Can Deep Clinical Models Handle Real-World Domain Shifts?

Jayaraman J. Thiagarajan\textsuperscript{1,\*}, Deepta Rajan\textsuperscript{2}, and Prasanna Sattigeri\textsuperscript{2}

\textsuperscript{1}Lawrence Livermore National Laboratory, \textsuperscript{2}IBM Research

Abstract

The hypothesis that computational models can be reliable enough to be adopted in prognosis and patient care is revolutionizing healthcare. Deep learning, in particular, has been a game changer in building predictive models, thereby leading to community-wide data curation efforts. However, due to the inherent variabilities in population characteristics and biological systems, these models are often biased to the training datasets. This can be limiting when models are deployed in new environments, particularly when there are systematic domain shifts not known a priori. In this paper, we formalize these challenges by emulating a large class of domain shifts that can occur in clinical settings, and argue that evaluating the behavior of predictive models in light of those shifts is an effective way of quantifying the reliability of clinical models. More specifically, we develop an approach for building challenging scenarios, based on analysis of disease landscapes, and utilize unsupervised domain adaptation to compensate for the domain shifts. Using the openly available MIMIC-III EHR dataset for phenotyping, we generate a large class of scenarios and evaluate the ability of deep clinical models in those cases. For the first time, our work sheds light into data regimes where deep clinical models can fail to generalize, due to significant changes in the disease landscapes between the source and target landscapes. This study emphasizes the need for sophisticated evaluation mechanisms driven by real-world domain shifts to build effective AI solutions for healthcare.

Introduction

The role of automation in healthcare and medicine is steered by both the ever-growing need to leverage knowledge from large-scale, heterogeneous information systems, and the hypothesis that computational models can actually be reliable enough to be adopted in prognosis and patient care. Deep learning, in particular, has been a game changer in this context, and recent successes in predictions with both aggregated electronic health records (EHR) as well as raw clinical measurements (e.g. ECG, EEG) have definitely provided strong evidence to support this hypothesis.

While designing computational models that can account for the complex biological phenomena and all possible interactions during disease evolution, e.g. cancer modeling (Szymanska et al. 2018), is not possible yet, building predictive models for patient health is a significant step in the direction of personalized care. This has resulted in community-wide curation of datasets, often comprised of a concise set of predictor variables carefully selected from routine tests and patient meta data. However, a not-so-trivial challenge in making statistical models useful in practice is to suitably prepare and transform patient data from an unseen target domain, in accordance with the custom data used for training. Furthermore, there is a fundamental trade-off in predictive modeling with healthcare data. While the complexity of the space of disease conditions naturally demands the expansion of the number of predictor variables, this often makes it extremely challenging to identify reliable correlations in the data, thus rendering the models highly specialized to the training datasets. In other words, despite the availability of large-sized datasets used for building models, we are still operating in the small data regime, wherein the curated data is not fully representative of what the model might encounter when deployed. Furthermore, existing deep learning solutions rarely provide meaningful uncertainties, thus making it difficult to distinguish between noisy correlations picked up by the model and a systematic shift in the data characteristics in a certain operating region.

In this paper, we formalize these challenges by emulating a large class of domain shifts that can occur in clinical settings, and argue that evaluating the behavior of predictive models in light of those shifts is an effective way of quantifying the effectiveness of models. Dealing with complex domain shifts is essential for adopting machine learning models in practice, and in the recent years, this problem has been well studied for even cases where the target domain does not have access to labeled data. Referred to as unsupervised domain adaptation, this class of techniques attempts to align the distributions of the source and target domains such that knowledge from labeled source domain data can be maximally utilized to design a model for the target domain. More specifically, adversarial training based approaches, e.g. domain adversarial neural networks (DANN) (Ganin et al. 2016) and virtual adversarial domain adaptation (VADA) (Shu et al. 2018), have been highly effective in...
domain shifts in computer vision.

In this paper, for the first time, we study the effectiveness of clinical models through unsupervised adaptation to real-world domain shifts. To this end, we use the openly available MIMIC-III EHR dataset (Goldberger et al. 2000) for phenotyping, to construct several scenarios (source-target pairs) that represent real-world domain shifts. For example, one possible scenario can constitute biases in the age distribution of patients in source and target domains. Subsequently, we build deep ResNet models, following their recent success in many clinical problems (Rajpurkar et al. 2017), along with unsupervised domain adaptation to build meaningful predictive models for the target domain. Using rigorous empirical studies, we obtain several key inferences. In many domain shifts, a pre-trained model from the source domain generalizes poorly to the target domain. In cases where the domain shift constituted a systematic shift in distribution of the predictor variables, for example train on patients with a combination of disease conditions and generalize to patients who present only a subset of those diseases, unsupervised domain adaptation produces highly effective predictive models. However, we find crucial cases where the domain shift is challenging for the source models to generalize, since the correlations picked up by the model do not apply to a large population of unobserved data, thus evidencing the small-data learning in clinical settings. In summary, this paper emphasizes the need for domain-shift based model evaluation to build reliable predictive models.

