Predicting Partially Observed Long-Term Outcomes with Adversarial Positive-Unlabeled Domain Adaptation

Mengying Yan MENGYING.YAN@DUKE.EDU
Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA; Duke AI Health, Duke University School of Medicine, Durham, NC, USA

Meng Xia MX41@DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, USA

Wei A. Huang WEI.HUANG@DUKE.EDU
Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA; Duke AI Health, Duke University School of Medicine, Durham, NC, USA

Chuan Hong CHUAN.HONG@DUKE.EDU
Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

Benjamin A. Goldstein BEN.GOLDSTEIN@DUKE.EDU
Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

Matthew M. Engelhard M.ENGELHARD@DUKE.EDU
Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA; Duke AI Health, Duke University School of Medicine, Durham, NC, USA

Abstract

Predicting long-term clinical outcomes often requires large-scale training data with sufficiently long follow-up. However, in electronic health records (EHR) data, long-term labels may not be available for contemporary patient cohorts. Given the dynamic nature of clinical practice, models that rely on historical training data may not perform optimally. In this work, we frame the problem as a positive–unlabeled domain adaptation task, where we seek to adapt from a fully labeled source domain (e.g., historical data) to a partially labeled target domain (e.g., contemporary data). We propose an adversarial framework that includes three core components: (1) Overall Alignment, to match feature distributions between source and target domains; (2) Partial Alignment, to map source negatives to unlabeled target samples; and (3) Conditional Alignment, to address conditional shift using available positive labels in the target domain. We evaluate our method on a benchmark digit classification task (SVHN-MNIST), and two real-world EHR applications: prediction of one-year mortality post COVID-19, and long-term prediction of neurodevelopmental conditions (NDC) in children. In all settings, our approach consistently outperforms baseline models and, in most cases, achieves performance close to an oracle model trained with fully observed labels.

Data and Code Availability This paper uses publicly available image datasets SVHN (Lecun et al., 1998) and MNIST (Netzer et al., 2011). This paper also uses EHR data from the Duke University Health System (DUHS) that cannot be made publicly available. Program code is publicly available and included in the supplementary material.

Institutional Review Board (IRB) This work was covered by our institution's IRB protocols: Pro00111373 and Pro00110219.

1. Introduction

Most clinical decision support tools are derived from electronic health records (EHRs) data (Alexiuk et al., 2024; Sutton et al., 2020; Goldstein et al., 2017a). In recent years, machine learning methods have been increasingly used to perform clinical risk prediction for a variety of outcomes, such as mortality, hospitalization, or disease (Shamout et al., 2021; Naemi et al., 2021; Shickel et al., 2018). Given their dense, granular nature, EHR data are well positioned for pre-

dicting near-term outcomes (Goldstein et al., 2017b). However, predicting long-term health outcomes has gained interest because early detection of risk can enable timely interventions and potentially improve patient prognosis (Nannan Panday et al., 2017).

One challenge in training models to predict long-term outcomes is the need for sufficient follow-up data to capture events far into the future. As such, researchers have typically utilized large epidemiological cohorts to develop clinical risk models for long-term outcomes (D'Agostino et al., 2008). Within an EHR system, however, long-term follow-up can be challenging to obtain due to the fluid nature of patient engagement with clinical care. Even when long-term follow-up is available, an additional challenge arises from potential data shifts, as models trained on historical data may not fully reflect current clinical practices or patient populations. Given the dynamic nature of clinical care, it is often desirable to use more contemporary data for training. This was highlighted during the COVID-19 pandemic, when there was a pressing need to develop models for longer-term outcomes using newly available data, for which there was limited follow-up.

We illustrate this scenario in (Figure 1). The left panel shows the typical training scenario for a historical population, predicting one-year outcomes in 2020 requires training data spanning the entire preceding year (2019). In contrast, the right panel shows the target population scenario, where the goal is to initiate training in early 2020 despite having only one month of follow-up data, rather than waiting for a full year of observations.

Our setting shares some similarities with the online learning paradigm, which focuses on incrementally updating models as new data arrives over time (Shalev-Shwartz, 2011). In our scenario, however, the primary challenge is not the sequential arrival of data *per se*, but rather the absence of long-term outcome labels for newer data points (*i.e.*, the target domain). In other words, we focus on updating our model for newer patients whose features have been observed, but whose labels have been only partially observed; whereas online learning focuses on continually updating models with new features and labels as they arrive.

This scenario is a special case of the outcome observability problem (Yan et al., 2022), where the absence of a long-term outcome is simply unobserved. Within the machine learning literature, this can be viewed as a positive and unlabeled (PU) problem

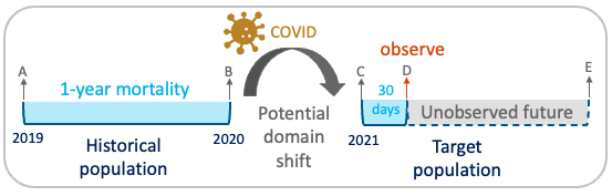

Figure 1: Motivating example illustrating the challenges of predicting long-term outcomes with partially observed outcomes. A-B: outcome observation window for the historical population (fully observed); D: the point at which we aim to deploy the model for the target population; C-D: available outcome observation window for the target population (partially observed); C-E: ideal outcome observation window for the target population.

(Elkan and Noto, 2008), in which the outcomes that are not observed as "positive" are unlabeled rather than definitively negative.

To address the lack of labeled data in a target population, we can leverage information from a different but related population (*e.g.*, a historical cohort in the same EHR system or an external dataset) for which long-term outcomes are already known. As illustrated in Figure 1, while the target population only has data for up to 30 days, a historical cohort may provide labels for one-year mortality. However, directly applying a model trained on the historical cohort to the target population may yield poor performance due to population shift (Subbaswamy and Saria, 2020). Such shifts can arise over time, across institutions, or among different disease categories, leading to significant bias if the source and target populations differ in their underlying characteristics.

In this work, we seek a method to adapt a model from a historical source cohort to a contemporary target cohort while accounting for the partial observability of outcomes in the target data. To achieve these goals, we propose an adversarial domain adaptation framework tailored to PU data in EHRs. Our method features three core components: (1) A overall alignment term that match overall feature distributions between source and target populations; (2) A partial alignment term that ensures source negatives have high likelihood under the target unlabeled distribu-

tion; (3) A conditional alignment to align the conditional distribution using available positive labels in the target domain.

Our contributions are as follows:

- We formalize the challenge of adapting EHR-based long-term outcome prediction models to temporal distribution shifts under limited follow-up.
- To address this challenge, we propose a novel adversarial positive-unlabeled domain adaptation approach that aligns a fully labeled source domain with a partially labeled target domain.
- We validate our method on one public benchmark and two real-world EHR datasets, demonstrating its practical effectiveness and robustness across diverse clinical settings.

2. Related Work

Domain adaptation (DA) strategies are designed to mitigate this problem by adapting a model from a source domain to a related but distinct target domain (Ben-David et al., 2010). By aligning feature distributions or representations across domains, DA can help existing models perform better on the new population of interest. Despite extensive research on domain adaptation in fields such as computer vision and natural language processing, directly applying these approaches to tabular EHR data can be challenging.

