Stable Learning for Out-Of-Distribution Healthcare Diagnosis Prediction

Yingtao Luo[1], Qiang Liu[2]
[1]University of Washington
[2]Chinese Academy of Sciences
yl3851@uw.edu,qiang.liu@nlpr.ia.ac.cn

ABSTRACT

Deep learning models have achieved promising disease prediction performance based on the web-based Electronic Health Records (EHR) of patients, however, learning statistical relations can be spurious to generalize to out-of-distribution data due to the strong correlation between procedures and diagnoses. Without the ability to distinguish between the two features, the model is incapable to learn the stable contribution of historical diagnoses and procedures to predict future diagnoses, which leads to untrustworthy representation learning. To address this problem, we propose to use a stable representation learning method called Causal Healthcare Embedding (CHE) that learns the stable contribution of diagnoses and procedures by removing their dependencies. Performing the sample reweighting technique, our method ensures a correct causal diagram for learning healthcare data and improves prediction generalization across different environments. Moreover, our proposed CHE method can be used as a flexible play-and-plug module that can enhance existing diagnosis prediction models. 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 for prediction without the confounding of unstable dependencies.

CCS CONCEPTS

• **Applied computing → Health care information systems**; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Healthcare informatics, causal inference, electronic health records, out-of-distribution

ACM Reference Format:
Yingtao Luo, Qiang Liu. 2018. Stable Learning for Out-Of-Distribution Healthcare Diagnosis Prediction. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 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. [7]. In the US, nearly 96% of hospitals had a digital electronic health record (EHR) in 2015 [7], which emphasizes the online disease prediction and prevention. EHR's comprehensive patient information (such as demographics, diagnoses, and procedures) provides valuable assistance for personal health status tracking and monitoring [11, 18, 38, 42, 66]. To predict the future diagnoses based on a patient's historical EHR, many deep learning models [12, 13, 17, 41] are proposed with promising accuracy to discover the statistical correlations in the training distribution for predictions under the I.I.D. hypothesis. However, the hypothesis does not hold true for most practical situations due to the distribution shifts such as data selection bias and confounding factors [5, 6, 23, 26]. For example, the patients' demographics in different hospitals may vary a lot. The interpretability of diagnosis prediction [40, 64] that characterizes the importance of each medical record in the final prediction result will also be unstable with shifting correlations across different environments. To this end, we are interested in discovering a stable correlation that reflects the causal effect of each causal feature across different environments, which is free of data biases introduced by the distribution shifts between training and inference. Learning the causation between causal features and the prediction can effectively increase the deep learning healthcare system trustworthiness for clinicians [48].

Causal inference and counterfactual predictions have been extensively studied in healthcare [32, 36, 51, 57]. Methods such as inverse propensity weighting [3, 24] are leveraged to weigh single observations to mimic the effects of randomization with respect to one variable of interest. Although these causal techniques are adopted for prediction tasks with few variables and single treatment, the diagnosis prediction by deep learning involves much more variables and treatments. The enormous space of combinations of different treatments in the sequential pattern brings challenges to accurately estimating the propensity of each treatment to simulate randomized controlled trials [1, 69]. Moreover, algorithms that aim at addressing distribution shifts such as domain generalization [44], causal transfer learning [49] and invariant causal prediction [47] cannot deal with distribution shifts unobserved in the training data. In the light of these problems, we are motivated to propose an algorithm that can estimate correct causation and make trustworthy healthcare predictions where the domains of training data are unknown without excessive computational complexity.

As shown in Fig. 1, the causal diagram of future diagnosis prediction in EHR consists of two sequences of causal features, i.e. "past

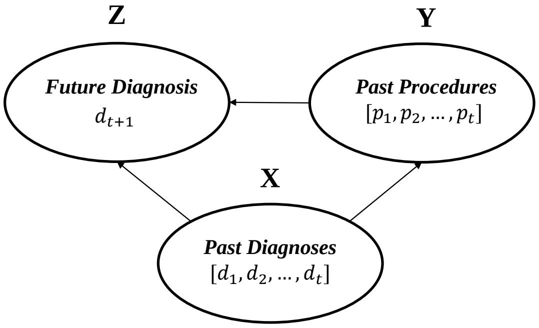

Figure 1: The causal diagram of diagnosis prediction in EHR.

diagnoses" (X) and "past procedures" (Y). By the ignorability assumption in causality [45], potential confounders are not discussed to make the scheme easier for understanding. Here, we discuss the scheme where the X and Y are correlated, thus the causal effect of each feature cannot be accurately estimated by statistical 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 statistical learning models to learn a stable relationship of $X \rightarrow Z$ and/or $Y \rightarrow Z$. As an instance, patients with diabetes usually take insulin, and diabetes may cause puffiness. With the strong correlation between diabetes and insulin, a statistical 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, which may result in basic-indemnificatory procedures and/or high-quality procedures for different patients with similar diagnoses. Therefore, models trained by one type of insurances may not generalize to other types of insurances. Accordingly, by decorrelating diagnoses and procedures of the EHR of patients, we have a larger chance to obtain more stable and causal diagnosis prediction models.

