An Iterative Self-Learning Framework for Medical Domain Generalization

Zhenbang Wu [1] **Huaxiu Yao** [2] **David M Liebovitz** [3] **Jimeng Sun** [1]
[1] University of Illinois Urbana-Champaign, {zw12, jimeng}@illinois.edu
[2] University of North Carolina at Chapel Hill, huaxiu@cs.unc.edu
[3] Northwestern University, david.liebovitz@nm.org

Abstract

Deep learning models have been widely used to assist doctors with clinical decision-making. However, these models often encounter a significant performance drop when applied to data that differs from the distribution they were trained on. This challenge is known as the domain shift problem. Existing domain generalization algorithms attempt to address this problem by assuming the availability of domain IDs and training a single model to handle all domains. However, in healthcare settings, patients can be classified into numerous latent domains, where the actual domain categorizations are unknown. Furthermore, each patient domain exhibits distinct clinical characteristics, making it sub-optimal to train a single model for all domains. To overcome these limitations, we propose SLDG, a self-learning framework that iteratively discovers decoupled domains and trains personalized classifiers for each decoupled domain. We evaluate the generalizability of SLDG across spatial and temporal data distribution shifts on two real-world public EHR datasets: eICU and MIMIC-IV. Our results show that SLDG achieves up to 11% improvement in the AUPRC score over the best baseline.

1 Introduction

Deep learning techniques have been increasingly popular in clinical predictive modeling with electronic health records (EHRs) [12, 11, 47, 58, 2]. However, these models typically assume that the training (source) data and testing (target) data share the same underlying data distribution (i.e., domain). This assumption can become problematic when models are applied to new domains, such as data from different hospitals or future time points [17, 59, 20, 37]. In these situations, domain shifts caused by variations in patient cohorts, clinical standards, and terminology adoption can significantly degrade the model's performance.

This paper aims to develop a clinical predictive model on the source data that effectively handles potential domain shifts when applied to the target data. Domain generalization (DG) [7] methods have been widely utilized to address such problems, including techniques like domain alignment [32, 24, 25, 45, 31, 51, 62], meta-learning [23, 27, 26, 5, 22, 29], and ensemble learning [9, 42, 43, 61]. However, when applied in healthcare settings, these methods encounter the following limitations:

- **Reliance on domain IDs.** Most DG methods depend on the presence of domain IDs, which indicate the domain to which each sample belongs, to guide the model training [24, 16, 4, 9, 42]. However, as shown in Fig. 1, patients can be divided into numerous latent domains based on features such as age, medical history, treatment, and symptoms. The actual categorization of these latent domains can be difficult to obtain and vary across different tasks [1, 49]. Consequently, existing DG methods often resort to broad domain categorizations, such as hospital or timestamp, which provide limited information [59, 17].

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

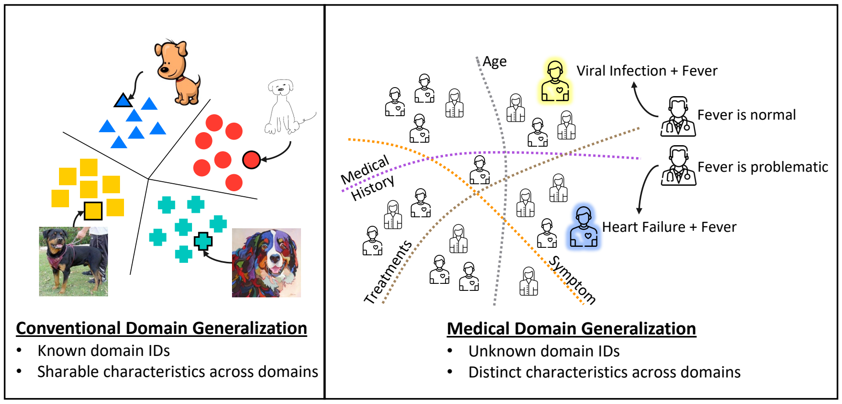

Figure 1: Conventional domain generalization methods typically rely on domain IDs and shared characteristics across domains to train a single generalized model. However, in the medical field, patients can be classified into numerous latent domains that are not directly observable. Additionally, each patient domain exhibits unique clinical characteristics, making it sub-optimal to train a single model for all domains.

- **Attempt to train a single model.** While some recent DG methods have attempted to alleviate the reliance on domain labels [62, 29], they try to train a single model that generalizes across all domains. However, patients from different domains possess distinct characteristics and require different treatment approaches [2, 58]. For example, as shown in Fig. 1, fever is considered a normal symptom for patients with viral infections as it helps stimulate the immune system. On the other hand, it can be a bad signal for patients with cardiovascular disease, leading to complications. Thus, training a single model for all domains is challenging and can lead to sub-optimal performance.

To overcome these limitations, we propose SLDG, a self-learning framework for domain generalization that iteratively discovers decoupled domains and trains customized classifiers for each discovered domain. Specifically, SLDG consists of the following iterative steps:

- **Decoupled domain discovery.** While domain labels are not initially available, we posit that they can be recovered by clustering the learned latent representations. However, identifying all fine-grained domains across various clinical features (e.g., demographics, diagnosis, and treatments) can be challenging. Instead, we propose to decouple these clinical features and discover the clusters separately for each type of feature. To achieve this, we maintain a distinct latent space for each type of features using a feature-specific patient encoder. Within each latent space, we perform hierarchical clustering independently to discover the domain categorizations. By adopting this approach, we effectively reduce the number of domains from exponential to linear to the number of feature types.

- **Domain-specific model customization.** To account for the unique characteristics of patients in different domains, our approach involves training customized classifiers for each domain. To ensure parameter efficiency, we extract domain representations from the learned clusters and utilize them to parameterize the domain-specific classifiers. For a given patient, we determine the closest domain by comparing the patient's representations with the domain representations, and subsequently select the corresponding classifier for accurate inference.

To assess the generalizability of SLDG across spatial and temporal shifts, we conduct experiments on two publicly available EHR datasets: eICU [38] and MIMIC-IV [18]. Our results demonstrate that SLDG outperforms the best baseline by up to 11% in terms of AUPRC score. We also conduct detailed analyses and ablation studies to investigate the factors contributing to the performance gain achieved by SLDG.

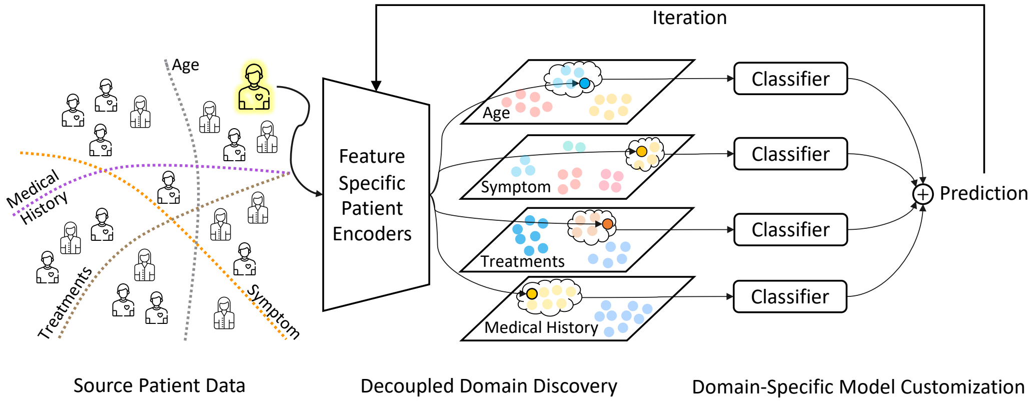

Figure 2: An illustration of the SLDG framework. The feature-specific patient encoder maps each patient into multiple latent spaces, with each space capturing patient characteristics from a specific perspective. Next, SLDG iteratively performs decoupled clustering to identify latent domains and learns domain-specific classifiers customized for each domain.

