HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2022 October 19.

Published in final edited form as:

J Biomed Inform. 2022 October ; 134: 104168. doi:10.1016/j.jbi.2022.104168.

$A_{DA}D_{IAG}$: Adversarial Domain Adaptation of Diagnostic Prediction with Clinical Event Sequences

Tianran Zhang[a,b,*], **Muhao Chen**[c], **Alex A.T. Bui**[a,b]

[a]Department of Bioengineering, UCLA, United States of America

[b]UCLA Medical & Imaging Informatics (MII), United States of America

[c]Department of Computer Science, University of Southern California, United States of America

Abstract

Early detection of heart failure (HF) can provide patients with the opportunity for more timely intervention and better disease management, as well as efficient use of healthcare resources. Recent machine learning (ML) methods have shown promising performance on diagnostic prediction using temporal sequences from electronic health records (EHRs). In practice, however, these models may not generalize to other populations due to dataset shift. Shifts in datasets can be attributed to a range of factors such as variations in demographics, data management methods, and healthcare delivery patterns. In this paper, we use unsupervised adversarial domain adaptation methods to adaptively reduce the impact of dataset shift on cross-institutional transfer performance. The proposed framework is validated on a next-visit HF onset prediction task using a BERT-style Transformer-based language model pre-trained with a masked language modeling (MLM) task. Our model empirically demonstrates superior prediction performance relative to non-adversarial baselines in both transfer directions on two different clinical event sequence data sources.

Keywords

Domain adaptation; Heart failure; Transformers; Clinical event sequence modeling

1. Introduction

Recent research has demonstrated the advantages of deep learning (DL) methods for diagnostic prediction using clinical temporal sequences [1–4]. Despite the reported

*Corresponding author at: Department of Bioengineering, UCLA, United States of America. tianranzhang@ucla.edu (T. Zhang).

CRediT authorship contribution statement

Tianran Zhang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Muhao Chen:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Alex A.T. Bui:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

improvements in predicting outcomes, these models' actual clinical impact still lags behind their projected potential. A critical reason for this unfilled promise is the inability to generalize findings beyond the development cohort/population [5,6]. Due to data availability and sharing restrictions, most existing models are only internally validated using same-source, in-distribution data similar to the development data (e.g., from the same institution). Such models tend to fail, if not suffer from lower performance on independent external test cases from other sources and in different distributions [7,8].

Existing work in transfer learning has thus explored ways to improve clinical model generalizability by utilizing EHRs from multiple sources [9–11]. One transfer learning method, domain adaptation (DA), leverages knowledge from a different but related domain to train models for decision making in a new target domain, given the same task in each domain both with varying distributions of data. This approach is particularly useful when the target domain lacks labeled data [10,12]. Typically, these DA approaches require target domain ground truth for model fine-tuning, which are often scarce in clinical practice. Markedly, in cross-dataset transfer learning, the representation taken directly from the source domain is not domain-adaptive and may still fail to generalize to new data. In contrast, more recent works on adversarial domain adaptation (ADA) adaptively learn a domain-invariant representation without requiring labels from the target domain. ADA combines adversarial training with discriminative feature learning to reduce the divergence between the source and target domain distribution, thus improving generalization performance [13]. Despite its successful use in myriad applications including bilingual sentiment classification [14], skin disease image classification [15], biological sequence classification [16] and clinical time series data classification [17,18], ADA has not yet been investigated for mitigating the domain shift problem in medical event sequence classification.

To handle domain shift in event sequence classification, we propose ADADIAG (Adversarial Domain-Adaptive Diagnostic Prediction), an unsupervised adversarial domain adaptation framework with a pre-trained language model (LM) for clinical event sequences, to reduce the effects of domain shift when adapting a diagnostic prediction model from source to target domain. In this study, we specifically focus on alleviating domain shift across *patient cohorts*, where "domains" stands for *datasets* extracted from different EHR systems. The two datasets used as source and target domains are (1) the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset [19,20]; and (2) data extracted from the UCLA Health Systems (hereafter referred to as UCLA data).

To demonstrate the utility of our proposed model, we adapt a heart failure (HF) onset prediction model trained on one patient cohort to another. Heart failure is one of the most frequent and serious conditions in the United States, contributing to one out of nine deaths [21]. For a patient with a developing set of symptoms but as of yet undiagnosed HF, it might take months or years before the next visit prior to HF is uncovered, during which time the disease progresses unchecked. For institutions with limited data availability/quality and/or model development resources – and hence, training a site-specific model is not a viable option – the ability ADADIAG offers in improving testing performance for externally trained models is especially meaningful. It can facilitate earlier detection and intervention by

providing accurate predictions of next-visit incidence even when no labeled data is available from the target cohort.

AdaDiag's contributions are twofold. First, we construct a pre-trained Transformer-based LM [22,23], fine-tuned for next-visit HF prediction on lab event sequences from one EHR dataset, and externally validate it on another dataset from a different institution. Our results show that although pre-trained LMs perform well when fine-tuned for the target task on the single data source, performance drops drastically when deployed to an institution with a shifted data distribution. Second, to address the generalizability issue against dataset shifts across institutions, we present an unsupervised domain adaptation framework for clinical event sequences that addresses the domain shift problem by learning a domain-invariant representation through an adversarial domain classifier. This approach can adapt to the unseen target domain data distribution without requiring any labels. Notably, when source and target domains are switched, superior performance in adversarial-based methods persists, showing robustness of our proposed framework across different source and domain data quality settings.

2. Related work

2.1. Clinical data representation

Medical events cover a wide array of clinical concepts, such as lab orders, medications, procedures, diagnoses, and myriad other observations. The management and storage of clinical event data pose *standardization* and *harmonization* challenges for transferring models between institutions. Events such as labs and medications are recorded under varying established and/or internal coding systems from each institution. Although endeavors are made to adapt events to a unified coding scheme (e.g., International Classification of Diseases (ICD); Logical Observation Identifiers Names and Codes (LOINC)) and/or ontology, manual mapping is often still needed in systems with local terminologies for data standardization. Data structures adopted in different systems create additional barriers to data harmonization. As an effort tackling this problem, researchers developed the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [24] across multiple observational databases within an organization to facilitate standardized analytics tools when conducting observational research. The OMOP CDM streamlines data extraction process across multiple observational data sources, where different logical organizations and physical formats coexist. It also harmonizes disparate coding systems to an established standard vocabulary to prepare for the integrated analysis with all sources. Although these efforts improve access to multi-source data, they do not resolve any underlying domain shift problem. As even with two event sequence datasets with the same format/coding system, site-specific characteristics such as patient demographics, disease prevalence and treatment patterns (e.g., procedure/lab/medication ordering habit, site policy), which cannot be explicitly standardized, still cause shifts in data distributions [25].

2.2. Diagnostic prediction over time

Modeling numerical clinical time series has been extensively investigated as a means to predict clinical outcomes [18,26–29]. There are fewer studies, however, that examine

clinical event sequence modeling, which is also a critical part of appreciating the diagnostic prediction problem. A number of earlier works have explored methods to model medical event sequences using word embedding based on the co-occurrence of event codes [1,30,31]. Farhan et al. [1] model clinical abnormal lab sequences to provide next-visit diagnostic prediction using Word2Vec (i.e., skip-gram/CBOW) embeddings [32]. A different representation learned using another word embedding algorithm, GloVe [33], is demonstrated to be effective on next-visit code/risk group prediction [30] and 30-day readmission prediction [31]. However, with each word (event) represented by a fixed vector, these static embedding approaches cannot take into account the varying meanings of a given medical event based on the different patient histories it occurs in.

Pre-trained LMs for EHR data.—In light of the rapid development of pre-trained deep LMs such as BERT [23] in natural language processing (NLP), recent research has tested LMs for clinical event sequence representation learning by drawing an analogy between word sequences (text) and event sequences [2,34–37]. Some works have applied gated recurrent unit (GRU)-based LMs and achieved superior performance over more naive baselines [36,37]. DoctorAI [36] explored representing disease/medication code sequences to predict medical codes appearing in future patient encounters. [37] extends [36] by building clinical event sequences that include labs and procedures, and by evaluating a range of shallow representation methods (e.g., Word2Vec) with logistic regression (LR) and gradient boosted trees (GBTs) for predicting mortality, long admission and other clinical outcomes. Following BERT's success in natural language, more recent studies utilized Transformer-based LMs trained on clinical event sequences to learn better representations that boost downstream task performance [2,34,35]. G-BERT [35] combined the power of graphical neural networks (GNN) and BERT by incorporating a medical ontology on top of a pre-trained LM to represent diagnosis code sequences more accurately for predicting medications. The BEHRT [2] and Med-BERT [34] studies pre-train a Transformer-based model from scratch on disease code sequences combined with structural information specific to the EHRs, achieving good fine-tuning performance on tasks such as prolonged length of stay (LOS) and disease prediction. In contrast to shallow embedding methods and other DL (e.g., recurrent neural network, RNN) methods, these Transformer-based models are able to distinguish and extract *different* semantic meanings of words based on their context, which corresponds with the different indications of a given medical event and observation of a disease trajectory.

