Improving Sepsis Prediction Model Generalization With Optimal Transport

Jie Wang JWANG3163@GATECH.EDU
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332

Ronald Moore RONALD.MOORE@EMORY.EDU
Department of Computer Science, Emory University, Atlanta, GA 30332

Yao Xie YAO.XIE@ISYE.GATECH.EDU
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332

Rishikesan Kamaleswaran RKAMALESWARAN@EMORY.EDU
Department of Biomedical Informatics, Emory University, Atlanta, GA 30332

Abstract

Sepsis is a deadly condition affecting many patients in the hospital. There have been many efforts to build models that predict the onset of sepsis, but these models tend to perform terribly when validated on external data from different hospitals due to distributional shifts in the data and insufficient samples from sepsis patients. To circumvent the curse from noisy and unbalanced samples, we develop a novel two-step approach for sepsis prediction: given feature-label points from the source domain and feature points from the target domain, to obtain a sepsis predictor that has satisfactory performance at the target domain. The proposed algorithm first learns how to transform sample points from the source domain to the target domain, and then applies the distributionally robust optimization (DRO) technique with the Sinkhorn distance and asymmetric cost function to reliably obtain a classifier with satisfactory out-of-sample performance. Connections between our proposed formulation and widely used classification models, i.e., DRO formulation with the Wasserstein distance and regularized logistic regression formulation, are also uncovered. Numerical experiments with synthetic and real datasets demonstrate the competitive performance of the proposed method.

Keywords: Domain Adaptation, Ethical AI, Optimal Transport, Sepsis Prediction

1. Introduction

Sepsis is a deadly condition affecting many patients in the hospital. The sepsis-3 definition defines sepsis as a dysfunctional host response to infection causing major organ failure and increasing the risk of death or major disability. A 2017 report published by the World Health Organization (WHO) (Organization et al., 2020) revealed that sepsis-related deaths accounted for roughly 20% of deaths worldwide. There is extensive literature on developing sepsis prediction models. However, many of these models performed poorly when validated on external data from different hospitals, primarily due to distributional shifts (Moor et al., 2021). To address this problem, domain adaptation techniques have been utilized to develop more robust frameworks. However, some of these domain adaptation methods still per-

form poorly on external data distributions (Guo et al., 2022).

In this work, we formulate the problem of sepsis prediction as the domain adaptation task, where features and labels are available only at the source domain, i.e., they are collected from a fixed hospital, but at the other hospital (target domain) only features are available. Even worse, samples for healthy patients are highly frequent while samples for sepsis patients are only rarely encountered in real-world datasets. In summary, there is a strong need for developing non-parametric framework for domain adaptation with noisy, high-dimensional, and unbalanced data.

A notable contribution to domain adaptation is the optimal transport-based framework (Flamary et al., 2017), which learns a transportation plan matching feature distributions between both domains and then obtains a predictor based on the estimated feature-label distribution on the target domain. Unfortunately, the estimation of data distribution is not reliable because collected samples are noisy and unbalanced so the estimated transportation plan is far from the ground truth optimal transport planning, especially for distributions corresponding to sepsis patients. Consequently, due to the distribution shift from the estimation step, the obtained predictor may not have satisfactory out-of-sample performance.

To address this issue, we propose a novel optimal transport-based domain adaptation two-step procedure leveraging the distributionally robust optimization (DRO) technique for robust and ethical sepsis prediction. First, we estimate the feature-label distribution at the target domain using the previous optimal transport-based algorithm (Flamary et al., 2017). Next, we propose a DRO model using Sinkhorn distance that jointly learns a feature-label distribution and a robust classifier so that the worst-case misclassification risk is optimized. The highlight of the proposed DRO formulation is that we use an asymmetric cost function that robustifies the minority group, i.e., samples corresponding to sepsis patients so that the mis-classification rate for correctly detecting sepsis disease (precision) is significantly reduced. Our contributions are summarized as follows:

- A two-step optimal transport-based strategy for domain adaptation is proposed. We leverage the idea of sample average approximation to solve the proposed formulation.
- Connections between our proposed formulation and classical Wasserstein DRO formulation and regularized logistic regression formulation are uncovered.
- Our proposed framework is examined using both synthetic and real datasets to demonstrate its competitive performance.

Notations Denote by $\mathbb{E}$ the expectation operator. For any positive integer N, define $[N] = \{1, 2, \ldots, N\}$. Fix a positive integer M, define $\delta_x = (\delta_{x,1}, \ldots, \delta_{x,M})$ as the M-vector of Kronecker deltas. For a measurable set $\mathcal{Z}$, denote by $\mathcal{M}(\mathcal{Z})$ the set of measures (not necessarily probability measures) on $\mathcal{Z}$, and $\mathcal{P}(\mathcal{Z})$ the set of probability measures on $\mathcal{Z}$. Denote by $\|x\|_A^2 := \left(x^{\mathrm{T}} A x\right)^{1/2}$ the weighted ℓ_2 norm with respect to the matrix A.

2. Related Work

2.1. Sepsis Prediction

Recently, there is a surge of interest in sepsis prediction with machine learning algorithms. Notable methodologies include ensemble learning (Barton et al., 2019; Goh et al., 2021), Bayesian learning (Nachimuthu and Haug, 2012; Brown et al., 2016), and

deep learning (Futoma et al., 2017b,a; Lin et al., 2018; Scherpf et al., 2019). Unfortunately, those models may perform poorly when validated on external data from different hospitals due to the shift in the data distribution between the training population and the testing population (Moor et al., 2021). A bad prediction model may result in risky or unethical medical treatment policies and severe consequences. As such, it is important to learn a reliable sepsis prediction model under the scenario of distribution shift.

