TRANSPORT-BASED TRANSFER LEARNING ON ELECTRONIC HEALTH RECORDS: APPLICATION TO DETECTION OF TREATMENT DISPARITIES

A PREPRINT

Wanxin Li
Department of Computer Science, University of British Columbia

Yongjin P. Park
BC Cancer Research, Part of Provincial Health Care Authority
Department of Pathology and Laboratory Medicine, University of British Columbia
Department of Statistics, University of British Columbia

Khanh Dao Duc
Department of Mathematics, University of British Columbia
Department of Computer Science, University of British Columbia

March 27, 2024

ABSTRACT

Many healthcare systems increasingly recognize the opportunities Electronic Health Records (EHRs) promise. However, EHR data sampled from different population groups can easily introduce unwanted biases, rarely permit individual-level data sharing, and make the data and fitted model hardly transferable across different population groups. In this paper, we propose a novel framework that leverages unbalanced optimal transport to facilitate the unsupervised transfer learning of EHRs between different population groups using a model trained in an embedded feature space. Upon deriving a theoretical bound, we find that the generalization error of our method is governed by the Wasserstein distance and unbalancedness between the source and target domains, as well as their labeling divergence, which can be used as a guide for binary classification and regression tasks. Our experiments, conducted on experimental datasets from MIMIC-III database, show that our transfer learning strategy significantly outperforms standard and machine learning transfer learning methods, with respect to accuracy and computational efficiency. Upon applying our framework to predict hospital duration for populations with different insurance plans, we finally find significant disparities across groups, suggesting our method as a potential tool to assess fairness in healthcare treatment.

Keywords Electronic Health Records · Transfer learning · Optimal transport · Treatment disparities

Data and Code Availability This paper uses the MIMIC-III dataset [Johnson et al., 2016], which is available on the PhysioNet repository [Moody et al., 2001]. The anonymized code repository is available at this link.

Institutional Review Board (IRB) This research does not require IRB approval.

1 Introduction

An Electronic Health Record (EHR) database is a digital platform that securely stores and manages comprehensive health information for individual patients, offering healthcare providers quick and efficient access to crucial medical data. Building a comprehensive and unbiased database of EHRs is a crucial first step to precision and personalized

medicine [Allen et al., 2012]. Comprehensive EHR databases generally achieve higher accuracy, avoiding potential issues of duplicate records and providing transparency among healthcare professionals, even across different healthcare systems [Menachemi and Collum, 2011]. Not just providing accurate information for each patient, a compendium of EHRs can serve as an important data set for augmenting human intelligence so that medical professionals can make informed decisions in everyday practice [Wagner et al., 2020].

In practice, it has been suggested that each EHR database was built for a different purpose to maximize its utility based on the needs [Kashani and Herasevich, 2015], making certain medical conditions highly, if not only, prevalent in specific studies and population groups [Woolf et al., 1955]. Having such an unequal distribution of medical conditions across different databases impedes our ability to diagnose rare conditions isolated within a specific population group. For example, alpha-1 antitrypsin deficiency, which affects the lungs and liver, is relatively uncommon in the general population but has higher prevalence rates in certain ethnic groups, such as those of Northern European descent [de Serres et al., 2007]. The variation in prevalence poses diagnostic challenges for patients in areas where the condition is less common. Moreover, unlike any other field of data science, data integration is practically not an option due to patient privacy and some hidden interests of stakeholders [Haas et al., 2011], which makes this problem even more challenging.

Overcoming this challenge introduces the need to transfer knowledge learned from data-rich population groups to data-rare population groups, which can be expressed as unsupervised *transfer learning* (TL) [Ganin and Lempitsky, 2015], in the case where labels are only available for the data-rich population group. Interestingly, methods that leverage tools from *Optimal Transport* (OT) have recently been proposed for unsupervised TL, as it aims to align and solve discrepancies between data distributions [Torres et al., 2021]. In the context of EHRs, unsupervised TL methods should also be careful of their specific characteristics, including heterogeneity, high-dimensionality and sparsity [Gupta et al., 2020]. In this paper, we hence introduce ***O**ptimal **T**ransport-based **T**ransfer learning for **E**lectronic **H**ealth **R**ecords*, *OTTEHR*, a novel method that leverages feature embeddings from EHRs and OT to perform unsupervised TL between unbalanced domains. Our contributions are as follows.

1. We introduce a method *OTTEHR* to enable unsupervised TL by applying barycentric projection from OT between unbalanced domains. Using experimental data from the MIMIC-III database, we show that our method outperforms standard and recent unsupervised TL methods with respect to accuracy and computational efficiency.
2. We establish a theoretical upper bound for the generalization error of *OTTEHR* and decompose it into a source error and labeling divergence terms that are universal, and a specific transport term, that allows us in practice to assess the suitability of our method on specific datasets.
3. Upon applying *OTTEHR* in the context of predicting patient duration in hospital using medical codes across different groups (e.g insurance), we show that our method allows us to quantify treatment disparities, suggesting potential applications to uncover treatment biases among subgroups and improving patient care.

2 Related Work

Application of OT-based transfer learning in healthcare To the best of our knowledge, our paper is the first OT-based TL study for EHRs where medical codes (e.g. ICD codes) are used as explanatory variables. OT-based semi-supervised and unsupervised transfer learning has been previously explored in the context of EHRs for sepsis prediction where vital signs and laboratory values are used as explanatory variables [Ding et al., 2023, Wang et al., 2022a]. More generally, Gautheron presented a method for TL using OT in the context of mapping prostate cancer between two magnetic resonance images generated by different scanners [Gautheron, 2017]. Chen *et al.* introduced a semi-supervised transfer learning approach that utilizes OT and frequency mixup to enhance the performance of electroencephalography (EEG)-based motor imagery recognition [Chen et al., 2022]. Liu *et al.* introduced a method that leverages TL and OT to enhance the performance of P300 detection, a widely used marker of cognitive function in brain-computer interfaces [Liu et al., 2020].

Medical record embedding As medical records often contain high-dimensional data, it is crucial to embed important information into low-dimensional vectors as a preprocessing step. Medical record embedding methods can be distinguished as supervised and unsupervised [Miotto et al., 2016]: Supervised methods estimate latent space for a specific task, such as predicting treatment outcomes, the duration of hospitalization, or socioeconomic impacts [Choi et al., 2017, Che et al., 2016, 2018], while unsupervised methods seek to locate patient data vectors, not aware of label information, but solely based on high-dimensional EHR information. In our proposed framework, we leverage Principle Component Analysis (PCA) to obtain some basic embedding.

Error upper bound for OT-based TL Ben-David *et al.* defined a formal model of transfer learning, also known as domain adaptation, in the case of binary classification [Ben-David et al., 2006], and derived a theoretical upper bound on the error of target data [Ben-David and Urner, 2014, Ben-David et al., 2010], which can be determined by various factors, including the error rates inherent to the source data, the divergence between the source and target distributions, and unwanted bias introduced by site-specific labeling mechanisms. Similarly, we extend the formal model of transfer learning to regression tasks and prove the upper bound on the error of the target data in the context of unbalanced OT. Courty et al. extended these concepts to explore the target error bound of balanced OT in a joint training process [Courty et al., 2017]. In comparison, our work focuses on a two-phase training process with unbalanced optimal transport. The two phase training process offers more stability and the ease of training. By separating the training into two phases, each loss function can be optimized more efficiently without the need to balance between them, which can be challenging in a joint loss function scenario [Malik et al., 2021, Li et al., 2022]. Also, in the context of EHRs, datasets of different groups are rarely balanced. In addition, the presence of clinical outliers for EHR may distort the OT plan. We utilized "unbalanced" OT to improve the robustness of our framework.

TL approaches In the context of standard statistical methods, Correlation Alignment (*CA*) [Sun et al., 2017], Transfer Component Analysis (*TCA*) [Pan et al., 2010], Euclidean space data alignment [He and Wu, 2019] and Geodesic Flow Kernel (*GFK*) [Gong et al., 2012] tackle TL by aligning statistical properties between source and target domains. These methods are suitable for both classification and regression tasks. In the context of machine learning, OT-based TL has been, to the best of our knowledge, exclusively been studied for classification tasks and more specifically tested on image classification tasks. OT-based transfer learning methods such as Joint Distribution Adaptation [Long et al., 2013], Deep Joint Distribution Optimal Transport (*deepJDOT*) [Damodaran et al., 2018], and Class-aware Sample Reweighting [Wang et al., 2023] leverages OT to reduce the Wasserstein distance between source and target domains, thereby aligning their distributions. Additionally, approaches such as Decomposed Transport Distance further refine these methodologies by introducing a decomposed distance metric within the optimal transport framework [Wang et al., 2022b]. On the other hand, non-OT-based TL has been studied for both classification and regression tasks. For classification tasks, moment-matching methods reduce the distribution discrepancies by matching statistics from two distinct distributions [Long et al., 2015, 2017, Maria Carlucci et al., 2017]. Adversarial learning methods minimize the distribution discrepancy by optimizing a selected function over the hypothesis space, concurrently learning feature representations to bridge the gap between domains [Ganin and Lempitsky, 2015, Tzeng et al., 2015, Ganin et al., 2016, Luo et al., 2017, Long et al., 2018, Zhang et al., 2019, Peng et al., 2019]. For regression tasks, most recent TL methods using Representation Subspace Distance (*RSD*) [Chen et al., 2021] and inverse GRAM matriecs (*daregram*) [Nejjar et al., 2023] learn a shared feature extractor by minimizing some discrepancies of source and target features.

3 Methods

3.1 Background

3.1.1 Optimal Transport

OT aims to solve a general transport problem where we consider moving data points in one distribution of mass to another at a minimal cost. In its discrete version, the OT problem can be formulated as follows: Consider two distributions μ_A and μ_B of point support $\mathbf{A} = \{\mathbf{a}_i \in \mathbb{R}^k,\ i = 1, \dots, n\}$ and $\mathbf{B} = \{\mathbf{b}_i \in \mathbb{R}^k,\ i = 1, \dots, m\}$ with probability mass functions ϕ_A and ϕ_B, respectively. We define a cost function $d : \mathbf{A} \times \mathbf{B} \to \mathbb{R}_{\geq 0}$ (e.g. the Euclidean distance between $\mathbf{a}_i$ with $\mathbf{b}_j$), and the associated objective function

$$\Phi_0(\pi) = \sum_{\substack{\mathbf{a}\in\mathbf{A}\\ \mathbf{b}\in\mathbf{B}}} d(\mathbf{a}, \mathbf{b})^p \pi(\mathbf{a}, \mathbf{b}), \tag{1}$$

where $\pi \in \mathrm{Mat}_{m,n}(\mathbb{R}_{\geq 0})$ and $p \in [1, +\infty]$. Minimizing this objective function with the marginal constraints $\sum_{\mathbf{b}\in\mathcal{B}} \pi(\mathbf{a}, \mathbf{b}) = \phi_A$ and $\sum_{\mathbf{a}\in\mathcal{A}} \pi(\mathbf{a}, \mathbf{b}) = \phi_B$ defines the classical OT problem. To deal with the computational cost of the classical OT formulation and handle distributions with different mass, *unbalanced entropy-regularized OT* was recently introduced [Pham et al., 2020] by adding additional regularization constraints to Equation 1 as

$$\Phi(\pi) = \Phi_0(\pi) - \frac{1}{\lambda}\mathcal{H}(\pi) + D_\varphi(\pi_A|\mu_A) + D_\varphi(\pi_B|\mu_B), \tag{2}$$

where $\mathcal{H}(\pi) = -\sum_{i,j} \pi(\mathbf{a}_i, \mathbf{b}_j) log(\pi(\mathbf{a}_i, \mathbf{b}_j))$, λ is the regularization parameter; $D_\varphi(\alpha, \beta)$ is the Csiszár $\varphi-$divergence and is given by, assuming that the discrete measures $\alpha = \sum_i^N \alpha_i \delta_{x_i}$ and $\beta = \sum_i^N \beta_i \delta_{x_i}$ share the same support

$\{x_i : i = 1, ..., N\}$,

$$D_\varphi(\alpha, \beta) = \sum_{\beta_i > 0} \varphi(\frac{\alpha_i}{\beta_i})\beta_i + \varphi'_\infty \sum_{\beta_i = 0} \alpha_i.$$

Minimizing Φ leads to define the *regularized unbalanced p-Wasserstein distance*, as

$$\mathcal{W}_p(\mu_A, \mu_B) = \min_{\pi \in \Pi(\mu_\mathcal{A}, \mu_\mathcal{B})} \Phi(\pi), \tag{3}$$

where Π is the set of positive matrices (that in this context define transport plans between μ_A and μ_B), $\pi_A = \sum_{\mathbf{b} \in \mathbf{B}} \pi(\mathbf{a}, \mathbf{b})$, $\pi_B = \sum_{\mathbf{a} \in \mathbf{A}} \pi(\mathbf{a}, \mathbf{b})$, and the minimizer of Equation (3) gives the *OT plan*. In the rest of the manuscript, we will work with $p = 1$, and will simply refer to the regularized unbalanced 1-Wasserstein distance as the *Wasserstein distance* for simplicity.

Upon finding π^*, and assuming $\sum_j \pi^*(\mathbf{a}_i, \mathbf{b}_j) > 0$, the transported features of $\mathbf{a}_i$ can then be derived by *barycentric projection*, as

$$T(\mathbf{a}_i) = \frac{\sum_j \pi^*(\mathbf{a}_i, \mathbf{b}_j)\mathbf{b}_j}{\sum_j \pi^*(\mathbf{a}_i, \mathbf{b}_j)}. \tag{4}$$

Remark: When $\sum_j \pi^*(\mathbf{a}_i, \mathbf{b}_j) = 0$, there is no mass transported from $\mathbf{a}_i$, and hence the barycentric projector is not defined.