To obtain a stable correlation structure between each variable and the final prediction, a strand of variable decorrelation technique [29, 30] is proposed for linear models. Its basic notion is to remove the dependencies between variables through a sample reweighting 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 X to Y (the arrows from each diagnosis to each procedure) 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 a challenge. 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) [19, 20] that measures the norm of cross-covariance from X to Y for feature decorrelation [4], which can align with the nonlinear neural models. Second, as pointed out by [69], treatments can be represented by latent factors as an alternative for estimating causality. While it is computationally expensive to calculate the binary sample weighting for all treatments[1], 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 removed throughout the training. In this way, the deep learning model can learn a stable correlation between each feature and the prediction, which generalizes to different environments without necessarily knowing the training data a priori.

To be noted, our proposed CHE method is a play-and-plug module for the diagnosis prediction task. It can be easily and adaptively incorporated with various diagnosis prediction models for EHR. 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 diagnoses and procedures by sample reweighting 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 play-and-plug module that can work well with various diagnosis prediction models in EHR. Therefore, the proposed method can be flexibly used in a variety of scenarios.
- Extensive experiments on two public datasets show that our method increases both NDCG and ACC of five state-of-the-art baselines by a significant margin, even when applying to data of different sources and/or distributions.

2 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.

2.1 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 [8, 11, 42, 66], disease progression [14, 59, 67], phenotyping [9, 33, 61], diagnosis prediction [15, 17, 39].

Owing to the sequential pattern of EHR data, Recurrent Neural Networks (RNNs) are naturally suitable, and Long Short-Term Memory (LSTM) [25] has been successfully applied. RETAIN [12] presents a reverse time attention model that preserves interpretability. Dipole [38] incorporates bidirectional RNN for making a prediction based on EHR. Camp [17] uses demography information in co-attention model for diagnosis prediction. ConCare [41] proposes to incorporate multi-head self-attention to model the sequential data of EHR. StageNet [18] integrates time intervals between visits into LSTM to model the stages of health conditions. INPREM [64] 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 [59] combines both long- and short-term information in EHR to make predictions. SETOR [46] utilizes ontological representation and neural ordinary equation for diagnosis prediction. Meanwhile, multi-sourced data is also considered in recent works [10, 65]. Besides, constructed on sequential prediction models, medical knowledge graphs are modeled to provide some prior knowledge for more accurate predictions [13, 39, 58, 62].

2.2 Counterfactual Prediction

Counterfactual learning [27] is an important direction of research in causal inference [43, 45]. Counterfactual learning can enable people to estimate the probability of counterfactual events and eventually identify the unbiased causal relations between events. The existing counterfactual learning approaches usually reweight samples based on propensity scores [2, 22, 35, 50], 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 [1, 21]. Recent works further extend counterfactual learning to the multi-level treatment [60] and bundle treatment [69] settings. Meanwhile, Permutation Weighting (PW) [1] conducts permutation on observed features for calculating propensity scores.

2.3 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 [29]. 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 [29, 30, 53]. Sample Reweighted Decorrelation Operator (SRDO) [54] generates some unobserved samples, and trains a binary classifier to get the probabilities of observation for reweighting the observed samples.

3 PRELIMINARY

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

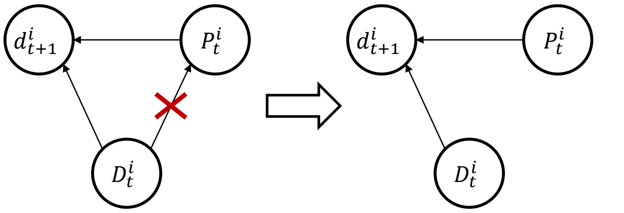

Figure 2: Decorrelation between past diagnoses D^i_t and past procedures P^i_t, for more accurate and stable prediction of future diagnosis d^i_{t+1}.

3.1 Problem Formulation

EHR data consists of a set of patients $V = \{v_1, v_2, ..., v_{|V|}\}$, and each patient 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 after t visits, there exists a historical diagnosis sequence $D^i_t = \left[d^i_1, d^i_2, ..., d^i_t\right]$ and a historical procedure sequence $P^i_t = \left[p^i_1, p^i_2, ..., p^i_t\right]$. Each diagnosis and procedure are M-dimensional multi-hot vector and N-dimensional multi-hot vector respectively, which means that $d^i_j \in \{0,1\}^M$ and $p^i_j \in \{0,1\}^N$. 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^i_t and P^i_t, we need to predict future diagnosis d^i_{t+1}.

3.2 Discussion

Several causal links in medical knowledge affect the causal relationships in EHR. First, diagnoses affect current and future procedures, in which doctors select procedures according to the patients' current and historical diseases based on medical experience. Second, both diagnoses and procedures have impacts on future diagnoses. It is easy to understand that diseases have development processes, which are accompanied by complications. Procedures, such as surgical operations and medication, affect the development and rehabilitation of diseases. Consequently, these relationships lead to a strong correlation between past diagnoses D^i_t and past procedures P^i_t. And both past diagnoses and past procedures have causal effects on future diagnosis, i.e., $D^i_t \longrightarrow d^i_{t+1}$ and $P^i_t \longrightarrow d^i_{t+1}$.