Most of the aforementioned methods have largely relied on their capability of learning better representations optimized solely on data from a single population and/or dataset. Such models suffer from lack of robustness under domain shift. Moreover, when using a source domain model on a target population encountered in clinical practice (e.g., testing), target domain labels may not be available for retraining for any number of reasons. Our study thus focuses on solving the challenging problem of unsupervised domain adaptation (UDA) on clinical event sequence data.

Unsupervised domain adaptation in medicine.—Work has been done on unsupervised domain adaptation for medical image analysis through cross-modality [38–40],

cross-vendor [41], and cross-site [42] adaptations. In other areas, UDA efforts have also been made in clinical NLP for negation detection [43], adapting detection algorithms across four corpora of clinical notes. In the context of EHR data modeling, where domains can be interpreted as patient populations, UDA can be used to improve the performance of machine learning on a target patient group by mitigating the domain shift between one and another, yet related patient population [27,29,44]. Most existing work on clinical domain adaptation using EHR data focus on modeling numerical time series, bridging the gap between patient groups with different age distributions and/or other disparities [18,27,29,45,46]. Building on earlier ADA works (e.g., domain adversarial neural network [47]) and advancements in generative adversarial networks (GAN) [48], Luo et al. [45] designed a Wasserstein GAN (WGAN [49])-based framework to improve cross-dataset transfer performance for electroencephalogram (EEG)-based emotion recognition. Purushotham et al. [18] take advantage of adversarial training and variational recurrent neural network (VRNN) [50] to learn latent temporal dependencies underlying EHR time series data adaptive across patient age groups. Similarly, [27] seeks to adversarially learn a domain-invariant representation of clinical time series for septic shock prediction with an LSTM-based framework, where domains are defined as patient groups divided by demographic attributes such as race, age, and gender. With a slightly different adversarial approach, [29] performed clinical time series augmentation by adding adversarial samples for improving the logistic regression (LR) model's generalizability across patient groups. Despite a similar focus on improving transportability across populations, these recent UDA studies are fundamentally different from earlier works that aim to extend the conclusions from randomized controlled trials (RCTs) [51], findings from epidemiology studies and public health decisions [52] to a distinct population with unknown outcomes. These studies [51,52] use statistical methods to analyze and account for population-level (demographic) changes. In contrast, using EHR-based clinical prediction models with new datasets is more challenging as clinical environments are less controlled than those of classical clinical studies [53]. In view of this, recent UDA methods aim at designing an EHR data representation learning scheme that can not only adjust for differences in cohort demographics, but also distribution shifts inherent to the data generation and collection process (e.g., different lab ordering patterns, policy shifts), which cannot be easily described and adjusted using classical statistical approaches.

3. Methods

We present AdaDiag, an adaptive deep learning framework designed to improve the unsupervised transfer performance on disease prediction tasks using clinical event sequences, moving from a labeled source domain to an unlabeled target domain. We first state the problem to be addressed and define the notations in Section 3.1, followed by an introduction to the AdaDiag framework with its main components detailed in Section 3.2. Sections 3.3 and 3.4 describe the two-stage training process of AdaDiag: (1) Transformer-based encoder pre-training, and (2) adversarial training.

3.1. Preliminary

Problem statement.—Predictive models derived from EHR data are often developed and validated on the same population, and yet show a great decline when deployed/tested

on external data due to dataset shift [25]. For instance, when a model trained on a national/multi-institutional dataset is used on data from a regional hospital, direct transfer performance may be sub-optimal due to site-specific data generation/storage processes. Here, we addresses the issue of transferring an event sequence diagnostic prediction model from a *source* dataset, where it was developed and trained, to another, *target* dataset, where it could be applied without requiring its disease labels, as an *unsupervised domain adaptation* problem.

The diagnostic prediction task seeks to estimate the likelihood of patients' disease onset based on their visit histories. For instance, the next-visit HF prediction task is based on predicting the *first* appearance of HF-related ICD-9/10 codes during the most recent visit, given the combined event history from all past visits of the patient. To differentiate between elements from the two data domains, we use superscripts *src* and *tgt* to indicate domain membership. For example, D^{src} and D^{tgt} represent the source and target domain. For a given patient i with a visit history X_i of n encounters $X_i = [x_1 \oplus x_2 \oplus \cdots \oplus x_n]$, each visit x_j consists of a sequence of events $x_j = [e_1 \oplus e_2 \oplus \cdots \oplus e_n] \in x_j$, with all events ordered sequentially by time. The next-visit disease label for event sequence X_i is denoted as $y_i \in \{0, 1\}$, which is available during training when $X_i \in D^{src}$. All sequences from D^{src} and D^{tgt} are assigned with domain labels $y_i' \in \{0, 1\}$.

3.2. *The* AdaDiag *framework*

As illustrated in Fig. 1, AdaDiag is a feed-forward network with two forward branches following the design in [14]. The network consists of three parts: (1) a joint feature extractor $\mathcal{F}$ that maps an input sequence X_i to a shared feature space $\mathcal{F}(X_i)$; (2) a diagnostic classifier $\mathcal{P}$ that predicts the label for X_i based on the feature representation $\mathcal{F}(X_i)$; and (3) a domain discriminator $\mathcal{Q}$ that also takes $\mathcal{F}(X_i)$ but predicts a label indicating domain identity (source/target) of X.

For improved performance, we pre-train a Transformer encoder as the feature extractor $\mathcal{F}$ to capture the contextualized information in the sequence. $\mathcal{F}$ feeds the sequence representation to $\mathcal{P}$, which is essentially a multi-layer perceptron (MLP) with a sigmoid output for binary diagnostic prediction. While trained with a different optimizer from $\mathcal{P}$'s, the domain discriminator $\mathcal{Q}$ is also an MLP, but ends with a linear layer to output a domain label [14]. During training, the diagnostic predictor $\mathcal{P}$ can only see disease labels from the source-domain dataset, whereas $\mathcal{Q}$ can observe (unlabeled) event sequences from both the source and target domain datasets.

The feature extractor $\mathcal{F}$ tries to learn a domain-invariant representation that aids in the prediction of the diagnostic predictor $\mathcal{P}$ as well as prevents the model from distinguishing features between different domains. The feature learned by $\mathcal{F}$ can be considered domain-invariant if a trained $\mathcal{Q}$ fails to distinguish between sequences from different domains. In this regard, $\mathcal{Q}$ is the adversarial component of the AdaDiag, as its target (distinguishing domains) goes against the overall goal of the AdaDiag framework on learning domain-invariant features. A well-trained $\mathcal{F}$ should be able to learn features that benefit the diagnostic prediction task, while keeping the domain identity as ambiguous as possible. Disease

prediction can be performed at inference time by running unlabeled target domain sequences through the feature extractor $\mathcal{F}$ and the diagnostic classifier $\mathcal{P}$. No disease labels from the target domain are required throughout the model development process. At inference time, an input sequence X_i is passed through sufficiently trained $\mathcal{F}$ and $\mathcal{P}$ to predict for the disease label y_i, while keeping the domain discriminator $\mathcal{Q}$ untouched.

3.3. Pre-training transformer encoder

Following the recent success in Transformer-based pre-trained language models [23,54,55] and their adaptations modeling EHR data [2,34], we construct a BERT-like architecture with six Transformer encoder layers, six attention heads, and an embedding dimension of 768 as the shared feature extractor $\mathcal{F}$ for a contextual representation that accounts for the entire disease progression process. Similar to [23], our modeling process also involved three special tokens: [CLS],[SEP], and [MASK]. [CLS] is a special symbol added in front of every input example, whose representation will be used as the final sequence representation in fine-tuning tasks; [SEP] is a special separator token, indicating the end of the input sequence. The MLM task is adopted as the pre-training task of the Transformer-based encoder, which seeks to recover randomly masked clinical events (represented by [MASK]) in given sequences. All unlabeled event sequences from both source and target domains are used for this process. Unlike language models (e.g. BERT) that processes subword or byte-pair sequences, our encoder treats individual LOINC codes as the minimal units, since a lab event code cannot be further divided into semantically meaningful sub-units. Fig. 2 illustrates the BERT-like input representation of our pre-trained Transformer-based model. As defined in the BERT paper [23], the input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings. As we do not differentiate segments within each input sequence, all segment embeddings are identical.

The MLM pre-training in our study follows a setting similar to the original BERT paper [23]. First, 15% of tokens in the sequences are randomly selected, and these chosen tokens will: (1) be replaced with the [MASK] token 80% of the time, (2) be replaced by another random tokens 10% of the time, and (3) stay unchanged the remaining 10% of the time. This mixed masking strategy was chosen to soften the discrepancy between pre-training and fine-tuning, as the [MASK] symbol will not appear during the fine-tuning stage [23]. For an input that contains one or more masked tokens, the model will generate the most likely substitution for each. We sampled 25% of all sequences for MLM evaluation, and trained the model for 100 epochs using the remaining sequences for predicting the masked token with cross-entropy loss. The best model was selected based on the lowest validation loss. In this process, the model captures the bidirectional context of each event in the sequence and accordingly learns a contextualized event representation.