For instance, contrastive-learning based DA methods have achieved success in vision tasks (Singh, 2021; Thota and Leontidis, 2021) by learning domain-invariant representations through drawing positive pairs close in representation space while pushing negative pairs apart. They align distributions between source and target domains by ensuring that similar examples remain close despite domain shifts. However, positive pairs are typically constructed through image augmentations—a practice that does not translate easily to the structure of tabular EHR data.

Adversarial learning-based methods, such as Domain-Adversarial Neural Networks (DANN) (Ganin et al., 2017) and Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al., 2017), offer a more promising route for tabular EHR settings. Indeed, several EHR risk prediction studies have successfully explored adversarial techniques (Zhang et al., 2022; Yu et al., 2021; Zhang et al., 2019) for domain adaptation in unsupervised settings. However, Zhang et al. (2022) and Zhang et al. (2019) do not account for the positive–unlabeled nature of our problem, and Yu et al. (2021) focus on generating adversarial samples to improve domain adaptation, rather than using adversarial training to align features between the source and target domains.

Domain adaptation methods for PU settings remain sparse; we are aware of only one existing approach (Sonntag et al., 2022). Our method therefore constitutes a new adversarial framework with partial distribution matching, providing a robust solution for predicting long-term outcomes in EHR data when labels are only partially observable.

3. Method

In this section, we first define our problem (Section 3.1), then describe how we pretrain the model on the source data (Section 3.2). Finally, we detail three components we used to align target features and source features (Section 3.3).

3.1. Problem Definition

We frame our problem as a positive and unlabeled (PU) domain adaptation problem. Let the source domain be $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{n_s}$, where each sample (x_i, y_i) has a fully observed label $y_i \in \{0, 1\}$. The target domain is $\mathcal{D}_t = \mathcal{D}_{tp} \cup \mathcal{D}_{tu}$, where $\mathcal{D}_{tp} = \{(x_i, y_i = 1)\}_{i=1}^{n_{tp}}$ contains observed positive samples, and $\mathcal{D}_{tu} = \{(x_i)\}_{i=1}^{n_{tu}}$ contains the unlabeled samples whose true label may be 0 or 1 but is unobserved (and treated as 0). Our objective is to learn a predictive model for the target domain by leveraging both the fully labeled source domain data ($\mathcal{D}_s$) and the PU target domain data ($\mathcal{D}_t$).

We build upon the Adversarial Discriminative Domain Adaptation (ADDA) framework (Tzeng et al., 2017) , which uses adversarial training to reduce the domain discrepancy. In the standard ADDA setting, a source encoder and a source classifier are first trained on labeled source data, then a target encoder is learned such that its feature representation is aligned with that of the source domain via an adversarial objective. However, ADDA assumes that the target data are unlabeled.

In our PU domain adaptation setting, we do have partial supervision on the target domain (i.e., some

target positives are observed). Therefore, we introduce additional alignment steps to effectively utilize these partially labeled target examples.

Our method involves: a source encoder M_s, a classifier $\mathcal{F}$, a target encoder M_t, and two discriminators D_1 and D_2 for different alignment objectives.

3.2. Source Model Pre-training

We begin with pre-training the source encoder M_s and the classifier $\mathcal{F}$ on the fully labeled source domain using a cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{CE} = &-\mathbb{E}_{(x,y)\sim(X_s,Y_s)}\big[y \log \mathcal{F}(M_s(x))\big] \\ &- \mathbb{E}_{(x,y)\sim(X_s,Y_s)}\big[(1-y)\log\big(1-\mathcal{F}(M_s(x))\big)\big]\end{aligned} \tag{1}$$

After this pre-training, we initialize the target encoder M_t using weights of M_s. In many real-world EHR applications, historical or pre-existing models are available and can be directly used for initialization and fixed thereafter.

3.3. Feature Alignment

To adapt the model from the source domain to the partially labeled target domain, we align the target and source feature distributions. As depicted in Figure 2, our alignment strategy includes three components: (1) Overall Alignment, (2) Partial Alignment and (3) Conditional Alignment.

These components are motivated by the need to align the entire target feature space with the source feature space (Overall Alignment); align the negative portion of the source distribution with the unlabeled portion of the target in a way that accommodates potential hidden positives in the unlabeled set (Partial Alignment); and ensure that the conditional distribution $P(Y \mid X)$ is also well-aligned by leveraging the known target positives (Conditional Alignment).

We train the target encoder M_t adversarially using two discriminators D_1 and D_2 to achieve overall and partial alignment (Sections 3.3.1 and 3.3.2), respectively. Then incorporate conditional alignment by optimizing M_t with a supervised loss of $\mathcal{F}(M_t(.))$ on D_{tp}. (Section 3.3.3). Finally, we summarize the overall objective in Section 3.3.4.

3.3.1. Overall Alignment

To align the feature distributions, we introduce a discriminator D_1, that learns to distinguish between features from the source encoder $M_s(x)$ and the target encoder $M_t(x)$. We train D_1 using a binary cross-entropy loss to classify domain indicators (source = 1 vs. target = 0):

$$\begin{aligned}\mathcal{L}_{D_1} = &-\mathbb{E}_{x\sim X_s}\big[\log D_1(M_s(x))\big] \\ &-\mathbb{E}_{x\sim X_t}\big[\log\big(1-D_1(M_t(x))\big)\big].\end{aligned} \tag{2}$$

This follows a standard adversarial setup in which D_1 is trained to separate the source and target distributions, while target encoder M_t is trained to prevent D_1 from doing so effectively by matching the source and target distributions in the representation space. This is done by employing the following inverted-label GAN loss (Tzeng et al., 2015) on the target domain:

$$\mathcal{L}_{M_t} = -\mathbb{E}_{x\sim X_t}\big[\log D_1(M_t(x))\big]. \tag{3}$$

This is a cross-entropy loss but with inverted domain indicators (target = 1). In other words, we treat the target features as if they were from the source domain to align the overall feature distribution of target with source. Overall alignment helps reduce overall domain discrepancy, ensuring that target features live in a feature space more similar to source features.

3.3.2. Partial Distribution Alignment

Although overall alignment brings the entire target distribution closer to the source, it treats all target samples uniformly. In PU learning, however, the unlabeled set $\mathcal{D}_{tu}$ contains both negative and positive samples, but we do not know which are truly positive. Simply forcing all target unlabeled samples to match source negatives can be detrimental, because some unlabeled target samples may be positive.

