Statistica Sinica Preprint No: SS-2025-0245	
Title	Doubly Robust Transfer Learning Under Sub-group Shift for Cohort-level Missing Indicator Covariates
Manuscript ID	SS-2025-0245
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0245
Complete List of Authors	Huali Zhao and Tianying Wang
Corresponding Authors	Tianying Wang
E-mails	tianyingw0905@outlook.com

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

Doubly Robust Transfer Learning Under Sub-group Shift for Cohort-Level Missing Indicator Covariates

Huali Zhao

School of Mathematics and Statistics, Huazhong University of Science and Technology

Tianying Wang*

Department of Statistics, Colorado State University

Abstract: Modern biomedical research increasingly relies on integrating multiple cohort studies, yet faces a critical challenge: indicator covariates such as smoking status, vaccination records, or diagnostic codes that are entirely absent in some cohorts due to differences in data collection protocols. This cohort-level missingness violates the assumptions underlying traditional missing data methods, as the complete absence of covariates across entire populations fundamentally differs from sporadic individual-level missingness. To address this gap, we develop a doubly robust transfer learning framework based on a novel sub-group shift assumption, which posits that the conditional distribution of the missing indicator given other variables remains stable across cohorts while allowing marginal distributions to vary. Our approach combines importance weighting with imputation in augmented estimating equations, achieving robustness to misspecification of either the density ratio model or the imputation model. We establish that the proposed

*Corresponding Author: Tianying.Wang@colostate.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

estimator is $n^{1/2}$-consistent and asymptotically normal under mild regularity conditions. Through extensive simulations and an application to UK Biobank data, we demonstrate superior performance compared to existing approaches. This work provides a rigorous framework for handling cohort-level missing indicators, addressing a pervasive challenge in large-scale biomedical data integration.

Key words and phrases: Completely missing, distribution shift, importance weight-ing, model heterogeneity.

1. Introduction

With the emergence of large-scale biomedical research programs, such as the Million Veteran Program (Gaziano et al., 2016), the UK Biobank (Bycroft et al., 2018), and the All of Us Research Program (Denny et al., 2019), researchers have unprecedented resources to explore complex relationships among genetics, lifestyle, environment, and health outcomes. Despite a shared goal of advancing scientific discovery, these initiatives vary considerably in their data collection priorities and protocols, creating substantial heterogeneity in the availability and types of information gathered (Schulz et al., 2020; Li and Zhang, 2025). Some initiatives emphasize lifestyle or behavioral factors, others prioritize wearable-device data, and still others focus on specific clinical outcomes. Such heterogeneity often results in datasets

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

with systematically missing covariates, impeding focused analyses.

Binary indicator variables constitute a substantial portion of potential covariates in contemporary biomedical datasets. In the UK Biobank's phenome catalogue, for instance, about 31% (836 out of 2687 analyzable variables) are simple binary indicators (Millard et al., 2019). Similarly, the OMOP standardized vocabularies, foundational to numerous electronic health record (EHR) systems, encompass over 10 million concepts, nearly two million of which represent medications or clinical conditions encoded as binary presence/absence flags (Reich et al., 2024). Crucially, binary covariates frequently suffer from structured, cohort-level missingness arising from administrative barriers, linkage limitations, or incompatible data standards, rather than random, individual-level gaps. We refer to this phenomenon as *completely missing at the cohort level*. For example, the UK Biobank provides linkage to primary-care records (e.g., diagnoses and prescriptions captured as binary read or SNOMED codes) for roughly 45% of participants, leaving 55% of the cohort systematically missing these GP-derived indicators due to differing coding systems (Allen et al., 2024). Similarly, U.S. Medicare claims databases lack direct measures of smoking status, relying instead on ICD-code-based proxies whose sensitivity compared to self-report is extremely low (<10%), effectively rendering smoking status

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

entirely absent for most beneficiaries (Desai et al., 2016).

This structural absence of key binary covariates is particularly problematic since indicators such as smoking status, alcohol consumption, and clinical diagnoses are known critical confounders in epidemiological studies (Reinau et al., 2014). Without proper handling of these cohort-level missing covariates, predictions, estimations, and causal inferences regarding treatment effects or disease progression may be substantially biased (Varewyck et al., 2016). Traditional missing data methods typically depend on partial covariate availability and assumptions like missing at random or not at random (Bang and Robins, 2005; Kennedy et al., 2019). These methods, however, become ineffective when an entire cohort lacks the covariate of interest, violating essential identifying assumptions. To our knowledge, no existing method explicitly addresses scenarios involving cohort-level missing binary covariates, representing a crucial methodological gap.

External data sources suggest transfer learning as a potential solution. Recent work has successfully applied transfer learning to missing outcome problems (Liu et al., 2023; Zhou et al., 2024; Cai et al., 2025). However, these methods fail for missing covariates due to fundamental differences in problem structure. First, under standard covariate shift assumptions, adapting between populations requires density ratios involving the miss-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

ing covariate that cannot be estimated without target population observations. Second, while missing outcomes enter regression models additively and permit direct imputation, missing covariates appear multiplicatively with parameters (e.g., $X\beta_x$ in linear models), creating non-separable estimation problems. This coupling between imputation and parameter estimation prevents the augmentation strategies that enable double robustness in existing frameworks. Moreover, Zhai and Han (2022) consider leveraging information from external studies that contain only a subset of the covariates available in the internal study of interest. We refer interested readers to Zhai and Han (2022); Li et al. (2024), and their references for a comprehensive overview of data integration and transfer learning frameworks that leverage external populations to improve parameter estimation and handle blockwise missing problems in the target population.

To this end, we propose a **D**oubly **R**obust **T**ransfer **L**earning method to correct the bias caused by **co**mpletely **m**issing **b**inary covariates, referred to as "DRTL-comb". Our contributions are threefold:

1. Cohort-level completely missing binary covariates are often overlooked in the literature, leading to biased inferences for other variables. To address this gap, we propose a novel transfer learning framework that incorporates imputation terms for missing covariates into importance-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

weighted estimating equations, achieving double robustness. This work offers a fresh perspective on missing data studies and advances transfer learning methodologies for addressing missing data issues.

2. We introduce a new sub-group shift assumption, allowing us to effectively use complete data from the source population. Unlike conventional covariate and label shifts (Kpotufe and Martinet, 2021; Plassier et al., 2023; Lee et al., 2025), our sub-group shift assumes the conditional distribution of covariate X given $(Y, \boldsymbol{Z})$ remains invariant, while allowing shifts in the joint distribution of outcomes Y and sub-covariates $\boldsymbol{Z}$. This assumption is less restrictive than label shift and provides an innovative perspective on transfer learning.

3. To correct biases caused by sub-group shifts, we augment our estimating equations with both importance sampling weights and imputation terms, forming an enhanced transfer learning algorithm. We prove that the DRTL-comb estimator is $n^{1/2}$-consistent, asymptotically normal, and doubly robust when either the density ratio model or the imputation model is correctly specified.

While our methodological framework explicitly addresses the important scenario of cohort-level missing binary covariates, the fundamental princi-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

ples and estimating equations readily extend to continuous covariates. Such generalizations are conceptually straightforward but require further theoretical elaboration; see Section 6 for additional discussion.

The rest of the paper is organized as follows: Section 2 introduces the problem setup and details the proposed method and algorithm. Section 3 establishes the consistency and asymptotic normality of the proposed method. The method's performance is demonstrated through simulation studies (Section 4) and an application using UK Biobank data (Section 5). Finally, Section 6 concludes with a discussion.

2. Methodology

2.1 Problem statement

For brevity, we hereafter use *completely missing* to refer specifically to *completely missing at the cohort level*, as discussed in Section 1. Since our methodology exclusively addresses covariates that are systematically absent for entire populations rather than sporadically missing at the individual level, this terminology is unambiguous in our context.

