Biometrics, 2024, 80(4), ujae144
https://doi.org/10.1093/biomtc/ujae144
Biometric Methodology

Robust and flexible learning of a high-dimensional classification rule using auxiliary outcomes

Muxuan Liang [1,*], Jaeyoung Park [2], Qing Lu [1], Xiang Zhong [3]

[1]Department of Biostatistics, University of Florida, Gainesville, FL 32611, United States, [2]School of Global Health Management and Informatics, University of Central Florida, Orlando, FL 32816, United States, [3]Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, United States

[*]Corresponding author: Muxuan Liang, Department of Biostatistics, University of Florida, 2004 Mowry Road, 5th Floor CTRB, Gainesville, FL 32611, USA (muxuan.liang@ufl.edu).

ABSTRACT

Correlated outcomes are common in many practical problems. In some settings, one outcome is of particular interest, and others are auxiliary. To leverage information shared by all the outcomes, traditional multi-task learning (MTL) minimizes an averaged loss function over all the outcomes, which may lead to biased estimation for the target outcome, especially when the MTL model is misspecified. In this work, based on a decomposition of estimation bias into two types, within-subspace and against-subspace, we develop a robust transfer learning approach to estimating a high-dimensional linear decision rule for the outcome of interest with the presence of auxiliary outcomes. The proposed method includes an MTL step using all outcomes to gain efficiency and a subsequent calibration step using only the outcome of interest to correct both types of biases. We show that the final estimator can achieve a lower estimation error than the one using only the single outcome of interest. Simulations and real data analysis are conducted to justify the superiority of the proposed method.

KEYWORDS: auxiliary outcomes; classification; high-dimensional data; multi-task learning; transfer learning.

1 INTRODUCTION

With the adoption of electronic health records and medical information systems, datasets increasingly massive in volume and diverse in variable categories have been leveraged for knowledge discovery and clinical decision support. In some datasets, in addition to the patient outcome of primary interest, multiple relevant health outcomes are available. In this paper, we denote these relevant outcomes as auxiliary outcomes, and our goal is to study how to safely use these auxiliary outcomes to help predict a binary target outcome in a classification framework with high-dimensional linear decision rules.

Our motivating example is predicting whether the improvement in hip joint functions fails to achieve the minimal clinical importance difference (MCID) after total hip arthroplasty (THA). The Hip disability and Osteoarthritis Outcome Scores for Joint Replacement (HOOS JR) survey is a commonly used instrument to measure THA patients' health outcomes. Predicting whether the change of the overall score measured in preoperative and postoperative surveys exceeds the MCID can help inform whether surgery is necessary. However, it is a challenging task due to the large variability of the reported outcomes and the high event rates of achieving MCID (ie, imbalanced labels). In addition to the overall score, the questionnaire also collects disease-specific information that measures the improvement in various perspectives, including pain, sleep, fatigue, and function (Katakam et al., 2022; Kunze et al., 2020). This motivates us to explore, whether we can leverage these related questionnaire items–auxiliary outcomes to facilitate target outcome prediction (ie, whether the overall score change exceeds MCID).

To model related outcomes jointly, multi-task learning (MTL) has emerged, aiming to exploit commonalities and differences across outcomes (Caruana, 1997). In MTL, it is typically assumed that some parameters are similar across tasks (Ando and Zhang, 2005; Argyriou et al., 2007; Bakker and Heskes, 2003; Maurer et al., 2013; Titsias and Lázaro-Gredilla, 2011; Yu et al., 2005; Zhang et al., 2008; Zhu et al., 2011), or these tasks bear a shared sparsity structure (Gong et al., 2013; 2014; Hernández-Lobato and Hernández-Lobato, 2013; Lounici et al., 2009; Obozinski et al., 2008; Rao et al., 2013; Wang et al., 2016; Yang et al., 2009). Subsequently, a common feature representation can be learned through MTL, and this approach has been widely applied in many fields (Li et al., 2014; Liu et al., 2017; 2015; Mrkšić et al., 2015; Shinohara, 2016; Zhang et al., 2016, 2014). In MTL, since outcomes are equally important, the objective function to be minimized is the averaged loss across all tasks. Different from MTL, we only address the performance of predicting the target outcome. The decision rule learned in MTL driven by the averaged loss might be biased towards predicting the auxiliary outcomes rather than the target outcome, ie, the jointly learned decision rule may not perform well when

Received: March 22, 2023; Revised: September 3, 2024; Accepted: November 19, 2024

predicting the target outcome. Thus, our objective is to develop a robust learning approach that is capable of exploiting commonalities and differences across outcomes with guaranteed performance in target outcome prediction.

Focusing on the performance of predicting the target outcome, a commonly used approach is transfer learning (Olivas et al., 2009). Transfer learning aims to improve the performance of target learners on target domains by transferring the knowledge contained in different but related source domains (Zhuang et al., 2020). Recently, Li et al. (2022) and Bastani (2021) addressed transfer learning problems in high-dimensional linear regressions; Tian and Feng (2023) addressed transfer learning problems in high-dimensional generalized linear models. In their proposed procedures, they (1) adopt a common working model for all auxiliary outcomes; (2) assume the contrast between the parameters in the target model and those in the auxiliary models are sufficiently close in l_1 or l_0 norm.

However, these assumptions are easily violated in many practical settings such as our motivating example. First, the auxiliary outcomes are related but different, and thus, they are not likely to share the same model. Second, the requirement regarding the contrast between the parameters in the target model and those in the auxiliary models can be restrictive for classification problems. For instance, considering both the target and auxiliary outcomes follow logistic regression models, if the parameters in the target model are twice as large as those in the auxiliary models, the contrast of the two sets of parameters is not necessarily small in l_1 or l_0 norm. However, from the perspective of classification problems, the optimal decision boundaries are identical for the target and auxiliary outcomes. Thus, there is a need for a more flexible learning approach that efficiently utilizes the possible similarity between decision boundaries, rather than focusing on the contrast of parameters, especially for classification problems.

In this work, we develop a robust and flexible learning approach using auxiliary outcomes to aid the estimation of a high-dimensional linear decision rule for the target outcome. Specifically, we propose a two-stage procedure. In the first stage, a common linear representation of the covariates is learned with all auxiliary outcomes using MTL to gain efficiency by borrowing relevant information from auxiliary outcomes. In the second stage, a calibration procedure is performed to reduce or correct the bias induced in the first stage to ensure the robustness of the estimator for the target outcome prediction. Compared with the existing literature, our contributions are the following. In the first stage, different from Li et al. (2022), Tian and Feng (2023), where the working models for auxiliary outcomes share similar coefficients and intercepts, we posit different decision rules (or models) for different outcomes to accommodate possible heterogeneity. In the second stage, instead of assuming that the contrast between the parameters in models for auxiliary outcomes and the target outcome enjoys a small l_1 norm or a sparse l_0 norm, we define a novel concept of within-subspace bias and against-subspace bias, and we only assume that the minimal against-subspace bias is sparse in l_0 norm or small in l_1 norm, which is a weaker condition than those in Bastani (2021), Li et al. (2022), Tian and Feng (2023). Theoretically, we show that the proposed estimator always has an estimation error comparable to that of using only the target outcome, even if the conditions in Li et al. (2022), Tian and Feng (2023) are violated. Especially, we show that with the presence of many weakly dependent outcomes, our proposed method can also lead to a convergence rate faster than the derived rate in Bastani (2021), Li et al. (2022), Tian and Feng (2023) and faster than using only the target outcome.

The rest of the paper is organized as follows. Section 2 introduces the proposed method. In Section 3, we investigate the theoretical properties of the proposed method. In Section 4, we conduct simulations to compare our method with other methods, especially MTL and methods in Li et al. (2022). In Section 5, we apply the proposed method to the motivating study for THA patients. We present a discussion and concluding remarks in Section 6.