To address this, we introduce a second discriminator D_2 focusing on aligning source negatives ($\mathcal{D}_{sn}$) with unlabeled target data ($\mathcal{D}_{tu}$), while still permitting positives in the unlabeled set to deviate if needed. We introduce a KL-divergence-based loss that encourages the negative portion of the source distribution to align with a corresponding portion of the unlabeled target distribution, while allowing the support of the unlabeled target samples, which include positives, to extend beyond that of the source negative samples:

$$\mathcal{L}_{D_2} = KL\big(p_{sn} \parallel p_{tu}\big) \tag{4}$$

where p_{sn} and p_{tu} denote the distributions of source negatives and target unlabeled data, respectively. Ghimire et al. (2021) showed that KL-divergence of probability density function $p(x)$ and $q(x)$ is given by

$$KL\big(p(x) \parallel q(x)\big) = \mathbb{E}_{x\sim p(x)}\Big[\sigma^{-1}D(x)\Big], \tag{5}$$

where D is a GAN type discriminator, and $\sigma^{-1}(x)$ is the logit function given by $\log\left(\frac{p(x)}{1-p(x)}\right)$. Thus in our case,

$$\mathcal{L}_{D_2} = \mathbb{E}_{x\sim p_{sn}(x)}\Big[\sigma^{-1}D_2(x)\Big]. \tag{6}$$

In practice, it can be estimated by

$$\frac{1}{n_{sn}}\sum_{x\in\mathcal{D}_{sn}}\sigma^{-1}\big(D_2(M_s(x))\big). \tag{7}$$

We again adopt an adversarial objective for the target encoder M_t, but only on the unlabeled target samples. We use a loss similar to (3):

$$\mathcal{L}_{\text{part}M_t} = -\mathbb{E}_{x\sim X_{tu}}\big[\log D_2\big(M_t(x)\big)\big]. \tag{8}$$

Here, we treat the unlabeled target features as if they were from the source negatives, thereby encouraging their representations to match.

3.3.3. Conditional Alignment

Even with overall and partial alignment, the conditional distribution $P(Y|X)$ may still differ between source and target domains. This is often referred to as conditional shift, can severely degrade performance if not addressed (Zhao et al., 2019).

To address the conditional shift, we leverage the observed target positives $\mathcal{D}_{tp}$. Specifically, we require M_t to classify these known positives correctly under the source-trained classifier $\mathcal{F}$. Hence, we introduce a supervised loss on $\mathcal{D}_{tp}$. Since all observed labels $y = 1$, the cross-entropy loss can be written as:

$$\mathcal{L}_{\text{sup}M_t} = -\mathbb{E}_{x\sim X_{tp}}\big[\log \mathcal{F}\big(M_t(x)\big)\big]. \tag{9}$$

This ensures that the learned target representations produce strong positive predictions in the target domain. Such conditional alignment leverages the label information in the target domain to reduce model shift and improve classification performance.

3.3.4. Overall Training Objective

Putting it all together, the overall training process consists of training two discriminators (D_1, D_2) and training the target encoder M_t jointly. We set the parameters of D_1 and D_2 to minimize $\mathcal{L}_{D_1}$ and $\mathcal{L}_{D_2}$, and set the parameters of M_t to minimize the combined loss $\mathcal{L}_{\text{comb}M_t}$ for M_t that includes an inverted-label GAN loss on all target, an inverted-label GAN loss on target unlabeled, and a supervised loss on target positives:

$$\mathcal{L}_{\text{comb}M_t} = \lambda_1\mathcal{L}_{M_t} + \lambda_2\mathcal{L}_{\text{part}M_t} + \lambda_3\mathcal{L}_{\text{sup}M_t}. \tag{10}$$

The proposed framework is shown in Figure 2. By building upon ADDA and incorporating partial distribution alignment for negative classes and conditional alignment via limited positive target labels, our framework addresses the unique challenges of PU domain adaptation.

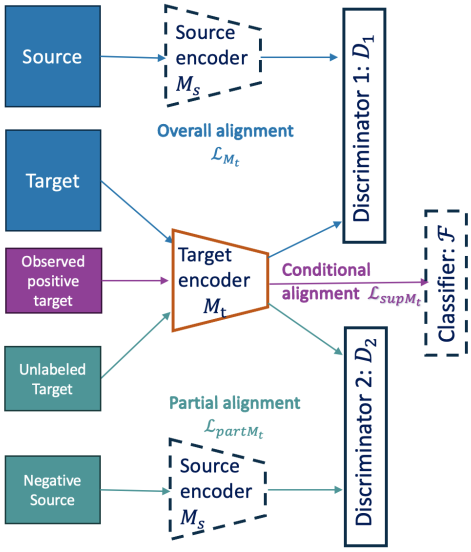

Figure 2: Illustration of the proposed framework with three alignment components: overall alignment, partial alignment, and conditional alignment. The source encoder and classifier are from the pre-training step.

4. Experiments

To illustrate our method, we first conduct a *proof of concept* analysis on the SVHN and MNIST datasets.

We then move to our two motivating real-world EHR based examples. We compare the proposed method to: (1) An oracle model that trains on target data with fully observed labels; (2) A baseline model trained exclusively on the (historical) source data without adaptation; (3) A model trained only on the partially observed target data; (4) Unsupervised ADDA.

To assess the contributions of each alignment component, we also conduct an ablation study for each application. We compare several variants of our proposed method: (1) Only using overall alignment ($\mathcal{L}_{M_t}$ (ADDA)); (2) Only using overall and partial alignments ($\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t}$); (3) Only using overall and conditional alignments ($\mathcal{L}_{M_t} + \mathcal{L}_{\text{sup}M_t}$); (4) Only using conditional alignments ($\mathcal{L}_{\text{sup}M_t}$); (5) Using all three components ($\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t} + \mathcal{L}_{\text{sup}M_t}$).

4.1. Proof of Concept: Digits Classification

We first illustrate our proposed method on a well-known domain adaptation task involving the SVHN and MNIST datasets. We use the SVHN dataset as the source domain and a stylized variant of MNIST (drawn from MNIST-C (Mu and Gilmer, 2019)) as the target domain. We randomly assign half of the MNIST target samples to the canny-edge style and the other half to the motion-blur style.

Since we focus on a binary classification task, we restrict our dataset to two digits: 3 (positive) and 5 (negative). We then create a PU setting in the target MNIST data by assigning observed positive labels with probability p_{obs1} for the canny-edge images, and p_{obs2} for the motion-blur images. This design generates different label-selection mechanisms, allowing us to test the robustness of our method under varying degrees of label observability.

4.2. Real Data Application 1: One-year Mortality Prediction After COVID-19

Our first real-world application aims to predict one-year mortality following emergency department (ED) visits, using data shortly after the COVID-19 pandemic commenced. This is a critical task given that ED patients often exhibit elevated risk of mortality (Gunnarsdottir and Rafnsson, 2006) and early identification of high-risk patients can facilitate timely interventions and potentially improve clinical outcomes (Hung et al., 2017; Dundar et al., 2016). While ample historical is available, given the changing nature of clinical care after COVID-19, it is reasonable to presume that there was meaningful domain-shift, therefore any model implemented in current clinical practice should be trained primarily on *post* COVID-19 data (Figure 1).