2.2. Domain Adaptation

Various approaches in literature are proposed to tackle the domain adaptation problem, the key of which is to reduce the mismatch between the source and target domain distributions. Classical regularized methods (Azizzadenesheli et al., 2019) have been implemented in domain adaptation frameworks. Deep learning-based algorithms (Venkataramani et al., 2018; Zhang et al., 2019a; Alves et al., 2018; Zhang et al., 2019b; Khoshnevisan and Chi, 2020, 2021; Zhu et al., 2022) can further improve the model performance due to the flexibility in data fitting and surprising predictions for unseen data of neural network functions. Domain adaption based on modern statistical distance functions such as maximum mean discrepancy (MMD) and Wasserstein distance has recently achieved much attention (Deng et al., 2021; Balagopalan et al., 2020), due to their flexibility and reliability for quantifying the discrepancy between distributions from different domains with data. As pointed out in Guo et al. (2022), the main shortcoming of the foregoing methodologies is that the prediction is insufficiently robust so that it may not generalize well for unseen data from the target domain, especially in applications from healthcare.

2.3. Optimal Transport and DRO

Optimal transport (OT) is a flexible way to quantify discrepancy between two probability distributions. It thereby serves as a suitable performance measure for data-driven domain adaptation tasks. Besides, there have been many variants of optimal transport to improve the computation and prediction performance beyond the regular optimal transport among which the most famous one is the so-called (entropy)-regularized optimal transport (Altschuler et al., 2017; Alaya et al., 2019; Feydy et al., 2019; Mensch and Peyré, 2020; Daniels et al., 2021). It is defined by regularizing the original mass transportation problem with a relative entropy penalty on the transport mapping. Since the convergence analysis of an efficient algorithm for solving such a problem is attributed to the mathematician Sinkhorn (Sinkhorn, 1964), the associated distance function is also named the Sinkhorn distance (Cuturi, 2013). It has been used in several important applications due to its computational efficiency and satisfactory statistical performance guarantees, including generative modeling (Genevay et al., 2018; Petzka et al., 2018; Luise et al., 2018; Patrini et al., 2020) and dimensionality reduction (Lin et al., 2020; Wang et al., 2021a, 2022; Huang et al., 2021). In this work, we apply Sinkhorn distance in the healthcare setting, specifically for reliable sepsis prediction.

Distributionally robust optimization (DRO) provides a principled approach to solve the decision-making problem under uncertainty, by seeking a minimax robust optimal decision that minimizes the expected loss under the most adverse distribution within a given set of relevant distributions, called ambiguity set. The popular OT-based DRO model constructs such an ambiguity set as a probability ball using the Wasserstein distance, which incorporates

the geometry of sample space, and thereby is suitable for comparing distributions with non-overlapping supports and hedging against data perturbations (Gao and Kleywegt, 2016). On the one hand, Wasserstein DRO has a finite-dimensional convex formulation under stringent conditions of the loss function (Shafieezadeh Abadeh et al., 2015; Mohajerin Esfahani and Kuhn, 2017). On the other hand, it has nice statistical performance guarantees both asymptotically (Blanchet et al., 2019, 2021b,a) and non-asymptotically (Gao, 2020; Chen and Paschalidis, 2018; Shafieezadeh-Abadeh et al., 2019). In recent literature, it has been applied in a variety of applications in operations research (Blanchet et al., 2018; Wang et al., 2021d,c; Kuhn et al., 2019; Wang and Xie, 2022).

3. Problem Setup and Formulation

Let $\{x_i^s, y_i^s\}_{i=1}^{N_s}$ be the training sample set generated from the *source domain*, where x_i^s stands for the i-th feature vector in $\mathbb{R}^d$ and y_i^s stands for the i-th label in $\{0, 1\}$. Also, let $\{x_i^t\}_{i=1}^{N_t}$ be the training sample set generated from the *target domain*, where x_i^t stands for the i-th feature vector. Our objective is to develop a classifier for domain adaptation such that, based on training sets $\{x_i^s, y_i^s\}_{i=1}^{N_s}$ and $\{x_i^t\}_{i=1}^{N_t}$, it gives prediction on new coming feature samples from the target domain. Traditional classification approaches are not applicable because i) the labels from the target domain are not available, and ii) the source domain and target domain may have non-overlapping supports. To address this issue, we propose an intuitive two-step strategy to probe the distributional region of the feature-label pair in the target domain.

3.1. Step 1: Interpolation

First, we formulate an optimal-transport based estimator of the data distribution in the target domain, following the step in existing literature (Flamary et al., 2017). Denote by the empirical distributions of feature vectors from source and target domains as

$$\mu_s = \frac{1}{N_s}\sum_{i=1}^{N_s}\delta_{x_i^s}, \quad \mu_t = \frac{1}{N_t}\sum_{i=1}^{N_t}\delta_{x_i^t}.$$

An optimal transport mapping for moving from the source to the target domain can be obtained by solving the following linear optimization problem:

$$\hat{\gamma} = \arg\min_{\gamma\in\Gamma} \sum_{i,j}\gamma_{i,j}c(x_i^s, x_j^t), \qquad (1)$$

where the constraint

$$\Gamma \triangleq \left\{\gamma \in \mathbb{R}_+^{N_s\times N_t} : \gamma\mathbf{1} = \mu_s, \gamma^{\mathrm{T}}\mathbf{1} = \mu_t\right\},$$

and the entry $c_{i,j} \triangleq c(x_i^s, x_j^t)$ quantifies the discrepancy between the i-th sample from the source domain and the j-th sample from the target domain. Specially, one may add entropic regularization to problem (1) to accelerate computation using Sinkhorn's algorithm (Cuturi, 2013), or label-based regularization to improve the classification performance. See (Flamary et al., 2017, Section 4) for detailed discussion.

After obtaining this transport mapping, the i-th sample x_i^s is moved to samples from the target domain $\{x_i^t\}_{j=1}^{N_t}$ according to probability $\{\gamma_{i,j}\}_{j=1}^{N_t}$. For $i \in [N_s]$, we compute a hard transformation of the source sample x_i^s using the following barycentric mapping:

$$\widehat{x_i^s} = \arg\min_{x\in\mathbb{R}^d} \sum_{j=1}^{N_t}\hat{\gamma}_{i,j}c(x, x_j^t). \qquad (2)$$

As a consequence, we formulate an empirical distribution from feature-label pairs $\{z_i^s\}_{i=1}^{N_s}$ with $z_i^s = (\widehat{x_i^s}, y_i^s)$, denoted as $\widehat{\mathbb{P}}$. Actually, $\widehat{\mathbb{P}}$ serves as the distributional estimate of the feature-label pair in the target domain.