2 Preliminaries

In EHR data, a patient's hospital visit is represented by a sequence of events, denoted as $x = [e_1, e_2, \ldots, e_m]$, where m is the total number of events in the visit. Each event e characterizes features of a certain type t, such as diagnosis, prescription, and lab tests. This mapping is denoted by a function $T(e) : \mathcal{E} \rightarrow \mathcal{T}$, where $\mathcal{E}$ and $\mathcal{T}$ denote the sets of all events and types, respectively. For example, a patient visit can be [Acute embolism (I82. 40), Atrial fibrillation (I48.91), Ultrasound (76700), CT scan (G0296), ECG (93042), Heparin IV (5224)]. And each event corresponds to a specific type (e.g., diagnosis, procedure, medication). The main objective of clinical predictive modeling is to predict the occurrence of future events, such as 15-day hospital readmission and 90-day mortality, denoted as $y \in \{+, -\}$, based on the patient's current visit x.

Existing clinical prediction works typically train a model $f_\phi(\cdot)$ with parameter ϕ by minimizing a loss function $l(\cdot)$ on source training data sampled from distribution P_{tr}, as in Eq. (1),

$$\arg\min_{\phi} \mathbb{E}_{(x,y)\sim P_{tr}}[l(f_\phi(x), y)], \tag{1}$$

with the hope that the trained model can perform well on the target test data distributed according to P_{te}. However, in real-world settings, the source and target distributions can differ due to spatial and temporal shifts, i.e., $P_{tr} \neq P_{te}$. Consequently, the model trained on source data may experience a drop in performance when applied to the target data.

3 The SLDG Approach

In this paper, our goal is to train a model $f_\phi(\cdot)$ on source data P_{tr} that can generalize to target data P_{te} despite potential domain shifts. Existing DG algorithms face limitations due to their reliance on domain IDs and attempts to train a single model for all domains. To overcome these limitations, we propose to iteratively discover latent domains and train customized classifiers for each domain. However, we face the challenge of dealing with a large number of latent domains, which not only makes domain discovery difficult but also results in an exponential increase in the number of model parameters with respect to the number of feature types. In the following sections, we will describe how our method SLDG addresses this challenge through decoupled domain discovery and domain-specific model customization. Additionally, we will introduce the training and inference strategy. Fig. 2 illustrates the SLDG framework.

3.1 Decoupled Domain Discovery

Although domain labels are not initially available, we hypothesize that the domain information is encoded in the learned latent representations and can be recovered with the clustering technique. However, patients can be categorized into thousands of latent domains, determined by various features such as age, medical history, treatment, and symptoms. For instance, a patient can fall into the fine-grained domain of *older male patients with a history of smoking and a diagnosis of type 2 diabetes*. Identifying all such fine-grained domains can be challenging, as clustering methods may either overlook smaller domains or result in an excessive number of domains that would inflate the number of parameters in subsequent steps. To address this, we propose decoupling these clinical features and independently discovering clusters for each feature type. For example, the patient above can simultaneously belong to the decoupled domains of *older*, *male*, *history of smoking*, and *diagnosis of type 2 diabetes*. This approach effectively reduces the number of domains from exponential to linear with respect to the number of feature types.

Concretely, we maintain a distinct latent space for each type of feature. When given an input patient visit x, `SLDG` maps it to the latent space corresponding to the feature type $t \in \mathcal{T}$ using a feature-specific patient encoder $E_t(\cdot)$, as in Eq. (2),

$$\mathbf{h}_t \coloneqq E_t(x), \;\; \mathbf{h}_t \in \mathbb{R}^h, \tag{2}$$

where h denotes the hidden dimension. Next, within each latent space of type t, `SLDG` gathers all patient representations $\{\mathbf{h}_t^{(i)}\}_{i=1}^{N_{tr}}$, where N_{tr} is the number of source training data, and performs clustering to discover the domain categorizations, as in Eq. (3),

$$\mathbf{M}_t \coloneqq \text{Cluster}(\{\mathbf{h}_t^{(i)}\}_{i=1}^{N_{tr}}), \;\; \mathbf{M}_t \in \{0,1\}^{N_{tr} \times K_t}, \tag{3}$$

where K_t represents the number of discovered domains in the latent space of type t. $\mathbf{M}_t$ denotes the learned domain assignment, where $\mathbf{M}_t[i,k]$ is equal to one if and only if (i.f.f.) the patient $x^{(i)}$ is assigned to the k-th domain. We will describe this procedure in detail in the following.

Feature-Specific Patient Encoding. This module is responsible for mapping each patient into multiple latent spaces, each capturing the patient's health status of a specific feature type. This enables subsequent modules to decouple the representations of different feature types. For a patient's hospital visit x with a list of events $[e_1, \ldots, e_m]$, `SLDG` computes the contextualized representation for each event by applying the embedding function $E(\cdot)$, as in Eq. (4),

$$[\mathbf{e}_1, \ldots, \mathbf{e}_m] = E([e_1, \ldots, e_m]), \;\; \mathbf{e}_j \in \mathbb{R}^h, \tag{4}$$

where $\mathbf{e}_j$ is the contextualized representation for event e_j with dimension h. We model $E(\cdot)$ using a three-layer Transformer [48] framework. To ensure that there are no unseen events in the target data, we initialize the event embedding look-up table with ClinicalBERT [3] embeddings of the event name and then project it down to our hidden dimension of size h. The embedding look-up table is fixed during training.

Next, `SLDG` aggregates the contextualized event representations $[\mathbf{e}_1, \ldots, \mathbf{e}_m]$ based on their types, such as family history, diagnosis, and treatments. For each type $t \in \mathcal{T}$, the type-specific representation $\mathbf{h}_t$ is computed by averaging the representations of all events of that type, as in Eq. (5),

$$\mathbf{h}_t = \text{Average}(\{\mathbf{e}_j \mid T(e_j) = t\}_{j=1}^{m}), \;\; \mathbf{h}_t \in \mathbb{R}^h, \tag{5}$$

where $T(e_i)$ indicates the type of event e_i. If no events belong to a certain type, the pooled sequence representation is used as a substitute. Consequently, each patient's hospital visit is represented by a set of vectors $\{\mathbf{h}_t\}_{t \in \mathcal{T}}$, with each vector capturing the patient's health status from a specific type of events. These decoupled patient representations are then utilized to perform per-feature-type domain clustering, described next.

Hierarchical Domain Clustering. This module is responsible for clustering patient representations in each latent space to discover latent domains, enabling subsequent modules to customize the classifier for each domain. In the previous step, we obtain a set of patient representations $\{\mathbf{h}_t^{(i)}\}_{i=1}^{N_{tr}}$ for each latent space of type t. To perform clustering, standard clustering techniques such as k-Means and Gaussian Mixture Model (GMM) require specifying the number of clusters, which is less ideal as the number of clusters can be difficult to choose and may vary across latent spaces. Inspired by GEORGE [46], we adopt a fully automated hierarchical clustering technique by monitoring the Silouette score [40].

Specifically, in each latent space, SLDG first applies UMAP [30] for dimensionality reduction. Then, it runs k-Means with $k \in \{2, \dots, 10\}$ to identify the optimal number of clusters based on the highest Silhouette score. Subsequently, SLDG further split each cluster into five sub-clusters. However, only sub-clusters surpassing the Silhouette score of the original cluster and containing at least 500 patients are retained. The final number of clusters in the latent space of type t is denoted as K_t. The cluster assignment is represented by a binary matrix $\mathbf{M}_t$ of size $N_{tr} \times K_t$, where $\mathbf{M}_t[i, k]$ is set to one i.f.f. the patient $x^{(i)}$ is assigned to the k-th cluster. This automated approach allows us to effectively select the number of clusters in each latent space, balancing between discovering overly coarse or fine-grained clusters.