The source population ($\mathcal{S}$) consists of response Y, a binary covariate X, and a set of p-dimensional covariates $\boldsymbol{Z} = (1, Z_1, \cdots, Z_{p-1})^\top$. While the target population ($\mathcal{T}$) consists of Y and $\boldsymbol{Z} = (1, Z_1, \cdots, Z_{p-1})^\top$ only.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

The binary covariate X is completely missing in the target population. For $\iota \in \{\mathcal{T}, \mathcal{S}\}$, we propose the sub-group shift assumption,

$$p^{(\iota)}_{Y,X,\boldsymbol{Z}}(y, x, \boldsymbol{z}) = p^{(\iota)}_{Y,\boldsymbol{Z}}(y, \boldsymbol{z}) p_{X|Y,\boldsymbol{Z}}(x \mid y, \boldsymbol{z}), \tag{2.1}$$

where $p^{(\iota)}_{Y,\boldsymbol{Z}}$ represents the joint probability density of $(Y, \boldsymbol{Z})$ in the population ι and $p_{X|Y,\boldsymbol{Z}}$ is the density of X conditional on $(Y, \boldsymbol{Z})$, which is the same across the two populations. The assumption of a shared $p_{X|Y,\boldsymbol{Z}}$ across populations is plausible in practice. For example, the target data include dietary intake, sex, age ($\boldsymbol{Z}$), and body mass index (BMI, Y), but lack information on smoking status (X). In contrast, the source data contain all variables $(X, Y, \boldsymbol{Z})$. Compared with the assumption of a shared $p_{X|Y}$ in Lee et al. (2025), it is more reasonable to assume that $p(x \mid y, \boldsymbol{z})$ is shared across the two datasets because conditioning on additional covariates $\boldsymbol{Z}$ accounts for relevant demographic and behavioral differences, thereby reducing the variability in the distribution of X across populations.

Remark 1. From the transfer learning perspective, our assumption (2.1) is weaker and less restrictive than the commonly used label shift assumption in the literature in the context of completely missing outcomes (Lee et al., 2025), which assumes $p^{(\iota)}_{Y,X,\boldsymbol{Z}}(y, x, \boldsymbol{z}) = p^{(\iota)}_{Y}(y) p_{X,\boldsymbol{Z}|Y}(x, \boldsymbol{z} \mid y),\ \iota \in \{\mathcal{T}, \mathcal{S}\}$. Here, $p^{(\iota)}_{Y}$ denotes the probability density of Y in the population ι, and

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

$p_{X,\boldsymbol{Z}|Y}$ is the joint density of $(X, \boldsymbol{Z})$ conditional on Y, which remains consistent across the two populations. From a practical perspective, our subgroup shift assumption is therefore more robust than label shift, because any setting that satisfies label shift automatically satisfies our weaker subgroup shift assumption, as $p_{X|Y,\boldsymbol{Z}} = p_{X,\boldsymbol{Z}|Y}/p_{\boldsymbol{Z}|Y}$.

From the missing data perspective, our assumption (2.1) aligns with the missing at random (MAR) mechanism $(X \perp S \mid Y, \boldsymbol{Z})$ by treating X as observed when $S = 1$ and missing when $S = 0$. However, the traditional MAR assumption focuses on the overall population of interest, often overlooking potential distributional shifts in $(Y, \boldsymbol{Z})$. In contrast, our subgroup shift assumption accommodates heterogeneity by allowing distinct joint distributions of $(Y, \boldsymbol{Z})$ across sub-populations.

We are interested in inferring the coefficients of Y regressing on the covariates $X, \boldsymbol{Z}$ in the target population $\mathcal{T}$ under a *working linear model*

$$E_{\mathcal{T}}(Y \mid X, \boldsymbol{Z}) = X\beta_x + \boldsymbol{Z}^\top \boldsymbol{\beta}_z, \tag{2.2}$$

where E_ι is the expectation operator on the population ι for $\iota \in \{\mathcal{T}, \mathcal{S}\}$, and $\boldsymbol{\beta} := (\beta_x, \boldsymbol{\beta}_z^\top)^\top \in \mathbb{R}^{1+p}$ collects the coefficients associated with X and $\boldsymbol{Z}$. Our goal is to estimate $\boldsymbol{\beta}_0 := (\beta_{x0}, \boldsymbol{\beta}_{z0}^\top)^\top$, the solution to the estimating

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

equation in $\mathcal{T}$,

$$E_{\mathcal{T}}\{(X, \boldsymbol{Z}^\top)^\top (Y - X\beta_x - \boldsymbol{Z}^\top \boldsymbol{\beta}_z)\} = \mathbf{0}. \tag{2.3}$$

Let $\{(Y_i, \boldsymbol{Z}_i), i \in \mathcal{I}_{\mathcal{T}}\}$ and $\{(Y_i, X_i, \boldsymbol{Z}_i), i \in \mathcal{I}_{\mathcal{S}}\}$ denote the observed data from the target and source populations with $n_{\mathcal{T}} = |\mathcal{I}_{\mathcal{T}}|$ and $n_{\mathcal{S}} = |\mathcal{I}_{\mathcal{S}}|$, respectively. A simple approach to estimate $\boldsymbol{\beta}_0$ is to directly solve an empirical estimating equation for Eq (2.3) using the source data. However, it will lead to inconsistent estimation because of the potential misspecification of model (2.2) and the sub-group shift, which means that even if Eq (2.2) holds, $E_{\mathcal{S}}\{(X, \boldsymbol{Z}^\top)^\top (Y - X\beta_x - \boldsymbol{Z}^\top \boldsymbol{\beta}_z)\}$ may not be zero.

2.2 Two preliminary methods

To motivate our DRTL-comb method, we first present two preliminary methods: the importance weighting (IW) method and the imputation (IM) method.

The first method is to use the complete data information in the source population. We define the density ratio as

$$\omega_0(Y, \boldsymbol{Z}) = p^{\mathcal{T}}_{Y,\boldsymbol{Z}}(Y, \boldsymbol{Z}) / p^{\mathcal{S}}_{Y,\boldsymbol{Z}}(Y, \boldsymbol{Z}). \tag{2.4}$$

Here, we assume an overlap (or positivity) condition such that $\omega_0(Y, \boldsymbol{Z}) < C$ for some finite constant C. Based on our sub-group shift assumption,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

an intuitive method is to incorporate importance sampling weighting and estimate $\boldsymbol{\beta}_0$ using $\widehat{\boldsymbol{\beta}}_{\mathrm{IW}} := (\widehat{\beta}_{x,\mathrm{IW}}, \widehat{\boldsymbol{\beta}}_{z,\mathrm{IW}}^\top)^\top$, the solution to the weighted estimating equation

$$\frac{1}{n_{\mathcal{S}}} \sum_{i \in \mathcal{I}_{\mathcal{S}}} \widehat{\omega}(Y_i, \boldsymbol{Z}_i)(X_i, \boldsymbol{Z}_i^\top)^\top (Y_i - X_i \beta_x - \boldsymbol{Z}_i^\top \boldsymbol{\beta}_z) = 0, \tag{2.5}$$

where $\widehat{\omega}(y, \boldsymbol{z})$ is the estimator of $\omega(y, \boldsymbol{z})$, a working model for the density ratio $\omega_0(y, \boldsymbol{z})$.

Another intuitive method is to impute the absent X in the target data and then plug it into the estimating equation (2.3). Denote the imputation $m_0(y, \boldsymbol{z}) : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}$ as

$$m_0(y, \boldsymbol{z}) = E(X \mid Y = y, \boldsymbol{Z} = \boldsymbol{z}), \tag{2.6}$$

which is well defined because $E_{\mathcal{S}}(X \mid Y = y, \boldsymbol{Z} = \boldsymbol{z}) = E_{\mathcal{T}}(X \mid Y = y, \boldsymbol{Z} = \boldsymbol{z})$ based on Eq (2.1). It is also natural to estimate $\boldsymbol{\beta}_0$ using $\widehat{\boldsymbol{\beta}}_{\mathrm{IM}} := (\widehat{\beta}_{x,\mathrm{IM}}, \widehat{\boldsymbol{\beta}}_{z,\mathrm{IM}}^\top)^\top$, the solution to the imputation estimating equations

$$\frac{1}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}_{\mathcal{T}}} \widehat{m}(Y_i, \boldsymbol{Z}_i)(Y_i - \beta_x - \boldsymbol{Z}_i^\top \boldsymbol{\beta}_z) = 0, \tag{2.7}$$

$$\frac{1}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}_{\mathcal{T}}} \boldsymbol{Z}_i \{Y_i - \widehat{m}(Y_i, \boldsymbol{Z}_i)\beta_x - \boldsymbol{Z}_i^\top \boldsymbol{\beta}_z\} = 0, \tag{2.8}$$

where $\widehat{m}(y, \boldsymbol{z})$ is the estimator of $m(y, \boldsymbol{z})$, a working model for the imputation $m_0(y, \boldsymbol{z})$. Here, we use the fact that $X_i^2 = X_i$, which allows us to

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

simplify $X_i(Y_i - X_i\beta_x - \boldsymbol{Z}_i^\top \boldsymbol{\beta}_z)$ to $X_i(Y_i - \beta_x - \boldsymbol{Z}_i^\top \boldsymbol{\beta}_z)$, yielding a clearer form of Eq (2.7).