However, the strong correlation between diagnoses and procedures hinders statistical machine learning models from learning correct causal relationships for future diagnosis. Without the ability to remove the dependency between D^i_t and P^i_t, statistical models may optimize to unstable relationships of $D^i_t \longrightarrow d^i_{t+1}$ and $P^i_t \longrightarrow d^i_{t+1}$ with unstable model weights according to 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 always occur together, it is hard for statistical models to identify whether diabetes or insulin causes complications. On the other hand, if the correlation between D^i_t and P^i_t is eliminated, the

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

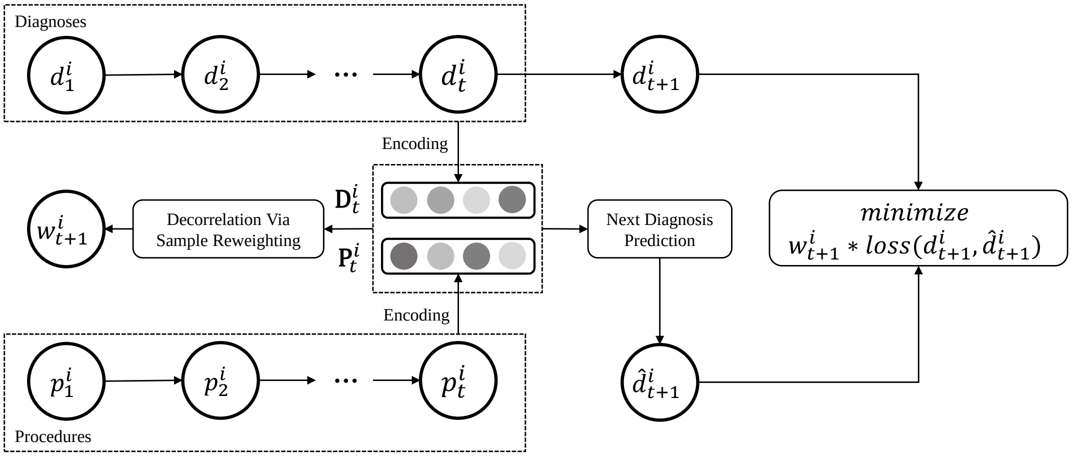

Figure 3: The schematic diagram of decorrelation between past diagnoses D_t^i and past procedures P_t^i via sample reweighting.

model can estimate the importance of diabetes and insulin for future diagnoses independently. 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.

The above phenomenon is easy to understand in linear models, considering that multicollinearity problem has been widely studied in linear models [16]. Diagnosis prediction models are mostly based on deep learning approaches, thus the discussion should be done in the embedding space. We first encode above diagnoses and procedures into an embedding space by any EHR encoders

$$\mathbf{D}_t^i = Encoder\left(D_t^i\right), \tag{1}$$

$$\mathbf{P}_t^i = Encoder\left(P_t^i\right), \tag{2}$$

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

$$\mathbf{d}_{t+1}^i = f\left(\mathbf{D}_t^i, \mathbf{P}_t^i\right). \tag{3}$$

Considering the strong correlation between D_t^i and P_t^i, we can predict one of them based on another as

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

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

where $g_{P\to D}(\cdot)$ and $g_{D\to P}(\cdot)$ are deep learning models, $\varepsilon_{P\to D}\left(\mathbf{D}_t^i\right)$ and $\varepsilon_{D\to P}\left(\mathbf{P}_t^i\right)$ are residual terms of the prediction, $\varepsilon_{P\to D}\left(\mathbf{D}_t^i\right) \neq \mathbf{D}_t^i$ and $\varepsilon_{D\to P}\left(\mathbf{P}_t^i\right) \neq \mathbf{P}_t^i$. Then, Eq. (3) can rewritten as

$$\mathbf{d}_{t+1}^i = f\left(g_{P\to D}\left(\mathbf{P}_t^i\right) + \varepsilon_{P\to D}\left(\mathbf{D}_t^i\right), \mathbf{P}_t^i\right), \tag{6}$$

or

$$\mathbf{d}_{t+1}^i = f\left(\mathbf{D}_t^i, g_{D\to P}\left(\mathbf{D}_t^i\right) + \varepsilon_{D\to P}\left(\mathbf{P}_t^i\right)\right). \tag{7}$$

Considering $g_{P\to D}(\cdot)$ and $g_{D\to P}(\cdot)$ are both deep learning models, so that their parameters can be merged with $f(\cdot)$. Thus, there shall be another deep learning model $f'(\cdot)$ that satisfies

$$\mathbf{d}_{t+1}^i = f'\left(\varepsilon_{P\to D}\left(\mathbf{D}_t^i\right), \mathbf{P}_t^i\right), \tag{8}$$

or

$$\mathbf{d}_{t+1}^i = f'\left(\mathbf{D}_t^i, \varepsilon_{D\to P}\left(\mathbf{P}_t^i\right)\right). \tag{9}$$