Deep Diagnostic Models with Clinical Time-Series

This section discusses challenges in building data-driven solutions for healthcare, current art in deep learning based clinical modeling, mainly using EHR data, as well as an overview of evaluation strategies used in practice.

Predictive Modeling: Clinical time-series modeling broadly focuses on problems such as anomaly detection, tracking progression of specific diseases or detecting groups of diseases by formulating them as binary classification tasks. Modeling multi-variate clinical time-series is often constrained by inherent data-related challenges such as long range temporal dependencies, missing measurements, and varying sampling rates. Deep learning provides a data-driven approach to exploiting relationships between a large number of input measurements while producing robust diagnostic predictions (Miotto et al. 2016). In particular, Recurrent Neural Networks (RNN) have become a de-facto solution for dealing with sequential data (Choi et al. 2016) in tasks such as intensive care unit (ICU) readmission (Lin et al. 2018), discovering phenotypes (Lipton et al. 2015), mapping clinical notes to ICD-9 codes (Huang, Osorio, and Sy 2018), detecting cardiac diseases (Rajan and Thiagarajan 2018), etc. The work by (Purushotham et al. 2017) extensively benchmarked the effectiveness of deep learning in analyzing EHR data. More recently, convolutional networks (Rajpurkar et al. 2017), and attention models (Song et al. 2018) have become highly effective alternatives to RNN-based solutions. Further, Differential Neural Computers (DNC) have been used for treatment recommendations, wherein very long term dependencies need to be modeled. In addition, prior-knowledge through medical ontologies can be utilized to further improve representation learning (Choi et al. 2017), (Che and Liu 2017). Despite the availability of a wide-variety of solutions, deep Residual Networks with 1-D convolutions (ResNet-1D), have been empirically deemed to be a popular solution for modeling sequence data (Bai, Kolter, and Koltun 2018). In this work, all domain adaptation studies are carried out using this architecture.

Model Evaluation: Selecting unbiased and meaningful metrics is crucial for validation of diagnostic models; however there is no consensus on the optimal metric for all applications. In addition, the problem formulation greatly impacts model evaluation and in-turn its clinical utility (Johnson, Pollard, and Mark 2017), (Sherman et al. 2017). Most commonly, results tend to shed light on correct predictions over the amount of misses. Examples include true positive rate (sensitivity), true negative rate (specificity) and positive/negative predictive values. These metrics provide point-estimates for disease detection without much focus on the false negatives and false positives. Other commonly reported metrics are AUROC, AUPRC and F1-scores, which do provide a somewhat global perspective (Rajkomar et al. 2018). However, these metrics do not inform about their behavior with systematic shifts in the population characteristics, which are more severe in the clinical domain compared to any other application. Consequently, rigorous evaluation of models under constrained scenarios is necessary to understand model generalizability, especially given that it is challenging to obtain representative data in healthcare.

Proposed Work: Gaining insights into disease progression and the relationships between different disease conditions is a treatment imperative. The key to ensuring computational models are truly effective in diagnosis is to incorporate the ability to understand dependencies between diseases, which we refer to as the disease landscape. For example, using domain experts to group diseases based on their chances of co-occurring in patients could be a potential approach towards organizing the space of diseases. However, such domain-knowledge dependent, rule-based systems designed by medical experts fail to provide a scalable and generalizable solution. Interestingly, the disease landscape can vary significantly across different parts of the population, thus rendering the pre-trained predictive models ineffective when deployed in a new environment. Hence, we propose to explicitly design challenging domain shifts, driven by the changes in disease landscapes across data subsets constructed using constraints from the real-world, in order to rigorously evaluate the broad applicability of modeling solutions. To the best of our knowledge, this is the first effort to study the model performances using the structure in the outcome space.