However, the validity of the IW estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{IW}}$ heavily relies on the model specification of $\omega(y, \boldsymbol{z})$ for $\omega_0(y, \boldsymbol{z})$. Similarly, the validity of the IM estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{IM}}$ heavily relies on the model specification of $m(y, \boldsymbol{z})$ for $m_0(y, \boldsymbol{z})$. Our simulation results show that both preliminary estimators perform poorly when either working model $\omega(y, \boldsymbol{z})$ or working model $m(y, \boldsymbol{z})$ is misspecified or poorly estimated, respectively (see Tables 1-2).

2.3 The DRTL-comb method

To overcome the limitations of the IW and IM methods, we propose a novel DRTL-comb method. The key is to effectively integrate the importance weighting equation (2.5) and the imputation estimating equations (2.7)-(2.8), and to avoid the deficiency of using only the IW or IM method. We first exploit the imputation model to impute the missing X in the target population and then augment the importance weighting equation (2.5) with

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

the imputed X, leading to the following augmented estimating equations,

$$\begin{aligned}\widehat{U}_{\mathrm{DR}}(\boldsymbol{\beta}) := & \frac{1}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}_{\mathcal{T}}} \widehat{m}(Y_i, \boldsymbol{Z}_i)(Y_i - \beta_x - \boldsymbol{Z}_i^{\top} \boldsymbol{\beta}_z) \\ & + \frac{1}{n_{\mathcal{S}}} \sum_{i \in \mathcal{I}_{\mathcal{S}}} \widehat{\omega}(Y_i, \boldsymbol{Z}_i)\{X_i - \widehat{m}(Y_i, \boldsymbol{Z}_i)\}(Y_i - \beta_x - \boldsymbol{Z}_i^{\top} \boldsymbol{\beta}_z) = 0, \end{aligned} \tag{2.9}$$

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\mathrm{DR}}(\boldsymbol{\beta}) := & \frac{1}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}_{\mathcal{T}}} \boldsymbol{Z}_i\{Y_i - \widehat{m}(Y_i, \boldsymbol{Z}_i)\beta_x - \boldsymbol{Z}_i^{\top} \boldsymbol{\beta}_z\} \\ & + \frac{1}{n_{\mathcal{S}}} \sum_{i \in \mathcal{I}_{\mathcal{S}}} \widehat{\omega}(Y_i, \boldsymbol{Z}_i)\boldsymbol{Z}_i\{\widehat{m}(Y_i, \boldsymbol{Z}_i) - X_i\}\beta_x = \mathbf{0}. \end{aligned} \tag{2.10}$$

A key observation used in the construction of Eq (2.9) is that $X_i^2 = X_i$, which can simplify $X_i(Y_i - X_i\beta_x - \boldsymbol{Z}_i^{\top}\boldsymbol{\beta}_z)$ as $X_i(Y_i - \beta_x - \boldsymbol{Z}_i^{\top}\boldsymbol{\beta}_z)$, leading to a clear and meaningful form of Eq (2.9). We denote the solution to Eqs (2.9)-(2.10) as the DRTL-comb estimator $\widehat{\boldsymbol{\beta}} := (\widehat{\beta}_x, \widehat{\boldsymbol{\beta}}_z^{\top})^{\top}$ for $\boldsymbol{\beta}_0$.

The idea of the construction of Eqs (2.9)-(2.10) aligns with the existing literature on doubly robust estimators for the average treatment effect in causal inference studies (Bang and Robins, 2005), as well as doubly robust estimators in transfer learning settings with missing outcomes or labels in the target data (Liu et al., 2023; Zhou et al., 2025). It is worth emphasizing that the scenario of completely missing binary covariates is more challenging than that of missing outcomes, due to the distinct roles that covariates and outcomes play in the estimating equations, which significantly adds more

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

challenges to constructing our estimating equations (2.9)-(2.10) to achieve desirable double robustness. Moreover, compared to the scenarios of Liu et al. (2023); Zhou et al. (2025), the establishment of our asymptotic theoretical results becomes more challenging due to: the differing statuses of the covariates X (completely missing) and $\boldsymbol{Z}$ (observed) in the target population, which requires solving two estimating equations simultaneously to determine the coefficients of X and $\boldsymbol{Z}$, and the complexity of the estimating equations (2.9)-(2.10), which introduces additional terms that need to be bounded.

2.4 Algorithm

In this section, we first provide the estimation strategies for nuisance models: density ratio model $\omega(Y, \boldsymbol{Z})$ and imputation model $m(Y, \boldsymbol{Z})$. Then we present the DRTL-comb algorithm.

For the estimation of the density ratio, the *working* density ratio model for $\omega_0(y, \boldsymbol{z})$ can be modeled as

$$\omega(y, z) = \exp(y\eta_y + \boldsymbol{z}^\top \boldsymbol{\eta}_z),$$

where $\eta_y \in \mathbb{R}$ and $\boldsymbol{\eta}_z \in \mathbb{R}^p$ are nuisance parameters. When $\omega(y, \boldsymbol{z})$ is correctly specified, we have $E_{\mathcal{S}}\{\omega(Y, \boldsymbol{Z})(Y, \boldsymbol{Z}^\top)^\top\} = E_{\mathcal{T}}\{(Y, \boldsymbol{Z}^\top)^\top\}$. Hence,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

we can estimate $\widehat{\boldsymbol{\eta}} := (\widehat{\eta}_y, \widehat{\boldsymbol{\eta}}_z^\top)^\top$ by

$$\widehat{\boldsymbol{\eta}} = \arg \min_{\eta_y \in \mathbb{R}, \boldsymbol{\eta}_z \in \mathbb{R}^p} \Big\{ n_{\mathcal{S}}^{-1} \sum_{i \in \mathcal{I}_{\mathcal{S}}} \exp(Y_i \eta_y + \boldsymbol{Z}_i^\top \boldsymbol{\eta}_z) - \ n_{\mathcal{T}}^{-1} \sum_{i \in \mathcal{I}_{\mathcal{T}}} (Y_i \eta_y + \boldsymbol{Z}_i^\top \boldsymbol{\eta}_z) \Big\}.$$

Then,

$$\widehat{\omega}(Y_i, \boldsymbol{Z}_i) = \exp(Y_i \widehat{\eta}_y + \boldsymbol{Z}_i^\top \widehat{\boldsymbol{\eta}}_z), \tag{2.11}$$

For the estimation of the imputation model, the *working* imputation model for $m_0(y, \boldsymbol{z}) = E(X \mid Y = y, \boldsymbol{Z} = \boldsymbol{z})$ can be modeled as

$$m(y, \boldsymbol{z}) = g(y\gamma_y + \boldsymbol{z}^\top \boldsymbol{\gamma}_z),$$

where the link function $g(a) = 1/(1+e^{-a})$, and $\gamma_y \in \mathbb{R}, \boldsymbol{\gamma}_z \in \mathbb{R}^p$ are nuisance parameters. We can estimate $\widehat{\boldsymbol{\gamma}} := (\widehat{\gamma}_y, \widehat{\boldsymbol{\gamma}}_z^\top)^\top$ by

$$\widehat{\boldsymbol{\gamma}} = \arg \min_{\gamma_y \in \mathbb{R}, \boldsymbol{\gamma}_z \in \mathbb{R}^p} n_{\mathcal{S}}^{-1} \sum_{i \in \mathcal{I}_{\mathcal{S}}} \big\{ - X_i (Y_i \gamma_y + \boldsymbol{Z}_i^\top \boldsymbol{\gamma}_z) + \ G(Y_i \gamma_y + \boldsymbol{Z}_i^\top \boldsymbol{\gamma}_z) \big\},$$

where $G(a) = \int_0^a g(u) du$. Then,

$$\widehat{m}(Y_i, \boldsymbol{Z}_i) = g(Y_i \widehat{\gamma}_y + \boldsymbol{Z}_i^\top \widehat{\boldsymbol{\gamma}}_z). \tag{2.12}$$

The finite-sample estimators $\widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\gamma}}$ converge to well-defined population values under our model specifications. To establish the theoretical properties of the DRTL-comb estimator, we formally define two population-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

level parameters $\bar{\boldsymbol{\eta}} := (\bar{\eta}_y, \bar{\boldsymbol{\eta}}_z^\top)^\top$ and $\bar{\boldsymbol{\gamma}} := (\bar{\gamma}_y, \bar{\boldsymbol{\gamma}}_z^\top)^\top$ as

$$\bar{\boldsymbol{\eta}} = \arg \min_{\eta_y \in \mathbb{R}, \boldsymbol{\eta}_z \in \mathbb{R}^p} \big[E_{\mathcal{S}}\{\exp(Y\eta_y + \boldsymbol{Z}^\top \boldsymbol{\eta}_z)\} - E_{\mathcal{T}}(Y\eta_y + \boldsymbol{Z}^\top \boldsymbol{\eta}_z) \big],$$

$$\bar{\boldsymbol{\gamma}} = \arg \min_{\gamma_y \in \mathbb{R}, \boldsymbol{\gamma}_z \in \mathbb{R}^p} E_{\mathcal{S}}\big\{ -X(Y\gamma_y + \boldsymbol{Z}^\top \boldsymbol{\gamma}_z) + G(Y\gamma_y + \boldsymbol{Z}^\top \boldsymbol{\gamma}_z) \big\}.$$

Then, denote $\bar{\omega}(Y, \boldsymbol{Z}) = \exp(Y\bar{\eta}_y + \boldsymbol{Z}^\top \bar{\boldsymbol{\eta}}_z)$ and $\bar{m}(Y, \boldsymbol{Z}) = g(Y\bar{\gamma}_y + \boldsymbol{Z}^\top \bar{\boldsymbol{\gamma}}_z)$.

