Journal of Machine Learning Research 25 (2024) 1-40 Submitted 11/22; Revised 9/24; Published 10/24

Transfer Learning with Uncertainty Quantification: Random Effect Calibration of Source to Target (RECaST)

Jimmy Hickey JHICKEY@NCSU.EDU
Department of Statistics
North Carolina State University

Jonathan P. Williams JWILLI27@NCSU.EDU
Department of Statistics, North Carolina State University
Centre for Advanced Study, Norwegian Academy of Science and Letters

Emily C. Hector EHECTOR@NCSU.EDU
Department of Statistics
North Carolina State University

Editor: Amos Storkey

Abstract

Transfer learning uses a data model, trained to make predictions or inferences on data from one population, to make reliable predictions or inferences on data from another population. Most existing transfer learning approaches are based on fine-tuning pre-trained neural network models, and fail to provide crucial uncertainty quantification. We develop a statistical framework for model predictions based on transfer learning, called *RECaST*. The primary mechanism is a Cauchy random effect that recalibrates a source model to a target population; we mathematically and empirically demonstrate the validity of our RECaST approach for transfer learning between linear models, in the sense that prediction sets will achieve their nominal stated coverage, and we numerically illustrate the method's robustness to asymptotic approximations for nonlinear models. Whereas many existing techniques are built on particular source models, RECaST is agnostic to the choice of source model, and does not require access to source data. For example, our RECaST transfer learning approach can be applied to a continuous or discrete data model with linear or logistic regression, deep neural network architectures, etc. Furthermore, RECaST provides uncertainty quantification for predictions, which is mostly absent in the literature. We examine our method's performance in a simulation study and in an application to real hospital data.

Keywords: Bayesian transfer learning, Electronic health records, Informative Bayesian prior, Model calibration.

1. Introduction

The use of artificial intelligence and machine learning (ML) is frequently limited in practice by a shortage of available training data and insufficient computational resources. To address these difficulties, transfer learning has developed as a powerful idea for leveraging the resources at leading institutions such as research hospitals (e.g., institutions having high quality data, exceptional research clinicians, high performance computing environments, etc.) to facilitate implementation of ML technologies in resource scarce settings such as small or rural hospitals. Developments in transfer learning methodologies are necessary to overcome resource allocation inequities, and they will likely drive the next decade of innovation in ML technologies.

Transfer learning consists broadly of two elements. The first is one or more *target* population(s) of interest that are associated with data sets for which there are resource limitations preventing the training of sophisticated models (e.g., a small hospital). The second is a *source* population (or populations) that is separate but in some way related to the target population. The source is associated with extensive data and/or resources for training sophisticated ML models. The premise of transfer learning is to use trained source models to aid in the training of target models. The source and targets are each composed of two components: a *domain*, denoted $\mathcal{D}$, and a *task*, denoted $\mathcal{T}$. A domain $\mathcal{D} := \{\mathcal{X}, P(x)\}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(x)$ over $x \in \mathcal{X}$. A task $\mathcal{T} := \{\mathcal{Y}, P(y \mid x)\}$ is composed of a label space $\mathcal{Y}$ and a conditional distribution $P(y \mid x)$ over $y \in \mathcal{Y}$ given $x \in \mathcal{X}$. Traditional ML is described by the source and target sharing the same domain, $\mathcal{D}_S = \mathcal{D}_T$, and sharing the same task, $\mathcal{T}_S = \mathcal{T}_T$. Transfer learning problems arise when the source and target domains and/or the source and target tasks are similar but different. We propose a new Bayesian transfer learning framework termed Random Effect Calibration of Source to Target (RECaST) for source and target data sets that share the same outcome space but possibly have different feature-to-outcome mappings.

1.1 Our Contributions

Early efforts in transfer learning focused on using labeled data to learn about unlabeled data from the *same* population (see Joachims (1999) and Vapnik (2009) for examples). In contrast, modern transfer learning methods explore how knowledge from one source domain can be applied to a *different* target domain. In this spirit, we consider transfer learning in the supervised learning problem that dominates ML applications. Our proposed method uses information from the source and target features and labels to build a predictive model that can be applied to obtain predictions of labels for new target data features of interest. The use of target labels is common across transfer learning and is sometimes referred to as *inductive* transfer learning (Pan and Yang, 2010). For example, a method is proposed in Donahue et al. (2014) to generalize a model built on ImageNet data for use on different labeled target data sets. A neural network is fine-tuned in Shao et al. (2019) to identify and classify machine faults. In Goussies et al. (2014), a decision forest is proposed that uses mixed information gain and label propagation to improve image and gesture recognition in the target domain.

RECaST is a Bayesian framework applied to the transfer learning setting where the feature-to-outcome mappings $P(y \mid x)$ may differ between the source and target. For example, source and target hospitals might record largely the same patient data features, but nuances in clinician practices/procedures, inconsistencies in data quality, population disparities, etc. may affect the suitability of using the source mapping as the target mapping. RECaST uses an estimated source model in tandem with the target data to estimate the distributions of a random effect that links the two domains. It then uses the estimated posterior distribution of the random effect parent parameters to construct a posterior predictive distribution of the outcome variable associated with a new target feature. The posterior predictive credible sets obtained through RECaST deliver critical quantification of prediction uncertainty that is lacking in most existing frameworks.

Two primary advantages of RECaST are its scalability, requiring estimation of only 2-3 parameters with no tuning parameters, and that it is agnostic to the source model specification. Importantly, RECaST only requires the source model and parameter estimates, not the source data itself; this is an immense benefit to applications with privacy concerns, such as with medical data. Further, we show that RECaST is asymptotically valid in the canonical case of distinct source and target Gaussian linear models, in that the coverage of prediction sets are guaranteed to asymptotically achieve their stated nominal level of significance.

To evaluate our proposed RECaST approach, we design synthetic simulation studies with both continuous and binary response data reflecting a variety of difficulty levels of transfer learning problems. Next, we investigate the performance of RECaST in real data simulations that arise by permuting real patient data from the multi-center eICU Collaborative Research Database (Pollard et al., 2018). A variety of both point-valued and set-valued prediction metrics are considered, including the empirical coverage of prediction sets. The performance of RECaST is compared to other state-of-the-art transfer learning approaches, including other source-free methods that do not require the source data while learning the target model. These include freeze-unfreeze approaches that are popular for neural networks, as well as a method based on adapting a random forest built on the source data to the target data (Gu et al., 2022). In some cases, it may be possible to have access to both the source and target data. As such we also compare RECaST to methods that require both data sets during training. These include an adversarial learning method (Shen et al., 2018), a method based on penalized GLMs (Tian and Feng, 2022), and a popular weighting approach used on clinical data (Wiens et al., 2014).

The remainder of our paper is organized as follows. We discuss related works in transfer learning in Section 2. In Section 3, we develop the theoretical basis for RECaST and its uncertainty quantification. We then develop Bayesian parameter estimation and prediction procedures in both the continuous and binary response cases in Sections 4 and 5, respectively. We conduct extensive simulation studies in Section 6 by exploring transfer learning problems of a range of difficulties. Section 7 considers a real data analysis for predicting shock in ICU data. Section 8 concludes. Proofs and computational details are provided in the Appendix. Throughout the paper we keep to the convention in the statistical literature of using $(\cdot)$ for innermost grouping followed by $\{\cdot\}$ and finally $[\cdot]$. Thus, an expression with many nested parentheses respects the ordering $[\{(\cdot)\}]$.

2. Related Work

General survey papers on transfer learning topics include Pan and Yang (2010); Lu et al. (2015); Weiss et al. (2016); Dube et al. (2020). For hospital disease risk and mortality prediction problems, Wiens et al. (2014), Gong et al. (2015), and Desautels et al. (2017) propose transfer learning approaches based on training algorithms using a learned weighted combination of source and target patient observations. These methods learn many parameters and require access to the source data. RECaST may at first glance appear similar to density ratio estimation, a common approach to transfer learning. Density ratio transfer learning methods, such as the one described in Stojanov et al. (2019), seek to learn the relationship between the source and target data via a ratio of their densities; however, these methods require joint access to the source and target data – a limitation avoided by RECaST. In Paul et al. (2016), Raghu et al. (2019), and Ahishakiye et al. (2021), approaches are considered to improve classification accuracy for medical imaging tasks using pre-trained deep neural networks (DNNs) on the ImageNet database (Deng et al., 2009). In the context of ICU patient monitoring, in Shickel et al. (2021) a data augmenting-based transfer learning approach is built for fitting a single-layer recurrent neural network trained on electronic health records (EHR) and wearable device data. Their model is limited in scope to only predicting the binary response of successful versus unsuccessful discharge from a hospital. Implemented in Gao and Cui (2021) is a transfer learning strategy for precision medicine in survival analysis with clinical omics data sets via freezing layers of a pre-trained Cox neural network. Developed in Lee et al. (2012) is a method using support vector machines to predict surgical mortality. Another approach, from Gu et al. (2023), is to generate additional synthetic target data from a source data set and adjust for heterogeneity in order to predict extreme obesity from medical records and genomics data. An example of low-dimensional representation transfer learning is given in Maurer et al. (2015), and *online* transfer learning is considered in Zhao et al. (2014); Wu et al. (2017). These applied methods are useful in modeling specific pieces of EHR data for prediction, but lack uncertainty quantification. Additionally, some require the learning of many parameters and access to the entire source data set.

Bayesian transfer learning adaptations include Baxter (1998), Raina et al. (2006), Wohlert et al. (2018), Bueno et al. (2020), Chandra and Kapoor (2020), Yang et al. (2020), Zhou et al. (2020), Abba et al. (2023); all except Baxter (1998) and Raina et al. (2006) are based on priors specified from neural network models fitted to source data sets. A posterior distribution fitted to a source DNN model is used as a prior on the parameters for the target task in Wohlert et al. (2018), and the model is trained using mean field variational Bayes (for a reference on variational Bayes, see Zhang et al., 2017). Boosting approaches to transfer learning are considered by Freund and Schapire (1999), Dai et al. (2007), and Desautels et al. (2017). In Abba et al. (2023), a penalized complexity prior between the source and target tasks is considered. While uncertainty quantification for predictions in transfer learning applications is mostly absent in the literature, approximate inference from Bayesian neural networks is used in Roy et al. (2022) to quantify uncertainty in parameter estimates and predictions to account for misaligned feature distributions. This approach, referred to as U-SFAN, is related to our RECaST framework in that it is source-free, but it requires that the source model is a neural network. Another difference is that U-SFAN focuses on using uncertainty in the source domain to guide uncertainty quantification in the target model, whereas RECaST provides uncertainty quantification directly based on the target predictions themselves.

It is important to note the difference between source-free transfer learning methods and “source-free domain adaption” (SFDA) methods: RECaST aims to use labels in the target domain in tandem with a model built in the source domain to learn about the target domain. SFDA methods, in contrast, have neither access to source data nor target labels, and often proceed by learning pseudo-labels for the target data. A comprehensive survey of SFDA approaches is given in Li et al. (2024). These surveyed strategies are predominantly non-model based, purely empirical, and lack a unified underlying framework. Moreover, those that focus on fine-tuning pre-trained neural network models on a target data set require the source model to be a neural network, and often fail to provide crucial uncertainty quantification.

Recently, there have been efforts to investigate theoretical properties of transfer learning approaches. For instance, a learning method based on LASSO for high-dimensional penalized linear regression is considered in Li et al. (2022), while diminishing the effect of *negative transfer*. Negative transfer occurs when including source data negatively impacts the performance on target data. In a similar setting, asymptotically valid confidence intervals for generalized linear model (GLM) parameters in high-dimensional transfer learning problems are established in Tian and Feng (2022). This technique is adapted to a more complicated federated transfer learning setting in Li et al. (2023). A parameter is defined in Cai and Wei (2021) to calculate an "effective sample size" to quantify total amount of information that can be transferred when the source and target conditional distributions differ. This approach is extended in Reeve et al. (2021), where assumptions are relaxed on the relationship between the source and target conditional distributions. Hector and Martin (2024) propose and study the inferential properties of an information-driven shrinkage estimator that is robust to heterogeneity between source and target feature-to-label mappings but assumes this mapping is of the same parametric form. These methods offer more mathematically rigorous motivations, but are restrictive in their modeling options. Such restrictions are eliminated in our proposed framework.

3. RECaST Framework

Our transfer learning problem is defined by the following four assumptions: (i) there is a well-developed structural component of the prediction model for the source domain denoted by $f(\boldsymbol{\theta}_S, x_S)$ which represents the relationship between the features and parameters; (ii) there exist ample source data for estimating the parameter(s) $\boldsymbol{\theta}_S$; (iii) $\mathcal{X}_S = \mathcal{X}_T$, and the structural component of the target prediction model, denoted by $g(\boldsymbol{\theta}_T, x_T)$, is believed to be *similar* to $f(\boldsymbol{\theta}_S, x_T)$; and (iv) there does not exist sufficient target data for reliably estimating the parameter(s) $\boldsymbol{\theta}_T$. We hereafter refer to $f(\boldsymbol{\theta}_S, x_S)$ and $g(\boldsymbol{\theta}_T, x_T)$ as *structural components* of their respective models. The notion of *similarity* will be defined in the construction of our RECaST framework for transfer learning, presented next.

Denote the forward data-generating representations of $P(y_S \mid \boldsymbol{x}_S)$ and $P(y_T \mid \boldsymbol{x}_T)$, respectively, by

$$\begin{aligned} Y_S &= h\big\{f(\boldsymbol{\theta}_S, \boldsymbol{x}_S), U_S\big\} \quad \text{and} \\ Y_T &= h\big\{g(\boldsymbol{\theta}_T, \boldsymbol{x}_T), U_T\big\}, \end{aligned} \tag{1}$$

where $\mathcal{X}_S = \mathcal{X}_T = \mathbb{R}^p$, and U_T and U_S are independent and identically distributed auxiliary random variables. We give two examples of the h function (for continuous and binary response examples), but the h function is much more general. It is to be understood as any scalar-valued function that relates the covariates to the auxiliary random variable in the fashion of a data generating equation. In fact, in the case of a continuous random variable, the h function can be taken to be the inverse cumulative distribution function, by the probability integral transform. For example, if

$$\begin{aligned} f(\boldsymbol{\theta}_S, \boldsymbol{x}_S) &= \boldsymbol{x}_S^\top \boldsymbol{\theta}_S, \\ h(\boldsymbol{x}_S^\top \boldsymbol{\theta}_S, U_S) &= \boldsymbol{x}_S^\top \boldsymbol{\theta}_S + U_S, \quad \text{and} \\ U_S &\sim \mathcal{N}(0,1), \end{aligned}$$

then $Y_S \sim \mathcal{N}(\boldsymbol{x}_S^\top \boldsymbol{\theta}_S, 1)$. Or in the case of binary outcome data, for example, if

$$\begin{aligned} f(\boldsymbol{\theta}_S, \boldsymbol{x}_S) &= \text{expit}(\boldsymbol{x}_S^\top \boldsymbol{\theta}_S), \\ h(\boldsymbol{x}_S^\top \boldsymbol{\theta}_S, U_S) &= \mathbf{1}\{U_S < \text{expit}(\boldsymbol{x}_S^\top \boldsymbol{\theta}_S)\}, \quad \text{and} \\ U_S &\sim \text{Uniform}(0,1), \end{aligned}$$

then $Y_S \sim \text{Bernoulli}\{\text{expit}(\boldsymbol{x}_S^\top \boldsymbol{\theta}_S)\}$, where $\text{expit}(z) := e^z/(1+e^z)$. The *similarity* between the source and target that makes this a formulation of a transfer learning problem is determined by how well the structural component $f(\boldsymbol{\theta}_S, \boldsymbol{x}_T)$ of the source model approximates the structural component $g(\boldsymbol{\theta}_T, \boldsymbol{x}_T)$ of the target model.