After this estimator is obtained, a natural approach used in literature is to train a classifier $f_\theta(\cdot)$ to minimize the following risk function, in which we specify the nominal distribution $\mathbb{P}$ as $\widehat{\mathbb{P}}$, called the *sample average approximation approach (SAA)*:

$$\mathcal{R}(\mathbb{P}, \theta) \triangleq \mathbb{E}_{z\sim\mathbb{P}}\big[f_\theta(z)\big],$$

where the loss function

$$f_\theta(z) = \log\left(1 + \exp\left(-y \cdot \theta^{\mathrm{T}} x\right)\right),$$

with $z \triangleq (x, y)$ being a given feature-label pair. Since the distributional estimate $\widehat{\mathbb{P}}$ can be quite different from the underlying true distribution from the target domain, directly training a classifier based on $\widehat{\mathbb{P}}$ could lead to serious out-of-sample disappointment. In other words, the obtained classifier may not perform well for new coming testing samples from the target domain, which is similar to the overfitting phenomenon studied in statistics (Smith and Winkler, 2006).

3.2. Step 2: Robustification via DRO

The out-of-sample disappointment phenomenon in Step 1 motivates us to consider the robustification step. In contrast to the SAA model, we consider the following distributionally robust formulation to learn a classifier $f_\theta(\cdot)$:

$$\min_{\theta} \left\{\max_{\mathbb{P}\in\mathcal{P}} \mathcal{R}(\mathbb{P}, \theta)\right\}, \tag{3}$$

where the goal is to pick an optimal classifier so that the worst-case risk is minimized. The ambiguity set $\mathcal{P}$ contains a class of candidate distributions on the predictor-response pair, which is constructed using the nominal distribution $\widehat{\mathbb{P}}$. The construction step of this ambiguity set is the following:

$$\mathcal{P} = \big\{\mathbb{P} :\ \mathcal{W}_\eta(\mathbb{P}, \widehat{\mathbb{P}}) \le \rho\big\}.$$

Here we take the function $\mathcal{W}_\eta(\cdot, \cdot)$ to be the Sinkhorn distance. See its formal definition as the following.

Definition 1 (Sinkhorn Distance) *Let $\mathcal{Z}$ be a measurable set. Consider distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Z})$, and let $\mu, \nu \in \mathcal{M}(\mathcal{Z})$ be two reference measures such that $\mathbb{P} \ll \mu$, $\mathbb{Q} \ll \nu$. For regularization parameter $\epsilon \ge 0$, the* Sinkhorn distance *between two distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as*

$$\mathcal{W}_\eta(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \big\{\mathbb{E}_{(z,z')\sim\gamma}[d(z, z')] + \eta H(\gamma \mid \mu \otimes \nu)\big\},$$

where $\Gamma(\mathbb{P}, \mathbb{Q})$ denotes the set of joint distributions whose first and second marginal distributions are $\mathbb{P}$ and $\mathbb{Q}$ respectively, $d(x, y)$ denotes the cost function, and $H(\gamma \mid \mu \otimes \nu)$ denotes the relative entropy of γ with respect to the product measure $\mu \otimes \nu$:

$$H(\gamma \mid \mu\otimes\nu) = \mathbb{E}_{(z,z')\sim\gamma} \log\left(\frac{\mathrm{d}\gamma(z, z')}{\mathrm{d}\mu(z)\,\mathrm{d}\nu(z')}\right).$$

This robustification step brings the following benefits for domain adaptation: i) The estimator of feature-label distribution for the target domain seems noisy and unreliable, while the DRO formulation further provides a data-driven estimation of this distribution, which usually leads to the improvement of the out-of-sample performance (Lin et al., 2022). ii) The ambiguity set is constructed in a data-driven manner using Sinkhorn distance, which naturally incorporates the geometry of the sample space and alleviates the over-conservativeness of the traditional Wasserstein uncertainty set thanks to entropic regularization. Furthermore, when specifying the transport cost in Sinkhorn distance as special asymmetric functions, the predictor can make better prediction for samples from minorities (Hui et al., 2021), which alleviates the curse of unbalanced data samples. iii) Finally, from the optimization point of view, the proposed formulation can be efficiently solved using the first-order method (Wang et al., 2021b), which is scalable especially for large-sample and high-dimensional scenarios.

4. Discussions for Robustification

It is worth mentioning that the current formulation (3) in the robustification step is not tractable, because the inner maximization problem requires taking into account uncountably many candidate distributions within the ambiguity set $\mathcal{P}$, and candidate distributions are supported in infinite-dimensional space. In this section, we provide a strong dual reformulation to equivalently convert this problem into a finite-dimensional optimization problem and present approximation algorithm to find the robust classifer in (3). Also, we will provide the connection between the DRO model (3) with other formulations studied in machine learning literature.

Convex Dual Reformulation of (3): For a pair of data points $z := (x, y)$ and $z' := (x', y')$, we specify the asymmetric cost function $d(z, z') = \|x - x'\|^2_{A(y)} + \kappa 1\{y \neq y'\}$ and the reference measure ν to be the Lebesgue measure, where the matrix $A(y) = \big(L1\{y = 1\} + 1\{y = 0\}\big) \cdot I_d$. Leveraging the strong duality result in (Wang et al., 2021b, Theorem 1), the minimax problem (3) can be equivalently formulated as a single minimization problem:

$$\min_{\theta,\lambda>0} \left\{ F(\theta, \lambda) \triangleq \lambda\overline{\rho} + \frac{\lambda\eta}{N_s} \sum_{i=1}^{N_s} \log\left(\mathbb{E}_{\mathbb{Q}_i} e^{f_\theta(z_i^s)/(\lambda\eta)}\right) \right\}, \quad (4)$$

where we define the constant

$$\overline{\rho} = \rho + \frac{\eta}{N_s} \sum_{i=1}^{N_s} \log\left(\int e^{-d(z_i^s, z)/\eta}\, \mathrm{d}z\right)$$

and for $i \in [N_s]$, define the kernel probability distribution

$$\frac{\mathrm{d}\mathbb{Q}_i(z)}{\mathrm{d}z} = \frac{e^{-d(z_i^s, z)/\eta}}{\int e^{-d(z_i^s, z')/\eta}\, \mathrm{d}z'}.$$

It is worth mentioning that such a reformulation holds for a broader class of loss functions, cost functions and reference measures. In this task, we only consider the restrictive choice for the simplicity of discussion.