Dataset Description

All experiments for the proposed study are carried out using the MIMIC-III dataset (Johnson et al. 2016), which is the largest publicly available database of de-identified EHR that includes clinical time-series collected from patients in the
ICU. We used version 1.4 of the database which includes a variety of data types such as diagnostic codes, survival rates, length of stay, and more all acquired from 46,520 subjects over a period of about ten years. More recently, this dataset has been benchmarked for different predictive tasks resulting in a cohort of 33,798 unique patients with a total of 42,276 hospital admissions and ICU stays (Harutyunyan et al. 2017). For our study, we focus on the task of acute care phenotyping that involves retrospectively predicting the likely disease conditions for each patient given all their ICU stay information.

Preparation of such large amounts of diverse content involves organizing each patient’s data into episodes containing both time-series events as well as episode-level outcomes like diagnosis, mortality, age etc. Here, the time-series data includes a total of 17 measurements such as capillary refill rate, systolic, diastolic and mean blood pressure, glucose, heart rate, oxygen saturation, temperature, height, weight, pH, and Glasgow Coma Scale (GCS) parameters like eye opening, motor and verbal response.

Phenotyping is typically a multi-class problem as each patient is diagnosed with several disease conditions. However, in this work we design a number scenarios for representing real-world domain shifts by posing binary classification problems to detect the presence or absence of subsets of related conditions from the disease landscape. Each scenario is comprised of a pair of source and target domains that are chosen based on a scenario-specific constraint (e.g. gender distribution). The dataset contains 25 disease categories, 12 of which are critical such as respiratory/renal failure, 8 are chronic conditions such as diabetes, atherosclerosis, with the remaining 5 being ‘mixed’ conditions such as liver infections and complications during surgery. We construct disease landscapes on data subsets from scenario-specific constraints, and accordingly build the source and target domains. The list of domain-shifts considered and the corresponding dataset sizes are shown in Table 1

### Table 1: List of Source and Target domains designed to represent a wide-range of real-world domain shifts.

| Domain Shift                                                                 | Scenario                                      | Dataset Size   |
|------------------------------------------------------------------------------|-----------------------------------------------|----------------|
| Unobserved disease conditions appear in the target                           | Resp-to-CardiacRenal                          | 18,323         |
|                                                                               | Cerebro-to-CardiacRenal                       | 20,295         |
| Age bias in the source and target                                           | Older-to-Younger                              | 24,835         |
|                                                                               | Younger-to-Older                              | 17,067         |
| Gender bias in the source and target                                        | Male-to-Female                                | 23,425         |
|                                                                               | Female-to-Male                                | 18,477         |
| Racial bias in the source and target                                        | White-to-Minority                             | 29,817         |
|                                                                               | Dual-to-Single                                | 13,708         |
| Target presents combinations of conditions observed in source               | Single-to-Dual                                | 15,099         |
| Train with noisy labels in the source                                       | 10% Label Flips                              | 18,323         |
|                                                                               | 20% Label Flips                              | 18,323         |
| Target data collected at a different sampling rate                           | Sampling Rate Change                         | 18,323         |
| Target data contains subset of source measurements                          | Missing-Measurements                         | 13,708         |
| Presence of adversarial samples in the target                               | Adversaries                                  | 24,835         |

building machine learning solutions targeted to the healthcare domain presents a huge opportunity towards digitizing medicine. Accordingly, the performance of novel deep learning based clinical models is becoming increasingly effective in disease diagnosis and treatment planning. Though these models are heavily data-driven, rarely do we study how these solutions actually adapt to domain-shifts in terms of population-specific biases and the factors influencing dependencies between disease conditions. As a result, these highly parameterized models can be easily biased towards the dominant population. In this section, we provide details about a multitude of such scenarios that are motivated by the implicit biases and imbalances present in clinical data. We study changes in the disease landscape across the shifts in population characteristics, based on attributes of clinical time-series datasets, and create scenarios to study the adaptability of clinical models. Note that, in our study, we assume that we have access to labeled data only in the source domain, and the data available in the target domain is unlabeled. A primary contribution of this work is the careful design of datasets that encompass a variety of domain shifts often faced in clinical workflows.