Accordingly, transfer learning should be effective if $\beta := g(\boldsymbol{\theta}_T, \boldsymbol{x}_T)/f(\boldsymbol{\theta}_S, \boldsymbol{x}_T) \approx 1$, and sufficient source data is available for reliable estimation of $\boldsymbol{\theta}_S$; in fact, the source and target models are identical if $\beta = 1$. Assuming $f(\boldsymbol{\theta}_S, \boldsymbol{x}_T) \neq 0$ *almost surely* (a.s.), it follows a.s. that

$$Y_{T,i} = h\big\{\beta_i \cdot f(\boldsymbol{\theta}_S, \boldsymbol{x}_{T,i}), U_{T,i}\big\}, \tag{2}$$

for $i \in \{1, \ldots, n_T\}$, where $Y_{T,1}, \ldots, Y_{T,n_T}$ is an independent sample of n_T target labels with associated features $\boldsymbol{x}_{T,1}, \ldots, \boldsymbol{x}_{T,n_T}$, and $\beta_i := g(\boldsymbol{\theta}_T, \boldsymbol{x}_{T,i})/f(\boldsymbol{\theta}_S, \boldsymbol{x}_{T,i})$. The identity given by Equation (2) is further motivated by the fact that, for first-order approximations of the source and target models, if we assume $\boldsymbol{x}_{T,1}, \ldots, \boldsymbol{x}_{T,n_T} \overset{\text{iid}}{\sim} \mathcal{N}_p(\mathbf{0}, \boldsymbol{I}_p)$, then by Lemma 1 (a well-known result for which we provide a proof in Appendix A, for convenience), $\beta_i = (\boldsymbol{x}_{T,i}^\top \boldsymbol{\theta}_T)/(\boldsymbol{x}_{T,i}^\top \boldsymbol{\theta}_S) \sim \text{Cauchy}(\delta, \gamma)$, with

$$\begin{aligned} \delta &= \frac{\boldsymbol{\theta}_T^\top \boldsymbol{\theta}_S}{\|\boldsymbol{\theta}_S\|^2}, \quad \text{and} \\ \gamma &= \frac{1}{\|\boldsymbol{\theta}_S\|^2} \sqrt{\|\boldsymbol{\theta}_S\|^2 \|\boldsymbol{\theta}_T\|^2 - (\boldsymbol{\theta}_T^\top \boldsymbol{\theta}_S)^2}. \end{aligned}$$

Lemma 1 *For any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, if $\boldsymbol{x} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{I}_p)$ then $(\boldsymbol{x}^\top \boldsymbol{a})/(\boldsymbol{x}^\top \boldsymbol{b}) \sim$ Cauchy(δ, γ), with $\delta = \boldsymbol{a}^\top \boldsymbol{b}/\|\boldsymbol{b}\|^2$ and $\gamma = \|\boldsymbol{b}\|^{-2} \sqrt{\|\boldsymbol{b}\|^2 \|\boldsymbol{a}\|^2 - (\boldsymbol{a}^\top \boldsymbol{b})^2}$.*

That being so, while Equation (2) is motivated by a first-order approximation, $f(\boldsymbol{\theta}_S, \boldsymbol{x}_S)$ and $g(\boldsymbol{\theta}_T, \boldsymbol{x}_T)$ need *not* share the same structure to implement the RECaST framework described by Equation (2). In fact, Equation (2) does not make any account of $g(\boldsymbol{\theta}_T, \boldsymbol{x}_T)$; it only assumes that the source model and parameters are available with the target data.

In practice, we assume without loss of generality that features have been centered and scaled to have mean zero and unit variance. Central limit theory supports the Gaussian approximation for more complex, nonlinear models (i.e., for the large p scenarios that characterize modern ML approaches). Specifically, appealing to the Lyapunov or Lindeberg central limit theorem gives Gaussian approximations for the distributions of $\boldsymbol{x}_{T,i}^\top \boldsymbol{\theta}_S / \|\boldsymbol{\theta}_S\|^2$ and $\boldsymbol{x}_{T,i}^\top \boldsymbol{\theta}_T / \|\boldsymbol{\theta}_T\|^2$. For more general assumptions on f and g, first-order approximations motivate $f(\boldsymbol{\theta}_S, \boldsymbol{x}_{T,i}) \approx \boldsymbol{x}_{T,i}^\top \boldsymbol{\theta}_S$ and $g(\boldsymbol{\theta}_T, \boldsymbol{x}_{T,i}) \approx \boldsymbol{x}_{T,i}^\top \boldsymbol{\theta}_T$. The edge case with $\gamma \to \infty$ describes a situation in which there is no link between the source and target domains. Assuming $\gamma < \infty$, the RECaST model specified by Equation (2) with random effect $\beta_i \sim \text{Cauchy}(\delta, \gamma)$ fully characterizes the *similarity* between the source and target domains. In addition to being the exact distribution in the linear model case with Gaussian features, the Cauchy distribution also provides benefit through its heavy tails. This attribute allows β_i to capture large disparities between source and target data sets, improving the frequentist coverage of resulting prediction sets.

Estimating parameters of Cauchy distributions is a notoriously difficult problem since the heavy tails allow outlying events to happen with relatively high probability (Schuster, 2012). Some estimation procedures focus on estimating solely the location parameter (Zhang, 2010) or the scale parameter (Kravchuk and Pollett, 2012), but rarely both. Fegyverneki (2013) explores the trade-off between using simple robust estimators, for both parameters, which are less asymptotically efficient than the maximum likelihood estimators. Recently, limit theorems are established in Akaoka et al. (2022) for quasi-arithmetic means for point estimation in cases where the strong law of large numbers fails, such as with Cauchy random variables. The fact that the Cauchy distribution appears in our work speaks to the difficulty of a transfer learning problem.

There are three primary advantages of our RECaST transfer learning model formulation in Equation (2) with random effect $\beta_i \sim \text{Cauchy}(\delta, \gamma)$. First, regardless of the complexity of the source model (e.g., $f(\boldsymbol{\theta}_S, \cdot)$ could represent a DNN with millions of parameters in $\boldsymbol{\theta}_S$ trained on extensive source data), RECaST only ever requires estimation of the parameters δ and γ, and perhaps a scale parameter associated with $U_{T,i}$ through $h(\cdot, U_{T,i})$. Existing transfer learning methods require either estimation of $\boldsymbol{\theta}_T$ (often via fine-tuning from an estimate of $\boldsymbol{\theta}_S$) or learning of $n_T + n_S$ weights for pooling the source and target data, where n_S is the number of source training labels. The scalability of our approach cannot be overstated. Second, RECaST needs no source data, only requiring the estimated source parameters $\widehat{\boldsymbol{\theta}}_S$. Such a feature is vital in applications such as with medical data where privacy constraints place legal and ethical barriers to accessing certain data sets. Third, RECaST naturally facilitates uncertainty quantification of target label predictions via the construction of prediction sets. The following two sections propose a Bayesian framework for estimation of the posterior predictive distribution of target labels in the continuous and binary response settings, respectively.

4. Continuous Response Data

4.1 Model and Estimation

Assume that $Y_{S,1}, \ldots, Y_{S,n_S}$ and $Y_{T,1}, \ldots, Y_{T,n_T}$ are mutually independent, continuous random variables generated according to source and target models, respectively, as expressed in Equation (1). Also assume that an estimator $f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x})$ is available for any feature vector $\boldsymbol{x} \in \mathcal{X}_S = \mathcal{X}_T$, where $\widehat{\boldsymbol{\theta}}_S$ is an estimator of $\boldsymbol{\theta}_S$ based on $Y_{S,1}, \ldots, Y_{S,n_S}$. In the continuous response setting, a natural choice for the h function in the RECaST model, defined by

Equation (2), is the Gaussian innovation formulation,

$$Y_{T,i} = \beta_i \cdot f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i}) + \sigma \cdot U_{T,i},$$

independently for $i \in \{1, \dots, n_T\}$, where $U_{T,i} \sim \mathcal{N}(0,1)$, $\sigma > 0$ is a scaling parameter to be learned from the target data, and $\beta_i \sim \text{Cauchy}(\delta, \gamma)$.

We specify a canonical prior on (δ, γ, σ) as

$$\pi(\delta, \gamma, \sigma) = \mathcal{N}(\delta \mid 1, \sigma_\delta^2) \cdot \log\mathcal{N}(\gamma \mid a, b) \cdot \log\mathcal{N}(\sigma \mid c, d).$$

The prior distributions are standard for shape and scale parameters. The hyperparameters for σ can be chosen based on prior information about the target domain. The hyperparameters for δ and γ can be chosen based on prior information about similarity between the source and target data. If the domains are known to be very similar, then the prior on δ may be centered near 1 with a small variance and the prior on γ may be chosen to have a mode near 0 with a small variance. This will result in a prior favoring δ and γ values that encourage $\beta = g(\boldsymbol{\theta}_T, \boldsymbol{x}_T)/f(\boldsymbol{\theta}_S, \boldsymbol{x}_T)$ values of 1, which indicates a similar source and target. In practice, to demonstrate the robustness of the RECaST framework and to cover a broad range of transfer learning settings, we choose hyperparameter values that induce diffuse priors. See Appendix C for more details.

A posterior distribution of the parameters (δ, γ, σ) can be expressed as

$$\begin{aligned}
&\pi\big(\delta, \gamma, \sigma \mid y_{T,1}, \dots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big) \\
&= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \pi\big(\delta, \gamma, \sigma, \beta_1, \dots, \beta_{n_T} \mid y_{T,1}, \dots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big)\ d\beta_1 \dots d\beta_{n_T} \\
&\propto \pi(\delta, \gamma, \sigma) \cdot \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{i=1}^{n_T} \Big[\mathcal{N}\big\{y_{T,i} \mid \beta_i f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i}), \sigma^2\big\} \cdot \text{Cauchy}(\beta_i \mid \delta, \gamma)\Big]\ d\beta_1 \dots d\beta_{n_T} \\
&= \pi(\delta, \gamma, \sigma) \cdot \prod_{i=1}^{n_T} \int_{\mathbb{R}} \mathcal{N}\big\{y_{T,i} \mid \beta_i f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i}), \sigma^2\big\} \cdot \text{Cauchy}(\beta_i \mid \delta, \gamma)\ d\beta_i \\
&= \pi(\delta, \gamma, \sigma) \cdot \prod_{i=1}^{n_T} \int_{\mathbb{R}} \mathcal{N}\big\{\beta_i f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i}) \mid y_{T,i}, \sigma^2\big\} \cdot \text{Cauchy}(\beta_i \mid \delta, \gamma)\ d\beta_i \\
&= \pi(\delta, \gamma, \sigma) \cdot \prod_{i=1}^{n_T} \int_{\mathbb{R}} \mathcal{N}\bigg\{\beta_i \ \bigg|\ \frac{y_{T,i}}{f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})}, \frac{\sigma^2}{f^2(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})}\bigg\} \cdot \frac{\text{Cauchy}(\beta_i \mid \delta, \gamma)}{\mid f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i}) \mid}\ d\beta_i,
\end{aligned} \tag{3}$$

where the univariate integrals in the last expression can be evaluated numerically. Next, the posterior predictive distribution of the label $\widetilde{Y}_T$ associated with some new target feature vector $\widetilde{\boldsymbol{x}}_T$ can be derived as the marginal distribution of

$$\begin{aligned}
&\pi\big(\widetilde{y}_T, \widetilde{\beta}, \sigma, \delta, \gamma \mid y_{T,1}, \dots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big) \\
&= \mathcal{N}\{\widetilde{y}_T \mid \widetilde{\beta} f(\widehat{\boldsymbol{\theta}}_S, \widetilde{\boldsymbol{x}}_T), \sigma^2\} \cdot \pi\big(\widetilde{\beta}, \sigma, \delta, \gamma \mid y_{T,1}, \dots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big) \\
&= \mathcal{N}\{\widetilde{y}_T \mid \widetilde{\beta} f(\widehat{\boldsymbol{\theta}}_S, \widetilde{\boldsymbol{x}}_T), \sigma^2\} \cdot \text{Cauchy}(\widetilde{\beta} \mid \delta, \gamma) \cdot \pi\big(\delta, \gamma, \sigma \mid y_{T,1}, \dots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big).
\end{aligned} \tag{4}$$

4.2 Remarks on Implementation

To estimate the posterior distribution given in Equation (3), we implement a random walk Metropolis-Hastings algorithm, numerically solving the univariate integrals with the `Julia` package `QuadGK` (Johnson, 2013). Furthermore, by expressing these integrals as expectations with respect to a Gaussian distribution (i.e., the final expression in Equation (3)), we show that they are numerically equivalent to definite integrals from -39 to 39. See Appendix B for the mathematical details of this bound. This substantially reduces the computational overhead for the numerical integration.

We detail our implementation of the Metropolis-Hastings algorithm in Appendix C. The chosen number of iterations and length of the burn-in period can be adjusted based on computational resources. Because Metropolis-Hastings evaluates the likelihood for all target data points for each iteration, the computational complexity is $\mathcal{O}(n_T \cdot n_{\text{iterations}})$, with $n_{\text{iterations}}$ the number of Metropolis-Hastings iterations. The fact that n_T is assumed to be small for transfer learning problems mitigates concerns about scalability. Posterior predictive credible sets can be constructed as usual in Bayesian inference, from the highest posterior density regions calculated via the empirical quantiles of the sampled posterior predictive values.

In Algorithm 1, we propose a procedure for drawing samples from the posterior predictive distribution described by Equation (4). Again take $\widetilde{\boldsymbol{x}}_T$ to be the feature vector for a new target data point with label $\widetilde{Y}_T$. With the learned posterior distribution of (δ, γ, σ), we are able to sample from the posterior predictive distribution of $\widetilde{Y}_T$. We first sample n_{post} (δ, γ, σ) triplets from the posterior distribution. For *each* of these triplets, we sample n_β β's from a Cauchy distribution with location and scale parameters corresponding to the δ and γ sampled from the posterior. Finally, for *each* sampled β we sample n_Y $\widetilde{y}_T$'s from the normal distribution with mean and variance determined by $\widetilde{\boldsymbol{x}}_T$, the sampled β, and the sampled σ. This gives a total of $n_{\text{post}} \cdot n_\beta \cdot n_Y$ samples from the posterior predictive distribution for each new target observation. These samples are used to construct the posterior predictive credible sets as described in Algorithm 1 with a computational complexity of $\mathcal{O}(n_{\text{post}} \cdot n_\beta \cdot n_Y)$. We discuss our choices for these parameters in Appendix C. We showcase the effectiveness of these proposed computational strategies in a variety of simulation scenarios in Section 6.2.

Algorithm 1 RECaST posterior predictive sampling: continuous response data

```
Input: $\widetilde{\boldsymbol{x}}_T$, samples from $\pi(\delta, \gamma, \sigma \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S)$, and sample sizes $n_{\text{post}}$, $n_\beta$, and $n_Y$
Output: A sample of values from $\pi(\widetilde{y}_T \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S)$
  for $i \leftarrow 1$ to $n_{\text{post}}$ do
    $\delta, \gamma, \sigma \leftarrow \text{random}\{\pi(\delta, \gamma, \sigma \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S)\}$
    for $j \leftarrow 1$ to $n_\beta$ do
      $\widetilde{\beta} \leftarrow \text{random}\{\text{Cauchy}(\delta, \gamma)\}$
      for $k \leftarrow 1$ to $n_Y$ do
        $\widetilde{Y}_T \leftarrow \text{random}[\mathcal{N}\{\widetilde{\beta} f(\widehat{\boldsymbol{\theta}}_S, \widetilde{\boldsymbol{x}}_T), \sigma^2\}]$
      end for
    end for
  end for
```

4.3 Theoretical Guarantees

In this section, we establish the asymptotic validity of our proposed posterior predictive credible sets in the case of linear source and target models with independent Gaussian innovations. Here, asymptotic validity means that the empirical coverage of a $1-\alpha$ level prediction credible set attains $1-\alpha$ level coverage, asymptotically in n_T, as described by the result of Theorem 3. Our mathematical proof of this result and of all supporting results are organized in Appendix A.