Optimization Algorithm: Since the objective function in (4) involves a nonlinear transformation of expectation with respect to a continuous distribution, it is challenging to evaluate or optimize the objective function in general. We apply the idea of *sample average approximation* to solve this DRO formulation. For each $i \in [N_s]$, we generate independent and identically distributed (i.i.d.) random samples $\{z_{i,j}^s\}_{j=1}^m$ from the kernel probability distribution $\mathbb{Q}_i$. Next, we consider the following formulation:

$$\min_{\theta,\lambda>0} \left\{ \widehat{F}(\theta, \lambda) \triangleq \lambda\overline{\rho} + \frac{\lambda\eta}{N_s} \sum_{i=1}^{N_s} \log\left(\frac{1}{m} \sum_{j=1}^{m} e^{f_\theta(z_{i,j}^s)/(\lambda\eta)}\right) \right\}. \quad (5)$$

In comparison with the objective in (4), the new objective is obtained by replacing the inner expectation with the sample mean with respect to generated random samples. It is worth mentioning that as the sample size $m \to \infty$, under mild assumptions, the optimal classifier in (5) will converge to that in (4). Also, the new formulation can be solved efficiently using first-order gradient method. Alternatively, one can check the formulation (5) is equivalent to a conic programming problem based on (Wang et al., 2021b, Corollary 1). Hence, the new formulation is conveniently solvable using interior point method based on off-the-shelf solvers such as CVX (Grant and Boyd, 2014).

Connections with Other Models: As the regularization parameter $\eta \to 0$, by (Wang et al., 2021b, Remark 1), one can check that the formulation (3) reduces to

$$\min_{\theta} \left\{ \max_{\mathbb{P}} \mathcal{R}(\mathbb{P}, \theta) : \mathcal{W}(\mathbb{P}, \widehat{\mathbb{P}}) \leq \rho \right\}, \quad (6)$$

where $\mathcal{W}(\cdot,\cdot)$ denotes the standard optimal transport distance. Hence, our proposed model can be viewed as a smoothed version of the Wasserstein DRO formulation. However, solving the Wasserstein DRO formulation can be computationally challenging in general, while Algorithm 4 presents an efficient optimization algorithm for the Sinkhorn DRO formulation with a provable convergence rate that is sample size independent. Also, when specifying the cost function $d(z,z') = \|x - x'\| + \infty 1\{y \neq y'\}$, the Wasserstein DRO formulation in (6) can be exactly reformulated as the following norm-regularized problem (Gao et al., 2017):

$$\min_{\theta} \ \mathcal{R}(\widehat{\mathbb{P}}, \theta) + \rho\|\theta\|_*, \tag{7}$$

where $\|\cdot\|_*$ is the dual norm of the norm function $\|\cdot\|$. Formulation (3) is therefore a softened version of the standard regularized logistic regression model.

5. Experiment on Synthetic Dataset

In this section, we provide a toy example to describe how our two-step procedure works. We generate a 2-dimensional dataset, in which each class has 30 sample points. Here we take

$$x_i^s \mid y_i^s = 0 \sim \mathcal{N}\left(\begin{pmatrix}-4\\9\end{pmatrix}, 0.3I_2\right),$$
$$x_i^s \mid y_i^s = 1 \sim \mathcal{N}\left(\begin{pmatrix}-6\\5\end{pmatrix}, 0.3I_2\right),$$
$$x_i^t \mid y_i^t = 0 \sim \mathcal{N}\left(\begin{pmatrix}4\\-2\end{pmatrix}, 0.3I_2\right),$$
$$x_i^t \mid y_i^t = 1 \sim \mathcal{N}\left(\begin{pmatrix}7\\-5\end{pmatrix}, 0.3I_2\right),$$

The visualization is provided in Fig. 1-(a), from which we can see the target domain is a rotation of the source domain.

Next, we plot the optimal transport mapping obtained in the interpolate procedure in Fig. 1-(b). The gray line corresponds to the transportation mapping. Due to the entropy regularization, one can see each sample from the source domain is transported to multiple points in the target domain.

We formulate the barycentric mapping according to the formulation (2). The visualization is provided in Fig. 1-(c). Since each point from the source domain is *deterministically* transported, we now obtain the estimators of feature-label pair from the target domain. Compared with the ground truth plot in Fig. 1, one can see that the estimators may have ten points with wrong labels.

Also, we plot the naive classifer and robust classifier obtained from the formulation (3) in Fig. 1-(c). As we can see, the in-sample mis-classification risk for robust classifer is 40%. However, as demonstrated in Fig. 1-(d), the out-of-sample mis-classification risk for robust classifer is 10%. This suggests that our robustification step greatly improves the performance of domain adaptation.

6. Experiment with Sepsis Data

6.1. Experiment Settings

When evaluating our proposed algorithms, we use real data collected from encounters at Emory University Hospital and Grady Hospital in the year 2016. We extract features that contain vital signs and laboratory values in this experiment, while variables related to demographic information are excluded in an effort to mitigate bias. See more details on those variables in Table 1. There exist non-negligible discrepancies in data distributions between the two hospitals due to biases in medical planning and devices. During the data preprocessing step, missing values are imputed by forward-filling vital signs up to 12 hours and lab values up to 36 hours. Any remaining missing values are imputed using the global median value for that variable.