3.4. Adversarial training

AdaDiag aims at learning features from event sequences that are simultaneously beneficial to disease risk discrimination and cross-domain generalization. This goal can be achieved by adversarially optimizing on two discriminative tasks: disease prediction and domain discrimination. Like two-player game training from GANs, the adversarial training scheme of AdaDiag is formed as a minimax problem. Specifically, we need to find a set of

parameters that *minimize* the disease prediction loss and at the same time *maximize* the domain discriminator loss.

As a result, adversarial training reduces the disparity between the marginal distributions of the source and target features, $P_{\mathcal{F}}^{src}$ and $P_{\mathcal{F}}^{tgt}$, over the shared feature space $\mathcal{F}(x)$:

$$P_{\mathcal{F}}^{src} \triangleq P\left(\mathcal{F}(x) \mid x \in D^{src}\right)$$

$$P_{\mathcal{F}}^{tgt} \triangleq P\left(\mathcal{F}(x) \mid x \in D^{tgt}\right)$$

To learn domain-invariant features, AdaDiag trains $\mathcal{F}$ to make distributions of $P_{\mathcal{F}}^{src}$ and $P_{\mathcal{F}}^{tgt}$ to be as close as possible to improve cross-domain generalization. Intuitively, if a well-trained $\mathcal{D}$ cannot determine the domain membership of the extracted features by $\mathcal{F}$ between the source and target domains, the features are domain-invariant.

In earlier works on adversarial domain adaptation (e.g., DANN [47], ADDA [56]), features are learned to confuse a classifier through different adversarial losses. Some [56] use the traditional GAN loss that can be deemed as minimizing the Jensen–Shannon (J–S) divergence between the source and target feature distributions, $P_{\mathcal{F}}^{src}$ and $P_{\mathcal{F}}^{tgt}$. When the learned features fail to mix distributions from both domains, gradient vanishing can occur if traditional probability-based loss measures such as cross-entropy or J–S divergence are used [57]. This situation might be better served by instead minimizing the Wasserstein distance [58], which appears to maintain gradient stability even when two distributions are far apart [49]. Specifically, we minimize the Wasserstein distance W between $P_{\mathcal{F}}^{src}$ and $P_{\mathcal{F}}^{tgt}$ over other alternatives [57] due to its stability on parameter selection as argued in [14,49], which is defined as follows:

$$W\left(P_{\mathcal{F}}^{src}, P_{\mathcal{F}}^{tgt}\right) = \inf_{\gamma \sim \Pi\left(P_{\mathcal{F}}^{src}, P_{\mathcal{F}}^{tgt}\right)} \mathbb{E}_{\left(x^{src}, x^{tgt}\right) \sim \gamma} \left[\left\| x^{src} - x^{tgt} \right\|\right] \tag{1}$$

where $\Pi\left(P_{\mathcal{F}}^{src}, P_{\mathcal{F}}^{tgt}\right)$ denotes all possible joint distributions of source and target distributions, $P_{\mathcal{F}}^{src}$ and $P_{\mathcal{F}}^{tgt}$. As Eq. (1)'s minimum is computationally intractable, its Kantorovich–Rubinstein duality form is usually used in practice [59]:

$$W\left(P_{\mathcal{F}}^{src}, P_{\mathcal{F}}^{tgt}\right) = \sup_{\|g\|_L \leq 1} \mathbb{E}_{f(x) \sim P_{\mathcal{F}}^{src}} [g(f(x))] - \mathbb{E}_{f(x') \sim P_{\mathcal{F}}^{tgt}} [g(f(x'))] \tag{2}$$

The supremum is over functions g where g is 1-Lipschitz continuous. For simplicity, we denote this as $\|g\|_L \leq 1$. Note that the function g is 1-Lipschitz continuous if and only if $|g(x) - g(y)| \leq |x - y|$, for all x and y. In our case, $\mathcal{D}$ serves as the function g in Eq. (2). Following

[14], to make $\mathcal{Q}$ a 1-Lipschitz continuous function, all parameters in $\mathcal{Q}$ are clipped to a fixed range, $[-c, c]$, at the end of each $\mathcal{Q}$ optimization step. The minimax optimization process of adversarial training involves two learning objectives: the domain discriminator objective J_q, and the disease classification objective J_p. The model is trained for these two objectives in an alternating fashion.

First, the discriminator $\mathcal{Q}$ is trained by maximizing the discriminator loss with $\mathcal{F}$ and $\mathcal{P}$ parameters fixed. The domain discriminator objective J_q is an approximation of the Wasserstein distance between the data distributions of the two domains. At the $\mathcal{Q}$ optimization step, it seeks to maximize J_q by updating its parameters in θ_q:

$$J_q(\theta_q) = W\left(P_{\mathcal{F}}^{src}, P_{\mathcal{F}}^{tgt}\right) = \max_{\theta_q}\left[\mathop{\mathbb{E}}_{\mathcal{F}(x)\sim P_{\mathcal{F}}^{src}}[\mathcal{Q}(\mathcal{F}(x))] - \mathop{\mathbb{E}}_{\mathcal{F}(x')\sim P_{\mathcal{F}}^{tgt}}[\mathcal{Q}(\mathcal{F}(x'))]\right] \tag{3}$$

Next, the disease classifier is optimized by minimizing the disease classification loss with the discriminator $\mathcal{Q}$ fixed. The disease classification objective J_p, parameterized by θ_p, aims to minimize the binary cross-entropy loss $L_p(\hat{y}, y)$:

$$J_p(\theta_p) = \min_{\theta_p} \mathop{\mathbb{E}}_{(x,y)}\left[L_p(\mathcal{P}(\mathcal{F}(x)), y)\right] \tag{4}$$

$L_p(\hat{y}, y)$ is defined as the negative log-likelihood of correctly predicting the binary disease label:

$$L_p(\hat{y}, y) = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)$$

where $\hat{y}_i$ is the next-visit disease onset prediction for the ith patient in the $\mathcal{P}$ output, y_i is the corresponding disease label, and output size is the number of predicted values/patients in the $\mathcal{P}$ output.

Lastly, serving for both discriminative tasks, the joint feature extractor $\mathcal{F}$ seeks to minimize the disease classification loss J_p as well as the Wasserstein distance J_q:

$$J_f = \min_{\theta_f}\left[J_p(\theta_f) + \lambda J_q(\theta_f)\right] \tag{5}$$

where λ is a hyperparameter that balances the losses of $\mathcal{P}$ and $\mathcal{Q}$

4. Experiments

Here, we describe the datasets as well as their pre-processing procedures (Section 4.1), introduce implementation details of AdaDiag (Section 4.2), define baseline models and AdaDiag variants (Section 4.3), and present domain adaptation results on the next-visit HF onset prediction task (Section 4.4).

4.1. Experimental setup

Abnormal lab sequences[1] from two data sources, UCLA and MIMIC-IV, are used to predict next-visit HF onset. Compared to disease codes, which are usually unordered within visits, time-stamped lab events capture more fine-grained temporal dynamics within and between visits. The distribution of lab sequences across institutions could differ for a number of reasons, including demographics, mismatched ordering patterns, and policy changes—all of which contribute to domain shift that may limit cross-data model generalization. In this section, experiments are setup to address this challenging case. We briefly introduce the two EHR datasets and describe methods used to extract abnormal lab events with corresponding disease labels. In addition, differences in the two datasets are discussed, which indicate possible domain shifts as a result of disparities in the data generation and curation processes.

UCLA dataset.—We selected adult (≥ 18 years old at initial admission time) patients who had at least one intensive care unit (ICU) stay at the UCLA Health System between 2013-03-01 and 2021-03-01, extracting all abnormal lab events from all in-patient visits within this time window along with their associated disease codes.

MIMIC-IV dataset.—MIMIC-IV data (version 1.0) [19] includes de-identified records from Beth Israel Deaconess Medical Center (BIDMC) for over 60,000 patients admitted to an ICU or the emergency department between 2008 and 2019.

Similar pre-processing steps are performed on lab events and diagnosis codes extracted from the two initial cohorts. To maximize clinical utility of our developed models, we focus on predicting *unseen* HF occurrences. Specifically, we excluded all encounters after the initial HF onset (if any), and use only encounters before (not including) the onset encounter as the sequence used for disease prediction. Patients with only one encounter remaining are removed as next-visit diagnosis prediction requires at least two visits. For a given patient i, abnormal lab events from all visits *before* the most recent visit (post-filtering) x_n are ordered by time and concatenated as the prediction input $X_i = [x_1 \oplus x_2 \oplus \cdots \oplus x_{n-1}]$, with its HF label defined as y_i, a binary indicator for having at least one HF diagnose (i.e., 3-digit ICD-9 code of *428* or ICD-10 code of *I50*) associated with $\{x_j\}$, $j \in \{1, 2, 3, \ldots, n-1\}$.