### Approach:
As mentioned in the previous section, our scenarios are based on posing phenotyping as a binary classification problem, i.e., detection of presence or absence of a selected set of disease conditions. Given that it is likely that each patient can be associated with multiple diseases, we first construct the disease landscape within a constrained population, as determined by the scenario. To this end, we utilize information sieve (Steeg and Galstyan 2016), a recent information-theoretic technique for identifying latent factors in data that maximally describe the total correlation between variables. Denoting a set of multivariate random variables
by \( X = \{X_i\}_{i=1}^N \), dependencies between the variables can be quantified using the multivariate mutual information, also known as total correlation (TC), which is defined as

\[
TC(X) = \sum_{i=1}^{N} H(X_i) - H(X),
\]

where \( H(X_i) \) denotes the marginal entropy. This quantity is non-negative and zero only when all the \( X_i \)'s are independent. Further, denoting the latent source of all dependence in \( X \) by \( Y \), we obtain \( TC(X|Y) = 0 \). Therefore, we solve the problem of searching for a feature \( Y \) that minimizes \( TC(X|Y) \). Equivalently, we can define the reduction in \( TC \) after conditioning on \( Y \) as \( TC(X; Y) = TC(X) - TC(X|Y) \). In our context, \( X \) corresponds to the set of disease outcomes, represented as binary variables. The optimization begins with \( X \), constructs \( Y_0 \) to maximize \( TC(X; Y_0) \). Subsequently, it computes the remainder information not explained by \( Y_0 \), and learns another feature, \( Y_1 \), that infers additional dependence structure. This procedure is repeated for \( k \) layers until the entirety of multivariate mutual information in the data is explained. This hierarchical decomposition represents the disease landscape, revealing the complex dependencies between outcome variables.

For example, the force-based layout in Figure 1 provides a holistic view of the landscape for the entire MIMIC-III dataset used in our experiments. Here, each circle corresponds to a latent factor and their sizes indicate the amount of total correlation explained by that factor, and the thickness of edges reveal contributions of the diseases to that factor. Such an organization of the disease conditions enables us to identify groups of related diseases, and by selecting non-overlapping groups, we can create scenarios where there is systematic shift between source and target. For example, one can construct a scenario where the goal is to detect diseases in cluster 2, while patients with disease conditions from cluster 0 are observed only in the target domain. Patients who presented disease conditions that are uncorrelated to cluster 0 are considered as negatives. This analysis can be carried out for every scenario, by splitting the population based on specific attributes (see Table 1) that are all potential sources of domain-shift between source and target. In the rest of this section, we describe the scenarios in detail.

(i) **Unobserved Diseases**: In order to accurately detect if a data sample is diseased or not, in an ideal case the training data must consist of enough examples of every possible disease. Nonetheless, models are trained to confidently detect a subset of diseases present in an available dataset. In a situation where such a model has to be deployed in a new environment, say at a different hospital, it is likely that the population contains disease conditions or their variants previously unseen by the model. To study the generalization of models to such unobserved conditions, we create two scenarios, each with a different set of diseases in the source domain, while the target domain was fixed to be source conditions plus a set of new diseases not observed in the source. The Resp-to-CardiacRenal scenario is made up of source data containing diseases from cluster 1 in Figure 1. Whereas the corresponding target data includes source diseases in addition to conditions from cluster 0, which make up the set of new unseen diseases. Correspondingly, the Cerebro-to-CardiacRenal scenario has source data with diseases pertinent to cluster 2, and target data to include this source disease list and the same unseen diseases from cluster 0 as mentioned in the Resp-to-CardiacRenal scenario.

(ii) **Age Bias**: Due to the 'small data' situation in clinical ML, curated datasets can be imbalanced based on population characteristics. In this scenario, we create a dataset biased by patient age and formulate two scenarios. One referred Older-to-Younger where the source data comprises of all patients who are 60 years and above, and the target data includes remaining patients who are below 60 years of age. While, the second scenario named Younger-to-Older represents the reverse condition with source containing patients younger than 60. Once the population is divided into two groups based on the chosen age criteria, we construct the corresponding disease landscapes and pick a cluster of dis-
eases in the source landscape for detection. However, due to the domain shift, the landscape can change significantly in the target. Hence, in the target domain new correlations that emerge with respect to the source diseases are included to the list of diseases to be detected.

(iii) Gender Bias: Similar to the previous case, in this scenario we explore the domain shift caused by biases in gender. We create two scenarios, Male-to-Female, i.e. source data consists only of patients who identify as male and target has only patients who are female. While, the second scenario Female-to-Male is comprised of female-only source data and male-only target data.

(iv) Racial Bias: An imbalance in the distribution of patients’ ethnicities is another likely cause of a domain shift. To model such a shift, we create a scenario White-to-Minority that contains prevalent racial groups such as white American, Russian, and European, that we deem to be majority, in the source data, while constructing the target data comprising of patients belonging to minority racial groups like Hispanic, South America, African, Asian, Portuguese, and others marked as unknown in the MIMIC-III dataset.