We use data from our institution's EHR system to define two cohorts: a historical source cohort (all ED encounters in 2018) with fully observed one-year mortality, and a target cohort (all ED encounters in 2021). Although we have full observability for the target cohort here, we simulate situations such that we only observe shorter outcomes in the target data, as if it were immediately after the COVID-19 outbreak. Specifically, we set up the target data to observe 7-day, 30-day, and 90-day mortality, which correspond to 17.3%, 31.1% and 51.7% actuarial outcomes respectively. Predictors including demographics, vital signs and disease diagnosis, with a total of 107 features. Table 7 in Appendix A summarizes the baseline characteristics of the historical and target cohorts. We selected a sample size of 20,000 from each cohort as the target training data and source training data, and an additional 10,000 samples from the 2021 cohort as the target test data, ensuring no overlap in patients between the training and test sets.

4.3. Real Data Application 2: Long-term prediction of neurodevelopmental conditions

Neurodevelopmental conditions (NDCs) include a range of conditions such as autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), motor delay (MD), language delay (LD), and developmental delays (DD) that typically manifest in early childhood (Thapar et al., 2017). Early detection of these conditions is crucial for initiating services to impact long-term outcomes (Cioni et al., 2016). In this application, our goal is to use clinical features available at 18 months old to predict whether a child will have an NDC diagnosis by 5.5 years of age (typical school age and 4 years in the future). Given the long-term nature of the outcome, to properly train a model in 2024, we could only use data as late as 2019. Ideally, we want to incorporate more contemporary data that captures evolving diagnostic standards and care practices. The proposed method will allow us to train the model with more contemporary yet partially observed outcome data.

Following the cohort definition and feature engineering in Huang et al. (2024), we use embeddings

Table 1: AUROC for digit classification (3vs.5) task on stylized MNIST data with 95% bootstrap confidence intervals. Obs1 and Obs2 are labeling probabilities for canny edges and motion blur styles, respectively. Oracle model: trained on MNIST with fully observed labels; Source model: trained on SVHN without adaptation; Target model: trained on the partially observed MNIST. ADDA: unsupervised baseline DA method. Bolded values indicate the best performance for each scenario.

Model		Oracle (full obs)	Source (w/o adapt)	Target (part obs)	ADDA	Proposed
Obs1	Obs2					
10%	1%	0.999 (0.999,1.000)	0.908 (0.898,0.916)	0.916 (0.912,0.938)	0.956 (0.951,0.963)	**0.980** (0.974,0.981)
10%	5%	0.999 (0.999,1.000)	0.908 (0.898,0.916)	0.933 (0.922,0.962)	0.951 (0.951,0.963)	**0.991** (0.990,0.994)
1%	1%	0.999 (0.999,1.000)	0.908 (0.898,0.916)	0.816 (0.801,0.856)	0.935 (0.951,0.963)	**0.975** (0.971,0.982)

of diagnosis and procedure codes extracted from the first 18 months of life. We divide the dataset into a historical source population (children born in 2014) and a target population (children born in 2017). We assume that, as of the current time, we only observe outcomes for up to 6 months, 1 year, or 2 years from the index date in the target population. These observation windows correspond to 14.3%, 30.4%, and 46.7% of full label availability, respectively. Table 8 in Appendix A summarizes the characteristics of the historical and target cohorts at the index encounter.

4.4. Hyperparameter Settings

We set the parameters for the combined loss function as $\lambda_1 = 1$, $\lambda_2 = 1$, and $\lambda_3 = 1$, corresponding to the overall alignment, partial alignment, and conditional alignment components, respectively. Model training was performed using the Adam optimizer with a learning rate of 0.001 for the discriminators (D_1, D_2) and the target encoder (M_t), and a batch size of 128. Each model was trained for 100 epochs. A sensitivity analysis of λ_1, λ_2, and λ_3 is presented in the Results section to assess the robustness of model performance to different hyperparameter settings.

5. Results

5.1. Digits Classification

Table 1 shows the AUROC averaged over 10 runs for the digits classification application. Despite having very few labeled samples in the target domain, our proposed method achieves the highest AUROC across all three label observability conditions, significantly outperforming baseline methods.

5.2. Real Data Application 1: One-year Mortality Prediction After COVID-19

Table 2 summarizes the AUROC results for predicting one-year mortality after an ED encounter under various approaches and observation windows. As a reference, if we had access to the full one-year outcome data in the target, the best performing model (i.e., oracle model) would achieve an AUROC of 0.907. Directly applying the model trained solely on historical data yields an AUROC of 0.869, reflecting the distribution mismatch between source and target. Training on the partially observed target data alone (i.e., using only the first 90, 30, or 7 days of outcomes) gives lower AUROC than the oracle model. Nonetheless, performance improves as the outcome observation window increases (from 7 days to 90 days). Using the unsupervised ADDA alignment boosts performance, demonstrating the benefit of adversarial domain adaptation for this task. Our proposed approach achieves the highest AUROC across all observation windows.

Notably, with just 90 days of outcome data—about half of the full observation period—we already attain an AUROC of 0.909, which is close to and even slightly above the 0.907 achieved by the oracle model. This indicates that partial and conditional alignment, in addition to overall alignment, effectively transfers

Table 2: AUROC results for one-year mortality prediction on target data with 95% bootstrap confidence intervals, given partial observations of the target outcome (90 days, 30 days, or 7 days). Event rates and outcome observability levels for each scenario are reported. Oracle model: trained on post-COVID target data with fully observed labels; Historical source model: trained on historical data without adaptation; Target model: trained exclusively on the partially observed target. ADDA: unsupervised baseline DA method. Bolded values indicate the best performance for each scenario.

Model		Oracle (full obs)	Historical (w/o adapt)	Target (part obs)	ADDA	Proposed
Event rate	Observe window					
3.6%	90 days	0.907 (0.894,0.915)	0.869 (0.854,0.882)	0.883 (0.866,0.892)	0.885 (0.879,0.902)	**0.909** (0.899,0.918)
2.2%	30 days	0.907 (0.894,0.915)	0.869 (0.854,0.882)	0.858 (0.842,0.872)	0.885 (0.879,0.902)	**0.903** (0.889,0.911)
1.2%	7 days	0.907 (0.894,0.915)	0.869 (0.854,0.882)	0.820 (0.801,0.835)	0.885 (0.879,0.902)	**0.900** (0.884,0.907)

knowledge from the historical source data and leverages the labels in target data. We have also included the results of the AUPRC and F1 score in Table 10 and Table 11 in Appendix B.

Overall, these results demonstrate that our proposed method with only limited outcome can be close to the performance of an oracle model trained with full one-year outcome labels in the target data.

5.3. Real Data Application 2: Long-term prediction of neurodevelopmental conditions

As shown in Table 3, models using historical source data without adaptation and model trained on partially observed target data both fall notably short of the oracle performance. Baseline unsupervised ADDA does improve the AUROC to 0.786 as it aligns the overall features of target with the historical data. Our proposed method gives the best AUROC among all three limited observability scenarios. Even in the most limited scenario (6 months of available outcome, i.e., 14.3% of the full data), the proposed model obtains an AUROC of 0.813, nearly matching the oracle model's 0.818. We have also included the results of the AUPRC and F1 score in Table 10 and Table 11 in Appendix B.