Based on the doubly robust estimating equations (2.9)-(2.10) and the estimation of nuisance models (2.11)-(2.12), we summarize the DRTL-comb method as Algorithm 1.

Further, we discuss some alternative working nuisance models for $\omega_0(y, \boldsymbol{z})$ and $m_0(y, \boldsymbol{z})$, and we provide the cross-fitted version of the DRTL-comb method when these nuisance functions are estimated using flexible non-parametric or machine learning methods; see Supplement S1 for details.

3. Theoretical Properties

In this section, we establish the consistency and asymptotic validity of the DRTL-comb estimator under mild assumptions. For any vector $\boldsymbol{a}$, let $\|\boldsymbol{a}\|_2$ represent its ℓ_2 norm. Assume that the dimensionality of $\boldsymbol{Z}$, p, is fixed, and $n_{\mathcal{S}}/n_{\mathcal{T}} = O(1)$. Define $\breve{g}(a) = \dot{g}(g^{-1}(a))$, and define the $(1+p)$-order

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

Algorithm 1 The DRTL-comb algorithm

Input: Target data $\{(Y_i, \boldsymbol{Z}_i), i \in \mathcal{I}_{\mathcal{T}}\}$ and source data $\{(Y_i, X_i, \boldsymbol{Z}_i), i \in \mathcal{I}_{\mathcal{S}}\}$.

Step 1: Compute the estimated density ratio $\{\widehat{\omega}(Y_i, \boldsymbol{Z}_i), i \in \mathcal{I}_{\mathcal{S}}\}$ using both target and source data via Eq (2.11).

Step 2: Compute the estimated imputation $\{\widehat{m}(Y_i, \boldsymbol{Z}_i), i \in \mathcal{I}_{\mathcal{S}} \cup \mathcal{I}_{\mathcal{T}}\}$ using the source data only via Eq (2.12).

Step 3: Solve the augmented estimating equations (2.9)-(2.10), and obtain the DRTL-comb estimator $\widehat{\boldsymbol{\beta}}$.

Output: The DRTL-comb estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$.

information matrix

$$\boldsymbol{J}_{\boldsymbol{\beta}} := -\begin{pmatrix} E\{\frac{\partial U(\boldsymbol{\beta})}{\partial \beta_x}\} & E\{\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_z}\} \\ E\{\frac{\partial \boldsymbol{V}(\boldsymbol{\beta})}{\partial \beta_x}\} & E\{\frac{\partial \boldsymbol{V}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_z}\} \end{pmatrix},$$

where $E\{\partial U(\boldsymbol{\beta})/\partial \boldsymbol{\beta}_z\} = \left[E\{\partial \boldsymbol{V}(\boldsymbol{\beta})/\partial \beta_x\}\right]^{\top}$, and

$$\begin{aligned} E\left\{\frac{\partial U(\boldsymbol{\beta})}{\partial \beta_x}\right\} &= E_{\mathcal{S}}[\bar{\omega}(Y, \boldsymbol{Z})\{\bar{m}(Y, \boldsymbol{Z}) - X\}] - E_{\mathcal{T}}\{\bar{m}(Y, \boldsymbol{Z})\}, \\ E\left\{\frac{\partial \boldsymbol{V}(\boldsymbol{\beta})}{\partial \beta_x}\right\} &= E_{\mathcal{S}}[\bar{\omega}(Y, \boldsymbol{Z})\boldsymbol{Z}\{\bar{m}(Y, \boldsymbol{Z}) - X\}] - E_{\mathcal{T}}\{\boldsymbol{Z}\bar{m}(Y, \boldsymbol{Z})\}, \\ E\left\{\frac{\partial \boldsymbol{V}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_z}\right\} &= -E_{\mathcal{T}}(\boldsymbol{Z}\boldsymbol{Z}^{\top}). \end{aligned}$$

Let $[\boldsymbol{J}_u, \boldsymbol{J}_v] = \boldsymbol{J}_{\boldsymbol{\beta}}^{-1}$ with $\boldsymbol{J}_u \in \mathbb{R}^{1+p}$ and $\boldsymbol{J}_v \in \mathbb{R}^{(1+p)\times p}$. Thus, $\boldsymbol{J}_{\boldsymbol{\beta}}, \boldsymbol{J}_u, \boldsymbol{J}_v$ are independent of $\boldsymbol{\beta}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

Assumption 1 (Regularity conditions)**.** There exists a constant $C_L > 0$ such that $|\dot{g}(a) - \dot{g}(b)| \leq C_L|a - b|$ for any $a, b \in \mathbb{R}$. $\boldsymbol{\beta}_0$ belongs to a compact space. $(Y_i, \boldsymbol{Z}_i)$ has bounded support and has a continuously differentiable density on populations $\mathcal{S}$ and $\mathcal{T}$. There exists a constant $C_U > 0$ such that $E_\iota\{\bar{\omega}^4(Y, \boldsymbol{Z}) + Y^4 + Y^{16} + \|\boldsymbol{Z}\|_2^4 + \|\boldsymbol{Z}\|_2^{16}\} < C_U$ for $\iota \in \{\mathcal{T}, \mathcal{S}\}$. The information matrix $\boldsymbol{J}_{\boldsymbol{\beta}_0}$, $E_{\mathcal{S}}\{\bar{\omega}(Y, Z)(Y, \boldsymbol{Z}^\top)^\top(Y, \boldsymbol{Z}^\top)\}$, and $E_{\mathcal{S}}[\dot{g}\{(Y, \boldsymbol{Z}^\top)\bar{\boldsymbol{\gamma}}\}(Y, \boldsymbol{Z}^\top)^\top(Y, \boldsymbol{Z}^\top)]$ have all eigenvalues bounded away from 0 and ∞.

Assumption 2 (Specification of the nuisance models)**.** At least one of the following two conditions holds: (i) $\omega_0(Y, \boldsymbol{Z}) = \exp(Y\eta_{y0} + \boldsymbol{Z}^\top\boldsymbol{\eta}_{z0})$ for some η_{y0} and $\boldsymbol{\eta}_{z0}$; or (ii) $m_0(Y, \boldsymbol{Z}) = g(Y\gamma_{y0} + \boldsymbol{Z}^\top\boldsymbol{\gamma}_{z0})$ for some γ_{y0} and $\boldsymbol{\gamma}_{z0}$.

Assumption 1 is reasonable and commonly used for asymptotic analysis of M-estimation such as logistic regression (van der Vaart, 2000, Chapter 5). Assumption 1 on the compactness of the domain of $(Y_i, \boldsymbol{Z}_i)$ could be relaxed to accommodate unbounded covariates with regular tail behaviors. Assumption 2 assumes that at least one nuisance model is correctly specified, which provides the conditions for the double robustness of the proposed method.

We first establish the consistency of the DRTL-comb estimator as follows. All proofs are presented in Supplement S2 and Supplement S3.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

Theorem 1. *Under Assumptions 1-2, it holds that*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = o_p(1).$$

Theorem 1 establishes the consistency of the DRTL-comb estimator and reveals its double robustness property. Unlike the preliminary IW and IM methods, which require correct specification of their respective nuisance models, our estimator remains consistent as long as either the density ratio model $\omega(Y, Z)$ or the imputation model $m(Y, Z)$ is correctly specified. This double robustness is crucial in practice, as it provides protection against model misspecification, a common challenge in transfer learning applications.

Having established consistency, we now derive the asymptotic distribution of the DRTL-comb estimator. We consider general linear combinations of the parameter vector to accommodate various inferential goals. For any $\boldsymbol{c} \in \mathbb{R}^{1+p}$, we establish the asymptotic normality of $\boldsymbol{c}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ in Theorem 2. Let $\|\boldsymbol{c}\|_2 = 1$, and denote $n = n_{\mathcal{S}}$ for brevity.