To this end, we can conclude that, due to the strong correlation between D_t^i and P_t^i, the model can either rely on D or P to make diagnosis prediction, which is unstable especially when generalizing to out-of-distribution data where the correlation between D_t^i and P_t^i changes. Thus, we need to decorrelate past diagnoses and past procedures. Once they are not correlated, we shall obtain $g_{P\to D}\left(\mathbf{P}_t^i\right) = 0$, $g_{D\to P}\left(\mathbf{D}_t^i\right) = 0$, $\varepsilon_{P\to D}\left(\mathbf{D}_t^i\right) = \mathbf{D}_t^i$ and $\varepsilon_{D\to P}\left(\mathbf{P}_t^i\right) = \mathbf{P}_t^i$. Then, the contribution of D_t^i and P_t^i to predicting d_{t+1}^i can be free of the interference of $D_t^i \to P_t^i$. Accordingly, as illustrated in Fig. 2, we plan to decorrelate $D_t^i \to P_t^i$ to learn more accurate, stable and causal diagnosis prediction models.

Another perspective regarding the reason why feature decorrelation works falls in the assumption that features at each timestamp are either stable or unstable [30]. For example, there may not be a causal relationship between the patient's trauma in a traffic accident ten years ago and his diabetic complications, but the unstable coexistence of the two events in the training data hinders the model from learning from the stable feature of his recent diabetes. Performing variable decorrelation for each feature at each timestamp for each patient removes the dependencies between stable features and unstable features, which encourages the model to learn from stable features that can generalize to different environments in any scenario. Because the training data is sampled from the patients in the real world, the data samples are inevitably biased, which could lead to the cases discussed above.

4 METHODOLOGY

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

We remove the dependency between D_t^i and P_t^i by sample weighting to make a stable prediction. Consider the measurable, positive definite kernel k of both variables and the corresponding Reproducing Kernel Hilbert Space (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:

$$\begin{aligned}\langle h_D, \Sigma_{DP} h_P \rangle = \mathbb{E}_{DP}[h_D(\mathbf{D}_t^i) h_P(\mathbf{P}_t^i)] \\ -\mathbb{E}_D[h_D(\mathbf{D}_t^i)]\mathbb{E}_P[h_P(\mathbf{P}_t^i)]. \end{aligned} \tag{10}$$

As proved by [63], if the product of k_D and k_P is characteristic, $\mathbb{E}[k_D(\mathbf{D}_t^i, \mathbf{D}_t^i)] < \infty$ and $\mathbb{E}[k_P(\mathbf{P}_t^i, \mathbf{P}_t^i)] < \infty$, we have

$$\Sigma_{DP} = 0 \Longleftrightarrow D \perp P, \tag{11}$$

which means that if $\mathbf{D}_t^i$ cannot be transformed into $\mathbf{P}_t^i$ via a nonlinear operator, the two variables are independent.

The squared Hilbert-Schmidt norm of the cross-covariance operator Σ_{DP} can be calculated by Hilbert-Schmidt Independence Criterion (HSIC) [19, 20]. Here, we use the quick unbiased calculation of HSIC in the embedding space by

$$HSIC(\mathbf{D}, \mathbf{P}) = \frac{1}{|V| \cdot (t^i - 1)} \sum_{i=1}^{|V|} \sum_{j=1}^{t^i - 1} HSIC(\mathbf{d}_j^i, \mathbf{p}_j^i). \tag{12}$$

Specifically, if h denotes the hidden dimensionality, we can consider calculating the HSIC of each $\mathbf{d}_j^i \in \mathbb{R}^h$ and $\mathbf{p}_j^i \in \mathbb{R}^h$ by

$$HSIC(\mathbf{d}_j^i, \mathbf{p}_j^i) = \frac{1}{(n-1)^2} Tr(K_d J K_p J), \tag{13}$$

where Tr is the trace of a matrix, $J = 1/n$, K_D and K_P are any kernel matrices. We can consider RBF kernel to calculate

$$K_d(x_1, x_2) = \exp(\frac{\|x_1 - x_2\|_2^2}{\sigma^2}), \tag{14}$$

where $x_1, x_2 \in \mathbf{d}_j^i \in \mathbb{R}^h$ represent the values in different dimensions of the latent $\mathbf{d}_j^i \in \mathbf{D}_t^i$. Similarly, there is $K_p(x_1, x_2)$ where $x_1, x_2 \in \mathbf{p}_t^i \in \mathbb{R}^h$ represent different dimensions of the latent representation of $\mathbf{p}_j^i \in \mathbf{P}_t^i$. The kernel tricks of $K_d, K_p \in \mathbb{R}^{h \times h}$ can approximately calculate HSIC rapidly. In this way, for each time's visit by each patient, the cross-covariance in the embedding space from diagnosis to procedure can be measured by HSIC.