3.2 Domain-Specific Model Customization

To accommodate the unique characteristics of patients in different domains, we propose to train customized classifiers for each decoupled domain. Given an input patient visit x and its multi-vector representations $\{\mathbf{h}_t\}_{t \in \mathcal{T}}$, SLDG computes the predicted probability o of a specific event occurring by employing a weighted combination of domain-specific classifiers in each latent space $t \in \mathcal{T}$, as in Eq. (6),

$$o := \frac{1}{|\mathcal{T}|} \sum\nolimits_{t \in \mathcal{T}} \sum\nolimits_{k=1}^{K_t} \underbrace{G_{t,k}(\mathbf{h}_t)}_{\text{gate}} \cdot \underbrace{C_{t,k}(\mathbf{h}_t)}_{\text{classifier}}, \;\; o \in \mathbb{R}, \tag{6}$$

where $C_{t,k}(\cdot)$ refers to the customized classifier for the discovered domain k in the latent space of type t, while $G_{t,k}(\cdot)$ corresponds to the gating function. In the following, we will elaborate on how SLDG leverages the clustering results to efficiently parameterize the domain-specific classifier and effectively determine the gating weights.

To efficiently parameterize the domain-specific classifier $C_{t,k}(\cdot)$ for the k-th discovered domain in the latent space of type t, we define two learnable weight vectors of size h: $\mathbf{w}^{+}_{t,k}$ and $\mathbf{w}^{-}_{t,k}$, which represent the prototypes of the positive and negative classes, respectively. The predicted probability of a specific event occurring is computed based on the relative distance between the patient representation $\mathbf{h}_t$ and the positive and negative prototypical weights, as in Eq. (7),

$$C_{t,k}(\mathbf{h}_t) = \frac{\exp(-d(\mathbf{w}^{+}_{t,k}, \mathbf{h}_t))}{\sum_{* \in \{+,-\}} \exp(-d(\mathbf{w}^{*}_{t,k}, \mathbf{h}_t))}, \;\; C_{t,k}(\mathbf{h}_t) \in \mathbb{R}, \tag{7}$$

where $d(\cdot, \cdot)$ is the Euclidean distance. To facilitate efficient learning, we initialize the two prototypical weight vectors $\mathbf{w}^{+}_{t,k}$ and $\mathbf{w}^{-}_{t,k}$, with the average representations of patients from the corresponding classes, as in Eq. (8),

$$\text{Init}(\mathbf{w}^{*}_{t,k}) = \text{Average}(\{\mathbf{h}^{(i)}_t \mid (M_t[i,k] = 1) \wedge (y^{(i)} = *)\}_{i=1}^{N_{tr}}), \;\; * \in \{+, -\}, \tag{8}$$

where $M_t[i,k] = 1$ includes only patients assigned to the k-th domain in the latent space of type t.

We adopt a similar approach to the gating function $G_t(\cdot)$. For each discovered domain k in the latent space of type t, we introduce a learnable prototypical weight vector $\mathbf{w}_{t,k} \in \mathbb{R}^h$. The gating weights are determined based on the distance between the patient representation $\mathbf{h}_t$ and the corresponding prototypical weights, as in Eq. (9),

$$G_t(\mathbf{h}_t) = \text{Softmax}(\{-d(\mathbf{w}_{t,k}, \mathbf{h}_t)\}_{k=1}^{K_t}), \;\; G_t(\mathbf{h}_t) \in \mathbb{R}^{K_t}, \tag{9}$$

where the prototypical weight vectors $\{\mathbf{w}_{t,k}\}_{k=1}^{K_t}$ are initialized as the average representations of patients in that domain, as in Eq. (10),

$$\text{Init}(\mathbf{w}_{t,k}) = \text{Average}(\{\mathbf{h}^{(i)}_t \mid M_t[i,k] = 1\}_{i=1}^{N_{tr}}), \;\; k = 1, \dots, K_t. \tag{10}$$

3.3 Training and Inference

To train SLDG, we begin by utilizing a pre-trained patient encoder [1] for decoupled domain discovery. Then, we iteratively update the model weights and re-generate the clusters every 20 epochs. In each

[1] In practice, we pre-train the patient encoder on the same clinical predictive task for 40 epochs. This means that SLDG undergoes a total of 100 epochs of training, which is consistent with other baselines.

iteration, we re-initialize the classifier and gating parameters. This iterative process is repeated three times to enhance the model's performance. During training, we minimize the binary cross-entropy loss. For inferencing, given a target patient visit, SLDG first maps it to multiple decoupled latent spaces with the feature-specific patient encoders $E_t(\cdot)$. Subsequently, in each latent space of type t, the gating function $G_t(\cdot)$ determines the weight combinations used to aggregate the predictions from domain-specific classifiers $\{C_{t,k}(\cdot)\}_{k=1}^{K_t}$. The final prediction is obtained by averaging the predictions from all latent domains. The pseudocode of SLDG can be found in Appx. A.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate SLDG on two publicly available real-world EHR datasets: eICU [38] and MIMIC-IV [18], which are described as follows:

- **eICU** [38] covers over 200K visits for 139K patients admitted to the intensive care unit (ICU) in one of the 208 hospitals across the United States. The data was collected between 2014 and 2015. The 208 hospitals can be further categorized into four groups based on their location (Midwest, Northeast, West, and South). We use age, gender, and ethnicity as patient demographic information, and leverage the diagnosis, treatment, medication, and lab tables to gather patient visit information.
- **MIMIC-IV** [18] covers over 431K visits for 180K patients admitted to the ICU in the Beth Israel Deaconess Medical Center. The data was collected between 2008 to 2019. The approximate actual year of each admission is revealed as one of the four-year groups (2008-2010, 2011-2013, 2014-2016, and 2017-2019). We use age, gender, and ethnicity as patient demographic information, and leverage diagnoses, procedures, and prescriptions to gather patient admission information.

We elaborate on the cohort selection process and provide comprehensive dataset statistics in Appx. B.1. In the end, we extract 149227 visits from 116075 patients in the eICU dataset, and 353238 visits from 156549 patients in the MIMIC-IV dataset.

Clinical Predictive Tasks. We focus on two common clinical predictive tasks: (1) Readmission prediction, which aims to determine whether a patient will be readmitted within the next 15 days following discharge. (2) Mortality prediction, which aims to predict whether a patient will pass away upon discharge in the eICU setting, or within 90 days after discharge in the MIMIC-IV setting. A detailed explanation of this setting can be found in Appx. B.2.

Data Split. We evaluate the performance of our model across spatial gaps using the eICU dataset. For this purpose, we select the target testing data as the group (Midwest) that demonstrated the largest performance gap in a pilot study. The remaining groups (Northeast, West, and South) are used as the source training data. To assess the model's performance across temporal gaps, we utilize the MIMIC-IV dataset. Patients admitted after 2014 are used as the target testing data, while all preceding patients are included in the source training data. We elaborate more on the data split in Appx. B.3.

Baselines. We compare SLDG against three categories of baselines. (1) The first category consists of naive baselines, including **Oracle**, trained directly on the target data, and **Base**, trained solely on the source data. (2) The second category comprises DG methods that require domain IDs. These include **DANN [16]** and **MLDG [22]**, which use coarse regional and temporal groups as the domain definition, and **ManyDG [56]**, which treats each patient as a unique domain. (3) The last category consists of DG methods that do not rely on domain IDs, including **IRM [4]**, **MMLD [29]**, and **DRA [14]**. A detailed explanation of all the baselines can be found in Appx. B.4.

