VecoCare: Visit Sequences-Clinical Notes Joint Learning for Diagnosis Prediction in Healthcare Data

Yongxin Xu[1,3] , **Kai Yang**[4] * , **Chaohe Zhang**[1,3] , **Peinie Zou**[1,3] , **Zhiyuan Wang**[1,3] ,
Hongxin Ding[1,3] , **Junfeng Zhao**[1,3] * , **Yasha Wang**[1,2] * and **Bing Xie**[1,3]
[1] Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China
[2] National Engineering Research Center For Software Engineering, Peking University, Beijing, China
[3] School of Computer Science and School of Software & Microelectronics,
Peking University, Beijing, China
[4] Zhongguancun Laboratory, Beijing, China
xuyx@stu.pku.edu.cn, {zhaojf, wangyasha, xiebing}@pku.edu.cn, yangkai@mail.zgclab.edu.cn

Abstract

Due to the insufficiency of electronic health records (EHR) data utilized in practical diagnosis prediction scenarios, most works are devoted to learning powerful patient representations either from structured EHR data (e.g., temporal medical events, lab test results, etc.) or unstructured data (e.g., clinical notes, etc.). However, synthesizing rich information from both of them still needs to be explored. Firstly, the heterogeneous semantic biases across them heavily hinder the synthesis of representation spaces, which is critical for diagnosis prediction. Secondly, the intermingled quality of partial clinical notes leads to inadequate representations of to-be-predicted patients. Thirdly, typical attention mechanisms mainly focus on aggregating information from similar patients, ignoring important auxiliary information from others. To tackle these challenges, we propose a novel visit sequences-clinical notes joint learning approach, dubbed VecoCare. It performs a Gromov-Wasserstein Distance (GWD)-based contrastive learning task and an adaptive masked language model task in a sequential pre-training manner to reduce heterogeneous semantic biases. After pre-training, VecoCare further aggregates information from both similar and dissimilar patients through a dual-channel retrieval mechanism. We conduct diagnosis prediction experiments on two real-world datasets, which indicates that VecoCare outperforms state-of-the-art approaches. Moreover, the findings discovered by VecoCare are consistent with the medical researches.

1 Introduction

With the widespread adoption of electronic healthcare information systems in various healthcare institutions, many deep learning models have been developed to leverage electronic health records (EHR) data for important medical applications [Yang *et al.*, 2017; Ma *et al.*, 2020; Yang *et al.*, 2021; Feng *et al.*, 2021; Ma *et al.*, 2021; Zhang *et al.*, 2021; Zhang *et al.*, 2022; Ma *et al.*, 2022; Xu *et al.*, 2023]. Diagnosis prediction is one of these vitally important healthcare scenarios. It predicts the future diagnoses of patients based on their historical sequences of clinical events, such as the diagnoses, lab tests, medications, etc.

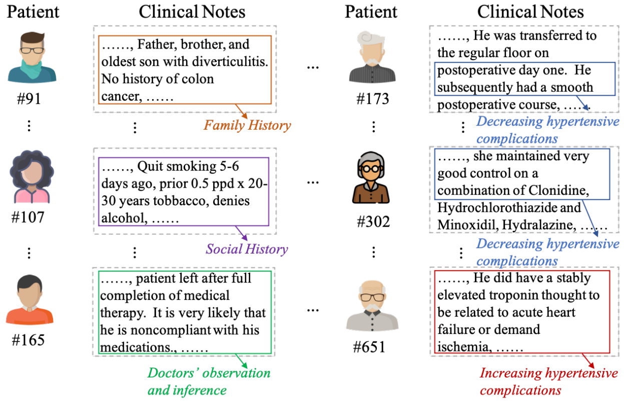
Patient
Clinical Notes
Patient
Clinical Notes
#91
……, Father, brother, and oldest son with diverticulitis. No history of colon cancer, ……
Family History
#107
……, Quit smoking 5-6 days ago, prior 0.5 ppd x 20-30 years tobbacco, denies alcohol, ……
Social History
#165
……, patient left after full completion of medical therapy. It is very likely that he is noncompliant with his medications., ……
Doctors' observation and inference
#173
……, He was transferred to the regular floor on postoperative day one. He subsequently had a smooth postoperative course, ……
Decreasing hypertensive complications
#302
……, she maintained very good control on a combination of Clonidine, Hydrochlorothiazide and Minoxidil, Hydralazine, ……
Decreasing hypertensive complications
#651
……, He did have a stably elevated troponin thought to be related to acute heart failure or demand ischemia, ……
Increasing hypertensive complications

Figure 1: Examples of unstructured clinical notes in EHR data.

Due to the data insufficiency in practical scenarios, making full use of the information contained in various types of data becomes the focus of deep learning model design. From the perspective of the type of EHR data, some try to leverage the information of structured EHR data (e.g., static baseline information, lab test values, historical diagnoses, etc.) to perform the diagnosis prediction [Ma *et al.*, 2017; Choi *et al.*, 2017; Ma *et al.*, 2018; Luo *et al.*, 2020; Tan *et al.*, 2022]. While some works delve into mining the unstructured clinical notes (i.e., texts written by clinical experts with assessments and concerns regarding patients' clinical conditions) [Thapa *et al.*, 2022]. As shown in Figure 1, compared with structured EHR data recording medical events and physiological signals, unstructured clinical notes describe fine-grained information about the patients' family history, social

*Corresponding Author.

history, the doctors' observations and inferences, etc., which are critically complementary to structured data.

However, synthesizing rich information from structured and unstructured EHR data is still marginally studied for diagnosis prediction. In this work, we exploit the potential synergies between structured EHR data and unstructured clinical notes to form a harmonious representation space for accurate diagnosis predictions. Although seemingly straightforward, implementing this intuition will encounter the following challenges:

C1. It is challenging to synthesize two representation spaces due to data heterogeneity. Some works integrate keywords of the clinical notes based on the similarity of medical events and textual clinical notes for diagnosis prediction [Lu *et al.*, 2021]. However, structured data are discrete codes or continuous values that quantitatively describe changes in a patient's health status. In contrast, unstructured clinical notes are mainly clinicians' qualitative descriptions in the natural language form. They are incompatible in data distributions, metric measures, and semantic representation spaces [Bronstein *et al.*, 2010; Roostaiyan *et al.*, 2017]. With typical attention mechanisms, it is challenging to reduce heterogeneous semantic biases across two types of information for the joint representation of optimal diagnosis results.

C2. It is tough to learn robust representations due to the intermingled quality of partial clinical notes. For the sake of differences in recorders or the urgency of patients' conditions, some patients' clinical notes suffer from misspellings, poor grammar, non-standard abbreviations and insufficient information [Burke *et al.*, 2014], leading the model to learn inadequate representations and resulting in sub-optimal performance. Therefore, it is necessary to refer to relevant information of other patients by exploiting the potential synergies and correlations among them to strengthen the utilization of low-quality clinical notes.

