

# Rank-based transfer learning for high-dimensional survival data with application to sepsis data

Nan Qiao<sup>1</sup>, Haowei Jiang<sup>1</sup>, and Cunjie Lin<sup>2</sup>

<sup>1</sup>School of Statistics, Renmin University of China, Beijing 100872, China

<sup>2</sup>Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing 100872, China [lincunjie@ruc.edu.cn](mailto:lincunjie@ruc.edu.cn)

## Abstract

Sepsis remains a critical challenge due to its high mortality and complex prognosis. To address data limitations in studying MSSA sepsis, we extend existing transfer learning frameworks to accommodate transformation models for high-dimensional survival data. Specifically, we construct a measurement index based on C-index for intelligently identifying the helpful source datasets, and the target model performance is improved by leveraging information from the identified source datasets via performing the transfer step and debiasing step. We further provide an algorithm to construct confidence intervals for each coefficient component. Another significant development is that statistical properties are rigorously established, including  $\ell_1/\ell_2$ -estimation error bounds of the transfer learning algorithm, detection consistency property of the transferable source detection algorithm and asymptotic theories for the confidence interval construction. Extensive simulations and analysis of MIMIC-IV sepsis data demonstrate the estimation and prediction accuracy, and practical advantages of our approach, providing significant improvements in survival estimates for MSSA sepsis patients.

**Keywords:** High-dimensional survival data; MIMIC sepsis cohort; Smooth concordance index; U-estimates.

## 1 Introduction

Sepsis, a life-threatening systemic inflammatory response syndrome caused by infection (Evans et al., 2021), affects approximately 50 million people worldwide annually and carries high mortality rates ranging from 15% to over 50% (Fleischmann-Struzek et al., 2020; Rudd et al., 2020). Due to its high incidence and complex prognosis, sepsis consumes a significant amount of medical resources and incurs substantial expenses. For example, in the USA, sepsis is the most common cause of in-hospital deaths and costs more than \$24 billion annually, accounting for 13% of healthcare expenditures (Paoli et al., 2018). In recent years, advancements in

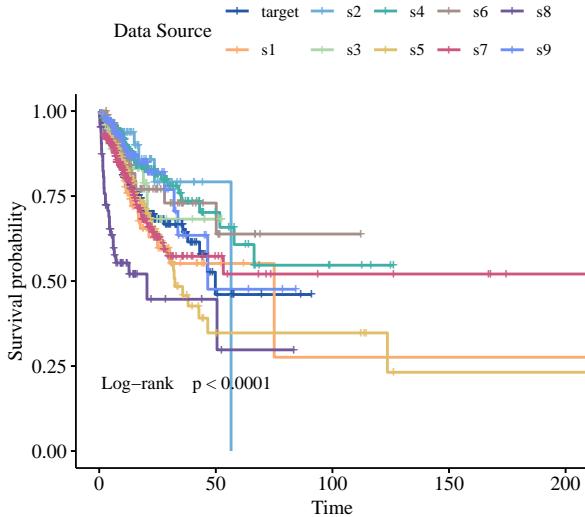


Figure 1: KM curves of different datasets

sepsis guidelines, timely administration of effective antibiotics, and comprehensive management treatments have led to a decrease in sepsis mortality, but the rate remains alarmingly high (Luhr et al., 2019).