3.1.2 Theoretical framework for Transfer Learning

We extend the theoretical framework of TL for binary classification [Ben-David et al., 2010] to include regression tasks as follows: Consider two spaces, a *source space*, represented by a distribution μ_S on *source domain* D_S and a labeling function f_S, and a *target space*, represented by a distribution μ_T on *target domain* D_T and a labeling function f_T. The label domain is $\{0, 1\}$ for binary classification and $\mathbb{R}$ for regression. We define a binary classifier/regressor a function $h : D_S \to \{0, 1\}$ or $\mathbb{R}$, so the probability that the binary classifier/regressor h disagrees with a labeling function f according to the distribution μ_S is

$$\epsilon_S(h, f) = \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu_\mathcal{S}} [|h(\mathbf{x}) - f(\mathbf{x})|],$$

where $|\cdot|$ is the l_1 norm metric on $\mathbb{R}$. In particular, the source error of h is $\epsilon_S(h) = \epsilon_S(h, f_S)$; and similarly, we define the target error, also known as the generalization error of h, as

$$\epsilon_T(h) = \epsilon_T(h, f_T). \tag{5}$$

Then, the problem of TL via OT aims to fine-tune a regressor h^*, learned from the source domain, into h' so that the resulting target error $\epsilon_T(h')$ is minimized.

3.2 Our Approach: *OTTEHR*

We propose our method *OTTEHR*, short for *Optimal Transport-based Transfer Learning for Electronic Health Records*, to address TL problems in EHR data using OT. We set up our notations as follows. Let $\mathcal{X}$ be the input domain and $\mathcal{Y}$ be the label domain. Let $(\mathbf{X}^S, \mathbf{Y}^S)$ denote the pair of source features and source labels and $(\mathbf{X}^T, \mathbf{Y}^T)$ denote the pair of target features and target labels, such that $\mathbf{X}^{\{S,T\}} = \{\mathbf{x}_i \in \mathbb{R}^q\}_{i=1}^{\{n,m\}}$ where n and m are the numbers of points in source and target domains and $\mathbf{Y}^S = \{y_i^S \in \mathbb{R} \text{ or } \in \{0, 1\}\}$. We preprocess the numerical medical codes into indicator arrays so that the lth coordinate in $\mathbf{x}_i$ denotes the presence of the lth code, using 0 or 1. We assume the unknown target labels $\mathbf{Y}^T$ are also in the same space as $\mathbf{Y}^S$.

OTTHER proceed in four main steps, also shown in Figure 1:

(1) Feature embedding Feature embedding is important to process EHRs as *(i)* EHRs are of high dimension, and the classifier/regressor h^* in step (2) and the OT plan π^* in step (3) cannot be learned well without proper dimensions; *(ii)* EHRs are also sparse, and feature embedding can be done effectively to resolve *(i)*.

To perform feature embedding, we run PCA on source features, that yields a mapping $g : \mathbb{R}^q \to \mathbb{R}^k$ where $k \ll q$, where k was chosen to explain approximately 75% of variances in source domains. Applying g to source features and target features then yields source embedding and target embedding as

$$D_S = \{z_i^S = g(\mathbf{x}_i^S) \in \mathbb{R}^k\}_{i=1}^n, \text{ and}$$
$$D_T = \{z_i^T = g(\mathbf{x}_i^T) \in \mathbb{R}^k\}_{i=1}^m, \text{ respectively.}$$

Note that our procedure allows to replace PCA by more advanced dimensionality reduction methods, which we mentioned in Section 6.

(2) Model training We train a classification or regression model $h^* : D_S \to Y_S$ using the source embedding D_S and source labels Y_S.

(3) Domain transportation We transport all target embeddings in D_T onto the source embedding domain D_S using equation 4, so the set of transported target embeddings is

$$\hat{D}_T = \{\hat{z}_i^T = T(g(\mathbf{x}_i^T)) \in \mathbb{R}^k\}_{i=1}^m.$$

Note that in the context of EHRs, datasets of different groups are rarely balanced. In addition, the presence of clinical outliers for EHR may distort the OT plan. Hence, we utilize "unbalanced" OT to improve the robustness of *OTTEHR*. Since the model training and domain transportation are performed separately, our framework is a two-phase training process compared to some previous works [Courty et al., 2017, Damodaran et al., 2018].

(4) Model prediction The source model is applied to the transported target embedding to obtain the "projected" target labels. We predict the label, $\hat{y}_i^T$ for all $\mathbf{x}_i^T \in \mathbf{X}^T$, using

$$\hat{y}_i^T = h^*(\hat{z}_i^T).$$

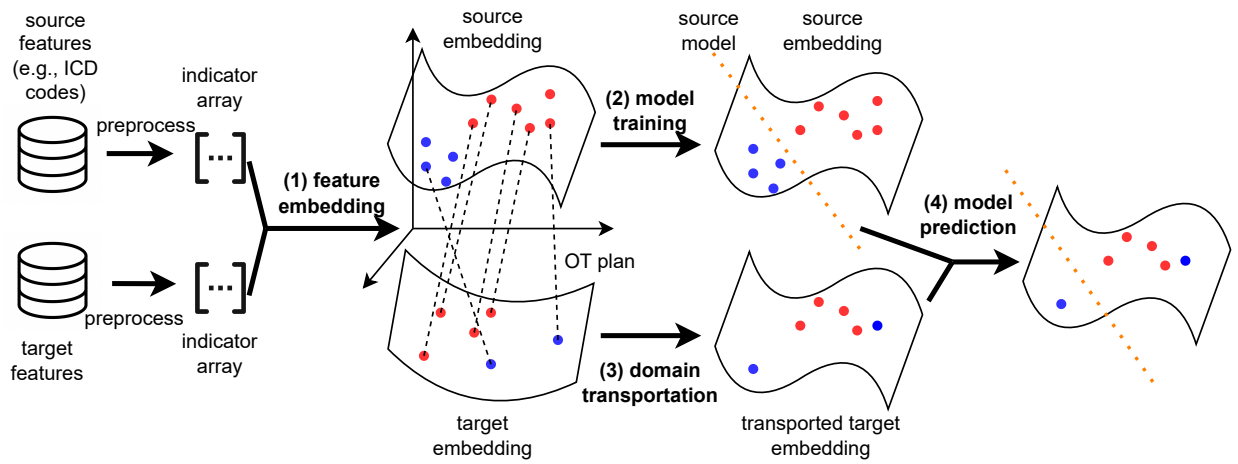

Figure 1: The framework of *OTTEHR*. The four major steps (1) feature embedding, (2) model training, (3) domain transportation and (4) model prediction are in bold. Details of these four steps can be found in Section 3.2. The indicator arrays consist of 0s and 1s where the $(i, l)^{th}$ entry indicates the presence of the lth medical code for the ith source/target feature. The orange dotted lines denote the trained source classifier from step (2). The black dotted lines denote the OT plan learned from step (3). The colors of the points on the embedding spaces represent different classes. Note that when using *OTTEHR*, *(i)* the labels for the target features are not required and they present here for the visualization purpose; *(ii)* the source model can be either a classifier or a regressor.

3.3 Implementation details

We implemented *OTTEHR* in Python 3, with the source code available at this link. To solve OT problems, our package depends on the *POT* library [Reinger, 2021]. We ran our experiments on a workstation with 32 central processing units powered by AMD Ryzen Threadripper 2950X 16-Core Processor, 125GB of RAM, and x86_64 Ubuntu 20.04.5 LTS. We used Euclidean distance (l_2 norm) as the cost function for OT and Kullback-Leibler divergence as the mass divergence D_φ. For OT hyperparameters, we set the entropic regularization parameter λ to 0.1 and the mass regularization parameter to 1.

4 Experiments

In this section, we first introduce our datasets derived from MIMIC-III database [Johnson et al., 2016], and then describe the set up of our experiments.

4.1 Dataset

The MIMIC-III database is a large, open-source database comprising anonymized health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 [Johnson et al., 2016]. It is a relational database consisting of 26 tables consisting of 8,922 unique ICD codes. We focused on the patient table and admission table. The patient table contains 46,520 patients, with 26,121 males and 20,399 females; each patient is attached with patient ID, gender, expire flag, date of birth, etc. The admission table contains 58,976 admissions, and each admission is attached with patient ID, admission type, admission time, and discharge time.

We first merged the admission table with the patient table by indexing unique patient IDs. For each admission, we calculated the duration in the hospital by taking the difference between the discharge time and the admission time (in seconds). We then treated the presence of each ICD code as the explanatory variable and used the duration in the hospital as the response. Hence, we divided the admissions into source and target by insurance groups or marital status groups. Insurance groups include "Self_Pay," "Private," "Government," "Medicare," "Medicaid." Marital status groups include "Separated," "Divorced," "Married," "Widowed," "Single."

4.2 Experimental Setup

We ran *OTTEHR* to predict duration in hospital using linear regression models on all possible pairs within insurance (marital status) groups, considering one group as the source and the other as the target, resulting in 40 distinct experiments (i.e., $5 \times 4 \times 2$). We validated our choice of group by visualizing in Figure S1 and Figure S2 the t-distributed Stochastic Neighbor Embedding (t-SNE) [Van der Maaten and Hinton, 2008] of the feature embeddings between pairwise groups for all insurance and marital status groups, showing that most of the distinct groups form their own clusters, and differences in the feature embedding space between each pair of source and target groups are significant. For each pair of source and target groups, we randomly sampled 120 admissions (training) from the source group and 100 (testing) admissions from the target group, and we conducted the same type of experiment repeatedly 100 times. The number of ICD codes varies from 700 to 900 for each experiment depending on the selected admissions. During the feature embedding step, we convert these ICD codes into a 50-dimensional space ($k = 50$). We analyzed the empirical relationship between the target error of *OTTEHR* and other terms presented in the upper bound in Section 5.1. Then, we benchmarked *OTTEHR* against existing TL methods.

5 Results

In this section, we first derive a theoretical upper bound on the target error for our method, as introduced in the methods Section 3.1.2. Using datasets from the MIMIC-III database, we validate the theoretical upper bound with empirical estimates obtained from our experiments. We then benchmark *OTTEHR*'s performance against its competitors, focusing on accuracy and computational cost. Finally, we use *OTTEHR* to reveal notable differences in duration in hospital based on insurance plans, suggesting *OTTEHR* as a potential tool to quantify treatment disparities.

5.1 Upper Bound for Binary Classification and Regression

To study the theoretical accuracy of *OTTEHR*, we derived an upper bound on the target error in the case of binary classification and regression. Our analysis focuses on Lipschitz continuous models. In practice, many models, including

for EHRs, satisfy this condition [Goldstein et al., 2017, Clegg et al., 2016, Harutyunyan et al., 2019] or can be approximated as Lipschitz continuous [Bartlett et al., 2017, Nair and Hinton, 2010, Ioffe and Szegedy, 2015], and assuming Lipschitz continuity or linearity is common in theoretical studies of transfer learning [Tong et al., 2021, Tian and Feng, 2023, Cai et al., 2021].

As we adapt embeddings from the target domain to the source domain using the barycentric operator T, the fine-tuning of the learned function on the source h^* yields $h' = h^* \circ T$. Assuming $\sum_y \pi^*(x, y) \neq 0$ (see Equation (4)), we can control the target error by the following theorem.

Theorem 1 (Upper bound for target error)**.** *Let μ_T (μ_S) be a discrete target (source) distribution defined on a domain $D_T(D_S)$ and with probability mass function $\phi_T(\phi_S)$. Let $h' = h^* \circ T$, where h^* is Lipschitz continuous and T is the barycentric projection (Equation* (4)*). The target error $\epsilon_T(h')$ defined by Equation* (5) *is bounded by:*

$$\epsilon_T(h') \leq \underbrace{\epsilon_T(h^*, f_S)}_{\text{source model error on the target domain}}$$
$$\underbrace{+ K\{\mathcal{W}_1(\mu_T, \mu_S) + M \sum_{x \in D_T} |\sum_{y \in D_S} \pi^*(x, y) - \phi_T(x)|\}}_{\text{transport}}$$
$$+ \underbrace{\mathbb{E}_{x \sim \mu_T}\{|f_T(x) - f_S(x)|\}}_{\text{labeling divergence}},$$

where K is the Lipschitz continuous constant for h^*, π^* is the OT plan and $M = \max_{x \in D_T, y \in D_S} \|y - x\|$.

The proof can be found in Appendix A. Ignoring the defined constants M and K, we interpret the upper bound as follows. The first term is the source model error evaluated on the target domain, illustrated by Figure S3. The second "transport term" is composed of two parts: *(i)* $\mathcal{W}_1(\mu_T, \mu_S)$ that denotes the Wasserstein between μ_T and μ_S (see Equation (3)), and *(ii)* $\sum_{x \in D_T} |\sum_{y \in D_S} \pi^*(x, y) - \phi_T(x)|\}$ that denotes the variance between mass at x and the transported mass from x to D_S, indicating the "unbalancedness" of the OT plan (the larger $D_\phi(\pi_A|\mu_A)$ and $D_\phi(\pi_B|\mu_B)$ in Equation (2), the larger this term will be). The last "labeling divergence" term $\mathbb{E}_{x \sim \mu_T}[|f_T(x) - f_S(x)|]$ denotes the divergence between the source and target true labeling functions.

Note that when OT is used rather than unbalanced OT, we have the marginal constraints $\sum_{x \in D_T} \pi^*(x, y) = \phi_T(x)$, resulting in eliminating unbalancedness term from Theorem 1. Also note that when h^* is linear, a special case of Lipschitz continuity, K in the theorem can be substituted by $|||h^*||| = \sup_{\|x\|=1, x \in D_T} \|h^*(x)\|$.

This generalization error bound provides a guide for analyzing and interpreting the performance of our method. More precisely, if the transport term is significantly smaller compared to the source model error and labeling divergence, then using OT to map individuals to another group should be a suitable option for the transfer learning task. Note that the estimation of the labeling divergence requires access to ground-truth target labels y_i^T to estimate the labeling function on the target domain f_T, which might not be immediately available in practice. Nevertheless, we can mitigate this issue if we have partial access to the target labels or prior knowledge of f_T.