C3. It is difficult to include all valuable patients in the reference range only by typical methods such as attention mechanisms. When complementing low-quality clinical notes with other patients' information, a natural intuition is to focus on significant samples through dot-product attention [Vaswani *et al.*, 2017]. In this way, larger weights are assigned to more similar patients (e.g., in Figure 1, patient #173 and patient #302 are both patients with progressively decreasing hypertensive complications). According to clinicians' practical diagnosis experiences, patients with extremely different or even opposite healthcare statuses (e.g., in Figure 1, patient #651 is a patient with a progressive increase in hypertensive complications) are also essential for diagnosis predictions [Unay and Ekin, 2011; Jia *et al.*, 2020], which is ignored by existing approaches.

To cope with these challenges, we put forward `VecoCare`, a novel visit sequences-clinical notes joint learning approach to fuse them into a representation space seamlessly for accurate diagnosis prediction. Specifically, our main contributions are summarized as follows:

- To solve the challenge **C1**, we propose two sequential pre-training tasks to bridge heterogeneous semantic biases across structured EHR data and clinical notes. Firstly, `VecoCare` incorporates a novel Gromov-Wasserstein Distance (GWD)-based contrastive learning task to learn a consistent semantic representation space by maximizing the agreement between visit sequence-clinical note pairs. Secondly, `VecoCare` further reduces heterogeneous semantic biases with an adaptive masked language model task based on a novel global-local fusing encoder, which can adaptively balance the influence of both types of information.
- Addressing challenge **C2** and **C3**, `VecoCare` employs a dual-channel retrieval mechanism to aggregate important auxiliary information from similar and dissimilar patients for more comprehensive representations.
- We conduct extensive experiments on two real-world EHR datasets, which show that `VecoCare` outperforms all state-of-the-art models in different evaluation metrics, including approaches that incorporate task-specific external knowledge. Besides, the medical findings discovered by `VecoCare` are also in accord with medical literature and can provide valuable medical insights or explanations.

2 Related Work

Over the past decade, there has been lots of works focusing on diagnosis prediction with deep learning models, with the vast majority of these methods focusing on mining structured data. One category of methods attempts to focus on capturing contextual dependencies between patient visit sequences. For example, T-LSTM [Baytas *et al.*, 2017] handles irregular time intervals by enabling time decay. RETAIN [Choi *et al.*, 2016] proposes a two-level attention mechanism based on recurrent neural networks (RNN). Dipole [Ma *et al.*, 2017] employs a combination of bi-directional RNN and attention mechanisms to predict diagnoses of patients' next visits. LSAN [Ye *et al.*, 2020] and HiTANet [Luo *et al.*, 2020] employ the self-attention mechanism to capture the temporal patterns. Furthermore, T-ContextGGAN [Xu *et al.*, 2022] and Chet [Lu *et al.*, 2022] build a graph structure according to the patient's medical history, and use graph neural networks (GNN) to learn the representation. Another mainstream category tries to focus on leveraging external medical knowledge graphs to improve representation learning. For example, GRAM [Choi *et al.*, 2017] and KAME [Ma *et al.*, 2018] both incorporate the hierarchy of disease codes to enhance learning. CGL [Lu *et al.*, 2021] constructs a graph with both medical knowledge and personal clinical observations, thence employing a collaborative graph learning method to learn the representations. MedPath [Ye *et al.*, 2021] utilizes the personalized knowledge graph to assist prediction. MetaCare++ [Tan *et al.*, 2022] incorporates domain knowledge of hierarchical and syndromic relations between various diseases. Based on the modeling of structured data, some existing works try to aggregate important information from unstructured clinical notes. For example, CGL [Lu *et al.*, 2021] integrates keywords of the clinical notes based on the similarity of medical events and textual clinical notes. However, they marginally investigated the heterogeneous semantic biases, the intermingled quality of partial clinical notes, and the difficulties in including valuable samples.

3 Problem Formulation

EHR Dataset. In this paper, EHR data consist of structured time-ordered visit sequences and unstructured clinical notes of patients. Let $\mathcal{C} = \left\{c_1, c_2, \ldots, c_{|\mathcal{C}|}, c_*\right\}$ be the entire set of codes used in an EHR dataset, where $|\mathcal{C}|$ is the number of medical codes. Following [Luo *et al.*, 2020], we also denote a special code c_* to represent the whole patient data. For the i-th patient, the visit sequence is defined as $\mathbf{X}_i = \left[\mathbf{x}_i^{cls}, \mathbf{x}_i^1, \mathbf{x}_i^2, \cdots, \mathbf{x}_i^T\right]$, where the t-th visit is denoted by a multi-hot vector $\mathbf{x}_i^t \in \{0,1\}^{|\mathcal{C}|}$. The k-th element of one visit vector is set to 1 if it contains the medical code c_k. $\mathbf{x}_i^{cls}$ denotes the [CLS] token which only contains the special code c_*. As for the clinical notes, let $\mathcal{N} = \left\{n_1, n_2, \ldots, n_{|\mathcal{N}|}\right\}$ be the dictionary of clinical notes, where $|\mathcal{N}|$ is the number of words. Following [Lu *et al.*, 2021], we select the notes $\mathbf{O}_i^T$ from the patient's T-th visit as another input because it already contains a summary of the observations from the previous T visits. For convenience, we drop the superscript T in $\mathbf{O}_i^T$ in the rest of this paper. $\mathbf{O}_i$ can be represented as $\mathbf{O}_i = \left[\mathbf{o}_i^{cls}, \mathbf{o}_i^1, \mathbf{o}_i^2, \cdots, \mathbf{o}_i^M\right]$, where M is the length of the clinical notes, $\mathbf{o}_i^m$ is represented by a binary vector $\{0,1\}^{|\mathcal{N}|}$, and $\mathbf{o}_i^{cls}$ is the [CLS] token.

Diagnosis Prediction. In this paper, our predictive objective is presented as a diagnosis prediction task, which is a multi-label classification problem. Given the previous T visits $\mathbf{X}_i$ and the clinical notes $\mathbf{O}_i$ of the i-th patient, the task is to predict a binary vector $\mathbf{y}_i \in \{0,1\}^{|\mathcal{Y}|}$, which represents the possible diagnoses in the $(T+1)$-th visit, where $|\mathcal{Y}|$ is the number of diagnoses.

4 Methodology

Figure 2 shows the architecture of VecoCare. It comprises the following sub-modules in a sequential manner: 1) GWD-based Contrastive Learning Module, 2) Adaptive Masked Language Model Module, 3) Dual-channel Retrieval Module.