2 LEARNING USING AUXILIARY OUTCOMES UNDER HETEROGENEOUS MODELS

Let $\boldsymbol{X} \in \mathbb{R}^p$ be a p-dimensional covariate vector excluding the intercept and $Y_0 \in \{1, -1\}$ be a univariate target outcome. We assume that some auxiliary outcomes are available along with the target outcome Y_0. We denote the auxiliary outcomes as $Y_1, Y_2, \cdots, Y_J \in \{1, -1\}$, where J is the number of auxiliary outcomes.

In our motivating example, the target outcome and auxiliary outcomes are available in the same dataset. There are other scenarios where the target outcome and auxiliary outcomes are not in the same dataset. For example, we may have a separate dataset containing only the auxiliary outcomes and covariates, denoted as the source-only dataset. To accommodate this scenario, we assume that we observe n samples in the target dataset where both the target outcome and auxiliary outcomes are available, ie, $\left\{(\boldsymbol{X}_i, Y_{0,i}, Y_{1,i}, \cdots, Y_{J,i})\right\}_{i=1}^{n}$; in addition, we observe $N - n$ samples in the source-only dataset where only auxiliary outcomes are available, ie, $\left\{(\boldsymbol{X}_i, Y_{1,i}, \cdots, Y_{J,i})\right\}_{i=n+1}^{N}$. We use $R_i = 0$ to indicate samples coming from the target dataset, and $R_i = 1$, from the source-only dataset. In this work, we consider a high-dimensional setting where $p > n$.

Learning a linear decision rule to predict the target outcome Y_0 using covariate vector $\boldsymbol{X}$ entails a classification problem. Empirical risk minimization (ERM) is often used to learn such a linear decision rule. Specifically, ERM minimizes a convex surrogate of the loss function, ie,

$$\min_{\boldsymbol{\theta}_0} \ell(\boldsymbol{\theta}_0) := \mathbb{E}\left[\phi\left\{Y_0(\boldsymbol{X}^\top \boldsymbol{\beta}_0 + c_0)\right\} \mid R = 0\right], \qquad (1)$$

where $\phi(\cdot)$ is a surrogate loss, $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, c_0)^\top$, $\boldsymbol{\beta}_0$ indicates the linear direction of $\boldsymbol{X}$ and c_0 indicates the intercept for predicting Y_0. By solving optimization problem (1), the decision rule, $d_0^*(\boldsymbol{X})$, with the form $d_0^*(\boldsymbol{X}) = \operatorname{sgn}\left(\boldsymbol{X}^\top \boldsymbol{\beta}_0^* + c_0^*\right)$, can be used for prediction purposes, where $\boldsymbol{\theta}_0^* = (\boldsymbol{\beta}_0^*, c_0^*)^\top$ is the minimizer of optimization problem (1). Our goal is to use the auxiliary outcomes to improve the estimation $\boldsymbol{\theta}_0^*$.

2.1 Step one: learn a linear representation using MTL

In this section, we introduce our proposed method, which consists of two steps. The first step is to learn a linear representation

using MTL incorporating the auxiliary outcomes. Denote the index set of auxiliary outcomes as $\mathcal{J} = \{1, 2, \cdots, J\}$.

In this work, we consider the following MTL method. We obtain a linear representation $\widehat{\boldsymbol{w}}_{\mathcal{J}}$ by solving

$$\min_{\boldsymbol{w}, \{c_j\}_{j\in\mathcal{J}}} \widehat{\mathbb{E}}_N \left[\sum_{j\in\mathcal{J}} \phi \left\{ Y_j(\boldsymbol{X}^\top \boldsymbol{w} + c_j) \right\} \right] + \lambda_N \|\boldsymbol{w}\|_1, \quad (2)$$

where λ_N is a tuning parameter and $\widehat{\mathbb{E}}_N[\cdot]$ is the empirical expectation of both the target and source-only datasets. In this procedure, we estimate J decision rules for $\{Y_j\}_{j\in\mathcal{J}}$, simultaneously. These decision rules are structured to learn a common parameter $\boldsymbol{w}$, which is the direction shared by all outcomes. In addition, the intercept represented by c_j's can be different for each outcome to accommodate possible heterogeneity. Leveraging information from auxiliary outcomes (and/or the source-only dataset), the estimator $\widehat{\boldsymbol{w}}_{\mathcal{J}}$ can approach $\boldsymbol{w}^*_{\mathcal{J}}$ with a low estimation error, where $\boldsymbol{w}^*_{\mathcal{J}}$ is the minimizer of $\min_{\boldsymbol{w}, \{c_j\}_{j\in\mathcal{J}}} \mathbb{E}\left[\sum_{j\in\mathcal{J}} \phi \left\{ Y_j(\boldsymbol{X}^\top \boldsymbol{w} + c_j) \right\}\right]$. Although the first step takes advantage of shared information across multiple outcomes, the estimator $\widehat{\boldsymbol{w}}_{\mathcal{J}}$ may be biased w.r.t. $\boldsymbol{\beta}^*_0$, especially when $\boldsymbol{w}^*_{\mathcal{J}}$ is biased w.r.t. $\boldsymbol{\beta}^*_0$.

Remark 1 *In our proposed MTL step, we primarily specify different intercepts to accommodate possible heterogeneity. Note that, we can allow any low-dimensional sub-vector of the coefficients to be different to accommodate heterogeneous effects. More detailed discussions can be found in the online Supplementary Materials.*

2.2 Step two: a novel calibration step

In this section, we present how to de-bias $\widehat{\boldsymbol{w}}_{\mathcal{J}}$ and construct an improved estimator for $\boldsymbol{\beta}^*_0$ through a novel calibration step. To start with, we decompose the bias of $\boldsymbol{w}^*_{\mathcal{J}}$, $\text{bias}(\boldsymbol{w}^*_{\mathcal{J}}) := \boldsymbol{w}^*_{\mathcal{J}} - \boldsymbol{\beta}^*_0 = (1-\gamma)\boldsymbol{w}^*_{\mathcal{J}} - \boldsymbol{\delta}$, where $\boldsymbol{\delta} := \boldsymbol{\beta}^*_0 - \gamma \boldsymbol{w}^*_{\mathcal{J}}$, ie, $\boldsymbol{\beta}^*_0 = \gamma \boldsymbol{w}^*_{\mathcal{J}} + \boldsymbol{\delta}$. The first term in this decomposition, $(1-\gamma)\boldsymbol{w}^*_{\mathcal{J}}$, is along the direction of $\boldsymbol{w}^*_{\mathcal{J}}$, and thus, we refer to it as the *within-subspace bias*; the remaining term $\boldsymbol{\delta}$ is referred to as the *against-subspace bias*. Notably, $\boldsymbol{\beta}^*_0$ is unknown. To determine appropriate γ and $\boldsymbol{\delta}$, we leverage the fact that $\boldsymbol{\beta}^*_0$ minimizes $\mathbb{E}\left[\phi\left\{Y_0(\boldsymbol{X}^\top \boldsymbol{\beta}_0 + c_0)\right\}\right]$. Thus, we replace $\boldsymbol{\beta}_0$ by $\gamma \boldsymbol{w}^*_{\mathcal{J}} + \boldsymbol{\delta}$ and propose to solve

$$\min_{\boldsymbol{\delta}, \gamma, c_0} \mathbb{E}\left[\phi\left\{Y_0(\boldsymbol{X}^\top \boldsymbol{\delta} + \gamma \boldsymbol{X}^\top \boldsymbol{w}^*_{\mathcal{J}} + c_0)\right\}\right]. \quad (3)$$

The loss function in (3) incorporates two adjustments to $\boldsymbol{w}^*_{\mathcal{J}}$, which corresponds to the within-subspace bias and against-subspace bias. First, we calibrate the scaling parameter γ along the subspace generated by $\boldsymbol{w}^*_{\mathcal{J}}$. The term γ identifies the within-subspace bias. For instance, if $\boldsymbol{w}^*_{\mathcal{J}} = 2\boldsymbol{\beta}^*_0$, then, setting $\gamma = 1/2$ can eliminate such a bias. Second, we calibrate the subspace generated by $\boldsymbol{w}^*_{\mathcal{J}}$ using $\boldsymbol{\delta}$. This calibration accounts for the against-subspace bias. If $\boldsymbol{w}^*_{\mathcal{J}} = \boldsymbol{\beta}^*_0 - \boldsymbol{e}$, then setting $\boldsymbol{\delta} = \boldsymbol{e}$ can account for such a bias, where $\boldsymbol{e} = (1, 0, \cdots, 0)^\top$.