Theorem 2. *Under Assumptions 1-2, it holds that*

$$\begin{aligned}\sqrt{n}\boldsymbol{c}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \frac{\sqrt{n}}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}_{\mathcal{T}}} F_i^{\mathcal{T}} + \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_{\mathcal{S}}} F_i^{\mathcal{S}} + \sqrt{n}\boldsymbol{\xi}_\gamma^\top(\widehat{\bar{\boldsymbol{\gamma}}} - \bar{\boldsymbol{\gamma}}) \\ + \sqrt{n}\boldsymbol{\xi}_\eta^\top(\widehat{\bar{\boldsymbol{\eta}}} - \bar{\boldsymbol{\eta}}) + o_p(1), \qquad (3.13)\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

where

$$F_i^{\mathcal{T}} = \boldsymbol{c}^\top \boldsymbol{J}_u \bar{m}_i (Y_i - \beta_{x0} - \boldsymbol{Z}_i^\top \boldsymbol{\beta}_{z0}) + \boldsymbol{c}^\top \boldsymbol{J}_v \boldsymbol{Z}_i (Y_i - \bar{m}_i \beta_{x0} - \boldsymbol{Z}_i^\top \boldsymbol{\beta}_{z0}),$$

$$F_i^{\mathcal{S}} = \bar{\omega}_i \{\boldsymbol{c}^\top \boldsymbol{J}_u (X_i - \bar{m}_i)(Y_i - \beta_{x0} - \boldsymbol{Z}_i^\top \boldsymbol{\beta}_{z0}) + \boldsymbol{c}^\top \boldsymbol{J}_v \boldsymbol{Z}_i (\bar{m}_i - X_i)\beta_{x0}\},$$

and $\boldsymbol{\xi}_\gamma = \boldsymbol{\xi}_\gamma^u + \boldsymbol{\xi}_\gamma^v$, $\boldsymbol{\xi}_\eta = \boldsymbol{\xi}_\eta^u + \boldsymbol{\xi}_\eta^v$ *with*

$$\begin{aligned}
\boldsymbol{\xi}_\gamma^u &= E_{\mathcal{T}}\{\boldsymbol{c}^\top \boldsymbol{J}_u \breve{g}(\bar{m})(Y, \boldsymbol{Z}^\top)^\top (Y - \beta_{x0} - \boldsymbol{Z}^\top \boldsymbol{\beta}_{z0})\} \\
&\quad - E_{\mathcal{S}}\{\bar{\omega} \boldsymbol{c}^\top \boldsymbol{J}_u \breve{g}(\bar{m})(Y, \boldsymbol{Z}^\top)^\top (Y - \beta_{x0} - \boldsymbol{Z}^\top \boldsymbol{\beta}_{z0})\}, \\
\boldsymbol{\xi}_\gamma^v &= -E_{\mathcal{T}}\{\boldsymbol{c}^\top \boldsymbol{J}_v \boldsymbol{Z} \breve{g}(\bar{m})(Y, \boldsymbol{Z}^\top)^\top \beta_{x0}\} + E_{\mathcal{S}}\{\bar{\omega} \boldsymbol{c}^\top \boldsymbol{J}_v \boldsymbol{Z} \breve{g}(\bar{m})(Y, \boldsymbol{Z}^\top)^\top \beta_{x0}\}, \\
\boldsymbol{\xi}_\eta^u &= E_{\mathcal{S}}\{\bar{\omega} \boldsymbol{c}^\top \boldsymbol{J}_u (X - \bar{m})(Y, \boldsymbol{Z}^\top)^\top (Y - \beta_{x0} - \boldsymbol{Z}^\top \boldsymbol{\beta}_{z0})\}, \\
\boldsymbol{\xi}_\eta^v &= E_{\mathcal{S}}\{\bar{\omega} \boldsymbol{c}^\top \boldsymbol{J}_v \boldsymbol{Z} (Y, \boldsymbol{Z}^\top)^\top (\bar{m} - X)\beta_{x0}\},
\end{aligned}$$

Here, we write $\bar{\omega}_i = \bar{\omega}(Y_i, \boldsymbol{Z}_i)$, $\bar{m}_i = \bar{m}(Y_i, \boldsymbol{Z}_i)$, $\bar{\omega} = \bar{\omega}(Y, \boldsymbol{Z})$, *and* $\bar{m} = \bar{m}(Y, \boldsymbol{Z})$ *for brevity. Consequently,* $\sqrt{n}\boldsymbol{c}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ *weakly converges to a Gaussian distribution with mean zero and variance of order one.*

Theorem 2 establishes that the DRTL-comb estimator $\widehat{\boldsymbol{\beta}}$ is $n^{1/2}$-consistent and asymptotically normal under the specified modest assumptions, a result further supported by our subsequent numerical experiments. The asymptotic expansion reveals the distinct contributions from each data source and the double robustness property. The first two empirical terms form a centered influence-function contribution and converge weakly to a mean-zero

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

Gaussian distribution under Assumption 2. Moreover, when Assumption 2 (i) holds (correct density ratio specification), $\boldsymbol{\xi}_\gamma = \mathbf{0}$, removing the effect of $\widehat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}$. Conversely, when Assumption 2 (ii) holds (i.e., correct imputation model specification), $\boldsymbol{\xi}_\eta = \mathbf{0}$, eliminating the contribution from $\widehat{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}}$. Thus, the DRTL-comb estimator achieves $n^{1/2}$-consistency when either nuisance model is correctly specified, confirming its double robustness.

4. Simulation Studies

4.1 Data generation

We evaluate the finite sample performance of the proposed estimator with respect to estimation accuracy. Similar to Cai et al. (2025) regarding data generation and model specification, we fix the sum of the sample sizes of the source population ($n_{\mathcal{S}}$) and the target population ($n_{\mathcal{T}}$) such that $n_{\mathcal{S}} + n_{\mathcal{T}} = 2000$, and our generating mechanisms of S_i ensure that the sample size ratio of two populations $n_{\mathcal{T}}/n_{\mathcal{S}}$ remains within the range $(0.7, 0.9)$. We consider $\boldsymbol{Z} = (1, Z_1, Z_2)^\top$, that is, $p = 3$. We generate $Y_i \sim N(1, 1.5^2)$ and $(Z_{1,i}, Z_{2,i}) \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$ for $i \in \mathcal{I}_{\mathcal{S}} \cup \mathcal{I}_{\mathcal{T}}$, where $\boldsymbol{\Sigma} = (\sigma_{jl}) \in \mathbb{R}^{2\times 2}$ with

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

$\sigma_{jl} = 0.3^{|j-l|}$. We consider two models to generate X_i:

$$\mathrm{M}_{\mathbf{cor}} : \mathrm{logit}\{P(X_i = 1 \mid Y_i, \boldsymbol{Z}_i)\} = -1.2 + 0.8Y_i + 3.2Z_{1,i} - 3.2Z_{2,i},$$

$$\mathrm{M}_{\mathbf{mis}} : \mathrm{logit}\{P(X_i = 1 \mid Y_i, \boldsymbol{Z}_i)\} = -1.2 + 0.8Y_i + 3.2Z_{1,i} - 3.2Z_{2,i} + 3Y_iZ_{1,i},$$

where $\mathrm{logit}(a) = \log(a/(1-a))$ for given $a \in (0,1)$. The imputation model $m(y, \boldsymbol{z}) = g\{(y, 1, z_1, z_2)^\top \boldsymbol{\gamma}\}$ is correctly specified under $\mathrm{M}_{\mathbf{cor}}$ but misspecified under $\mathrm{M}_{\mathbf{mis}}$, as $\mathrm{M}_{\mathbf{mis}}$ includes the interaction term. We consider two models to generate a membership variable S_i to assign the ith observation to the source data when $S_i = 1$ and to the target data when $S_i = 0$:

$$\mathrm{W}_{\mathbf{cor}} : \ \mathrm{logit}\{P(S_i = 1 \mid Y_i, \boldsymbol{Z}_i)\} = 1 - 0.6Y_i - 0.5Z_{1,i} + 0.3Z_{2,i},$$

$$\mathrm{W}_{\mathbf{mis}} : \ \mathrm{logit}\{P(S_i = 1 \mid Y_i, \boldsymbol{Z}_i)\} = 1 - 0.6Y_i - 0.5Z_{1,i} + 0.3Z_{2,i} + 2Y_iZ_{1,i}.$$

The density ratio model $\omega(y, \boldsymbol{z}) = \exp\{(y, 1, z_1, z_2)^\top \boldsymbol{\eta}\}$ is correctly specified under $\mathrm{W}_{\mathbf{cor}}$ but misspecified under $\mathrm{W}_{\mathbf{mis}}$, as $\mathrm{W}_{\mathbf{mis}}$ includes the interaction term. Then, we have three different sets of configurations: (I) $\mathrm{M}_{\mathbf{cor}}$ and $\mathrm{W}_{\mathbf{cor}}$, (II) $\mathrm{M}_{\mathbf{mis}}$ and $\mathrm{W}_{\mathbf{cor}}$, and (III) $\mathrm{M}_{\mathbf{cor}}$ and $\mathrm{W}_{\mathbf{mis}}$.