5.4. Ablation study

For SVHN-MNIST example, we evaluate different components of the proposed method in Table 4. We report overall AUROC and subgroup AUROC for canny edges and motion blur styles. As shown in Table 4, adding partial alignment on top of overall alignment ($\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t}$) generally boosts performance, especially in cases with higher positive-label observability. Adding conditional alignment ($\mathcal{L}_{M_t} + \mathcal{L}_{\text{sup}M_t}$ or $\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t} + \mathcal{L}_{\text{sup}M_t}$) further increases both the overall and subgroup AUROC, leveraging the small pool of known positives. However, in these cases when observability is extremely low, conditional alignment alone may not help because the very small number of positive labels can lead the model to overemphasize the positive class. Overall, the combination of all three components (overall, partial, and conditional alignment) yields the best performance.

The AUROC results for prediction one-year mortality and predicting long-term NDC in Table 5 show that in contrast to the digit classification task, including our partial alignment term in addition to overall alignment does not substantially improve performance. This may be because overall adversarial alignment already reduces much of the distribution mismatch, and in contrast to the classification task, this is a risk prediction setting in which positive and negative distributions are not separable. On the other hand, adding conditional alignment significantly improves performance over overall alignment alone. Enforcing correct classification of known positives in the target domain helps reduce label shift and results in higher AUROC. However, using conditional

Table 3: AUROC results for long-term NDC prediction on target data with 95% bootstrap confidence intervals, given partial observations of the target outcome (6 months, 1 year, 2 years). Event rates and outcome observability levels for each scenario are reported. Oracle model: trained on target data with fully observed labels; Historical source model: trained on historical data without adaptation; Target model: trained exclusively on the partially observed target. ADDA: unsupervised baseline DA method. Bolded values indicate the best performance for each scenario.

Model		Oracle (full obs)	Historical (w/o adapt)	Target (part obs)	ADDA	Proposed
Event rate	Observe window					
10.0%	2 years	0.818 (0.788,0.840)	0.779 (0.754,0.808)	0.772 (0.729,0.787)	0.786 (0.761,0.815)	**0.814** (0.787,0.840)
6.5%	1 years	0.818 (0.788,0.840)	0.779 (0.754,0.808)	0.755 (0.725,0.783)	0.786 (0.761,0.815)	**0.813** (0.782,0.835)
3.1%	6 months	0.818 (0.788,0.840)	0.779 (0.754,0.808)	0.724 (0.695,0.754)	0.786 (0.761,0.815)	**0.813** (0.781,0.830)

alignment in isolation is not sufficient: Without adversarial alignment to anchor the representations, the model can overemphasize the positive class, harming performance on negatives. Apart from AUROC, we observed that adding conditional alignment can increase the calibration-in-the-large in Appendix B. Because conditional alignment shifts predictions toward positive outcomes, this underscores the importance of balancing all alignment components.

Additionally, details on training time for each method and dataset are provided in Appendix C.

5.5. Sensitivity Analysis

To assess the robustness of our method to the choice of hyperparameters, we conducted a sensitivity analysis by varying the values of the parameters λ_1, λ_2, and λ_3 in the three applications. These parameters control the overall alignment, partial alignment and conditional alignment components in the combined loss function of the target encoder. For each application, we evaluate model performance by varying one λ at a time (values set to 1, 2, and 10) while keeping the other two fixed. The results, summarized in Table 6, show that AUROC remains relatively stable across different λ values, suggesting that the method is not overly sensitive to the exact choice of hyperparameters.

6. Discussion

In this paper, we address the challenge of developing a clinical prediction model for long-term outcomes before the long-term outcomes are fully observed in the target data. This question is particularly relevant when one wants to account for changing clinical practices over time. We formulated this setting as a positive–unlabeled domain adaptation problem, which incorporates the dual tasks of adaptation across source and target domains to mitigate population shift, and partial label observability to handle unlabeled data that may still be positive in the future.

To tackle these challenges, we proposed a new adversarial framework that extends ADDA with three key components: (1) Overall Alignment to align the overall feature distribution between source and target; (2) Partial Alignment to match source negatives with the unlabeled portion of target data; and (3) Conditional Alignment to directly utilize the limited positive labels in the target domain.

Through experiments on a benchmark digit classification task (SVHN to stylized MNIST) and two real-world EHR applications (one-year mortality prediction and long-term NDC prediction), we demonstrated that the proposed method consistently outperforms baseline models. Notably, our approach achieves performance comparable to a model trained with fully observed labels (oracle model) in most scenarios.

In practice, this framework can be integrated into existing clinical workflows, allowing health systems to

Table 4: AUROC comparison across different components for digits classification. Subgroup performance is reported for canny-edge and motion-blur styles. $\mathcal{L}_{M_t}$: overall alignment; $\mathcal{L}_{\text{part}M_t}$: partial alignment; $\mathcal{L}_{\text{sup}M_t}$: conditional alignment.

Model		$\mathcal{L}_{M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t} + \mathcal{L}_{\text{sup}M_t}$
Obs	Style					
	both	0.956	0.965	0.981	0.5	0.980
10%	canny edges	0.934	0.957	0.984	0.5	0.989
1%	motion blur	0.973	0.975	0.969	0.5	0.972
	both	0.951	0.963	0.987	0.5	0.991
10%	canny edges	0.934	0.960	0.994	0.5	0.995
5%	motion blur	0.971	0.964	0.981	0.5	0.987
	both	0.935	0.955	0.966	0.5	0.975
1%	canny edges	0.902	0.939	0.969	0.5	0.989
1%	motion blur	0.969	0.973	0.963	0.5	0.967

use historical cohorts to train robust models that predict long-term outcomes even when the current population only has short-term outcome data available. By aligning feature representations and considering available positive labels, our method can facilitate long-term risk prediction and potentially improve patient outcomes.

Although our motivating example focuses on using historical data, the proposed method can be generalized to consider any fully labeled *source* dataset, applied to a partially observed *target* dataset. For example, we may have source data emanating from an external cohort such as larger payer databases, clinical trials, or epidemiological cohorts. In these settings, it is reasonable to assume that there is some degree of domain shift to a target health system data context. Moreover, if the outcome can take place over an extended observation window, it is unlikely to be fully observed. This work therefore fits into a larger framework regarding the use of external data sources to address observability challenges in EHR-based datasets (Yan et al., 2025).

Despite promising results, our approach has two primary limitations. First, using conditional alignment can lead to larger calibration error, therefore additional re-calibration strategies may be needed. Second, achieving balance among the three alignment components is non-trivial, and determining the optimal weight requires careful tuning. Since the target data does not have fully observed labels, we could not perform cross-validation to tune hyperparameters and therefore used fixed values for the λ terms. Future research could explore ways to perform hyperparameter searching in settings with partially observed outcomes.

Overall, this work highlights the importance of domain adaptation and positive–unlabeled learning as complementary strategies for bridging the gap between historical datasets and current patient populations in need of timely, long-term risk predictions. We hope our findings encourage further exploration of adaptive techniques that can make the most of limited outcome labels in clinical practice.

Acknowledgments

This work was supported by the Autism Centers of Excellence Award P50HD093074 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), National Institutes of Health.