Suppose that $Y_{S,j}$ follows a Gaussian distribution centered at $\boldsymbol{x}_{S,j}^\top \boldsymbol{\theta}_S$, independently for $j \in \{1, \ldots, n_S\}$. In the class of transfer learning problems we consider, it is assumed that consistent or meaningful estimators are available for all source model parameters, and that ample data/resources are available for estimating them. Accordingly, assume that n_S is sufficiently large such that $\boldsymbol{\theta}_S$ is regarded as known. Next, assume that $Y_{T,1}, \ldots, Y_{T,n_T} \overset{\text{iid}}{\sim} \mathcal{N}(\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T, \sigma^2)$, for some feature vector $\widetilde{\boldsymbol{x}} \in \mathcal{X}_T = \mathcal{X}_S$, and $\boldsymbol{\theta}_T$ unknown. Leveraging the RECaST transfer learning framework, the likelihood function of (δ, γ) can be expressed as

$$L(\delta, \gamma \mid y_{T,1}, \ldots, y_{T,n_T}, \beta_1, \ldots, \beta_{n_T}) = \prod_{i=1}^{n_T} \Big[\mathcal{N}\big\{y_{T,i} \mid \beta_i \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S, \sigma^2\big\} \cdot \text{Cauchy}(\beta_i \mid \delta, \gamma)\Big]. \quad (5)$$

We investigate the asymptotic coverage of prediction sets constructed from the RECaST posterior predictive distribution with plugin maximum likelihood estimators (MLEs) $\widehat{\delta}$ and $\widehat{\gamma}$ for δ and γ, respectively:

$$\pi(\widetilde{y}_T, \widetilde{\beta} \mid y_1, \ldots, y_{n_T}) = \mathcal{N}(\widetilde{y}_T \mid \widetilde{\beta} \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S, \sigma^2) \cdot \text{Cauchy}(\widetilde{\beta} \mid \widehat{\delta}, |\widehat{\gamma}|).$$

This is the same as considering maximum a posteriori (MAP) estimators for δ and γ with a flat prior $\pi(\delta, \gamma) \propto 1$, and the choice of prior is not so meaningful in the $n_T \to \infty$ setting. Recall that in the RECaST framework the $\beta_1, \ldots, \beta_{n_T}$ that appear in the likelihood function in Equation (5) are iid Cauchy(δ, γ) random effects. Nonetheless, we demonstrate with Lemma 2 that the MLEs $\widehat{\delta}$ and $\widehat{\gamma}$ converge in probability to fixed points such that

$$\pi(\widetilde{Y}_T, \widetilde{\beta} \mid y_1, \ldots, y_{n_T}) \approx \mathcal{N}(\widetilde{Y}_T \mid \widetilde{\beta} \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S, \sigma^2) \cdot \mathbf{1}\left\{\widetilde{\beta} = \frac{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T}{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S}\right\} = \mathcal{N}(\widetilde{Y}_T \mid \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T, \sigma^2),$$

as desired. This fact leads to our main theoretical result, Theorem 3, which establishes the asymptotic validity of $1-\alpha$ level RECaST prediction sets of the form $[a_{n_T}^\alpha, b_{n_T}^\alpha]$, with

$$\begin{aligned} a_{n_T}^\alpha &:= \Phi^{-1}(\alpha/2) \cdot \sigma + \widetilde{\beta} \cdot \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S \quad \text{and} \\ b_{n_T}^\alpha &:= \Phi^{-1}(1-\alpha/2) \cdot \sigma + \widetilde{\beta} \cdot \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S, \end{aligned}$$

for any $\alpha \in (0,1)$ and $\widetilde{\beta} \sim \text{Cauchy}(\widehat{\delta}, |\widehat{\gamma}|)$.

Lemma 2 *Assuming $Y_{T,1}, \ldots, Y_{T,n_T} \overset{iid}{\sim} \mathcal{N}(\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T, \sigma^2)$ and $\beta_1, \ldots, \beta_{n_T} \overset{iid}{\sim}$ Cauchy(δ, γ), independently, the MLEs of δ and γ for Equation* (5) *satisfy*

$$\widehat{\delta} \longrightarrow \frac{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T}{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S} \quad \text{and} \quad \widehat{\gamma} \longrightarrow 0$$

in probability as $n_T \to \infty$.

Theorem 3 *Assume that $\widetilde{Y}_T \sim \mathcal{N}(\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T, \sigma^2)$. Then, for any $\alpha \in (0,1)$,*

$$P\Big(\widetilde{Y}_T \in [a^\alpha_{n_T}, b^\alpha_{n_T}]\Big) = \int_{a^\alpha_{n_T}}^{b^\alpha_{n_T}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\widetilde{y}_T - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T)^2} d\widetilde{y}_T \longrightarrow 1-\alpha$$

in probability as $n_T \to \infty$.

In Section 6, we provide empirical evidence that RECaST achieves near nominal coverage even in more practical, small n_T settings, trained on target data that arise from both linear and non-linear models. In the empirical investigations in Section 6, we relax the assumptions of known σ and the availability of repeated samples from a fixed feature vector $\widetilde{\boldsymbol{x}}$.

5. Binary Response Data

5.1 Model and Estimation

Assume that $Y_{S,1}, \ldots, Y_{S,n_S}$ and $Y_{T,1}, \ldots, Y_{T,n_T}$ are mutually independent, Bernoulli random variables generated according to source and target models, respectively, as expressed in Equation (1). Also assume that an estimator $f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x})$ is available for any feature vector $\boldsymbol{x} \in \mathcal{X}_S = \mathcal{X}_T$, where $\widehat{\boldsymbol{\theta}}_S$ is an estimator of $\boldsymbol{\theta}_S$ based on $Y_{S,1}, \ldots, Y_{S,n_S}$. In the binary response setting, a natural choice for the h function in the RECaST model, defined by Equation (2), is the logistic model formulation,

$$Y_{T,i} = \mathbf{1}\big[U_{T,i} < \operatorname{expit}\{\beta_i \cdot f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})\}\big],$$

with $U_{T,i} \sim \text{Uniform}(0,1)$ independently for $i \in \{1, \ldots, n_T\}$ and $\beta_i \sim \text{Cauchy}(\delta, \gamma)$.

As in the continuous setting, the RECaST posterior distribution of the parameters can be constructed as

$$\begin{aligned}
&\pi\big(\delta, \gamma \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big) \\
&= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \pi\big(\delta, \gamma, \beta_1, \ldots, \beta_{n_T} \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big)\ d\beta_1 \ldots d\beta_{n_T} \\
&\propto \pi(\delta,\gamma) \cdot \prod_{i=1}^{n_T} \int_{\mathbb{R}} \text{Bernoulli}\big[y_{T,i} \mid \operatorname{expit}\{\beta_i f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})\}\big] \cdot \text{Cauchy}(\beta_i \mid \delta, \gamma)\ d\beta_i,
\end{aligned}$$

and the posterior predictive distribution of the label $\widetilde{Y}_T$ associated with some new target feature vector $\widetilde{\boldsymbol{x}}_T$ can be derived as the marginal distribution of

$$\begin{aligned}
&\pi\big(\widetilde{y}_T, \widetilde{\beta}, \delta, \gamma \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big) \\
&= \text{Bernoulli}\big[\widetilde{y}_T \mid \operatorname{expit}\{\widetilde{\beta} f(\widehat{\boldsymbol{\theta}}_S, \widetilde{\boldsymbol{x}}_T)\}\big] \cdot \text{Cauchy}(\widetilde{\beta} \mid \delta, \gamma) \cdot \pi\big(\delta, \gamma \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big). \quad (6)
\end{aligned}$$

We specify a canonical prior on (δ, γ) as

$$\pi(\delta, \gamma) = \mathcal{N}(\delta \mid 1, \sigma_\delta^2) \cdot \log\mathcal{N}(\gamma \mid a, b),$$

with diffuse choices of the hyperparameters σ_δ, a, b. A similar description to that in Section 3.1 of the choice of priors holds here.

A $1-\alpha$ level RECaST prediction credible set, denoted $\Gamma^{\alpha}_{n_T}$, for binary response values is constructed as

$$\Gamma^{\alpha}_{n_T} = \begin{cases} \{0\}, & \text{if} \quad \widetilde{p} < 1-\widetilde{p} \quad \text{and} \quad 1-\alpha \leq 1-\widetilde{p} \\ \{1\}, & \text{if} \quad 1-\widetilde{p} \leq \widetilde{p} \quad \text{and} \quad 1-\alpha \leq \widetilde{p} \\ \{0,1\}, & \text{else,} \end{cases} \tag{7}$$

where $\widetilde{p} := \pi\big(\widetilde{y}_T = 1 \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big)$.

5.2 Remarks on Implementation

The RECaST transfer learning computations in the binary response setting follow analogously to those described in Section 4.2. For completeness, Algorithm 2 specifies the procedure we propose for drawing samples from the posterior predictive distribution described by Equation (6).

Algorithm 2 RECaST posterior predictive sampling: binary response data

```
Input: $\widetilde{\boldsymbol{x}}_T$, samples from $\pi\big(\delta, \gamma \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big)$, and sample sizes $n_{\text{post}}$, $n_\beta$, and $n_Y$
Output: A sample of values from $\pi\big(\widetilde{y} \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big)$
  for $i \leftarrow 1$ to $n_{\text{post}}$ do
      $\delta, \gamma \leftarrow \text{random}\big\{\pi\big(\delta, \gamma \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big)\big\}$
      for $j \leftarrow 1$ to $n_\beta$ do
          $\widetilde{\beta} \leftarrow \text{random}\big\{\text{Cauchy}(\delta, \gamma)\big\}$
          for $k \leftarrow 1$ to $n_Y$ do
              $\widetilde{Y}_T \leftarrow \text{random}\Big(\text{Bernoulli}\big[\text{expit}\big\{\widetilde{\beta} f(\widehat{\boldsymbol{\theta}}_S, \widetilde{\boldsymbol{x}}_T)\big\}\big]\Big)$
          end for
      end for
  end for
```

6. Simulation Study

6.1 Objectives and Setup

In this section, we examine the finite sample performance of RECaST through simulations on synthetic data. We consider continuous and binary responses with source models corresponding to linear (RECaST LM) and logistic (RECaST GLM) regression, respectively, as well as a DNN (RECaST DNN) source model for both response types. We assess the empirical coverage with respect to the nominal coverage level of the prediction sets. If the method is calibrated, the empirical coverage will match the nominal significance level. We use the terms *empirical coverage* and *observed coverage* interchangeably.

We generate the synthetic data from linear and logistic regressions with source parameter vector $\boldsymbol{\theta}_S$ and target parameter vector $\boldsymbol{\theta}_T$, with $p = 50$ features (including an intercept). The features are generated from the standard Gaussian distribution, $\boldsymbol{x}_{S,i}, \boldsymbol{x}_{T,j} \sim \mathcal{N}_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1})$. We fix the source data generating parameters $\boldsymbol{\theta}_S$. The source data generating parameters are set to $\boldsymbol{\theta}_S = (-\boldsymbol{a}, \boldsymbol{b})$ where $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{25}$ have components independently sampled from Uniform$(0.75, 5)$ and then fixed for all simulations. The *similarity* of source and target domains is controlled by choosing the value of $\sigma_{\text{TL}} > 0$ in constructing $\boldsymbol{\theta}_T = \boldsymbol{\theta}_S + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \sigma^2_{\text{TL}} \boldsymbol{I}_p)$. We consider values of $\sigma^2_{\text{TL}} \in \{0, 0.25, 1, 4\}$. Setting $\sigma^2_{\text{TL}} = 0$ corresponds to $\boldsymbol{\theta}_T = \boldsymbol{\theta}_S$, i.e., no difference between the source and target distributions. Since the source parameters lie within $[-5, -0.75] \cup [0.75, 5]$, a variance of $\sigma^2_{\text{TL}} = 4$ allows for significant differences between $\boldsymbol{\theta}_T$ and $\boldsymbol{\theta}_S$. We fix the source sample size at $n_S = 1000$, and vary the target sample size n_T to examine performance when $p < n_T$ ($n_T = 100, 250$), p is near n_T ($n_T = 40, 60$), and $p > n_T$ ($n_T = 20$). We simulate 300 source and target data sets for each of these 20 combinations of σ^2_{TL} and n_T values, and implement the estimation procedures described in Sections 4.2 and 5.2. See Appendix C for additional details about the specifics of our implementations.

We compare to a linear model baseline (LM) which is built only on the target data. Another baseline for comparison is constructed from training a DNN on the target data, without any transfer learning, and we compare RECaST to other state-of-the-art transfer learning approaches. We build a DNN on the source data and fine-tune the last layer on the target data (Unfreeze DNN); this is often referred to as *freezing* the weights of the source DNN and *unfreezing* the last layer. See Appendix D for details on this procedure. Other state-of-the-art transfer learning approaches that we compare RECaST to include TransRF (Gu et al., 2022), a source-free method that adapts a random forest model built in the source domain to target data, and glmtrans (Tian and Feng, 2022), which is based on penalized GLMs and designed to mitigate the impact of negative transfer. Unlike RECaST and TransRF, glmtrans requires the source data to be available during the training of the model. In the continuous setting, we compare to the source-free methods outlined by Tripuraneni et al. (2021). We compare to both their first order method (MTL FO) and their method of moments approach (MTL MoM). Note that while this method does not require the source data when learning the target model, it does require that the source parameters were learned following their formulation whereas RECaST is agnostic to the choice of source model. In the binary setting, we compare RECaST to the regularized logistic regression (Wiens) approach of Wiens et al. (2014). This approach uses the combined source and target EHR data to build a regularized model for disease prediction – similar to the real data application we consider in Section 7, but with the disadvantage that Wiens requires access to the source data (while RECaST does not). In the binary setting, we also compare RECaST to the adversarial transfer learning approach WDGRL (Shen et al., 2018), which also requires access to the source data.

Throughout this section, all DNN training proceeds by setting aside a portion of the training data to be used as a calibration data set. The final DNN parameters are chosen from the epoch with the minimum calibration loss to improve generalizability to out-of-sample test sets. Additional details/specifications for our DNN training procedures are provided in Appendix D.

6.2 Continuous Response Results

Table 1 and Figure 1 summarize the performance of the prediction uncertainty quantification provided by our RECaST framework implementations. Table 1 presents the empirical coverage for 95% nominal level prediction sets for each simulation setting. Recall that the empirical coverage should ideally match the nominal significance for a given level; an empirical coverage greater than the nominal coverage level corresponds to a conservative interval estimate. RECaST methods consistently provide empirical coverage at or slightly above nominal levels, supporting the use of RECaST for inference on out-of-sample target domain predictions. Additionally, Figure 1 plots empirical versus nominal coverage for the $\sigma^2_{\text{TL}} = 0.25$, $n_T = 100$ and $\sigma^2_{\text{TL}} = 4$, $n_T = 20$ settings at a grid of nominal levels. The empirical coverages consistently achieve the associated nominal levels or are slightly conservative.

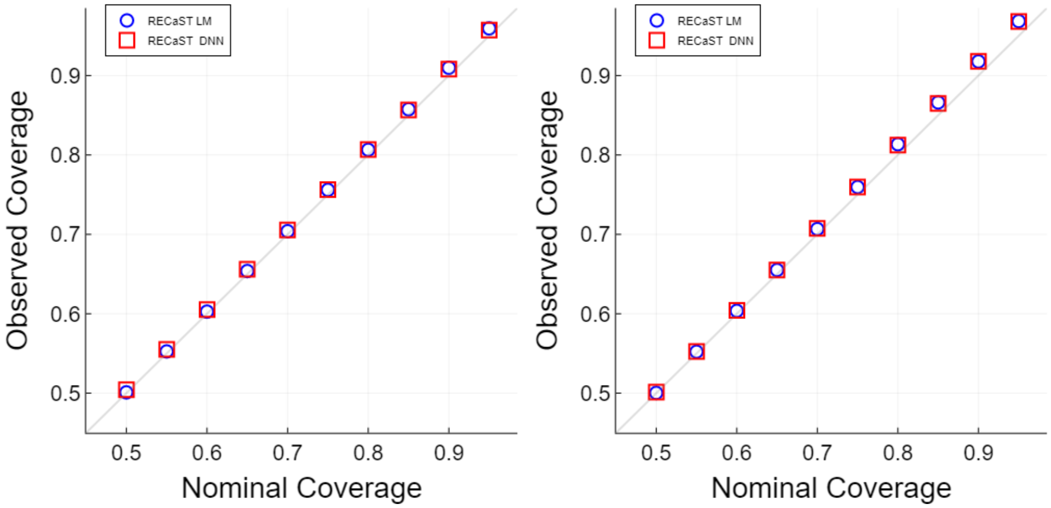

Figure 1: Reliability curves of the nominal coverage versus the empirical coverage, averaged over 300 source and target data sets for each setting; the out-of-sample test sets each contain 250 observations. The left panel shows an easy setting: $n_T = 100$ and $\sigma^2_{\text{TL}} = 0.25$. The right panel shows a difficult setting: $n_T = 20$ and $\sigma^2_{\text{TL}} = 4$.

Out-of-sample root mean squared errors (RMSEs) for all methods, averaged over 300 source and target data sets are presented in Table 2. The LM provides the best prediction when the sample size is large since in this case it correctly specifies the data generating model and has enough data to estimate the parameters. There is a large decrease in performance, noted by the increase in RMSE, when $n_T < p$ and a generalized inverse has to be used for parameter estimation. As expected, the performance of DNN deteriorates as the target sample size decreases. Note that the baseline DNN is overparameterized, which leads to it having higher RMSEs than the baseline LM.

