preprocessing, the average number of visits for the 1970 selected patients is 3.69, the average number of codes in a visit is 13.23, the total number of unique ICD-9 codes in diagnoses is 3320, and the total number of unique ICD-9 codes in procedures is 988.
- **MIMIC-IV Dataset** We use diagnoses and procedures data from the MIMIC-IV database[3] [57], which contains patients admitted between 2008 and 2019. Patients with less than three admission records are excluded. After this preprocessing, the average number of visits for the 10023 selected patients is 4.64, the average number of codes in a visit is 14.12, the total number of unique ICD-9 codes in diagnoses is 6274, and the total number of unique ICD-9 codes in procedures is 1973.

B. Baseline Models

We apply our method on the following baselines for the overall evaluation of diagnosis prediction accuracy. For a fair comparison, all models are used with adaptation to our task

[2] https://physionet.org/content/mimiciii/1.4/

[3] https://physionet.org/content/mimiciv/0.4/

TABLE II

WITH RANDOM DATA DIVISION, PERFORMANCES OF BASEMODELS, PW+BASEMODELS AND CHE+BASEMODELS. BEST PERFORMANCES ARE INDICATED BY BOLD FONTS. THE IMPROVEMENT INDICATES THE RELATIVE INCREASE OF CHE+BASEMODEL OVER BASEMODEL. $*$ DENOTES SIGNIFICANT IMPROVEMENT OF CHE+BASEMODEL, MEASURED BY T-TEST WITH P-VALUE$<$ 0.01, OVER BASEMODEL AND PW+BASEMODEL.

Approach	MIMIC-III				MIMIC-IV				Average
	NDCG@10	NDCG@20	ACC@10	ACC@20	NDCG@10	NDCG@20	ACC@10	ACC@20	
LSTM	0.2648	0.2712	0.1779	0.2597	0.3469	0.3386	0.2167	0.3084	0.2730
PW+LSTM	0.2669	0.2724	0.1791	0.2605	0.3488	0.3394	0.2177	0.3040	0.2736
CHE+LSTM	**0.2756***	**0.2809***	**0.1853***	**0.2690***	**0.3589***	**0.3496***	**0.2246***	**0.3186***	**0.2828***
Improv	4.079%	3.577%	4.160%	3.581%	3.459%	3.249%	3.646%	3.307%	3.590%
RETAIN	0.3409	0.3413	0.2305	0.3261	0.4095	0.3946	0.2568	0.3533	0.3316
PW+RETAIN	0.3436	0.3449	0.2316	0.3241	0.4120	0.3981	0.2580	0.3527	0.3331
CHE+RETAIN	**0.3545***	**0.3579***	**0.2353***	**0.3354***	**0.4231***	**0.4085***	**0.2630***	**0.3614***	**0.3424***
Improv	3.989%	4.864%	2.082%	2.852%	3.321%	3.523%	2.414%	2.293%	3.257%
Dipole	0.3071	0.3104	0.2075	0.2959	0.3801	0.3710	0.2379	0.3352	0.3056
PW+Dipole	0.3072	0.3110	0.2077	0.2965	0.3860	0.3754	0.2388	0.3376	0.3075
CHE+Dipole	**0.3371***	**0.3385***	**0.2213***	**0.3149***	**0.4054***	**0.3932***	**0.2523***	**0.3529***	**0.3270***
Improv	9.769%	8.842%	6.651%	6.421%	6.656%	5.984%	6.053%	5.280%	7.003%
Concare	0.2963	0.2979	0.1949	0.2793	0.3748	0.3615	0.2346	0.3226	0.2952
PW+Concare	0.2972	0.2980	0.1952	0.2798	0.3720	0.3602	0.2335	0.3234	0.2948
CHE+Concare	**0.3068***	**0.3121***	**0.2076***	**0.2935***	**0.3876***	**0.3760***	**0.2444***	**0.3371***	**0.3081***
Improv	3.544%	4.767%	6.516%	5.084%	3.415%	4.011%	4.177%	4.495%	4.370%
Stagenet	0.3364	0.3379	0.2284	0.3222	0.3979	0.3853	0.2513	0.3471	0.3258
PW+Stagenet	0.3343	0.3362	0.2267	0.3210	0.3960	0.3861	0.2529	0.3476	0.3251
CHE+Staegnet	**0.3432***	**0.3467***	**0.2315***	**0.3295***	**0.4064***	**0.3976***	**0.2559***	**0.3541***	**0.3331***
Improv	2.021%	2.604%	1.357%	2.266%	2.136%	3.192%	1.830%	2.017%	2.241%

where only historical diagnoses and procedures are available. Side information like ontology and temporal intervals is not fused, thus the performances may not necessarily match the ones reported in the original papers.