The decomposition of $\text{bias}(\boldsymbol{w}^*_{\mathcal{J}})$, ie, $\text{bias}(\boldsymbol{w}^*_{\mathcal{J}}) = (1-\gamma)\boldsymbol{w}^*_{\mathcal{J}} - \boldsymbol{\delta}$ provides multiple options to adjust for possible bias. For each choice of γ, we can obtain a corresponding $\boldsymbol{\delta}$ that leads to a unique decomposition of the bias. For example, when $\gamma = 1$, the corresponding $\boldsymbol{\delta} = \boldsymbol{\beta}^*_0 - \boldsymbol{w}^*_{\mathcal{J}}$; when $\gamma = 1/2$, the corresponding $\boldsymbol{\delta} = \boldsymbol{\beta}^*_0 - \boldsymbol{w}^*_{\mathcal{J}}/2$. Both $\gamma = 1$ and $\gamma = 1/2$ lead to a specification of $\boldsymbol{\delta}$ such that $\boldsymbol{\beta}^*_0 = \gamma \boldsymbol{w}^*_{\mathcal{J}} + \boldsymbol{\delta}$. However, under different choices of γ, the $\boldsymbol{\delta}$'s may be different in terms of their l_0 and l_1 norms, resulting in different levels of difficulties in estimating them. For example, the contrast, $\boldsymbol{w}^*_{\mathcal{J}} - \boldsymbol{\beta}^*_0$, may not be sparse in l_0 norm nor small in l_1 norm. In this case, the contrast $\boldsymbol{w}^*_{\mathcal{J}} - \boldsymbol{\beta}^*_0$ may not be easy to estimate. Among all possible decompositions, the γ's that can lead to a sparse (l_0 norm) or a small (l_1 norm) against-subspace bias, $\boldsymbol{\delta}$, are preferable. For ease of exposition, we focus on the $\boldsymbol{\delta}$ with the least l_1 norm. The results under l_0 norm can be found in the online Supplementary Materials.

Denote the set of $\boldsymbol{\delta}$'s with the least l_1 norm as $\boldsymbol{\delta}^*$. To pin down the γ such that $\boldsymbol{\delta} \in \boldsymbol{\delta}^*$, we propose a special treatment: we first separate the space of $\boldsymbol{\delta}$ into several domains such that in each domain, the solution is unique; then, we select the final estimator through a cross-fitting procedure. Below we introduce how these domains are defined, and show that, at least one solution to (3) in these domains satisfies that $\boldsymbol{\delta} \in \boldsymbol{\delta}^*$.

Remark 2 *If we only focus on the $\boldsymbol{\delta}$ with the least l_1 norm, we can directly solve (3) with a lasso penalty and this special treatment is not required. However, if we focus on the $\boldsymbol{\delta}$ with the least l_0 norm, we need to solve (3) with a l_0 penalty, which is not trivial. The proposed procedure provides a unified approach with theoretical guarantees regardless of how sparsity or scale of against-subspace bias is defined.*

We construct the following domain $\Gamma_k = \left\{\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_p)^\top : \delta_k = 0\right\}$, where $k = 1, \cdots, p$. Let $S^*_{\mathcal{J}}$ be the set of indexes of the non-zero coefficients of $\boldsymbol{w}^*_{\mathcal{J}}$. Due to the strict convexity of ϕ and the assumption that the coordinates of $\boldsymbol{X}$ are not linearly dependent, for any $k \in S^*_{\mathcal{J}}$, there exists a unique γ such that $\boldsymbol{\beta}^*_0 - \gamma \boldsymbol{w}^*_{\mathcal{J}} \in \Gamma_k$. This implies that the objective function in (3) on each Γ_k has a unique minimizer, for any $k \in S^*_{\mathcal{J}}$.

Lemma 1 further implies that to determine the γ such that $\boldsymbol{\delta} \in \boldsymbol{\delta}^*$, we only need to solve the optimization problem (3) within each domain.

Lemma 1 *There exists a $k \in S^*_{\mathcal{J}}$ such that the minimizer of the optimization problem (3) in the domain Γ_k is the minimizer of the optimization problem (3) with $\boldsymbol{\delta} \in \boldsymbol{\delta}^*$.*

Motivated by this, we consider a set of optimization problems

$$\min_{\boldsymbol{\delta}\in\Gamma_k, \gamma, c_0} \widehat{\mathbb{E}}_n \left[\phi\left\{Y_0(\boldsymbol{X}^\top \boldsymbol{\delta} + \gamma \boldsymbol{X}^\top \widehat{\boldsymbol{w}}_{\mathcal{J}} + c_0)\right\}\right] + \widetilde{\lambda}_n \|\boldsymbol{\delta}\|_1 \quad (4)$$

for $k \in S$, where $\widetilde{\lambda}_n$ is a tuning parameter and S is a set of prespecified indices. The objective function in (4) is the empirical version of that in (3), and the domain of $\boldsymbol{\delta}$ is constrained to a set Γ_k. In each Γ_k, the solution of optimization (4) is unique for $k \in S$; when $S^*_{\mathcal{J}} \subset S$, the optimization is guaranteed to identify $\boldsymbol{\delta}$ such that $\boldsymbol{\delta} \in \boldsymbol{\delta}^*$. In our implementation, $\widetilde{\lambda}_n$ is chosen via cross-validation and S is chosen as the index set of nonzero coefficients in $\widehat{\boldsymbol{w}}_{\mathcal{J}}$.

To select the final estimator among the different domains of $\boldsymbol{\delta}$, we propose a cross-fitting procedure. First, we split the entire