Evaluation Metrics. Both readmission prediction and mortality prediction are binary classification tasks. To evaluate the performance of the models, we calculate the Area Under the Precision-Recall Curve (AUPRC) and the Area Under the Receiver Operating Characteristic Curve (AUROC) scores. For each metric, we report the average scores and standard deviation by performing bootstrapping

Table 1: Results of domain generalization on the eICU and MIMIC-IV datasets. An asterisk (*) indicates that SLDG achieves a significant improvement over the best baseline method, with a p-value smaller than 0.05. The experimental results demonstrate that SLDG exhibits robustness against spatial (eICU) and temporal (MIMIC-IV) domain shifts.

Method	eICU				MIMIC-IV			
	Readmission		Mortality		Readmission		Mortality	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
Oracle	0.219 (0.01)	0.677 (0.01)	0.271 (0.01)	0.839 (0.01)	0.282 (0.01)	0.693 (0.00)	0.428 (0.00)	0.898 (0.01)
Base	0.104 (0.02)	0.510 (0.01)	0.230 (0.01)	0.803 (0.01)	0.237 (0.01)	0.665 (0.01)	0.374 (0.01)	0.861 (0.00)
DANN	0.135 (0.01)	0.538 (0.01)	0.245 (0.01)	0.808 (0.01)	0.247 (0.01)	0.673 (0.01)	0.380 (0.02)	0.873 (0.02)
MLDG	0.104 (0.01)	0.525 (0.01)	0.224 (0.01)	0.797 (0.01)	0.205 (0.01)	0.637 (0.02)	0.360 (0.01)	0.857 (0.01)
ManyDG	0.150 (0.01)	0.549 (0.01)	0.259 (0.01)	0.814 (0.01)	0.249 (0.01)	0.676 (0.01)	0.388 (0.01)	0.880 (0.01)
IRM	0.136 (0.01)	0.538 (0.01)	0.252 (0.02)	0.811 (0.01)	0.242 (0.00)	0.668 (0.01)	0.387 (0.01)	0.876 (0.01)
MMLD	0.167 (0.01)	0.578 (0.00)	0.256 (0.01)	0.818 (0.01)	0.250 (0.02)	0.679 (0.01)	0.393 (0.01)	0.887 (0.01)
DRA	0.148 (0.01)	0.551 (0.01)	0.249 (0.01)	0.810 (0.01)	0.246 (0.01)	0.670 (0.01)	0.387 (0.01)	0.875 (0.01)
SLDG	**0.186 (0.01)***	**0.623 (0.01)***	**0.268 (0.01)***	**0.824 (0.01)***	**0.274 (0.01)***	**0.690 (0.01)***	**0.416 (0.00)***	**0.899 (0.01)***

1000 times. Additionally, we conduct independent two-sample t-tests to assess whether SLDG achieves a significant improvement over the baseline methods.

Implementation Details. For all baselines, we use the same Transformer [48] architecture as the backbone encoder. Patient demographics features (age, gender, and ethnicity) are embedded with an embedding look-up table. We also embed the timestamps with sinusoidal positional encoding. The medical, patient demographics, and temporal embeddings are added together to form the overall sequence embedding. All models are trained for 100 epochs, and the best model is selected based on the AUPRC score monitored on the source validation set. For SLDG, UMAP [30] from UMAP-learn [41] is used with 2 components, 10 neighbors, and 0 minimum distance; and k-Means from Scikit-learn [35] is used with the default hyper-parameter. Further information regarding the detailed implementations can be found in Appx. B.5.

4.2 Main Results

Table 1 presents the domain generalization results on the eICU [38] and MIMIC-IV [18] datasets. Firstly, we observe a significant performance gap between the Oracle and Base methods, indicating the presence of substantial spatial and temporal domain gaps. This supports the use of the DG setting. Notably, the readmission tasks exhibit larger domain gaps, which is reasonable since hospitals across different locations and timestamps may have varying criteria for patient readmission. Secondly, we note that the two DG methods, DANN [16] and MLDG [22], utilizing coarse domain partitions such as region and timestamp, achieve minimal or no improvements. This outcome is expected because the domain partitions are too coarse, making it challenging to identify consistent domain features. In comparison, ManyDG [56] achieves better performance by considering each individual patient as a unique domain. Among the remaining three baseline methods that do not rely on domain IDs, IRM [4] demonstrates the slightest improvement. DRA [14] performs better due to the usage of multi-head networks, which share a similar intuition as SLDG. MMLD [29] attains the highest performance among all baselines, showcasing the advantages of explicit domain discovery. Lastly, SLDG outperforms baselines for all tasks. Specifically, in terms of the AUPRC score, SLDG achieves an 11% relative improvement in eICU readmission prediction, 3% in eICU mortality prediction, 10% in MIMIC-IV readmission prediction, and 6% in MIMIC-IV mortality prediction.

4.3 Quantitative Analysis

This section provides quantitative analyses to elucidate the performance enhancements achieved by SLDG. The analyses encompass the evaluation of clustering results, ablation studies on the clustering algorithm, the impact of the number of clusters and iterations, and a runtime comparison.

Evaluation of clustering results. First, we evaluate the domain recovery ability of DG methods that do not rely on domain IDs, namely MMLD [29], DRA [14], and the proposed SLDG. Since the actual latent domain categorizations are unavailable, we assess the separability of the learned clustering results with the Silhouette score [40]. Note that the reported Silhouette score is calculated on the testing set, while the hyper-parameters are chosen based on the Silhouette score on the training set. As depicted in Fig. 3, DRA achieves the lowest score, which aligns with expectations as it solely learns latent domain categorizations through multi-head networks without explicit clustering. In contrast, MMLD generates more distinct clusters due to its iterative clustering and training setup. However, it still necessitates manual specification of the number of clusters. In comparison, SLDG obtains the highest score using an automated hierarchical clustering technique.

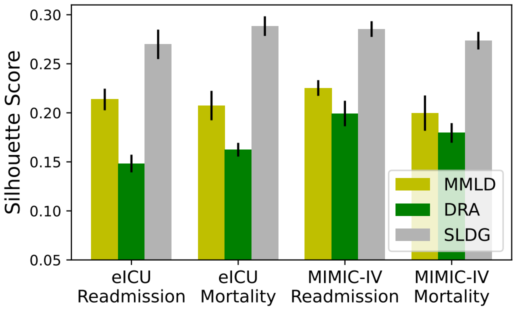

Figure 3: Performance of latent domain clustering. A higher Silhouette score indicates improved cluster separability.

Influence of the clustering algorithm. Next, we assess the influence of different clustering algorithms. Naive k-Means and GMM require manual specification of the number of clusters, which we set to the same value as SLDG's. The results can be found in Tab. 2. We observe that naive k-Means and GMM achieve similar perform similarly to the best baseline methods in Tab. 1. This outcome is reasonable since the success of SLDG relies on both the accurate discovery of latent domains and customized models for each domain. Naive clustering techniques often fail to identify subtle yet important latent domains. In contrast, SLDG, utilizing the automatic hierarchical clustering technique, achieves the highest score.

Table 2: Ablation study on the influence of the clustering algorithm.