(v) Disease Subsets: In general, there exists differences in diagnostic procedures for detection of various diseases. As a common practice, lab tests for some two diseases could be conducted together based on the likelihood of them co-occurring in an ICU setting, giving rise to a dataset containing patients with both diseases. In practice, these models would be required to detect the presence or absence of one of the diseases explicitly. Therefore, we design a scenario called Dual-to-Single, where we create a dataset to test the ability of models trained on patients who are associated with two disease conditions simultaneously to a target domain containing patients who have only either of the two. Here, we divide patients into four groups, namely: those that have only cardiac diseases, only renal diseases, neither cardiac or renal diseases, and, both cardiac and renal diseases. We do not construct the landscapes for this case, instead, patients that have neither cardiac or renal related diseases are treated as class 0, while the rest are considered as class 1.

(vi) Disease Supersets: Similar to the previous domain shift characterized by disease subsets, an alternate situation could occur with models having to adapt to predicting patients that have two diseases, while having been trained on patients who only had one disease. Such a scenario named Single-to-Dual is explored using a source dataset comprising of patients diagnosed as having either cardiac only or respiratory only disease, and a target dataset with patients that have both cardiac and respiratory-related diseases.

(vii) Noisy Labels: Variabilities in diagnoses between different experts is common in clinical settings (Valizadegan, Nguyen, and Hauskrecht 2013). Consequently, it is important to understand the impact of those variabilities on behavior of the model, when adopted to a new environment. We extend the Resp-to-CardiacRenal scenario by adding uncertainties to the diagnostic labels in the source domain and study its impact on the target domain. In particular, we simulate two such scenarios by randomly flipping 10% and 20% of the labels in the source to create 10% Label Flips and 20% Label Flips scenarios.

(viii) Sampling Rate Change: Another commonly observed issue with time-series data across sources is the variability in sampling rates, which affects amount of information captured and corresponding analysis results in a chosen window of time. The Sampling Rate Change scenario encapsulates such a domain shift, where we create a source-target pair by extending the Resp-to-CardiacRenal scenario to comprise of measurements collected at two different sampling rates. The source contains time-series sampled at 96 hours while the target is sampled at 48 hours.

(ix) Missing Measurements: Clinical time-series is routinely plagued by missing measurements during deployment, often due to resource or time limitations, and can be typically handled by either imputing values or removing those measurements entirely from the dataset. We demonstrate this domain shift by extending the Dual-to-Single scenario, wherein we assume that the set of measurements, pH, Temperature, Height, Weight, and all Verbal Response GCS parameters, were not available in the target dataset. We impute the missing measurements with zero and this essentially reduces the dimensionality of valid measurements from 76 to 41 in the target set.

(x) Adversarial Examples: Machine learning models including deep neural networks have been shown to be prone to adversarial perturbations, i.e. systematic changes in the measurements that can significantly alter the prediction. In this scenario, we test the robustness of VADA under adversarial perturbations. Specifically, we obtain adversarial samples from the source-only trained model, and simulate a realistic case where a deployed model is updated periodically using unsupervised domain adaptation. An adversary poisons the incoming unlabeled target data used for the next round of adaptation using the available source data and the current deployed model. In particular, we use the Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014) to demonstrate the effect of poisoning the target data on the adaptation performance in the Older-to-Younger scenario described before. FGSM perturbs the input example using the sign of the loss function gradient as

\[ x^{adv} = x + \epsilon \text{ sign} (\nabla_x L(x, y)) . \] (2)

The adversarial examples generated by FGSM lie at the surface of the \( \ell_{\infty} \) ball of radius \( \epsilon \) around the input \( x \).

Methods

In this section, we describe the predictive model used in our study and the techniques employed for performing domain adaptation. As described earlier, we assumed that we have access to only unlabeled data in the target, and hence we resort to unsupervised domain adaptation.

Architecture: The ResNet-1D model used in our study is comprised of 7 residual blocks which transform the raw time-series input into feature representations. Each residual block is made up of two 1-D convolutional and batch normalization layers with a kernel size of 3 and about 64 or 256 filters, a dropout layer and a leaky-ReLU activation layer. Further, a 1-D max pooling layer is used after the third and seventh residual block to aggregate the features. In addition,
the parameters chosen for training include: an Adam optimizer with a learning rate of 0.001, and a batch size of 128 with an equal balance of positive and negative samples.