target dataset into M folds. Denote the index set of the m-th fold as $\mathcal{I}_m$ and the dataset excluding the m-th fold as $\mathcal{I}_m^c$. For each fold $m \in \{1, \cdots, M\}$ and each $k \in S$, we calculate the minimizer of optimization (4) using the data in $\mathcal{I}_m^c$. Denote its minimizer as $\widehat{\boldsymbol{\delta}}_{\mathcal{I}_m^c}(k)$, $\widehat{\gamma}_{\mathcal{I}_m^c}(k)$, $\widehat{c}_{\mathcal{I}_m^c}(k)$. Subsequently, we have $\widehat{\boldsymbol{\beta}}_{\mathcal{I}_m^c}(k) = \widehat{\boldsymbol{\delta}}_{\mathcal{I}_m^c}(k) + \widehat{\gamma}_{\mathcal{I}_m^c}(k)\widehat{\boldsymbol{w}}_{\mathcal{J}}$. Then, we calculate the loss of $\widehat{\boldsymbol{\beta}}_{\mathcal{I}_m^c}(k)$ and $\widehat{c}_{\mathcal{I}_m^c}(k)$ using the data in $\mathcal{I}_m$ and denote the calculated loss as $L_{\mathcal{I}_m}(k)$, ie, $L_{\mathcal{I}_m}(k) = \widehat{\mathbb{E}}_{\mathcal{I}_m}\left[\phi\left\{Y_0(\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_{\mathcal{I}_m^c}(k) + \widehat{c}_{\mathcal{I}_m^c}(k))\right\}\right]$ and $\widehat{\mathbb{E}}_{\mathcal{I}_m}[\cdot]$ is the empirical average of the data in $\mathcal{I}_m$. We repeat this procedure for each $m \in \{1, \cdots, M\}$ and each $k \in S$. Finally, we calculate the averaged loss $L(k) = \sum_{m=1}^M L_{\mathcal{I}_m}(k)/M$, and choose the k^* that minimizes $L(k)$ among all $k \in S$. Then, we calculate the final estimator $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{c}_0$ by $\widehat{\boldsymbol{\beta}}_0 = \sum_{m=1}^M \widehat{\boldsymbol{\beta}}_{\mathcal{I}_m^c}(k^*)/M$, and $\widehat{c}_0 = \sum_{m=1}^M \widehat{c}_{\mathcal{I}_m^c}(k^*)/M$. In our simulation and real data analysis, we choose $M = 2$ for ease of computation. A summary of the entire algorithm can be found in the online Supplementary Materials.

3 THEORETICAL PROPERTIES

To provide theoretical support for the proposed method, we investigate the convergence rate of the proposed estimator. The proof of all the lemmas, theorems, and corollaries can be found in the online Supplementary Materials.

We first introduce some additional notations. We denote the index set of the non-zero coordinates of $\boldsymbol{\beta}_0^*$ as S^*. The cardinality of S^* is denoted as s^*. We also define α and h_b such that

$$\operatorname{var}\left[\sum_{j\in\mathcal{J}} Y_j\phi'\left\{Y_j(\boldsymbol{X}^\top\boldsymbol{w}_{\mathcal{J}}^* + c_j^*)\right\}X_k \mid \boldsymbol{X}\right] \le CJ^\alpha,$$

$$\sup_{j\in\mathcal{J}}\left|E\left[Y_j\phi'\left\{Y_j(\boldsymbol{X}^\top\boldsymbol{w}_{\mathcal{J}}^* + c_j^*)\right\}X_k \mid \boldsymbol{X}\right]\right| \le Ch_b,$$

for all $k \in \{1, 2, \cdots, p\}$, where $\phi'(\cdot)$ is the first order derivative of $\phi(\cdot)$, X_k is the k-th covariate, and C is a sufficiently large constant. Assuming $\phi'(\cdot)$ is bounded, these inequalities automatically hold with $\alpha = 2$ and $h_b = O(1)$. Under certain assumptions, these inequalities may hold with $\alpha < 2$ and $h_b \to 0$ (or $h_b = 0$). For example, when Y_j's are mutually independent conditional on $\boldsymbol{X}$, we can take $\alpha = 1$; when

$$P(Y_j = 1 \mid \boldsymbol{X})/\left\{1 - P(Y_j = 1 \mid \boldsymbol{X})\right\} = \phi'\left(-\boldsymbol{X}^\top\boldsymbol{w}_{\mathcal{J}}^* - c_j^*\right)/\phi'\left(\boldsymbol{X}^\top\boldsymbol{w}_{\mathcal{J}}^* + c_j^*\right), \quad (5)$$

we can take $h_b = 0$. Note that, when $\phi(\cdot)$ is a logistic loss, model (5) is equivalent to logistic models with the same coefficients and different intercepts for auxiliary outcomes. From these examples, we can see that α controls the mutual dependence between Y_j's, conditional on $\boldsymbol{X}$, and h_b controls the bias of $P(Y_j = 1 \mid \boldsymbol{X})$ w.r.t. model (5). Thus, by incorporating α and h_b, our theoretical results can accommodate dependent Y_j's and model misspecification w.r.t. model (5). A detailed discussion can be found in the online Supplementary Materials. Furthermore, we define

$$h \equiv \inf_{\gamma}\sup_{j\in\mathcal{J}}\left\|\gamma\boldsymbol{w}_j^* - \boldsymbol{\beta}_0^*\right\|_1, \quad (6)$$

where $\boldsymbol{w}_j^*$ and c_j^* are the minimizer of $\min_{\boldsymbol{w},c_j}\mathbb{E}\left[\phi\left\{Y_j(\boldsymbol{X}^\top\boldsymbol{w} + c_j)\right\}\right]$.

Remark 3 *The definition (6) implies a relationship for the cosine angle between $\boldsymbol{w}_j^*$ and $\boldsymbol{\beta}_0^*$ (decision boundaries differ up to an intercept). Specifically, we have*

$$\sup_{j\in\mathcal{J}}\left|(\boldsymbol{\beta}_0^*)^\top\boldsymbol{w}_j^*\right|/\left(\left\|\boldsymbol{\beta}_0^*\right\|_2\left\|\boldsymbol{w}_j^*\right\|_2\right) \ge \left(\left\|\boldsymbol{\beta}_0^*\right\|_2 - h\right)/\left(\left\|\boldsymbol{\beta}_0^*\right\|_2 + h\right).$$

To investigate the theoretical property of $\widehat{\boldsymbol{\beta}}_0$, we introduce the following assumptions.

Assumption 1 *There is a constant C_1 such that $\|\boldsymbol{X}\|_\infty$, $\sup_{\boldsymbol{X},j}|\boldsymbol{X}^\top\boldsymbol{w}_j^*|$, and $|c_j^*|$'s are upper bounded by C_1 with probability 1.*

Assumption 2 *Define $\widetilde{\boldsymbol{X}} = (1, \boldsymbol{X})$. There is a positive constant $\lambda_{\min}$ such that the smallest eigenvalue of $\mathbb{E}\left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^\top\right)$ is lower bounded by $\lambda_{\min}$.*

Assumption 3 *We assume that $\sup_{\boldsymbol{X}}|\boldsymbol{X}^\top\boldsymbol{\beta}_0^*| \le C_2$, and $|c_0^*| \le C_2$ with probability 1, where C_2 is a constant. We also require that the ratios between the k-th coefficients in $\boldsymbol{\beta}_0^*$ and $\boldsymbol{w}_{\mathcal{J}}^*$ are bounded for any $k \in S$. We also assume that $s^*\log p/n \to 0$.*

Assumption 4 *Define $\widetilde{\boldsymbol{X}}_k = (\boldsymbol{X}^\top\boldsymbol{w}_{\mathcal{J}}^*, \boldsymbol{X}_{-k}, 1)$, where $\boldsymbol{X}_{-k}$ is the vector of covariates $\boldsymbol{X}$ excluding the kth covariate. We assume that there is a positive constant $\widetilde{\lambda}_{\min}$ such that the smallest eigenvalue of $\mathbb{E}\left(\widetilde{\boldsymbol{X}}_k\widetilde{\boldsymbol{X}}_k^\top\right)$ is lower bounded by $\widetilde{\lambda}_{\min}$ for all $k \in S$.*

Assumptions 1 and 3 impose a uniform upper bound on the design matrix for technical simplicity. Assumptions 2 and 4 impose a uniform lower bound for the eigenvalues of the design matrix to ensure the identifiability of $\boldsymbol{\beta}_0^*$, c_0^*, $\boldsymbol{\delta}$, and γ for all $k \in S$. In Assumption 4, we also assume that the ratios between the k-th coefficients in $\boldsymbol{\beta}_0^*$ and $\boldsymbol{w}_{\mathcal{J}}^*$ are bounded, for any $k \in S$. To achieve this, we could specify S such that extremely small coefficients in $\widehat{\boldsymbol{w}}_{\mathcal{J}}$ are excluded. In addition, we assume that $s^*\log p/n \to 0$. When $s^* = O(n^\kappa)$ and $N = O(n^{1/(1-\kappa)})$ with $0 \le \kappa < 1$, this assumption requires that $p = o\{\exp\left(N^{1-2\kappa}\right)\}$.