Interestingly, the RECaST RMSE values remain consistent for each value of σ^2_{TL}, regardless of sample size, suggesting that RECaST is appropriate even when the target sample size is so small as to preclude a target-only analysis. Meanwhile, Unfreeze DNN exhibits an increase in RMSE for each value of σ^2_{TL} as n_T decreases. As source and target become more dissimilar, both Unfreeze DNN and RECaST exhibit similar increases in RMSE. In fact, with $n_T = 250$ and $\sigma^2_{\text{TL}} = 4$, the target-only DNN outperforms both RECaST methods. This setting is the most prone to negative transfer: the target sample size is large enough to learn meaningful DNN parameters, *and* the source and target data distributions differ greatly, making transfer difficult. We see this phenomenon with the target only LM as well; with a sample size of $n_T = 40$, both RECaST methods outperform the LM except for when the source and target are most dissimilar. When $n_T = 20$, the RECaST methods outperform the LM in all settings. This highlights a situation where transfer learning is necessary because the target domain lacks sufficient data to efficiently estimate the target parameters, even with a correctly specified model.

The MTL FO and MTL MoM both see increases in RMSE as the source and target become more dissimilar and see a larger increase in RMSE as the target sample size decreases. Interestingly, in this simulation these two methods have the same performance when there are more target sample points than there are features. While in some settings with larger target sample sizes the Unfreeze DNN slightly outperforms RECaST, it has larger standard errors and fails to provide uncertainty quantification. We find that TransRF sometimes performs well but with high RMSE variance. We were not able to evaluate TransRF when the target sample size was 20 as the software gave NA values instead of predictions without an accompanying error message. While glmtrans sometimes has smaller RMSE than RECaST, recall that it requires access to the source data and that only RECaST provides uncertainty quantification for predictions.

n_T	σ^2_{TL}	RECaST LM	RECaST DNN
250	0	96(1.8)	94(1.9)
	0.25	95(1.9)	95(1.9)
	1	95(1.9)	95(1.8)
	4	95(2.0)	95(1.9)
100	0	96(1.8)	94(2.1)
	0.25	96(1.8)	96(2.0)
	1	96(1.8)	96(1.8)
	4	96(1.8)	96(1.9)
60	0	97(2.2)	94(2.6)
	0.25	96(1.9)	96(1.8)
	1	96(1.8)	96(1.8)
	4	96(1.8)	96(1.8)
40	0	97(2.1)	94(3.3)
	0.25	96(2.4)	96(2.4)
	1	96(2.6)	96(2.5)
	4	96(2.8)	96(2.8)
20	0	98(1.8)	95(3.0)
	0.25	97(2.6)	97(2.8)
	1	97(2.6)	97(2.8)
	4	97(2.7)	97(2.9)

Table 1: Empirical coverage (standard error) at the 95% nominal level, averaged over 300 source and target data sets for each setting; the out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

n_T	σ^2_{TL}	LM	DNN	RECaST LM	RECaST DNN	Unfreeze DNN	TransRF	glmtrans	MTL FO	MTL MoM
250	0	0.57(0.03)	2.8(0.38)	0.52(0.027)	1.2(0.090)	0.58(0.038)	14(1.5)	0.56(0.026)	1.9(0.07)	1.9(0.07)
	0.25	0.57(0.03)	2.9(0.37)	3.6(0.43)	3.8(0.4)	2.8(0.42)	14(1.4)	0.56(0.027)	1.9(0.5)	1.9(0.5)
	1	0.57(0.03)	3.1(0.43)	7.1(0.86)	7.2(0.84)	5.5(0.90)	14(1.5)	0.56(0.027)	2.0(1.0)	2.0(1.0)
	4	0.57(0.03)	3.7(0.52)	14(1.7)	14(1.7)	11(1.8)	17(2.6)	0.56(0.027)	2.5(1.7)	2.5(1.7)
100	0	0.71(0.07)	8.9(1.6)	0.52(0.022)	1.2(0.095)	0.81(0.095)	22(12)	0.69(0.047)	2.4(0.2)	2.4(0.2)
	0.25	0.71(0.06)	9.1(1.3)	3.6(0.42)	3.8(0.40)	3.2(0.57)	28(75)	0.73(0.068)	2.4(0.7)	2.4(0.7)
	1	0.71(0.06)	9.4(1.3)	7.1(0.85)	7.2(0.83)	6.3(1.1)	23(16)	0.74(0.075)	2.5(1.2)	2.5(1.2)
	4	0.71(0.06)	11(1.52)	14(1.7)	14(1.7)	13(2.1)	31(34)	0.74(0.073)	3.1(2.1)	3.1(2.1)
60	0	1.3(0.27)	14(2.5)	0.52(0.025)	1.2(0.11)	1.5(0.29)	46(73)	0.75(0.05)	4.0(0.87)	4.0(0.88)
	0.25	1.3(0.23)	13(1.6)	3.6(0.43)	3.8(0.42)	3.7(0.78)	48(99)	1.2(0.21)	4.0(1.2)	4.0(1.2)
	1	1.3(0.23)	14(1.8)	7.1(0.87)	7.2(0.86)	6.8(1.3)	54(85)	1.7(0.38)	4.2(2.1)	4.2(2.1)
	4	1.3(0.23)	16(2.6)	14(1.8)	14(1.8)	13(2.0)	51(59)	3.0(0.85)	5.1(3.5)	5.1(3.5)
40	0	10(2.1)	17(2.6)	0.52(0.024)	1.2(0.089)	1.8(0.61)	74(64)	0.78(0.063)	11(2.1)	10(1.9)
	0.25	11(1.8)	17(2.4)	3.6(0.41)	3.8(0.40)	4.1(1.1)	62(74)	2.5(0.51)	11(2.1)	10(1.9)
	1	11(2.0)	18(2.5)	7.2(0.83)	7.3(0.83)	7.6(2.2)	69(110)	4.7(1.1)	11(2.4)	11(2.2)
	4	12(2.0)	20(2.9)	14(1.7)	14(1.7)	14(3.0)	150(540)	8.9(2.3)	13(3.0)	13(2.9)
20	0	18(1.8)	21(1.8)	0.54(0.03)	1.2(0.11)	2.5(2.2)	-	0.81(0.078)	18(2.0)	17(1.6)
	0.25	18(1.8)	21(1.8)	3.7(0.44)	3.9(0.42)	4.7(2.5)	-	3.4(0.39)	18(1.9)	18(1.7)
	1	18(1.9)	22(2.0)	7.3(0.90)	7.4(0.90)	8.5(3.5)	-	6.7(0.8)	19(2.1)	18(1.9)
	4	21(2.4)	24(2.7)	15(1.8)	15(1.8)	16(4.0)	-	6.7(0.8)	21(2.5)	20(2.4)

Table 2: Out-of-sample RMSE (standard error) averaged over 300 source and target data sets for each setting; the out-of-sample test sets each contain 250 observations.

6.3 Binary Response Results

Table 3 shows that RECaST procedures, again, provide near nominal coverages with low standard errors across sample sizes in the binary response setting.

n_T	σ^2_{TL}	DNN	RECaST GLM	RECaST DNN	Wiens	Unfreeze DNN	TransRF	glmtrans	WDGRL
250	0	84(9.5)	95(0.78)	96(0.090)	100(0)	91(8.4)	89(12)	98(4.4)	95(4.4)
	0.25	89(7.5)	95(0.82)	96(0.13)	100(0)	93(6.8)	88(11)	98(2.6)	94(4.8)
	1	91(6.2)	95(0.65)	96(0.14)	100(0)	95(4.5)	87(9.5)	98(3.6)	86(6.0)
	4	93(6.4)	95(0.40)	95(0.39)	99(1.5)	95(4.8)	86(11)	97(4.1)	75(7.6)
100	0	80(12)	96(1.1)	96(0.25)	100.0(0)	90(8.7)	68(22)	96(6.4)	96(3.4)
	0.25	83(11)	95(1.3)	96(0.27)	100(0)	92(7.8)	78(3.0)	94(6.2)	93(4.2)
	1	88(8.2)	95(1.2)	96(0.35)	100(4.6)	94(5.9)	69(18)	94(8.5)	89(5.7)
	4	92(6.2)	95(0.81)	95(0.90)	95(13)	94(10.0)	49(20)	83(4.6)	89(2.3)
60	0	80(13)	95(1.2)	96(0.54)	100(0.0)	92(6.4)	65(19)	93(9.3)	95(3.7)
	0.25	77(17)	95(1.3)	96(0.67)	100(0.0)	91(8.1)	64(22)	88(12)	95(3.1)
	1	80(21)	95(1.1)	95(0.867)	100(0.49)	94(6.1)	63(19)	88(15)	90(4.7)
	4	84(15)	95(0.59)	95(1.0)	96.8(4.7)	93(8.7)	58(19)	89(11)	80(6.9)
40	0	68(23)	95(1.6)	96(0.86)	100(0.0)	89(11)	60(20)	88(14)	95(5.5)
	0.25	72(20)	95(1.6)	96(0.99)	100(0.0)	90(7.9)	55(25)	81(16)	95(4.1)
	1	76(19)	94(1.5)	95(1.2)	100(0.53)	89(8.7)	59(19)	85(14)	89(5.9)
	4	77(25)	94(1.4)	94(1.1)	97(3.2)	90(7.2)	63(20)	78(14)	78(7.4)
20	0	67(22)	95(1.1)	96(0.78)	100(0.0)	85(17)	-	86(15)	-
	0.25	75(16)	95(1.1)	95(0.98)	100(0.0)	86(14)	-	68(15)	-
	1	75(16)	95(0.84)	95(1.1)	100(0.55)	86(17)	-	63(18)	-
	4	72(13)	95(0.47)	94(0.99)	98(1.5)	80(18)	-	66(17)	-

Table 3: Empirical coverage (standard error) at the 95% nominal level, averaged over 300 source and target data sets for each setting; the out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

Compared to the other approaches, RECaST provides substantial inferential advantages that are robust to small target sample sizes and large dissimilarity between source and target. Recall from Equation (7) that prediction sets in the binary response setting are determined entirely by the Bernoulli probability of observing label 1. Thus, we can construct prediction sets for the DNN, Unfreeze DNN, and Wiens methods, as well. When a method fails to discriminate between the two labels at level $1-\alpha$ (e.g., when the Bernoulli probability of success and failure are *both below* $1-\alpha$), then the prediction set must include both labels to attain the $1-\alpha$ level. In such cases, as observed for the Wiens method in various settings in Table 3, the prediction set achieves 100% empirical coverage, but is unhelpful for prediction.

Table 4 provides the area under the receiver operator characteristic curve (AUC) for all methods and simulation settings. In all settings except one, RECaST DNN outperforms all other methods. We see similar patterns here as in the continuous setting. The RECaST models consistently report the highest AUC, with low standard errors across sample sizes. In contrast, the AUC of DNN and Unfreeze DNN drastically declines as n_T decreases. As expected, the AUC of all transfer learning methods decreases as the difficulty of the problem increases with larger values of σ^2_{TL}. RECaST DNN and WDGRL frequently outperform other methods; however, WDGRL requires access to the source data, an important limitation that is unrealistic in many applications. WDGRL crashed with a sample size of $n_T = 20$, so we are unable to evaluate its performance in these settings.

The benefits to coverage properties and predictive performance of the RECaST method are especially important in the binary response case. This demonstrates that RECaST can be used even when the linearity assumption of Lemma 1 is violated.

n_T	σ^2_{TL}	DNN	RECaST GLM	RECaST DNN	Wiens	Unfreeze DNN	TransRF	glmtrans	WDGRL
250	0	95(1.7)	98(2.1)	98(0.61)	80(3.5)	97(1.2)	66(13)	97(2.0)	97(0.95)
	0.25	95(1.6)	97(2.3)	98(0.89)	80(3.9)	97(1.2)	69(10)	96(1.6)	97(1.3)
	1	94(1.5)	93(3.8)	96(1.5)	79(3.9)	95(1.7)	66(11)	97(1.8)	95(1.5)
	4	95(1.7)	84(5.5)	89(2.8)	76(4.0)	89(3.0)	67(12)	97(1.6)	88(3.2)
100	0	85(7.9)	96(2.2)	98(0.64)	81(4.2)	96(2.2)	47(19)	87(9.3)	98(0.66)
	0.25	83(9.6)	95(2.7)	97(1.0)	80(4.1)	95(1.9)	18(14)	81(5.3)	97(1.0)
	1	84(8.4)	92(3.8)	95(1.4)	79(4.4)	93(2.4)	48(20)	81(5.5)	95(1.4)
	4	82(11)	83(4.7)	89(3.1)	74(4.3)	87(4.8)	49(20)	83(4.6)	89(2.3)
60	0	72(13)	96(1.9)	98(1.0)	80(4.3)	94(5.2)	49(20)	83(4.6)	89(2.3)
	0.25	74(11)	94(2.5)	97(1.4)	80(4.3)	94(2.34)	36(21)	74(5.1)	97(0.78)
	1	75(10)	90(3.5)	95(1.7)	78(4.1)	91(6.0)	33(20)	74(5.3)	95(1.9)
	4	72(11)	83(4.0)	89(3.3)	73(4.7)	84(8.0)	29(18)	75(5.6)	89(2.4)
40	0	68(11)	96(1.6)	98(1.1)	80(3.8)	94(4.5)	27(16)	83(16)	97(1.1)
	0.25	68(11)	94(2.2)	97(1.3)	80(4.0)	92(6.7)	19(15)	67(4.9)	97(1.2)
	1	65(12)	90(3.0)	95(1.9)	78(3.9)	89(7.9)	32(18)	69(5.7)	95(1.7)
	4	67(12)	82(4.1)	89(3.5)	74(4.2)	80(12)	31(18)	69(4.9)	89(3.2)
20	0	60(8.7)	96(1.7)	97(1.4)	80(3.9)	89(10)	-	81(18)	-
	0.25	61(9.1)	94(2.1)	97(1.9)	79(4.2)	87(13)	-	62(5.1)	-
	1	60(9.5)	90(2.7)	94(2.5)	77(4.5)	83(14)	-	60(5.0)	-
	4	62(8.2)	82(3.5)	88(3.0)	72(5.0)	77(11)	-	63(5.0)	-

Table 4: Out-of-sample AUC (standard error) averaged over 300 source and target data sets for each setting; the out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

6.4 Robustness of RECaST

6.4.1 Over-Parameterized RECaST DNN

In all previous simulations, the true data generating mechanisms are linear or logistic models. To test the robustness of RECaST, we now consider a more complex case where data are generated from neural networks. We generate data from a neural network with a densely connected input layer of size $\ell_1 = (p, 10)$ and then pass through a ReLU activation function to an output layer of size $\ell_2 = (10, 1)$, where there are $p = 50$ features generated as described in Section 6.1. For the binary response data, we append a sigmoid activation function to the end of the output layer. While the source and target data generating networks share architectures, we consider two relationships between the source and target neural network parameters.