4.1 GWD-based Contrastive Learning Module

Base Encoder. Given a sparse binary visit vector $\mathbf{x}_i^t$ and a sparse binary word vector $\mathbf{o}_i^m$, we first encode them to a relatively dense space using two linear functions as follows:

$$\begin{aligned} \mathbf{v}_i^t &= \mathbf{W}_v \mathbf{x}_i^t + \mathbf{b}_v, \\ \mathbf{s}_i^m &= \mathbf{W}_s \mathbf{o}_i^m + \mathbf{b}_s, \end{aligned} \tag{1}$$

where $\mathbf{W}_v \in \mathbb{R}^{d_v \times (|\mathcal{C}|+1)}$, $\mathbf{b}_v \in \mathbb{R}^{d_v}$, $\mathbf{W}_s \in \mathbb{R}^{d_s \times (|\mathcal{N}|+1)}$, and $\mathbf{b}_s \in \mathbb{R}^{d_s}$ are trainable parameters. As a result, the data of each patient can be represented by $\mathbf{V}_i = \left[\mathbf{v}_i^{cls}, \mathbf{v}_i^1, \mathbf{v}_i^2, \cdots, \mathbf{v}_i^T\right]$ and $\mathbf{S}_i = \left[\mathbf{s}_i^{cls}, \mathbf{s}_i^1, \mathbf{s}_i^2, \cdots, \mathbf{s}_i^M\right]$. To explicitly capture the global interactions within the visit sequence and to obtain a time-aware contextual representation, we utilize a time-aware Transformer [Luo *et al.*, 2020] f^v as the base visit encoder to extract visit contextual features $\mathbf{H}_i$:

$$\left[\mathbf{h}_i^{cls}, \mathbf{h}_i^1, \mathbf{h}_i^2, \cdots, \mathbf{h}_i^T\right] = f^v\left(\left[\mathbf{v}_i^{cls}, \mathbf{v}_i^1, \mathbf{v}_i^2, \cdots, \mathbf{v}_i^T\right]\right). \tag{2}$$

For the clinical notes, we adopt a Transformer [Vaswani *et al.*, 2017] f^s as the text encoder backbone to encode notes contextual features $\mathbf{U}_i$:

$$\left[\mathbf{u}_i^{cls}, \mathbf{u}_i^1, \mathbf{u}_i^2, \cdots, \mathbf{u}_i^M\right] = f^s\left(\left[\mathbf{s}_i^{cls}, \mathbf{s}_i^1, \mathbf{s}_i^2, \cdots, \mathbf{s}_i^M\right]\right). \tag{3}$$

To map two types of contextual representations to a joint embedding space, we then employ two distinct non-linear projection layers (g_v and g_s) to convert $\mathbf{h}_i^t$ and $\mathbf{u}_i^m$ into normalized lower-dimensional embeddings $\tilde{\mathbf{h}}_i^t \in \mathbb{R}^d$ and $\tilde{\mathbf{u}}_i^m \in \mathbb{R}^d$, respectively.

To learn a consistent semantic representation space for alignment and measurement, a natural idea is to maximizing the agreement between true visit sequences-clinical notes pairs versus random pairs via contrastive learning [He *et al.*, 2020]. However, due to the heterogeneity of the data, the different metric-measure spaces present significant challenges when attempting to reliably evaluate the similarity between these two types of contextual representations. The Gromov-Wasserstein Optimal Transport is employed to minimize the transportation cost between two distributions by directly comparing the metric spaces, rather than evaluating samples across these spaces [Peyré *et al.*, 2016; Mémoli, 2011]. Essentially, this framework focuses on the distances between pairs of points within each domain and assesses how these distances correspond to those in alternate domains. As a result, we compute the Gromov-Wasserstein distance (GWD) between the two representation distributions and employ this distance as a loss function to further optimize the representation learning. Specifically, let $\mathbf{p}_h$ and $\mathbf{p}_u$ represent discrete distributions of two types of contextual representations, where $\mathbf{p}_h = \left\{p_h^0, p_h^1, \ldots, p_h^T\right\}$ and $\mathbf{p}_u = \left\{p_u^0, p_u^1, \ldots, p_u^M\right\}$, $\sum_{j=0}^{T} p_h^j = \sum_{k=0}^{M} p_u^k = 1$. The GWD between the two discrete distributions $\mathbf{p}_h$, $\mathbf{p}_u$ can be defined as:

$$D_{gw}(\mathbf{p}_h, \mathbf{p}_u) = \min_{\boldsymbol{T} \in \pi(\mathbf{p}_h, \mathbf{p}_u)} \sum_{j,j',k,k'} \boldsymbol{T}_{jk} \boldsymbol{T}_{j'k'} \hat{c}\left(\tilde{\mathbf{h}}_i^j, \tilde{\mathbf{u}}_i^k, \tilde{\mathbf{h}}_i^{j'}, \tilde{\mathbf{u}}_i^{k'}\right), \tag{4}$$

where $\pi(\mathbf{p}_h, \mathbf{p}_u)$ denotes all the joint distributions, $\boldsymbol{T}$ represents the transport plan between two types of features and $\boldsymbol{T}_{jk}$ denotes the amount of mass shifted from p_h^j to p_u^k. In addition, $\hat{c}\left(\tilde{\mathbf{h}}_i^j, \tilde{\mathbf{u}}_i^k, \tilde{\mathbf{h}}_i^{j'}, \tilde{\mathbf{u}}_i^{k'}\right) = \left\| d\left(\tilde{\mathbf{h}}_i^j, \tilde{\mathbf{h}}_i^{j'}\right) - d\left(\tilde{\mathbf{u}}_i^k, \tilde{\mathbf{u}}_i^{k'}\right) \right\|_2$ is the cost function to measure the distance between different metric measure spaces where the distance $d\left(\tilde{\mathbf{h}}_i^j, \tilde{\mathbf{h}}_i^{j'}\right) = \exp\left(-\cos\left(\tilde{\mathbf{h}}_i^j, \tilde{\mathbf{h}}_i^{j'}\right)/\tau_g\right)$ of two features is measured based on the cosine similarity, τ_g is the temperature parameter.