		Missing	Overall	Emory	Grady	P-Value
n			90374	66712	23662	
Sepsis, n (%)	No	0	81036 (89.7)	60265 (90.3)	20771 (87.8)	<0.001
	Yes		9338 (10.3)	6447 (9.7)	2891 (12.2)	
Daily Weight (kg), median [Q1,Q3]		202	82.0 [68.0,89.0]	82.0 [68.6,88.5]	76.3 [67.6,89.4]	<0.001
Height (cm), median [Q1,Q3]		12990	168.4 [162.6,175.3]	168.4 [165.1,175.3]	170.2 [161.5,175.3]	0.012
Pulse, median [Q1,Q3]		1596	81.0 [72.0,90.0]	80.0 [72.0,89.5]	82.0 [73.0,92.0]	<0.001
Temperature (Celsius), median [Q1,Q3]		2026	36.6 [36.5,36.8]	36.6 [36.5,36.8]	36.7 [36.5,36.9]	<0.001
Non-Invasive Systolic Blood Pressure, median [Q1,Q3]		1715	126.0 [115.0,139.0]	125.0 [114.0,138.0]	129.0 [117.0,141.0]	<0.001
Invasive Systolic Blood Pressure, median [Q1,Q3]		83528	128.8 [113.0,145.9]	128.0 [112.0,145.0]	133.0 [117.0,149.0]	<0.001
Invasive Mean Arterial Pressure, median [Q1,Q3]		83424	83.5 [75.0,94.0]	82.2 [74.0,93.0]	89.0 [80.0,98.0]	<0.001
Non-Invasive Mean Arterial Pressure, median [Q1,Q3]		8270	88.0 [80.0,97.0]	87.0 [79.0,96.0]	90.0 [83.0,99.0]	<0.001
Best Mean Arterial Pressure, median [Q1,Q3]		8221	88.0 [80.3,97.0]	87.0 [80.0,96.0]	90.0 [82.5,99.0]	<0.001
Non-Invasive Diastolic Blood Pressure, median [Q1,Q3]		1719	70.5 [64.0,78.0]	69.0 [63.0,76.0]	74.0 [67.0,82.0]	<0.001
Invasive Diastolic Blood Pressure, median [Q1,Q3]		83540	63.0 [55.0,72.0]	62.0 [54.0,70.0]	70.0 [62.0,78.0]	<0.001
Unassisted Respiratory Rate, median [Q1,Q3]		1666	18.0 [18.0,18.0]	18.0 [18.0,18.0]	18.0 [18.0,18.0]	0.460
End Tidal CO2, median [Q1,Q3]		90046	33.0 [27.0,38.0]	nan [nan,nan]	33.0 [27.0,38.0]	nan
Base Excess, median [Q1,Q3]		85186	-1.9 [-4.2,0.6]	-2.7 [-4.8,-0.7]	-1.0 [-3.7,1.6]	<0.001
Bicarbonate (HCO3), median [Q1,Q3]		17297	25.0 [23.0,27.0]	25.0 [23.0,27.0]	25.0 [23.0,27.0]	0.003
FiO2, median [Q1,Q3]		77309	0.3 [0.2,0.4]	0.3 [0.2,0.4]	0.3 [0.3,0.4]	<0.001
pH, median [Q1,Q3]		80082	7.4 [7.3,7.4]	7.4 [7.3,7.4]	7.4 [7.3,7.4]	0.002
Partial Pressure of Carbon Dioxide (PaCO2), median [Q1,Q3]		80054	39.0 [34.8,44.0]	38.4 [34.0,43.0]	40.0 [35.0,45.0]	<0.001
Oxygen Saturation (SaO2), median [Q1,Q3]		82171	97.1 [95.1,98.5]	96.9 [95.0,98.1]	100.0 [97.0,100.0]	<0.001
Aspartate Aminotransferase (AST), median [Q1,Q3]		31680	23.0 [17.0,35.0]	23.0 [18.0,34.0]	23.0 [16.0,37.0]	<0.001
Blood Urea Nitrogen (BUN), median [Q1,Q3]		17661	14.0 [10.0,21.0]	15.0 [10.0,22.0]	13.0 [9.0,19.0]	<0.001
Alkaline Phosphatase, median [Q1,Q3]		31688	77.0 [60.0,104.0]	78.0 [60.0,105.0]	75.0 [58.0,102.0]	<0.001
Calcium, median [Q1,Q3]		17375	8.7 [8.3,9.1]	8.7 [8.3,9.0]	8.8 [8.4,9.2]	<0.001
Chloride, median [Q1,Q3]		17609	104.0 [101.0,106.0]	104.0 [101.0,107.0]	103.0 [101.0,106.0]	<0.001
Creatinine, median [Q1,Q3]		17556	0.9 [0.7,1.2]	0.9 [0.7,1.2]	0.9 [0.7,1.2]	<0.001
Direct Bilirubin, median [Q1,Q3]		68264	0.1 [0.1,0.2]	0.2 [0.1,0.4]	0.1 [0.1,0.2]	<0.001
Glucose, median [Q1,Q3]		16103	111.0 [97.0,136.0]	113.0 [98.0,139.0]	106.0 [93.0,129.0]	<0.001
Lactic Acid, median [Q1,Q3]		72063	1.5 [1.1,2.1]	1.4 [1.0,1.8]	2.0 [1.5,2.6]	<0.001
Magnesium, median [Q1,Q3]		44429	1.9 [1.8,2.1]	2.0 [1.8,2.1]	1.9 [1.8,2.1]	<0.001
Phosphorus, median [Q1,Q3]		59747	3.3 [2.8,3.9]	3.3 [2.8,3.9]	3.4 [2.9,3.9]	0.198
Potassium, median [Q1,Q3]		20086	3.9 [3.7,4.2]	3.9 [3.7,4.2]	4.0 [3.8,4.3]	<0.001
Total Bilirubin, median [Q1,Q3]		34180	0.6 [0.4,0.9]	0.6 [0.4,0.9]	0.5 [0.4,0.8]	<0.001
Troponin, median [Q1,Q3]		64861	0.0 [0.0,0.1]	0.0 [0.0,0.1]	0.0 [0.0,0.0]	<0.001
Hematocrit, median [Q1,Q3]		20104	33.8 [29.3,38.3]	33.2 [28.9,37.5]	35.3 [30.3,39.9]	<0.001
Hemoglobin, median [Q1,Q3]		20025	11.0 [9.4,12.6]	10.8 [9.3,12.3]	11.6 [9.8,13.1]	<0.001
Partial Prothrombin Time (PTT), median [Q1,Q3]		69279	29.8 [27.1,33.4]	30.5 [27.6,34.4]	28.9 [26.3,32.0]	<0.001
White Blood Cell Count, median [Q1,Q3]		20275	8.4 [6.2,11.2]	8.4 [6.2,11.2]	8.5 [6.3,11.3]	0.001
Fibrinogen, median [Q1,Q3]		85998	308.0 [221.0,426.0]	275.0 [203.0,391.2]	369.0 [280.8,463.0]	<0.001
Platelets, median [Q1,Q3]		20348	217.5 [170.0,276.0]	219.0 [171.0,279.0]	214.0 [169.0,270.9]	<0.001