Matthew Engelhard is supported by grant K01MH127309 from the National Institute of Mental Health (NIMH).

Health Data Science at Duke is supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1 TR002553. The Duke AI Health Data Science Fellowship Program is supported by the above grant, the Duke Department of Biostatistics & Bioinformatics, and Duke AI Health. The Duke Protected Analytics Computing Environment (PACE) program is supported by the above grant and by Duke University Health System. The

Table 5: AUROC comparison in the ablation studies for one-year mortality prediction and long-term NDC prediction with different observation windows with 95% bootstrap confidence intervals. $\mathcal{L}_{M_t}$: overall alignment; $\mathcal{L}_{\text{part}M_t}$: partial alignment; $\mathcal{L}_{\text{sup}M_t}$: conditional alignment.

Model	$\mathcal{L}_{M_t}$(ADDA)	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t} + \mathcal{L}_{\text{sup}M_t}$
1-year mortality					
Observe window					
90 days	0.885	0.887	0.908	0.834	0.909
	(0.879,0.902)	(0.867,0.892)	(0.899,0.917)	(0.811,0.848)	(0.899,0.918)
30 days	0.885	0.885	0.905	0.836	0.903
	(0.879,0.902)	(0.876,0.897)	(0.893,0.914)	(0.817,0.830)	(0.889,0.911)
7 days	0.885	0.885	0.903	0.842	0.900
	(0.879,0.902)	(0.871,0.891)	(0.891,910)	(0.819,0.852)	(0.884,0.907)
long-term NDC					
Observe window					
2 years	0.786	0.786	0.814	0.587	0.805
	(0.761,0.815)	(0.769,0.821)	(0.789,0.842)	(0.561,0.625)	(0.787,0.840)
1 year	0.786	0.786	0.813	0.599	0.804
	(0.761,0.815)	(0.768,0.819)	(0.786,0.838)	(0.577,0.642)	(0.782,0.835)
6 months	0.786	0.783	0.813	0.604	0.798
	(0.761,0.815)	(0.755,0.812)	(0.783,0.835)	(0.589,0.656)	(0.781,0.830)

Table 6: Sensitivity analysis of AUROC across different λ_1, λ_2, and λ_3 values in different applications.

(λ_1, λ_2, λ_3)	(1,1,1)	(2,1,1)	(10,1,1)	(1,2,1)	(1,10,1)	(1,1,2)	(1,1,10)
1-year mortality							
Observe 30 days	0.903	0.898	0.895	0.900	0.901	0.901	0.891
Long-term NDC							
Observe 1 year	0.813	0.790	0.787	0.807	0.792	0.804	0.811
Digits classification							
Obs1=10%, Obs2=5%	0.991	0.989	0.992	0.898	0.987	0.993	0.992

content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Mackenzie Alexiuk, Heba Elgubtan, and Navdeep Tangri. Clinical Decision Support Tools in the Electronic Medical Record. *Kidney International Reports*, 9(1):29–38, January 2024. ISSN 2468-0249. doi: 10.1016/j.ekir.2023.10.019. URL https://www.sciencedirect.com/science/article/pii/S2468024923015590.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, May 2010. ISSN 1573-0565. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Giovanni Cioni, Emanuela Inguaggiato, and Giuseppina Sgandurra. Early intervention in neurodevelopmental disorders: underlying neural mechanisms. *Developmental Medicine & Child Neurology*, 58(S4):61–66, 2016. ISSN 1469-8749. doi: 10.1111/dmcn.13050. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/dmcn.13050. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/dmcn.13050.

Ralph B. D'Agostino, Ramachandran S. Vasan, Michael J. Pencina, Philip A. Wolf, Mark Cobain, Joseph M. Massaro, and William B. Kannel. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, 117(6):743–753, February 2008. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.107.699579.

Zerrin Defne Dundar, Mehmet Ergin, Mehmet A. Karamercan, Kursat Ayranci, Tamer Colak, Alpay Tuncar, Basar Cander, and Mehmet Gul. Modified Early Warning Score and VitalPac Early Warning Score in geriatric patients admitted to emergency department. *European Journal of Emergency Medicine: Official Journal of the European Society for Emergency Medicine*, 23(6):406–412, December 2016. ISSN 1473-5695. doi: 10.1097/MEJ.0000000000000274.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 213, Las Vegas, Nevada, USA, 2008. ACM Press. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401920. URL http://dl.acm.org/citation.cfm?doid=1401890.1401920.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, pages 189–209. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58346-4 978-3-319-58347-1. doi: 10.1007/978-3-319-58347-1_10. URL http://link.springer.com/10.1007/978-3-319-58347-1_10. Series Title: Advances in Computer Vision and Pattern Recognition.

Sandesh Ghimire, Aria Masoomi, and Jennifer Dy. Reliable Estimation of KL Divergence using a Discriminator in Reproducing Kernel Hilbert Space. In *Advances in Neural Information Processing Systems*, volume 34, pages 10221–10233. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/54a367d629152b720749e187b3eaa11b-Abstract.html.

Benjamin A. Goldstein, Ann Marie Navar, Michael J. Pencina, and John P. A. Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 24(1):198–208, January 2017a. ISSN 1527-974X. doi: 10.1093/jamia/ocw042.

Benjamin A. Goldstein, Michael J. Pencina, Maria E. Montez-Rath, and Wolfgang C. Winkelmayer. Predicting mortality over different time horizons: which data elements are needed? *Journal of the American Medical Informatics Association: JAMIA*, 24(1):176–181, January 2017b. ISSN 1527-974X. doi: 10.1093/jamia/ocw057.

O. S. Gunnarsdottir and V. Rafnsson. Mortality of the users of a hospital emergency department. *Emergency Medicine Journal*, 23(4):269–273, April 2006. ISSN 1472-0205, 1472-0213. doi: 10.1136/emj.2005.026690. URL https://emj.bmj.com/

content/23/4/269. Publisher: British Association for Accident and Emergency Medicine Section: Original Article.

Wei A. Huang, Matthew Engelhard, Marika Coffman, Elliot D. Hill, Qin Weng, Abby Scheer, Gary Maslow, Ricardo Henao, Geraldine Dawson, and Benjamin A. Goldstein. A conditional multi-label model to improve prediction of a rare outcome: An illustration predicting autism diagnosis. *Journal of Biomedical Informatics*, 157:104711, September 2024. ISSN 1532-0464. doi: 10.1016/j.jbi.2024.104711. URL https://www.sciencedirect.com/science/article/pii/S1532046424001291.

Shang-Kai Hung, Chip-Jin Ng, Chang-Fu Kuo, Zhong Ning Leonard Goh, Lu-Hsiang Huang, Chih-Huang Li, Yi-Ling Chan, Yi-Ming Weng, Joanna Chen-Yeen Seak, Chen-Ken Seak, and Chen-June Seak. Comparison of the Mortality in Emergency Department Sepsis Score, Modified Early Warning Score, Rapid Emergency Medicine Score and Rapid Acute Physiology Score for predicting the outcomes of adult splenic abscess patients in the emergency department. *PloS One*, 12(11):e0187495, 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0187495.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791. URL https://ieeexplore.ieee.org/document/726791. Conference Name: Proceedings of the IEEE.