Sepsis is triggered by the body's extreme response to infection caused by bacteria, fungi, viruses, or parasites. In sepsis, the most prevalent bacteria that can cause serious clinical consequences are *Staphylococcus aureus* (*S. aureus*) and *Escherichia coli* (*E. coli*) (Faix, 2013), while *S. aureus* is a leading cause of bloodstream infections in hospitals. Among these, Methicillin-Susceptible *Staphylococcus aureus* (MSSA) and methicillin-resistant *Staphylococcus aureus* (MRSA) have received a lot of attention due to the high prevalence, significant morbidity, and mortality (Kourtis, 2019). Studies also show that *Enterococcus* and *Pseudomonas* can induce a severe inflammatory response (Marra et al., 2006), while Gram-negative sepsis is also common in clinical study (Parker and Watkins, 2001). Besides, other causes of sepsis were also recorded in the MIMIC-IV database (<https://mimic.mit.edu>). When we only focus on one type of sepsis, say MSSA, the specific data are often inadequate due to the general focus on broader categories of sepsis or antibiotic-resistant strains like MRSA. This may lead to unsatisfactory analysis results, especially when using complex survival models with right-censored data and high-dimensional covariates. In such cases, it is smart to borrow information from other types of sepsis due to their similarities in symptoms, diagnosis, treatment, and prevention. For example, Kavanagh (2019) found that some classical risk factors associated with MRSA infections were also related to MSSA. Also, many studies are conducted to establish differences between clinical, laboratory, and outcomes of MSSA and MRSA infections, which prompts us to be careful when using information from other datasets.

Transfer learning, which aims to improve the target task's performance by transferring the knowledge contained in different but related source domains (Radhakrishnan et al., 2023), is a promising machine learning methodology for solving the above problem. Given the limited data on MSSA sepsis, traditional models based

solely on target data often lack accuracy and reliability. By incorporating auxiliary information from sepsis with similar pathophysiological mechanisms, transfer learning is expected to help identify critical features that enhance the understanding and prediction of MSSA sepsis outcomes. Nine different sepsis sources (Table 1) are available for transferring, but we are not quite clear about which ones are useful, since MSSA and other sources share the characteristics of causing severe bacterial infections but differ in bacterial types and antibiotic resistance profiles. A preliminary analysis of the sepsis data (For more details of this dataset, see Section 4) from MIMIC-IV provides strong evidence of both similarities and differences among different types of sepsis (Figure 1). In particular, MSSA is susceptible to methicillin, while MRSA is resistant. Other bacterial sources, such as Streptococcus, Enterococcus, E. coli, and Pseudomonas, vary in Gram stain characteristics, pathogenic mechanisms, treatment protocols, and survival probabilities. Hence, not all sepsis sources are useful for transfer learning due to the differences in resistance profiles and pathogenic mechanisms. This necessitates identifying only helpful datasets to improve parameter estimation and predictive accuracy for MSSA sepsis.

In recent years, transfer learning approaches have achieved remarkable success in various fields, such as computer vision (Wang and Deng, 2018), natural language processing (Pruksachatkun et al., 2020), sentiment analysis (Liu et al., 2019), image analysis (Zhang et al., 2015), and the bioinformatics fields (Petegrosso et al., 2017). Numerous methodological frameworks have been developed for transfer learning. One key approach is addressing distribution shift (Uehara et al., 2020; Mo et al., 2021; Wu and Yang, 2023; Chu et al., 2023) by leveraging summary information, such as moments, from the source population to improve the estimates of the target population. However, traditional distribution shift studies typically need data distribution assumptions between target and source datasets and thus usually require low-dimensional settings, which can limit their practical applicability. Alternatively, several studies have focused on parameter-transfer learning for different statistical models under high-dimensional setting. Li et al. (2022) proposed a multi-source transfer learning framework for high-dimensional linear models. Tian and Feng (2023) extended this framework to generalized linear models and developed a consistent procedure for identifying transferable sources. Qiao et al. (2023) conducted transfer learning algorithms for high-dimensional quantile regression (QR) models with the technique of convolution-type smoothing. In semiparametric frameworks, Hu and Zhang (2023) developed a model averaging approach to transfer parameter information from source models to the target model for prediction. Li et al. (2023) addressed time-varying differences in regression coefficients and baseline hazard functions between a target and a source by developing a transfer learning approach in the Cox proportional hazards model. However, despite these advancements, challenges remain, particularly in survival analysis with high-dimensional predictors and multiple sources, where estimators can become unstable, especially with small sample sizes.