In our first set of simulations, as in Section 6.1, we take the parameters of the source neural network to be $\boldsymbol{\theta}_S \overset{\text{iid}}{\sim} U(-1, 1)$ and define the parameters of the target neural network as $\boldsymbol{\theta}_T = \boldsymbol{\theta}_S + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}_{p\times 10+10\times 1}(\mathbf{0}, 0.025\boldsymbol{I})$. For a continuous outcome, Table 5 shows that the RECaST DNN methods have the lowest RMSE for all sample sizes. Both RECaST methods outperform the target-only DNN across all settings, even when the target sample size is large ($n_T = 250$). The glmtrans method performs similarly to RECaST LM but worse than RECaST DNN. For all sample sizes, the RECaST framework produces wide posterior predictive intervals with 100% observed coverage for the 95% nominal confidence level – see Table 12 in Appendix E. This greater than nominal coverage demonstrates RECaST will be conservative but reliable. Indeed, the observed over-coverage is safer than narrower intervals centered around incorrect values with below nominal coverage. For a binary outcome, Table 6 reveals that both RECaST methods outperform the target-only DNN for all sample sizes. This shows robustness to negative transfer. The performance of the RECaST methods is stable across target sample sizes in this setting, with stable AUCs and standard errors, whereas other methods degrade in performance as the target sample size decreases. Table 7 shows the empirical coverages of each method at the 75% nominal level. Only the RECaST GLM, RECaST DNN, and Wiens methods provide conservative coverage values for all sample sizes whereas the other methods tend to under-cover the true labels as the target sample size decreases.

n_T	LM	DNN	RECaST LM	RECaST DNN	Unfreeze DNN	TransRF	glmtrans	MTL FO	MTL MoM
250	2.9(0.18)	3.1(0.22)	2.9(0.16)	2.1(0.15)	3(0.22)	3.9(0.53)	2.7(0.15)	3.1(0.21)	3.1(0.2)
100	3.8(0.36)	4.2(0.51)	2.9(0.16)	2.1(0.15)	3.3(0.37)	5.5(2.6)	2.7(0.18)	3.8(0.37)	3.9(0.38)
60	6.7(1.6)	5.1(0.56)	2.9(0.16)	2.1(0.15)	3.7(0.65)	12(10)	2.8(0.19)	6.3(1.3)	6.4(1.4)
40	6.1(1.1)	5.4(0.56)	2.9(0.16)	2.1(0.15)	4(0.67)	170(560)	2.8(0.18)	7.4(1.4)	6.8(1.5)
20	4.8(0.38)	5.8(0.44)	3(0.21)	2.2(0.17)	4.6(0.93)	-	2.9(0.26)	7.2(0.83)	4.9(0.42)

Table 5: Out of sample RMSE (standard error) averaged over 300 source and target data sets when the generating models are neural networks and the target model parameters are generated as $\boldsymbol{\theta}_T = \boldsymbol{\theta}_S + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.025\boldsymbol{I})$. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

n_T	DNN	RECaST GLM	RECaST DNN	Wiens	Unfreeze DNN	TransRF	glmtrans	WDGRL
250	85(3.3)	92(1.6)	91(1.9)	75(4)	89(2.8)	64(13)	86(2.5)	89(1.7)
100	73(9.5)	92(1.6)	91(1.8)	75(3.5)	85(5.9)	46(18)	76(4.4)	89(1.7)
60	66(9.7)	92(1.7)	91(2)	75(4.4)	81(9)	29(22)	71(6.9)	89(2.2)
40	62(9)	92(1.8)	91(1.8)	76.0(3.5)	78(11)	24(16)	67(8)	89(1.9)
20	58(7.9)	92(1.8)	91(2)	76.0(3.6)	72(14)	-	56(6.4)	-

Table 6: Out-of-sample AUC (standard error) averaged over 300 source and target data sets when the generating models are neural networks and the target model parameters are generated as $\boldsymbol{\theta}_T = \boldsymbol{\theta}_S + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.025\boldsymbol{I})$. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

6.4.2 Orthogonal Source and Target Data Generating Model Parameters

In our second set of simulations, we set the source and target weight matrices to be orthogonal, i.e., $\boldsymbol{\theta}_S^\top \boldsymbol{\theta}_T = \mathbf{0}$. For a continuous outcome, Table 8 shows that RECaST again provides consistent predictive performance across target sample sizes. For small sample sizes, both RECaST methods outperform the target-only LM and DNN. The unfreeze DNN and glmtrans also perform well, but we mention again that they do not provide uncertainty quantification of predictions. Table 13 in Appendix E shows that RECaST provides conservative coverage intervals which, again, is a safe feature in this difficult transfer learning setting. For a binary outcome, Table 9 shows that RECaST again outperforms the target-only DNN in realistic settings where the target sample size is small. The RECaST methods have consistent AUCs across target sample sizes whereas other methods deteriorate as the sample size decreases. Table 10 shows that only RECaST GLM, RECaST DNN, and the Wiens method provide conservative uncertainty quantification for all target sample sizes at the 75% nominal level.

Overall, the results presented in this section show that RECaST is robust to negative transfer under more complex data generating mechanisms. In all cases, the RECaST methods outperformed the target-only DNN while boasting conservative predictive coverage intervals when the target sample size is small. In Appendix F we explore other relationships between the source and target data when the data generating mechanism is a (generalized) linear model. These include orthogonality of source and target parameters and the target data having more features than the source.

7. eICU Data

The eICU Collaborative Research Database (Pollard et al., 2018) is a publicly available database of ICU encounters across multiple hospitals in the United States, making it well-suited for imitating transfer learning settings using real data. In the spirit of the transfer learning application in Wiens et al. (2014), we focus on correctly diagnosing physiological shock for newly admitted ICU patients. We define a binary response variable as the indicator of the event that a patient experienced shock upon ICU admission, using a combination of Internal Classification of Diseases 10 (ICD-10) codes: R57 Shock, not elsewhere classified; R58 Hemorrhage, not elsewhere classified; or R65.21 Severe sepsis with septic shock.

n_T	DNN	RECaST GLM	RECaST DNN	Wiens	Unfreeze DNN	TransRF	glmtrans	WDGRL
250	68(13)	98(2)	97(4.1)	88(6.6)	71(13)	72(12)	80(11)	70(7.9)
100	63(12)	95(6.1)	94(6.7)	86(7.5)	70(13)	58(15)	72(12)	64(16)
60	59(12)	90(14)	91(10)	88(7.6)	66(14)	53(26)	75(16)	69(12)
40	58(12)	87(15)	87(13)	88(5.2)	63(15)	61(13)	69(17)	67(15)
20	55(12)	81(17)	81(18)	89(6.6)	59(14)	-	57(17)	-

Table 7: Empirical coverage (standard error) at the 75% nominal level for a binary response, averaged over 300 source and target data sets when the generating models are neural networks and the target model parameters are generated as $\boldsymbol{\theta}_T = \boldsymbol{\theta}_S + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.025\boldsymbol{I})$. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

n_T	LM	DNN	RECaST LM	RECaST DNN	Unfreeze DNN	TransRF	glmtrans	MTL FO	MTL MoM
250	0.93(0.43)	1.1(0.19)	2.9(0.91)	3(0.98)	1.6(0.45)	3.5(1.1)	1(0.41)	1.3(0.69)	1.3(0.69)
100	1.2(0.56)	2.3(0.51)	2.8(0.89)	3(0.97)	1.8(0.52)	5.9(3.2)	1.3(0.52)	1.6(0.86)	1.6(0.86)
60	2.2(1.1)	3.2(0.88)	2.8(0.89)	3(0.96)	2.1(0.62)	7.7(6.3)	1.9(0.7)	2.6(1.4)	2.7(1.3)
40	2.8(0.94)	3.8(0.97)	2.8(0.89)	3(0.97)	2.4(0.8)	21(24)	2.3(0.61)	4.5(1.3)	3.4(1.4)
20	3.7(1.1)	4.5(1.1)	2.9(0.91)	3.1(1)	2.9(0.89)	-	2.6(0.83)	6.6(1.1)	3.8(1.1)

Table 8: Out of sample RMSE (standard error) averaged over 300 source and target data sets when the source and target neural network weight matrices are orthogonal. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

Features are limited to baseline variables measured at admission. While the simulations of Section 6 explicitly link the source and target data through the data generation process, the similarity between source and targets defined in our eICU data application is unknown.

We consider 19 features including patient demographics, Acute Physiology Score III variables, and Glasgow Coma Scale test. Descriptions of these features can be found in Table 22 in Appendix H. The data consist of measurements on 45,945 patients across 156 unique hospitals. Only 700 of these patients were diagnosed with shock upon admission. No individual hospital had enough positive cases to be reliably used as a source data set. To curate a balanced data set, we take all 700 patients with shock and randomly sample an additional 700 patients with no shock. Next, 80% of the hospitals associated with our sampled 1,400 patients are randomly selected to define the 'source hospital'. The source data set consists of all ICU encounters at the 'source hospital'. Of the remaining 20% of hospitals, half are randomly assigned to the 'target training hospital', and the other half define a 'target testing hospital'. Notice that this procedure splits hospitals rather than patients; the source data set may not consist of 80% of patients. The target training and target testing data sets typically contain 80 to 130 patients each.

We repeat the described sampling procedure 300 times, to imitate 300 transfer learning scenarios from real data. A logistic regression model and a DNN model are trained on each of the 300 source data sets, and all previously considered binary response transfer learning methods are implemented on the target data sets. To boost the performance of the source DNN model, the architecture of the DNN is chosen from a set of candidate architectures by

n_T	DNN	RECaST GLM	RECaST DNN	Wiens	Unfreeze DNN	TransRF	glmtrans	WDGRL
250	94(1.8)	76(17)	87(11)	72(6.7)	90(6.3)	69(9.5)	95(2.1)	89(11)
100	84(7.2)	77(17)	87(11)	67(7.4)	87(7.9)	43(16)	79(5.6)	87(11)
60	74(10)	78(16.0)	87(11)	64(9)	83(11)	33(19)	66(11)	83(14)
40	68(10)	79(15)	87(11)	64(9.7)	81(13)	24(17)	61(8.4)	87(12)
20	60(9.6)	83(12)	87(12)	65(11)	77(14)	-	55(5.7)	-

Table 9: Out-of-sample AUC (standard error) averaged over 300 source and target data sets when the source and target neural network weight matrices are orthogonal. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

n_T	DNN	RECaST GLM	RECaST DNN	Wiens	Unfreeze DNN	TransRF	glmtrans	WDGRL
250	68(18)	100(0)	98.0(1.8)	77(29)	70(14)	73(9.5)	79(9.7)	65(19)
100	65(14)	100(0)	99.0(2.2)	80(25)	69(15)	58(13)	72(10)	64(17)
60	63(12)	88(3.5)	94.0(6.2)	80(23)	67(15)	57(16)	68(16)	53(17)
40	60(13)	89(9.9)	89(15)	75(20)	63(15)	61(20)	63(18)	67(15)
20	56(12)	81(15)	81(17)	78(22)	61(15)	-	56(15)	-

Table 10: Empirical coverage (standard error) at the 75% nominal level for a binary response, averaged over 300 source and target data sets when the source and target neural network weight matrices are orthogonal. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

maximizing AUC, averaged over 100 of the source data sets; additional details are provided in Appendix D. In Figure 2, we report the empirical coverage and AUC.

Because the real data generating model is unknown we consider two additional target-only models to test for negative transfer. We compare to a GLM and a Gaussian process (GP) trained only on the target data. In this setting, they perform worse than all of the transfer learning methods, with the GLM achieving an AUC of 0.606 and the GP achieving an AUC of 0.512. Plots for the TransRF, glmtrans, and WDGRL methods can be found in Appendix G. The AUCs of glmtrans and WDGRL were 0.68 and 0.708, respectively. RECaST has similar predictive performance to Wiens and WDGRL but without requiring access to the source data, and it outperforms the DNN and Unfreeze DNN approaches. Pairing RECaST with either the logistic regression or DNN source models produced near optimal average AUC, with respect to the average AUC values of 0.704 and 0.708, respectively, for the source logistic regression model and source DNN model. Figure 2 also demonstrates that RECaST generally produces prediction sets that achieve their nominal level of coverage for target test response values, even for non-linear models with non-Gaussian data, whereas the other approaches do not.

In addition to splitting the data into source and target by hospital, we explore making this division based on other features. First, we take the target data to be all patients aged 51 and under. This split resulted in approximately 20% of the patients in the target data and 80% in the source data. Second, we take the target data to be all patients aged 55 and under. This age was chosen because 20% of the patients *that experienced shock* are

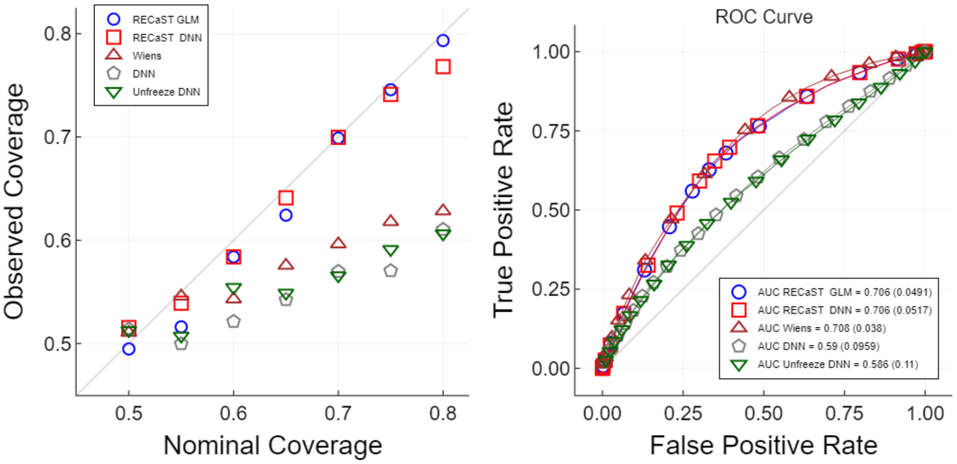

Figure 2: The left panel displays the reliability curve of the nominal versus empirical out-of-sample coverage of prediction sets averaged over 300 target-testing data sets; the right panel reports the out-of-sample receiver operating characteristic (ROC) curve averaged pointwise over 300 target-testing data sets. The legend also reports the AUC (standard error) averaged over the same 300 target-testing data sets. Note that we cut the reliability curve at a nominal coverage of 0.8 because there are very few observations with higher coverage, undermining the reliability of coverage estimation at higher nominal levels.

aged 55 and under. Third, we take the target data to be all female patients, which account for about 45% of the data. This more even split between source and target will be a good test for negative transfer. Finally, we take the target data to be all patients who are not Caucasian, corresponding to roughly 20% of the data.

Table 11 shows the average AUC, AUC standard error, and average empirical coverage at the 80% nominal level summarized over 300 target training and testing data sets. While the standard errors are large, we see that the average AUC of RECaST is larger than that of the target-only methods in all but one setting. The only instance in which RECaST has smaller AUC is when the target data consist of the female patients. This may be due to the similar sample sizes between the source and target for this particular setting, as we demonstrated in the synthetic data simulations that RECaST is most advantageous when the target sample size is small. The RECaST AUCs are within a standard error of Wiens, glmtrans, and WDGRL, but RECaST does not require access to the source data. We see that the empirical coverages for the RECaST method are near the 80% nominal value; the Wiens and glmtrans methods are more conservative when the data are split by age. The TransRF method reports coverage lower than the 80% nominal level in all settings. This analysis demonstrates a general use case for RECaST as a clinical tool across a broad range of scenarios.

8. Concluding Remarks

The RECaST framework is adaptable to virtually any source model that makes predictions, and can accommodate both continuous and binary responses. The source data themselves

	Age $\leq$ 51	Age $\leq$ 55	Female	Non-Caucasian
Target only GLM	71(6.9) [0.74]	71(6.0) [0.73]	70(4.2) [0.74]	67(6.4) [0.72]
Target only GP	66(16) [0.78]	67(13) [0.73]	64(11) [0.67]	61(11) [0.74]
Target only DNN	69(8.1) [0.72]	69(7.0) [0.69]	68(5.2) [0.70]	67(8.4) [0.69]
RECaST GLM	73(6.8) [0.78]	72(6.2) [0.78]	69(4.5) [0.77]	71(6.0) [0.84]
RECaST DNN	73(6.8) [0.82]	72(6.1) [0.78]	69(4.5) [0.79]	71(6.1) [0.83]
Unfreeze DNN	69(9.0) [0.75]	69(7.8) [0.73]	68(5.2) [0.72]	66(8.4) [0.70]
Wiens	73(6.5) [0.85]	73(5.7) [0.87]	70(4.5) [0.79]	71(6.6) [0.87]
glmtrans	71(7.1) [0.85]	71(5.5) [0.84]	70(4.7) [0.76]	66(7.2) [0.82]
TransRF	54(14) [0.69]	59(13) [0.71]	66(7.6) [0.72]	56(12) [0.69]
WDGRL	72(7.2) [0.76]	71(6.7) [0.73]	70(3.8) [0.74]	73(6.4) [0.80]

Table 11: Out-of-sample AUC (standard error) [empirical coverage at the 80% nominal level] averaged over 300 target training and testing data sets for each target data setting of the eICU data. All reported values are multiplied by 100.

are not required, which is a significant advantage when legal or ethical barriers to access of source data sets exist, e.g., due to privacy concerns. Unlike other transfer learning methods, RECaST always provides uncertainty quantification through prediction sets. Our conclusions are supported by both theoretical justifications and performance in simulation studies on synthetic and real data using linear and two-layer neural network source models.