Based on GWD, for the i-th visit sequences-clinical notes contextual features pair ($\tilde{\mathbf{H}}_i$,$\tilde{\mathbf{U}}_i$) in a mini-batch, we alternate between treating two distinct feature types as queries and keys to learn the correct pairings. This results in a pair of symmetric, temperature-normalized InfoNCE losses [Oord *et al.*, 2018] that optimize the preservation of mutual information between the authentic pairs in the latent space:

$$\begin{aligned} \ell_i^{\mathrm{v2t}} &= -\log \frac{\exp\left(-D_{gw}\left(\tilde{\mathbf{H}}_i, \tilde{\mathbf{U}}_i\right)/\tau\right)}{\sum_{k=1}^{B} \exp\left(-D_{gw}\left(\tilde{\mathbf{H}}_i, \tilde{\mathbf{U}}_k\right)/\tau\right)}, \\ \ell_i^{\mathrm{t2v}} &= -\log \frac{\exp\left(-D_{gw}\left(\tilde{\mathbf{U}}_i, \tilde{\mathbf{H}}_i\right)/\tau\right)}{\sum_{k=1}^{B} \exp\left(-D_{gw}\left(\tilde{\mathbf{U}}_i, \tilde{\mathbf{H}}_k\right)/\tau\right)}, \end{aligned} \tag{5}$$

where B is the batch size and τ is the instance-level temperature hyper-parameter. The overall objective of GWD-based

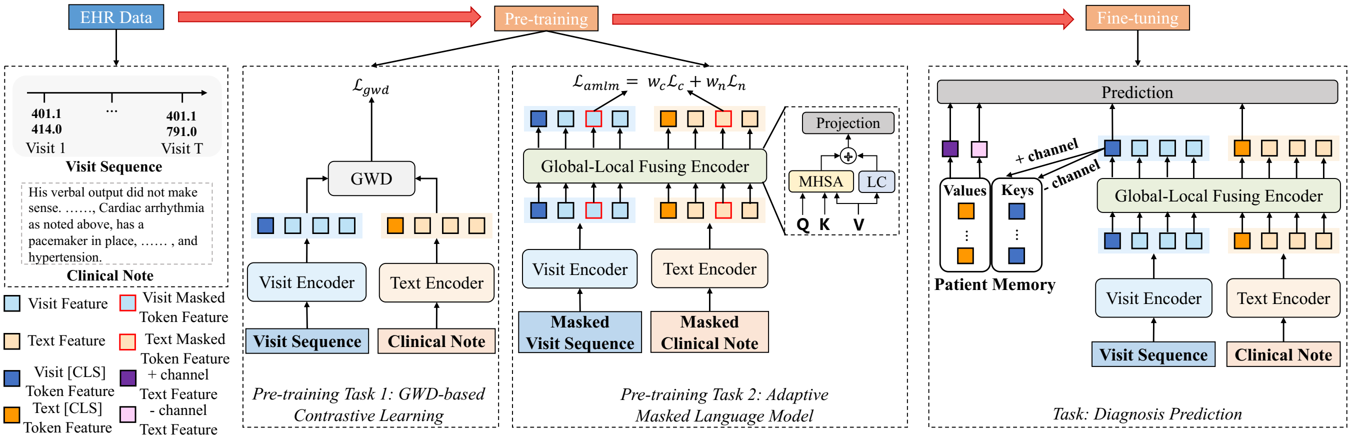

Figure 2: The Framework of VecoCare

contrastive learning is the average of the two losses:

$$\mathcal{L}_{\text{gwd}} = \frac{1}{2Tr}\sum_{i=1}^{Tr}\left(\ell_i^{\text{v2t}} + \ell_i^{\text{t2v}}\right), \tag{6}$$

where Tr is the total number of visit sequences-clinical notes pairs.

4.2 Adaptive Masked Language Model Module

After a pre-training phase based on the GWD-based contrastive learning task, VecoCare learns a consistent semantic representation space. To perform deeper feature information aggregation and integration, we propose a novel global-local fusing encoder to enable the coupling of global (self-attention) and local (convolution) information. Transformers can effectively capture global long-range dependencies. However, on the one hand, there are temporal patterns of disease changes in neighboring visits [Ye *et al.*, 2020], and on the other hand, there are many information-rich local text fragments (e.g., past medical history, social history, etc.) in the clinical notes. The inductive bias of convolutional neural networks (CNN) can help us aggregate local features. Specifically, given the sequence of features $\tilde{\mathbf{z}}^0 = \left[\tilde{\mathbf{H}}_i; \tilde{\mathbf{U}}_i\right]$ encoded by pre-trained base encoders and projection layers, we propose to add an convolution operation on the values when calculating self-attention:

$$\begin{aligned}\tilde{\mathbf{z}}^l &= \text{LayerNorm}\left(\tilde{\mathbf{z}}^{l-1} + \text{MHSA}\left(\tilde{\mathbf{z}}^{l-1}\right) + \text{Conv}\left(\tilde{\mathbf{z}}_v^{l-1}\right)\right), \\ \bar{\mathbf{z}}^l &= \text{LayerNorm}\left(\tilde{\mathbf{z}}^l + \text{FFN}\left(\tilde{\mathbf{z}}^l\right)\right),\end{aligned} \tag{7}$$

where $l = 1, \ldots, L$ refers to the number of such stacked layers. MHSA refers to the Multi-Head Self-Attention, FFN refers to a feed forward network, and LayerNorm is the layer normalization. Following [Vaswani *et al.*, 2017], $\tilde{\mathbf{z}}_v^{l-1}$ is the value projected directly from $\tilde{\mathbf{z}}^{l-1}$. After L layers, we can obtain the representation $\bar{\mathbf{z}}^L = \left[\bar{\mathbf{H}}_i; \bar{\mathbf{U}}_i\right]$ which captures global and local token-wise interactions.

Then, we propose an adaptive masked language model (AMLM) task which randomly masks word tokens and medical code tokens and predicts the ground-truth labels from the output of the global-local fusing encoder. It integrates the context of other medical code tokens and textual tokens:

$$\mathcal{L}_c = \sum_{c_i \in \mathcal{C}_{\text{mask}}} -\log p\left(c_i\right), \quad \mathcal{L}_n = \sum_{n_i \in \mathcal{N}_{\text{mask}}} -\log p\left(n_i\right), \tag{8}$$

where $\mathcal{C}_{\text{mask}}$ is the set of masked medical code tokens, $\mathcal{N}_{\text{mask}}$ is the set of masked textual tokens, $p\left(c_i\right)$ and $p\left(n_i\right)$ denote the probability of predicting the original token. We adopt the same masking strategy and prediction method as BERT [Devlin *et al.*, 2018].