To facilitate a successful transfer, standardization of lab codes is needed so that the sequences from two local systems speak the same “language”. We convert the local lab codes to a unified vocabulary, LOINC. UCLA Health has mappings from its local codes to LOINC for almost all available labs. In contrast, MIMIC-IV has no LOINC mappings for lab items in its microbiology events table, so we extracted raw lab sequences only from the *labevents* table using the dictionary file provided to map from local labs to LOINC codes. After removing all labs that are not mapped to LOINC codes, 96.7%[2] of the *labevents* occurrences in the extracted MIMIC-IV sequences remain. The pre-processed UCLA data has sequences for 18,736 patients with 1218 unique LOINC codes, while the

[1]Abnormal lab sequences include only lab tests with flagged/abnormal results.
[2]In MIMIC-IV (v1.0), less than 17% (269/1630) of codes in the *d_labitems* mapping file are mapped to LOINC, covering 90.3% of all occurrences. We combined the local to LOINC lab mappings provided in MIMIC-III (v1.4) to map a larger percentage of labs to LOINC.

MIMIC-IV LOINC sequences have 27,782 patients with 272 unique codes. There are 139 shared LOINC codes in both vocabularies. The difference in vocabulary coverage is a result of multiple reasons from data generation (e.g., lab availability, ordering patterns, policy changes) to curation processes (e.g., incomplete mapping process), and is one cause of domain shift. There are other differences in the two datasets that may indicate potential shifts in data distribution. As shown in Table 1, MIMIC-IV patients have on average fewer visits, shorter sequence lengths, a higher proportion of females, and a higher HF onset rate compared to the UCLA patients.

We conduct domain adaptation experiments in two directions: from UCLA to MIMIC-IV, and from MIMIC-IV to UCLA, both assuming no labels available in the target domain. The two datasets can each serve as D^{src} or D^{tgt}. 80% of the sequences from D^{src} are randomly selected and used for training, while the rest are used for model selection based on the validation area under the receiver operator characteristic (AUROC) curve metric. We maintain the same splits when UCLA and MIMIC-IV each serves as D^{src} for comparability. The best model is reported with AUROC and precision–recall area under the curve (pr-AUC) on all D^{tgt} sequences. We note here that this is deemed as zero-shot HF prediction, as no D^{tgt} labels are involved during the model development phase.

4.2. Implementation details

We first pre-train AdaDiag with the MLM objective to learn the network parameters in the joint feature encoder $\mathcal{F}$ that can predict the masked lab event tokens, on all unlabeled sequences from D^{src} and D^{tgt}. Similar to the setup reported in Med-BERT [34], we find that Transformers with six layers and six attention heads in the pre-trained model is the best architecture. Different from Med-BERT, we choose to use a hidden size of 768 (same for all Transformer-based encoders in this paper). The pre-trained model is then fine-tuned on the HF onset prediction and domain discrimination tasks with an adversarial objective. $\mathcal{F}$ and $\mathcal{P}$ are optimized jointly using AdamW [60] with a learning rate of 1E-5 and a weight decay of 0.01. A linear learning rate scheduler is used in all experiments. To balance the learning speed of the two alternatingly optimized adversarial objectives, Q is trained using a *separate* AdamW optimizer with a learning rate of 5E-4 and a weight decay of 0.01. To ensure that the discriminator $\mathcal{Q}$ satisfies the 1-Lipchitz constraint of the Wasserstein objective [49], the weights of $\mathcal{Q}$ are clipped within [−0.01,0.01] at the end of each training step of $\mathcal{Q}$, following the values used in [14]; the adversarial objective weight parameter λ from Eq. (5) is adjusted to 0.2 for all adversarial experiments. A fixed sequence length is chosen to be 1024 with sequences truncated/padded from the left, considering the fact that recent event history is more relevant to the upcoming disease onset. We train AdaDiag variants and the baselines and select the best model based on the validation AUROC metric from D^{src}. When transferring from UCLA data to MIMIC-IV data, the selected AdaDiag architecture has six layers in the shared feature extractor $\mathcal{F}$ (taken from the pre-trained Transformer model), zero layers in the disease classifier $\mathcal{P}$ ($\mathcal{P}$ is simply an output layer in this case) and two layers in the domain discriminator $\mathcal{Q}$. When the transfer is conducted reversely (from MIMIC-IV to UCLA), the best architecture has zero layers in the domain discriminator, with other settings remaining the same. The domain adaptation performance of AdaDiag

is reported through AUROC and pr-AUC metrics on the entire D^{tgt} dataset. AdaDiag was implemented using Huggingface [61] based on PyTorch [62].

4.3. Baseline models

GRU/bi-GRU encoder with skip-gram embedding.—A pre-trained Transformer encoder is used in AdaDiag. While in non-Transformer-based clinical event sequence models [36,63–65], the immediate sequence representation is provided using shallow word embedding methods (e.g., skip-gram/CBOW models in Word2Vec), before being fed into encoders like long short-term memory (LSTM) networks/GRUs or convolutional neural networks (CNN) to learn a final representation. More advanced models apply bidirectional RNNs (i.e., bi-LSTM or bi-GRU) to better capture the temporal dependencies of clinical visits and improve model interpretability. We implemented the first two baseline models as a one-layer GRU encoder and a one-layer bi-GRU encoder. Their initial sequence encoding is provided by a skip-gram algorithm pre-trained on all sequences from D^{src} and D^{tgt}, with a window size of 20 and a embedding dimension of 768, which is the same as the dimension of the pre-trained Transformer encoder. The encoded features are directly passed to a linear output layer with Sigmoid activation to provide HF prediction, for which an Adam optimizer with a learning rate of 1E-3 is used.

Pre-trained transformer.—Recent studies have shown the effectiveness of pre-trained Transformer-based encoders on learning better event sequence representation compared to RNN-based methods, achieving improved performance when fine-tuned on downstream tasks [2,34]. An intuitive baseline is applying the non-adversarial version of AdaDiag with pre-trained Transformer encoder fine-tuned on D^{src} directly to D^{tgt}. For fair comparison, the pre-trained encoder prior to fine-tuning is the same as the one used in AdaDiag, which is pre-trained with the parameters of six layers, six attention heads, and a hidden dimension of 768.

Untrained transformer model.—To understand the added value of pre-training to model generalizability, we compare the performance of the fine-tuned Transformer against the fine-tuned Transformer with no pre-training, where the latter is defined with the same architecture as the former but has randomly initialized layers. All the above baseline models discussed thus far are non-adversarial. We also report the results of the adversarial version of the fine-tuned, untrained Transformer model to demonstrate how adversarial training can be beneficial in another model setting, and to further illustrate the utility of pre-training in adversarially trained models.

4.4. Results

MIMIC-IV to UCLA transfer.—We first implement AdaDiag and baseline models for adapting from MIMIC-IV to UCLA data, given the fact that the former has a larger population and is publicly available. This is a more realistic scenario considering the data sharing restrictions of institution-specific datasets, as training AdaDiag requires data access from the source domain. In this setting, we emulate the situation where a local hospital system (UCLA) deploys models developed on public data from external institutions (BIDMC) to inform decisions. As shown in Table 2, all baseline models performed

well on MIMIC-IV validation data. However, when tested on the UCLA sequences, they experienced drastic drops in both metrics. Over all other baselines, the Transformer baseline model performed best on MIMIC-IV and UCLA datasets, while the GRU model with skip-gram embedding performed the worst. Using metrics reported in Table 2, we created a graph visualizing relative performance loss for all models in Fig. 3. In comparison to their non-adversarial counterparts, the two adversarial models had less performance loss from the cross-data transfer. AdaDiag achieved superior predictive performance (highlighted in gray) on UCLA data in comparison to all non-adversarial baselines and its adversarial variant (i.e., AdaDiag without pre-training). Specifically, compared with the best baseline model, non-adversarial pre-trained Transformer, AdaDiag's adversarial training boosted AUROC and pr-AUC on the UCLA data by 4.0%[3] and 4.1%. When no pre-training was performed, adversarial training boosted Transformer model's performance by 3.8% in AUROC and 8.1% in pr-AUC. These observations brought us to the conclusion that adversarial training benefits the Transformer-based models' generalization performance, while not greatly sabotaging their performance on the source domain. The untrained Transformer encoder baseline (fine-tuned on MIMIC-IV data) did not outperform pre-trained Skip-gram embedding with bi-GRU encoder when tested on MIMIC-IV data. Adding pre-training to it significantly improved its AUROC by 9.5% and pr-AUC by 15.0%, achieving superior performance relative to the bi-GRU with Skip-gram embedding baseline. In addition, pre-training was able to improve the AUROC and pr-AUC on target domain by 8.9% and 10.7% when added to the adversarial variant of the untrained Transformer baseline. These improvements show the importance of pre-training on increasing model's generalization performance on new datasets.