Table 1: Vital Signs and Laboratory Value Statistics for Emory and Grady Patient Encounters

For each experiment trial, we randomly split the data into labeled training samples from the source domain, unlabeled training samples from the target domain, and labeled testing samples from the target domain. The performance of a given sepsis predictor is quantified as the classification results on unlabeled training samples and labeled testing samples. Since sepsis prediction is a high-stake mission-critical task, we use four metrics as classification statistics: *sensitivity, specificity, precision, and accuracy*. The higher those four metrics are, the better performance the obtained predictor has. We perform 200 independent experiment trials and discuss the details of experiment results in the next subsection.

We compare our proposed framework with the following baseline approaches in literature:

- (Basic-OT): the basic optimal transport algorithm in (Flamary et al., 2017, Section 3);
- (Reg-OT): the label-regularized optimal transport algorithm in (Flamary et al., 2017, Section 4);
- (FDA): feature-level domain adaptation in Kouw et al. (2016);

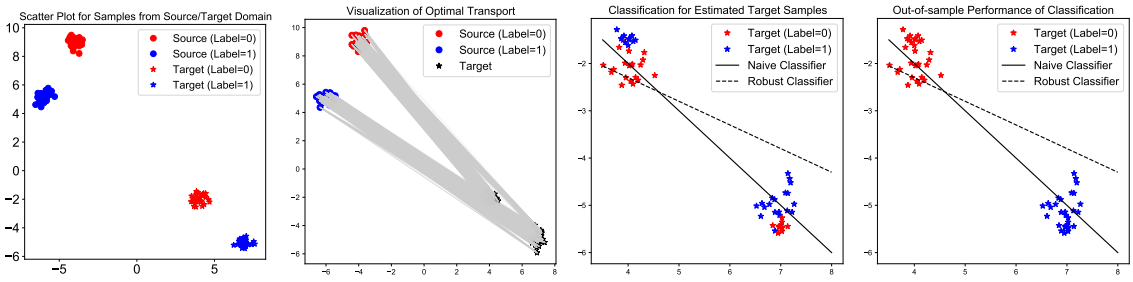

Figure 1: (a) Scatter plot of sample points; (b) Visualization of Optimal Transport; (c) Plot for classification on estimated target samples; (d) Plot for classification on ground truth of target samples.

- (SAS): subspace aligned classifier in Fernando et al. (2014);
- (TCS): transfer component classifier (TCS) in Pan et al. (2010).

All hyper-parameters are tuned by cross-validation based on training samples from the source domain.

6.2. Experiment Results

The detailed classification results together with the basic information of the real dataset are summarized in Table 2. In particular, we report the performance of domain adaptation for transforming from Emory hospital data to that of Grady hospital, and transforming from Grady to Emory hospital. From Table 2(a) and Table 2(b), we can see our proposed SDRO algorithm outperforms the other baseline approaches for almost all metrics in all scenarios. Especially, the baseline approaches have very small precision for two scenarios, while our proposed algorithm greatly improves this metric, indicating that it performs well for samples coming from minorities, i.e., patients with sepsis disease. Also, for the task of domain adaptation for Grady $\rightarrow$ Emory, all approaches have very small classification accuracy. One possible explanation is that the corresponding optimal transport mapping may lead to highly noisy labels at the target domain. However, our proposed algorithm still improves the classification accuracy since the DRO technique can deal with noisy data with satisfactory out-of-sample performance.

7. Conclusion

In this work, we proposed a two-step optimal transport-based strategy for the task of domain adaptation with applications to sepsis prediction. The proposed algorithm first learns how to transform sample points from the source domain to the target domain. To deal with the challenge of noisy and unbalanced samples, the algorithm next applies the distributionally robust optimization technique with the Sinkhorn distance and asymmetric cost function to obtain a reliable classifier with satisfactory out-of-sample performance. The connection between our proposed formulation and widely used classification models, i.e., DRO formulation with the Wasserstein distance and regularized logistic regression formulation, was also uncovered. Numerical experiments on synthetic and real datasets demonstrated the competitive performance of this algorithm.

Table 2: Results of domain adaptation with several optimal transport-based approaches. Each experiment is repeated for 20 independent trials, and 95% confidence intervals of classification results are reported for different approaches.