Inspired by feature decorrelation techniques [53, 54], we propose to minimize HSIC by sample reweighting to mitigate the dependency between diagnoses and procedures in the embedded space. We use $w^i \in \mathbb{R}^t$ to denote the sample weights of a patient, in which w_j^i denotes the weight for each time's visit of a patient. We denote the weighted samples as $\mathbf{WD}$ and $\mathbf{WP}$, where the weighted samples of patient i at timestamp t, i.e., $\mathbf{WD}_t^i$ and $\mathbf{WP}_t^i$, are

$$\mathbf{WD}_t^i = w^i \, \mathbf{D}_t^i = [w_1^i \, \mathbf{d}_1^i, w_2^i \, \mathbf{d}_2^i, ..., w_t^i \, \mathbf{d}_t^i], \tag{15}$$

$$\mathbf{WP}_t^i = w^i \, \mathbf{P}_t^i = [w_1^i \, \mathbf{p}_1^i, w_2^i \, \mathbf{p}_2^i, ..., w_t^i \, \mathbf{p}_t^i]. \tag{16}$$

To minimize the correlation between diagnoses and procedures, we propose to optimize w with HSIC as follows

$$w^* = \underset{w}{\operatorname{argmin}} \, HSIC(\mathbf{WD}, \mathbf{WP}). \tag{17}$$

Overall, we iteratively optimize the weighted loss and the HSIC by

$$Enc_{n+1}, Prd_{n+1} = \underset{Enc, Prd}{\operatorname{argmin}} \sum_{i=1}^{|V|} \sum_{t=1}^{t^i - 1} w_t^i(n) \mathbf{L}_t^i, \tag{18}$$

where

$$\mathbf{L}_t^i = L(Prd(Enc(D_t^i), Enc(P_t^i)), d_{t+1}^i), \tag{19}$$

and

$$w(n+1) = \underset{w}{\operatorname{argmin}} \, \epsilon \cdot HSIC(wEnc_{n+1}(D), wEnc_{n+1}(P)). \tag{20}$$

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

Eq. (18) and Eq. (20) are optimized iteratively, meaning that we first optimize the neural network and then optimize the HSIC for each iteration. Every two subsequences D_t^i and P_t^i of length t 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 $\mathbf{D}_t^i$ and $\mathbf{P}_t^i$ 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. The convergence of both HSIC and weighted cross-entropy loss ensures that the fine-tuned model maps diagnoses and procedures into an embedding space where each diagnosis is independent of each procedure.

The time complexity of calculating HSIC only grows with the hidden dimensionality h. By naive algorithms, the multiplication of K_d and K_p is $O(h^3)$, and the calculation of trace is also $O(h^3)$. For deep learning models, compared to the number of data samples $|V|$ and the number of timestamps t, h is a hyperparameter and is thus trivial. 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. Therefore, using traditional counterfactual weighting to estimate propensity scores can be very challenging. This reflects that our proposed module is a way to achieve causal prediction without excessive complexity.

5 EXPERIMENTS

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

5.1 Datasets

We evaluate our proposed sequential counterfactual 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

Table 1: Performances of BaseModels, PW+BaseModels and CHE+BaseModels with random data division. 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			
	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
PW+LSTM	0.2669	0.2724	0.1791	0.2605	0.3488	0.3394	0.2177	0.3040
CHE+LSTM	**0.2756***	**0.2809***	**0.1853***	**0.2690***	**0.3589***	**0.3496***	**0.2246***	**0.3186***
Improv %	4.079%	3.577%	4.160%	3.581%	3.459%	3.249%	3.646%	3.307%
RETAIN	0.3409	0.3413	0.2305	0.3261	0.4095	0.3946	0.2568	0.3533
PW+RETAIN	0.3436	0.3449	0.2316	0.3241	0.4120	0.3981	0.2580	0.3527
CHE+RETAIN	**0.3545***	**0.3579***	**0.2353***	**0.3354***	**0.4231***	**0.4085***	**0.2630***	**0.3614***
Improv %	3.989%	4.864%	2.082%	2.852%	3.321%	3.523%	2.414%	2.293%
Dipole	0.3071	0.3104	0.2075	0.2959	0.3801	0.3710	0.2379	0.3352
PW+Dipole	0.3072	0.3110	0.2077	0.2965	0.3860	0.3754	0.2388	0.3376
CHE+Dipole	**0.3308***	**0.3342***	**0.2189***	**0.3120***	**0.4054***	**0.3932***	**0.2523***	**0.3529***
Improv %	7.717%	7.668%	5.494%	5.441%	6.656%	5.984%	6.053%	5.280%
Concare	0.2963	0.2979	0.1949	0.2793	0.3748	0.3615	0.2346	0.3226
PW+Concare	0.2972	0.2980	0.1952	0.2798	0.3720	0.3602	0.2335	0.3234
CHE+Concare	**0.3068***	**0.3121***	**0.2076***	**0.2935***	**0.3876***	**0.3760***	**0.2444***	**0.3371***
Improv %	3.544%	4.767%	6.516%	5.084%	3.415%	4.011%	4.177%	4.495%
Stagenet	0.3364	0.3379	0.2284	0.3222	0.3979	0.3853	0.2513	0.3471
PW+Stagenet	0.3343	0.3362	0.2267	0.3210	0.3960	0.3861	0.2529	0.3476
CHE+Staegnet	**0.3432***	**0.3467***	**0.2315***	**0.3295***	**0.4064***	**0.3976***	**0.2559***	**0.3541***
Improv %	2.021%	2.604%	1.357%	2.266%	2.136%	3.192%	1.830%	2.017%

(MIMIC-III) database[2] [28], 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 Medical Information Mart for Intensive Care (MIMIC-IV) database[3] [28], which contains patients admitted to an ICU or the emergency department between 2008 and 2019. Patients who had 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.