5.2 Predictive performance is dominated by the labeling divergence

To demonstrate the potential of our method to be applied to real data, we ran our method using an experimental dataset of ICD codes (see Section 4.1). More precisely, we focused on how the duration of hospitalization can be predicted by medical codes for different insurance or marital status groups (as detailed in Section 4.2), as a simple but relevant test case [Goldman and Smith, 2002, Umberson, 1987]. First, we compared the target errors of *OTTEHR*, guided by Theorem 1. Note that in this context, using the same model for estimating $\hat{f}_S$ as h^* leads to $h^*(x) = \hat{f}_S(x)$ for all $x \in D_T$, and $\epsilon_T(h^*, \hat{f}_S) = 0$, so it was not necessary to study the relationship between the target error and the source model error on the target domain, leaving the transport term and the labeling divergence to be studied.

For insurance group experiments, as shown in Figure 2, we separately plotted the target error against the transport term and against the labeling divergence with the results for all the insurance group experiments combined. We observed a strong correlation between the target error and labeling divergence, with a Pearson correlation coefficient (PCC) of 0.70 (Figure 2.B), in contrast to a weaker correlation between the target error and the transport term, with a PCC of

0.16 (Figure 2.A). Marital status group experiments yields similar patterns, with a PCC of 0.63 for the transport term (Figure S4.B) and 0.09 for the labeling divergence (Figure S4.A). In both cases, we explain these results by the large difference in orders of magnitude between the labeling divergence (10^8) and the transport term (10^0). Further analysis of pairwise insurance (marital status) groups confirmed consistent trends (see Figures S5 to S8), indicating that in our experiments, the target error is overall dominated by the labeling divergence. As previously discussed in Section 5.1, since the transport term is significantly smaller than the labeling divergence, *OTTEHR* is suitable for solving these transfer learning tasks.

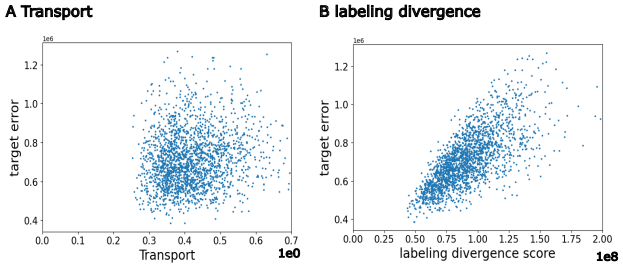

Figure 2: Bound analysis for all insurance group experiments. The relationship between target error and (**A**) transport term with a PCC of 0.70, and (**B**) labeling divergence with a PCC of 0.16, combining the results for all pairwise insurance group experiments.

5.3 Benchmarking of accuracy and computation time

To assess the empirical performance of *OTTEHR*, we benchmarked *OTTEHR* against the standard statistical methods *Transfer Component Analysis (TCA)* [Pan et al., 2010], *Correlation Analysis (CA)* [Sun et al., 2017], *Geodesic Flow Kernel (GFK)* [Gong et al., 2012], machine learning OT-based method *deepJDOT* [Damodaran et al., 2018], and machine learning non-OT-based methods *Representation Subspace Distance (RSD)* [Chen et al., 2021] and *inverse Gram matrices (daregram)* [Nejjar et al., 2023] (also see Section 2) on the predictions for target groups for the transfer learning tasks detailed in Section 4.2 using mean absolute error (*MAE*) and root mean square error (*RMSE*) [Chai and Draxler, 2014].

For insurance group experiments, violin plots in Figure 3 shows the log-transformed *MAE* and *RMSE*, with smaller values indicating better performance. We notably observed that *OTTEHR*'s median *MAE*/*RMSE* is smaller than those of *TCA*, *TCA*, *GFK*, *RSD* and *daregram* with comparable standard deviations. Although *OTTEHR*'s median *MAE*/*RMSE* is slightly larger than that of *deepJDOT*, its standard deviation is significantly smaller. We also provided detailed results in Tables S1 and S2 for pairwise insurance groups, showing the medians and standard deviations of *MAE*/*RMSE* for all the methods, and outperformance ratios of *OTTEHR* to other methods. Overall, *OTTEHR* achieved a 14% to 28% reduction in median *MAE/RMSE* compared to *TCA*, *CA*, *GFK* and *RSD* and *daregram* and a 55% to 68% reduction in the standard deviation of the log-transformed *MAE/RMSE* compared to *deepJDOT*. Similar trends are observed in marital status group experiments, with *OTTEHR* outperforming *TCA*, *CA*, *GFK* and *RSD* and *daregram* by 14% to 26% in median *MAE/RMSE* and by 59% to 76% in the standard deviation of the log-transformed *MAE/RMSE* (refer to Figure S9 and tables S3 and S4 for details).

In addition to accuracy, we compared *OTTEHR* with competing methods in terms of average computation time per experiment. Table S5 shows *OTTEHR* is the fastest among *TCA, CA, GFK, deepJDOT, RSD* and *daregram*. Specifically,*OTTEHR*'s runtime is on par with *TCA, GFK, RSD* and *daregram*, 24.8 ($20116.7/810.8$) times faster than that of *CA*, and 42.7 ($34642.6/810.8$) times faster than that of *deepJDOT*.

5.4 *OTTEHR* reveals treatment disparities based on insurance plans

After confirming the empirical performance of *OTTEHR*, we applied it to quantify treatment disparities based on insurance plans and predict the potential impact of transitioning between different plans. We focused on Medicaid, a collaborative program that assists individuals with limited income and resources [Gruber, 2003], Medicare, that provides coverage to people aged 65 and older, as well as to younger individuals with certain disabilities and diseases [Finkelstein and McKnight, 2008], and Private insurance, offered by various companies, that offers greater flexibility in healthcare services [Cutler and Gruber, 1997]. People often switch from Medicaid and Medicare to private insurance due to factors such as reductions in federal matching funds or when their preferred healthcare providers fall out of their Medicare network [Foundation, 2023, Kaiser Family Foundation, 2023, 2021]. Conversely, transitions to Medicaid and Medicare

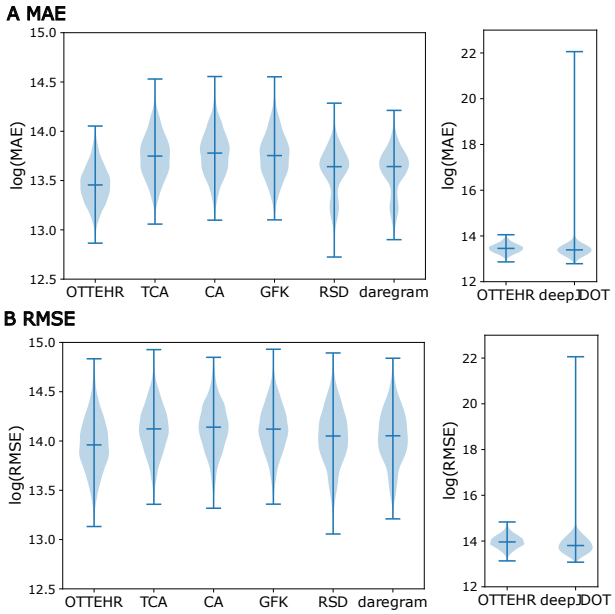

Figure 3: Benchmark results for all insurance group experiments. Violin plots of the log of (**A**) *MAE* and (**B**) *RMSE* between projected duration and observed duration on target admissions for insurance groups using *OTTEHR*, *TCA*, *CA*, *GFK*, *RSD* and *daregram*. The blue bars denote the medians. The heights of the violin plots denote the variability, positively correlated with the standard deviations. Smaller *MAE* and *RMSE* values indicate better performance.

from private insurance can result from changes in age, income, or health status Long et al. [2014], Baicker and Chandra [2006].

Specifically, we considered the projected duration obtained from transfer learning to be significantly reduced if it is at least 300 hours (12.5 days) shorter than the originally observed duration. In Figure 4, we showed kernel density estimate plot (KDE) of projected duration versus the original duration in hours for admissions transitioning (**A**) from Medicaid to private insurance, (**B**) from private insurance to Medicaid, (**C**) from Medicare to private insurance, and (**D**) from private insurance to Medicare, with blue dots denoting admissions with significantly reduced durations, where we observed very different proportions of admissions with significantly reduced durations. Our findings indicate that 13.1% of admissions would result in significantly reduced durations when transitioning from Medicaid to private insurance, compared to only 9.5% with such reductions when transitioning from private insurance to Medicaid, making a difference of 3.6%. Similarly, 9.0% of admissions would result in significantly reduced durations when transitioning from Medicare to private insurance, compared to 10.5% with such reductions when transitioning from private insurance to Medicaid, making a difference of 1.5%. The larger difference of bi-directional transition process between private insurance and Medicaid suggests more disparities between private insurance and Medicaid compared to those between private insurance and Medicare. Such results can be generalized to more insurance groups (e.g., Government), as shown in Figure S10, where percentages of admissions with significantly reduced durations are shown across all pairs of insurance groups.

6 Discussion

This paper presents *OTTEHR*, an OT-based unsupervised TL framework for EHRs. While biased models can lead to incorrect diagnoses, treatments, and healthcare decisions [Chen et al., 2023, Mittermaier et al., 2023], *OTTEHR* can potentially alleviate these biases by leveraging OT when comparing different population groups. Our study more precisely establishes a theoretical upper bound for the generalization error. Interestingly, we decomposed this bound into some general terms (namely the source error and the labeling divergence) that are shared by any transfer learning method, and a specific transport term, that we can use in practice to evaluate the suitability of our method on real data, as shown in our application to the MIMIC-III dataset. We also note that all these terms are computable (as we did in Figures 2 and S4 to S8) or can be estimated if we have limited access to target labels or some prior knowledge about the target domain's labeling function. Overall, our benchmarking suggests that in the context of EHRs, aligning probability distributions between the source and target domains can be highly effective. Upon focusing on predicting

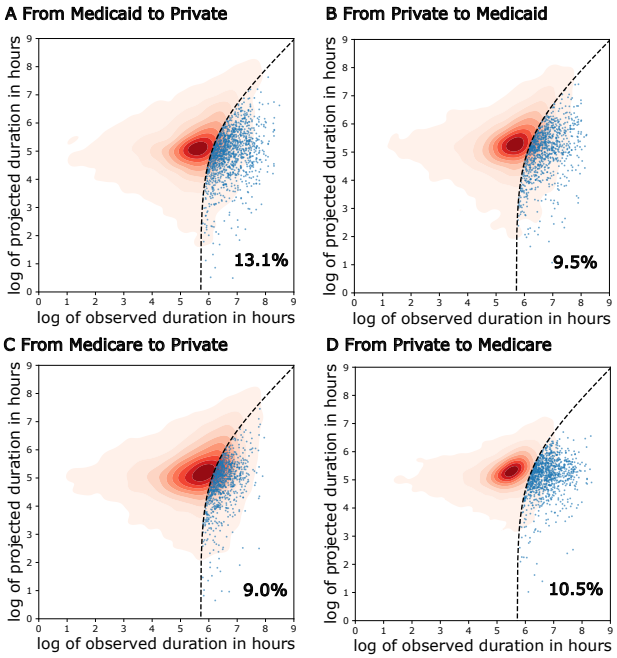

Figure 4: Admissions with significantly reduced durations of stay in hospital when transitioning between private insurance and Medicaid/Medicare. The KDE plot of the log of projected duration in hours versus the log of observed duration in hours for admissions transitioning (**A**) from Medicaid to private insurance, (**B**) from private insurance to Medicaid, (**C**) from Medicare to private insurance, and (**D**) from private insurance to Medicare. The dotted black lines denote projected duration = observed duration -300. The blue dots denote admissions with significantly reduced duration of stay in hospital, where the projected duration is at least 300 hours (12.5 days) less than the observed duration. The annotated percentages are the proportions of significantly reduced durations.

duration in hospital, we also detected significant differences underlying treatment disparities across insurance groups, suggesting our method's potential for uncovering treatment biases among subgroups and improving patient care.

In conclusion, we can list several potential future directions. While we have shown that *OTTEHR* enables knowledge transfer between datasets, it would be interesting to further evaluate *OTTEHR* on other relevant regression and classification tasks and other demographic factors. These include predicting the time interval between consecutive visits [Poole et al., 2016] and mortality rates [Goodacre et al., 2006], using appropriate and potentially larger datasets [Pader et al., 2021, Sudlow et al., 2015]. Also, extending our method to semi-supervised transfer learning [Wei et al., 2019] and designing a unified model that simultaneously solves feature embedding and classification problems [Song et al., 2017] could improve predictive performance on target domains with limited labeled data. From a theoretical perspective, it would be beneficial to extend the upper bound theorem, to include other tasks, such as, multi-class classification, non-continuous Lipschitz models, and other OT-metrics that can be more suitable for transfer learning across different EHR systems [Séjourné et al., 2021]. Finally, there are several other potential areas for improving our current approach, including reducing its computational complexity to handle larger datasets, optimizing the embedding with more complex manifold learning techniques, and integrating heterogeneous information, such as laboratory results and doctor's notes.

Acknowledgments We acknowledge the help from Zhenyuan Zhang and Greg d'Eon for discussing the proof for Theorem 1.

References

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

George B Moody, Roger G Mark, and Ary L Goldberger. Physionet: A web-based resource for the study of physiologic signals. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):70–75, 2001.

Naomi Allen, Cathie Sudlow, Paul Downey, Tim Peakman, John Danesh, Paul Elliott, John Gallacher, Jane Green, Paul Matthews, Jill Pell, et al. UK biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, 1(3):123–126, 2012.

Nir Menachemi and Taleah H Collum. Benefits and drawbacks of electronic health record systems. *Risk Management and Healthcare Policy*, pages 47–55, 2011.