(a)
Domain adaptation for Emory → Grady

		Precision	Recall	F_1 Score	Accuracy
Basic-OT	Train (Unlabeled)	$.155 \pm .069$	$.009 \pm .006$	$.018 \pm .010$	$.737 \pm .028$
	Test (Labeled)	$.135 \pm .052$	$.008 \pm .003$	$.015 \pm .006$	$.734 \pm .028$
Reg-OT	Train (Unlabeled)	$.194 \pm .057$	$.008 \pm .005$	$.015 \pm .015$	$.731 \pm .024$
	Test (Labeled)	$.104 \pm .034$	$.010 \pm .004$	$.018 \pm .009$	$\mathbf{.735 \pm .031}$
FDA	Train (Unlabeled)	$.128 \pm .042$	$.010 \pm .006$	$.018 \pm .012$	$.715 \pm .019$
	Test (Labeled)	$.097 \pm .041$	$.007 \pm .003$	$.013 \pm .003$	$.727 \pm .025$
SAS	Train (Unlabeled)	$.127 \pm .043$	$.009 \pm .004$	$.017 \pm .011$	$.729 \pm .034$
	Test (Labeled)	$.128 \pm .041$	$.014 \pm .006$	$.025 \pm .008$	$.733 \pm .051$
TCS	Train (Unlabeled)	$.150 \pm .034$	$.010 \pm .003$	$.018 \pm .010$	$.734 \pm .027$
	Test (Labeled)	$.112 \pm .029$	$.015 \pm .003$	$.027 \pm .009$	$.722 \pm .035$
SDRO	Train (Unlabeled)	$\mathbf{.211 \pm .075}$	$\mathbf{.011 \pm .004}$	$\mathbf{.021 \pm .032}$	$\mathbf{.739 \pm .067}$
	Test (Labeled)	$\mathbf{.269 \pm .087}$	$\mathbf{.017 \pm .007}$	$\mathbf{.032 \pm .003}$	$.733 \pm .029$
Number of Predictors		39			
Labeled Size		16712			
Unlabeled Size		13662			
Testing Size		10000			

(b)
Domain adaptation for Grady → Emory

		Precision	Recall	F_1 Score	Accuracy
Basic-OT	Train (Unlabeled)	$.307 \pm .079$	$.051 \pm .014$	$.088 \pm .023$	$.531 \pm .020$
	Test (Labeled)	$.311 \pm .058$	$.050 \pm .008$	$.087 \pm .014$	$.527 \pm .016$
Reg-OT	Train (Unlabeled)	$.324 \pm .059$	$.061 \pm .004$	$.106 \pm .013$	$.526 \pm .017$
	Test (Labeled)	$.343 \pm .071$	$.063 \pm .006$	$.106 \pm .011$	$.523 \pm .008$
FDA	Train (Unlabeled)	$.365 \pm .064$	$.053 \pm .009$	$.093 \pm .019$	$.546 \pm .025$
	Test (Labeled)	$.258 \pm .049$	$.049 \pm .005$	$.082 \pm .007$	$.530 \pm .018$
SAS	Train (Unlabeled)	$.300 \pm .049$	$.060 \pm .007$	$.100 \pm .014$	$.532 \pm .031$
	Test (Labeled)	$.372 \pm .054$	$.064 \pm .008$	$.109 \pm .007$	$.523 \pm .010$
TCS	Train (Unlabeled)	$.338 \pm .047$	$.057 \pm .006$	$.098 \pm .015$	$.532 \pm .034$
	Test (Labeled)	$.382 \pm .043$	$.063 \pm .006$	$.093 \pm .003$	$.527 \pm .008$
SDRO	Train (Unlabeled)	$\mathbf{.383 \pm .009}$	$\mathbf{.066 \pm .004}$	$\mathbf{.112 \pm .007}$	$\mathbf{.562 \pm .007}$
	Test (Labeled)	$\mathbf{.388 \pm .041}$	$\mathbf{.065 \pm .006}$	$\mathbf{.111 \pm .011}$	$\mathbf{.554 \pm .003}$
Number of Predictors		39			
Labeled Size		13662			
Unlabeled Size		16712			
Testing Size		50000			

8. Acknowledgement

Y. Xie and J. Wang were supported by an NSF CAREER CCF-1650913, NSF DMS-2134037, CMMI-2015787, DMS-1938106, and DMS-1830210, and a grant from Emory Hospital. R. Kamaleswaran and Y. Xie were supported by an NIH Supplemental Award under the grant number 3R01GM139967-02S1. R. Kamaleswaran was supported by the National Institutes of Health under Award Numbers R01GM139967 and UL1TR002378. R. Kamaleswaran was also supported by Surgical Critical Care Initiative, funded through the Department of Defense's Health Program—Joint Program Committee 6/Combat Casualty Care (USUHS HT9404-13-1-0032 and HU0001-15-2-0001).

References

Mokhtar Z. Alaya, Maxime Berar, Gilles Gasso, and Alain Rakotomamonjy. Screening Sinkhorn Algorithm for Regularized Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 32, pages 12191–12201, Dec 2019.

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30, pages 1961–1971, 2017.

Tiago Alves, Alberto Laender, Adriano Veloso, and Nivio Ziviani. Dynamic Prediction of ICU Mortality Risk Using Domain Adaptation. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1328–1336, Dec 2018.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019.

Aparna Balagopalan, Jekaterina Novikova, Matthew B. A. Mcdermott, Bret Nestor, Tristan Naumann, and Marzyeh Ghassemi. Cross-Language Aphasia Detection using Optimal Transport Domain Adaptation. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116, pages 202–219, Apr 2020.

Christopher Barton, Uli Chettipally, Yifan Zhou, Zirui Jiang, Anna Lynn-Palevsky, Sidney Le, Jacob Calvert, and Ritankar Das. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Computers in Biology and Medicine*, 109:79–84, Jun 2019.

Jose Blanchet, Lin Chen, and Xun Yu Zhou. Distributionally robust mean-variance portfolio selection with wasserstein distances. *arXiv preprint arXiv:1802.04885*, February 2018.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, October 2019.

Jose Blanchet, Karthyek Murthy, and Viet Anh Nguyen. Statistical analysis of wasserstein distributionally robust estimators. *arXiv preprint arXiv:2108.02120*, August 2021a.

Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in wasserstein distributionally robust estimation. *arXiv preprint arXiv:1906.01614*, March 2021b.

Samuel M. Brown, Jason Jones, Kathryn Gibb Kuttler, Roger K. Keddington, Todd L. Allen, and Peter Haug. Prospective evaluation of an automated

method to identify patients with severe sepsis or septic shock in the emergency department. *BMC Emergency Medicine*, 16:31, Aug 2016.

Ruidi Chen and Ioannis Ch. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13):1–48, August 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, volume 26, pages 2292–2300, Dec 2013.