- **LSTM:** [35] A recurrent neural network with long-short term gating mechanism.
- **RETAIN:** [7] A two-level neural model based on reverse time attention for healthcare.
- **Dipole:** [3] An attention-based bidirectional recurrent neural network for healthcare.
- **Concare:** [10] A self-attention model with cross-head decorrelation to capture health context for healthcare.
- **StageNet:** [6] A deep learning model with stage-aware LSTM and convolutional modules for health prediction.

We denote above models as **BaseModels**, and we incorporate them with the **CHE** method as **CHE+BaseModels**.

In addition, counterfactual learning methods such as inverse propensity weighting are also frequently used. Therefore, we also conduct experiment on the counterfactual Permutation Weighting (**PW**) technique [25] for comparison. PW conducts permutation on observed features for calculating propensity scores. While the historical EHR in the dataset is regarded as positive samples, we randomly generate negative samples that do not exist in the dataset and estimate their propensity scores. The combination of various ICD-9 codes in a sequence is a large space as discussed in Section 3.2. We generate 10x larger number of negative samples to make the propensity estimation as accurate as possible. We also incorporate PW with the above BaseModels, and name them as **PW+BaseModels**. For each PW+BaseModel, the propensity scores are calculated via training the corresponding BaseModel with permutation.

C. Settings

We conduct two diagnosis prediction experiments. In the first experiment, we aim at evaluating the performance of our proposed method when training data and test data are divided randomly by patients to approximately simulate I.I.D. distributions. Following prior works [3], [7], We randomly divide the dataset into the training, validation and testing set in a 0.75:0.1:0.15 ratio. In the second experiment, considering the insurance type, such as Medicare, Medicaid and Private, may affect procedures for similar diagnoses, we evaluate the performance when training and test data are divided by the type of insurances to simulate the scenario of OOD generalization. Here, we divide all the Medicare data into the training and validation set in a 0.7:0.3 ratio and use the Private/Other (MIMIC-III/MIMIC-IV) data as the test set.

Common hyperparameters used by all models in the experiments include learning rate, batch size, hidden dimension, dropout rate. CHE+BaseModel has a special hyperparameter: weighting coefficient ϵ. These hyperparameters are tuned with an appropriate range to obtain the optimal evaluation metrics on the validation set for each individual model. The range of learning rate is $\{$1e-2, 3e-3, 1e-3$\}$, the range of batch size is $\{16, 32, 64, 128, 256\}$, the range of hidden dimension is $\{16, 32, 64\}$, the range of dropout rate is $\{0,1, 0.5\}$, the range of coefficient is $\{0.1, 0.3, 1, 3, 10\}$. We apply early-stopping so the training will stop if the validation metrics do not increase in twenty epochs and the test performance will be recorded. All results are averaged under five different random seeds and recorded in four significant figures.

TABLE III
Under out-of-distribution data, performances of BaseModels, PW+BaseModels and CHE+BaseModels. Best performances are indicated by bold fonts. The improvement indicates the relative increase of CHE+BaseModel over BaseModel. $*$ denotes significant improvement of CHE+BaseModel, measured by t-test with p-value$<$ 0.01, over BaseModel and PW+BaseModel.