Theorem 1 *Under Assumptions 1–4, taking $\lambda_N \gg |\mathcal{J}|\sqrt{\log|\mathcal{J}|/N}$, and*

$$\lambda_N \ge \sqrt{2(|\mathcal{J}|^\alpha + |\mathcal{J}|^2h_b^2)\log p/N} \vee |\mathcal{J}|\log p/N, \quad \text{and}$$

$\widetilde{\lambda}_n \asymp \sqrt{\log p/n}$, we have

$$\max\left(\left\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0^*\right\|_2^2, \left\|\widehat{c}_0 - c_0^*\right\|_2^2\right) \lesssim \left(s^*\lambda_N^2/|\mathcal{J}|^2 + \lambda_N C_\Sigma h/|\mathcal{J}|\right) \wedge (C_\Sigma h)^2 + \left(\widetilde{\lambda}_n^2 + \widetilde{\lambda}_n h_\delta^* \wedge (h_\delta^*)^2\right) + \log(|S| \vee n)/n$$

with probability approaching to 1, where h^*_δ is the minimizer of $\min_\gamma \|\boldsymbol{\beta}^*_0 - \gamma \boldsymbol{w}^*_{\mathcal{J}}\|_1$ *s.t.* $\boldsymbol{\beta}^*_0 - \gamma \boldsymbol{w}^*_{\mathcal{J}} \in \cup_{k\in S}\Gamma_k$, *and* C_Σ *is defined in Lemma 2 in online Supplementary Materials.*

The resultant rate in Theorem 1 is structured as the sum of three terms. The first term, $\left(s^*\lambda_N^2/|\mathcal{J}|^2 + \lambda_N C_\Sigma h/|\mathcal{J}|\right) \wedge (C_\Sigma h)^2$, is related to the estimation error of $\widehat{\boldsymbol{w}}_{\mathcal{J}}$. The C_Σ reflects the heterogeneity in the design matrix in (generalized) linear models, which is assumed to be a constant in Tian and Feng (2023) by their Assumption 4 or Li et al. (2022) by their Condition 4. The second term, $\widetilde{\lambda}_n^2 + \widetilde{\lambda}_n h^*_\delta \wedge \left(h^*_\delta\right)^2$, is associated with the minimal against-subspace bias of $\boldsymbol{w}^*_{\mathcal{J}}$. The third term, $\log(|S| \vee n)/n$, accounts for the variability of selecting k^* in the algorithm.

Compared with the convergence rate of using only n samples and the target label Y_0, ie, $O_p\left(s^* \log p/n\right)$, the convergence rate shown in Theorem 1 can be faster. For example, when $N \gg n$ and $C_\Sigma h \ll s^*\sqrt{N \log p/n^2}$, the first term is smaller than $s^* \log p/n$. For the second term, when $h^*_\delta \ll s^*\sqrt{\log p/n}$, we have $\widetilde{\lambda}_n^2 + \widetilde{\lambda}_n h^*_\delta \ll s^* \log p/n$. The third term is always negligible compared with $s^* \log p/n$. Hence, when N is sufficiently large compared with n, $h^*_\delta \ll s^*\sqrt{\log p/n}$ can lead to a convergence rate faster than $O_p\left(s^* \log p/n\right)$. The convergence rate shown in Theorem 1 can also be faster than $O_p\left(s^* \log p/n\right)$ even if $N = n$. A detailed discussion can be found in the online Supplementary Materials.

Remark 4 *To achieve the requirement that* $C_\Sigma h \ll s^*\sqrt{N \log p/n^2}$ *and* $h^*_\delta \ll s^*\sqrt{\log p/n}$, *the choice of* $\mathcal{J}$ *is important. Without an appropriate selection of* $\mathcal{J}$, *the convergence rate of* $\widehat{\boldsymbol{\beta}}_0$ *is not necessarily faster than the convergence rate obtained using only the target label* Y_0 *(with sample size* n*); this phenomenon is referred to as the negative transfer (Tian and Feng, 2023). The transferable source detection algorithm proposed in Tian and Feng (2023) can also be applied in our proposed method to avoid a possible negative transfer (see the online Supplementary Materials for a detailed discussion).*

Remark 5 *Theorem 1 is established under definition (6). The definition of* h *enables an upper bound on the* l_1 *norm of* $\gamma \boldsymbol{w}^*_{\mathcal{J}} - \boldsymbol{\beta}^*_0$. *In the online Supplementary Materials, we derive another convergence rate using the* l_0 *norm of* $\gamma \boldsymbol{w}^*_{\mathcal{J}} - \boldsymbol{\beta}^*_0$.

4 SIMULATIONS

In this section, we conduct simulations to compare the performance of our proposed method (Calibrated) with other existing approaches (eg, MTL approaches and other transfer learning approaches). One of the comparison methods, referred to as the baseline approach, uses solely the target outcome and directly solves $\min_{\boldsymbol{\beta}_0, c_0} \widehat{\mathbb{E}}_n\left[\phi\left\{Y_0(\boldsymbol{X}^\top\boldsymbol{\beta}_0 + c_0)\right\}\right] + \lambda_n\|\boldsymbol{\beta}_0\|_1$,, where the logistic loss is chosen for $\phi(\cdot)$ and λ_n is tuned by cross-validation. The other approaches for comparison include a direct transfer learning approach and two MTL approaches. The direct transfer learning approach implements a modified algorithm, where one fixes $\gamma = 1$, and c_j's in Step one are assumed to be the same. This modified algorithm can be considered as an extension of the oracle Trans-Lasso Algorithm (TransferDirect) proposed in Li et al. (2022). The MTL approach 1 (MultiTask1) extends the algorithm proposed in Obozinski et al. (2008) using a logistic loss with a grouped lasso penalty. The MTL approach 2 (MultiTask2) solves $\min_{\boldsymbol{w}, \{c_j\}_{j\in\mathcal{J}\cup 0}} \widehat{\mathbb{E}}_n\left[\sum_{j\in\mathcal{J}} \phi\left\{Y_j(\boldsymbol{X}^\top\boldsymbol{\beta} + c_j)\right\}\right] + \lambda_n\|\boldsymbol{\beta}\|_1$. MultiTask2 shares a similar loss function as the MTL used in Step One for our proposed approach. Comparing the proposed method with the baseline approach, we can examine the performance gained from using the auxiliary outcome. Comparing the proposed method with direct transfer learning, we can see the benefit of the proposed method over the existing transfer learning approaches. By comparing the two MTL approaches, we can examine the difference between transfer learning and MTL approaches when focusing on target label prediction.