Norman Mu and Justin Gilmer. MNIST-C: A Robustness Benchmark for Computer Vision, June 2019. URL http://arxiv.org/abs/1906.02337. arXiv:1906.02337 [cs].

Amin Naemi, Thomas Schmidt, Marjan Mansourvar, Mohammad Naghavi-Behzad, Ali Ebrahimi, and Uffe Kock Wiil. Machine learning techniques for mortality prediction in emergency departments: a systematic review. *BMJ Open*, 11(11):e052663, November 2021. ISSN 2044-6055, 2044-6055. doi: 10.1136/bmjopen-2021-052663. URL https://bmjopen.bmj.com/content/11/11/e052663. Publisher: British Medical Journal Publishing Group Section: Emergency medicine.

R. S. Nannan Panday, T. C. Minderhoud, N. Alam, and P. W. B. Nanayakkara. Prognostic value of early warning scores in the emergency department (ED) and acute medical unit (AMU): A narrative review. *European Journal of Internal Medicine*, 45:20–31, November 2017. ISSN 1879-0828. doi: 10.1016/j.ejim.2017.09.027.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Shai Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2011. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000018. URL http://www.nowpublishers.com/article/Details/MAL-018.

Farah Shamout, Tingting Zhu, and David A. Clifton. Machine Learning for Clinical Outcome Prediction. *IEEE Reviews in Biomedical Engineering*, 14:116–126, 2021. ISSN 1941-1189. doi: 10.1109/RBME.2020.3007816. URL https://ieeexplore.ieee.org/abstract/document/9134853. Conference Name: IEEE Reviews in Biomedical Engineering.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, September 2018. ISSN 2168-2208. doi: 10.1109/JBHI.2017.2767063.

Ankit Singh. CLDA: Contrastive Learning for Semi-Supervised Domain Adaptation. In *Advances in Neural Information Processing Systems*, volume 34, pages 5089–5101. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/288cd2567953f06e460a33951f55daaf-Abstract.html.

Jonas Sonntag, Gunnar Behrens, and Lars Schmidt-Thieme. Positive-Unlabeled Domain Adaptation. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, October 2022. doi: 10.1109/DSAA54385.2022.10032409. URL

https://ieeexplore.ieee.org/document/10032409/?arnumber=10032409.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, April 2020. ISSN 1465-4644. doi: 10.1093/biostatistics/kxz041. URL https://doi.org/10.1093/biostatistics/kxz041.

Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3:17, February 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0221-y. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7005290/.

Anita Thapar, Miriam Cooper, and Michael Rutter. Neurodevelopmental disorders. *The Lancet Psychiatry*, 4(4):339–346, April 2017. ISSN 2215-0366, 2215-0374. doi: 10.1016/S2215-0366(16)30376-5. URL https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(16)30376-5/fulltext. Publisher: Elsevier.

Mamatha Thota and Georgios Leontidis. Contrastive Domain Adaptation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2209–2218, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-899-4. doi: 10.1109/CVPRW53098.2021.00250. URL https://ieeexplore.ieee.org/document/9522966/.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous Deep Transfer Across Domains and Tasks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4068–4076, December 2015. doi: 10.1109/ICCV.2015.463. URL https://ieeexplore.ieee.org/document/7410820. ISSN: 2380-7504.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial Discriminative Domain Adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.316. URL http://ieeexplore.ieee.org/document/8099799/.

Mengying Yan, Michael J Pencina, L Ebony Boulware, and Benjamin A Goldstein. Observability and its impact on differential bias for clinical prediction models. *Journal of the American Medical Informatics Association*, 29(5):937–943, May 2022. ISSN 1527-974X. doi: 10.1093/jamia/ocac019. URL https://doi.org/10.1093/jamia/ocac019.

Mengying Yan, Hwanhee Hong, Jonathan Wilson, and Benjamin A Goldstein. Estimating the Observability of an Outcome from an Electronic Health Records Dataset Using External Data. *American Journal of Epidemiology*, page kwaf013, January 2025. ISSN 0002-9262. doi: 10.1093/aje/kwaf013. URL https://doi.org/10.1093/aje/kwaf013.

Yiqin Yu, Pin-Yu Chen, Yuan Zhou, and Jing Mei. Adversarial Sample Enhanced Domain Adaptation: A Case Study on Predictive Modeling with Electronic Health Records, January 2021. URL http://arxiv.org/abs/2101.04853. arXiv:2101.04853 [cs].

Tianran Zhang, Muhao Chen, and Alex A.T. Bui. AdaDiag: Adversarial Domain Adaptation of Diagnostic Prediction with Clinical Event Sequences. *Journal of biomedical informatics*, 134:104168, October 2022. ISSN 1532-0464. doi: 10.1016/j.jbi.2022.104168. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9580228/.

Yuan Zhang, Xi Yang, Julie Ivy, and Min Chi. Time-aware Adversarial Networks for Adapting Disease Progression Modeling. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–11, June 2019. doi: 10.1109/ICHI.2019.8904698. URL https://ieeexplore.ieee.org/document/8904698. ISSN: 2575-2634.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On Learning Invariant Representations for Domain Adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7523–7532. PMLR, May 2019. URL https://proceedings.mlr.press/v97/zhao19a.html. ISSN: 2640-3498.

Appendix A. Summary Tables

Table 7 presents the one-year mortality outcome rate and baseline characteristics for the historical (source) and contemporary (target) cohorts in the one-year mortality prediction after COVID-19 application.

Table 8 summarizes baseline demographic and clinical characteristics at the index encounter, along with long-term outcome rates, stratified by historical (source) and contemporary (target) cohorts in the neurodevelopmental conditions (NDC) application.

Appendix B. Additional Real Data Application Results

We also evaluated calibration-in-the-large for both one-year mortality prediction and long-term NDC prediction using different variants of the proposed method. Table 9 shows that adding conditional alignment can increase mean calibration in these two risk prediction tasks. Because conditional alignment shifts predictions toward positive outcomes, these results underscore the importance of balancing all alignment components.

Table 10 summarizes the AUPRC results for predicting one-year mortality and long-term neurodevelopmental conditions across different methods and observation windows. The proposed method consistently outperforms baseline approaches, achieving the highest AUPRC in most settings. In all cases, the AUPRC substantially exceeds the outcome prevalence. Notably, for long-term NDC prediction, the proposed method AUPRC closely approaches that of the oracle model.

Table 11 shows the F1 score results for the two applications. We selected the threshold as the positive prevalence plus 0.1 to account for potential calibration error. Given the class imbalance and low positive rates, the proposed method achieves high F1 scores and consistently outperforms all baseline approaches.

Appendix C. Computation Time

To provide more transparency regarding the computational cost of our methods, we report the average training time for each method across all datasets used in this study. The adversarial domain adaptation approach introduces additional training complexity due to the iterative updates of both the encoder and discriminator components. However, we observed that the total training time remained manageable with all three components. Table 12 summarizes the training time (in seconds) for each method-dataset combination. All experiments were conducted on a server with NVIDIA RTX A5000 GPU, and training was considered complete upon convergence or after a fixed number of epochs.