5.2 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 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:** [25] A recurrent neural network with long-short term gating mechanism.
- **RETAIN:** [12] A two-level neural model based on reverse time attention for healthcare.
- **Dipole:** [38] An attention-based bidirectional recurrent neural network for healthcare.
- **Concare:** [41] A self-attention model that uses cross-head decorrelation to capture health context for healthcare.
- **StageNet:** [18] A deep learning model with stage-aware LSTM and convolutional modules for health risk 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 in scenarios where the number of treatments is small. Except for the above baselines, we also conduct experiment on the counterfactual Permutation Weighting (**PW**) technique [1] for comparison. PW conducts permutation on observed features for calculating propensity scores. While the historical EHR in the original dataset is naturally regarded as positive samples, we randomly generate negative samples that do not exist in the dataset and estimate their propensity scores via PW learning. Because the combination of various ICD-9 codes in a

[2] https://physionet.org/content/mimiciii/1.4/
[3] https://physionet.org/content/mimiciv/0.4/

sequence is a large space as discussed in Section 3.2, we generate as many negative samples as possible (ten times larger than the original dataset) to make sure that the propensity estimation is as accurate as possible. We also incorporate PW with the above BaseModels, and name them as **PW+BaseModels**.

5.3 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 [12, 38], 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 data and test data are divided by the type of insurances to simulate the scenario of out-of-distribution 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 including learning rate, batch size, hidden dimension, dropout rate. And 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}. The optimization 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. Therefore, the rounding error is within 5×10^{-5}.

5.4 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 [13, 38, 64], 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 $k \in [10, 20]$.

5.5 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 1 shows performance comparison with random data division. The results show that CHE can provide improvements under all metrics on both datasets. Overall, the NDCG@k relatively increases by 4.15%, and ACC@k relatively increases by 3.70% on average for all datasets and 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 100 to 1000 times. Because the dependency between past diagnoses and past procedures is minimized, unstable features without causal links to the future diagnosis will not interfere with the model to learn from causal features. Deep learning models will tend to weigh the contributions of stable features to the future diagnosis prediction independently.

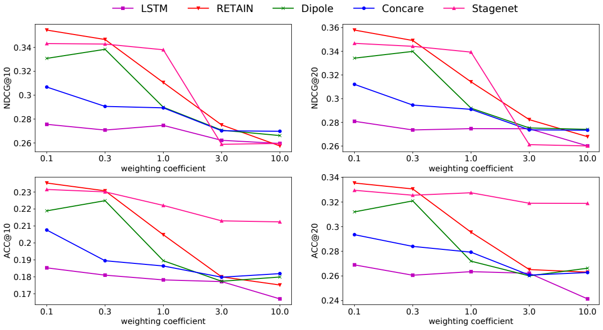

Figure 4: Performances of CHE+BaseModels with different weighting coefficients on the MIMIC-III dataset with random data division.

Second, in Table 2, 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 NDCG@k by 8.20% and relative increase of ACC@k by 9.20% on average. 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 stable features and estimate their contributions to the prediction independently. Because stable features have causal links to future diagnosis, they 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.

5.6 Hyperparameter Study

Because the selection of weighting coefficient ϵ is crucial for the optimization convergence, and ϵ is the special hyperparameter of our proposed CHE method, we report the diagnosis prediction performances with different $\epsilon \in \{0.1, 0.3, 1, 3, 10\}$ in the hyperparameter study. 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 sensitive to the selection of coefficient. A too-large ϵ will cause a sharp drop in prediction performance, and $\epsilon \in \{0.1, 0.3\}$ has the optimal performance.

Table 2: Performances of BaseModels, PW+BaseModels and CHE+BaseModels under out-of-distribution data. 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			
	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
PW+LSTM	0.2075	0.2102	0.1413	0.2012	0.4468	0.4221	0.2525	0.3598
CHE+LSTM	**0.2182***	**0.2184***	**0.1492***	**0.2064***	**0.4649***	**0.4390***	**0.2692***	**0.3792***
Improv %	4.948%	3.409%	5.070%	4.718%	5.779%	5.809%	6.868%	6.189%
RETAIN	0.2385	0.2447	0.1615	0.2364	0.5195	0.4859	0.3019	0.4173
PW+RETAIN	0.2396	0.2443	0.1614	0.2362	0.5251	0.4888	0.3042	0.4195
CHE+RETAIN	**0.2503***	**0.2589***	**0.1687***	**0.2492***	**0.5413***	**0.5061***	**0.3126***	**0.4329***
Improv %	4.948%	5.803%	4.458%	5.415%	4.196%	4.157%	3.544%	3.738%
Dipole	0.2287	0.2367	0.1482	0.2264	0.4741	0.4469	0.2770	0.3873
PW+Dipole	0.2402	0.2481	0.1567	0.2354	0.4727	0.4433	0.2796	0.3865
CHE+Dipole	**0.2729***	**0.2772***	**0.1782***	**0.2631***	**0.5176***	**0.4905***	**0.3067***	**0.4280***
Improv %	19.33%	17.11%	20.24%	16.21%	9.175%	9.756%	10.72%	10.51%
Concare	0.2139	0.2229	0.1398	0.2139	0.4910	0.4616	0.2857	0.3974
PW+Concare	0.2122	0.2216	0.1414	0.2146	0.4947	0.4628	0.2878	0.4002
CHE+Concare	**0.2199***	**0.2310***	**0.1521***	**0.2264***	**0.5270***	**0.4944***	**0.3087***	**0.4267***
Improv %	2.805%	3.634%	8.798%	5.844%	7.332%	7.106%	8.050%	7.373%
Stagenet	0.2149	0.2224	0.1456	0.2171	0.5722	0.5423	0.3418	0.4754
PW+Stagenet	0.2145	0.2230	0.1451	0.2153	0.5830	0.5522	0.3499	0.4874
CHE+Staegnet	**0.2199***	**0.2310***	**0.1521***	**0.2264***	**0.6861***	**0.6567***	**0.4269***	**0.5879***
Improv %	2.327%	3.867%	4.464%	4.284%	19.91%	21.10%	24.90%	23.66%