Let $\boldsymbol{\beta}_{U_0}$ be the coefficients related to the latent variable U_0 for the target outcome. We generate experimental data following the simulation scenarios below:

(i) We set $n = N$. Let $\boldsymbol{\beta}_{U_0} = (1, -1, 1, -1, 0, \cdots, 0, 0.5, -0.5, 2, -2, 0.5, 0.5, 0, \cdots, 0)^\top$ and $U_0 = 5G(\boldsymbol{X}^\top\boldsymbol{\beta}_{U_0}) + 0.2\epsilon_{U_0}$, where ϵ_{U_0} follows a standard normal distribution. The function $G(\cdot)$ is the cumulative distribution function of a standard normal distribution. Set $\widetilde{U} = 5G(\boldsymbol{X}^\top\boldsymbol{\beta}_{\widetilde{U}}) + 0.2\epsilon_{\widetilde{U}}$, where $\epsilon_{\widetilde{U}}$ follows a standard normal distribution, and the q-th coordinate of $\boldsymbol{\beta}_{\widetilde{U}}$ satisfies that $\boldsymbol{\beta}_{\widetilde{U},q} = \boldsymbol{\beta}_{U_0,q}$ for $q \neq 2, 4$ and $\boldsymbol{\beta}_{\widetilde{U},2} = \boldsymbol{\beta}_{\widetilde{U},4} = 1$. The target outcome is generated by setting $Y_0 = \text{sgn}\left(U_0 - u_{0,1/4}\right)$, where $u_{0,1/4}$ is the first quartile of U_0. We further introduce a weighting parameter ω and generate the auxiliary outcome Y_1 by setting $Y_1 = \text{sgn}\left(U_1 - u_{1,3/4}\right)$, where $U_1 = (1-\omega)U_0 + \omega\widetilde{U}$, and $u_{1,3/4}$ is the third quartile of U_1.

(ii) We set $n = 0.2N$ and $\boldsymbol{\beta}_{U_0} = (1, -1, 1, -1, 0, \cdots, 0)^\top$. We generate U_0 based on a binomial distribution $B\left\{8, G(\boldsymbol{X}^\top\boldsymbol{\beta}_{U_0})\right\}$, where the number of trials equals 8 and the success probability equals $G(\boldsymbol{X}^\top\boldsymbol{\beta}_{U_0})$. Then, we corrupt this U_0: when $U_0 \leq 3$, we set $U_1 = U_0 + B(3, \omega)$; when $U_0 > 4$, we set $U_1 = U_0 - B(3, \omega)$. The target outcome is set as $Y_0 = 1\{U_0 > 0\}$; the auxiliary outcomes are set as $Y_j = 1\left\{U_1 - (2j-1)\right\}$, where $j = 1, 2, 3, 4$.

Scenarios (i) and (ii) both involve a parameter ω that controls the relevancy between auxiliary and target outcomes. The relevance between auxiliary and target outcomes gradually decreases with the increase of ω. Specifically, with the increase of ω, $\boldsymbol{w}^*_{\mathcal{J}}$ involves more against-subspace bias. When $\omega = 0$, because $U_0 = U_1$ in both settings, we can show that $\boldsymbol{\beta}^*_0 = \gamma\boldsymbol{w}^*_{\mathcal{J}}$ for some γ under a Gaussian design.

For covariate vector $\boldsymbol{X}$, we have the following two designs. In Design I, the covariate vector X follows Gaussian distribution $N(\boldsymbol{0}, \boldsymbol{I}_p)$. In Design II, we first generate a p-dimensional vector following $N(\boldsymbol{0}, \boldsymbol{\Sigma}_p)$, where the (l, k)th coordinate of $\boldsymbol{\Sigma}_p$ is

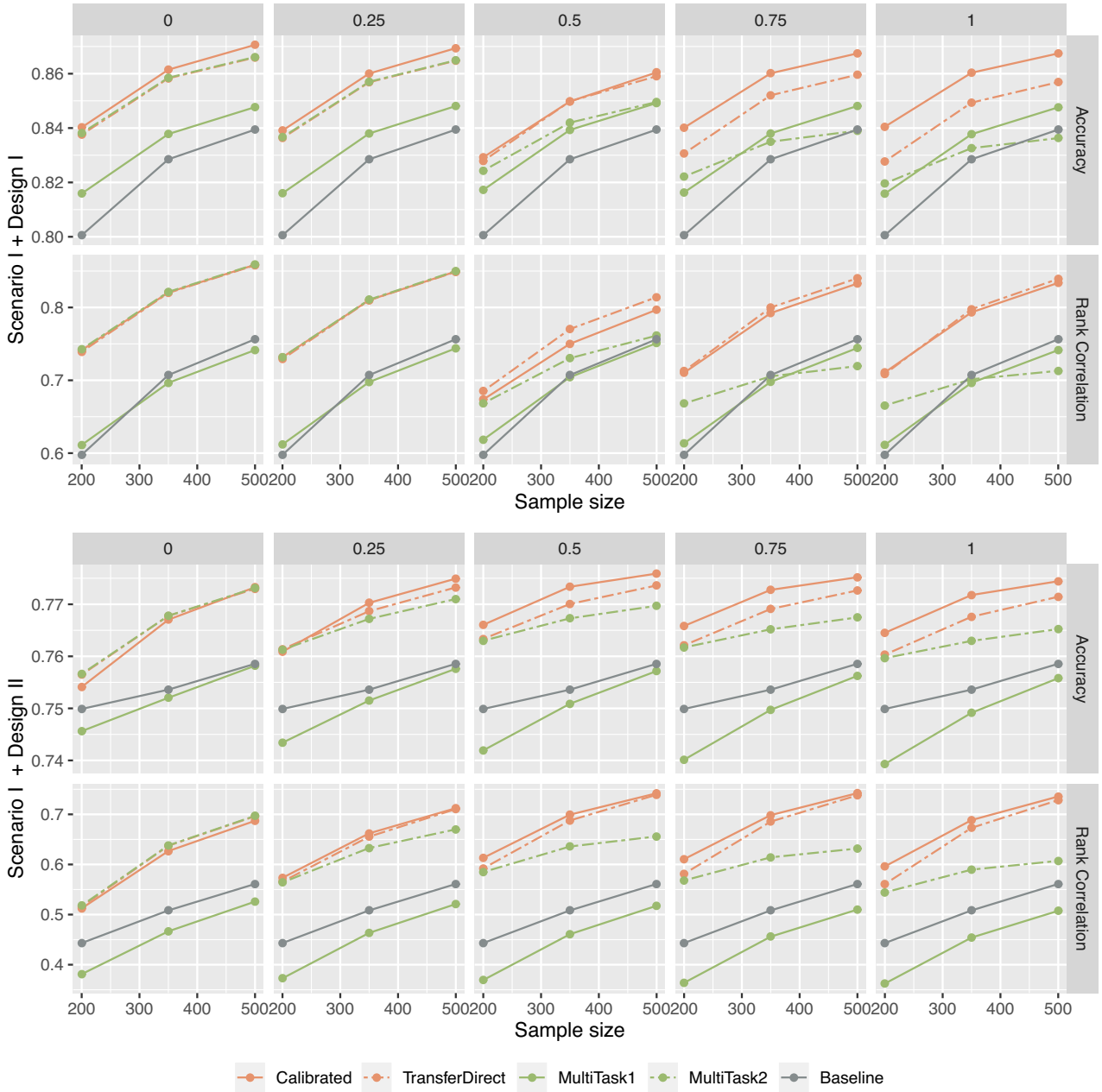

FIGURE 1 Simulation results for Scenario I with the change of sample sizes and ω.

$0.5^{|l-k|}$; then, for $l = 1, \cdots, \lfloor p/4 \rfloor$, we replace the $4l$-th coordinates in the generated vector with a binary variable. This binary variable is 1 if and only if the generated coordinate is greater than 0. Compared with Design I, Design II has correlated and discrete covariates. We test our methods using both designs for Scenarios (i) and (ii).