Tyler Wagner, FNU Shweta, Karthik Murugadoss, Samir Awasthi, AJ Venkatakrishnan, Sairam Bade, Arjun Puranik, Martin Kang, Brian W Pickering, John C O'Horo, et al. Augmented curation of clinical notes from a massive ehr system reveals symptoms of impending covid-19 diagnosis. *Elife*, 9:e58227, 2020.

Kianoush Kashani and Vitaly Herasevich. Utilities of electronic medical records to improve quality of care for acute kidney injury: past, present, future. *Nephron*, 131(2):92–96, 2015.

Barnet Woolf et al. On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19(4): 251–253, 1955.

Frederick J de Serres, Ignacio Blanco, et al. Pi s and pi z alpha-1 antitrypsin deficiency worldwide. a review of existing genetic epidemiological data. *Monaldi Archives for Chest Disease*, 67(4), 2007.

Sebastian Haas, Sven Wohlgemuth, Isao Echizen, Noboru Sonehara, and Günter Müller. Aspects of privacy for electronic health records. *International Journal of Medical Informatics*, 80(2):e26–e31, 2011.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

Luis Caicedo Torres, Luiz Manella Pereira, and M Hadi Amini. A survey on optimal transport for machine learning: Theory and applications. *ArXiv:2106.01963*, 2021.

Vagisha Gupta, Shelly Sachdeva, and Subhash Bhalla. A novel deep similarity learning approach to electronic health records data. *IEEE Access*, 8:209278–209295, 2020.

Ruiqing Ding, Yu Zhou, Jie Xu, Yan Xie, Qiqiang Liang, He Ren, Yixuan Wang, Yanlin Chen, Leye Wang, and Man Huang. Cross-hospital sepsis rarly detection via semi-supervised optimal transport with self-paced ensemble. *IEEE Journal of Biomedical and Health Informatics*, 2023.

Jie Wang, Ronald Moore, Yao Xie, and Rishikesan Kamaleswaran. Improving sepsis prediction model generalization with optimal transport. In *Machine Learning for Health*, pages 474–488. PMLR, 2022a.

Léo Gautheron. Domain adaptation using optimal transport: Application to prostate cancer mapping. Master's thesis, Jean Monnet University, 2017.

Peiyin Chen, He Wang, Xinlin Sun, Haoyu Li, Celso Grebogi, and Zhongke Gao. Transfer learning with optimal transportation and frequency mixup for EEG-based motor imagery recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2866–2875, 2022.

Zhenjie Liu, Qiang Qiu, Jun Li, Lizhe Wang, and Antonio Plaza. Geographic optimal transport for heterogeneous data: Fusing remote sensing and social media. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6935–6945, 2020.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6(1):1–10, 2016.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine Learning for Healthcare conference*, pages 286–305. PMLR, 2017.

Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for ICU outcome prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 371. American Medical Informatics Association, 2016.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Shai Ben-David and Ruth Urner. Domain adaptation–can quantity compensate for quality? *Annals of Mathematics and Artificial Intelligence*, 70:185–202, 2014.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach Learn*, 79:151–175, 2010.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Farjad Malik, Simon Wouters, Ruben Cartuyvels, Erfan Ghadery, and Marie-Francine Moens. Two-phase training mitigates class imbalance for camera trap image classification with cnns. *arXiv preprint arXiv:2112.14491*, 2021.

Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. Improving word translation via two-stage contrastive learning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4353–4374, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.299. URL `https://aclanthology.org/2022.acl-long.299`.

Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pages 153–171, 2017.

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, 2010.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2019.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2012.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2200–2207, 2013.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463, 2018.

Shengsheng Wang, Bilin Wang, Zhe Zhang, Ali Asghar Heidari, and Huiling Chen. Class-aware sample reweighting optimal transport for multi-source domain adaptation. *Neurocomputing*, 523:213–223, 2023.

Bilin Wang, Shengsheng Wang, Zhe Zhang, Xin Zhao, and Zihao Fu. Decomposed-distance weighted optimal transport for unsupervised domain adaptation. *Applied Intelligence*, 52(12):14070–14084, 2022b.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105. PMLR, 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. Autodial: Automatic domain alignment layers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5067–5075, 2017.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. *Advances in Neural Information Processing Systems*, 30, 2017.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*, pages 5102–5112. PMLR, 2019.

Xinyang Chen, Sinan Wang, Jianmin Wang, and Mingsheng Long. Representation subspace distance for domain adaptation regression. In *International Conference on Machine Learning*, pages 1749–1759, 2021.

Ismail Nejjar, Qin Wang, and Olga Fink. Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11744–11754, 2023.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *International Conference on Machine Learning*, pages 7673–7682. PMLR, 2020.

Royce Reinger. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (11), 2008.

Benjamin A Goldstein, Ann Marie Navar, and Rickey E Carter. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *European Heart Journal*, 38(23): 1805–1814, 2017.

Andrew Clegg, Chris Bates, John Young, Ronan Ryan, Linda Nichols, Elizabeth Ann Teale, Mohammed A Mohammed, John Parry, and Tom Marshall. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age and Ageing*, 45(3):353–360, 2016.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. pmlr, 2015.

Xinyi Tong, Xiangxiang Xu, Shao-Lun Huang, and Lizhong Zheng. A mathematical framework for quantifying transferability in multi-source transfer learning. *Advances in Neural Information Processing Systems*, 34:26103–26116, 2021.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.

Guanyu Cai, Lianghua He, MengChu Zhou, Hesham Alhumade, and Die Hu. Learning smooth representation for unsupervised domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4181–4195, 2021.

Dana P Goldman and James P Smith. Can patient self-management help explain the ses health gradient? *Proceedings of the National Academy of Sciences*, 99(16):10929–10934, 2002.

Debra Umberson. Family status and health behaviors: Social control as a dimension of social integration. *Journal of health and social behavior*, pages 306–319, 1987.

Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae). *Geoscientific Model Development*, 7(1):1525–1534, 2014.

Jonathan Gruber. Medicaid. In *Means-tested transfer programs in the United States*, pages 15–78. University of Chicago Press, 2003.

Amy Finkelstein and Robin McKnight. What did medicare do? the initial impact of medicare on mortality and out of pocket medical spending. *Journal of Public Economics*, 92(7):1644–1668, 2008.

David M Cutler and Jonathan Gruber. Medicaid and private insurance: evidence and implications. *Health Affairs*, 16 (1):194–200, 1997.

Kaiser Family Foundation. What happens after people lose medicaid coverage? *KFF (Kaiser Family Foundation)*, 2023. Accessed: 2024-02-15.

Kaiser Family Foundation. 10 things to know about the unwinding of the medicaid continuous enrollment requirement, 2023. Accessed: 2024-02-15.

Kaiser Family Foundation. Half of all eligible medicare beneficiaries are now enrolled in private medicare advantage plans, 2021. Accessed: 2024-02-15.

Sharon K Long, Genevieve M Kenney, Stephen Zuckerman, Dana E Goin, Douglas Wissoker, Fredric Blavin, Linda J Blumberg, Lisa Clemans-Cope, John Holahan, and Katherine Hempstead. The health reform monitoring survey: Addressing data gaps to provide timely insights into the affordable care act. *Health Affairs*, 33(1):161–167, 2014.

Katherine Baicker and Amitabh Chandra. The labor market effects of rising health insurance premiums. *Journal of Labor Economics*, 24(3):609–634, 2006.

Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6):719–742, 2023.

Mirja Mittermaier, Marium M Raza, and Joseph C Kvedar. Bias in ai-based models for medical applications: challenges and mitigation strategies. *NPJ Digital Medicine*, 6(1):113, 2023.

Sarah Poole, Shaun Grannis, and Nigam H Shah. Predicting emergency department visits. *AMIA Summits on Translational Science Proceedings*, 2016:438, 2016.

S Goodacre, J Turner, and Jon Nicholl. Prediction of mortality among emergency medical admissions. *Emergency Medicine Journal*, 23(5):372–375, 2006.

Joy Pader, Robert B Basmadjian, Dylan E O'Sullivan, Nicole E Mealey, Yibing Ruan, Christine Friedenreich, Rachel Murphy, Edwin Wang, May Lynn Quan, and Darren R Brenner. Examining the etiology of early-onset breast cancer in the canadian partnership for tomorrow's health (CanPath). *Cancer Causes & Control*, 32(10):1117–1128, 2021.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779, 2015.

Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2019.

Tiecheng Song, Jianfei Cai, Tianqi Zhang, Chenqiang Gao, Fanman Meng, and Qingbo Wu. Semi-supervised manifold-embedded hashing with joint feature representation and classifier learning. *Pattern Recognition*, 68:99–110, 2017.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

A Proof of Theorem - Upper Bound for Binary Classification and Regression

Let μ_T (μ_S) be a discrete target (source) distribution defined on a domain $D_T(D_S)$ and with probability mass function $\phi_T(\phi_S)$. Let $h' = h^* \circ T$, where h^* is Lipschitz continuous and T is the barycentric projection (Equation (4)). The target error $\epsilon_T(h')$ defined by Equation (5) is bounded by

$$\begin{aligned}
&\epsilon_T(h') \leq \epsilon_T(h^*, f_S) \\
&+ K\{\mathcal{W}_1(\mu_T, \mu_S) + M \sum_{x \in D_T} |\sum_{y \in D_S} \pi^*(x,y) - \phi_T(x)|\} \\
&+ \mathbb{E}_{x \sim \mu_T}\{|f_T(x) - f_S(x)|\},
\end{aligned}$$

where K is the Lipschitz continuous constant for h^*, π^* is the OT plan and $M = \max_{x \in D_T, y \in D_S} \|y - x\|$.

Proof. We first rewrite the target error as

$$\begin{aligned}
&\epsilon_T(h^* \circ T) = \epsilon_T(h^* \circ T, f_T) \\
&= \epsilon_T(h^* \circ T, f_T) + \epsilon_T(h^*, f_T) - \epsilon_T(h^*, f_T) \\
&+ \epsilon_T(h^*, f_S) - \epsilon_T(h^*, f_S) \\
&= \epsilon_T(h^*, f_S) + \underbrace{\epsilon_T(h^* \circ T, f_T) - \epsilon_T(h^*, f_T)}_{(*)} \\
&+ \underbrace{\epsilon_T(h^*, f_T) - \epsilon_T(h^*, f_S)}_{(**)}.
\end{aligned} \tag{6}$$

Since K is the Lipschitz constant for h^*, for all x, there exists $K > 0$ such that $\|h^* \circ T(x) - h^*(x)\| \leq K\|T(x) - x\|$. We now separately analyze $(*)$ and $(**)$. By the triangle inequality,

$$\begin{aligned}
&\epsilon_T(h^* \circ T, f_T) - \epsilon_T(h^*, f_T) \\
&= \mathop{\mathbb{E}}_{x \sim \mu_T}[\|h^* \circ T(x) - f_T(x)\|] - \mathop{\mathbb{E}}_{x \sim \mu_T}[\|h^*(x) - f_T(x)\|] \\
&= \mathop{\mathbb{E}}_{x \sim \mu_T}[\|h^* \circ T(x) - f_T(x)\| - \|h^*(x) - f_T(x)\|] \\
&\leq \mathop{\mathbb{E}}_{x \sim \mu_T}[\|h^* \circ T(x) - h^*(x)\|] \\
&\leq \mathop{\mathbb{E}}_{x \sim \mu_T}[K\|T(x) - x\|] \\
&= K \mathop{\mathbb{E}}_{x \sim \mu_T}[\|T(x) - x\|],
\end{aligned} \tag{7}$$

Let $\rho(x) = \frac{\phi_T(x)}{\sum_{y \in D_S} \pi^*(x,y)}$ ($\rho(x)$ is well defined since $\sum_{y \in D_S} \pi^*(x,y) \neq 0$). We then obtain

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x \sim \mu_T}[\|T(x) - x\|] \\
&= \sum_{x \in D_T} \|T(x) - x\| \phi_T(x) \\
&= \sum_{x \in D_T} (\|[\frac{1}{\sum_{y \in D_S} \pi^*(x,y)} \sum_{y \in D_S} y\pi^*(x,y)] - x\| \rho(x) \sum_{y \in D_S} \pi^*(x,y)) \\
&= \sum_{x \in D_T} \|\rho(x) \sum_{y \in D_S} y\pi^*(x,y) - \rho(x) \sum_{y \in D_S} \pi^*(x,y)x\| \\
&= \sum_{x \in D_T} \rho(x) \sum_{y \in D_S} \|(y - x)\pi^*(x,y)\| \\
&= \sum_{x \in D_T} (1 - (1 - \rho(x))) \sum_{y \in D_S} \|(y - x)\pi^*(x,y)\|.
\end{aligned} \tag{8}$$

We can thus further bound $\mathbb{E}_{x\sim\mu_T}[\|T(x)-x\|]$ as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mu_T}[\|T(x)-x\|] \leq \sum_{x\in D_T}\sum_{y\in D_S}\|y-x\|\pi^*(x,y)\\
&+\sum_{x\in D_T}|1-\rho(x)|\sum_{y\in D_S}\|y-x\|\pi^*(x,y)\\
&\leq \sum_{x\in D_T}\sum_{y\in D_S}\|y-x\|\pi^*(x,y)-\frac{1}{\lambda}\mathcal{H}(\pi^*)\\
&+D_\varphi(\pi_S^*|\phi_S)+D_\varphi(\pi_T^*|\phi_T)\\
&+\sum_{x\in D_T}|1-\rho(x)|\sum_{y\in D_S}\|y-x\|\pi^*(x,y)\\
&\leq \mathcal{W}(\mu_T,\mu_S)+M\sum_{x\in D_T}|1-\rho(x)|\sum_{y\in D_S}\pi^*(x,y),
\end{aligned} \tag{9}$$

where $M=\max_{x\in D_T, y\in D_S}\|y-x\|$. Finally,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mu_T}[\|T(x)-x\|]\\
&\leq \mathcal{W}(\mu_T,\mu_S)+M\sum_{x\in D_T}|1-\frac{\phi_T(x)}{\sum_{y\in D_S}\pi^*(x,y)}|\sum_{y\in D_S}\pi^*(x,y)\\
&=\mathcal{W}(\mu_T,\mu_S)+M\sum_{x\in D_T}|\sum_{y\in D_S}\pi^*(x,y)-\phi_T(x)|,
\end{aligned} \tag{10}$$

which yields the upper bound for $(*)$

$$\begin{aligned}
&\epsilon_T(h^*\circ T,f_T)-\epsilon_T(h^*,f_T)\\
&\leq K\{\mathcal{W}(\mu_T,\mu_S)+M\sum_{x\in D_T}|\sum_{y\in D_S}\pi^*(x,y)-\phi_T(x)|\}.
\end{aligned} \tag{11}$$

Considering $(**)$, we have by triangle inequality,

$$\begin{aligned}
&\epsilon_T(h^*,f_T)-\epsilon_T(h^*,f_S)\\
&=\mathop{\mathbb{E}}_{x\sim\mu_S}[|h^*(x)-f_T(x)|]-\mathop{\mathbb{E}}_{x\sim\mu_S}[|h^*(x)-f_S(x)|]\\
&\leq \mathop{\mathbb{E}}_{x\sim\mu_S}[|f_T(x)-f_S(x)|].
\end{aligned} \tag{12}$$

Plugging Equations (11) and (12) into Equation (6) yields

$$\begin{aligned}
&\epsilon_T(h^*\circ T)\leq \epsilon_T(h^*,f_S)\\
&+K\{\mathcal{W}(\mu_T,\mu_S)+M\sum_{x\in D_T}|\sum_{y\in D_S}\pi^*(x,y)-\phi_T(x)|\}\\
&+\mathop{\mathbb{E}}_{x\sim\mu_T}[|f_T(x)-f_S(x)|],
\end{aligned} \tag{13}$$

which completes the proof.