In this study, we propose a transfer learning algorithm to enhance model performance on a target survival dataset by leveraging information from other source datasets with similar but not exactly the same distributions. Instead of Cox and some parametric models, we consider the nonparametric transformation model due to its flexibility for modeling censored survival data (Song et al., 2007). This model

directly exploits the monotonic relationship between survival time and covariates while makes no parametric assumptions on either the transformation function or the error. It includes the Cox and many other models as special cases and is more robust against model misspecification. However, to achieve the transfer learning for high dimensional transformation models, we need to overcome the following difficulties: First, the partial rank (PR) estimation (Khan and Tamer, 2007), based on maximizing a discontinuous objective function, is often used to estimate the parameters for a single dataset, which is computationally expensive and practically impossible to compute in the case of high dimensional covariates, not to mention the case with multiple heterogeneous source datasets. Second, we need to intelligently identify the helpful source datasets to avoid the “negative transfer” situation where the target task is compromised by the poor source data which is dramatically different from the target cohort. But it is not easy to construct a measurement index for quantifying the usefulness of the source data due to the right censored data and unknown transformation function. Third, although transfer learning has been successfully applied to linear regression model or quantile regression model, the procedure and its theoretical understanding in the context of the survival transformation model based on U-estimation is significantly more complicated than the transfer learning based on M-estimation. Accordingly, new and more challenging theoretical and numerical developments for identifying informative sources and performing transfer learning with U-estimation are needed. Overall, this study can provide a practically useful new transfer learning approach for survival data with high dimensional covariates when some source data may not improve model performance and can even be harmful.

The rest of this article is organized as follows. Section 2 presents the notation, model, algorithm, and the corresponding theories are also provided. Section 3 contains the results of our simulation studies, evaluating empirical performance. In Section 4, we provide the data analysis results for the motivating example with MIMIC sepsis cohorts and interpret the findings. Concluding remarks and discussions are provided in Section 5. All proofs are in Supplementary Materials.

## 2 Methodology

### 2.1 Transformation Model

Let  $T$  be an uncensored survival time of interest measured from an initial event to the failure event, which is subject to right censoring. We denote the censoring time by  $C$  and the observed survival data are  $Y = \min(T, C)$  and  $\Delta = I(T \leq C)$ . Consider the following transformation model:

$$g(T) = \boldsymbol{\beta}^\top \mathbf{X} + \varepsilon, \quad (1)$$

where  $g(\cdot)$  is a monotone increasing function with an unspecified form,  $\varepsilon$  is a random error term with an unknown distribution  $F$ , which is independent of covariates  $\mathbf{X} \in \mathbb{R}^p$ , and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$  is a vector of unknown regression parameters with dimension  $p$ . For identifiability, the  $\ell_2$  norm of  $\boldsymbol{\beta}$  is restricted to 1, that is  $\|\boldsymbol{\beta}\|_2 = (\sum_{j=1}^p |\beta_j|^2)^{\frac{1}{2}} = 1$ . Without specifying the form of  $g(\cdot)$  and  $F$ , the transformation

model includes many popular models as special cases (Chen, 2002), such as the proportional hazard model, proportional odds model, and accelerated failure time model.

We first briefly review the approach for estimating parameters of model (1) with a single dataset consisting of  $n$  independent and identically distributed samples  $\mathcal{D} = \{Y_i, \Delta_i, \mathbf{X}_i\}_{i=1}^n$ . With unknown  $g(\cdot)$  and  $F$ , the most commonly used approach for estimating the coefficients  $\beta$  is PR estimation (Khan and Tamer, 2007) with the objective function

$$\hat{H}(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(Y_i > Y_j) I(\beta^\top \mathbf{X}_i > \beta^\top \mathbf{X}_j), \quad (2)$$