To compare the performance of different approaches, we generate a testing dataset with sample size $n = 10^4$ and calculate two scores. Let $\widehat{\mathbb{E}}_{\text{test}}[\cdot]$ be the empirical expectation calculated using the testing dataset. The first score is the accuracy. Given an estimated decision rule $\widehat{d}_0(\boldsymbol{X}) = \text{sgn}(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_0 + \widehat{c}_0)$, the accuracy is defined as $\widehat{\mathbb{E}}_{\text{test}}[1\{Y_0 = \widehat{d}_0(\boldsymbol{X})\}]$. The other score is the rank correlation. We calculate the rank correlation between $\boldsymbol{X}^\top \boldsymbol{\beta}_{U_0}$ and $\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}_0$ and use it as a proxy of the estimation error. In these simulations, we vary the sample size of the training dataset from $N = 200, 350$, to 500 and fix $p = 1000$. In Scenario (i), we change ω from 0 to 1 with an increment of 0.25. In Scenario (ii), we change ω from 0 to 0.3 with an increment of 0.1. We repeat each simulation setting for 500 times.

Figures 1 and 2 illustrate how the performance metrics vary with the increase of sample sizes and ω, for simulation Scenarios (i) and (ii), respectively. In Scenario (i), in terms of the accuracy and the rank correlation, the proposed method outperforms the baseline approach regardless of the change of sample sizes and ω. Compared with MultiTask1 and MultiTask2, our proposed method and TransferDirect are more robust w.r.t. the change of

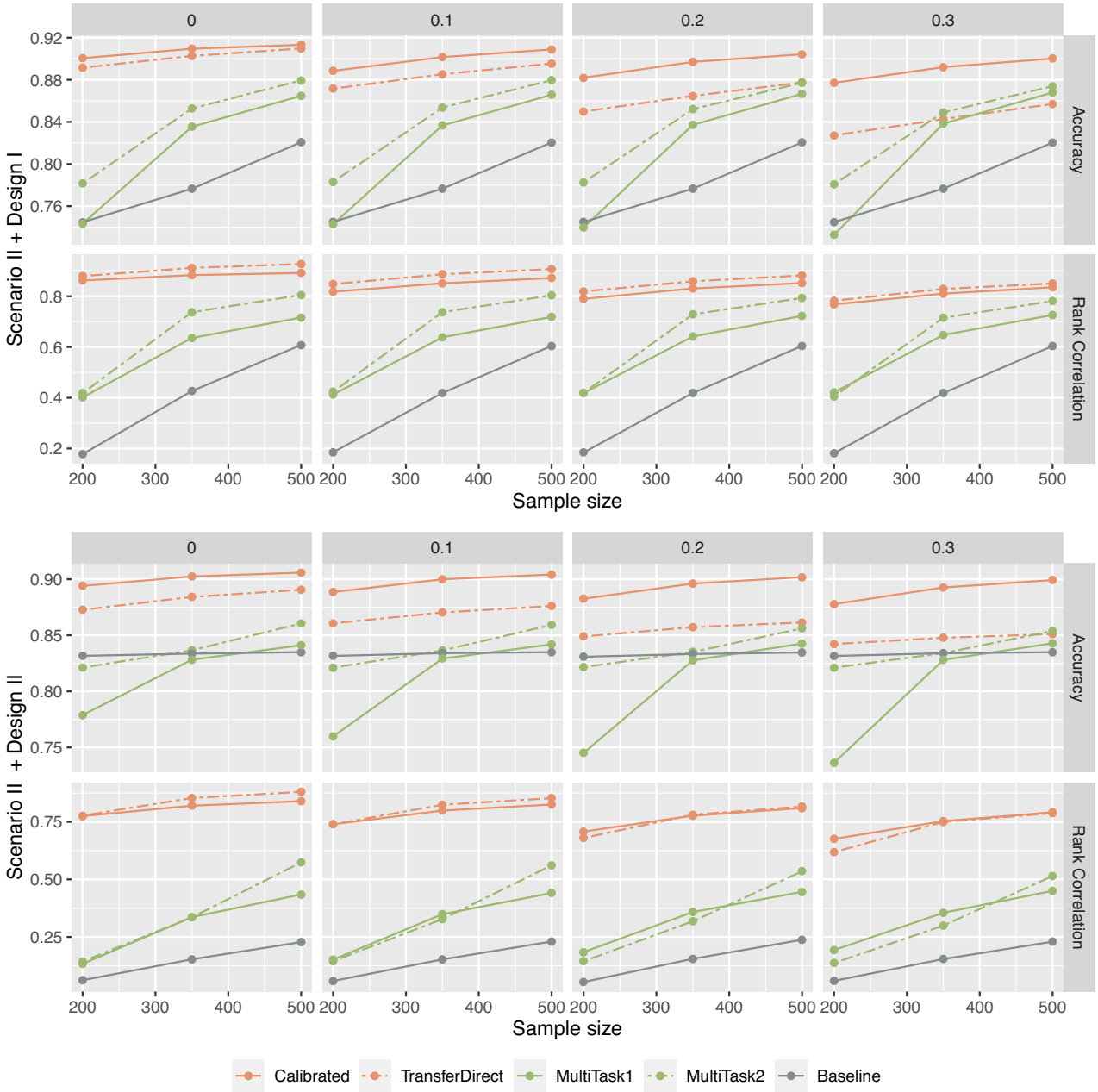

FIGURE 2 Simulation results for Scenario II with the change of sample sizes and ω.

ω; compared with TransferDirect, our proposed method shows great advantages in terms of prediction accuracy. In Scenario (ii), our proposed method also performs better than other methods regardless of the change of sample sizes and ω.

5 APPLICATION TO PREDICTING OUTCOMES AFTER THA

In this section, we apply our proposed method to predict the event of not achieving MCID in terms of HOOS JR scores for THA patients. In this dataset, we have 202 patients who underwent an index THA hospitalization, and we consider 13 variables, including race, Risk Assessment and Prediction Tool, and preoperative HOOS JR survey responses, as covariates. The target outcome is the event of not achieving the anchor-based MCID in their (overall) improvement (Fontana et al., 2019); $Y_0 = 1$, if the patient did not achieve the MCID and $Y_0 = -1$, otherwise. For the choice of auxiliary outcomes, several other survey questionnaires (eg, improvement in pain) are considered. If the survey outcome is continuous or ordinal, we calculate the quartiles and then define a binary outcome based on whether the original outcome surpasses each quartile or not, and use the transformed outcomes as auxiliary outcomes.

To compare different methods, we randomly split the dataset into a training dataset (70% of the entire dataset) and a testing dataset (30% of the entire dataset). We fit the proposed and other

TABLE 1 Comparison of the mean (standard error) of accuracy and the Area Under Receiver Operating Characteristic Curve (AUC) estimated from five methods by repeated sample-splittings of the real data.

Method	Accuracy	AUC
Calibrated	0.746 (0.003)	0.712 (0.004)
TransferDirect	0.740 (0.003)	0.701 (0.004)
MultiTask1	0.712 (0.003)	0.660 (0.004)
MultiTask2	0.739 (0.003)	0.713 (0.004)
Baseline	0.731 (0.003)	0.663 (0.004)

comparison methods on the training dataset and calculate the accuracy and the Area Under Receiver Operating Characteristic Curve (AUC) on the testing dataset. The accuracy reflects the estimation error attributed to both coefficients and intercept estimation, allowing for a scalar multiplier difference, ie, if two sets of coefficients and intercepts are proportional, their accuracy will be the same. The AUC calculated by using the coefficients with varying intercepts primarily reflects the estimation error of coefficients (up to a scalar difference). The entire procedure is repeated 500 times. The mean and standard error of the accuracy and AUC are reported in Table 1.

The results therein show that the proposed method achieves the highest accuracy compared with all other methods in terms of prediction accuracy; the proposed method performs comparable to MultiTask2 in terms of the AUC. This implies that the coefficients derived from MultiTask2 and the proposed method are similar (up to a scalar difference), but the intercept estimation from MultiTask2 is more biased. This also implies that the estimation error of MultiTask2 is mainly attributed to the within-subspace bias rather than the against-subspace bias. The online Supplementary Materials provide an additional application to MCID prediction where our proposed method achieves the highest AUCs among all methods.