Approach	MIMIC-III				MIMIC-IV				Average
	NDCG@10	NDCG@20	ACC@10	ACC@20	NDCG@10	NDCG@20	ACC@10	ACC@20	
LSTM	0.2082	0.2112	0.1420	0.1971	0.4395	0.4149	0.2519	0.3571	0.2771
PW+LSTM	0.2075	0.2102	0.1413	0.2012	0.4468	0.4221	0.2525	0.3598	0.2802
CHE+LSTM	**0.2182***	**0.2204***	**0.1492***	**0.2064***	**0.4649***	**0.4390***	**0.2692***	**0.3792***	**0.2933***
Improv	4.948%	4.356%	5.070%	4.718%	5.779%	5.809%	6.868%	6.189%	5.846%
RETAIN	0.2385	0.2447	0.1615	0.2364	0.5195	0.4859	0.3019	0.4173	0.3257
PW+RETAIN	0.2396	0.2443	0.1614	0.2362	0.5251	0.4888	0.3042	0.4195	0.3274
CHE+RETAIN	**0.2503***	**0.2589***	**0.1687***	**0.2492***	**0.5625***	**0.5260***	**0.3244***	**0.4519***	**0.3583***
Improv	4.948%	5.803%	4.458%	5.415%	8.277%	8.253%	7.453%	8.291%	6.939%
Dipole	0.2287	0.2367	0.1482	0.2264	0.4741	0.4469	0.2770	0.3873	0.3032
PW+Dipole	0.2402	0.2481	0.1567	0.2354	0.4727	0.4433	0.2796	0.3865	0.3078
CHE+Dipole	**0.2729***	**0.2772***	**0.1782***	**0.2631***	**0.5176***	**0.4905***	**0.3067***	**0.4280***	**0.3418***
Improv	19.33%	17.11%	20.24%	16.21%	9.175%	9.756%	10.72%	10.51%	12.73%
Concare	0.2139	0.2229	0.1398	0.2139	0.4910	0.4616	0.2857	0.3974	0.3033
PW+Concare	0.2122	0.2216	0.1414	0.2146	0.4947	0.4628	0.2878	0.4002	0.3044
CHE+Concare	**0.2253***	**0.2382***	**0.1521***	**0.2299***	**0.5270***	**0.4944***	**0.3087***	**0.4267***	**0.3253***
Improv	5.330%	6.864%	8.798%	7.480%	7.332%	7.106%	8.050%	7.373%	7.254%
Stagenet	0.2149	0.2224	0.1456	0.2171	0.5722	0.5423	0.3418	0.4754	0.3415
PW+Stagenet	0.2145	0.2230	0.1451	0.2153	0.5830	0.5522	0.3499	0.4874	0.3463
CHE+Staegnet	**0.2295***	**0.2373***	**0.1544***	**0.2320***	**0.6861***	**0.6567***	**0.4269***	**0.5879***	**0.4014***
Improv	6.794%	6.700%	6.044%	6.863%	19.91%	21.10%	24.90%	23.66%	17.54%

D. Evaluation Metrics

We adopt the top$k$ accuracy and normalized discounted cumulative gain (NDCG) to evaluate the diagnosis prediction performance. We use the same accuracy@k metric used in prior works [3], [8], [36], which is defined as the correct medical ICD-9 codes ranked in top$k$ divided by min(k, $|y_t|$), where $|y_t|$ is the number of ICD-9 codes in the (t+1)-th visit. NDCG@k further considers the normalization of gains and the ranking of correct medical codes, where codes with higher relevance will affect the final score more than those with lower relevance. In our experiments, we use $k \in [10, 20]$. We also provide an averaged metric of all the four metrics on two datasets to reflect the overall comparison to baselines.

E. Performance Comparison

In this subsection, we conduct performance comparison among BaseModels, PW+BaseModels and CHE+BaseModels, from two perspectives: datasets with random data division and out-of-distribution division respectively.

First, Table II shows performance comparison with random data division. The results show that CHE can provide improvements under all metrics on both datasets, as CHE is designed to remove spurious relationships for models to focus on causal features. Overall, the Averaged Metric of NDCG@k and ACC@k on two datasets increases by 4.092% compared to BaseModels. We use t-test with a p-value of 0.01 to evaluate the performance improvement and confirm that the improvements by CHE on all BaseModels are statistically significant. Moreover, PW+BaseModels slightly outperform BaseModels, but CHE+BaseModels can still significantly outperform PW+BaseModels. Using the optimization described in Eqs.(18-19), the HSIC is reduced by more than 1000 times. Because the dependency between past diagnoses and past procedures is minimized, non-causal features will not interfere with the model to learn from causal features for prediction.