Method	**eICU**				**MIMIC-IV**			
	Readmission		**Mortality**		**Readmission**		**Mortality**	
	AUPRC	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**
SLDG + k-Means	0.148 (0.01)	0.553 (0.01)	0.249 (0.01)	0.814 (0.00)	0.250 (0.01)	0.670 (0.01)	0.388 (0.01)	0.886 (0.01)
SLDG + GMM	0.143 (0.01)	0.549 (0.01)	0.240 (0.01)	0.808 (0.01)	0.250 (0.01)	0.688 (0.01)	0.390 (0.00)	0.888 (0.00)
SLDG	**0.186 (0.01)***	**0.623 (0.01)***	**0.268 (0.01)***	**0.824 (0.01)***	**0.274 (0.01)***	**0.690 (0.01)***	**0.416 (0.01)***	**0.899 (0.01)***

Influence of the number of clusters and iterations. Next, we analyze the impact of the number of clusters and iterations on the eICU readmission prediction task. We also compare our results with the best-performing baseline, MMLD [29]. The results can be found in Fig. 4. The upper panel of the figure shows that the model's performance initially improves with an increasing number of clusters. This improvement can be attributed to the finer granularity of clustering, which enables better identification of domains and customization of experts. However, as the number of clusters continues to increase, the model's performance starts to decline. This decline is caused by the growing number of model parameters, making training more challenging and leading to overfitting on suspicious samples. Similarly, the trend observed in the lower panel of the figure for the number of iterations aligns with the number of clusters. The performance initially improves as

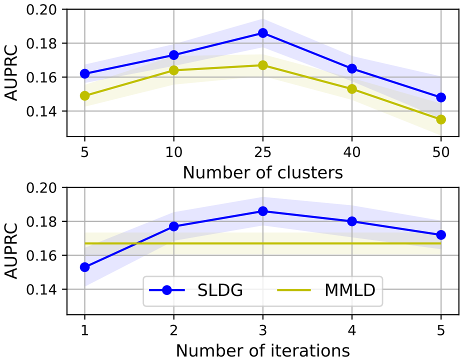

Figure 4: Results on the influence of the number of clusters and iterations.

the number of iterations increases, allowing the model to learn more from the data. However, after a certain point, the model starts overfitting on specific clusters, leading to decreased performance.

Runtime comparison. Lastly, we compare the training time of SLDG with the naive Base baseline. All runtimes are measured on a single NVIDIA A6000 GPU. The results can be found in Tab. 3. The use of the UMAP [30] dimensionality reduction technique enables SLDG to perform clustering quickly. As a result, the training time overhead of SLDG is reasonably low overall (18% on eICU and 20% on MIMIC-IV) compared to the significant performance improvement achieved (up to 79% relative improvement on AUPRC score on eICU and up to 15% on MIMIC-IV).

Table 3: Runtime comparison.

Method	**eICU**	**MIMIC-IV**
Base	67 min	93 min
SLDG	79 min	112 min

4.4 Case Study

Latent Space	**Frequent Clinical Events from the Top-2 Clusters**
Diagnosis	Sepsis, Infection, Kidney failure
	Cardiac Arrest, Congestive Heart Failure
Treatment	Heart valve procedures, Cardiovascular monitoring
	Ventilation, Respiratory intubation
Medication	Propofol, Lorazepam, Fentanyl
	Amiodarone, Noradrenaline

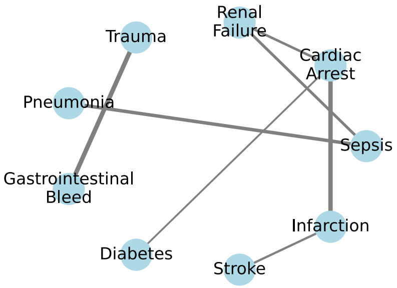

Figure 5: Left: Common clinical events observed in two largest domains identified from each latent space. Right: Learned similarity among the identified domains within the diagnostic latent space.

For the case study, we first examine the recovered domains within each latent space. The results are presented in the left panel of Fig. 5. It is evident that the common events identified within the same domains are consistent. For instance, *propofol*, *lorazepam*, and *fentanyl* is frequently used together, serving the purpose of anesthesia (pre-surgery), sedation (in-surgery), and pain management (post-surgery). Furthermore, the right panel of Figure 5 illustrates a visualization of the learned domain similarity, i.e., the distances between the domain prototypical weights $\{\mathbf{w}_{t,k}\}_{k=1}^{K_t}$. Notably, our method (SLDG) successfully captures meaningful relationships among the latent domains. For instance, a strong relationship is learned between *pneumonia* and *sepsis*. In clinical practice, when pneumonia is severe or if the infection spreads beyond the lungs, it can enter the bloodstream and trigger a systemic response, leading to sepsis [8]. Another example is the observed strong association between *cardiac arrest* and *infarction*. In practice, if a significant portion of the heart muscle is damaged, it can disrupt the heart's electrical system, potentially leading to cardiac arrest [19].

5 Related Work

Domain Generalization The goal of DG is to learn a model using data from multiple source domains in order to achieve effective generalization to a distinct target domain [7]. To achieve this, domain alignment approaches try to match the feature distributions among multiple source domains with techniques such as moments minimization [32, 25], contrastive learning [31], adversarial learning [25, 45], regularizers [4, 24], and augmentation [62, 50, 44, 60]. Meta-learning frameworks have also been utilized to simulate new domain scenarios during training [23, 27, 26, 5, 22, 29]. Additionally, domain-specific model ensemble techniques have been employed [9, 42, 43, 61]. However, these conventional DG methods assume the availability of domain IDs, which may not be feasible in healthcare settings where patients can belong to numerous unobserved domains.

Recent advancements in DG have attempted to alleviate the reliance on domain IDs [29, 62, 14, 10, 33]. MMLD [29] is the most relevant prior work to ours. It simultaneously discovers latent domains and learns domain-invariant features through adversarial learning. However, it focuses on training a single model that generalizes across all domains. Given that patients from different domains exhibit distinct characteristics and require different treatment approaches [2, 58], training a single model for all domains poses challenges and can result in sub-optimal performance.

Clinical Predictive Modeling. The main objective of clinical predictive modeling is to predict the occurrence of future events, such as 15-day hospital readmission and 90-day mortality, based on existing patient information. Deep learning models have been widely used in clinical predictive modeling with EHR data [55, 57]. These models are designed to capture temporal patterns in patient data [11, 36, 6, 28], model structural information in medical codes [13, 52], augment the model using pre-training [39], or leverage patient similarities for better decision making [58, 2]. However, these models typically assume an unchanged test domain and may suffer from degraded performance with domain shift. To address this issue, AutoMap [53] solves the feature space shift issue by learning an auto-mapping function without considering any distribution shift. MedLink [54] aggregates de-identified patient data from different sites to enable joint training. ManyDG [56] tackles patient covariate shift by treating each patient as a unique domain and disentangling domain variant and invariant features. However, maintaining a large number of domains is unnecessary, as similar patients often exhibit similar clinical behavior and can share a common domain.

6 Conclusion

Clinical predictive models often exhibit degraded performance when applied to data from new regions or future periods due to distribution shifts. To address this, we propose SLDG, a self-learning framework that iteratively identifies decoupled domains and trains customized classifiers for each domain. We evaluate SLDG on two medical datasets, and our results show that it outperforms all baseline methods. In addition, we provide detailed qualitative analyses and case studies to support our findings.

Acknowledgments

This work was supported by NSF award SCH-2205289, SCH-2014438, and IIS-2034479. This project has been funded by the Jump ARCHES endowment through the Health Care Engineering Systems Center.

References

[1] Davies Adeloye, Stephen Chua, Chinwei Lee, Catriona Basquill, Angeliki Papana, Evropi Theodoratou, Harish Nair, Danijela Gasevic, Devi Sridhar, Harry Campbell, Kit Yee Chan, Aziz Sheikh, and Igor Rudan. Global and regional estimates of COPD prevalence: Systematic review and meta-analysis. *Journal of global health*, 5(2):020415, December 2015. ISSN 2047-2978 2047-2986. doi: 10.7189/jogh.05.020415. Place: Scotland.

[2] Henrique Aguiar, Mauro Santos, Peter Watkinson, and Tingting Zhu. Learning of cluster-based feature importance for electronic health record time-series. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 161–179. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/aguiar22a.html`.

[3] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL `https://aclanthology.org/W19-1909`.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[5] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/647bba344396e7c8170902bcf2e15551-Paper.pdf`.

[6] 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*, KDD '17, page 65–74, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3097997. URL `https://doi.org/10.1145/3097983.3097997`.

[7] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL `https://proceedings.neurips.cc/paper_files/paper/2011/file/b571ecea16a9824023ee1af16897a582-Paper.pdf`.