We note that when h^* is linear, Equation (7) can be rewritten in the following way:

$$\begin{aligned}
&\epsilon_T(h^*\circ T,f_T)-\epsilon_T(h^*,f_T)\\
&\leq \mathop{\mathbb{E}}_{x\sim\mu_T}[\|h^*\circ T(x)-h^*(x)\|]\\
&=\mathop{\mathbb{E}}_{x\sim\mu_T}[\|h^*(T(x)-(x)\|]\\
&\leq \mathop{\mathbb{E}}_{x\sim\mu_T}[|||h^*|||\|T(x)-x\|]\\
&=|||h^*|||\mathop{\mathbb{E}}_{x\sim\mu_T}[\|T(x)-x\|],
\end{aligned} \tag{14}$$

where $|||h^*||| = \sup_{\|x\|=1, x \in D_T} \|h^*(x)\|$.

In this case, the theorem can be rewritten as:

$$\underbrace{\epsilon_T(h')}_{\text{target error}} \leq \underbrace{\epsilon_T(h^*, f_S)}_{\text{source model error on the target domain}}$$
$$+ \underbrace{|||h^*|||\{\mathcal{W}_1(\mu_T, \mu_S) + M \sum_{x \in D_T} | \sum_{y \in D_S} \pi^*(x,y) - \phi_T(x)|\}}_{\text{transport}}$$
$$+ \underbrace{\mathbb{E}_{x \sim \mu_T}\{|f_T(x) - f_S(x)|\}}_{\text{labeling divergence}},$$

where $|||h^*||| = \sup_{\|x\|=1, x \in D_T} \|h^*(x)\|$.

□

B Supplementary Figures

B.1 Feature embedding differences

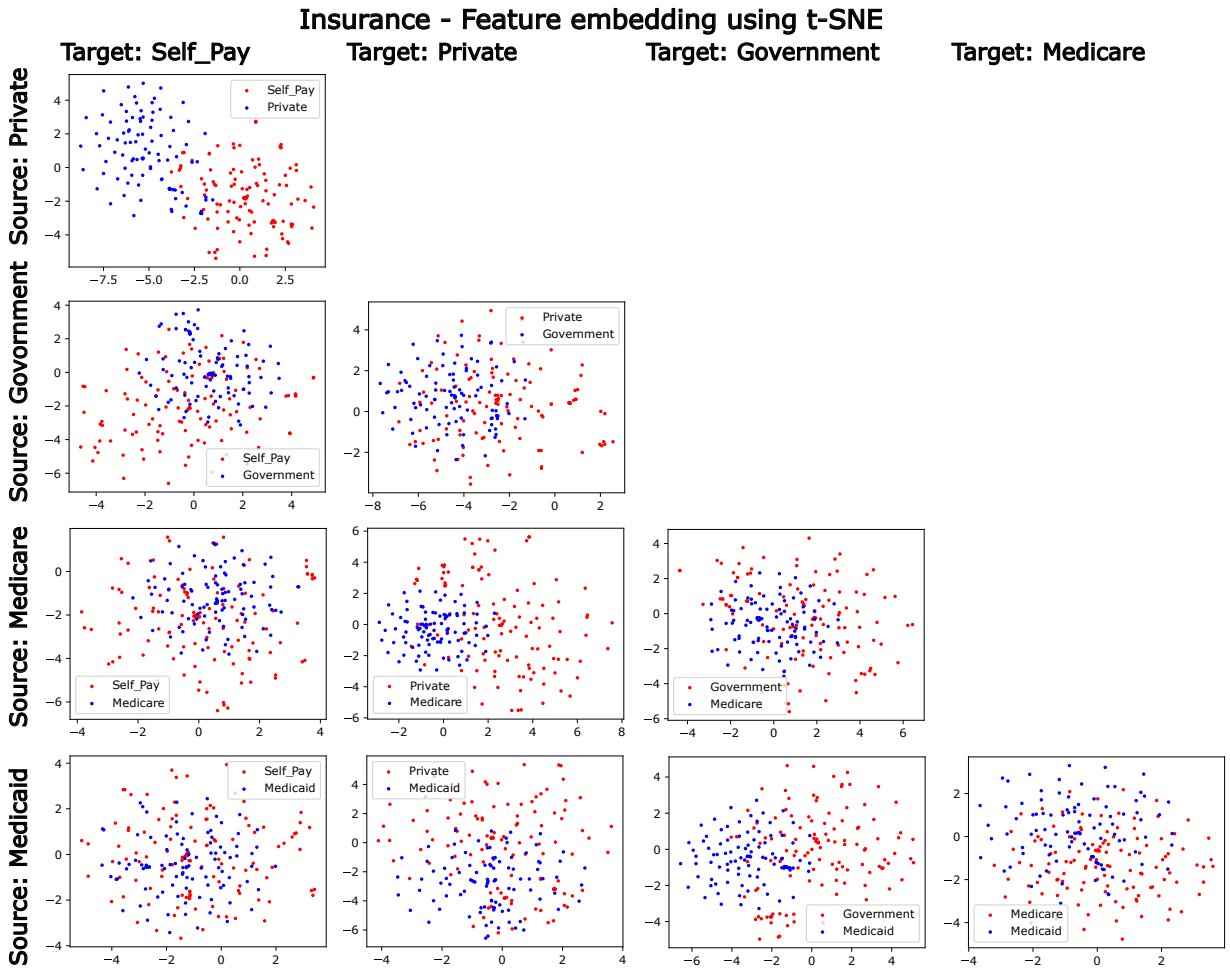

Figure S1: Feature embedding for pairwise marital status groups using t-SNE. Marital status groups include "Self_Pay," "Private," "Government," "Medicare," and "Medicaid." Most of the groups form their own clusters.

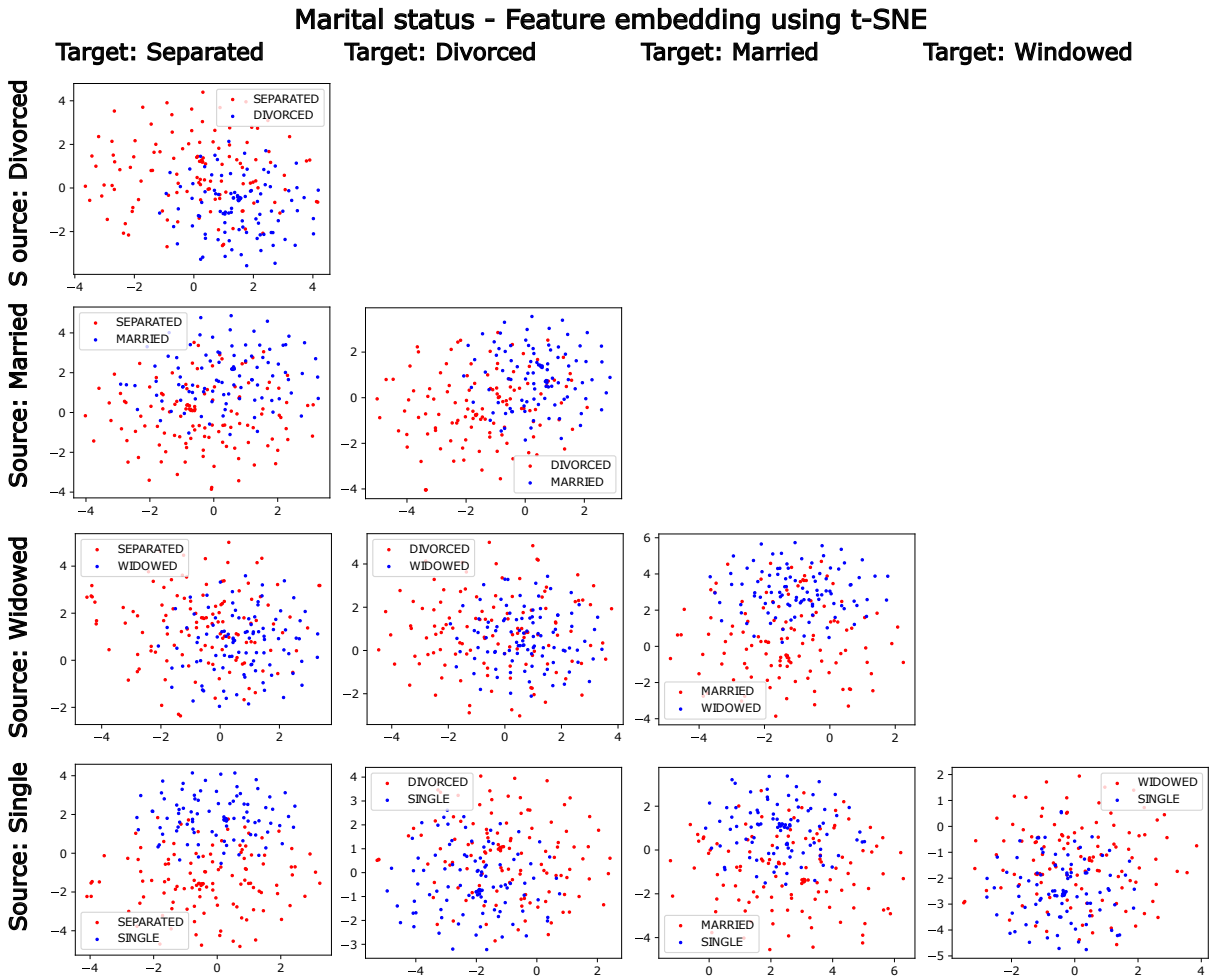

Figure S2: Feature embedding for pairwise marital status groups using t-SNE. Marital status groups include "Separated," "Divorced," "Married," "Widowed," and "Single." Most of the groups form their own clusters.

B.2 Illustration of $\epsilon_T(h^*, f_S)$ in Theorem 1

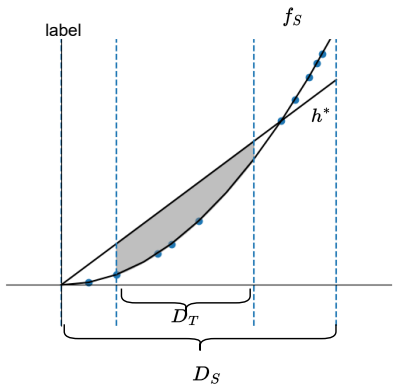

Figure S3: Illustration of $\epsilon_T(h^*, f_S)$. D_S and D_T are source and target embedding spaces. The blue dots denote the source embeddings and source labels. f_S is the ground-truth labeling function for source embedding features and source labels. h^* is the source model trained by source embedding features and source labels. The gray area denotes $\epsilon_T(h^*, f_S)$.

B.3 Bound analysis

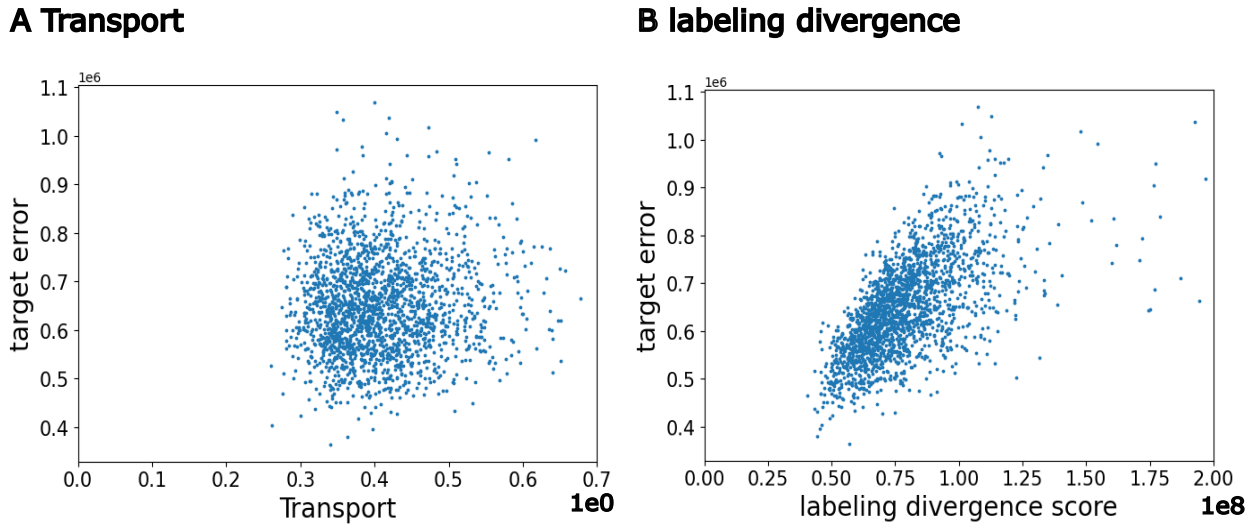

Figure S4: Bound analysis for all marital status group experiments. The relationship between target error and (**A**) transport term with a PCC of 0.63, and (**B**) labeling divergence with a PCC of 0.09, combining the results for all pairwise marital status group experiments.