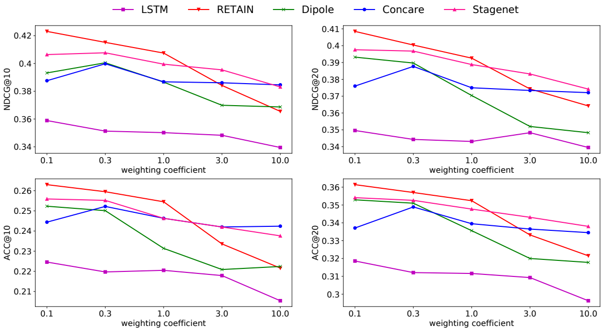

Figure 5: Performances of CHE+BaseModels with different weighting coefficients on the MIMIC-IV dataset with random data division.

Table 3: Feature interpretations 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	0.2475

5.7 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), a causal and stable model should rely on related diseases such as diabetes instead of other less related features to make prediction. In Table 3, we show the EHR of a patient in the MIMIC-III dataset and the feature interpretations, contribution of the features to the prediction, in CHE+Dipole and Dipole. Specifically, we apply gradient backpropagation for calculating feature interpretations [31, 34, 52, 55, 68].

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}, {41401, 4280, 25050, 36201, 99591, 5856, 25060, 3572}. 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.

6 CONCLUSION

In this work, we focus on the problem of unstable and unreliable 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 true 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 reweighting technique. To be noted, our proposed CHE method can be used as a play-and-plug module. Experiments show that, CHE can significantly improve the performances of diagnosis prediction models.

REFERENCES

[1] David Arbour, Drew Dimmery, and Arjun Sondhi. 2021. Permutation weighting. In *ICML*. 331–341.

[2] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46, 3 (2011), 399–424.

[3] Peter C Austin and Elizabeth A Stuart. 2015. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine* 34, 28 (2015), 3661–3679.

[4] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. 2020. Learning de-biased representations with biased representations. In *ICML*. 528–539.

[5] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2019. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. In *ICLR*.

[6] M Alan Brookhart, Til Stürmer, Robert J Glynn, Jeremy Rassen, and Sebastian Schneeweiss. 2010. Confounding control in healthcare database research: challenges and potential approaches. *Medical Care* 48, 6 0 (2010), S114.

[7] Dustin Charles, Meghan Gabriel, and Michael F Furukawa. 2013. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2014. *ONC data brief* 9 (2013), 1–9.

[8] 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 *ICDM*. 787–792.

[9] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep computational phenotyping. In *KDD*. 507–516.

[10] Chacha Chen, Junjie Liang, Fenglong Ma, Lucas Glass, Jimeng Sun, and Cao Xiao. 2021. UNITE: Uncertainty-based Health Risk Prediction Leveraging Multi-sourced Data. In *The Web Conference*. 217–226.

[11] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In *SDM*. 432–440.

[12] Edward Choi, Mohammad Taha Bahadori, Joshua A Kulas, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In *NeurIPS*. 3512–3520.

[13] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning. In *KDD*. 787–795.

[14] Edward Choi, Nan Du, Robert Chen, Le Song, and Jimeng Sun. 2015. Constructing disease network and temporal progression model via context-sensitive hawkes process. In *ICDM*. 721–726.

[15] Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. 2020. Learning the graphical structure of electronic health records with graph convolutional transformer. In *AAAI*. 606–613.

[16] Donald E Farrar and Robert R Glauber. 1967. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics* (1967), 92–107.

[17] Jingyue Gao, Xiting Wang, Yasha Wang, Zhao Yang, Junyi Gao, Jiangtao Wang, Wen Tang, and Xing Xie. 2019. Camp: Co-attention memory networks for diagnosis prediction in healthcare. In *ICDM*. 1036–1041.

[18] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In *The Web Conference*. 530–540.

[19] Daniel Greenfeld and Uri Shalit. 2020. Robust learning with the hilbert-schmidt independence criterion. In *ICML*. 3759–3768.

[20] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. 2007. A kernel statistical test of independence.. In *NeurIPS*. 585–592.