[8] Klaus F. Bodmann. Current guidelines for the treatment of severe pneumonia and sepsis. *Chemotherapy*, 51(5):227–233, Aug 2005. doi: 10.1159/000087452. URL `https://doi.org/10.1159/000087452`.

[9] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22405–22418. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/bcb41ccdc4363c6848a1d760f26c28a0-Paper.pdf`.

[10] Zhihong Chen, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Feiyue Huang, and Xinyu Jin. Generalizable Representation Learning for Mixture Domain Face Anti-Spoofing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1132–1139, May 2021. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i2.16199. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16199`.

[11] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3512–3520, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[12] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 08 2016. ISSN 1067-5027. doi: 10.1093/jamia/ocw112. URL `https://doi.org/10.1093/jamia/ocw112`.

[13] 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*, KDD '17, page 787–795, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098126. URL `https://doi.org/10.1145/3097983.3098126`.

[14] Lucas Deecke, Timothy Hospedales, and Hakan Bilen. Latent domain learning with dynamic residual adapters, 2020.

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/finn17a.html`.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, jan 2016. ISSN 1532-4435.

[17] Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Alistair E. W. Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific Reports*, 12(1):2726, February 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-06484-1. URL `https://www.nature.com/articles/s41598-022-06484-1`.

[18] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL `https://doi.org/10.1038/s41597-022-01899-x`.

[19] J. Kim, E. So, H. J. Kim, K. S. Seo, and M. H. Karm. Cardiac arrest due to an unexpected acute myocardial infarction during head and neck surgery: A case report. *J Dent Anesth Pain Med*, 18(1):57–64, 2018. doi: 10.17245/jdapm.2018.18.1.57. URL `https://doi.org/10.17245/jdapm.2018.18.1.57`.

[20] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/koh21a.html`.

[21] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization, 2017.

[22] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

[23] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[24] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[25] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11682. URL `https://ojs.aaai.org/index.php/AAAI/article/view/11682`.

[26] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3915–3924. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/li19l.html`.

[27] Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 475–485, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59713-9.

[28] 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*, KDD '17, page 1903–1911, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098088. URL `https://doi.org/10.1145/3097983.3098088`.

[29] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11749–11756, Apr. 2020. doi: 10.1609/aaai.v34i07.6846. URL `https://ojs.aaai.org/index.php/AAAI/article/view/6846`.

[30] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[31] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[32] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page I–10–I–18. JMLR.org, 2013.

[33] Li Niu, Wen Li, and Dong Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2774–2783, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7298894. URL `http://ieeexplore.ieee.org/document/7298894/`.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] X. Peng, G. Long, T. Shen, S. Wang, J. Jiang, and C. Zhang. Bitenet: Bidirectional temporal encoder network to predict medical outcomes. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 412–421, Los Alamitos, CA, USA, nov 2020. IEEE Computer Society. doi: 10.1109/ICDM50108.2020.00050. URL `https://doi.ieeecomputersociety.org/10.1109/ICDM50108.2020.00050`.

[37] Christian S. Perone, Pedro Ballester, Rodrigo C. Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194: 1–11, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.03.026. URL `https://www.sciencedirect.com/science/article/pii/S1053811919302034`.

[38] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, September 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.178. URL `https://doi.org/10.1038/sdata.2018.178`.

[39] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86, May 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00455-y. URL `https://doi.org/10.1038/s41746-021-00455-y`.

[40] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL `https://www.sciencedirect.com/science/article/pii/0377042787901257`.

[41] Tim Sainburg, Leland McInnes, and Timothy Q Gentner. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907, 2021.

[42] Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *Pattern Recognition*, 135:109115, 2023. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2022.109115. URL `https://www.sciencedirect.com/science/article/pii/S0031320322005957`.

[43] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 68–83, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58542-6.

[44] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1Dx7fbCW`.

[45] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[46] Nimit Sohoni, Jared A. Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[47] Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S. Hertzberg, and Carl Yang. 4sdrug: Symptom-based set-to-set small and safe drug recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3970–3980, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539089. URL `https://doi.org/10.1145/3534678.3539089`.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[49] Claus F. Vogelmeier, Kenneth R. Chapman, Marc Miravitlles, Nicolas Roche, Jørgen Vestbo, Chau Thach, Donald Banerji, Robert Fogel, Francesco Patalano, Petter Olsson, Konstantinos Kostikas, and Jadwiga A. Wedzicha. Exacerbation heterogeneity in COPD: subgroup analyses from the FLAME study. *International journal of chronic obstructive pulmonary disease*, 13: 1125–1134, 2018. ISSN 1178-2005 1176-9106. doi: 10.2147/COPD.S160011. Place: New Zealand.

[50] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/1d94108e907bb8311d8802b48fd54b4a-Paper.pdf`.

[51] Ziqi Wang, Marco Loog, and Jan van Gemert. Respecting domain relations: Hypothesis invariance for domain generalization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9756–9763, 2021. doi: 10.1109/ICPR48806.2021.9412797.

[52] Zhenbang Wu, Huaxiu Yao, Zhe Su, David M Liebovitz, Lucas M Glass, James Zou, Chelsea Finn, and Jimeng Sun. Knowledge-driven new drug recommendation, 2022.

[53] Zhenbang Wu, Cao Xiao, Lucas M. Glass, David M. Liebovitz, and Jimeng Sun. Automap: Automatic medical code mapping for clinical prediction model deployment. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 505–520, Cham, 2023. Springer International Publishing. ISBN 978-3-031-26390-3.

[54] Zhenbang Wu, Cao Xiao, and Jimeng Sun. Medlink: De-identified patient health record linkage. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 2672–2682, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599427. URL `https://doi.org/10.1145/3580305.3599427`.

[55] Chao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, Oct 2018. doi: 10.1093/jamia/ocy068.

[56] Chaoqi Yang, M Brandon Westover, and Jimeng Sun. ManyDG: Many-domain generalization for healthcare applications. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=lcSfirnflpW`.

[57] Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin P. Danek, and Jimeng Sun. Pyhealth: A deep learning toolkit for healthcare applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5788–5789, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599178. URL `https://doi.org/10.1145/3580305.3599178`.

[58] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1): 715–723, May 2021. doi: 10.1609/aaai.v35i1.16152. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16152`.

[59] Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In

Proceedings of the Conference on Health, Inference, and Learning, pages 279–290, Virtual Event USA, April 2021. ACM. ISBN 978-1-4503-8359-2. doi: 10.1145/3450439.3451878. URL `https://dl.acm.org/doi/10.1145/3450439.3451878`.

[60] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 561–578, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58517-4.

[61] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021. doi: 10.1109/TIP.2021.3112012.

[62] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=6xHJ37MVxxp`.

Contents of Appendix

A Pseudocode of SLDG

Algorithm 1: Training and Inference for SLDG.

```
// Training
Require: Source training data from $P_{tr}$
Pre-train patient encoder $E(\cdot)$ on the same task with binary cross-entropy loss for 40 epochs
for iteration ranging from 1 to 3 do
    Perform decoupled domain discovery with the encoder $E(\cdot)$ by Eq. (2), (3)
    Initialize gating and classifier weights $\mathbf{w}_{t,k}$, $\mathbf{w}^{+}_{t,k}$, $\mathbf{w}^{-}_{t,k}$ by Eq. (10), (8)
    for epoch ranging from 1 to 20 do
        for each patient $(x, y) \sim P_{tr}$ do
            Obtain decoupled patient representations $\{\mathbf{h}_t\}_{t \in \mathcal{T}}$ by Eq. (4), (5)
            Compute domain-specific predictions $C_{t,k}(\mathbf{h}_t)$ by Eq. (7)
            Compute gating weights $G_t(\mathbf{h}_t)$ by Eq. (9)
            Obtain final prediction $o$ by Eq. (6)
            Update model parameters with binary cross-entropy loss
        end for
    end for
end for
// Inference
Require: Target testing data from $P_{te}$
for each patient $(x, y) \sim P_{tE}$ do
    Obtain decoupled patient representations $\{\mathbf{h}_t\}_{t \in \mathcal{T}}$ by Eq. (4), (5)
    Compute domain-specific predictions $C_{t,k}(\mathbf{h}_t)$ by Eq. (7)
    Compute gating weights $G_t(\mathbf{h}_t)$ by Eq. (9)
    Obtain final prediction $o$ by Eq. (6)
end for
```

B Additional Experimental Setup

B.1 Datasets

For both datasets, we select our cohorts by filtering out visits of patients younger than 18 or older than 89 years old, visits that last longer than 10 days, and visits with data from less than 3 or more than 256 timestamps. In the case of the eICU dataset, we additionally exclude visits lasting shorter than 12 hours, as the predictions are made 12 hours after admission. Similarly, for the MIMIC-IV dataset, we exclude visits where the patient ultimately passed away, as the predictions are made upon discharge. Tab. 4 provides detailed statistics of the two datasets.