The RECaST framework may be extended in several directions to accommodate the complexity of EHR data. Broadening RECaST to handle differing feature spaces between source and target hospitals would allow for it to be applied in more general settings. As EHR databases are updated, it would be useful to perform online transfer learning. Patient clinical notes are also frequently available in EHR data and have been used by other transfer learning approaches (e.g., Si and Roberts, 2020). However, transfer learning approaches that combine quantitative and text features to create a unified patient representation are currently lacking. Another promising direction is to study RECaST framework formulations for multi-class classification. One such formulation would be to specify the h function in Equation (2) as

$$h\big\{f(\boldsymbol{\theta}_S, \boldsymbol{x}_S), U_S\big\} = \sum_{k=1}^{K} k \cdot \mathbf{1}\big[U_S \in \Delta_k\{f(\boldsymbol{\theta}_S, \boldsymbol{x}_S)\}\big],$$

where K is the number of classes and $U_S \sim \text{Uniform}(\Delta)$ with $\Delta_1, \ldots, \Delta_K$ – all functions of $f(\boldsymbol{\theta}_S, \boldsymbol{x}_S)$ – being triangular regions that form a partition of the simplex Δ over the multi-class outcome space (e.g., see, Jacob et al., 2021; Williams, 2021).

Acknowledgments

The authors thank the reviewers and action editor for their valuable feedback, which led to an improved manuscript. The first author's contribution to this work was funded by the U. S. National Institutes of Health T32 grant HL079896. The last author is partially supported by the U. S. National Science Foundation, grant DMS–2337943. Research reported in this publication was supported by the National Heart, Lung, And Blood Institute of the National Institutes of Health under Award Number R56HL155373. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A. Proofs

Proof [Proof of Lemma 1] It is well-established (see, e.g., Hinkley, 1969) that if $V \sim \mathcal{N}(0, \sigma_V^2)$ and $W \sim \mathcal{N}(0, \sigma_W^2)$ with correlation coefficient ρ, then

$$\frac{V}{W} \sim \text{Cauchy}\Big(\frac{\rho\sigma_V}{\sigma_W}, \frac{\sigma_V}{\sigma_W}\sqrt{1-\rho^2}\Big). \tag{8}$$

Accordingly, since

$$\begin{bmatrix} \boldsymbol{a}^\top \\ \boldsymbol{b}^\top \end{bmatrix} \boldsymbol{x} \sim \mathcal{N}\Bigg(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{a}^\top\boldsymbol{a} & \boldsymbol{a}^\top\boldsymbol{b} \\ \boldsymbol{b}^\top\boldsymbol{a} & \boldsymbol{b}^\top\boldsymbol{b} \end{bmatrix} \Bigg),$$

it follows that $\boldsymbol{x}^\top\boldsymbol{a} \sim \mathcal{N}(0, \boldsymbol{a}^\top\boldsymbol{a})$, $\boldsymbol{x}^\top\boldsymbol{b} \sim \mathcal{N}(0, \boldsymbol{b}^\top\boldsymbol{b})$, and $\rho = (\boldsymbol{a}^\top\boldsymbol{b})/(\|\boldsymbol{b}\|\,\|\boldsymbol{a}\|)$. The result follows from Equation (8) by taking $V = \boldsymbol{x}^\top\boldsymbol{a}$ and $W = \boldsymbol{x}^\top\boldsymbol{b}$. ■

Before proceeding directly to the proof of Lemma 2, the following necessary supporting result is stated and proved.

Lemma 4 *The MLEs of* γ *and* δ *for Equation* (5)*, respectively, are*

$$\begin{aligned} \widehat{\gamma} &= \frac{\sum_{i=1}^{n_T}(v_i - \overline{v})(y_i - \overline{y}_T)}{\widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta}_S \sum_{i=1}^{n_T}(v_i - \overline{v})^2} \quad \textit{and} \\ \widehat{\delta} &= \frac{\overline{y}_T}{\widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta}_S} - \overline{v} \cdot \widehat{\gamma}, \end{aligned}$$

where $v_i = (\beta_i - \delta)/\gamma$ *for* $i \in \{1, \ldots, n_T\}$, $\overline{v} := \sum_{i=1}^{n_T} v_i/n_T$ *and* $\overline{y}_T := \sum_{i=1}^{n_T} y_{T,i}/n_T$.

Proof [Proof of Lemma 4] After the change of variables $v_i = (\beta_i - \delta)/\gamma$ for $i \in \{1, \ldots, n_T\}$, the likelihood function in Equation (5) takes the form

$$\prod_{i=1}^{n_T} \Big[\mathcal{N}\big\{ y_{T,i} \mid (\gamma v_i + \delta)\widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta}_S, \sigma^2 \big\} \cdot \text{Cauchy}(v_i \mid 0, 1) \Big].$$

Taking partial derivatives with respect to δ and γ gives the first-order conditions

$$\begin{aligned} \sum_{i=1}^{n_T} \Big\{ \frac{y_{T,i}}{\widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta}_S} - \gamma v_i - \delta \Big\} &= 0 \\ \sum_{i=1}^{n_T} \Big\{ \frac{y_{T,i}}{\widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta}_S} - \gamma v_i - \delta \Big\} v_i &= 0. \end{aligned}$$

Solving this system yields the MLEs in Lemma 4. ■

Proof [Proof of Lemma 2] With the assumptions that $Y_{T,1}, \dots, Y_{T,n_T} \overset{\text{iid}}{\sim} \mathcal{N}(\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T, \sigma^2)$ independent of $V_1, \dots, V_{n_T} \overset{\text{iid}}{\sim} \text{Cauchy}(0,1)$, first, define the following notations:

$$\boldsymbol{Y} := \begin{pmatrix} Y_{T,1} \\ \vdots \\ Y_{T,n_T} \end{pmatrix}, \quad \overline{\boldsymbol{Y}} := \overline{Y}_T \cdot \mathbf{1}_{n_T}, \quad \overline{Y}_T := \frac{1}{n_T} \sum_{i=1}^{n_T} Y_{T,i},$$

and

$$\boldsymbol{V} := \begin{pmatrix} V_1 \\ \vdots \\ V_{n_T} \end{pmatrix}, \quad \overline{\boldsymbol{V}} := \overline{V} \cdot \mathbf{1}_{n_T}, \quad \overline{V} := \frac{1}{n_T} \sum_{i=1}^{n_T} V_i,$$

where $\mathbf{1}_{n_T}$ is an n_T-dimensional column vector with every component having value 1.

By the Cauchy-Schwarz inequality,

$$|\widehat{\gamma}| = \frac{\left| \sum_{i=1}^{n_T} (V_i - \overline{V})(Y_i - \overline{Y}_T) \right|}{\left| \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S \right| \sum_{i=1}^{n_T} (V_i - \overline{V})^2} \le \frac{\left\| \boldsymbol{Y} - \overline{\boldsymbol{Y}} \right\|_2 \left\| \boldsymbol{V} - \overline{\boldsymbol{V}} \right\|_2}{\left| \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S \right| \left\| \boldsymbol{V} - \overline{\boldsymbol{V}} \right\|_2^2} = \frac{\left\| \boldsymbol{Y} - \overline{\boldsymbol{Y}} \right\|_2}{\left| \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S \right| \left\| \boldsymbol{V} - \overline{\boldsymbol{V}} \right\|_2},$$

where $\|\cdot\|_2$ is the Euclidean norm. We first need to establish the fact that square-root sums of independent, centered, and squared Cauchy random variables grow in value at the rate of at least $n_T^{\alpha + \frac{1}{2}}$ for any $\alpha \in (0, 1/2)$. Accordingly, for any $\varepsilon > 0$ and any $\alpha \in (0, 1/2)$,

$$\begin{aligned}
P\Big(\left\| \boldsymbol{V} - \overline{\boldsymbol{V}} \right\|_2 < n_T^{\alpha + \frac{1}{2}} \varepsilon^{-1} \Big) &= P\Big(\left\| \boldsymbol{V} - \overline{\boldsymbol{V}} \right\|_2^2 < n_T^{2\alpha+1} \varepsilon^{-2} \Big) \\
&= P\Big(\sum_{i=1}^{n_T} V_i^2 - n_T \overline{V}^2 < n_T^{2\alpha+1} \varepsilon^{-2} \Big) \\
&\le P\Big(\sum_{i=1}^{n_T} V_i^2 - n_T^{1+\alpha} < n_T^{2\alpha+1} \varepsilon^{-2} \Big) + P\Big(- n_T \overline{V}^2 < -n_T^{1+\alpha} \Big) \\
&= P\Big(\sum_{i=1}^{n_T} V_i^2 < n_T^{2\alpha+1} \varepsilon^{-2} + n_T^{1+\alpha} \Big) + P\Big(|\overline{V}| > n_T^{\frac{\alpha}{2}} \Big) \\
&\le P\Big(\sum_{i=1}^{n_T} V_i^2 < n_T^{2\alpha+1} \{ \varepsilon^{-2} + 1 \} \Big) + 2F_V\Big(- n_T^{\alpha/2} \Big), \qquad (9)
\end{aligned}$$

where $F_V(\cdot)$ is the Cauchy$(0,1)$ distribution function. The first term vanishes for any $\alpha \in (0, 1/2)$ as $n_T \to \infty$ by Lemma 2.1 in Eicker (1985), and the second term vanishes as $n_T \to \infty$ by the definition of a distribution function.

Next, in order to show the convergence of both MLEs, we need that $n_T^{\alpha/2}\widehat{\gamma} \to 0$ in probability as $n_T \to \infty$. Our argument goes as follows. For any $\varepsilon > 0$ and any $\alpha \in (0, 1/2)$,

$$\begin{aligned}
P\Big(|\widehat{\gamma}| > n_T^{-\alpha/2}\varepsilon\Big) &\leq P\left(\frac{\|\boldsymbol{Y} - \overline{\boldsymbol{Y}}\|_2}{|\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S|\,\|\boldsymbol{V} - \overline{\boldsymbol{V}}\|_2} > \frac{n_T^{(1+\alpha)/2}}{n_T^{(1+\alpha)/2}} \frac{\varepsilon}{n_T^{\alpha/2}}\right) \\
&\leq P\left(\frac{\|\boldsymbol{Y} - \overline{\boldsymbol{Y}}\|_2}{|\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S|} > n_T^{(1+\alpha)/2}\right) + P\left(\frac{1}{\|\boldsymbol{V} - \overline{\boldsymbol{V}}\|_2} > \frac{\varepsilon}{n_T^{\alpha+1/2}}\right) \\
&= P\left(\frac{\|\boldsymbol{Y} - \overline{\boldsymbol{Y}}\|_2^2}{\sigma^2} > \frac{|\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S|^2}{\sigma^2} n_T^{1+\alpha}\right) + P\Big(\|\boldsymbol{V} - \overline{\boldsymbol{V}}\|_2 < n_T^{\alpha+\frac{1}{2}}\varepsilon^{-1}\Big).
\end{aligned}$$

Denoting $S := \|\boldsymbol{Y} - \overline{\boldsymbol{Y}}\|_2^2/\sigma^2 \sim \chi^2_{n_T-1}$, and applying the Chernoff bound to the first quantity in the last expression gives, for any $t < 1/2$,

$$P\left(S > \frac{|\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S|^2}{\sigma^2} n_T^{1+\alpha}\right) \leq (1 - 2t)^{-(n_T - 1)/2} \exp\left\{-\frac{t|\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S|^2}{\sigma^2} n_T^{1+\alpha}\right\}.$$

Choosing $t = 1/4$ yields the bound

$$P\Big(|\widehat{\gamma}| > n_T^{-\alpha/2}\varepsilon\Big) \leq e^{-n_T^{1+\alpha}\cdot\frac{1}{2}\left(\frac{1}{2\sigma^2}|\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S|^2 - n_T^{-\alpha} + n_T^{-1-\alpha}\right)} + P\Big(\|\boldsymbol{V} - \overline{\boldsymbol{V}}\|_2 < n_T^{\alpha+\frac{1}{2}}\varepsilon^{-1}\Big).$$

Thus, by Equation (9), it follows that $n_T^{\alpha/2}\widehat{\gamma} \to 0$ in probability as $n_T \to \infty$. This fact implies that $\widehat{\gamma} \to 0$ in probability as $n_T \to \infty$, and is needed to prove the asymptotic convergence of $\widehat{\delta}$, next.

Since $Y_{T,1}, \ldots, Y_{T,n_T} \overset{\text{iid}}{\sim} \mathcal{N}(\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T, \sigma^2)$, it follows that $\overline{Y}_T = \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T + \sigma n_T^{-\frac{1}{2}} U$, where $U \sim \mathcal{N}(0, 1)$. That being so, for any $\varepsilon > 0$ and any $\alpha \in (0, 1/2)$,

$$\begin{aligned}
P\left(\left|\widehat{\delta} - \frac{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T}{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S}\right| > \varepsilon\right) &= P\left\{\left|\frac{1}{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S}(\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T + \sigma n_T^{-\frac{1}{2}} U) - \overline{V}\widehat{\gamma} - \frac{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T}{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S}\right| > \varepsilon\right\} \\
&= P\left(\left|\frac{\sigma}{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S} n_T^{-\frac{1}{2}} U - \overline{V}\widehat{\gamma}\right| > \varepsilon\right) \\
&\leq P\left(\left|\frac{\sigma}{\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S} n_T^{-\frac{1}{2}} U\right| > \frac{\varepsilon}{2}\right) + P\Big(|\overline{V}\widehat{\gamma}| > \frac{\varepsilon}{2}\Big) \\
&= 2\Phi\Big\{-n_T^{\frac{1}{2}} \cdot \varepsilon|\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S|/(2\sigma)\Big\} + P\Big(|\overline{V}| > n_T^{\alpha/2}/2\Big) + P\Big(|\widehat{\gamma}| > n_T^{-\alpha/2}\varepsilon\Big) \\
&= 2\Phi\Big\{-n_T^{\frac{1}{2}} \cdot \varepsilon|\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S|/(2\sigma)\Big\} + 2F_V\Big(-n_T^{\alpha/2}/2\Big) + P\Big(|\widehat{\gamma}| > n_T^{-\alpha/2}\varepsilon\Big),
\end{aligned}$$

where $\Phi(\cdot)$ is the standard Gaussian distribution function. The first two terms in the last expression vanish by the definition of a distribution function, and the third term vanishes by the same because we previously established that $n_T^{\alpha/2}\widehat{\gamma} \to 0$ in probability as $n_T \to \infty$. Hence, $\widehat{\delta} \to \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T/(\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S)$ in probability as $n_T \to \infty$. ■

Proof [Proof of Theorem 3] Our argument begins with direct evaluation of the probability that $\widetilde{Y}_T \sim \mathcal{N}(\widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T, \sigma^2)$ is contained in the interval $[a^\alpha_{n_T}, b^\alpha_{n_T}]$, and it finishes by applying the result of Lemma 2.

$$\begin{aligned} P\Big(\widetilde{Y}_T \in [a^\alpha_{n_T}, b^\alpha_{n_T}]\Big) &= \int_{a^\alpha_{n_T}}^{b^\alpha_{n_T}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\widetilde{y}_T - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T)^2} d\widetilde{y}_T \\ &= \Phi\Big(\frac{b^\alpha_{n_T} - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T}{\sigma}\Big) - \Phi\Big(\frac{a^\alpha_{n_T} - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T}{\sigma}\Big), \end{aligned}$$

where $\Phi(\cdot)$ is the standard Gaussian distribution function. We will first demonstrate that $\Phi(W) \to 1 - \alpha/2$, with

$$\begin{aligned} W &:= \frac{b^\alpha_{n_T} - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T}{\sigma} \\ &= \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big) + \frac{1}{\sigma}\Big(\widetilde{\beta} \cdot \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T\Big) \\ &\sim \text{Cauchy}\Big\{\Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big) + \frac{1}{\sigma}\Big(\widehat{\delta} \cdot \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T\Big),\ \Big|\frac{\widehat{\gamma}}{\sigma} \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S\Big|\Big\} \end{aligned}$$

since $\widetilde{\beta} \sim \text{Cauchy}(\widehat{\delta}, |\widehat{\gamma}|)$.