In order to mitigate the risk of our model developing an over-dependence on particular feature types throughout the pre-training phase, we utilize a technique that actively fine-tunes the loss weights, ensuring an equitable distribution of influence across the various feature types. To evaluate the extent to which the model fits the target, following [Athalye *et al.*, 2018], we define a relative loss:

$$\tilde{\mathcal{L}}_*(t) = \mathcal{L}_*(t)/\mathcal{L}_*(0), \tag{9}$$

where $\mathcal{L}_*(t)$ and $\mathcal{L}_*(0)$ represent the loss values at time t and at the initial time 0, respectively. A smaller relative loss demonstrates a more rapid convergence of the model to the target. Inspired by [Zhao *et al.*, 2022], we dynamically adjust the loss weights with each update in relation to the relative loss to ensure that both $\mathcal{L}_c$ and $\mathcal{L}_n$ experience a relatively equitable decline throughout the updating procedure:

$$w_c(t) = \frac{m_w\left[\tilde{\mathcal{L}}_c(t)\right]^\beta}{\left[\tilde{\mathcal{L}}_n(t)\right]^\beta + \left[\tilde{\mathcal{L}}_c(t)\right]^\beta} + \left(1 - m_w\right) w_c(t-1), \tag{10}$$

$$w_n(t) = 1 - w_c(t), \tag{11}$$

where $w_*(t)$ is the loss weight at time t, β is the hyper-parameter controlling the degree of updating the loss weight, and m_w is the momentum coefficient hyper-parameter. The final AMLM loss is calculated as follows:

$$\mathcal{L}_{\text{amlm}}(t) = w_c(t)\mathcal{L}_c(t) + w_n(t)\mathcal{L}_n(t). \tag{12}$$

With the AMLM loss adjustment, VecoCare can further reduces heterogeneous semantic biases.

4.3 Dual-channel Retrieval Module

After performing the above two pre-training stages, now we obtain the contextual representations $\bar{\mathbf{H}}_i = \left[\bar{\mathbf{h}}_i^{cls}, \bar{\mathbf{h}}_i^1, \bar{\mathbf{h}}_i^2, \cdots, \bar{\mathbf{h}}_i^T\right]$ and $\bar{\mathbf{U}}_i = \left[\bar{\mathbf{u}}_i^{cls}, \bar{\mathbf{u}}_i^1, \bar{\mathbf{u}}_i^2, \cdots, \bar{\mathbf{u}}_i^M\right]$ from the pre-trained global-local fusing encoder. In order to simulate the process of a doctor recalling cases related to the current patient to assist clinical analysis in real world, `VecoCare` maintains a patient key-value memory bank, which consists of the visit sequences representations and the clinical notes representations of patients in the training set. Specifically, we denote the patient memory bank $\mathbf{G}$ as a vectorized indexable dictionary as follows:

$$\mathbf{G} = \left\{\bar{\mathbf{h}}_j^{cls} : \bar{\mathbf{u}}_j^{cls}\right\}_{j=1}^{Tr}, \tag{13}$$

where Tr is the training set size, $\mathbf{G}$ is initialized at the beginning of training and updated at each training epoch. For clarity, we use $\mathbf{G}_k = \left[\bar{\mathbf{h}}_1^{cls}; \bar{\mathbf{h}}_2^{cls}; \cdots; \bar{\mathbf{h}}_{Tr}^{cls}\right] \in \mathbb{R}^{Tr \times d}$ to denote the key vectors and $\mathbf{G}_v = \left[\bar{\mathbf{u}}_1^{cls}; \bar{\mathbf{u}}_2^{cls}; \cdots; \bar{\mathbf{u}}_{Tr}^{cls}\right] \in \mathbb{R}^{Tr \times d}$ to denote the value vectors.

For patients with limited information, a natural idea is to rely on information from other patients with similar health conditions. Nevertheless, information from other patients with extremely different or even opposite healthcare statuses can also be useful for medical prediction and treatment prognosis in healthcare [Unay and Ekin, 2011; Jia *et al.*, 2020]. We contend that conventional attention mechanisms are inadequate for modeling negative correlations. Specifically, according to [Vaswani *et al.*, 2017], an attention function can be described as mapping a query and a set of key-value pairs to an output. A typical implementation involves calculating the dot products between the query and the keys, followed by applying a softmax function to derive the weights. The limitation is that the softmax function would treat the negative correlation between query and key as inconsequential. (e.g., executing a dot product and softmax operation on two vectors with opposite directions yields the smallest weight). To address this issue, we introduce a dual-channel retrieval mechanism, comprising both positive and negative channels. For the positive channel, we use the traditional dot-produt attention [Vaswani *et al.*, 2017]. For the negative channel, we capture the negative correlation by inverting the values of the dot-produt between the key and query, and output the opposite of resulting weights after the softmax function:

$$\begin{aligned} \alpha^{pos} &= \operatorname{Softmax}\left(\bar{\mathbf{h}}_i^{cls}\mathbf{G}_k^\top\right), \\ \alpha^{neg} &= -\operatorname{Softmax}\left(-\bar{\mathbf{h}}_i^{cls}\mathbf{G}_k^\top\right), \end{aligned} \tag{14}$$

where $\alpha^{pos}, \alpha^{neg} \in \mathbb{R}^{Tr}$. In order to fully capture relationships between different perspectives, we use the weights obtained from the positive and negative channels separately to obtain the weighted sum of the values:

$$\bar{\mathbf{u}}_p = \alpha^{pos}\mathbf{G}_v, \quad \bar{\mathbf{u}}_n = \alpha^{neg}\mathbf{G}_v, \tag{15}$$

where $\bar{\mathbf{u}}_p, \bar{\mathbf{u}}_n \in \mathbb{R}^d$. Finally, we concatenate $\bar{\mathbf{h}}_i^{cls}$, $\bar{\mathbf{u}}_i^{cls}$, $\bar{\mathbf{u}}_p$, and $\bar{\mathbf{u}}_n$ as the output vector $\mathbf{r} \in \mathbb{R}^{4d}$ for the patient: $\mathbf{r} = \bar{\mathbf{h}}_i^{cls} \| \bar{\mathbf{u}}_i^{cls} \| \bar{\mathbf{u}}_p \| \bar{\mathbf{u}}_n$. We then use a simple linear layer with a sigmoid activation function on the model output $\mathbf{r}$ to calculate the predicted probability $\mathbf{y}_i'$. In this case, the cross-entropy loss is applied as the loss function:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B}\left(\mathbf{y}_i^\top \log\left(\mathbf{y}_i'\right) + \left(1-\mathbf{y}_i\right)^\top \log\left(1-\mathbf{y}_i'\right)\right), \tag{16}$$

where B is the batch size. $\mathbf{y}_i' \in [0,1]^{|\mathcal{Y}|}$ is the predicted probability, and $\mathbf{y}_i \in \{0,1\}^{|\mathcal{Y}|}$ is the ground truth.

5 Experiments

In this section, we will show the experimental results including performance comparison, ablation studies and analysis to validate the predictive power and interpretability of `VecoCare`. The source code is available at [1].

5.1 Experimental Setup

Datasets

- **MIMIC-III Dataset** We conduct diagnosis prediction on ICU data from the publicly available Medical Information Mart for Intensive Care (MIMIC-III) database [Johnson *et al.*, 2016]. In this study, we select patients with at least two visits. We use the diagnoses of the last visit as labels and incorporate the remaining visits within a 365-day prediction window as features. For the clinical notes, following [Lu *et al.*, 2021], we select the notes from the last visit in input features and filter out notes of type ”Discharge summary” for fair prediction. We study an 135-class classification problem that predicts the diseases of a patient by categorizing ICD-9 codes into 135 broader medical groups using the Unified Medical Language System [Ho *et al.*, 2014].
- **RWH Dataset.** Another dataset we use is an EHR dataset from a real-world hospital. The cleaned dataset consists of 10,408 patients with 458,952 visits. We adopt the same pre-processing method as used for the MIMIC-III dataset.