UCLA to MIMIC-IV transfer.—Given that labels from both datasets are readily available, we can verify if conclusions from MIMIC-IV to UCLA transfer still hold true with a different setup: transferring from a source data (UCLA) with smaller dataset but larger event vocabulary than the target data (MIMIC-IV). Models trained on UCLA data were tested on MIMIC-IV (Table 3), showing steep declines in AUROCs and pr-AUCs. AdaDiag had the least performance loss while the GRU+Skip-gram embedding model exhibited the most (Fig. 4). When comparing the two adversarial models with their non-adversarial counterparts, we found that they experienced smaller relative performance loss. Thus, the same conclusion from our previous experiments (i.e., MIMIC-IV to UCLA transfer) persists: adversarial training helps model generalize when transferring across datasets. Table 3 also shows that with 3.4% gain in AUROC and 4.3% gain in pr-AUC, AdaDiag (highlighted in gray) outperformed the best non-adversarial baseline on MIMIC-IV data, while maintaining a comparable source domain validation performance on UCLA data. Both adversarial models outperformed their non-adversarial baselines, indicating that their zero-shot adaptation was enhanced by adversarial training without significant negative impact on their source domain performances. In this transfer setting, pre-training also played a major role, as was the case in UCLA to MIMIC-IV. In baseline models, the untrained Transformer performed worse than the bi-GRU model with skip-gram embedding; pre-training boosted its performance by 6.6% in AUROC and 14.8% in pr-AUC. In adversarial models, pre-

[3]This percentage was calculated for *relative improvement*, same as below.

training improved the model performance on MIMIC-IV by 5.7% in AUROC and 12.2% in pr-AUC.

In both adaptation settings, the GRU encoder with skip-gram embedding was significantly less effective on learning features generalizable across datasets than other bi-GRU and Transformer-based methods, which is consistent with results reported in previous studies [34] and is possibly due to its left-to-right recurrent learning scheme and inability of learning bidirectional/contextual representations.

T-SNE visualization of feature distributions.—To compare and contrast the impact of adversarial training, we use t-SNE [66] for dimensionality reduction and visualize the feature distributions generated by Transformer encoders from different models/training stages in 2D. Fig. 5 shows distributions of representations of pre-trained Transformer models before (Fig. 5(a)) and after fine-tuning (Figs. 5(b) and 5(d)); and adversarially trained pre-trained Transformer models (i.e., AdaDiag) (Figs. 5(c) and 5(e)) in both transfer directions. In Fig. 5(a), sequence representations from the two domains are far away from each other, showing through MLM pre-training alone is not sufficient to bridge the gap between UCLA and MIMIC-IV data. Train-on-source-only models are built on top of the pre-trained Transformer model and fine-tuned on the disease classification task. Their encoders' new mappings (Figs. 5(b) and 5(d)) brought features from the two domains slightly closer, while remaining fairly separate from each other.

Visualization of self-attention in transformer encoders.—The self-attention mechanism of the Transformer layers is able to capture complex relationships between lab events, adding interpretability to our model. Fig. 6 shows an analysis of self-attention in AdaDiag (MIMIC-IV to UCLA)'s Transformer encoder. Based on the approach presented by [67], we analyze attention-based patterns for two patients, referred to as A (Fig. 6(a)) and B (Fig. 6(b)), from the UCLA cohort. For each patient, the abnormal lab events are presented as two identical columns, with events ordered chronologically from top to bottom. By passing the sequences through the six Transformer encoder layers, each with six attention heads, a 'headview' for each head from each layer is generated depicting attention weights of all events in the sequence, given an event of interest. An abnormal lab (in gray, on the right) is linked with all lab events in the same sequence, with the shades of the color block/edges reflecting the attention weights/degrees of association. For patient A, whose history consists of a short sequence of events from the same encounter, a strong association is found between the B-type natriuretic peptide (BNP) test (elevated value observed in heart failure) and the alanine aminotransferase (ALT) test (used to check for liver damage). The ALT test is also weakly associated with aspartate aminotransferase (AST) (another test for tissue damage in organs like liver/heart) and other red blood cell related tests such as erythrocyte distribution width (EDW) and erythrocyte count. In patient B's history, which spanned across multiple visits, the AST test is closely related to several blood tests that are relevant to red blood cells and their ability to transmit oxygen in blood: hemoglobin/hematocrit concentration, mean corpuscular hemoglobin concentration (MCHC), and erythrocyte count. This example especially shows AdaDiag's ability of extracting long term dependencies in a multi-visit sequence. The associations between

ALT and BNP (patient A) could help uncover new patterns when evaluating liver damage as early signs of heart failure due to the complex cardiohepatic interactions [68]. The relationships between ALT or AST and erythrocyte related tests (patient A/B) might indicate the underlying linkage between conditions such as anemia and organ (e.g., liver/heart) tissue damage. Identifying such self-attention patterns has enabled more profound understanding of AdaDiag's functionality and extended its potential into discovering new knowledge that has clinical relevance.

5. Conclusion

Improved generalizability of clinical predictive models is essential to achieving widespread clinical application under the constraint of low training resources. In this paper, we address the dataset shift issue that has prevented successful cross-dataset applications by enforcing domain-invariant representations through unsupervised adversarial training. We introduce a novel Transformer-based adversarial domain adaptation framework that transfers an event sequence diagnostic prediction model from a *source* domain, where it was developed and trained, to another *target* domain where it could be applied without requiring the disease labels. Its utility was demonstrated on next-visit HF onset prediction in two transfer settings, using two large clinical event datasets: from MIMIC-IV to UCLA, and UCLA to MIMIC-IV. While the RNN or Transformer-based non-adversarial baselines suffered greatly when tested on unseen sequences from the target data, adversarial training was found to be effective in improving the Transformer-based model's performance on unseen targets, maintaining similar accuracy on the source data. We also highlighted the importance of pre-training in the ablation studies with an untrained Transformer model, showing that pre-training in conjunction with adversarial training led to an increased generalization power for AdaDiag. A t-SNE plot illustrating the effect of adversarial training on feature distributions is presented for mixing two distributions with originally large differences from two domains. With the help of the Transformer encoder, the interpretability of the self-attention patterns learned within AdaDiag was visualized using the Bertviz [67] tool, which showed clinically meaningful associations between abnormal lab events from a given patient's history. This also allows AdaDiag to explore formerly unknown patterns for medical knowledge discovery.

AdaDiag's application goes beyond HF prediction; other clinical prediction tasks, such as length of stay (LOS) and mortality prediction, also require cross-population generalizability, and are part of our research plans for the future. Future directions we would like to explore include: (1) extending AdaDiag to other types of clinical events such as medication and diagnostic codes; (2) applying AdaDiag to correct temporal dataset shift; and (3) training AdaDiag without access to source domain data.

For reproducing findings in AdaDiag, our codes and model development details can be found at https://github.com/tianranzhang/AdaDiag.

Acknowledgments

This work was supported by NIH R01 CA226079. We thank Dr. Mindy Ross for reviewing the manuscript and providing clinical insights; Dr. Chelsea J.-T. Ju for consultation on analysis; Roger Zou for proof reading the manuscript.

References

[1]. Farhan W, Wang Z, Huang Y, Wang S, Wang F, Jiang X, A predictive model for medical events based on contextual embedding of temporal sequences, JMIR Med. Inf 4 (2016).

[2]. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G, BEHRT: Transformer for electronic health records, Sci. Rep 10 (2020).

[3]. Choi E, Xu Z, Li Y, Dusenberry MW, Flores G, Xue Y, Dai AM, Learning the graphical structure of electronic health records with graph convolutional transformer, in: AAAI, 2020.

[4]. Zhang T, Chen M, Bui AAT, Diagnostic prediction with sequence-of-setsrepresentation learning for clinical events, Artificial Intelligence in Medicine. Conference on Artificial Intelligence in Medicine 12299 (2020) 348–358.

[5]. Riley R, Ensor J, Snell KIE, Debray T, Altman D, Moons K, Collins G, External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges, BMJ 353 (2016).

[6]. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado GC, King D, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (2019).

[7]. Johnson AEW, Pollard T, Naumann T, Generalizability of predictive models for intensive care unit patients, in: Machine Learning for Health (ML4H) Workshop At NeurIPS 2018, 2018.

[8]. Saria S, Subbaswamy A, Tutorial: Safe and reliable machine learning, 2019, ArXiv abs/1904.07204.

[9]. Dubois S, Romano N, Jung K, Shah N, Kale DC, The effectiveness of transfer learning in electronic health records data, in: ICLR, 2017.

[10]. Sun Z, Peng S, Yang Y, Wang X, Li F, A general fine-tuned transfer learning model for predicting clinical task acrossing diverse EHRs datasets, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 490–495.

[11]. Gupta P, Malhotra P, Narwariya J, Vig L, Shroff GM, Transfer learning for clinical time series analysis using deep neural networks, J. Healthc. Inf. Res 4 (2020) 112–137.

[12]. Desautels T, Calvert J, Hoffman J, Mao Q, Jay M, Fletcher GS, Barton C, Chettipally U, Kerem Y, Das R, Using transfer learning for improved mortality prediction in a data-scarce hospital setting, Biomed. Inf. Insights 9 (2017).