**Domain Adaptation:** The goal of unsupervised domain adaptation is to learn a good predictive model for the target domain \((x \sim P_t(x))\) using labeled source examples \((\{(x, y) : x \sim P_s(x)\})\) and unlabeled target examples. While subspace alignment strategies have been well studied for domain adaptation, more recently Domain Adversarial Neural Networks (DANN) (Ganin et al. 2016) have become popular. DANN utilizes adversarial training in order to achieve domain alignment, while simultaneously constructing a predictive model in the source domain. In other words, we attempt to learn representations that are domain invariant but can preserve discriminative properties useful for the predictive modeling. Specifically, DANN uses an adversary or a discriminator \(D\) and matches the distribution of features for the source and target domains,

\[
L_D(g\# P_s, g\# P_t) := \sup_D \mathbb{E}_{x \sim P_s} \ln D(g(x)) + \mathbb{E}_{x \sim P_t} \ln (1 - D(g(x))),
\]

where \(g\) is a measurable feature generator mapping and \(g\# P\) denotes the push-forward of a distribution \(P\). The overall objective of domain adaptation can be formulated as

\[
\min_{g \in \mathcal{G}, h \in \mathcal{H}} L_y(f; P_s) + \lambda_D L_D(g\# P_s, g\# P_t),
\]

where \(h\) is the prediction function, \(L_y(f; P_s) := \mathbb{E}_{x, y \sim P_s}[y \ln f(x)]\) is the cross-entropy loss for the labeled source examples (assuming \(f\) outputs the probabilities of classes and \(y\) is the label vector), \(L_D(\cdot, \cdot)\) is the loss term measuring the distance between the two distributions. This term results in minimizing the variational form of the Jensen-Shannon (JS) divergence (Goodfellow et al. 2014) between source and target feature distributions. While this formulation relies on the assumption that alignment of the domains will enable the use of \(h\) for both domains, there is evidence it can fail in practice (Shu et al. 2018). Hence, more recently, Virtual Adversarial Domain Adaptation (VADA), which extends DANN by employing additional regularization in the form of the conditional entropy \(L_{ce}(\cdot, \cdot)\), was proposed (Shu et al. 2018). In particular, the regularization term can be defined as

\[
L_{ce}(f; P_t) := -\mathbb{E}_{x \sim P_t} [f(x)^T \ln f(x)],
\]

Minimizing this cost on the unlabeled target samples pushes the predictor boundaries away from the high density regions. In addition, VADA incorporates virtual adversarial training (VAT) to avoid unlabeled samples which can manifest in predictor model with high capacity. This is achieved using a cost function to ensure that predictions within a neighborhood in the target domain is smooth.

\[
L_{vat}(f; P_t) := \mathbb{E}_{x \sim P_t} \left[\max_{\|r\| \leq \epsilon} D_h(f(x) + r) \right]
\]

Hence, the overall objective function of VADA can be obtained as:

\[
\min_{g \in \mathcal{G}, h \in \mathcal{H}} \min_{f = h\# g} L_y(f; P_s) + \lambda_D L_D(g\# P_s, g\# P_t) + \lambda_s L_{vat}(f; P_s)
+ \lambda_t(L_{ce}(f; P_t) + L_{vat}(f; P_t)),
\]

**Results and Discussion**

In this section, we present empirical results from our study – performances of the ResNet-1D model trained on the source data when subjected to perform predictions on a variety of target scenarios listed in Table 1. In order to obtain a comprehensive evaluation of the models, we considered three summary metrics, namely weighted average AUROC, AUPRC and the Matthews Correlation Coefficient (MCC). While AUROC and AUPRC are commonly used in clinical model evaluation, MCC is considered to be a balanced measure for binary classification problems, even when the

![Figure 2: Summary of results from our empirical study on generalizability of deep clinical models with complex domain-shifts.](image-url)
Table 2: Comparison of generalization performance obtained using models trained with only the labeled source data, and the models from unsupervised domain adaptation (VADA) with the target data as well. We report three metrics, AUROC, AUPRC and MCC, for each of the scenarios.