6 DISCUSSION

In this work, we develop a robust and flexible learning approach to improving decision rule estimation using auxiliary outcomes. Our approach involves a two-step procedure that takes advantage of the information provided by auxiliary outcomes and retains robustness against the bias introduced by auxiliary outcomes. Our novel bias decomposition allows for weaker required conditions and achieves superior performance against existing approaches.

One possible extension is to propose a transfer learning approach under a more relaxed condition. The proposed estimator achieves a fast rate when $\inf_\gamma \sup_{j\in\mathcal{J}} \left\|\gamma \boldsymbol{w}_j^* - \boldsymbol{\beta}_0^*\right\|_1 \ll s^*\sqrt{\log p/n}$, which is less restrictive than the requirement in Li et al. (2022), Tian and Feng (2023). A more mild condition is, for example, that $\sup_{j\in\mathcal{J}} \inf_\gamma \left\|\gamma \boldsymbol{w}_j^* - \boldsymbol{\beta}_0^*\right\|_1 \ll s^*\sqrt{\log p/n}$ or $\sup_{j\in\mathcal{J}} \inf_\gamma \left\|\gamma \boldsymbol{w}_j^* - \boldsymbol{\beta}_0^*\right\|_2 \ll \sqrt{s^*\log p/n}$. It could be interesting to see a transfer learning approach that achieves faster convergence rates under these milder conditions. In addition, the proposed method can also be modified to suit various practical needs. For example, we may use a distributed learner (Duan et al., 2022) to overcome the communication barrier in the first step. This communication barrier comes from the fact that the datasets from different owners (eg, hospitals) cannot be pooled on a single machine due to privacy regulations (eg, HIPAA on sharing medical records).

ACKNOWLEDGMENTS

The authors thank Dr. Jiwei Zhao, the AE, and two referees for their helpful comments. Muxuan Liang and Jaeyoung Park are the co-first authors of this paper.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Appendices, referenced in Sections 3–5, are available with this paper at the Biometrics website on Oxford Academic. The R codes for implementing the proposed methods are also posted online with this article.

FUNDING

None declared.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The data that support the findings in this paper are available from the corresponding author upon reasonable request.

REFERENCES

Ando, R. K. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6, 1817–1853.

Argyriou, A., Pontil, M., Ying, Y. and Micchelli, C. (2007). A spectral regularization framework for multi-task structure learning. In *Advances in Neural Information Processing Systems*, 20, New York, USA: Curran Associates, Inc.

Bakker, B. and Heskes, T. (2003). Task clustering and gating for Bayesian multitask learning. *Journal of Machine Learning Research*, 4, 83–99.

Bastani, H. (2021). Predicting with proxies: transfer learning in high dimension. *Management Science*, 67, 2964–2984.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28, 41–75.

Duan, R., Ning, Y. and Chen, Y. (2022). Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 109, 67–83.

Fontana, M. A., Lyman, S., Sarker, G. K., Padgett, D. E. and MacLean, C. H. (2019). Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clinical Orthopaedics and Related Research*, 477, 1267.

Gong, P., Ye, J. and Zhang, C. (2013). Multi-stage multi-task feature learning. *Journal of Machine Learning Research*, 14, 2979–3010.

Gong, P., Zhou, J., Fan, W. and Ye, J. (2014). Efficient multi-task feature learning with calibration. In *Proceedings of the 20th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, pp. 761–770.

Hernández-Lobato, D. and Hernández-Lobato, J. M. (2013). Learning feature selection dependencies in multi-task learning. In *Advances in Neural Information Processing Systems*, 26, New York, USA: Curran Associates, Inc.

Katakam, A., Karhade, A. V., Collins, A., Shin, D., Bragdon, C., Chen, A. F. et al. (2022). Development of machine learning algorithms to predict achievement of minimal clinically important difference for the KOOS-PS following total knee arthroplasty. *Journal of Orthopaedic Research*, 40, 808–815.

Kunze, K. N., Karhade, A. V., Sadauskas, A. J., Schwab, J. H. and Levine, B. R. (2020). Development of machine learning algorithms to predict clinically meaningful improvement for the patient-reported health state after total hip arthroplasty. *Journal of Arthroplasty*, 35, 2119–2123.

Li, S., Cai, T. T. and Li, H. (2022). Transfer learning for high-dimensional linear regression: prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B*, 84, 149–173.

Li, S., Liu, Z.-Q. and Chan, A. B. (2014). Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 482–489. IEEE Computer Society: DC, USA.

Liu, P., Qiu, X. and Huang, X. (2017). Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1, 1–10.

Liu, W., Mei, T., Zhang, Y., Che, C. and Luo, J. (2015). Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3707–3715.

Lounici, K., Pontil, M., Tsybakov, A. B. and Van De Geer, S. (2009). Taking advantage of sparsity in multi-task learning, arXiv, arXiv:0903.1468, preprint: not peer reviewed.

Maurer, A., Pontil, M. and Romera-Paredes, B. (2013). Sparse coding for multitask and transfer learning. In *Proceedings of the 30th International Conference on Machine Learning*, 343–351, New York, USA: The Association for Computing Machinery.

Mrkšić, N., Ó Séaghdha, D., Thomson, B., Gašić, M., Su, P.-H., Vandyke, D. et al. (2015). Multi-domain dialog state tracking using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2, 794–799.

Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2008). High-dimensional union support recovery in multivariate regression. In *Advances in Neural Information Processing Systems*, 21,New York, USA: Curran Associates.

Olivas, E. S., Guerrero, J. D. M., Martinez-Sober, M., Magdalena-Benedito, J. R., Serrano, L. et al. (2009). *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques*. Hershey, PA, USA: IGI Global.

Rao, N., Cox, C., Nowak, R. and Rogers, T. (2013). Sparse overlapping sets lasso for multitask learning and its application to fmri analysis. In *Advances in Neural Information Processing Systems*, 26, New York, USA: Curran Associates, Inc.

Shinohara, Y. (2016). Adversarial multi-task learning of deep neural networks for robust speech recognition.In *Interspeech*, 2369–2372.

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

Titsias, M. and Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in Neural Information Processing Systems*, 24, 2339–2347, Curran Associates, Inc, New York, USA.

Wang, X., Bi, J., Yu, S., Sun, J. and Song, M. (2016). Multiplicative multitask feature learning. *Journal of Machine Learning Research*, 17, 2820–2852.

Yang, X., Kim, S. and Xing, E. P. (2009). Heterogeneous multitask learning with joint sparsity constraints. In *Advances in Neural Information Processing Systems*, 22, 2151–2159. New York, USA: Curran Associates, Inc.

Yu, K., Tresp, V. and Schwaighofer, A. (2005). Learning Gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine Learning*, 1012–1019. New York, USA: Tthe Association for Computing Machinery.

Zhang, J., Ghahramani, Z. and Yang, Y. (2008). Flexible latent variable models for multi-task learning. *Machine Learning*, 73, 221–242.

Zhang, W., Li, R., Zeng, T., Sun, Q., Kumar, S. et al. (2016). Deep model based transfer and multi-task learning for biological image analysis. *IEEE Transactions on Big Data*, 6, 322–333.

Zhang, Z., Luo, P., Loy, C. C. and Tang, X. (2014). Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, 94–108. Switzerland: Springer Cham.

Zhu, J., Chen, N. and Xing, E. (2011). Infinite latent SVM for classification and multi-task learning. In *Advances in Neural Information Processing Systems*, 24, New York, USA: Curran Associates, Inc.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H. et al. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109, 43–76.

Received: March 22, 2023; Revised: September 3, 2024; Accepted: November 19, 2024