B.2 Clinical Predictive Tasks

We focus on two common clinical predictive tasks: readmission prediction and mortality prediction.

In the case of the eICU dataset, the predictions are made 12 hours after admission. Readmission prediction aims to determine whether a patient will be readmitted within the next 15 days following discharge. Mortality prediction, on the other hand, aims to predict whether a patient will pass away upon discharge. The overall prevalence for these tasks is 15% for readmission and 4% for mortality.

For the MIMIC-IV dataset, the predictions are made at the time of discharge. Similar to the eICU dataset, the readmission prediction task is defined as predicting whether a patient will be readmitted within 15 days after discharge. To prevent information leakage, the mortality prediction task for MIMIC-IV is defined as predicting whether a patient will pass away within 90 days after discharge. The overall prevalence for these tasks is 14% for readmission and 4% for mortality.

Table 4: Dataset statistics.

Item	eICU	MIMIC-IV
#Patients	116075	156549
#Admissions	149227	353238
Readmission Rate	0.15	0.14
Mortality Rate	0.04	0.04
	Region: Midwest	**Year: 2008-2010**
#Patients	29767	37328
#Admissions	35989	56433
Readmission Rate	0.10	0.14
Mortality Rate	0.03	0.04
Age	62	56
Gender	F: 0.46, M: 0.54	F: 0.53, M: 0.47
Race	African American: 0.09, Asian: 0.01, Caucasian: 0.83, Hispanic: 0.01, Native American: 0.01, Other: 0.04	African American: 0.15, Asian: 0.03, Caucasian: 0.71, Hispanic: 0.06, Native American: 0.00, Other: 0.04
Average #Events	90.01	31.87
	Region: Northeast	**Year: 2011-2013**
#Patients	5886	39125
#Admissions	6958	62586
Readmission Rate	0.17	0.15
Mortality Rate	0.06	0.04
Age	62	57
Gender	F: 0.44, M: 0.56	F: 0.53, M: 0.47
Race	African American: 0.03, Asian: 0.01, Caucasian: 0.92, Hispanic: 0.01, Native American: 0.00, Other: 0.03	African American: 0.17, Asian: 0.03, Caucasian: 0.66, Hispanic: 0.07, Native American: 0.00, Other: 0.07
Average #Events	104.54	35.19
	Region: South	**Year: 2014-2016**
#Patients	27584	41737
#Admissions	33033	64592
Readmission Rate	0.11	0.14
Mortality Rate	0.04	0.04
Age	62	57
Gender	F: 0.46, M: 0.54	F: 0.52, M: 0.48
Race	African American: 0.21, Asian: 0.01, Caucasian: 0.68, Hispanic: 0.05, Native American: 0.00, Other: 0.03	African American: 0.17, Asian: 0.04, Caucasian: 0.66, Hispanic: 0.06, Native American: 0.00, Other: 0.07
Average #Events	84.28	36.53
	Region: West	**Year: 2017-2019**
#Patients	17670	40496
#Admissions	19803	63654
Readmission Rate	0.29	0.14
Mortality Rate	0.04	0.04
Age	63	58
Gender	F: 0.45, M: 0.55	F: 0.52, M: 0.48
Race	African American: 0.05, Asian: 0.03, Caucasian: 0.77, Hispanic: 0.05, Native American: 0.02, Other: 0.08	African American: 0.17, Asian: 0.04, Caucasian: 0.65, Hispanic: 0.06, Native American: 0.00, Other: 0.07
Average #Events	85.29	36.95

B.3 Data Split

The eICU dataset comprises data collected from hospitals across the United States, while the MIMIC-IV dataset spans a period of ten years. Therefore, we utilize the eICU dataset to evaluate the model's performance across spatial gaps, and the MIMIC-IV dataset to assess its performance across temporal gaps.

For the eICU dataset, we divide it into four spatial groups based on regions: Midwest, Northeast, West, and South. Each group is then split into 70% for training, 10% for validation, and 20% for testing. We evaluate the gap between groups by comparing the performance of the backbone model trained on data from within the same group and data from outside the group. The target testing data is selected as the group (Midwest) that exhibits the largest performance gap, while the remaining groups (Northeast, West, and South) are used as the source training data.

Regarding the MIMIC-IV dataset, we divide it into four temporal groups: 2008-2010, 2011-2013, 2014-2016, and 2017-2019. Each group is further split into training, validation, and testing sets with a ratio of 70%, 10%, and 20% respectively. We consider patients admitted after 2014 as the target testing data, while all preceding patients are included in the source training data.

B.4 Baselines

We first compare SLDG to two naive baselines.

- **Oracle:** We directly train a backbone model on the training set of the target domain, select the best model on the target validation set, and evaluate its performance on the target testing set. This model is trained with in-domain data and can be seen as a upper bound for all domain generalization method.
- **Base:** We train a backbone model on the training set of the source domain, select the best model on the source validation set, and evaluate its performance on the target testing set. This model is trained with out-domain data and should act as a performance lower bound.

We then compare SLDG to both classic and recent domain generalization methods. For a fair comparisons, all the methods below are trained on the source training set, selected on the source validation set, and tested on the target testing set.

- **DANN [16]:** Domain-Adversarial Neural Networks leverage a domain classifier and a gradient reversal layer to extract domain-invariant representations. This method uses the coarse regional and temporal groups as the domain definition.
- **MLDG [22]:** Meta-Learning for Domain Generalization adopts the Model-Agnostic Meta-Learning (MAML) [15] framework and simulates the new domain scenario during training. This method also uses the coarse regional and temporal groups as the domain definition.
- **ManyDG [56]:** Many-Domain Generalization disentangles domain-variant and invariant features through mutual reconstruction and orthogonal projection. This method treats each patient as a unique domain.
- **IRM [4]:** Invariant Risk Minimization learns domain-invariant representations by minimizing a bound on the expected generalization error under domain shifts. It acts as a regularizer and does not require domain IDs.
- **MMLD [29]:** Domain Generalization using a Mixture of Multiple Latent Domains iteratively assigns pseudo domain labels via clustering and trains the domain-invariant feature extractor through adversarial learning. This method does not rely on domain IDs.
- **DRA [14]:** Latent Domain Learning with Dynamic Residual Adapters uses layer-wise multi-head correction networks with a gating mechanism and residual connection to enhance model learning. This method does not rely on domain IDs.