Both datasets are fully anonymized and carefully sanitized before our access. We screen patients with clinical notes information during their medical visits. For each note, we use the first 512 words, while the rest are cut off for computational efficiency.

Baselines

To compare `VecoCare` with state-of-the-art models, we select the following models as baselines:

- **T-LSTM** [Baytas *et al.*, 2017] introduces a time decay mechanism in LSTM.
- **Dipole** [Ma *et al.*, 2017] uses bidirectional RNNs and three attention mechanisms to predict patient visit information.
- **GRAM** [Choi *et al.*, 2017] utilizes medical ontologies to derive code representation learning.
- **KAME** [Ma *et al.*, 2018] predicts patients' future diagnoses based on a knowledge attention mechanism.
- **HiTANet** [Luo *et al.*, 2020] incorporates time-awareness into the self-attention component.
- **CGL** [Lu *et al.*, 2021] proposes collaborative graph learning for enhanced utilization of external medical knowledge.

[1] https://github.com/xyxpku/VecoCare

	MIMIC-III				RWH			
Methods	mi-AUROC	ma-AUROC	mi-AUPRC	ma-AUPRC	mi-AUROC	ma-AUROC	mi-AUPRC	ma-AUPRC
T-LSTM	.8817(.008)	.5452(.020)	.3880(.018)	.1229(.009)	.8030(.005)	.6092(.007)	.3068(.007)	.1260(.010)
Dipole	.8990(.003)	.6851(.014)	.4480(.009)	.1705(.007)	.8112(.004)	.7012(.006)	.3189(.012)	.1515(.007)
GRAM	.8382(.002)	.5215(.012)	.4452(.006)	.1576(.016)	.7819(.003)	.6622(.005)	.3267(.002)	.1375(.003)
KAME	.8887(.011)	.6165(.011)	.4586(.020)	.1725(.008)	.8147(.002)	.7157(.003)	.3405(.006)	.2195(.010)
HiTANet	.9057(.006)	.7176(.022)	.5195(.008)	.2011(.012)	.8229(.011)	.7527(.021)	.3639(.008)	.2231(.010)
CGL	.9081(.005)	.7201(.015)	.5298(.009)	.2114(.010)	.8862(.003)	.7809(.010)	.4133(.009)	.2725(.007)
Chet	.9136(.003)	.7309(.017)	.5372(.007)	.2328(.013)	.8921(.006)	.7913(.009)	.4179(.007)	.2803(.014)
MetaCare++	.9122(.005)	.7235(.009)	.5395(.005)	.2188(.015)	.8994(.007)	.7859(.011)	.4142(.015)	.2711(.006)
VecoCare$_{g-}$	.9097(.002)	.7284(.010)	.5264(.009)	.2239(.012)	.9048(.004)	.8180(.008)	.4219(.011)	.2956(.004)
VecoCare$_{a-}$	.8989(.001)	.6879(.008)	.5030(.002)	.1867(.005)	.8961(.017)	.7953(.029)	.3940(.066)	.2739(.053)
VecoCare$_{d-}$	.9109(.003)	.7416(.016)	.5404(.018)	.2394(.016)	.9053(.004)	.8175(.008)	.4229(.006)	.3001(.007)
VecoCare	**.9179(.001)**	**.7687(.008)**	**.5592(.005)**	**.2646(.012)**	**.9137(.007)**	**.8356(.014)**	**.4336(.013)**	**.3107(.007)**

Table 1: Results for the diagnosis prediction task on MIMIC-III and RWH dataset.

- **Chet** [Lu *et al.*, 2022] attempts to learn on dynamic disease graphs from patient visit history.
- **MetaCare++** [Tan *et al.*, 2022] introduces a clinical meta-learner to capture temporal relations among patient visits.

For a fair comparison, we take the clinical notes as additional input, and after obtaining the notes contextual representation via the base text encoder of VecoCare, we concatenate it with the patient representation obtained by baseline methods to perform the prediction. Moreover, we carefully tuned the hyper-parameters of the baselines on the validation set using the grid-search strategy to ensure their best performance. We also conduct the following ablation studies:

- VecoCare$_{g-}$ removes the GWD-based contrastive learning module from VecoCare.
- VecoCare$_{a-}$ removes the adaptive masked language model module from VecoCare.
- VecoCare$_{d-}$ removes the dual-channel retrieval module from VecoCare. It directly concatenates $\bar{\mathbf{h}}_i^{cls}$ and $\bar{\mathbf{u}}_i^{cls}$ to perform the final task.

Evaluation Metrics and Strategy

We assess the performance with four widely used evaluation metrics: micro-averaged of the area under ROC curve (mi-AUROC), macro-averaged AUROC (ma-AUROC), micro-averaged of the area under the precision-recall curve (mi-AUPRC) and macro-averaged AUPRC (ma-AUPRC). AUPRC serves as the most informative and the primary evaluation metric when handling highly imbalanced and skewed datasets like the real-world EHR data [Davis and Goadrich, 2006]. The datasets are divided into the training set, validation set, and test set with a proportion of 0.70:0.15:0.15. Both mean and standard deviation of test performance are reported.

5.2 Experimental Results

Table 1 presents the experimental results of the baseline methods as well as VecoCare on the two datasets. The number in () denotes the standard deviation. The results indicate that VecoCare exhibits a notable and persistent advantage over other baseline methods, especially ma-AUPRC, which is the most informative and the primary evaluation metric. We find that VecoCare outperforms CGL which utilizes keywords from the clinical notes to assist diagnosis prediction. It demonstrates the effectiveness of reducing heterogeneous semantic biases and utilizing auxiliary information from important similar and dissimilar patients. Besides, it is worth mentioning that VecoCare significantly outperforms GRAM, KAME, CGL, and MetaCare++ even without any task-specific external knowledge priors used in these methods. It further validates the significance of synthesizing rich information from structured and unstructured EHR data.

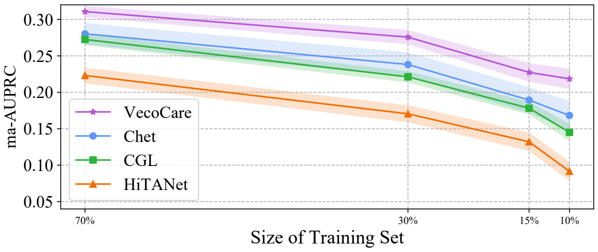

Figure 3: Performance comparison of VecoCare and several baselines with different amount of training data on RWH dataset.