| Scenario               | Source Only | VADA          |
|------------------------|-------------|---------------|
|                        | AUROC | AUPRC | MCC | AUROC | AUPRC | MCC |
| Resp-to-CardiacRenal    | 0.74  | 0.85  | 0.37 | 0.73  | 0.84  | 0.37 |
| Cerebro-to-CardiacRenal | 0.69  | 0.85  | 0.22 | 0.67  | 0.83  | 0.28 |
| Sampling Rate Change    | 0.67  | 0.79  | 0.32 | 0.75  | 0.85  | 0.35 |
| 10% Label Flips         | 0.71  | 0.81  | 0.29 | 0.76  | 0.86  | 0.40 |
| 20% Label Flips         | 0.69  | 0.80  | 0.34 | 0.62  | 0.76  | 0.30 |
| Older-to-Younger        | 0.65  | 0.63  | 0.17 | 0.64  | 0.61  | 0.21 |
| Adversaries             | 0.65  | 0.63  | 0.20 | 0.65  | 0.63  | 0.23 |
| Younger-to-Older        | 0.62  | 0.90  | 0.09 | 0.61  | 0.90  | 0.13 |
| Male-to-Female          | 0.67  | 0.69  | 0.27 | 0.69  | 0.71  | 0.29 |
| Female-to-Male          | 0.67  | 0.85  | 0.25 | 0.71  | 0.86  | 0.23 |
| White-to-Minority       | 0.64  | 0.65  | 0.19 | 0.66  | 0.67  | 0.22 |
| Dual-to-Single          | 0.65  | 0.82  | 0.19 | 0.67  | 0.83  | 0.22 |
| Missing-Measurements    | 0.62  | 0.80  | 0.15 | 0.65  | 0.82  | 0.20 |
| Single-to-Dual          | 0.77  | 0.76  | 0.37 | 0.84  | 0.82  | 0.49 |

classes are of different sizes. The Matthews Correlation Coefficient is computed as follows:

\[
\text{MCC} = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}
\]

where \(tp\), \(tn\), \(fp\), \(fn\) denote the total number of true positives, true negatives, false positives and false negatives from the binary classification task. The appropriateness of a metric depends entirely on the requirements of an application; for example in some cases false negatives can be less detrimental than false positives, while in other cases reducing the number of false positives can be highly critical.

Table 2 details the performance of both approaches for all the scenarios. The first and crucial observation is that, in most cases, unsupervised domain adaptation produces significant improvements to the generalization performance, and this corroborates the results reported in the literature for computer vision tasks. However, the improved generalization can still be insufficient for deploying the model in practice, thus emphasizing regimes of high uncertainty for the chosen model. Figure 2 illustrates a summary of our study, indicating which domain shifts could be handled by deep clinical models and which ones could not. More interestingly, this inference can change drastically depending on the metric of choice, which is a well-known issue in the healthcare domain. Note, we do not tune the hyper-parameters for training the VADA models for each scenario separately; instead we set the \(\lambda_p\), \(\lambda_s\) and \(\lambda_t\) parameters to 0.01, which worked reasonably well across all scenarios.

**AUPRC metric:** With the standard AUPRC metric (Figure 2(b)), we observe that the racial bias (White-to-Minority) was challenging to handle, while the gender bias (Male-to-Female), and age bias (Older-to-Younger) were partially difficult to generalize for the deep ResNet-1D model. Interestingly, the model generalizes exceedingly well with disease landscape changes in Younger-to-Older and Female-to-Male scenarios. Also, in practice, it is known that quality of labels are central to the success of predictive models. However, surprisingly, we observe that limited noise in the labels (10% Label Flips) results in improved generalization (compared to Resp-to-CardiacRenal). However, the performance degrades with the increase in the amount of noise (20% Label Flips). Another striking observation is the robustness of deep models to sampling rate changes, missing measurements, and inclusion of unobserved disease conditions in the target data. Finally, we expected the behavior of VADA to be affected by the presence of adversarial examples in the target dataset; but our results showed that presence of adversaries slightly improved the performance.

**MCC metric:** Broadly, we observe that the AUROC and AUPRC demonstrate a similar trend in our case, since the positive and negative classes are well balanced. However, MCC tends to produce different insights, since both false positives and false negatives must be low simultaneously for the score to be high. As it can be seen from Figure 2(a), the model was the most effective in Single-to-Dual, where it could detect combinations of conditions. In contrast to the AUROC metric, with MCC, we observe that the age bias can affect the model’s generalization drastically, and the performance goes down for the Dual-to-Single scenario. However, there is agreement between the metrics in model’s ability to handle sampling rate change and label noise. Finally, the deep model is surprisingly very effective (in terms of both \(fp\) and \(fn\)) with inclusion of unobserved diseases in the target domain, which is conventionally believed to be an extremely challenging task.

In summary, the proposed approach of studying model adaptability with complex domain-shifts produces interest-
ing insights into the robustness and weaknesses of deep clinical models. We believe that this general protocol can be broadly applied for evaluating machine learning models in healthcare, and can easily integrate domain knowledge into the evaluation process.