Second, in Table III, we further show performance comparison with out-of-distribution data. Because we use Medicare insurance data for training and Private/Other insurance data for testing, the training and test sets are not I.I.D. Therefore, this poses additional challenges to the generalization of deep learning models. The results show that CHE has relatively greater improvements on all metrics against BaseModels than the results with random data division, with relative increase of the averaged metric of NDCG@k and ACC@k by 10.06%. And the significant test shows that, CHE+BaseModel significantly outperforms BaseModel and PW+BaseModel. This demonstrates our claim that the proposed CHE encourages models to rely on causal features and estimate their contributions to the prediction independently, as causal features are always useful to make disease diagnosis predictions regardless of the data distribution shifts. Moreover, the results also show that the counterfactual PW approach does not always increase the diagnosis prediction accuracy, which might be due to the inaccurate estimation of propensity scores as discussed in Section 3.2. We notice that the increase of metric values with the OOD data is more obvious for CHE+Stagenet than for other models. Based on our observation, it could be that the collaborative training of HSIC and cross-entropy loss is especially well optimized by the CHE strategy for CHE+Stagenet. We will further discuss how the CHE behaves and whether it can guide model optimization in the Visualization part.

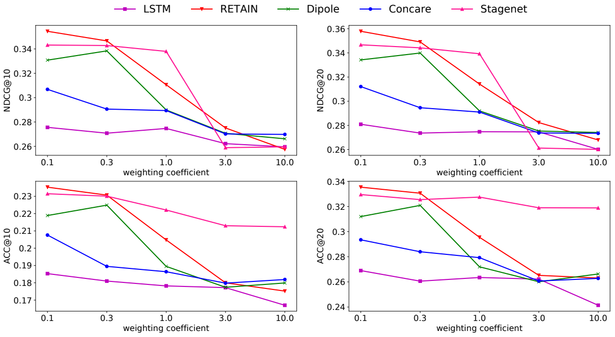

Fig. 4. Performances of CHE+BaseModels with different weighting coefficients on the MIMIC-III dataset with random data division.

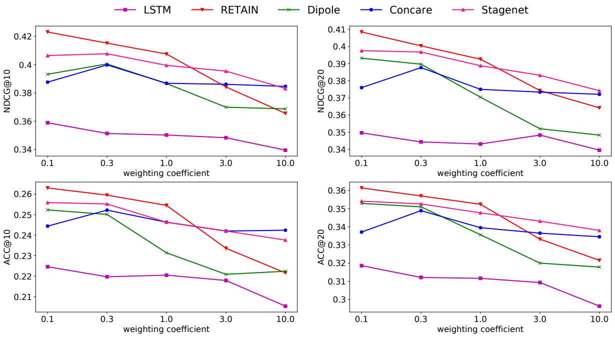

Fig. 5. Performances of CHE+BaseModels with different weighting coefficients on the MIMIC-IV dataset with random data division.

F. Hyperparameter Study

Because the selection of weighting coefficient ϵ is crucial for the optimization convergence, we report the model performances with different $\epsilon \in \{0.1, 0.3, 1, 3, 10\}$. Other hyperparameters are fixed. Fig. 4 and Fig.5 show that smaller ϵ is better for all models on both datasets. In general, the diagnosis prediction accuracy is still sensitive to the selection of coefficient. A too-large ϵ will cause a sharp drop in prediction performance. For example, when $\epsilon \in \{3, 10\}$, the NDCG@10 and ACC@10 drop by more than 30% for MIMIC-III dataset. $\epsilon \in \{0.1, 0.3\}$ has the optimal performance.