We further conduct the ablation studies to examine the design of VecoCare. The superior performance of VecoCare than the VecoCare$_{g-}$ and VecoCare$_{a-}$ verifies that reducing heterogeneous semantic biases can significantly improve model performance. Moreover, VecoCare outperforms VecoCare$_{d-}$, demonstrating that it is also effective to incorporate information from other relevant patients' clinical notes.

5.3 Analysis

Robustness Against Data Insufficiency

To further explore the benefits of VecoCare in situations with limited data availability, we assess its robustness in the face of inadequate training samples. Specifically, we simulate data scarcity by reducing the RWH dataset's training set from 70% to 30%, 15%, and 10%, while maintaining a fixed test set for unbiased comparison. We conduct experiments under these conditions, and the ma-AUPRC (± std.) is plotted in Figure 3. As shown in Figure 3, VecoCare

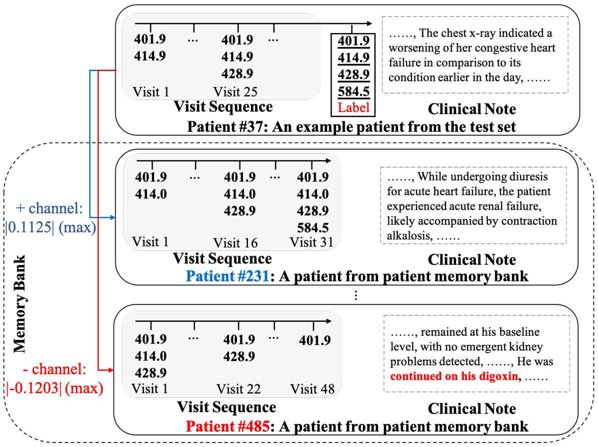
401.9
414.9
...
401.9
414.9
428.9
...
Visit 1
Visit 25
Visit Sequence
401.9
414.9
428.9
584.5
Label
......, The chest x-ray indicated a worsening of her congestive heart failure in comparison to its condition earlier in the day, ...
Clinical Note
Patient #37: An example patient from the test set
+ channel:
|0.1125| (max)
Memory Bank
- channel:
|-0.1203| (max)
401.9
414.0
...
401.9
414.0
428.9
...
401.9
414.0
428.9
584.5
Visit 1
Visit 16
Visit 31
Visit Sequence
......, While undergoing diuresis for acute heart failure, the patient experienced acute renal failure, likely accompanied by contraction alkalosis, ...
Clinical Note
Patient #231: A patient from patient memory bank
⋮
401.9
414.0
428.9
...
401.9
428.9
...
401.9
Visit 1
Visit 22
Visit 48
Visit Sequence
......, remained at his baseline level, with no emergent kidney problems detected, ..., He was continued on his digoxin, ...
Clinical Note
Patient #485: A patient from patient memory bank

Figure 4: An example from the RWH testing set showing the relevant patient information retrieval results.

consistently surpasses the chosen baseline models across all settings. Moreover, as the training set size diminishes, the performance of baseline methods declines more rapidly than that of VecoCare, resulting in a wider performance gap. Impressively, when utilizing only 10% of the data for training, VecoCare still attains a ma-AUPRC of 0.2185 on the RWH dataset, significantly outpacing the baseline methods. This highlights the excellent performance and robustness of VecoCare in scenarios with insufficient data.

Case Study for Relevant Patient Information Retrieval

Now in this section we use case studies on the RWH datasets to show the ability of VecoCare to retrieve relevant patient information. In this example, Figure 4 shows the data of a patient diagnosed with hypertension (401.9), chronic ischemic heart disease (414.9), heart failure (428.9), and acute renal failure (584.5) at the to-be-predicted visit. We also show two other patients selected in the training set from the patient memory bank $\mathbf{G}$ with the highest attention weights (absolute value) in the positive and negative channels, respectively. Although the three patients have suffered from the same hypertension disease (401.9), the disease processes of patient #231 and patient #485 are different and opposite. Specifically, we can observe that the hypertension complications of the patient from the testing set (i.e., patient #37) and patient #231 are increasing. However, patient #485 experiences a stepwise reduction in hypertensive complications. It demonstrate that VecoCare is adept at concentrating on patients with analogous disease trajectories while simultaneously paying close attention to those with vastly different or even opposing health statuses, which is helpful for the patient's future treatment prognosis. For example, doctors can refer to the treatment procedures in the clinical notes of patient #485 to mitigate the current patient's condition with digoxin.

Case Study for Attention Analysis of the Fusing Encoder

To further illustrate the interpretability and reasonableness of the proposed VecoCare, we conduct case studies to visualize and interpret the attention weights learned by the global-local fusing encoder. Figure 5 shows the data of two patients diagnosed with diabetes mellitus at the to-be-predicted visit and VecoCare successfully predicts the outcome. The average attention weights of one head calculated by the self-attention module are shown. The ordinates of the two figures are the *Query* token [CLS] $\mathbf{x}_i^{cls}$ and the abscissas are the *Key* word tokens in the clinical notes. The darker boxes mean that the patient's overall representation is more concerned with the word, and vice versa. For convenience, we visualize only the most noticed word tokens. In the first example, we can observe that VecoCare attaches the highest importance to "Nonalcoholic" and "steatohepatitis", aligning with the medical research [Loomba *et al.*, 2012]. As for another example, We can observe that VecoCare pays more attention to "family", "history", and "diabetes". According to medical research [Valdez, 2009], a family history of diabetes is a major risk factor for this disease. Thus, VecoCare can effectively capture important clinical notes context and reflect them in the form of attention weights for reliable explanations.

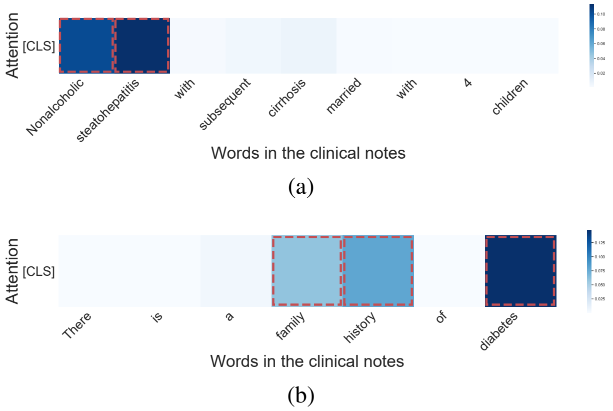

Figure 5: Attention weight visualization for two patients diagnosed with diabetes mellitus from the MIMIC-III testing set.