[13]. Xu W, He J, Shu Y, Transfer learning and deep domain adaptation, in: Aceves-Fernandez MA (Ed.), Advances and Applications in Deep Learning, IntechOpen, Rijeka, 2020, 10.5772/intechopen.94072.

[14]. Chen X, Sun Y, Athiwaratkun B, Cardie C, Weinberger KQ, Adversarial deep averaging networks for cross-lingual sentiment classification, Trans. Assoc. Comput. Linguist 6 (2018) 557–570.

[15]. Gu Y, Ge Z, Bonnington C, Zhou J, Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification, IEEE J. Biomed. Health Inf 24 (2020) 1379–1393.

[16]. Lin R, Zeng X, Kitani K, Xu M, Adversarial domain adaptation for cross data source macromolecule in situ structural classification in cellular electron cryo-tomograms, Bioinformatics 35 (2019) i260 – i268. [PubMed: 31510673]

[17]. Tonutti M, Ruffaldi E, Cattaneo A, Avizzano C, Robust and subject-independent driving manoeuvre anticipation through domain-adversarial recurrent neural networks, Robot. Auton. Syst 115 (2019) 162–173.

[18]. Purushotham S, Carvalho W, Nilanon T, Liu Y, Variational recurrent adversarial deep domain adaptation, in: ICLR, 2017.

[19]. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R, MIMIC-IV (version 1.0), in: PhysioNet, 2021.

[20]. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE, Physiobank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals., Circulation 101 23 (2000) E215–20. [PubMed: 10851218]

[21]. Virani S, Alonso A, Benjamin E, Bittencourt M, Callaway C, Carson A, Chamberlain A, Chang AR, Cheng S, Delling FN, Djoussé L, Elkind M, Ferguson J, Fornage M, Khan S, Kissela B, Knutson K, Kwan T, Lackland D, Lewis T, Lichtman J, Longenecker C, Loop M, Lutsey P, Martin S, Matsushita K, Moran A, Mussolino M, Perak AM, Rosamond W, Roth GA, Sampson U, Satou G, Schroeder E, Shah SH, Shay C, Spartano N, Stokes A, Tirschwell D, VanWagner L, Tsao C, Heart disease and Stroke statistics—2020 update: A report from the American heart association, in: Circulation, 2020.

[22]. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I, Attention is all you need, 2017, ArXiv abs/1706.03762.

[23]. Devlin J, Chang M-W, Lee K, Toutanova K, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL, 2019.

[24]. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie MJ, Defalco F, Londhe AA, Zhu VJ, Ryan PB, Feasibility and utility of applications of the common data model to multiple, disparate observational health databases, J. Am. Med. Inf. Assoc. : JAMIA 22 (2015) 553–564.

[25]. Subbaswamy A, Saria S, From development to deployment: dataset shift, causality, and shift-stable models in health AI, Biostatistics (2019).

[26]. Che Z, Liu Y, Deep learning solutions to computational phenotyping in health care, in: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 2017, pp. 1100–1109.

[27]. Zhang Y, Yang X, Ivy J, Chi M, Time-aware adversarial networks for adapting disease progression modeling, in: 2019 IEEE International Conference on Healthcare Informatics (ICHI), 2019, pp. 1–11.

[28]. Khoshnevisan F, Chi M, An adversarial domain separation framework for septic shock early prediction across EHR systems, in: 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 64–73.

[29]. Yu Y, Chen P-Y, Zhou Y, Mei J, Adversarial sample enhanced domain adaptation: A case study on predictive modeling with electronic health records, 2021, ArXiv abs/2101.04853.

[30]. Choi E, Bahadori MT, Searles E, Coffey C, Thompson M, Bost J, Tejedor-Sojo J, Sun J, Multi-layer Representation Learning for Medical Concepts, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[31]. Liu W, Singh K, Ryan A, Sukul D, Mahmoudi E, Waljee A, Stansbury C, Zhu J, Nallamothu B, Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding, 2019, BioRxiv.

[32]. Mikolov T, Chen K, Corrado GS, Dean J, Efficient estimation of word representations in vector space, in: ICLR, 2013.

[33]. Pennington J, Socher R, Manning CD, GloVe: Global vectors for word representation, in: EMNLP, 2014.

[34]. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D, Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, NPJ Digit. Med 4 (2021).

[35]. Shang J, Ma T, Xiao C, Sun J, Pre-training of graph augmented transformers for medication recommendation, in: IJCAI, 2019.

[36]. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J, Doctor AI: Predicting clinical events via recurrent neural networks, in: JMLR Workshop and Conference Proceedings, Vol. 56, 2016, pp. 301–318.

[37]. Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl S, Shah N, Language models are an effective patient representation learning technique for electronic health record data, 2020, ArXiv abs/2001.05295.

[38]. Dou Q, Ouyang C, Chen C, Chen H, Heng P, Unsupervised cross-modality domain adaptation of ConvNets for biomedical image segmentations with adversarial loss, in: IJCAI, 2018.

[39]. Chen C, Dou Q, Chen H, Qin J, Heng P-A, Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation, in: AAAI, 2019.

[40]. Wolterink JM, Dinkla AM, Savenije M, Seevinck PR, van den Berg CAT, Isgum I, Deep MR to CT synthesis using unpaired data, in: SASHIMI@MICCAI, 2017.

[41]. Yan W, Wang Y, Xia M, Tao Q, Edge-guided output adaptor: Highly efficient adaptation module for cross-vendor medical image segmentation, IEEE Signal Process. Lett 26 (2019) 1593–1597.

[42]. Wollmann T, Eijkman CS, Rohr K, Adversarial domain adaptation to improve automatic breast cancer grading in lymph nodes, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 582–585, 10.1109/ISBI.2018.8363643.

[43]. Miller T, Bethard S, Amiri H, Savova GK, Unsupervised domain adaptation for clinical negation detection, in: BioNLP, 2017.

[44]. Kouw WM, An introduction to domain adaptation and transfer learning, 2018, ArXiv abs/1812.11806.

[45]. Luo Y, Zhang S-Y, Zheng W-L, Lu B-L, Wgan domain adaptation for EEG-based emotion recognition, in: ICONIP, 2018.

[46]. Bao G, Zhuang N, Tong L, Yan B, Shu J, Wang L, Zeng Y, Shen Z, Two-level domain adaptation neural network for EEG-based emotion recognition, Front. Human Neurosci 14 (2020).

[47]. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky VS, Domain-adversarial training of neural networks, J. Mach. Learn. Res 17 (2016) 59:1–59:35.

[48]. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC, Bengio Y, Generative adversarial nets, in: NIPS, 2014.

[49]. Arjovsky M, Chintala S, Bottou L, Wasserstein GAN, 2017, ArXiv abs/1701.07875.

[50]. Chung J, Kastner K, Dinh L, Goel K, Courville AC, Bengio Y, A recurrent latent variable model for sequential data, in: NIPS, 2015.

[51]. Inoue K, Hsu W, Arah OA, Prosper AE, Aberle DR, Bui AAT, Generalizability and transportability of the national lung screening trial data: Extending trial results to different populations, in: Cancer Epidemiology, Biomarkers & Prevention, A Publication of the American Association for Cancer Research, Cosponsored By the American Society of Preventive Oncology, 2021.

[52]. Sauver JLS, Grossardt BR, Leibson CL, Yawn BP, Melton LJI, Rocca WA, Generalizability of epidemiological findings and public health decisions: an illustration from the rochester epidemiology project., Mayo Clinic Proc. 87 (2) (2012) 151–160.

[53]. Curth A, Thoral PJ, van den Wildenberg W, Bijlstra P, de Bruin D, Elbers PWG, Fornasa M, Transferring clinical prediction models across hospitals and electronic health record systems, in: PKDD/ECML Workshops, 2019.

[54]. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V, RoBERTa: A robustly optimized BERT pretraining approach, 2019, ArXiv abs/1907.11692.

[55]. Radford A, Narasimhan K, Improving language understanding by generative pre-training, 2018.

[56]. Tzeng E, Hoffman J, Saenko K, Darrell T, Adversarial discriminative domain adaptation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2962–2971.

[57]. Shen J, Qu Y, Zhang W, Yu Y, Adversarial representation learning for domain adaptation, 2017, ArXiv abs/1707.01217.

[58]. Rüschendorf L, The wasserstein distance and approximation theorems, Probab. Theory Related Fields 70 (1985) 117–129.

[59]. Villani C, Optimal transport: Old and new, 2008.

[60]. Loshchilov I, Hutter F, Decoupled weight decay regularization, in: ICLR, 2019.

[61]. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J, HuggingFace's transformers: State-of-the-art natural language processing, 2019, ArXiv abs/1910.03771.

[62]. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S, PyTorch: An imperative style, high-performance deep learning library, 2019, ArXiv abs/1912.01703.

[63]. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S, Deepr: A convolutional net for medical records, IEEE J. Biomed. Health Inf 21 (2017) 22–30.

[64]. Choi E, Schuetz A, Stewart WF, Sun J, Using recurrent neural network models for early detection of heart failure onset, J. Am. Med. Inf. Assoc. : JAMIA 24 (2017) 361–370.