Max Daniels, Tyler Maunu, and Paul Hand. Score-based Generative Neural Networks for Large-Scale Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 34, pages 12955–12965, 2021.

Fucheng Deng, Shikui Tu, and Lei Xu. Multi-source unsupervised domain adaptation for ECG classification. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 854–859, Dec 2021.

Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Subspace alignment for domain adaptation. *arXiv preprint arXiv:1409.5241*, October 2014.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, Apr 2019.

R Flamary, N Courty, D Tuia, and A Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 39:1853–1865, Sep 2017.

Joseph Futoma, Sanjay Hariharan, and Katherine Heller. Learning to detect sepsis with a multitask Gaussian process RNN classifier. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1174–1182, Aug 2017a.

Joseph Futoma, Sanjay Hariharan, Katherine Heller, Mark Sendak, Nathan Brajer, Meredith Clement, Armando Bedoya, and Cara O'Brien. An improved multi-output gaussian process rnn with real-time validation for early sepsis detection. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68, pages 243–254, Aug 2017b.

Rui Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*, October 2020.

Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, April 2016.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *arXiv:1712.06050 [cs.LG]*, 2017.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 2018.

Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel

Yu Heng Tan. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications*, 12:711, Jan 2021.

Michael Grant and Stephen Boyd. Cvx: Matlab software for disciplined convex programming, version 2.1, 2014.

Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Alistair E. W. Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific Reports*, 12: 2726, Feb 2022.

Minhui Huang, Shiqian Ma, and Lifeng Lai. A riemannian block coordinate descent method for computing the projection robust wasserstein distance. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4446–4455, July 2021.

Y Hui, J Xie, J Blanchet, and P Glynn. Empirical optimal transport projections with non-symmetric costs. *preprint*, Mar 2021.

Farzaneh Khoshnevisan and Min Chi. An Adversarial Domain Separation Framework for Septic Shock Early Prediction Across EHR Systems. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 64–73, Dec 2020.

Farzaneh Khoshnevisan and Min Chi. Unifying Domain Adaptation and Domain Generalization for Robust Prediction Across Minority Racial Groups. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 521–537, Sep 2021.

Wouter M. Kouw, Laurens J.P. van der Maaten, Jesse H. Krijthe, and Marco Loog. Feature-level domain adaptation. *Journal of Machine Learning Research*, 17 (171):1–32, June 2016.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, August 2019.

Chen Lin, Yuan Zhang, Julie Ivy, Muge Capan, Ryan Arnold, Jeanne M. Huddleston, and Min Chi. Early Diagnosis and Prediction of Sepsis Shock by Combining Static and Dynamic Information Using Convolutional-LSTM. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 219–228, Jun 2018.

Fengming Lin, Xiaolei Fang, and Zheming Gao. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control & Optimization*, 12:159, 2022.

Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael Jordan. Projection robust wasserstein distance and riemannian optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 9383–9397, December 2020.

Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with wasserstein distance. In *Advances in Neural Information Processing Systems*, December 2018.

Arthur Mensch and Gabriel Peyré. Online Sinkhorn: Optimal Transport distances from sample streams. In *Advances in Neural Information Processing Systems*, volume 33, pages 1657–1667, 2020.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, July 2017.

Michael Moor, Bastian Rieck, Max Horn, Catherine R. Jutzeler, and Karsten Borgwardt. Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. *Frontiers in Medicine*, 8: 607952, May 2021.

Senthil K. Nachimuthu and Peter J. Haug. Early Detection of Sepsis in the Emergency Department using Dynamic Bayesian Networks. *AMIA Annual Symposium Proceedings*, 2012:653–662, Nov 2012.

World Health Organization et al. *Global report on the epidemiology and burden of sepsis: current evidence, identifying gaps and future directions*. World Health Organization, Geneva, 2020.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22 (2):199–210, November 2010.

Giorgio Patrini, Rianne van den Berg, Patrick Forre, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In *Uncertainty in Artificial Intelligence*, pages 733–743, July 2020.

Henning Petzka, Asja Fischer, and Denis Lukovnikov. On the regularization of wasserstein GANs. In *International Conference on Learning Representations*, March 2018.

Matthieu Scherpf, Felix Gräßer, Hagen Malberg, and Sebastian Zaunseder. Predicting sepsis with a recurrent neural network using the MIMIC III database. *Computers in Biology and Medicine*, 113:103395, Oct 2019.

Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 28, December 2015.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, June 2019.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, June 1964.

James E Smith and Robert L Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52:311–322, Mar 2006.

Rahul Venkataramani, Hariharan Ravishankar, and Saihareesh Anamandra. Towards Continuous Domain adaptation for Healthcare, Dec 2018.

Jie Wang and Yao Xie. A data-driven approach to robust hypothesis testing using sinkhorn uncertainty sets. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 3315–3320, July 2022.

Jie Wang, Rui Gao, and Yao Xie. Two-sample test using projected wasserstein distance. In *2021 IEEE International Symposium on Information Theory (ISIT)*, July 2021a.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, September 2021b.

Jie Wang, Rui Gao, and Hongyuan Zha. Reliable off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2011.04102*, January 2021c.

Jie Wang, Zhiyuan Jia, Hoover Yin, and Shenghao Yang. Small-sample inferred adaptive recoding for batched network coding. In *2021 IEEE International Symposium on Information Theory (ISIT)*, July 2021d.

Jie Wang, Rui Gao, and Yao Xie. Two-sample test with kernel projected wasserstein distance. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 8022–8055, March 2022.

Yuan Zhang, Xi Yang, Julie Ivy, and Min Chi. ATTAIN: Attention-based Time-Aware LSTM Networks for Disease Progression Modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 4369–4375, Aug 2019a.

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, Jun 2019b.

Yuanda Zhu, Janani Venugopalan, Zhenyu Zhang, Nikhil K. Chanani, Kevin O. Maher, and May D. Wang. Domain Adaptation Using Convolutional Autoencoder and Gradient Boosting for Adverse Events Prediction in the Intensive Care Unit. *Frontiers in Artificial Intelligence*, 5:640926, Apr 2022.