6 Conclusions and Future Works

In this work, we propose a novel visit sequences-clinical notes joint learning method named VecoCare, which fully considers the fusion of clinical notes with structured data for accurate diagnosis prediction. In particular, VecoCare first reduces heterogeneous semantic biases by two novel sequentially executed pre-training tasks. After that, VecoCare utilizes a dual-channel retrieval mechanism to aggregate information from both similar and dissimilar patients, thus learning a more comprehensive representation. Experimental results on two real-world EHR datasets demonstrate the clear advantages of VecoCare over the state-of-the-art baselines even without any task-specific external knowledge priors. In the meanwhile, case studies further demonstrate the robustness and novel interpretability of VecoCare. Besides, the findings are in accord with experts and medical knowledge, which shows it can provide useful insights. In the future, we will explore incorporating longer and richer clinical notes as well as other data types in EHR to bring higher performance improvements.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62172011), the National Natural Science Foundation of China (No.62102007), and the Fundamental Research Funds for the Central Universities.

References

[Athalye *et al.*, 2018] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

[Baytas *et al.*, 2017] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74, 2017.

[Bronstein *et al.*, 2010] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3594–3601. IEEE, 2010.

[Burke *et al.*, 2014] Harry B Burke, Albert Hoang, Dorothy Becher, Paul Fontelo, Fang Liu, Mark Stephens, Louis N Pangaro, Laura L Sessums, Patrick O'Malley, Nancy S Baxi, et al. Qnote: an instrument for measuring the quality of ehr clinical notes. *Journal of the American Medical Informatics Association*, 21(5):910–916, 2014.

[Choi *et al.*, 2016] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

[Choi *et al.*, 2017] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795, 2017.

[Davis and Goadrich, 2006] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Feng *et al.*, 2021] Yujie Feng, Jiangtao Wang, Yasha Wang, and Sumi Helal. Completing missing prevalence rates for multiple chronic diseases by jointly leveraging both intra- and inter-disease population health data correlations. In *Proceedings of the Web Conference 2021*, pages 183–193, 2021.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[Ho *et al.*, 2014] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 115–124, 2014.

[Jia *et al.*, 2020] Zheng Jia, Xian Zeng, Huilong Duan, Xudong Lu, and Haomin Li. A patient-similarity-based model for diagnostic prediction. *International Journal of Medical Informatics*, 135:104073, 2020.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[Loomba *et al.*, 2012] Rohit Loomba, Maria Abraham, Aynur Unalp, Laura Wilson, Joel Lavine, Ed Doo, Nathan M Bass, and Nonalcoholic Steatohepatitis Clinical Research Network. Association between diabetes, family history of diabetes, and risk of nonalcoholic steatohepatitis and fibrosis. *Hepatology*, 56(3):943–951, 2012.

[Lu *et al.*, 2021] Chang Lu, Chandan K Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. *arXiv preprint arXiv:2105.07542*, 2021.

[Lu *et al.*, 2022] Chang Lu, Tian Han, and Yue Ning. Context-aware health event prediction via transition functions on dynamic disease graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4567–4574, 2022.

[Luo *et al.*, 2020] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 647–656, 2020.

[Ma *et al.*, 2017] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911, 2017.

[Ma *et al.*, 2018] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 743–752, 2018.

[Ma *et al.*, 2020] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 833–840, 2020.

[Ma *et al.*, 2021] Liantao Ma, Xinyu Ma, Junyi Gao, Xianfeng Jiao, Zhihao Yu, Chaohe Zhang, Wenjie Ruan, Yasha Wang, Wen Tang, and Jiangtao Wang. Distilling knowledge from publicly available online emr data to emerging epidemic for prognosis. In *Proceedings of the Web Conference 2021*, pages 3558–3568, 2021.

[Ma *et al.*, 2022] Xinyu Ma, Yasha Wang, Xu Chu, Liantao Ma, Wen Tang, Junfeng Zhao, Ye Yuan, and Guoren Wang. Patient health representation learning via correlational sparse prior of medical features. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[Mémoli, 2011] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Peyré *et al.*, 2016] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

[Roostaiyan *et al.*, 2017] Seyed Mahdi Roostaiyan, Ehsan Imani, and Mahdieh Soleymani Baghshah. Multi-modal deep distance metric learning. *Intelligent Data Analysis*, 21(6):1351–1369, 2017.

[Tan *et al.*, 2022] Yanchao Tan, Carl Yang, Xiangyu Wei, Chaochao Chen, Weiming Liu, Longfei Li, Jun Zhou, and Xiaolin Zheng. Metacare++: Meta-learning with hierarchical subtyping for cold-start diagnosis prediction in healthcare data. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 449–459, 2022.

[Thapa *et al.*, 2022] Nischay Bikram Thapa, Sattar Seifollahi, and Sona Taheri. Hospital readmission prediction using clinical admission notes. In *Australasian Computer Science Week 2022*, pages 193–199. 2022.

[Unay and Ekin, 2011] Devrim Unay and Ahmet Ekin. Dementia diagnosis using similar and dissimilar retrieval items. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1889–1892. IEEE, 2011.

[Valdez, 2009] Rodolfo Valdez. Detecting undiagnosed type 2 diabetes: family history as a risk factor and screening tool. *Journal of diabetes science and technology*, 3(4):722–726, 2009.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Xu *et al.*, 2022] Yuyang Xu, Haochao Ying, Siyi Qian, Fuzhen Zhuang, Xiao Zhang, Deqing Wang, Jian Wu, and Hui Xiong. Time-aware context-gated graph attention network for clinical risk prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[Xu *et al.*, 2023] Yongxin Xu, Xu Chu, Kai Yang, Zhiyuan Wang, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. Seqcare: Sequential training with external medical knowledge graph for diagnosis prediction in healthcare data. In *Proceedings of the ACM Web Conference 2023*, pages 2819–2830, 2023.

[Yang *et al.*, 2017] Kai Yang, Xiang Li, Haifeng Liu, Jing Mei, Guotong Xie, Junfeng Zhao, Bing Xie, and Fei Wang. Tagited: Predictive task guided tensor decomposition for representation learning from electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[Yang *et al.*, 2021] Kai Yang, Zhaojing Luo, Jinyang Gao, Junfeng Zhao, Beng Chin Ooi, and Bing Xie. Lda-reg: Knowledge driven regularization using external corpora. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[Ye *et al.*, 2020] Muchao Ye, Junyu Luo, Cao Xiao, and Fenglong Ma. Lsan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1753–1762, 2020.

[Ye *et al.*, 2021] Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. Medpath: Augmenting health risk prediction via medical knowledge paths. In *Proceedings of the Web Conference 2021*, pages 1397–1409, 2021.

[Zhang *et al.*, 2021] Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. Grasp: Generic framework for health status representation learning based on incorporating knowledge from similar patients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 715–723, 2021.

[Zhang *et al.*, 2022] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2418–2428, 2022.

[Zhao *et al.*, 2022] Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 585–602. Springer, 2022.