[65]. Ma T, Xiao C, Wang F, Health-ATM: A deep architecture for multifaceted patient health record representation and risk prediction, in: SDM, 2018.

[66]. van der Maaten L, Hinton GE, Visualizing data using t-SNE, J. Mach. Learn. Res 9 (2008) 2579–2605.

[67]. Vig J, A multiscale visualization of attention in the transformer model, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Florence, Italy, 2019, pp. 37–42, http://dx.doi.org/10.18653/v1/P19-3007, URL https://www.aclweb.org/anthology/P19-3007.

[68]. Hadi HE, Vincenzo AD, Vettor R, Rossato M, Relationship between heart disease and liver disease: A two-way street, Cells 9 (2020).

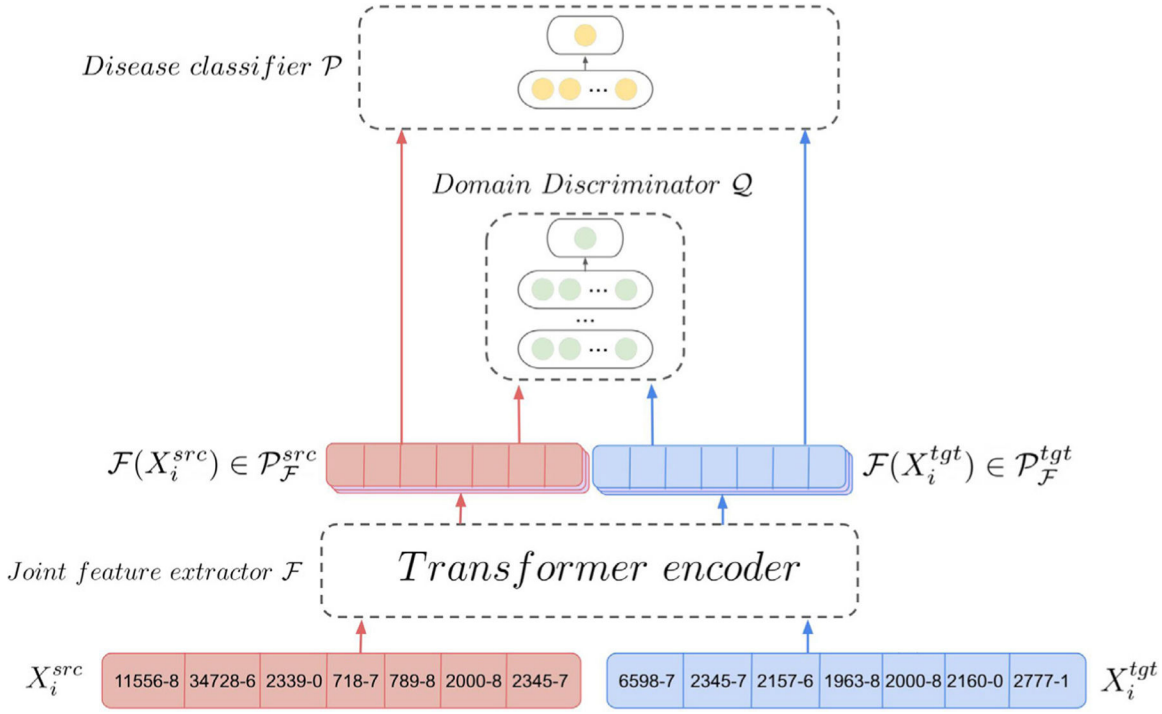


Fig. 1.

Illustration of the proposed AdaDiag framework, consisting of three modules: the joint feature extractor $\mathcal{F}$ that maps sequences from the source and target domain to a shared feature space, the classifier $\mathcal{P}$ that predicts next-visit HF onset and the discriminator $\mathcal{Q}$ for distinguishing source and target domain identity given the features from $\mathcal{F}$.

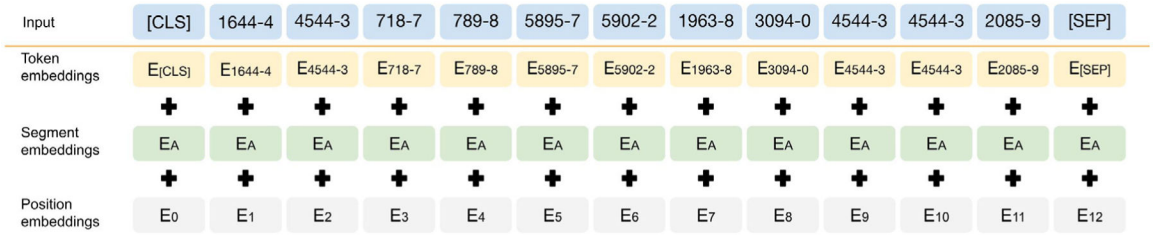


Fig. 2.

BERT-style input representation of the pre-trained Transformer-based model. As defined in the BERT paper [23], the input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Fig. 3.

Illustration of relative performance losses in all baseline and adversarial models, when adapting from MIMIC-IV to UCLA data, calculated as (source metric–target metric)/source metric × 100%.

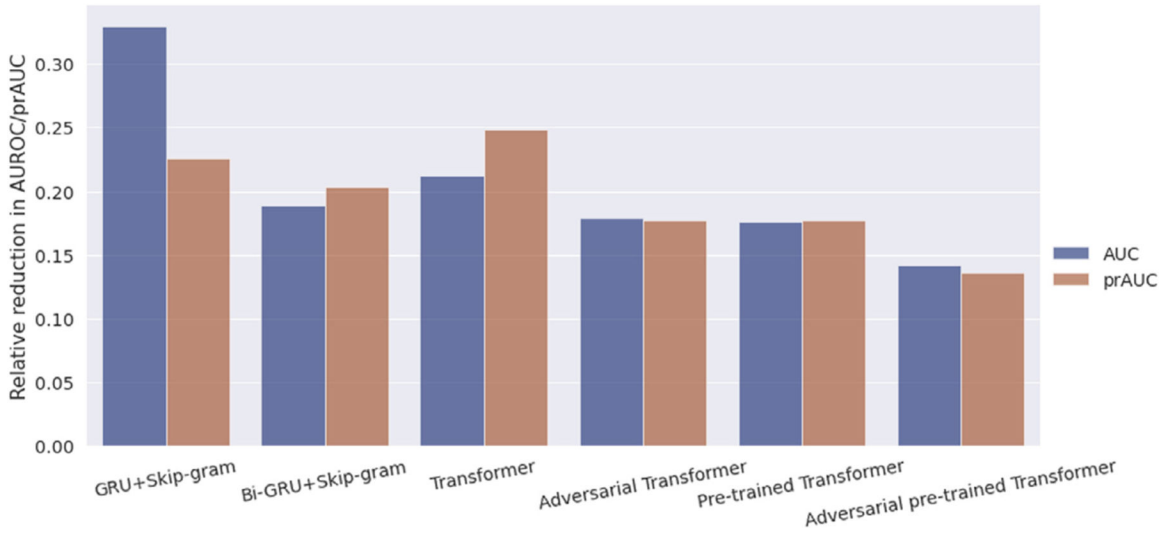
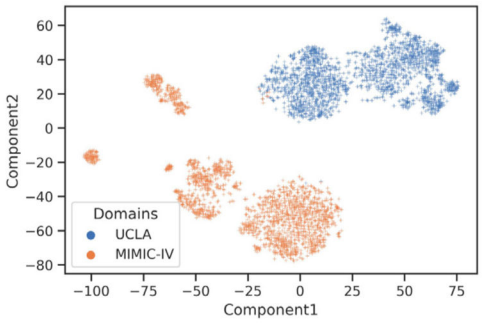
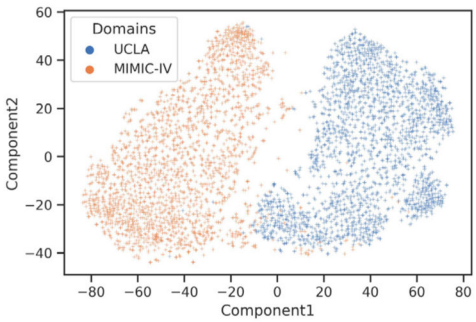


Fig. 4.

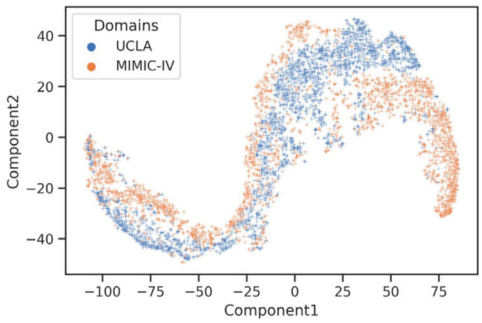
Illustration of relative performance losses in all baseline and adversarial models, when adapting from UCLA to MIMIC-IV data. Relative performance loss is calculated as (source metric–target metric)/source metric × 100%.