For each setting, we require benchmark values of the true parameter $\boldsymbol{\beta}_0$ to evaluate the performance of our estimator. Since this parameter is defined as the solution to the estimating equation (2.3) in the target population, we approximate it using a large-scale dataset. Specifically, we generate a dataset with 10^6 observations from the target population fol-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

lowing the data-generating processes described above, and obtain $\widehat{\boldsymbol{\beta}}^{oracle}$ by solving Eq (2.3) with fully observed X. This oracle estimate serves as the ground truth for calculating bias and root mean squared error (RMSE), as it represents what we could achieve with unlimited target data and no missing covariates. Then, 500 bootstrap samples for variance estimation and 500 simulation replications are generated to summarize the average performance measures. For the given estimators $\widehat{\beta}_0, \widehat{\beta}_x, \widehat{\beta}_{z_1}, \widehat{\beta}_{z_2}$ which correspond to the coefficients of the intercept and X, Z_1, Z_2 respectively, we report the empirical average bias, RMSE, standard error, and coverage rate of the nominal 95% confidence interval results for β_x and $\boldsymbol{\beta}_z = (\beta_0, \beta_{z_1}, \beta_{z_2})^\top$ in Tables 1-2.

4.2 Results

We compare the DRTL-comb estimator with the Naive estimator, which is obtained by directly regressing Y on $\boldsymbol{Z}$ while ignoring completely missing X in the target population; thus, there are no inference results for β_x. We also present the preliminary IW and IM estimators as two benchmark estimators. As mentioned in Section 1, there are no existing methods for handling completely missing binary covariates. Therefore, we do not include comparisons with other approaches. For the variance estimator of the Naive

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

method, we use the standard error of linear regression. For the variance estimators of the IW, IM, and DRTL-comb methods, we use bootstrap in practice, which appears to have better numerical performance than using the asymptotic variance estimated directly by the moment estimators as suggested by Liu et al. (2023).

As shown in Tables 1-2, the Naive method performs poorly in all configurations because it ignores the binary covariate X, which is related to $Y, \boldsymbol{Z}$ under $\mathrm{M}_{\mathbf{cor}}$ or $\mathrm{M}_{\mathbf{mis}}$. Specifically, the Naive estimators for $\boldsymbol{\beta}_z$ exhibit significantly larger bias and RMSE, with coverage rates almost approaching zero, compared to other methods in all configurations. This demonstrates that even when interest lies solely in inference for $\boldsymbol{Z}$, ignoring a completely missing binary covariate X that is associated with both Y and $\boldsymbol{Z}$ can lead to biased estimates. When both nuisance models are correct (configuration (I)), the two preliminary methods (IW and IM) and the DRTL-comb method demonstrate similar performance in terms of bias and RMSE. When the imputation model is misspecified (configuration (II)), IM exhibits a larger bias and RMSE than IW and DRTL-comb for most coefficients, whereas with a misspecified density ratio model (configuration (III)), IW shows a greater bias and RMSE than IM and DRTL-comb. However, DRTL-comb achieves almost unbiased point estimates for β_x and $\boldsymbol{\beta}_z$ in three configurations, show-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

Table 1: Point estimator results for β_x and $\boldsymbol{\beta}_z$.

	Bias				**RMSE**			
True	Naive	IW	IM	DRTL-comb	Naive	IW	IM	DRTL-comb
Configuration (I): $\mathbf{M_{cor}}$ and $\mathbf{W_{cor}}$								
$\beta_0 = 1.105$	0.572	-0.003	0.003	0.000	0.572	0.114	0.079	0.090
$\beta_x = 1.103$	/	0.000	0.000	0.005	/	0.204	0.112	0.140
$\beta_{z_1} = -0.437$	0.313	-0.007	-0.005	-0.007	0.313	0.117	0.061	0.064
$\beta_{z_2} = 0.392$	-0.319	0.006	0.003	0.004	0.322	0.120	0.059	0.063
Configuration (II): $\mathbf{M_{mis}}$ and $\mathbf{W_{cor}}$								
$\beta_0 = 1.263$	0.413	-0.002	0.120	0.045	0.416	0.127	0.146	0.116
$\beta_x = 0.854$	/	-0.003	-0.235	-0.089	/	0.210	0.268	0.199
$\beta_{z_1} = -0.427$	0.304	-0.007	0.132	0.047	0.308	0.124	0.146	0.096
$\beta_{z_2} = 0.219$	-0.147	0.011	-0.005	-0.002	0.156	0.101	0.058	0.059
Configuration (III): $\mathbf{M_{cor}}$ and $\mathbf{W_{mis}}$								
$\beta_0 = 0.618$	0.471	-0.036	0.005	0.005	0.474	0.080	0.071	0.073
$\beta_x = 0.999$	/	0.027	-0.004	-0.004	/	0.121	0.115	0.119
$\beta_{z_1} = -1.004$	0.252	1.486	0.002	0.002	0.256	1.487	0.054	0.056
$\beta_{z_2} = 0.339$	-0.310	-0.040	0.001	0.001	0.314	0.071	0.063	0.063

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

Table 2: Variance estimator results for β_x and $\boldsymbol{\beta}_z$.

	Standard Error				**Coverage Rate**			
True	Naive	IW	IM	DRTL-comb	Naive	IW	IM	DRTL-comb
Configuration (I): $\mathbf{M_{cor}}$ and $\mathbf{W_{cor}}$								
$\beta_0 = 1.105$	0.049	0.107	0.079	0.089	0.000	0.942	0.950	0.944
$\beta_x = 1.103$	/	0.186	0.114	0.134	/	0.920	0.960	0.940
$\beta_{z_1} = -0.437$	0.051	0.104	0.061	0.064	0.000	0.930	0.954	0.948
$\beta_{z_2} = 0.392$	0.051	0.103	0.060	0.064	0.000	0.920	0.964	0.950
Configuration (II): $\mathbf{M_{mis}}$ and $\mathbf{W_{cor}}$								
$\beta_0 = 1.263$	0.049	0.111	0.080	0.097	0.000	0.910	0.642	0.934
$\beta_x = 0.854$	/	0.187	0.124	0.161	/	0.932	0.498	0.918
$\beta_{z_1} = -0.427$	0.051	0.107	0.062	0.079	0.000	0.908	0.422	0.908
$\beta_{z_2} = 0.219$	0.051	0.091	0.057	0.057	0.188	0.922	0.944	0.942
Configuration (III): $\mathbf{M_{cor}}$ and $\mathbf{W_{mis}}$								
$\beta_0 = 0.618$	0.048	0.076	0.074	0.077	0.000	0.930	0.962	0.968
$\beta_x = 0.999$	/	0.122	0.120	0.128	/	0.950	0.966	0.966
$\beta_{z_1} = -1.004$	0.048	0.062	0.054	0.056	0.000	0.000	0.958	0.946
$\beta_{z_2} = 0.339$	0.047	0.058	0.062	0.063	0.000	0.878	0.944	0.944

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

ing its double robustness. For the variance estimator, DRTL-comb typically falls between the IW and IM methods, which indicates DRTL-comb will not introduce a larger standard error. Regarding the coverage rate, IW has poor coverage rates for β_{z_1}, β_{z_2} in configuration (III), and IM also has unsatisfactory coverage in configuration (II). However, DRTL-comb maintains a nominal coverage rate in most cases.

4.3 Additional simulation results

To better illustrate the robustness of the proposed method under three non-ideal settings, we also conduct additional simulations that consider (i) unbalanced binary X, (ii) reduced overlap between the target and source populations, and (iii) smaller sample size ratios between the target and source data. For case (i), we use a larger intercept (4 instead of the previous value of -1.2) in models $\mathrm{M}_{\mathbf{cor}}$ and $\mathrm{M}_{\mathbf{mis}}$ in Section 4.1. This modification increases the proportion of 1's in X to the range $(0.75, 0.90)$, compared with the earlier range of $(0.40, 0.50)$. For case (ii), we remove source samples with negative values of Y to create scenarios with reduced overlap in the marginal distributions of Y across the two populations compared to the settings in Section 4.1. This modification also results in reduced overlap in the joint distributions of $(Y, \boldsymbol{Z})$ between the two populations. For case (iii),

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

we adopt two new models for generating S_i compared to the data generation mechanism described in Section 4.1. See Supplement S4 for details.