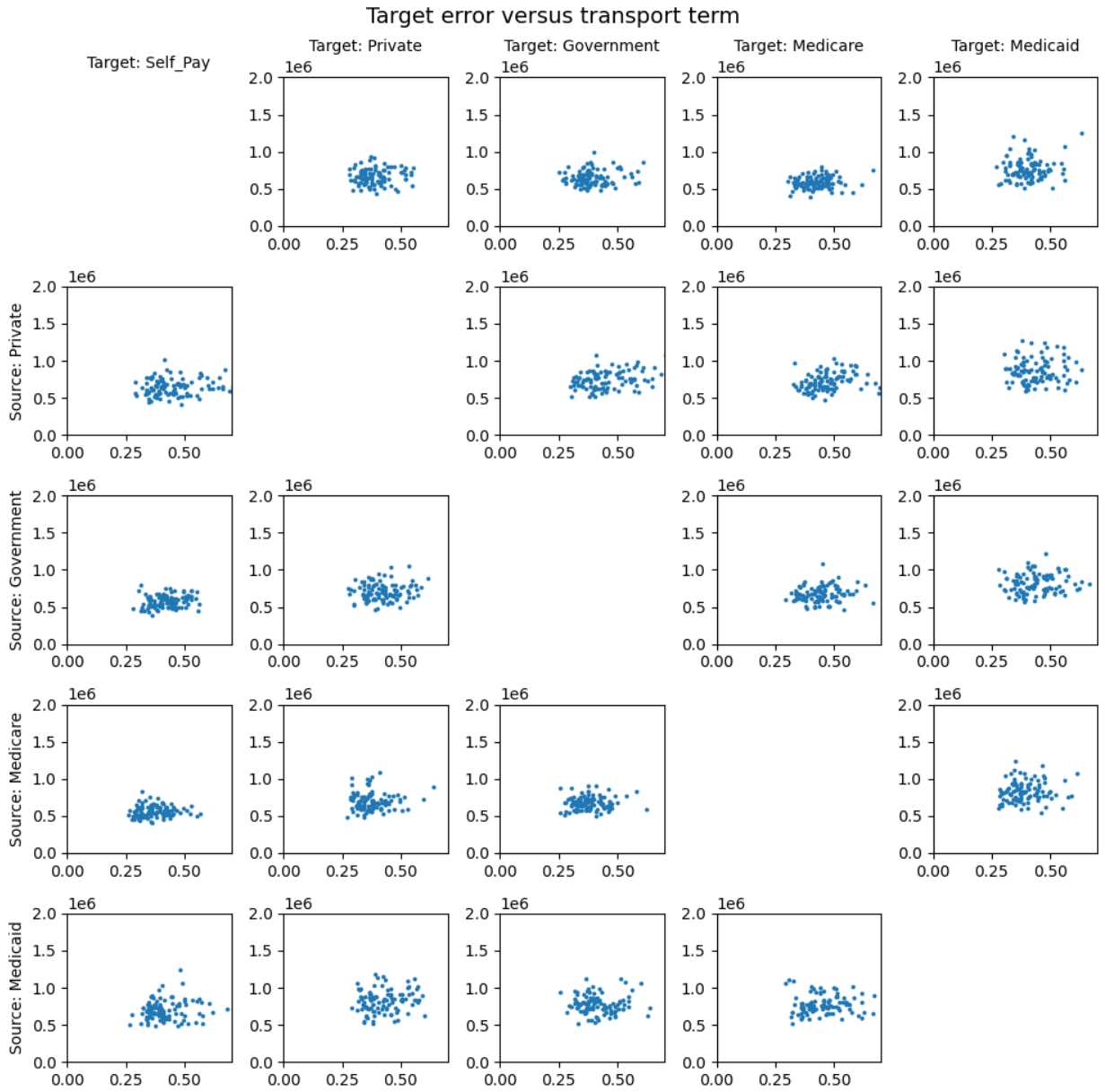

Figure S5: Bound analysis for pairwise insurance group experiments with respect to transport term. Target error versus transport term for pairwise insurance groups with an average PCC of 0.09.

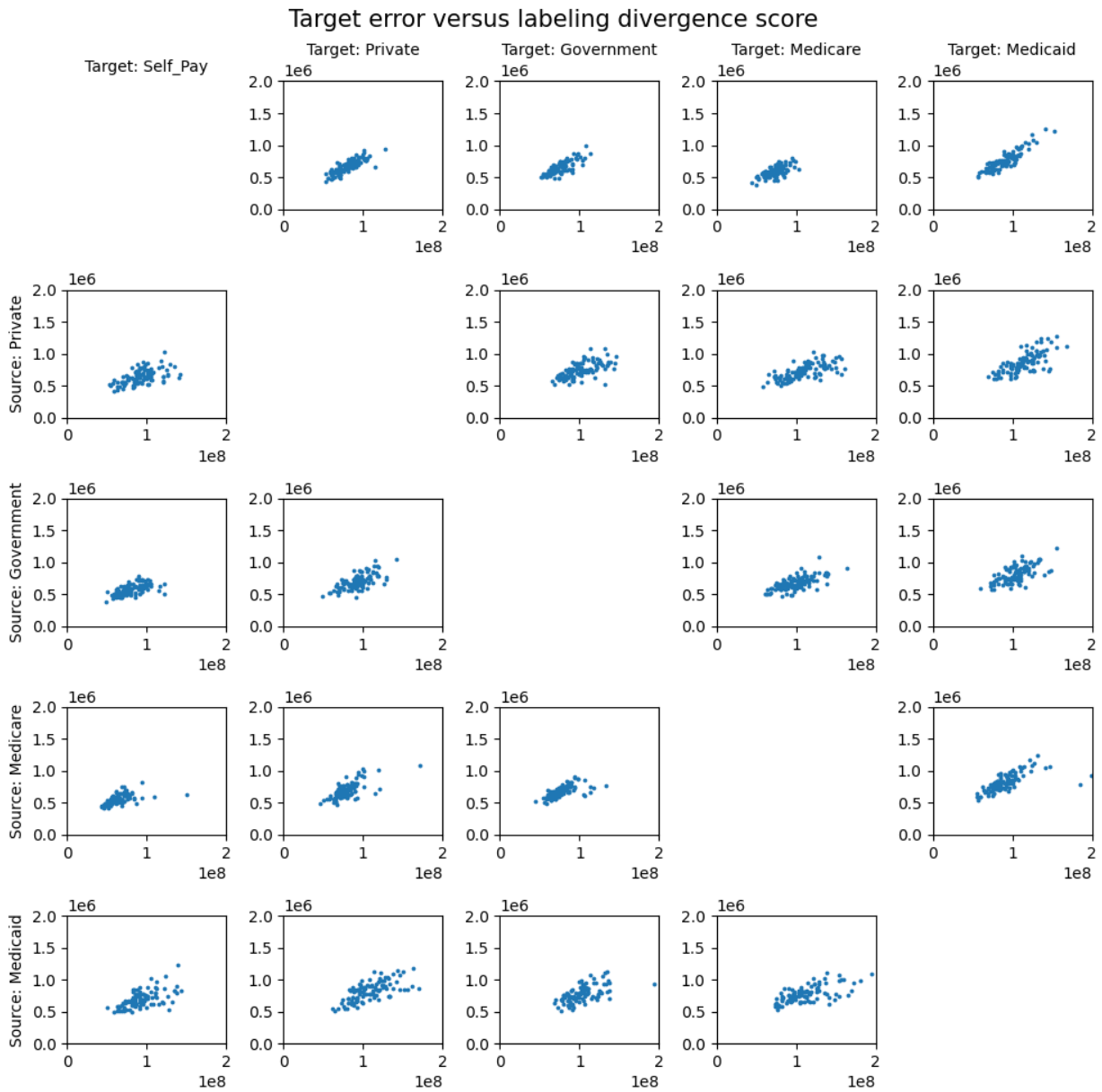

Figure S6: Bound analysis for pairwise insurance group experiments with respect to labeling divergence. Target error versus labeling divergence for pairwise insurance groups with an average PCC of 0.67.

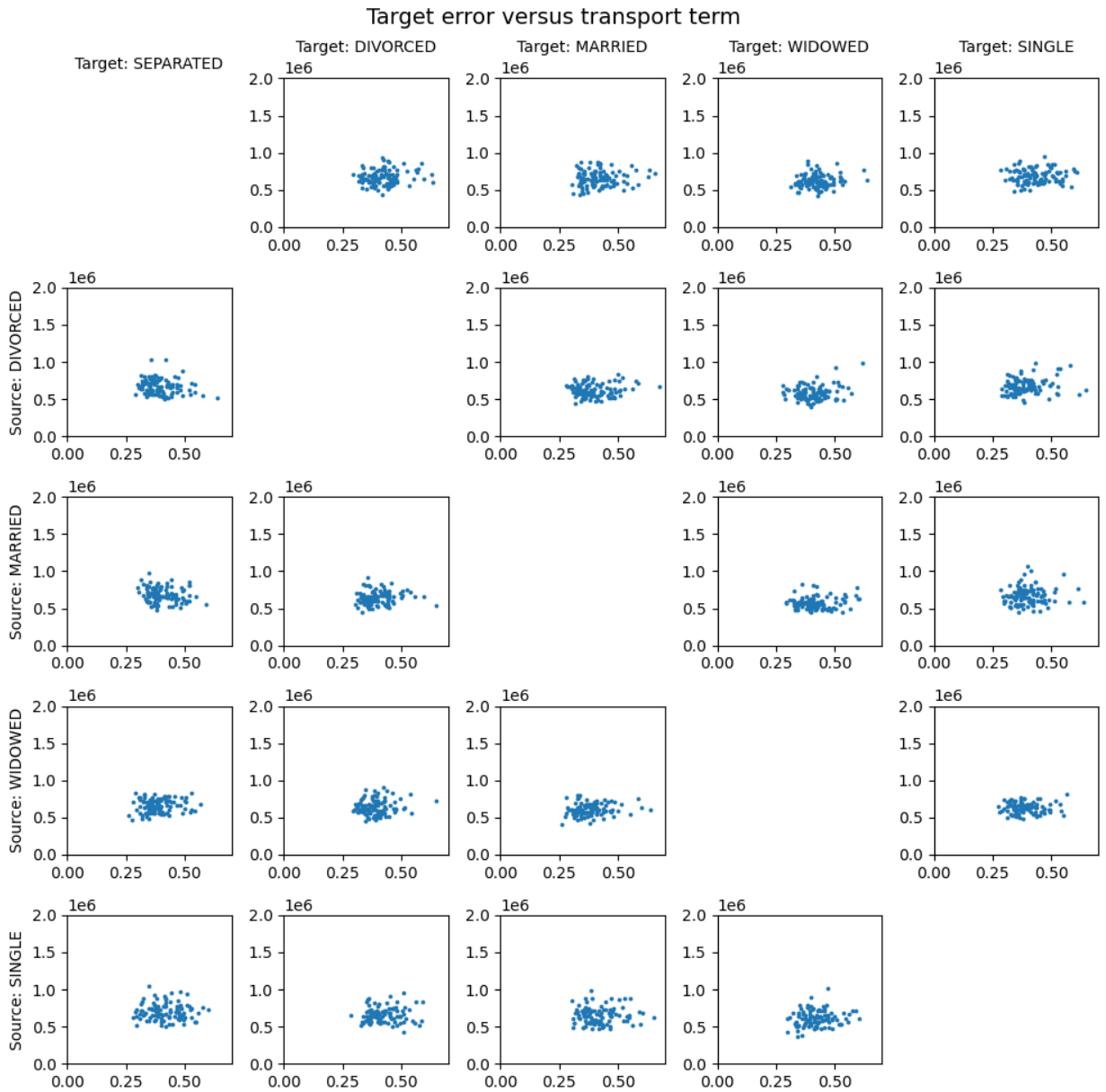

Figure S7: Bound analysis for pairwise marital status group experiments with respect to transport term. Target error versus transport term for pairwise marital status groups with an average PCC of 0.03.