References

[Bai, Kolter, and Koltun 2018] Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271.

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

[Choi et al. 2016] Choi, E.; Bahadori, M. T.; Schuetz, A.; Stewart, W. F.; and Sun, J. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference, 301–318.

[Choi et al. 2017] Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. Gram: graph-based attention model for healthcare representation learning. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 787–795. ACM.

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

[Choi et al. 2016] Choi, E.; Bahadori, M. T.; Schuetz, A.; Stewart, W. F.; and Sun, J. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference, 301–318.

[Choi et al. 2017] Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. Gram: graph-based attention model for healthcare representation learning. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 787–795. ACM.

[Goldberger et al. 2000] Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101(23):e215–e220.

[Goodfellow et al. 2014] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Advances in neural information processing systems, 2672–2680.

[Goodfellow, Shlens, and Szegedy 2014] Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[Harutyunyan et al. 2017] Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; and Galstyan, A. 2017. Multitask learning and benchmarking with clinical time series data. arXiv preprint arXiv:1703.07771.

[Huang, Osorio, and Sy 2018] Huang, J.; Osorio, C.; and Sy, L. W. 2018. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. arXiv preprint arXiv:1802.02311.

[Johnson et al. 2016] Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. Scientific data 3.

[Johnson, Pollard, and Mark 2017] Johnson, A. E.; Pollard, T. J.; and Mark, R. G. 2017. Reproducibility in critical care: a mortality prediction case study. In Machine Learning for Healthcare Conference, 361–376.

[Lin et al. 2018] Lin, Y.-W.; Zhou, Y.; Faghri, F.; Shaw, M. J.; and Campbell, R. H. 2018. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. bioRxiv 385518.

[Lipton et al. 2015] Lipton, Z. C.; Kale, D. C.; Elkan, C.; and Wetzell, R. 2015. Learning to diagnose with lstm recurrent neural networks. arXiv preprint arXiv:1511.03677.

[Miotto et al. 2016] Miotto, R.; Li, L.; Kidd, B. A.; and Dudley, J. T. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports 6:26094.

[Purushotham et al. 2017] Purushotham, S.; Meng, C.; Che, Z.; and Liu, Y. 2017. Benchmark of deep learning models on large healthcare mimic datasets. arXiv preprint arXiv:1710.08531.

[Rajan and Thiagarajan 2018] Rajan, D., and Thiagarajan, J. J. 2018. A generative modeling approach to limited channel ecg classification. arXiv preprint arXiv:1802.06458.

[Rajkomar et al. 2018] Rajkomar, A.; Oren, E.; Chen, K.; Dai, A. M.; Hajaj, N.; Hardt, M.; Liu, P. J.; Liu, X.; Marcus, J.; Sun, M.; et al. 2018. Scalable and accurate deep learning with electronic health records. npj Digital Medicine 1(1):18.

[Rajpurkar et al. 2017] Rajpurkar, P.; Hannun, A. Y.; Haghpanahi, M.; Bourn, C.; and Ng, A. Y. 2017. Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836.

[Sherman et al. 2017] Sherman, E.; Gurm, H.; Balis, U.; Owens, S.; and Wiens, J. 2017. Leveraging clinical timeseries data for prediction: A cautionary tale. In AMIA Annual Symposium Proceedings, volume 2017, 1571. American Medical Informatics Association.

[Shu et al. 2018] Shu, R.; Bui, H. H.; Narui, H.; and Ermon, S. 2018. A dirt-t approach to unsupervised domain adaptation. arXiv preprint arXiv:1802.08735.

[Song et al. 2018] Song, H.; Rajan, D.; Thiagarajan, J. J.; and Spanias, A. 2018. Attend and Diagnose: Clinical Time Series Analysis using Attention Models. Proceedings of AAAI 2018.

[Steeg and Galstyan 2016] Steeg, G. V., and Galstyan, A. 2016. The information sieve. In Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48, 164–172. JMLR. org.

[Szymanska et al. 2018] Szymanska, Z.; Cytowski, M.; Mitchell, E.; Macnamara, C. K.; and Chaplain, M. A. 2018. Computational modelling of cancer development and growth: modelling at multiple scales and multiscale modelling. Bulletin of Mathematical Biology.

[Valizadegan, Nguyen, and Hauskrecht 2013] Valizadegan, H.; Nguyen, Q.; and Hauskrecht, M. 2013. Learning classification models from multiple experts. Journal of biomedical informatics 46(6):1125–1135.