The simulation results (Tables S4.1-S4.6 in Supplement S4) show that DRTL-comb generally remains stable in most configurations. However, in the smaller $n_{\mathcal{T}}/n_{\mathcal{S}}$ setting with a misspecified imputation model, finite-sample bias and subnominal coverage can still occur. In case (i), the IM method shows similar performance to our previous simulation, whereas the IW method produces larger standard errors for most configurations, which is potentially due to the unbalanced binary X. In case (ii), the IM method again shows performance comparable to our previous simulation, whereas the IW method exhibits substantial bias and poor coverage across all configurations. This deterioration arises because removing source samples with negative Y reduces the population overlap, leading to an inaccurate estimation of the density ratio for the distribution of $(Y, \boldsymbol{Z})$ across the two populations. In case (iii), the results for the IM and IW methods are also consistent with those reported in Section 4.1.

5. Real Data Experiments with UK Biobank Data

Although completely missing covariates are common in practice (Section 1), they pose an empirical validation challenge: when a covariate is truly ab-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

sent, we cannot evaluate the sub-group shift assumption or assess estimator accuracy. To create an evaluable setting, we adopt a controlled approach by artificially removing known covariates from a fully observed dataset. This strategy enables us to indirectly observe evidence supporting our sub-group shift assumption through meaningful differences in coefficient estimates between populations. Moreover, because the withheld covariates remain observable in the original data, we have reliable benchmark estimates against which we can directly assess the accuracy and robustness of our proposed method. Such controlled benchmarking would be impossible if these covariates were genuinely absent from the start.

5.1 Data introduction

In this section, we evaluate the performance of the proposed method using the UK Biobank dataset, a well-known comprehensive biomedical data resource. To simulate a scenario where certain covariates are entirely missing in the target population but available in the source population, we partition the data into target and source groups, explicitly introducing the shifts between them. We then exclude a key covariate from the target population and compare the performance of our method against two baseline approaches (IW and IM). The Naive method is excluded due to its poor

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

performance in simulation studies. This artificially missing strategy ensures that the comparison reflects real-world applications while also providing benchmark results for the true $\boldsymbol{\beta}_0$ based on Eq (2.3), using fully observed X in the target population to measure bias in this real data setting.

Our response of interest Y is the body mass index (BMI), which is closely related to various health problems, such as type 2 diabetes, heart disease, stroke, and some types of cancer (Guh et al., 2009). We consider the smoking status as binary X with "never smoking" defined as 0 and "previous or current smoking" defined as 1. We consider 7909 white British participants with smoking status recorded. To define the target and source populations for our experiment, we treat all individuals with a negative polygenic risk score for BMI (prs.BMI) as the source population ($n_{\mathcal{S}} = 4713$) and all individuals with a positive prs.BMI as the target population ($n_{\mathcal{T}} = 3196$). This setup mimics a scenario in which the smoking-status variable is not recorded in the target population for individuals at higher genetic risk, positive prs.BMI, whereas it is recorded in the source population for individuals at lower genetic risk, negative prs.BMI. Such selective data collection based on disease risk is common in biomedical research. The covariates $\boldsymbol{Z}$ include total energy, sex, and age, consistent with variables commonly considered in BMI-related epidemiological studies

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

(Arem et al., 2013). Smoking status is related to age, sex, and BMI (Dare et al., 2015) and often missing in most studies due to its sensitive nature, privacy concerns, cultural differences, and other data collection limitations (Hedeker et al., 2007; Blankers et al., 2016). We use standardized total energy and age with zero mean and unit variance in the subsequent analysis, which makes the empirical column means of total energy and age included in $\boldsymbol{Z}$ zero. As descriptive evidence related to Eq (2.1), we first compare the marginal distribution of X across the two populations using Pearson's chi-squared test. This test does not reveal significant differences in the distributions of binary X between the source and target populations at the 0.05 significance level. We also fit the logistic regression of X on Y and $\boldsymbol{Z}$. The fitted regression coefficients are similar across the two populations; see results in Table S5.8, Supplement S5. These provide evidence to support our sub-group shift assumption.

5.2 Benchmark with observed X: empirical evidence of sub-group shift

We begin by presenting the linear regression results for two populations separately, using the observed X in the target data (Table 3). Comprehensive results are available in Table S5.7, Supplement S5. These estimates serve

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

as benchmarks for assessing the performance of the proposed estimators.

As shown in Table 3, there are notable differences in the estimated covariate coefficients and p-values across the two populations. The coefficient of energy for the target population is greater than that of the source population, suggesting an increased effect of energy on BMI among the target individuals. Furthermore, the effects of X, sex, and age on BMI also differ between the two populations, with coefficients highlighting significant heterogeneity. These differences demonstrate the heterogeneity between the source and target populations, as well as the limitations of directly analyzing target data using source data. Such disparities highlight the need for tailored methods to account for population-specific variations.

5.3 Results

We now artificially remove X from the target data to evaluate our method's performance. Table 4 presents the estimation results, including point estimator bias (computed relative to the benchmark estimates from Section 5.2 with fully observed X, denoted as "Bias"), bootstrap standard errors ($B = 500$, denoted as "SE"), p-values for testing $H_0 : \beta_j = 0$ (denoted as "p-value"), and 95% confidence intervals (denoted as "95%CI").

As shown in Table 4, the DRTL-comb method achieves the smallest ab-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

Table 3: Benchmark results with X (smoking status) observed in both populations. prs.BMI, polygenic risk score for BMI; Estimate, point estimator; p-value, p-value of hypothesis testing $H_0 : \beta_j = 0$.

Source (prs.BMI $<$ 0)			**Target (prs.BMI $\geq$ 0)**		
Covariate	Estimate	p-value	Covariate	Estimate	p-value
X	0.752	$< 5e-4$	X	0.774	$< 5e-4$
energy	0.074	0.201	energy	0.242	0.008
sex	0.888	$< 5e-4$	sex	1.004	$< 5e-4$
age	0.053	0.358	age	0.008	0.927

solute bias for all parameters, demonstrating superior robustness to model misspecification. In contrast, the preliminary IW and IM methods exhibit substantial bias for most covariates, likely due to misspecification in the density ratio model and the imputation model, respectively. Additionally, the IW method reports a non-significant result for the effect of total energy on BMI (p-value $= 0.107$), which contradicts findings in the existing epidemiological literature (Amatruda et al., 1993).

In summary, as consistently demonstrated in the simulation studies, our method outperforms other approaches and effectively corrects the bias

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

Table 4: Results for the target population with missing X.

Method	Covariate	Bias	SE	95%CI	p-value
IW	X	-0.081	0.244	(0.214,1.172)	0.005
	energy	-0.074	0.105	(-0.036,0.373)	0.107
	sex	-0.750	0.218	(-0.172,0.682)	0.242
	age	-0.100	0.113	(-0.313,0.129)	0.415
IM	X	0.413	0.182	(0.830,1.544)	$< 5e-4$
	energy	-0.005	0.094	(0.053,0.423)	0.012
	sex	-0.067	0.176	(0.591,1.283)	$< 5e-4$
	age	-0.008	0.094	(-0.183,0.184)	0.999
DRTL-comb	X	0.070	0.210	(0.432,1.257)	$< 5e-4$
(proposed)	energy	-0.004	0.095	(0.053,0.424)	0.012
	sex	-0.036	0.178	(0.621,1.317)	$< 5e-4$
	age	0.005	0.094	(-0.172,0.197)	0.893

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

caused by missing covariate information for both β_x and $\boldsymbol{\beta}_z$, ensuring valid statistical inference.

6. Discussion

This paper introduces an innovative approach utilizing transfer learning techniques to deal with completely missing binary covariates in the target data, which is a common and crucial challenge that is unaddressed by existing methods. While extensive methods exist for partially missing covariates and completely missing outcomes, the scenario of covariates that are systematically absent for entire populations has been overlooked. This is not merely a special case of existing methods: the mathematical structure of missing covariates in estimating equations creates unique challenges that invalidate standard approaches. Our doubly robust framework, built on a novel sub-group shift assumption, provides the first rigorous solution to this prevalent problem in modern biomedical data integration.