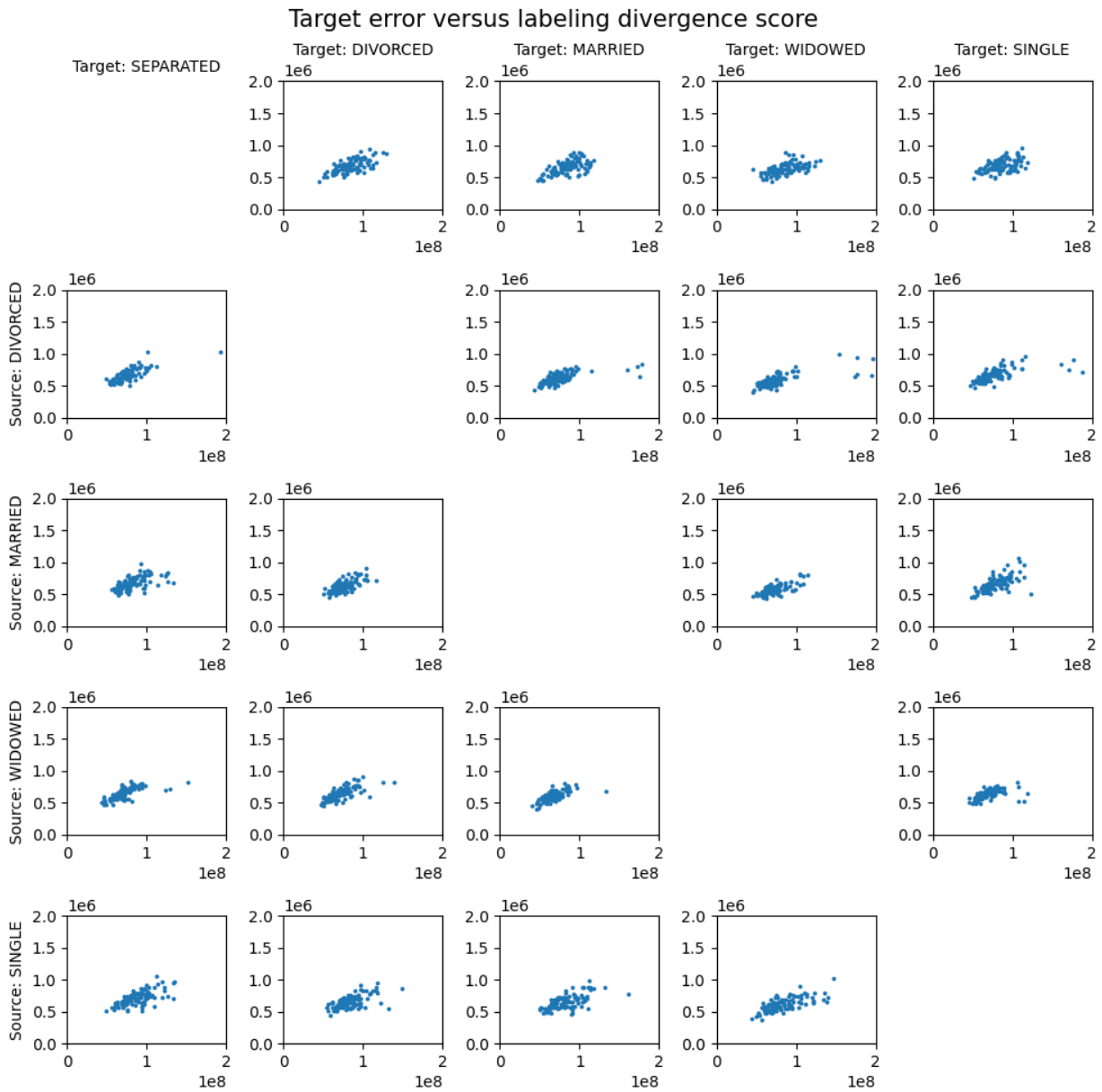

Figure S8: Bound analysis for pairwise marital status group experiments with respect to labeling divergence. Target error versus labeling divergence for pairwise marital status groups with an average PCC of 0.59.

B.4 Benchmark results for marital status group experiments

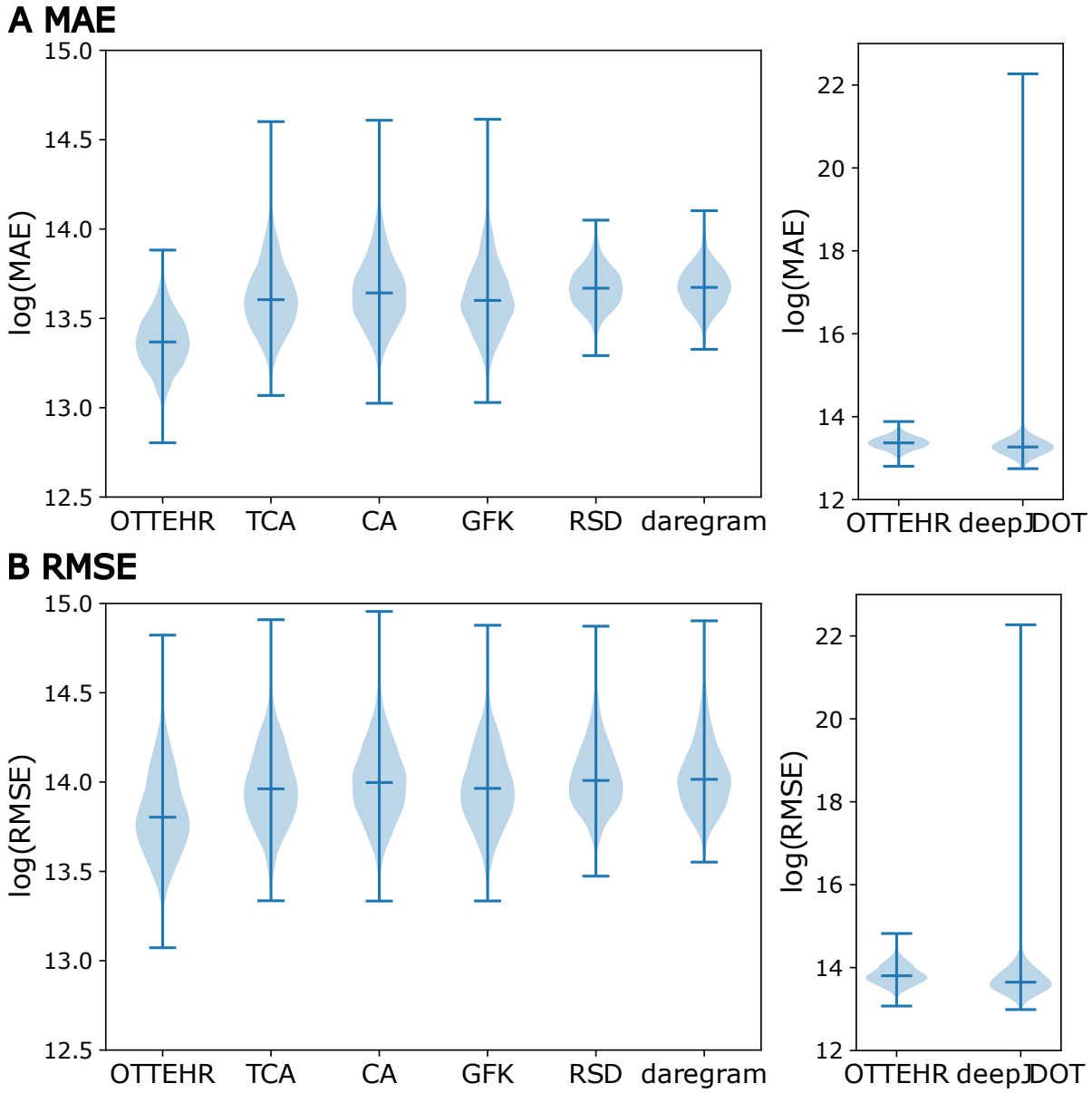

Figure S9: Benchmark results for all marital status group experiments. Violin plots of the log of (**A**) *MAE* and (**B**) *RMSE* between projected duration and observed duration on target admissions for marital status groups using *OTTEHR*, *TCA*, *CA*, *GFK*, *RSD* and *daregram*. The blue bars denote the medians. The heights of the violin plots denote the variability, positively correlated with the standard deviations. Smaller *MAE* and *RMSE* values indicate better performance.

B.5 Admissions with significantly reduced durations in hospital for all pairwase experiments

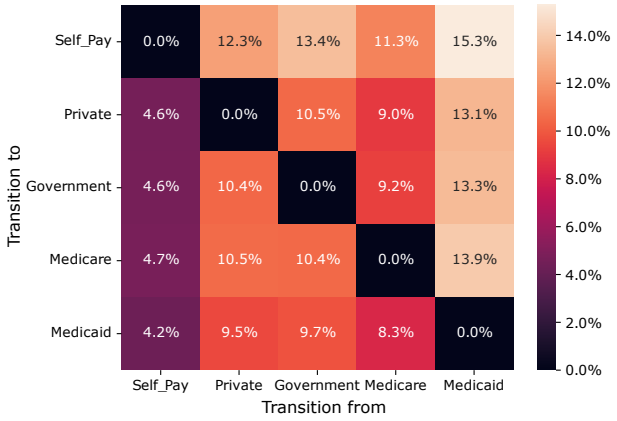

Figure S10: Percentages of admissions having significantly reduced durations in hospital for all pairwise insurance group experiments. For example, when transitioning to self paid insurance plans, 15.3% of the admissions on Medicaid would result in significantly reduced durations in hospital.

C Supplementary Tables

C.1 Benchmark results for pairwise group experiments

Table S1: Benchmark results for pairwise insurance group experiments. Medians and standard deviations of the log of *MAE* between projected duration and observed duration on target admissions for different insurance groups using *OTTEHR*, *TCA*, *CA*, *GFK*, *deepJDOT*, *RSD* and *daregram*. The outperformance ratio of *OTTEHR* to *TCA*/*CA*/*GFK*/*RSD*/*daregram* is defined as the percentage decrease in the median *MAE* from *TCA*/*CA*/*GFK* to *OTTEHR*. The outperformance ratio of *OTTEHR* to *deepJDOT* is defined as the percentage decrease in median of the log-transformed of standard deviation of *MAE* from *deepJDOT* to *OTTEHR*.

Source	Target	*OTTEHR*	*TCA*	*CA*	*GFK*	*deepJDOT*	*RSD*	*daregram*
Self_Pay	Private	13.40(0.16)	13.60(0.17)	13.63(0.18)	13.60(0.17)	13.32(0.18)	13.68(0.13)	13.66(0.14)
Self_Pay	Government	13.35(0.15)	13.55(0.16)	13.56(0.17)	13.56(0.16)	13.32(0.17)	13.64(0.12)	13.62(0.13)
Self_Pay	Medicare	13.28(0.14)	13.54(0.14)	13.58(0.15)	13.55(0.15)	13.13(0.47)	13.64(0.11)	13.67(0.10)
Self_Pay	Medicaid	13.53(0.18)	13.74(0.18)	13.73(0.18)	13.73(0.18)	13.46(0.29)	13.79(0.14)	13.78(0.12)
Private	Self_Pay	13.36(0.17)	13.78(0.19)	13.80(0.17)	13.78(0.20)	13.30(0.87)	13.21(0.13)	13.23(0.10)
Private	Government	13.52(0.16)	13.83(0.18)	13.86(0.19)	13.84(0.18)	13.44(0.31)	13.62(0.14)	13.61(0.11)
Private	Medicare	13.48(0.16)	13.95(0.21)	13.94(0.21)	13.94(0.21)	13.19(0.13)	13.63(0.10)	13.66(0.10)
Private	Medicaid	13.65(0.18)	13.98(0.17)	13.99(0.17)	13.98(0.17)	13.55(0.77)	13.82(0.13)	13.81(0.14)
Government	Self_Pay	13.28(0.14)	13.66(0.18)	13.69(0.18)	13.67(0.18)	13.25(0.12)	13.22(0.09)	13.25(0.12)
Government	Private	13.45(0.16)	13.78(0.17)	13.82(0.18)	13.78(0.17)	13.49(1.01)	13.66(0.14)	13.67(0.12)
Government	Medicare	13.42(0.15)	13.85(0.19)	13.84(0.19)	13.85(0.19)	13.17(0.80)	13.66(0.10)	13.66(0.09)
Government	Medicaid	13.62(0.16)	13.91(0.16)	13.93(0.17)	13.92(0.16)	13.56(0.55)	13.81(0.13)	13.82(0.14)
Medicare	Self_Pay	13.23(0.13)	13.44(0.18)	13.47(0.18)	13.44(0.18)	13.28(0.81)	13.22(0.12)	13.22(0.12)
Medicare	Private	13.44(0.17)	13.62(0.17)	13.65(0.18)	13.62(0.17)	13.45(1.06)	13.67(0.14)	13.68(0.12)
Medicare	Government	13.42(0.13)	13.58(0.17)	13.62(0.15)	13.57(0.17)	13.43(0.36)	13.64(0.13)	13.64(0.12)
Medicare	Medicaid	13.61(0.17)	13.74(0.20)	13.78(0.20)	13.74(0.20)	13.55(0.22)	13.81(0.11)	13.77(0.14)
Medicaid	Self_Pay	13.42(0.18)	13.77(0.19)	13.81(0.20)	13.78(0.19)	13.42(0.87)	13.23(0.12)	13.22(0.11)
Medicaid	Private	13.64(0.18)	13.88(0.20)	13.95(0.19)	13.87(0.20)	13.54(0.17)	13.66(0.16)	13.66(0.13)
Medicaid	Government	13.55(0.16)	13.87(0.17)	13.89(0.16)	13.87(0.18)	13.53(0.81)	13.63(0.13)	13.66(0.13)
Medicaid	Medicare	13.54(0.16)	13.92(0.21)	13.96(0.21)	13.92(0.21)	13.32(0.25)	13.66(0.10)	13.65(0.10)
Population Median		13.46(0.20)	13.75(0.23)	13.78(0.23)	13.75(0.23)	13.39(0.62)	13.64(0.24)	13.64(0.23)
Outperformance ratio of *OTTEHR*		N.A.	25.17	27.39	25.17	67.74	16.47	16.47

Table S2: Benchmark results for pairwise insurance group experiments. Medians and standard deviations of the log of *RMSE* between projected duration and observed duration on target admissions for different insurance groups using *OTTEHR*, *TCA*, *CA*, *GFK*, *deepJDOT*, *RSD* and *daregram*. The outperformance ratio of *OTTEHR* to *TCA*/*CA*/*GFK*/*RSD*/*daregram* is defined as the percentage decrease in the median *RMSE* from *TCA*/*CA*/*GFK* to *OTTEHR*. The outperformance ratio of *OTTEHR* to *deepJDOT* is defined as the percentage decrease in median of the log-transformed of standard deviation of *RMSE* from *deepJDOT* to *OTTEHR*.