For any $\epsilon > 0$,

$$\begin{aligned} P\Big(|\Phi(W) - (1 - \alpha/2)| < \epsilon\Big) &= P\Big(1 - \alpha/2 - \epsilon < \Phi(W) < 1 - \alpha/2 + \epsilon\Big) \\ &= P\Big\{\Phi^{-1}(1 - \alpha/2 - \epsilon) < W < \Phi^{-1}(1 - \alpha/2 + \epsilon)\Big\} \\ &= F_W\Big\{\Phi^{-1}(1 - \alpha/2 + \epsilon)\Big\} - F_W\Big\{\Phi^{-1}(1 - \alpha/2 - \epsilon)\Big\}, \end{aligned}$$

where $F_W(\cdot)$ is the Cauchy distribution function associated with W. Then,

$$F_W\Big\{\Phi^{-1}(1 - \alpha/2 + \epsilon)\Big\} = \frac{1}{2} + \frac{1}{\pi} \arctan\left\{\frac{c_1 - (\widehat{\delta} \cdot \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T)/\sigma}{|\widehat{\gamma} \cdot \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S|/\sigma}\right\},$$

with $c_1 := \Phi^{-1}(1 - \alpha/2 + \epsilon) - \Phi^{-1}(1 - \alpha/2) > 0$, and similarly,

$$F_W\Big\{\Phi^{-1}(1 - \alpha/2 - \epsilon)\Big\} = \frac{1}{2} + \frac{1}{\pi} \arctan\left\{\frac{c_2 - (\widehat{\delta} \cdot \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T)/\sigma}{|\widehat{\gamma} \cdot \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_S|/\sigma}\right\},$$

with $c_2 := \Phi^{-1}(1 - \alpha/2 - \epsilon) - \Phi^{-1}(1 - \alpha/2) < 0$. Accordingly, it follows by Lemma 2 that

$$F_W\Big\{\Phi^{-1}(1 - \alpha/2 + \epsilon)\Big\} \longrightarrow 1 \quad \text{and} \quad F_W\Big\{\Phi^{-1}(1 - \alpha/2 - \epsilon)\Big\} \longrightarrow 0$$

in probability as $n_T \to \infty$, and so

$$\Phi\Big(\frac{b^\alpha_{n_T} - \widetilde{\boldsymbol{x}}^\top \boldsymbol{\theta}_T}{\sigma}\Big) = \Phi(W) \longrightarrow 1 - \alpha/2$$

in probability as $n_T \to \infty$. A similar argument shows that

$$\Phi\Big(\frac{a^{\alpha}_{n_T} - \widetilde{\boldsymbol{x}}^{\top}\boldsymbol{\theta}_T}{\sigma}\Big) \longrightarrow \alpha/2,$$

in probability as $n_T \to \infty$, concluding the proof. ■

Appendix B. Bounding Continuous Integral

Recall the posterior distribution of the calibration parameters for the continuous response setting,

$$\begin{aligned}&\pi\big(\delta, \gamma, \sigma \mid y_{T,1}, \ldots, y_{T,n_T}, \widehat{\boldsymbol{\theta}}_S\big)\\ &\quad= \pi(\delta,\gamma,\sigma) \cdot \prod_{i=1}^{n_T} \int_{\mathbb{R}} \frac{\mathrm{Cauchy}(\beta_i \mid \delta, \gamma)}{\mid f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i}) \mid} \cdot \mathcal{N}\left\{\beta_i \;\middle|\; \frac{y_{T,i}}{f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})}, \frac{\sigma^2}{f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})^2}\right\} d\beta_i.\end{aligned}$$

Calculating this posterior requires the evaluation of n_T integrals over $\mathbb{R}$. For computational efficiency, we estimate the posterior by integrating over closed intervals. The incurred numerical error can be tuned to be lower than computer precision.

Performing the substitution $u_i = \big\{\beta_i - y_{T,i}/f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})\big\}/\big\{\sigma/|f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})|\big\}$ re-expresses the ith integral as

$$\begin{aligned}&\int_{\mathbb{R}} \frac{\mathcal{N}(u_i \mid 0,1)}{\sigma} \cdot \mathrm{Cauchy}\left[u_i \;\middle|\; \frac{|f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})|}{\sigma}\left\{\delta - \frac{y_{T,i}}{f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})}\right\}, \frac{|f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})|\gamma}{\sigma}\right] du_i\\ &\leq \frac{1}{\sigma}\Bigg(\int_{s_1}^{s_2} \mathcal{N}(u_i \mid 0,1) \cdot \mathrm{Cauchy}\left[u_i \;\middle|\; \frac{|f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})|}{\sigma}\left\{\delta - \frac{y_{T,i}}{f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})}\right\}, \frac{|f(\widehat{\boldsymbol{\theta}}_S, \boldsymbol{x}_{T,i})|\gamma}{\sigma}\right] du_i\\ &\qquad + \phi(s_1) + \phi(s_2)\Bigg),\end{aligned}$$

for any s_1 and s_2 satisfying $s_1 < 0 < s_2$, where $\phi(\cdot)$ is the standard Gaussian density function. Then choose s_1 and s_2 so that $\phi(s_1) + \phi(s_2)$ is as small as desired. For example, we set $s_1 = -39$ and $s_2 = 39$, giving $\phi(s_1)$ and $\phi(s_2)$ numerically equal to zero in the base `Julia` software (for comparison, $\phi(38) = 1.097 \times 10^{-314}$).

Appendix C. MCMC Implementation Details

Sections 4.2 and 5.2 detail the procedure for sampling from the posterior predictive distribution of a new observation. RECaST first estimates the joint posterior density of the re-calibration parameters (δ, γ, σ) in the linear model and (δ, γ) in the logistic model. We specify disperse priors $\delta \sim \mathcal{N}(1, 400)$, $\log(\gamma) \sim \mathcal{N}(0, 9)$, and in the continuous setting $\log(\sigma^2) \sim \mathcal{N}(0, 9)$. We run the Metropolis-Hastings estimation algorithm of the posterior distribution for 100,000 iterations with the initial 20,000 iterations used as a burn-in period to tune the proposal variance. The parameters from the final 50,000 iterations are used as the posterior distribution. Finally, $n_{\text{post}} = 300$ equally spaced triplets/pairs of this distribution are taken as a posterior sample to be used in the posterior predictive estimation, which we denote by $\{\delta_i, \gamma_i, \sigma_i\}_{i=1}^{300}$ and $\{\delta_i, \gamma_i\}_{i=1}^{300}$ in the linear and logistic models respectively. For each triplet/pair, a sample of $n_\beta = 300$ β's are taken from the Cauchy distribution, each used to generate $n_Y = 300$ samples from the posterior predictive distribution. This gives $300 \times 300 \times 300 = 27,000,000$ posterior predictive observations for each out-of-sample test point, $(Y_{T,\text{test}}, \widetilde{\boldsymbol{x}}_T)$.

Appendix D. Neural Network Training Procedure

The following procedure is used to train all neural networks considered: the source DNN, the DNN trained only on target data, and the Unfreeze DNN.

We initialize the weights using Xavier initialization (Glorot and Bengio, 2010). The network is trained for 2500 epochs using the ADAM optimizer and an MSE loss. A portion of the training data is set aside as an out-of-sample calibration set during training. At each epoch, the training and calibration loss are tracked. The final parameterization used is taken from the epoch with the lowest calibration loss to avoid overfitting.

The candidate architectures ranged from networks with 316 parameters to 11,641 parameters with varied number of layers, layer sizes, activation functions, and dropout proportions. The architecture described below was chosen as it had the best test set AUC on the eICU data of all considered architectures. We use a two layer neural network with layer sizes $\ell_1 = (p, 25)$ and $\ell_2 = (25, 1)$. These layers are connected with a Rectified Linear Unit (ReLU) activation function. In the binary response setting, the output of ℓ_2 is converted to a probability through a softmax activation function. For consistency, this architecture is also used for the simulated data analysis in Section 6.

The source neural network for RECaST learns parameters in both layers using only source data. The DNN network learns parameters in both layers using only the target data. The Unfreeze DNN network learns parameters in both layers first using only the source data. Then, the target data are processed through the same neural network, re-training parameters in the second layer and leaving the first layer unchanged from the values learned on the source data set.

Appendix E. Additional Tables for Section 6.4

Tables 12 and 13 present the coverage results for the simulations in Section 6.4.

Appendix F. Additional Robustness Results

Here, we consider the case where $\boldsymbol{\theta}_S$ and $\boldsymbol{\theta}_T$ are orthogonal. We take $\boldsymbol{\theta}_S$ to be the same $p = 50$ feature vector as in Section 6.1 and take $\boldsymbol{\theta}_T$ to be a vector in the null space of $\boldsymbol{\theta}_S$. The data are otherwise generated following Section 6.1 from a linear or logistic regression with a fixed source sample size of $n_S = 1000$ and a varying target sample size of $n_T \in \{20, 40, 60, 100, 250\}$.

In the continuous outcome case, we compare RECaST to the DNN, Unfreeze DNN, TransRF and glmtrans approaches. In the binary outcome case, we compare RECaST to

n_T	RECaST LM	RECaST DNN
250	100(0.4)	100(0.28)
100	100(0.33)	100(0.33)
60	100(0.35)	100(0.32)
40	100(0.29)	100(0.32)
20	100(0.37)	100(0.37)

Table 12: Empirical coverage (standard error) at the 95% nominal level for a continuous response, averaged over 300 source and target data sets when the when the target model parameters are generated as $\boldsymbol{\theta}_T = \boldsymbol{\theta}_S + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}_{p\times 10+10\times 1}(\mathbf{0}, 0.025\boldsymbol{I})$. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

n_T	RECaST LM	RECaST DNN
250	100(0.53)	100(0.24)
100	100(0.42)	100(0.68)
60	100(0.36)	100(0.3)
40	100(0.3)	100(0.93)
20	100(0.26)	100(0.25)

Table 13: Empirical coverage (standard error) at the 95% nominal level for a continuous response, averaged over 300 source and target data sets when the source and target neural network weight matrices are orthogonal. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

the target-only DNN, Unfreeze DNN, TransRF, glmtrans, WDGRL and Wiens methods. Table 14 shows the predictive performance of RECaST for the orthogonally misaligned source and target setting with a continuous response.

n_T	LM	DNN	RECaST LM	RECaST DNN	Unfreeze DNN	TransRF	glmtrans	MTL FO	MTL MoM
250	0.56(0.03)	0.85(0.061)	1.1(0.051)	1.1(0.051)	0.92(0.082)	0.63(0.066)	0.54(0.025)	0.6(0.03)	0.6(0.03)
100	0.71(0.06)	1.1(0.12)	1.1(0.050)	1.1(0.051)	1.1(0.19)	0.97(1.1)	0.55(0.031)	0.76(0.06)	0.76(0.06)
60	1.3(0.28)	1.3(0.13)	1.1(0.05)	1.1(0.051)	1.2(0.28)	1.8(2.0)	0.57(0.043)	1.3(0.26)	1.3(0.26)
40	1.2(0.22)	1.3(0.13)	1.1(0.056)	1.1(0.056)	1.5(0.40)	3.0(6.2)	0.6(0.066)	3.2(0.71)	1.4(0.29)
20	1.0(0.07)	1.4(0.14)	1.2(0.078)	1.2(0.073)	1.9(0.59)	-	0.66(0.08)	5.5(0.81)	1.0(0.07)

Table 14: Out of sample RMSE (standard error) averaged over 300 source and target data sets when the source data generating parameters are orthogonal to the target data generating parameters. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

When the target data are plentiful ($n_T = 100, 250$) the RMSE for the LM built solely on the target data outperforms the RECaST methods. This aligns with previous results where there is a large amount of data and a large discrepancy between the source and target (i.e., when transfer learning is not appropriate). As the number of target data points decreases, RECaST outperforms target-only DNN. These results further demonstrate the robustness of RECaST to negative transfer. Notice that glmtrans is also robust; in each of these scenarios glmtrans opted to not use the source data. The MTL MoM method also provides good predictive performance for all sample sizes without requiring access to the source data. Similar to the previous robustness tests, Table 15 shows that RECaST provides conservative predictive intervals resulting in over-coverage at the 95% level.

n_T	RECaST LM	RECaST DNN
250	100(0)	100(0)
100	100(0)	100(0)
60	100(0)	100(0)
40	100(0)	100(0)
20	100(0.024)	100(0.024)

Table 15: Empirical coverage (standard error) at the 95% nominal level, averaged over 300 source and target data sets when the source data generating parameters are orthogonal to the target data generating parameters. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

For a binary response, Table 16 shows the difficulty of this problems. All methods have very low AUCs, including the target-only DNN. For large sample sizes, glmtrans performs relatively well, again due to its ability to ignore the source data entirely and because the model matches the data generating mechanism. Table 17 shows that all methods have predictive coverages with very high standard errors, again displaying the difficulty of this problem.

Next we consider a setting in which the source feature space is a subset of the target feature space: $\mathcal{X}_S \subsetneq \mathcal{X}_T$. We assign 12 features to the true target data $\boldsymbol{X}_T$ but only 9 features to the true source data $\boldsymbol{X}_S$. The parameters are generated as $\boldsymbol{\theta}_T = (-\boldsymbol{a}, \boldsymbol{b})$ where $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^6$ have components independently sampled from Uniform$(0.75, 5)$, and $\boldsymbol{\theta}_S = [\theta_{T,1}, \ldots, \theta_{T,9}]$, the first nine components of $\boldsymbol{\theta}_T$. The responses, $\boldsymbol{Y}_S$ and $\boldsymbol{Y}_T$, are generated via linear or logistic regression with their respective feature vectors.

Table 18 shows that for a continuous response, every method has similar predictive performance when the target sample size is large. As the target sample size decreases, RECaST and glmtrans have the best performance, maintaining a stable RMSE value and outperforming the target-only DNN. This shows that RECaST is robust to negative transfer in this setting.

n_T	DNN	RECaST GLM	RECaST DNN	Wiens	Unfreeze DNN	TransRF	glmtrans	WDGRL
250	59(5.6)	51(3.1)	50(3.7)	49(3.6)	55(4.7)	57(9.1)	72(3.5)	49(3.7)
100	54(5.3)	50(3.5)	49(3.4)	45(3.7)	53(4.7)	37(14)	69(7.1)	50(3.5)
60	52(5.3)	50(3.8)	50(4.2)	43(3.4)	51(4.7)	25(16)	61(10)	48(3.6)
40	52(4.9)	50(3.5)	50(3.8)	42(2.9)	52(4)	23(16)	59(9.1)	49(4)
20	52(4.8)	50(3.5)	50(3.4)	41(4.1)	51(4.5)	-	56(9.5)	-

Table 16: Out-of-sample AUC (standard error) averaged over 300 source and target data sets when the source and target model parameter vectors are orthogonal. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

n_T	DNN	RECaST GLM	RECaST DNN	Wiens	Unfreeze DNN	TransRF	glmtrans	WDGRL
250	59(11)	100(0)	100(0)	53(10)	55(10)	61(13)	81(11)	52(17)
100	55(11)	83(29)	100(0)	52(11)	52(11)	54(13)	77(15)	47(18)
60	50(11)	89(19)	51(10)	48(11)	49(12)	52(15)	74(19)	47(17)
40	52(11)	45(16)	77(27)	47(11)	52(11)	52(27)	67(19)	48(10)
20	51(11)	60(27)	57(25)	51(12)	51(14)	-	59(13)	-

Table 17: Empirical coverage (standard error) at the 75% nominal level, averaged over 300 source and target data sets when the source and target model parameter vectors are orthogonal. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

Table 19 shows RECaST again provides conservative predictive intervals at the 95% level. We see similar results for a binary response outcome in Table 20. Both RECaST methods have stable AUCs as the target sample size decreases, outperforming the target-only DNN and the other transfer learning methods. WDGRL and glmtrans also perform well for larger sample sizes, but both require access to the source data while training. Table 21 shows that the RECaST and Wiens methods again provide conservative predictive coverage for all target sample sizes. WDGRL and glmtrans under-cover in some scenarios.

n_T	LM	DNN	RECaST LM	RECaST DNN	Unfreeze DNN	TransRF	glmtrans	MTL FO	MTL MoM
250	5.5(1.3)	5.8(1.4)	5.4(1.3)	5.4(1.3)	5.52(1.36)	6.5(1.5)	5.3(1.2)	6.3(1.2)	6.3(1.2)
100	5.7(1.3)	6.3(1.5)	5.4(1.3)	5.4(1.3)	5.68(1.35)	11.0(14.0)	5.4(1.2)	6.5(1.2)	6.5(1.2)
60	5.8(1.4)	6.8(1.7)	5.4(1.3)	5.4(1.3)	5.94(1.53)	14.0(14.0)	5.3(1.2)	6.7(1.3)	6.7(1.3)
40	6.2(1.5)	7.2(1.8)	5.4(1.4)	5.4(1.4)	6.06(1.76)	30.0(43.0)	5.4(1.3)	6.9(1.4)	6.9(1.4)
20	7.3(2.1)	8.2(1.9)	5.5(1.3)	5.5(1.4)	6.63(1.93)	-	5.6(1.5)	8.1(1.9)	8.1(1.9)

Table 18: The reported values are: average out-of-sample RMSE (standard deviation). These summaries are over all 300 different source and target data sets for each target sample size when the target data had more features than the source.

n_T	RECaST LM	RECaST DNN
250	100(0)	100(0)
100	100(0)	100(0)
60	100(0)	100(0)
40	100(0)	100(0)
20	100(0.022)	100(0.2)

Table 19: Empirical coverage (standard error) at the 95% nominal level, averaged over 300 source and target data sets when the target data had more features than the source. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

n_T	DNN	RECaST GLM	RECaST DNN	Wiens	Unfreeze DNN	TransRF	glmtrans	WDGRL
250	86(5.3)	89(4.5)	89(4.5)	74(7.9)	89(3.9)	60(16)	90(5.9)	89(6.5)
100	83(10)	89(4.4)	89(4.3)	75 (8.0)	84(10)	46(18)	88(4.1)	89(4.2)
60	75(13)	89(4.2)	89(4)	74(7.2)	79(12)	38(21)	86(3.8)	90(3.4)
40	73(11)	89(4.3)	89(4.3)	73(7.3)	78(14)	32(12)	82(9.2)	89(4.0)
20	75(7.9)	88(4.5)	88(4.3)	74(6.6)	81(9.6)	-	69(13)	-

Table 20: Out-of-sample AUC (standard error) averaged over 300 source and target data sets when the target data had more features than the source. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

n_T	DNN	RECaST GLM	RECaST DNN	Wiens	Unfreeze DNN	TransRF	glmtrans	WDGRL
250	70(14)	100(0)	98(1.6)	87(11)	75(13)	73(14)	71(12)	63(15)
100	65(7.6)	96(0)	95(7.3)	85(8.9)	73(12)	64(15)	72(11)	64(13)
60	67(16)	89(6.5)	90(12)	86(8.2)	70(14)	58(22)	79(13)	69(9.9)
40	60(11)	86(16)	88(11)	86(8.2)	72(13)	53(17)	74(15)	63(17)
20	59(12)	82(17)	80(19)	86(9.8)	67(19)	-	70(16)	-

Table 21: Empirical coverage (standard error) at the 75% nominal level, averaged over 300 source and target data sets when the target data had more features than the source. The out-of-sample test sets each contain 250 observations. All reported values are multiplied by 100.