G. Visualization

To better understand whether the proposed Causal Healthcare Embedding can truly make the model to learn causal features, we visualize the contribution of each feature to the prediction of future diagnosis. We know that the cause of diabetic retinopathy is diabetes. For patients with background diabetic retinopathy (ICD-9 code 36021), an ideal model should rely on related diseases such as diabetes to make prediction. In Table IV, we show the EHR of a patient in the MIMIC-III dataset and the feature interpretations, contributions of the features to the prediction, in CHE+Dipole and Dipole. Specifically, we apply gradient backpropagation for calculating feature interpretations [58]–[61]. The contributions of diagnosis $d^i_{j'}$ and procedure $p^i_{j'}$ for predicting d^i_{j+1} $(j' \leq j)$ can be calculated as $\frac{\partial \hat{d}^i_{j+1}}{\partial d^i_{j'}} = \frac{\partial \hat{d}^i_{j+1}}{\partial \mathbf{E}^{i,j'}_d} \left(\mathbf{E}^{i,j'}_d\right)^T$ and $\frac{\partial \hat{d}^i_{j+1}}{\partial p^i_{j'}} = \frac{\partial \hat{d}^i_{j+1}}{\partial \mathbf{E}^{i,j'}_p} \left(\mathbf{E}^{i,j'}_p\right)^T$ respectively, where $\hat{d}^i_{j+1} = Prd(Enc(D^i_j), Enc(P^i_j))$ is the prediction as in the loss function in Eq. (15).

TABLE IV
FEATURE CONTRIBUTIONS OF A PATIENT FROM MIMIC-III.

Feature	CHE+Dipole			Dipole		
	1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}
Diagnosis	0.6841	1.710	1.201	0.8997	0.1366	1.714
Procedure	0.3413	0.1777	0.358	0.5469	0.1733	1.248

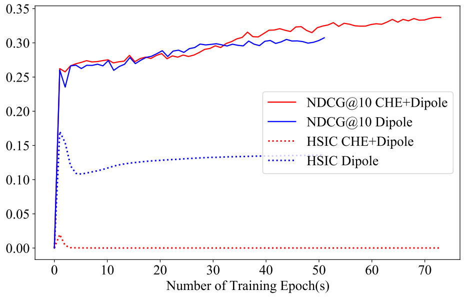

Fig. 6. The curves of HSIC and NDCG@10 on the test set when training CHE+Dipole and Dipole on the MIMIC-III dataset.

In this example, the diagnosis sequence is {4280, 5856}, {99592, 4280, 25060, 3572, V5861, V1251, 99662, 40391, 03811, 25050, 36201, 5856}, {03811, 5856, 99681, 42832}. The future diagnosis to be predicted is background diabetic retinopathy (36021). The first visit contains two diseases that appear frequently among people, i.e. congestive heart failure (4280) and end stage renal disease (5856). The second visit contains some highly related features, such as diabetic retinopathy (25050), diabetes with neurological manifestations (25060), and polyneuropathy in diabetes (3572). In the third visit, the complications of transplanted kidney (99681) might be related. Compared with Dipole, CHE+Dipole pays more attention to causal features, i.e., the second visit with many highly related diagnoses. Moreover, the contributions of diagnosis and procedure are less correlated.

We visualize the curves of HSIC and NDCG@10 on the test set in Figure 6, where the test samples are not weighted. Early stopping is adopted if the loss function does not decrease for 20 epochs. We find that CHE+Dipole has a much lower HSIC in inference and more training steps before early stopping with a higher NDCG@10 in the end. This aligns well with our observation that the CHE+Dipole prediction is more accurate, since the optimization path is guided by stable causality.

VI. CONCLUSION

In this work, we focus on the problem of unstable representation learning in deep learning-based diagnosis prediction models on EHR data. This is caused by the strong correlation between diagnoses and procedures in EHR, and it is hard for deep learning-based models to learn their causal relationships to future diagnosis. Accordingly, we propose a CHE method to learn causal representations for diagnosis prediction models, via removing dependencies between diagnoses and procedures by weighting technique. To be noted, our proposed CHE method can be used as a plug-and-play module. We demonstrate by extensive experiments on the sequential diagnosis and procedure features as examples that CHE can significantly improve the performances of diagnosis prediction models. The visualizations demonstrate that CHE is more causal than baselines and the optimization path is guided by causality, which marks a promising direction for healthcare. In future works, we will further explore the removals of spurious relationships in multimodal healthcare data for more real-world problems.

References

[1] D. Charles, M. Gabriel, and M. F. Furukawa, "Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2014," *ONC data brief*, vol. 9, pp. 1–9, 2013.