Source	Target	*OTTEHR*	*TCA*	*CA*	*GFK*	*deepJDOT*	*RSD*	*daregram*
Self_Pay	Private	14.03(0.23)	14.11(0.22)	14.15(0.23)	14.09(0.22)	13.99(0.25)	14.19(0.19)	14.17(0.21)
Self_Pay	Government	13.95(0.20)	14.02(0.19)	14.03(0.19)	14.03(0.19)	13.90(0.23)	14.11(0.17)	14.11(0.18)
Self_Pay	Medicare	13.71(0.20)	13.90(0.17)	13.89(0.17)	13.89(0.17)	13.64(0.45)	13.93(0.19)	13.97(0.16)
Self_Pay	Medicaid	14.15(0.25)	14.24(0.23)	14.24(0.22)	14.24(0.23)	14.16(0.30)	14.30(0.20)	14.29(0.17)
Private	Self_Pay	13.85(0.23)	14.08(0.21)	14.05(0.18)	14.07(0.21)	13.53(0.86)	13.63(0.19)	13.69(0.16)
Private	Government	14.05(0.23)	14.22(0.20)	14.22(0.21)	14.23(0.19)	13.82(0.31)	14.09(0.20)	14.08(0.17)
Private	Medicare	13.95(0.21)	14.27(0.21)	14.24(0.21)	14.25(0.21)	13.53(0.22)	13.95(0.14)	13.97(0.14)
Private	Medicaid	14.24(0.24)	14.38(0.20)	14.37(0.20)	14.37(0.19)	14.15(0.74)	14.33(0.19)	14.32(0.21)
Government	Self_Pay	13.72(0.22)	13.94(0.19)	13.97(0.18)	13.96(0.20)	13.49(0.17)	13.67(0.16)	13.66(0.17)
Government	Private	14.05(0.23)	14.18(0.20)	14.24(0.21)	14.19(0.19)	14.01(0.96)	14.15(0.21)	14.19(0.19)
Government	Medicare	13.88(0.22)	14.16(0.19)	14.16(0.19)	14.16(0.19)	13.58(0.78)	13.97(0.15)	13.96(0.18)
Government	Medicaid	14.18(0.21)	14.28(0.18)	14.33(0.18)	14.30(0.18)	14.13(0.54)	14.33(0.21)	14.35(0.20)
Medicare	Self_Pay	13.56(0.18)	13.71(0.20)	13.71(0.19)	13.70(0.20)	13.49(0.79)	13.67(0.19)	13.68(0.19)
Medicare	Private	13.98(0.24)	14.06(0.21)	14.09(0.22)	14.05(0.21)	13.97(0.99)	14.18(0.22)	14.17(0.19)
Medicare	Government	13.88(0.20)	13.96(0.20)	14.01(0.18)	13.96(0.20)	13.91(0.36)	14.10(0.17)	14.08(0.18)
Medicare	Medicaid	14.18(0.24)	14.21(0.24)	14.25(0.24)	14.22(0.24)	14.12(0.29)	14.33(0.17)	14.29(0.21)
Medicaid	Self_Pay	13.84(0.22)	14.07(0.19)	14.07(0.20)	14.06(0.19)	13.60(0.86)	13.71(0.17)	13.66(0.18)
Medicaid	Private	14.15(0.23)	14.26(0.21)	14.29(0.20)	14.25(0.21)	14.01(0.24)	14.18(0.23)	14.14(0.19)
Medicaid	Government	14.02(0.20)	14.18(0.18)	14.22(0.16)	14.18(0.18)	13.89(0.78)	14.10(0.18)	14.14(0.18)
Medicaid	Medicare	13.97(0.23)	14.22(0.21)	14.23(0.22)	14.21(0.21)	13.65(0.28)	13.95(0.18)	13.97(0.16)
Population Median		13.96(0.28)	14.12(0.25)	14.14(0.25)	14.12(0.25)	13.80(0.63)	14.05(0.29)	14.05(0.28)
Outperformance ratio of *OTTEHR*		N.A.	14.79	16.47	14.79	55.56	8.61	8.61

Table S3: Benchmark results for pairwise marital status group experiments. Medians and standard deviations of the log of *MAE* between projected duration and observed duration on target admissions for different marital status groups using *OTTEHR*, *TCA*, *CA*, *GFK*, *deepJDOT*, *RSD* and *daregram*. The outperformance ratio of *OTTEHR* to *TCA*/*CA*/*GFK*/*RSD*/*daregram* is defined as the percentage decrease in the median *MAE* from *TCA*/*CA*/*GFK* to *OTTEHR*. The outperformance ratio of *OTTEHR* to *deepJDOT* is defined as the percentage decrease in median of the log-transformed of standard deviation of *MAE* from *deepJDOT* to *OTTEHR*.

Source	Target	*OTTEHR*	*TCA*	*CA*	*GFK*	*deepJDOT*	*RSD*	*daregram*
Separated	Divorced	13.41(0.14)	13.69(0.19)	13.74(0.17)	13.70(0.19)	13.30(0.13)	13.73(0.12)	13.70(0.10)
Separated	Married	13.40(0.16)	13.71(0.19)	13.74(0.18)	13.71(0.19)	13.30(0.35)	13.64(0.11)	13.66(0.10)
Separated	Widowed	13.33(0.14)	13.71(0.18)	13.72(0.18)	13.71(0.18)	13.21(0.11)	13.59(0.09)	13.60(0.10)
Separated	Single	13.45(0.14)	13.72(0.18)	13.79(0.17)	13.72(0.18)	13.39(0.12)	13.67(0.12)	13.67(0.12)
Divorced	Separated	13.42(0.15)	13.58(0.16)	13.64(0.16)	13.58(0.16)	13.35(0.52)	13.74(0.09)	13.75(0.10)
Divorced	Married	13.33(0.14)	13.56(0.23)	13.60(0.23)	13.55(0.23)	13.24(0.28)	13.65(0.10)	13.66(0.11)
Divorced	Widowed	13.28(0.16)	13.55(0.26)	13.54(0.26)	13.54(0.26)	13.15(0.93)	13.62(0.10)	13.60(0.09)
Divorced	Single	13.40(0.15)	13.59(0.24)	13.64(0.24)	13.58(0.24)	13.33(1.20)	13.66(0.10)	13.67(0.12)
Married	Separated	13.42(0.14)	13.65(0.17)	13.69(0.14)	13.65(0.16)	13.32(0.13)	13.73(0.11)	13.74(0.10)
Married	Divorced	13.36(0.13)	13.59(0.14)	13.61(0.14)	13.59(0.14)	13.26(0.90)	13.67(0.11)	13.66(0.11)
Married	Widowed	13.25(0.13)	13.53(0.16)	13.56(0.16)	13.53(0.17)	13.12(0.92)	13.61(0.08)	13.61(0.09)
Married	Single	13.40(0.17)	13.63(0.17)	13.66(0.17)	13.62(0.17)	13.37(1.00)	13.64(0.12)	13.69(0.12)
Widowed	Separated	13.39(0.14)	13.54(0.18)	13.55(0.17)	13.55(0.17)	13.31(0.58)	13.73(0.10)	13.73(0.10)
Widowed	Divorced	13.34(0.15)	13.50(0.19)	13.55(0.19)	13.51(0.19)	13.24(0.50)	13.69(0.10)	13.67(0.11)
Widowed	Married	13.30(0.13)	13.46(0.15)	13.49(0.14)	13.45(0.15)	13.21(0.14)	13.64(0.11)	13.68(0.11)
Widowed	Single	13.37(0.11)	13.52(0.16)	13.56(0.18)	13.52(0.17)	13.31(0.86)	13.67(0.12)	13.67(0.11)
Single	Separated	13.46(0.15)	13.74(0.19)	13.76(0.19)	13.72(0.19)	13.32(0.22)	13.73(0.10)	13.73(0.12)
Single	Divorced	13.37(0.15)	13.69(0.18)	13.73(0.20)	13.69(0.18)	13.26(0.13)	13.70(0.10)	13.70(0.11)
Single	Married	13.38(0.17)	13.64(0.21)	13.68(0.21)	13.64(0.21)	13.23(0.57)	13.66(0.10)	13.67(0.11)
Single	Widowed	13.31(0.17)	13.65(0.23)	13.65(0.23)	13.64(0.24)	13.15(0.14)	13.59(0.10)	13.62(0.11)
Population Median		13.37(0.15)	13.60(0.20)	13.64(0.20)	13.60(0.20)	13.27(0.61)	13.67(0.11)	13.67(0.11)
Outperformance ratio of *OTTEHR*		N.A.	20.55	23.66	20.55	75.41	25.92	25.92

Table S4: Benchmark results for pairwise marital status group experiments. Medians and standard deviations of the log of *RMSE* between projected duration and observed duration on target admissions for different marital status groups using *OTTEHR*, *TCA*, *CA*, *GFK*, *deepJDOT*, *RSD* and *daregram*. The outperformance ratio of *OTTEHR* to *TCA*/*CA*/*GFK*/*RSD*/*daregram* is defined as the percentage decrease in the median *RMSE* from *TCA*/*CA*/*GFK* to *OTTEHR*. The outperformance ratio of *OTTEHR* to *deepJDOT* is defined as the percentage decrease in median of the log-transformed of standard deviation of *RMSE* from *deepJDOT* to *OTTEHR*.

Source	Target	*OTTEHR*	*TCA*	*CA*	*GFK*	*deepJDOT*	*RSD*	*daregram*
Separated	Divorced	13.81(0.26)	14.00(0.25)	14.05(0.23)	14.01(0.24)	13.64(0.25)	14.06(0.25)	14.02(0.20)
Separated	Married	13.83(0.24)	14.01(0.21)	14.07(0.20)	14.01(0.22)	13.63(0.36)	13.97(0.17)	13.99(0.17)
Separated	Widowed	13.77(0.24)	14.00(0.21)	14.03(0.21)	14.01(0.21)	13.52(0.19)	13.89(0.16)	13.89(0.16)
Separated	Single	13.88(0.21)	14.07(0.18)	14.10(0.19)	14.07(0.18)	13.78(0.23)	14.06(0.18)	14.04(0.21)
Divorced	Separated	13.85(0.24)	13.97(0.21)	14.01(0.21)	13.97(0.21)	13.91(0.51)	14.10(0.17)	14.17(0.18)
Divorced	Married	13.75(0.21)	13.89(0.25)	13.93(0.24)	13.88(0.24)	13.64(0.31)	14.01(0.17)	13.99(0.18)
Divorced	Widowed	13.66(0.27)	13.86(0.29)	13.87(0.28)	13.85(0.29)	13.47(0.91)	13.92(0.16)	13.89(0.16)
Divorced	Single	13.83(0.22)	13.94(0.27)	13.97(0.25)	13.92(0.26)	13.70(1.17)	14.07(0.17)	14.05(0.21)
Married	Separated	13.92(0.24)	14.03(0.21)	14.11(0.19)	14.05(0.20)	13.80(0.25)	14.12(0.19)	14.15(0.16)
Married	Divorced	13.80(0.27)	13.92(0.24)	13.95(0.24)	13.92(0.24)	13.65(0.91)	14.00(0.18)	13.99(0.24)
Married	Widowed	13.67(0.18)	13.85(0.17)	13.86(0.17)	13.85(0.17)	13.44(0.90)	13.90(0.16)	13.90(0.15)
Married	Single	13.85(0.25)	14.01(0.23)	14.02(0.22)	13.99(0.22)	13.80(0.95)	14.03(0.20)	14.06(0.20)
Widowed	Separated	13.95(0.24)	14.01(0.23)	14.02(0.23)	14.02(0.22)	13.84(0.57)	14.14(0.17)	14.08(0.17)
Widowed	Divorced	13.79(0.26)	13.89(0.25)	13.90(0.26)	13.87(0.25)	13.66(0.51)	13.99(0.19)	13.99(0.20)
Widowed	Married	13.75(0.22)	13.82(0.21)	13.87(0.20)	13.82(0.21)	13.61(0.25)	13.99(0.19)	14.01(0.19)
Widowed	Single	13.79(0.20)	13.89(0.19)	13.93(0.20)	13.88(0.20)	13.74(0.84)	14.07(0.19)	14.06(0.17)
Single	Separated	13.99(0.23)	14.12(0.20)	14.16(0.20)	14.12(0.20)	13.83(0.29)	14.18(0.19)	14.14(0.20)
Single	Divorced	13.82(0.23)	14.01(0.20)	14.06(0.22)	14.01(0.20)	13.63(0.27)	14.04(0.18)	14.04(0.21)
Single	Married	13.85(0.26)	14.02(0.23)	14.06(0.23)	14.01(0.23)	13.63(0.56)	14.01(0.16)	14.00(0.18)
Single	Widowed	13.71(0.24)	13.94(0.25)	13.96(0.25)	13.94(0.25)	13.47(0.25)	13.89(0.15)	13.91(0.18)
Population Median		13.80(0.25)	13.96(0.24)	14.00(0.23)	13.96(0.23)	13.65(0.62)	14.01(0.19)	14.02(0.20)
Outperformance ratio of *OTTEHR*		N.A.	14.79	18.13	14.79	59.68	18.94	19.75

C.2 Computation time

Table S5: Average computational time in seconds per experiment for *OTTEHR*, *TCA*, *CA*, *GFKdeepJDOT*, *RSD* and *daregram*.

Group	*OTTEHR*	*TCA*	*CA*	*GFK*	*deepJDOT*	*RSD*	*daregram*
Insurance	333.6	418.2	11387.2	453.5	17350.0	281.2	430.2
Marital status	477.2	561.2	8729.5	595.0	17292.6	251.0	436.9
Total	810.8	979.4	20116.7	1048.5	34642.6	532.2	436.9