While we have presented our theory, algorithm, and simulations explicitly for binary covariates, the underlying doubly robust transfer learning framework can naturally generalize to handle continuous or categorical covariates. To extend our approach, one would modify the estimating equations to accommodate additional conditional moments (e.g., second-order

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

moments for continuous covariates). However, fully developing theoretical guarantees in the continuous setting requires additional regularity assumptions and more complex derivations. Thus, we leave these important generalizations for future research, prioritizing here conceptual clarity and direct practical applicability, given the common occurrence and significant impact of completely missing binary covariates in real-world biomedical datasets.

For notational simplicity, we have focused on a single cohort-level missing binary covariate X. In practice, multiple binary indicators (e.g., smoking status, medication use, diagnostic codes) may be simultaneously missing in the target cohort. Under a multivariate sub-group shift assumption $p_{\mathcal{T}}(X \mid Y, \boldsymbol{Z}) = p_{\mathcal{S}}(X \mid Y, \boldsymbol{Z})$ with $X \in \{0,1\}^q$, our DRTL-comb framework extends naturally by replacing the scalar X and its conditional mean $m(Y, \boldsymbol{Z}) = E(X \mid Y, \boldsymbol{Z})$ with the vector X and the multivariate conditional mean $E(X \mid Y, \boldsymbol{Z}) \in \mathbb{R}^q$. The corresponding estimating equations and information matrix increase in dimension and require additional regularity and overlap conditions, as well as appropriate joint imputation models (e.g., multivariate logistic or copula-based approaches). For clarity and space, we leave a detailed multivariate development to future work.

While extensions to generalized linear models and interaction effects represent important future directions, they require fundamental modifica-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

tions to our framework due to the non-separability of missing covariates and parameters in non-linear settings. We leave these challenging but important extensions to future work, noting that our current linear framework already addresses a substantial portion of epidemiological analyses involving continuous health outcomes.

Supplementary Material

The online Supplementary Material contains the potential nuisance models, the cross-fitted version of the proposed method, detailed proofs of all technical results, and extended simulation and data analysis results. The R code is provided on GitHub at: https://github.com/tianyingw/DRTL-comb.

Acknowledgments

The authors thank the Editor, Associate Editor, and reviewers for insightful and constructive comments, which have greatly helped improve the paper. We thank all participants and researchers who contributed to the UK Biobank datasets (www.ukbiobank.ac.uk). This research has been conducted using the UK Biobank Resource under Application Number 207159.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

References

Allen, N. E., B. Lacey, D. A. Lawlor, J. P. Pell, J. Gallacher, L. Smeeth, P. Elliott, P. M. Matthews, R. A. Lyons, A. D. Whetton, A. Lucassen, M. E. Hurles, M. Chapman, A. W. Roddam, N. K. Fitzpatrick, A. L. Hansell, R. Hardy, R. E. Marioni, V. B. O'Donnell, J. Williams, C. M. Lindgren, M. Effingham, J. Sellors, J. Danesh, and R. Collins (2024). Prospective study design and data analysis in UK Biobank. *Science Translational Medicine 16*(729), eadf4428.

Amatruda, J. M., M. C. Statt, and S. L. Welle (1993). Total and resting energy expenditure in obese women reduced to ideal body weight. *The Journal of Clinical Investigation 92*(3), 1236–1242.

Arem, H., J. Reedy, J. Sampson, L. Jiao, A. R. Hollenbeck, H. Risch, S. T. Mayne, and R. Z. Stolzenberg-Solomon (2013). The Healthy Eating Index 2005 and risk for pancreatic cancer in the NIH–AARP study. *Journal of the National Cancer Institute 105*(17), 1298–1305.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Blankers, M., E. S. Smit, P. van der Pol, H. de Vries, C. Hoving, and M. van Laar (2016). The missing=smoking assumption: a fallacy in internet-based smoking cessation trials? *Nicotine & Tobacco Research 18*(1), 25–33.

Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

O. Delaneau, J. O'Connell, et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature 562*(7726), 203–209.

Cai, T., M. Li, and M. Liu (2025). Semi-supervised triply robust inductive transfer learning. *Journal of the American Statistical Association 120*(550), 1037–1047.

Dare, S., D. F. Mackay, and J. P. Pell (2015). Relationship between smoking and obesity: a cross-sectional study of 499,504 middle-aged adults in the UK general population. *PLoS One 10*(4), e0123579.

Denny, J. C., J. L. Rutter, D. B. Goldstein, A. Philippakis, J. W. Smoller, G. Jenkins, E. Dishman, J. L. McCauley, and All of Us Research Program Investigators (2019). The "All of Us" Research Program. *The New England Journal of Medicine 381*(7), 668–676.

Desai, R. J., D. H. Solomon, N. Shadick, C. Iannaccone, and S. C. Kim (2016). Identification of smoking using medicare data—a validation study of claims-based algorithms. *Pharmacoepidemiology and Drug Safety 25*(4), 472–475.

Gaziano, J. M., J. Concato, M. Brophy, L. Fiore, S. Pyarajan, J. Breeling, S. Whitbourne, J. Deen, C. Shannon, D. Humphries, et al. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology 70*, 214–223.

Guh, D. P., W. Zhang, N. Bansback, Z. Amarsi, C. L. Birmingham, and A. H. Anis (2009). The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. *BMC Public Health 9*, 88.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

REFERENCES

Hedeker, D., R. J. Mermelstein, and H. Demirtas (2007). Analysis of binary outcomes with missing data: missing=smoking, last observation carried forward, and a little multiple imputation. *Addiction 102*(10), 1564–1573.

Kennedy, E. H., J. A. Mauro, M. J. Daniels, N. Burns, and D. S. Small (2019). Handling missing data in instrumental variable methods for causal inference. *Annual Review of Statistics and Its Application 6*(1), 125–148.

Kpotufe, S. and G. Martinet (2021). Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics 49*(6), 3299–3323.

Lee, S.-h., Y. Ma, and J. Zhao (2025). Doubly flexible estimation under label shift. *Journal of the American Statistical Association 120*(549), 278–290.

Li, S. and L. Zhang (2025). Multi-dimensional domain generalization with low-rank structures. *Journal of the American Statistical Association 120*(552), 2522–2534.

Li, Y., Y. Wei, and M. Liu (2024). Adaptive learning with blockwise missing and semi-supervised data. *arXiv preprint arXiv:2405.18722*.

Liu, M., Y. Zhang, K. P. Liao, and T. Cai (2023). Augmented transfer regression learning with semi-non-parametric nuisance models. *Journal of Machine Learning Research 24*(293), 1–50.

Millard, L. A., M. R. Munafò, K. Tilling, R. E. Wootton, and G. Davey Smith (2019). MR-pheWAS with stratification and interaction: searching for the causal effects of smoking heaviness identified an effect on facial aging. *PLoS Genetics 15*(10), e1008353.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

Plassier, V., M. Makni, A. Rubashevskii, E. Moulines, and M. Panov (2023). Conformal prediction for federated uncertainty quantification under label shift. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 27907–27947. PMLR.

Reich, C., A. Ostropolets, P. Ryan, P. Rijnbeek, M. Schuemie, A. Davydov, D. Dymshyts, and G. Hripcsak (2024). OHDSI standardized vocabularies—a large-scale centralized reference ontology for international data harmonization. *Journal of the American Medical Informatics Association 31*(3), 583–590.

Reinau, D., C. Surber, S. S. Jick, and C. R. Meier (2014). Epidemiology of basal cell carcinoma in the United Kingdom: incidence, lifestyle factors, and comorbidities. *British Journal of Cancer 111*(1), 203–206.

Schulz, M.-A., B. T. T. Yeo, J. T. Vogelstein, J. Mourao-Miranada, J. N. Kather, K. Kording, B. Richards, and D. Bzdok (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications 11*(1), 4238.

van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge University Press.

Varewyck, M., S. Vansteelandt, M. Eriksson, and E. Goetghebeur (2016). On the practice of ignoring center-patient interactions in evaluating hospital performance. *Statistics in Medicine 35*(2), 227–238.

Zhai, Y. and P. Han (2022). Data integration with oracle use of external information from heterogeneous populations. *Journal of Computational and Graphical Statistics 31*(4), 1001–1012.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0245

REFERENCES

Zhou, D., M. Li, T. Cai, and M. Liu (2024). Model-assisted and knowledge-guided transfer regression for the underrepresented population. *arXiv preprint arXiv:2410.06484*.

Zhou, D., M. Liu, M. Li, and T. Cai (2025). Doubly robust augmented model accuracy transfer inference with high dimensional features. *Journal of the American Statistical Association 120*(549), 524–534.

Huali Zhao, School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China;

E-mail: zhl21@mails.tsinghua.edu.cn

Tianying Wang (corresponding author), Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, U.S.A.;

E-mail: Tianying.Wang@colostate.edu