[2] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," in *SDM*, 2016, pp. 432–440.

[3] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *KDD*, 2017, pp. 1903–1911.

[4] T. Ma, C. Xiao, and F. Wang, "Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction," in *SDM*, 2018, pp. 261–269.

[5] X. S. Zhang, F. Tang, H. H. Dodge, J. Zhou, and F. Wang, "Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records," in *KDD*, 2019, pp. 2487–2495.

[6] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun, "Stagenet: Stage-aware neural networks for health risk prediction," in *The Web Conference*, 2020, pp. 530–540.

[7] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "Retain: an interpretable predictive model for healthcare using reverse time attention mechanism," in *NeurIPS*, 2016, pp. 3512–3520.

[8] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: graph-based attention model for healthcare representation learning," in *KDD*, 2017, pp. 787–795.

[9] J. Gao, X. Wang, Y. Wang, Z. Yang, J. Gao, J. Wang, W. Tang, and X. Xie, "Camp: Co-attention memory networks for diagnosis prediction in healthcare," in *ICDM*, 2019, pp. 1036–1041.

[10] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, and J. Gao, "Concare: Personalized clinical feature embedding via capturing the healthcare context," in *AAAI*, 2020, pp. 833–840.

[11] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola, "Correcting sample selection bias by unlabeled data," *NeurIPS*, pp. 601–608, 2006.

[12] M. A. Brookhart, T. Stürmer, R. J. Glynn, J. Rassen, and S. Schneeweiss, "Confounding control in healthcare database research: challenges and potential approaches," *Medical Care*, vol. 48, no. 6 0, p. S114, 2010.

[13] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*, 2018.

[14] Y. Bengio, T. Deleu, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal, "A meta-transfer objective for learning to disentangle causal mechanisms," in *ICLR*, 2019.

[15] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 2009.

[16] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *ICML*, 2013, pp. 10–18.

[17] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, "Invariant models for causal transfer learning," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1309–1342, 2018.

[18] J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference by using invariant prediction: identification and confidence intervals," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

[19] K. Kuang, P. Cui, S. Athey, R. Xiong, and B. Li, "Stable prediction across unknown environments," in *KDD*, 2018, pp. 1617–1626.

[20] K. Kuang, R. Xiong, P. Cui, S. Athey, and B. Li, "Stable prediction with model misspecification and agnostic distribution shift," in *AAAI*, 2020, pp. 4485–4492.

[21] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, A. J. Smola *et al.*, "A kernel statistical test of independence." in *NeurIPS*, 2007, pp. 585–592.

[22] D. Greenfeld and U. Shalit, "Robust learning with the hilbert-schmidt independence criterion," in *ICML*, 2020, pp. 3759–3768.

[23] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh, "Learning de-biased representations with biased representations," in *ICML*, 2020, pp. 528–539.

[24] H. Zou, P. Cui, B. Li, Z. Shen, J. Ma, H. Yang, and Y. He, "Counterfactual prediction for bundle treatment," *NeurIPS*, 2020.

[25] D. Arbour, D. Dimmery, and A. Sondhi, "Permutation weighting," in *ICML*, 2021, pp. 331–341.

[26] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *ICDM*, 2017, pp. 787–792.

[27] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, "Constructing disease network and temporal progression model via context-sensitive hawkes process," in *ICDM*, 2015, pp. 721–726.

[28] Y. Zhang, "Attain: Attention-based time-aware lstm networks for disease progression modeling." in *IJCAI*, 2019.

[29] M. Ye, J. Luo, C. Xiao, and F. Ma, "Lsan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction," in *CIKM*, 2020, pp. 1753–1762.

[30] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, "Deep computational phenotyping," in *KDD*, 2015, pp. 507–516.

[31] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *KDD*, 2015, pp. 705–714.

[32] Z. Zeng, Y. Deng, X. Li, T. Naumann, and Y. Luo, "Natural language processing for ehr-based computational phenotyping," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 1, pp. 139–153, 2018.

[33] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "Kame: Knowledge-based attention model for diagnosis prediction in healthcare," in *CIKM*, 2018, pp. 743–752.