where  $I(\cdot)$  is the indicator function. Note that  $\hat{H}(\beta)$  is a second-order U-statistic quantifying correlation between the observed and fitted values, which differs fundamentally from M-estimation. And the indicator function is non-differentiable, posing more challenges for optimization. To tackle the computational problem, we consider the objective function with a smooth approximation:

$$\hat{\mathcal{L}}(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(Y_i > Y_j) S_n(\beta^\top \mathbf{X}_i - \beta^\top \mathbf{X}_j),$$

where  $S_n(x) = \frac{1}{1+\exp(-x/\sigma_n)}$  is used to approximate the indicator function  $I(x > 0)$ , and  $\sigma_n$  is strictly positive and decreasing and satisfies  $\lim_{n \rightarrow \infty} \sigma_n = 0$ . By maximizing  $\hat{\mathcal{L}}(\beta)$ , we can derive the smoothed partial rank (SPR) estimator, which has been shown to be asymptotically equivalent to the PR estimator (Song et al., 2007). Thus, the smoothing approach exhibits favorable theoretical properties and yields a computationally affordable estimate with no loss of efficiency.

With high dimensional covariates, a popular approach for performing variable selection and regularized estimation is to consider the penalized objective function

$$\hat{Q}_n(\beta) = -\hat{\mathcal{L}}(\beta) + p_\lambda(\beta),$$

where  $p_\lambda(\cdot)$  is the penalty function that depends on the tuning parameter  $\lambda$ . But it is computationally challenging to optimize  $\hat{Q}_n(\beta)$  due to its non-convexity. In this study, we propose using the Forward and Backward Stagewise (Fabs) algorithm (Shi et al., 2018), an effective computational solution to the penalized SPR estimation, to achieve satisfactory computational efficiency and accuracy.

## 2.2 Transfer Learning Method

In the following, we present our method for the transformation model in the framework of transfer learning more formally. Specifically, two types of datasets are observed: a target dataset  $\mathcal{D}^{(0)} = \left\{Y_i^{(0)}, \Delta_i^{(0)}, \mathbf{X}_i^{(0)}\right\}_{i=1}^{n_0}$  and  $K$  dependent source datasets with the  $k$ -th source denoted as  $\mathcal{D}^{(k)} = \left\{Y_i^{(k)}, \Delta_i^{(k)}, \mathbf{X}_i^{(k)}\right\}_{i=1}^{n_k}$ ,  $k = 1, \dots, K$ , where  $Y_i^{(k)} = \min(T_i^{(k)}, C_i^{(k)})$  and  $\Delta_i^{(k)} = I(T_i^{(k)} \leq C_i^{(k)})$ ,  $i = 1, \dots, n_k$ . We assume that the transformation models hold in the target and source datasets:

$$g(T_i^{(k)}) = \beta^{(k)\top} \mathbf{X}_i^{(k)} + \varepsilon_i^{(k)}, \quad k = 0, 1, \dots, K, \quad i = 1, \dots, n_k,$$

where  $\beta^{(k)} \in \mathbb{R}^p$  are possibly different coefficients of different cohorts and  $g(\cdot)$  is an unspecified monotone increasing function. Here, we assume that both target and source datasets share the same transformation function  $g(\cdot)$  for simplicity. On the one hand, in practice, the source domain used for transfer often possesses some similarity with the target domain, so sharing the same transformation function is a reasonable assumption. Under this assumption, model-based transfer learning is reduced to the transfer of parameters, which can simplify the problem. On the other hand, we focus on the estimation of parameters  $\beta$  and the estimation procedure does not involve the form of transformation function but only requires the function to be monotonic, thus different functions  $g_k(\cdot)$  are also allowed in the estimation.