Table 7: One-year mortality rate and baseline characteristics stratified by historical and contemporary data

	Historical (N=103,741)	**Contemporary (N=20,000)**
1-year mortality outcome rate		
Scenario 1	8.3%	3.6% (90 days)
Scenario 2	8.3%	2.2% (30 days)
Scenario 3	8.3%	1.2% (7 days)
Age (years), mean (sd)	46.0 (23.8)	46.7 (23.1)
Sex		
Female	41,837 (40.3%)	8,439 (42.2%)
Systolic Blood Pressure (mmHg), mean (sd)	135.2 (23.4)	133.1 (21.8)
Diastolic Blood Pressure (mmHg), mean (sd)	78.2 (13.9)	80.1 (14.6)
Oxygen Saturation (%), mean (sd)	97.9 (2.8)	98.3 (2.4)
Temperature (°F), mean (sd)	98.2 (1.9)	98.2 (2.1)
Pulse (bpm), mean (sd)	90.2 (21.8)	90.4 (20.7)
Respiration Rate, mean (sd)	19.0 (5.0)	18.8 (4.3)
Acuity Level		
Level 1	940 (0.9%)	192 (1.0%)
Level 2	27,683 (26.7%)	5,430 (27.2%)
Level 3	52,607 (50.7%)	10,428 (52.1%)
Level 4	19,483 (18.8%)	3,463 (17.3%)
Level 5	3,028 (2.9%)	487 (2.4%)
Local tumor	1,356 (1.3%)	483 (2.4%)
Metastatic Tumor	3,183 (3.1%)	726 (3.6%)
Diabetes w/ Complication	9,338 (9.0%)	2,798 (14.0%)
Diabetes wo Complication	18,845 (18.2%)	4,799 (24.0%)
Renal Disease	21,101 (20.3%)	7,052 (35.3%)

Table 8: Long-term NDC outcome rates and baseline characteristics at index encounter, stratified by historical and contemporary data

	Historical (N=4097)	**Contemporary (N=5479)**
Long-term NDC outcome rate		
Scenario 1	20.2%	10.0% (2 years)
Scenario 2	20.2%	6.5% (1 year)
Scenario 3	20.2%	3.1% (6 months)
Sex, n (%)		
Female	1980 (48.3%)	2656 (48.5%)
Male	2117 (51.7%)	2823 (51.5%)
Age (months), mean (SD)	17.86 (2.33)	18.03 (2.33)
Race/Ethnicity, n (%)		
Hispanic	483 (11.8%)	914 (16.7%)
Non-Hispanic Asian	191 (4.7%)	217 (4.0%)
Non-Hispanic Black	1251 (30.5%)	1577 (28.8%)
Non-Hispanic White	1729 (42.2%)	2088 (38.1%)
Other	443 (10.8%)	683 (12.5%)
Preferred Language, n (%)		
Arabic	15 (0.4%)	15 (0.3%)
English	3661 (89.8%)	4890 (89.2%)
Other	57 (1.4%)	43 (0.8%)
Spanish	364 (8.9%)	531 (9.7%)
Primary Payer, n (%)		
Commercial	2026 (49.5%)	2438 (44.5%)
Medicaid	1837 (44.9%)	2735 (49.9%)
Other Government	33 (0.8%)	106 (1.9%)
Uninsured	201 (4.9%)	200 (3.6%)
Birth weight (g), mean (SD)	3249.9 (623.4)	3261.0 (608.8)
ICU Admission, n (%)	288 (7.0%)	298 (5.4%)

Table 9: Mean calibration-in-the-large comparison for one-year mortality prediction and long-term NDC prediction with different observation windows. $\mathcal{L}_{M_t}$: overall alignment; $\mathcal{L}_{\text{part}M_t}$: partial alignment; $\mathcal{L}_{\text{sup}M_t}$: conditional alignment.

Model	$\mathcal{L}_{M_t}$(ADDA)	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t} + \mathcal{L}_{\text{sup}M_t}$
1-year mortality					
Observe window					
90 days	0.028	0.029	0.113	0.742	0.125
30 days	0.028	0.027	0.103	0.667	0.118
7 days	0.028	0.029	0.081	0.576	0.085
Long-term NDC					
Observe window					
2 years	0.004	0.006	0.126	0.776	0.144
1 year	0.004	0.005	0.106	0.759	0.120
6 months	0.004	0.005	0.083	0.766	0.098

Table 10: AUPRC results for one-year mortality prediction and long-term NDC prediction with different observation windows. Event rates and outcome observability levels for each scenario are reported. Oracle model: trained on post-COVID target data with fully observed labels; Historical source model: trained on historical data without adaptation; Target model: trained exclusively on the partially observed target. ADDA: unsupervised baseline DA method. Bolded values indicate the best performance for each scenario.

Model	Oracle (full obs)	Historical (w/o adapt)	Target (part obs)	ADDA	Proposed
1-year mortality					
90 days	0.477	0.409	0.443	0.412	0.425
30 days	0.477	0.409	0.382	0.412	0.416
7 days	0.477	0.409	0.304	0.312	0.403
Long-term NDC					
2 years	0.642	0.588	0.543	0.613	0.642
1 years	0.642	0.588	0.541	0.613	0.646
6 months	0.642	0.588	0.499	0.613	0.639

Table 11: F1 score results for one-year mortality prediction and long-term NDC prediction with different observation windows. Event rates and outcome observability levels for each scenario are reported. Oracle model: trained on post-COVID target data with fully observed labels; Historical source model: trained on historical data without adaptation; Target model: trained exclusively on the partially observed target. ADDA: unsupervised baseline DA method. Bolded values indicate the best performance for each scenario.

Model	Oracle (full obs)	Historical (w/o adapt)	Target (part obs)	ADDA	Proposed
1-year mortality					
90 days	0.176	0.169	0.030	0.171	0.391
30 days	0.176	0.169	0.059	0.171	0.321
7 days	0.176	0.169	0.008	0.171	0.275
Long-term NDC					
2 years	0.404	0.351	0.093	0.385	0.560
1 years	0.404	0.351	0.060	0.385	0.583
6 months	0.404	0.351	0.013	0.385	0.549

Table 12: Average running time(s) for each components. $\mathcal{L}_{M_t}$: overall alignment; $\mathcal{L}_{\text{part}M_t}$: partial alignment; $\mathcal{L}_{\text{sup}M_t}$: conditional alignment.

	Digit classification	1-year mortality	Long-term NDC
Source sample size	11,552	20,000	4,097
Target sample size	1,902	20,000	5,479
Model			
$\mathcal{L}_{M_t}$	217.66	973.48	14.10
$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t}$	253.14	1017.53	24.25
$\mathcal{L}_{M_t} + \mathcal{L}_{\text{sup}M_t}$	217.72	988.14	16.33
$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t} + \mathcal{L}_{\text{sup}M_t}$	261.88	1022.97	25.63