Appendix G. Comparative eICU Results

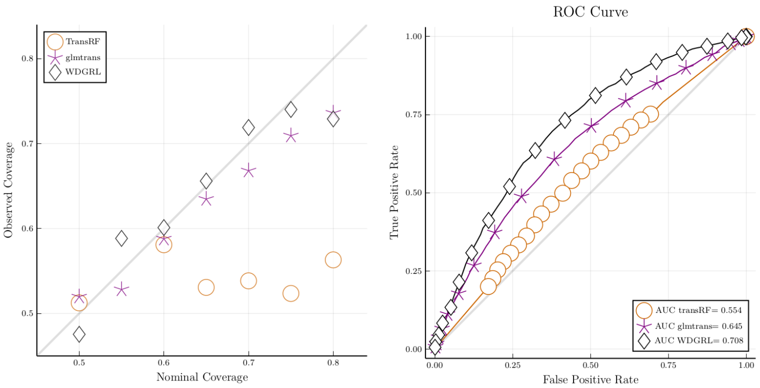

Figure 3: Results for TransRF, glmtrans, and WDGRL on the eICU data set. The left panel displays the reliability curve of the nominal versus empirical out-of-sample coverage of prediction sets averaged over 300 target-testing data sets; the right panel reports the out-of-sample receiver operating characteristic (ROC) curve averaged pointwise over 300 target-testing data sets. The legend also reports the AUC (standard error) averaged over the same 300 target-testing data sets. Note that we cut the reliability curve at a nominal coverage of 0.8 because there are very few observations with higher coverage, undermining the reliability of coverage estimation at higher nominal levels.

Appendix H. eICU Feature Descriptions

Variable	Description
Age	Age in years
Gender	Gender as either Male, Female, Unknown or Other
Ethnicity	Ethnicity as either Asian, Caucasian, African American, Native American, Hispanic or Other/Unknown
Weight	Weight upon admission
Temperature	Worst temperature measured from a midpoint of 38°C
White blood cell count	Worst white blood cell count from a midpoint of 11,500 white blood cells per microliter
Respiratory rate	Worst respiratory rate from a midpoint of 19 breaths per minute
Heart rate	Worst heart rate from a midpoint of 75 beats per minute
Hematocrit level	Worst hematocrit from a midpoint of 45.5%
Creatinine level	Worst serum creatinine from a midpoint of 1.0 milligrams per deciliter
Glucose level	Worst glucose from a midpoint of 130 milligrams per deciliter
Oxygen saturation	Oxygen saturation in the blood measured by a pulse oximeter
Dialysis	An indicator reporting if the patient is on dialysis
Intubated	An indicator reporting if the patient was intubated during the worst measurement of their arterial blood gas
Ventilated	Binary an indicator reporting if the patient was ventilated during the measurement worst respiratory rate
Eye	Eye score ranging from 1 to 4 on the Glasgow Coma Scale
Motor	Motor score ranging from 1 to 6 on the Glasgow Coma Scale
Verbal	Verbal score ranging from 1 to 3 on the Glasgow Coma Scale

Table 22: Descriptions of the features from the eICU Collaborative Research Database used in the shock data analysis.

References

Mohamed A. Abba, Jonathan P. Williams, and Brian J. Reich. A penalized complexity prior for deep Bayesian transfer learning with application to materials informatics. *Annals of Applied Statistics*, 17(4):3241 – 3256, 2023. doi: 10.1214/23-AOAS1759. URL https://doi.org/10.1214/23-AOAS1759.

Emmanuel Ahishakiye, Martin Bastiaan Van Gijzen, Julius Tumwiine, Ruth Wario, and Johnes Obungoloch. A survey on deep learning in medical image reconstruction. *Intelligent Medicine*, 1(3):118–127, 2021. ISSN 2667-1026. doi: https://doi.org/10.1016/j.imed.2021.03.003. URL https://www.sciencedirect.com/science/article/pii/S2667102621000061.

Yuichi Akaoka, Kazuki Okamura, and Yoshiki Otobe. Limit theorems for quasi-arithmetic means of random variables with applications to point estimations for the Cauchy distribution. *Brazilian Journal of Probability and Statistics*, 36(2):385 – 407, 2022. doi: 10.1214/22-BJPS531. URL `https://doi.org/10.1214/22-BJPS531`.

Jonathan Baxter. *Theoretical Models of Learning to Learn*, pages 71–94. Springer US, Boston, MA, 1998. ISBN 978-1-4615-5529-2. doi: 10.1007/978-1-4615-5529-2_4. URL `https://doi.org/10.1007/978-1-4615-5529-2_4`.

Angel Bueno, Carmen Benítez, Silvio De Angelis, Alejandro Díaz Moreno, and Jesús M. Ibáñez. Volcano-seismic transfer learning and uncertainty quantification with Bayesian neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2):892–902, 2020. doi: 10.1109/TGRS.2019.2941494.

T. Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Annals of Statistics*, 49(1):100 – 128, 2021. doi: 10.1214/20-AOS1949. URL `https://doi.org/10.1214/20-AOS1949`.

Rohitash Chandra and Arpit Kapoor. Bayesian neural multi-source transfer learning. *Neurocomputing*, 378:54–64, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.10.042. URL `https://www.sciencedirect.com/science/article/pii/S0925231219314213`.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 193–200, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933. doi: 10.1145/1273496.1273521. URL `https://doi.org/10.1145/1273496.1273521`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Thomas Desautels, Jacob Calvert, Jana Hoffman, Qingqing Mao, Melissa Jay, Grant Fletcher, Chris Barton, Uli Chettipally, Yaniv Kerem, and Ritankar Das. Using transfer learning for improved mortality prediction in a data-scarce hospital setting. *Biomedical Informatics Insights*, 9, 2017.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 647–655, Bejing, China, 22–24 Jun 2014. PMLR. URL `https://proceedings.mlr.press/v32/donahue14.html`.

Parijat Dube, Bishwaranjan Bhattacharjee, Elisabeth Petit-Bois, and Matthew Hill. *Improving Transferability of Deep Neural Networks*, pages 51–64. Springer International Publishing, Cham, 2020. ISBN 978-3-030-30671-7. doi: 10.1007/978-3-030-30671-7_4. URL `https://doi.org/10.1007/978-3-030-30671-7_4`.

F. Eicker. Sums of independent squared Cauchy variables grow quadratically: Applications. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 47(1):133–140, 1985. ISSN 0581572X. URL `http://www.jstor.org/stable/25050525`.

Sándor Fegyverneki. A simple robust estimation for parameters of cauchy distribution. *Miskolc Math. Notes*, 14(3):887–892, 2013.

Yoav Freund and Robert Schapire. A short introduction to boosting. *Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

Yan Gao and Yan Cui. Multi-ethnic survival analysis: Transfer learning with Cox neural networks. In *Survival Prediction-Algorithms, Challenges and Applications*, pages 252–257. PMLR, 2021.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

Jen J. Gong, Thoralf M. Sundt, James D. Rawn, and John V. Guttag. Instance weighting for patient-specific risk stratification models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 369–378, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783397. URL https://doi.org/10.1145/2783258.2783397.

Norberto A. Goussies, Sebastián Ubalde, and Marta Mejail. Transfer learning decision forests for gesture recognition. *Journal of Machine Learning Research*, 15(113):3847–3870, 2014. URL http://jmlr.org/papers/v15/goussies14a.html.

Tian Gu, Yi Han, and Rui Duan. A transfer learning approach based on random forest with application to breast cancer prediction in underrepresented populations. In *Pacific Symposium on Biocomuting 2023*, pages 186–197. World Scientific, 2022.

Tian Gu, Phil H. Lee, and Rui Duan. Commute: Communication-efficient transfer learning for multi-site risk prediction. *Journal of Biomedical Informatics*, 137:104243, 2023. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2022.104243. URL https://www.sciencedirect.com/science/article/pii/S1532046422002489.

Emily C. Hector and Ryan Martin. Turning the information-sharing dial: Efficient inference from different data sources. *Electronic Journal of Statistics*, 18(2):2974 – 3020, 2024. doi: 10.1214/24-EJS2265. URL https://doi.org/10.1214/24-EJS2265.

D. V. Hinkley. On the ratio of two correlated normal random variables. *Biometrika*, 56 (3):635–639, 12 1969. ISSN 0006-3444. doi: 10.1093/biomet/56.3.635. URL https://doi.org/10.1093/biomet/56.3.635.

Pierre E. Jacob, Ruobin Gong, Paul T. Edlefsen, and Arthur P. Dempster. A Gibbs sampler for a class of random convex polytopes. *Journal of the American Statistical Association*, 116(535):1181–1192, 2021. doi: 10.1080/01621459.2021.1881523. URL https://doi.org/10.1080/01621459.2021.1881523. PMID: 35340357.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 200–209, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Steven G. Johnson. QuadGK.jl: Gauss–Kronrod integration in Julia. `https://github.com/JuliaMath/QuadGK.jl`, 2013.

O. Y. Kravchuk and P. K. Pollett. Hodges-Lehmann scale estimator for Cauchy distribution. *Communications in Statistics - Theory and Methods*, 41(20):3621–3632, 2012. doi: 10.1080/03610926.2011.563016. URL `https://doi.org/10.1080/03610926.2011.563016`.

Gyemin Lee, Ilan Rubinfeld, and Zeeshan Syed. Adapting surgical models to individual hospitals using transfer learning. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 57–63, 2012. doi: 10.1109/ICDMW.2012.93.

Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–22, 2024. doi: 10.1109/TPAMI.2024.3370978.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 84(1):149—173, February 2022. ISSN 1369-7412. doi: 10.1111/rssb.12479. URL `https://doi.org/10.1111/rssb.12479`.

Sai Li, Tianxi Cai, and Rui Duan. Targeting underrepresented populations in precision medicine: A federated transfer learning approach. *Annals of Applied Statistics*, 17(4): 2970–2992, 2023.

Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, and Guangquan Zhang. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80:14–23, 2015.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17, 05 2015.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Rahul Paul, Samuel H. Hawkins, Yoganand Balagurunathan, Matthew B. Schabath, Robert James Gillies, Lawrence O. Hall, and Dmitry B. Goldgof. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography*, 2:388 – 395, 2016.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):1–13, 2018.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf`.

Rajat Raina, Andrew Y. Ng, and Daphne Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 713–720, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143934. URL `https://doi.org/10.1145/1143844.1143934`.

Henry W. J. Reeve, Timothy I. Cannings, and Richard J. Samworth. Adaptive transfer learning. *Annals of Statistics*, 49(6):3618 – 3649, 2021. doi: 10.1214/21-AOS2102. URL `https://doi.org/10.1214/21-AOS2102`.

Subhankar Roy, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, and Arno Solin. Uncertainty-guided source-free domain adaptation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 537–555, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19806-9.

S. Schuster. Parameter estimation for the cauchy distribution. In *2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 350–353, 2012.

Siyu Shao, Stephen McAleer, Ruqiang Yan, and Pierre Baldi. Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Transactions on Industrial Informatics*, 15 (4):2446–2455, 2019. doi: 10.1109/TII.2018.2864759.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11784. URL `https://ojs.aaai.org/index.php/AAAI/article/view/11784`.

Benjamin Shickel, Anis Davoudi, Tezcan Ozrazgat-Baslanti, Matthew Ruppert, Azra Bihorac, and Parisa Rashidi. Deep multi-modal transfer learning for augmented patient acuity assessment in the intelligent icu. *Frontiers in Digital Health*, 3, 2021. ISSN 2673-253X. doi: 10.3389/fdgth.2021.640685. URL `https://www.frontiersin.org/articles/10.3389/fdgth.2021.640685`.

Yuqi Si and Kirk Roberts. Patient representation transfer learning from clinical notes based on hierarchical attention network. *AMIA Summits on Translational Science Proceedings*, 2020:597, 2020.

Petar Stojanov, Mingming Gong, Jaime Carbonell, and Kun Zhang. Low-dimensional density ratio estimation for covariate shift correction. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3449–3458. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/stojanov19a.html.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 0(0):1–14, 2022. doi: 10.1080/01621459.2022.2071278. URL https://doi.org/10.1080/01621459.2022.2071278.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

Vladimir Vapnik. Transductive inference and semi-supervised learning. In *Semi-Supervised Learning*. IEEE, 2009.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4):699–706, 2014.

Jonathan P Williams. Discussion of "a Gibbs sampler for a class of random convex polytopes". *Journal of the American Statistical Association*, 116(535):1198–1200, 2021. doi: 10.1080/01621459.2021.1946405. URL https://doi.org/10.1080/01621459.2021.1946405.

Jennifer Wohlert, Andreas Munk, Sarthak Sengupta, and Felix Laumann. Bayesian transfer learning for deep networks. *viXra*, 2018.

Qingyao Wu, Hanrui Wu, Xiaoming Zhou, Mingkui Tan, Yonghui Xu, Yuguang Yan, and Tianyong Hao. Online transfer learning with multiple homogeneous or heterogeneous sources. *IEEE Transactions on Knowledge and Data Engineering*, 29(7):1494–1507, 2017. doi: 10.1109/TKDE.2017.2685597.

Han Yang, Shuangjian Jiao, and Peng Sun. Bayesian-convolutional neural network model transfer learning for image detection of concrete water-binder ratio. *IEEE Access*, 8: 35350–35367, 2020. doi: 10.1109/ACCESS.2020.2975350.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 11 2017. doi: 10.1109/TPAMI.2018.2889774.

Jin Zhang. A highly efficient L-estimator for the location parameter of the Cauchy distribution. *Computational statistics*, 25(1):97–105, 2010.

Peilin Zhao, Steven C.H. Hoi, Jialei Wang, and Bin Li. Online transfer learning. *Artificial Intelligence*, 216:76–102, 2014. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2014.06.003. URL https://www.sciencedirect.com/science/article/pii/S0004370214000800.

Changsheng Zhou, Jiangshe Zhang, Junmin Liu, Chunxia Zhang, Guang Shi, and Junying Hu. Bayesian transfer learning for object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):7705–7719, 2020. doi: 10.1109/TGRS.2020.2983201.