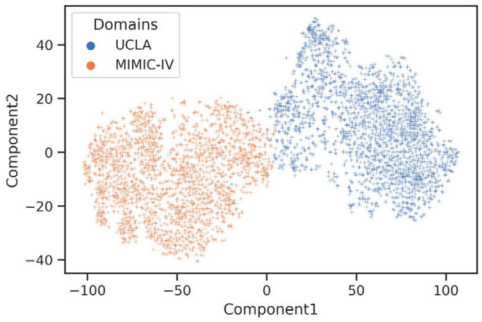
(a) Transformer model pre-trained on unlabeled MIMIC-IV and UCLA sequences.



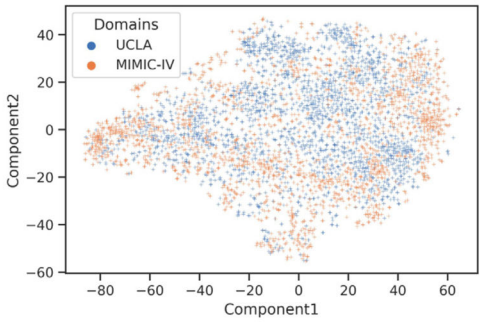
(b) Pre-trained Transformer model fine-tuned on MIMIC-IV data.



(c) Adversarial pre-trained Transformer model for MIMIC-IV to UCLA adaptation.



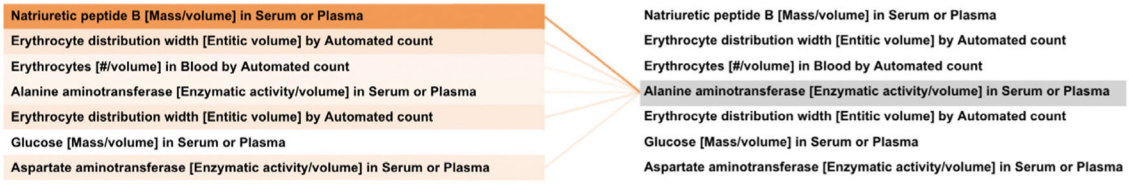
(d) Pre-trained Transformer model fine-tuned on UCLA data.



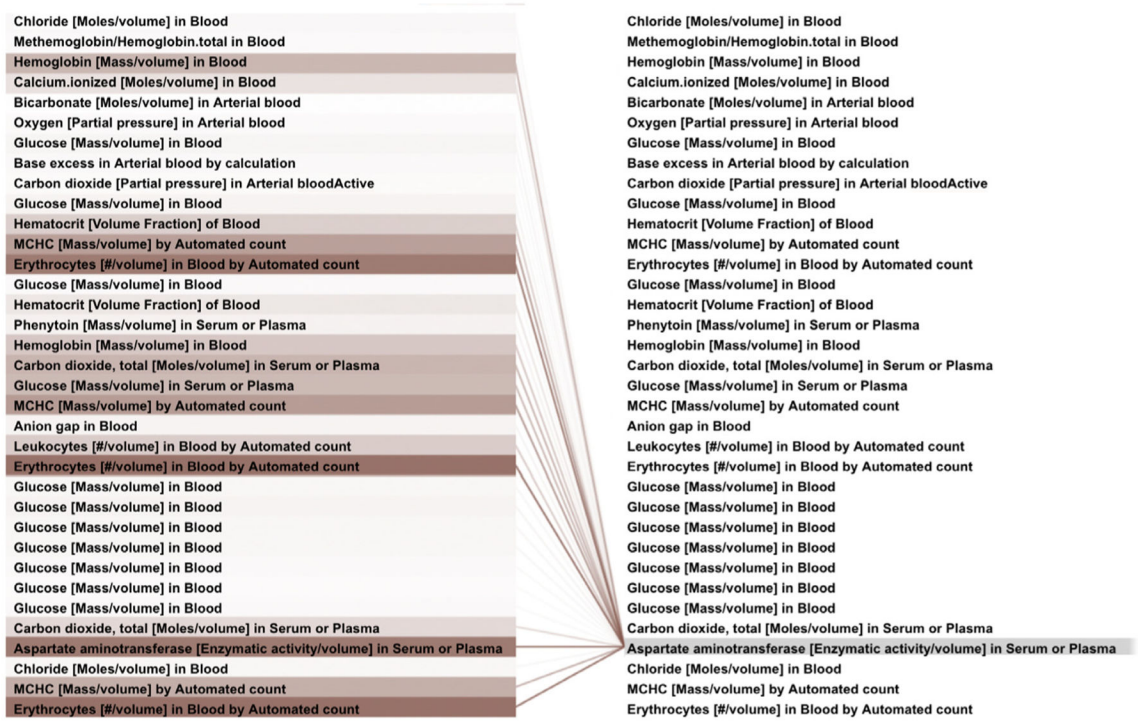
(e) Adversarial pre-trained Transformer model for UCLA to MIMIC-IV adaptation.

Fig. 5.

t-SNE visualizations of activations at the end of the Transformer feature encoders from different models/training stages. Neither pretraining nor fine-tuning were able to bridge the domain gap, whereas adversarial training mixed the distributions between the two datasets effectively.



(a) Abnormal lab event sequence for **patient A**. A strong association between abnormalities in alanine aminotransferase (ALT) and niatriuretic peptide B (BNP) is found.



(b) Abnormal lab event sequence for **patient B**. Abnormal values in hemoglobin or erythrocyte concentration in blood are strongly associated with elevated values in aspartate aminotransferase (AST).

Fig. 6.

Analysis of self-attention in AdaDiag's Transformer encoder layer for MIMIC-IV to UCLA adaptation. Colors of the edges corresponds to individual attention heads from the first Transformer layer (e.g., orange: the second head; brown: the sixth head), and shades of the edges/highlighted region indicate attention weights.

Table 1

Data summary of extracted cohorts from UCLA and MIMIC-IV dataset.

	UCLA	MIMIC-IV
Number of patients	18,736	27,782
Number of visits	283,502	145,961
Avg. number of visits per patient	15.1	5.3
Number of unique lab codes	1218	272
Avg. sequence length per patient	419.5	234.8
Female ratio	43.7%	51.4%
HF incidence rate	16.4%	27.7%

Table 2

MIMIC-IV to UCLA domain adaptation performance comparison. Metrics are reported with 95% confidence interval (CI). The highlighted row shows performance of the proposed model, AdaDiag. We performed statistical t-test to demonstrate that AdaDiag significantly outperformed the best baseline model, pre-trained Transformer baseline, on target domain AUROC and pr-AUC.

	Method	MIMIC-IV(D^{src})		UCLA(D^{tgt})	
		AUROC	pr-AUC	AUROC	pr-AUC
Baselines	GRU+Skip-gram	0.7671 ± .0143	0.5987 ± .0256	0.4628 ± .0114	0.1642 ± .0083
	Bi-GRU+Skip-gram	0.7918 ± .0139	0.6318 ± .0258	0.6623 ± .0096	0.2425 ± .0112
	Transformer	0.7997 ± .0133	0.6525 ± .0243	0.6222 ± .0114	0.2459 ± .0121
	Pre-trained Transformer	0.8000 ± .0134	0.6443 ± .0246	0.6816 ± .0104	0.2828 ± .0133
Adversarial	Transformer	0.7977 ± .0132	0.6468 ± .0244	0.6456 ± .0111	0.2659 ± .0129
	Pre-trained Transformer	0.7985 ± .0131	0.6374 ± .0251	0.7089 ± .0099 **	0.2944 ± .0133 *

For simplicity, we use superscript ** to indicate p value < 0.001, and * to denote p value < 0.01.

Table 3

UCLA to MIMIC-IV domain adaptation performance comparison. Metrics are reported with 95% CI. The highlighted row shows performance of the proposed model, AdaDiag. We performed statistical t-test to demonstrate that AdaDiag significantly outperformed the best baseline model, pre-trained Transformer baseline, on target domain AUROC and pr-AUC.

	Method	UCLA(D^{src})		MIMIC-IV(D^{tgt})	
		AUROC	pr-AUC	AUROC	pr-AUC
Baselines	GRU+Skip-gram	0.7640 ± .0206	0.3971 ± .0370	0.5120 ± .0076	0.3075 ± .0088
	Bi-GRU+Skip-gram	0.8058 ± .0188	0.5064 ± .0424	0.6540 ± .0071	0.4032 ± .0106
	Transformer	0.8004 ± .0197	0.5123 ± .0402	0.6309 ± .0073	0.3851 ± .0105
	Pre-trained Transformer	0.8167 ± .0182	0.5375 ± .0200	0.6727 ± .0072	0.4422 ± .0114
Adversarial	Transformer	0.8018 ± .0193	0.5126 ± .0418	0.6583 ± .0070	0.4110 ± .0108
	Pre-trained Transformer	0.8113 ± .0188	0.5336 ± .0399	0.6959 ± .0069 **	0.4610 ± .0115 **

For simplicity, we use superscript ** to indicate p value < 0.001, and * to denote p value < 0.01.