[34] E. Choi, Z. Xu, Y. Li, M. Dusenberry, G. Flores, E. Xue, and A. Dai, "Learning the graphical structure of electronic health records with graph convolutional transformer," in *AAAI*, 2020, pp. 606–613.

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

[36] X. Zhang, B. Qian, S. Cao, Y. Li, H. Chen, Y. Zheng, and I. Davidson, "Inprem: An interpretable and trustworthy predictive model for healthcare," in *KDD*, 2020, pp. 450–460.

[37] J. Luo, M. Ye, C. Xiao, and F. Ma, "Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records," in *KDD*, 2020, pp. 647–656.

[38] X. Peng, G. Long, T. Shen, S. Wang, and J. Jiang, "Sequential diagnosis prediction with transformer and ontological representation," *arXiv preprint arXiv:2109.03069*, 2021.

[39] X. Zhang, B. Qian, Y. Li, Y. Liu, X. Chen, C. Guan, and C. Li, "Learning robust patient representations from multi-modal electronic health records: A supervised deep learning approach," in *SDM*, 2021, pp. 585–593.

[40] C. Chen, J. Liang, F. Ma, L. Glass, J. Sun, and C. Xiao, "Unite: Uncertainty-based health risk prediction leveraging multi-sourced data," in *The Web Conference*, 2021, pp. 217–226.

[41] M. Zhang, C. R. King, M. Avidan, and Y. Chen, "Hierarchical attention propagation for healthcare representation learning," in *KDD*, 2020, pp. 249–256.

[42] M. Ye, S. Cui, Y. Wang, J. Luo, C. Xiao, and F. Ma, "Medpath: Augmenting health risk prediction via medical knowledge paths," in *The Web Conference*, 2021, pp. 1397–1409.

[43] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *ICML*, 2016, pp. 3020–3029.

[44] S. L. Morgan and C. Winship, *Counterfactuals and Causal Inference*. Cambridge University Press, 2015.

[45] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[46] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research*, vol. 46, no. 3, pp. 399–424, 2011.

[47] N. Hassanpour and R. Greiner, "Learning disentangled representations for counterfactual regression," in *ICLR*, 2019.

[48] M. J. Lopez and R. Gutman, "Estimation of causal effects with multiple treatments: a review and new ideas," *Statistical Science*, pp. 432–454, 2017.

[49] N. Hassanpour and R. Greiner, "Counterfactual regression with importance sampling weights." in *IJCAI*, 2019, pp. 5880–5887.

[50] J. Yoon, J. Jordon, and M. Van Der Schaar, "Ganite: Estimation of individualized treatment effects using generative adversarial nets," in *ICLR*, 2018.

[51] Z. Shen, P. Cui, J. Liu, T. Zhang, B. Li, and Z. Chen, "Stable learning via differentiated variable decorrelation," in *KDD*, 2020, pp. 2185–2193.

[52] Z. Shen, P. Cui, T. Zhang, and K. Kunag, "Stable learning via sample reweighting," in *AAAI*, 2020.

[53] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *CVPR*, 2021, pp. 5372–5382.

[54] S. Fan, X. Wang, C. Shi, P. Cui, and B. Wang, "Generalizing graph neural networks on out-of-distribution graphs," *arXiv preprint arXiv:2111.10657*, 2021.

[55] D. E. Farrar and R. R. Glauber, "Multicollinearity in regression analysis: the problem revisited," *The Review of Economic and Statistics*, pp. 92–107, 1967.

[56] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. Lanckriet, and B. Schölkopf, "Kernel choice and classifiability for rkhs embeddings of probability distributions." in *NIPS*, vol. 22, 2009, pp. 1750–1758.

[57] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[59] D. Smilkov, N. Thorat, B. Kim, F. Viegas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," in *ICML*, 2017.

[60] J. Li, X. Chen, E. Hovy, and D. Jurafsky, "Visualizing and understanding neural models in nlp," *arXiv preprint arXiv:1506.01066*, 2015.

[61] Q. Liu, Z. Liu, H. Zhang, Y. Chen, and J. Zhu, "Mining cross features for financial credit risk assessment," in *CIKM*, 2021.