The goal of this study is to transfer useful information from source data to improve the estimation of the target model, which is assumed to be  $\ell_0$ -sparse in the sense that  $\|\beta^{(0)}\|_0 = s \ll p$ . Besides, to avoid identifiability problems, we restrict  $\|\beta^{(k)}\|_2 = (\sum_{j=1}^p |\beta_j^{(k)}|^2)^{1/2} = 1$ . When transferring the information from source datasets, we allow the parameters  $\beta^{(k)}$  to be different from the target parameter  $\beta^{(0)}$ . However, when the target and source models are disparate, the information borrowed from the source datasets may negatively impact the estimation of the target model, a phenomenon known as “negative transfer”. Thus, to ensure effective transfer learning, we need that  $\beta^{(0)}$  are similar to  $\beta^{(k)}$ , in the sense that  $\|\beta^{(0)} - \beta^{(k)}\|_2 \leq h$  for some reasonably small  $h > 0$ . Here,  $h$  can be treated as a similarity measure. Given a reasonable  $h$ , the index of helpful datasets is denoted by  $\mathcal{A}_h = \{k : k = 1, \dots, K, \|\beta^{(0)} - \beta^{(k)}\|_2 \leq h\}$ . In practice, the value of  $h$  and the corresponding index  $\mathcal{A}_h$  are unknown, and identifying helpful datasets is crucial to ensure transfer gains.

For simplicity, we first consider that the set of helpful datasets  $\mathcal{A}_h$  is known and denote  $n_{\mathcal{A}_h} = \sum_{k \in \mathcal{A}_h} n_k$ , and  $n = n_0 + n_{\mathcal{A}_h}$ . In general framework of transfer learning,  $n_{\mathcal{A}_h} \gg n_0$ . We consider high-dimensional scenarios, i.e., allowing  $p > n$ . To estimate  $\beta^{(0)}$  more effectively and accurately, we propose to borrow information from the datasets  $\mathcal{A}_h$  in the following two steps:

**In the first step**, we perform an initial transfer estimation by fitting a transformation model with  $\ell_1$ -penalty, utilizing data from the target dataset and the helpful source datasets indexed by  $\mathcal{A}_h$ . That is to compute  $\hat{\mathbf{w}}^{\mathcal{A}_h} = \arg \min_{\mathbf{w} \in \mathbb{R}^p} \hat{Q}_n^{\mathcal{A}_h}(\mathbf{w})$ , where

$$\hat{Q}_n^{\mathcal{A}_h}(\mathbf{w}) = - \sum_{k \in \mathcal{A}_h \cup \{0\}} \frac{\alpha_k}{n_k(n_k - 1)} \sum_{i \neq l} \Delta_l^{(k)} I(Y_i^{(k)} > Y_l^{(k)}) S_n \left( \mathbf{w}^\top \mathbf{X}_i^{(k)} - \mathbf{w}^\top \mathbf{X}_l^{(k)} \right) + \lambda_{\mathbf{w}} \|\mathbf{w}\|_1, \quad (3)$$

in which,  $\alpha_k = n_k/n$ ,  $n = n_0 + \sum_{k \in \mathcal{A}_h} n_k$ ,  $\|\mathbf{w}\|_1 = \sum_{j=1}^p |\mathbf{w}_j|$ , and  $\lambda_{\mathbf{w}}$  is a tuning parameter controlling sparsity.

**In the second step**, we need to run a debiased estimation. As the source data may differ from the target, the estimator obtained in the first step is likely to be biased. Define  $\beta^{(0)} = \mathbf{w}^{\mathcal{A}_h} + \boldsymbol{\delta}^{\mathcal{A}_h}$ , where  $\boldsymbol{\delta}^{\mathcal{A}_h}$  quantifies the potential difference between the target coefficient  $\beta^{(0)}$  and fusion parameter  $\mathbf{w}^{\mathcal{A}_h}$ . In this step, we use only the target data  $\mathcal{D}^{(0)}$  to learn  $\boldsymbol{\delta}^{\mathcal{A}_h}$ , and impose an  $\ell_1$ -penalty to achieve data-driven debiasing. Specifically, calculate  $\hat{\boldsymbol{\delta}}^{\mathcal{A}_h} = \arg \min_{\boldsymbol{\delta} \in \mathbb{R}^p} \hat{Q}_{n_0}^{\mathcal{A}_h}(\boldsymbol{\delta})$  with