B.5 Implementation Details

For all baselines, we use the Transformer as the backbone encoder. The number of layers is 3, the embedding dimension is 128, the number of attention heads is 2. The event embedding look-up table is initialized with ClinicalBERT [3] embeddings of the event name and then project it down to 128 dimension with a linear layer. Patient demographics features (age, gender, and ethnicity) are separately embeded with another embedding look-up table. We also embed the timestamps with sinusoidal positional encoding. The medical, patient demographics, and temporal embeddings are added together to form the overall sequence embedding. For SLDG, UMAP [30] from UMAP-learn [41] is used with 2 components, 10 neighbors, and 0 minimum distance; and k-Means from Scikit-learn [35] is used with the default hyper-parameter. We apply a dropout of rate 0.2. We use Adam as the optimizer with a learning rate of 1e-4 and a weight decay of 1e-5. All models are trained for 100 epochs. The batch size is 256. We select the best model by monitoring the AUPRC score on the source validation set (except for the Oracle baseline, where we directly use the target validation set). We implement SLDG using PyTorch [34] 1.11 and Python 3.8. The model is trained on a CentOS Linux 7 machine with 128 AMD EPYC 7513 32-Core Processors, 512 GB memory, and eight NVIDIA RTX A6000 GPUs.

C Limitations and Broader Impacts

In terms of limitations, it is important to acknowledge that our work operates under the assumption that the target testing data still exhibit some similarities with the source training data. If there is a significant distribution shift, the knowledge acquired from the source training data may become irrelevant. In such cases, neither the DG baselines nor our proposed method can effectively address the problem. It would be more appropriate to explore transfer learning or train a new model to obtain a better solution. Further, we propose SLDG to tackle two main challenges: (1) unknown domain IDs and (2) distinct characteristics across domains. In the scenario when the domain IDs are given and clearly separable (e.g., photo, art painting, cartoon, and sketch in the PACS [21] dataset), SLDG 's domain discovery approach might be unnecessary. Existing DG methods directly utilizing the domain IDs could be a better solution.

In terms of broader impacts, our work tackles a practical and prevalent issue in healthcare known as the domain shift problem. We aim to inspire future research in this area: both by investigating the existence of domain shift under various scenarios, and by contributing to the development of effective solutions for this real-world challenge.

D Notations

Notation	Meaning
x	a patient's hospital visit
$[e_1, \ldots, e_m]$	sequence of m events
t	type of an event
$T(\cdot)$	mapping function from event to its type
$\mathcal{E}$	set of all events
$\mathcal{T}$	set of all event types
$y \in \{+, -\}$	label, i.e., the occurrence of a certain future event
$f_\phi(\cdot)$	overall clinical predictive model
ϕ	model parameter
$l(\cdot)$	loss function
P_{tr}, P_{te}	training and testing data distribution
$E_t(\cdot)$	feature-specific patient encoder for event type t
$\mathbf{h}_t$	patient representation in latent space of type t
h	hidden dimension
$\{\mathbf{h}_t^{(i)}\}_{i=1}^{N_{tr}}$	all patient representations in latent space of type t
N_{tr}	total number of training samples
K_t	number of discovered domains in the latent space of type t
$\mathbf{M}_t$	domain assignment matrix
$[\mathbf{e}_1, \ldots, \mathbf{e}_m]$	contextualized representation for event sequence $[e_1, \ldots, e_m]$
$E(\cdot)$	embedding function
$\{\mathbf{h}_t\}_{t \in \mathcal{T}}$	multi-vector representations for a single patient
$C_{t,k}(\cdot)$	customized classifier for the discovered domain k in the latent space of type t
$G_{t,k}(\cdot)$	the gating weight for the customized classifier $C_{t,k}(\cdot)$
o	model output
$\mathbf{w}_{t,k}^+, \mathbf{w}_{t,k}^-$	learnable prototype weight vectors of the positive and negative classes for the k-th discovered domain in the latent space of type t
$d(\cdot, \cdot)$	Euclidean distance
$\mathbf{w}_{t,k}$	learnable prototypical weight vector for the discovered domain k in the latent space of type t

E Additional Illustrations

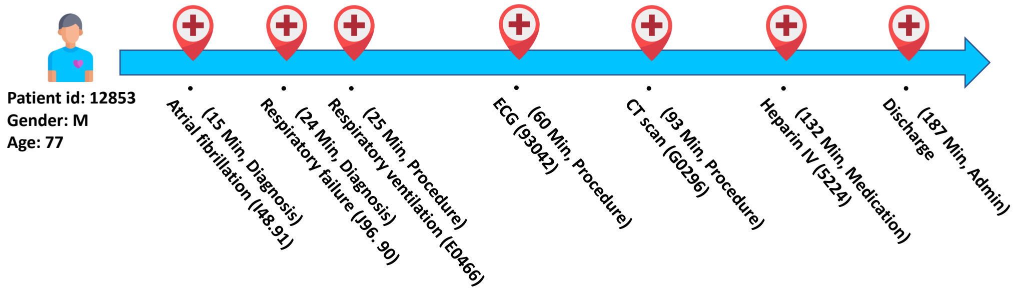

Figure 6: An illustration of the patient visit as input.

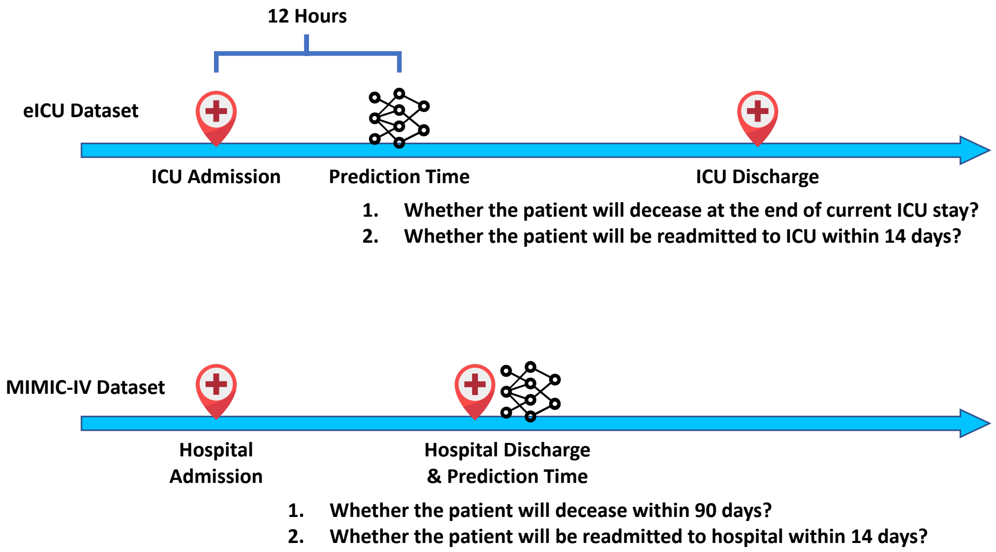
12 Hours
eICU Dataset
ICU Admission
Prediction Time
ICU Discharge
1. Whether the patient will decease at the end of current ICU stay?
2. Whether the patient will be readmitted to ICU within 14 days?
MIMIC-IV Dataset
Hospital Admission
Hospital Discharge & Prediction Time
1. Whether the patient will decease within 90 days?
2. Whether the patient will be readmitted to hospital within 14 days?

Figure 7: An illustration of the task definitions in the eICU and the MIMIC-IV datasets.

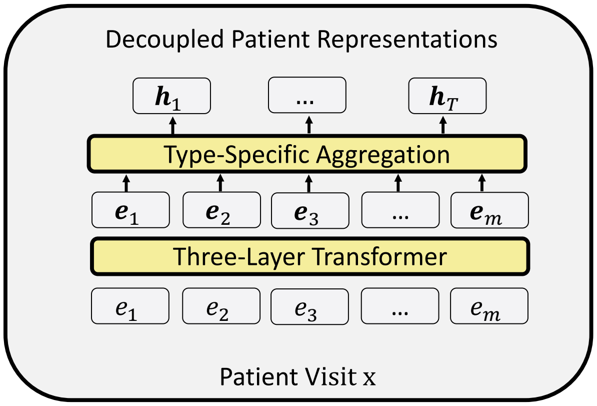

Figure 8: The architecture of the feature-specific patient encoder.