[21] Negar Hassanpour and Russell Greiner. 2019. CounterFactual Regression with Importance Sampling Weights.. In *IJCAI*. 5880–5887.

[22] Negar Hassanpour and Russell Greiner. 2019. Learning disentangled representations for counterfactual regression. In *ICLR*.

[23] Dan Hendrycks and Thomas Dietterich. 2018. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *ICLR*.

[24] Miguel A Hernán and James M Robins. 2006. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health* 60, 7 (2006), 578–586.

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[26] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. 2006. Correcting sample selection bias by unlabeled data. *NeurIPS* (2006), 601–608.

[27] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *ICML*. 3020–3029.

[28] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 1 (2016), 1–9.

[29] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. 2018. Stable prediction across unknown environments. In *KDD*. 1617–1626.

[30] Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. 2020. Stable prediction with model misspecification and agnostic distribution shift. In *AAAI*. 4485–4492.

[31] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066* (2015).

[32] Jiuyong Li, Saisai Ma, Thuc Le, Lin Liu, and Jixue Liu. 2016. Causal decision trees. *IEEE Transactions on Knowledge and Data Engineering* 29, 2 (2016), 257–271.

[33] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. 2015. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *KDD*. 705–714.

[34] Qiang Liu, Zhaocheng Liu, Haoli Zhang, Yuntian Chen, and Jun Zhu. 2021. Mining Cross Features for Financial Credit Risk Assessment. In *CIKM*.

[35] Michael J Lopez and Roee Gutman. 2017. Estimation of causal effects with multiple treatments: a review and new ideas. *Statist. Sci.* (2017), 432–454.

[36] Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. 2018. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics* 27, 1 (2018), 209–219.

[37] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *KDD*. 647–656.

[38] 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 *KDD*. 1903–1911.

[39] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *CIKM*. 743–752.

[40] Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *AAAI*. 825–832.

[41] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. 2020. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *AAAI*. 833–840.

[42] Tengfei Ma, Cao Xiao, and Fei Wang. 2018. Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction. In *SDM*. 261–269.

[43] Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and Causal Inference*. Cambridge University Press.

[44] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *ICML*. 10–18.

[45] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys* 3 (2009), 96–146.

[46] Xueping Peng, Guodong Long, Tao Shen, Sen Wang, and Jing Jiang. 2021. Sequential Diagnosis Prediction with Transformer and Ontological Representation. *arXiv preprint arXiv:2109.03069* (2021).

[47] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* (2016), 947–1012.

[48] Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2, 7 (2020), 369–375.

[49] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. 2018. Invariant models for causal transfer learning. *The Journal of Machine Learning Research* 19, 1 (2018), 1309–1342.

[50] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[51] Megan S Schuler and Sherri Rose. 2017. Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology* 185, 1 (2017), 65–73.

[52] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*. 618–626.

[53] Zheyan Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li, and Zhitang Chen. 2020. Stable learning via differentiated variable decorrelation. In *KDD*. 2185–2193.

[54] Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kunag. 2020. Stable learning via sample reweighting. In *AAAI*.

[55] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viegas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. In *ICML*.

[56] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert RG Lanckriet, and Bernhard Schölkopf. 2009. Kernel choice and classifiability for RKHS embeddings of probability distributions.. In *NIPS*, Vol. 22. 1750–1758.

[57] Mark J Van der Laan and Sherri Rose. 2011. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media.

[58] Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. MedPath: Augmenting Health Risk Prediction via Medical Knowledge Paths. In *The Web Conference*. 1397–1409.

[59] Muchao Ye, Junyu Luo, Cao Xiao, and Fenglong Ma. 2020. LSAN: Modeling Long-term Dependencies and Short-term Correlations with Hierarchical Attention for Risk Prediction. In *CIKM*. 1753–1762.

[60] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*.

[61] Zexian Zeng, Yu Deng, Xiaoyu Li, Tristan Naumann, and Yuan Luo. 2018. Natural language processing for EHR-based computational phenotyping. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16, 1 (2018), 139–153.

[62] Muhan Zhang, Christopher R King, Michael Avidan, and Yixin Chen. 2020. Hierarchical attention propagation for healthcare representation learning. In *KDD*. 249–256.

[63] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. 2021. Deep Stable Learning for Out-Of-Distribution Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5372–5382.

[64] Xianli Zhang, Buyue Qian, Shilei Cao, Yang Li, Hang Chen, Yefeng Zheng, and Ian Davidson. 2020. INPREM: An Interpretable and Trustworthy Predictive Model for Healthcare. In *KDD*. 450–460.

[65] Xianli Zhang, Buyue Qian, Yang Li, Yang Liu, Xi Chen, Chong Guan, and Chen Li. 2021. Learning Robust Patient Representations from Multi-modal Electronic Health Records: A Supervised Deep Learning Approach. In *SDM*. 585–593.

[66] Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. 2019. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *KDD*. 2487–2495.

[67] Yuan Zhang. 2019. ATTAIN: Attention-based Time-Aware LSTM Networks for Disease Progression Modeling.. In *IJCAI*.

[68] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *CVPR*. 2921–2929.

[69] Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. 2020. Counterfactual prediction for bundle treatment. *NeurIPS* (2020).