$$\hat{Q}_{n_0}^{\mathcal{A}_h}(\boldsymbol{\delta}) = \frac{-1}{n_0(n_0 - 1)} \sum_{i \neq l} \Delta_l I(Y_i^{(0)} > Y_l^{(0)}) S_n \left( (\hat{\mathbf{w}}^{\mathcal{A}_h} + \boldsymbol{\delta})^\top \mathbf{X}_i^{(0)} - (\hat{\mathbf{w}}^{\mathcal{A}_h} + \boldsymbol{\delta})^\top \mathbf{X}_l^{(0)} \right) + \lambda_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_1, \quad (4)$$

where  $\|\boldsymbol{\delta}\|_1 = \sum_{j=1}^p \delta_j$ , and  $\lambda_\delta$  is a tuning parameter. Then the final estimator is obtained by  $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_h}^{(0)} = \widehat{\mathbf{w}}^{\mathcal{A}_h} + \widehat{\boldsymbol{\delta}}^{\mathcal{A}_h}$ .

Here we use the Forward and Backward Stagewise (Fabs) algorithm (Shi et al., 2018) to solve (3) and (4), and choose the final solution according to the BIC criterion. Besides, the SPR estimator introduces a smoothing parameter  $\sigma_n$ , which will intuitively affect the convergence rate of the final estimate. In numerical study, we take  $\sigma_n = n^{-1/2}$  as suggested by Song et al. (2007) and Shi et al. (2018). Specifically, Shi et al. (2018) rigorously demonstrated that the SPR estimator with the Fabs algorithm is robust to moderate variations in  $\sigma_n$  and  $\sigma_n = 1/\sqrt{n}$  achieves stable performance across high-dimensional settings without requiring case-specific tuning. In fact, we can show that the smoothing bias (Tan et al., 2022; Qiao et al., 2023) is  $\mathcal{O}(\sigma_n^2)$  and then the estimation error bounds of the estimator will not be affected by  $\sigma_n$  when  $\sigma_n = o(n^{-1/4})$ . Simulation study in Supplementary Materials also shows that  $n^{-1/2}$  is a reasonable choice for  $\sigma_n$  in our setting.

As discussed above, the estimates in (3) and (4) are applicable only if the appropriate sources for transfer are known. Thus, we refer to this method as “Oracle-Trans”. In practice, however, transferring information from certain sources may not enhance the target model’s performance and can even degrade it. Thus, it is crucial to identify the beneficial datasets. Generally, a source dataset is considered helpful, or transferable, if incorporating its data improves the target model’s learning. Here we propose a novel and computationally feasible method to identify informative sources  $\widehat{\mathcal{A}}_h$  using concordance index (C-index) (Khan and Tamer, 2007) as the detection criterion. Note that maximizing (2) is equivalent to maximizing the C-index function:

$$C(\boldsymbol{\beta}; \mathcal{D}) = \frac{\sum_{i \neq j} \Delta_j I(Y_i > Y_j) I(\boldsymbol{\beta}^\top \mathbf{X}_i > \boldsymbol{\beta}^\top \mathbf{X}_j)}{\sum_{i \neq j} \Delta_j I(Y_i > Y_j)}, \quad (5)$$

which counts concordant pairs between the predicted and true outcomes and evaluates the overall performance of the fitted survival model. The value of C-index is between 0 and 1, and 1 means an ideal prediction. Thus, we use the C-index to evaluate the performance of the transfer learning.

Next, we propose to detect the informative sources using a three-fold cross-validation approach, which splits the target samples into three parts, performs transfer learning using two parts, and calculates the C-index with the other part:

Step 1. Split the target data: Randomly divide the target dataset  $\mathcal{D}^{(0)}$  into three folds  $\mathcal{D}^{(0)[r]} = \left\{ Y_i^{(0)[r]}, \Delta_i^{(0)[r]}, \mathbf{X}_i^{(0)[r]} \right\}$  for  $r = 1, 2, 3$ .

Step 2. Calculate the threshold: For the  $r$ -th fold, use  $\mathcal{D}^{(0)[r]}$  as the testing set, and the other two folds,  $\mathcal{D}^{(0)}/\mathcal{D}^{(0)[r]}$ , as the training set. We calculate the estimate  $\widehat{\mathbf{w}}^{(0)[r]}$  by minimizing (3) using the training data  $\mathcal{D}^{(0)}/\mathcal{D}^{(0)[r]}$  but without using any source dataset, and then calculate the C-index on the testing set  $\mathcal{D}^{(0)[r]}$ , i.e.,  $\widehat{C}^{(0)[r]} = C(\widehat{\mathbf{w}}^{(0)[r]}; \mathcal{D}^{(0)[r]})$ . Finally, we take the average  $\widehat{C}^{(0)} = 1/3 \sum_{r=1}^3 \widehat{C}^{(0)[r]}$  as the threshold.

Step 3. Calculate the C-index for each source dataset: We run a transfer estimation using each source dataset to calculate the C-index. For  $k = 1, \dots, K$  and  $r = 1, 2, 3$ , we obtain a transfer estimate  $\widehat{\mathbf{w}}^{(k)[r]}$  by minimizing (3) using  $\mathcal{D}^{(k)} \cup \{\mathcal{D}^{(0)}/\mathcal{D}^{(0)[r]}\}$ . Then, we calculate the C-index on the testing dataset  $\mathcal{D}^{(0)[r]}$ ,

i.e.,  $\widehat{C}^{(k)[r]} = C(\widehat{\mathbf{w}}^{(k)[r]}; \mathcal{D}^{(0)[r]})$ , and then take the average of the three-fold cross-validation,  $\widehat{C}^{(k)} = 1/3 \sum_{r=1}^3 \widehat{C}^{(k)[r]}$ ,  $k = 1, 2, \dots, K$  as the C-index for each source data.

Step 4. Select the informative sources via C-index: If  $\widehat{C}^{(k)} > \widehat{C}^{(0)}$ , we conclude that the  $k$ -th source dataset is beneficial and include  $k$  in  $\widehat{\mathcal{A}}_h$ .

Here, we choose three-fold cross-validation as in Tian and Feng (2023). To further explore the effect of different numbers of folds in cross-validation, we conduct an additional simulation study in Supplementary Materials. We find that increasing the number of folds has minimal impact on the estimation accuracy but the computational time increases significantly. Therefore, three-fold cross-validation provides a reasonable and efficient trade-off in our setting. We summarize our method in Algorithm 1. With the selected informative sources, we can implement transfer learning by solving the optimization problems (3) and (4) with  $\mathcal{A}_h$  replaced by  $\widehat{\mathcal{A}}_h$ . We denote the resulting estimator as  $\widehat{\beta}_{\text{Auto}}^{(0)}$  and the method is termed “Auto-Trans” to emphasize its ability to automatically identify the informative source datasets.

---

**Algorithm 1** Detection of informative sources

---

**Require:** Target dataset  $\mathcal{D}^{(0)}$  and  $K$  source datasets  $\mathcal{D}^{(k)}, k = 1, 2, \dots, K$ .

**Ensure:**  $\widehat{\mathcal{A}}_h$ , the estimate of  $\mathcal{A}_h$ .

Divide the target dataset  $\mathcal{D}^{(0)}$  randomly into three folds, i.e.,  $\mathcal{D}^{(0)} = \bigcup_{r=1}^3 \mathcal{D}^{(0)[r]}$ .

**for**  $r = 1, 2, 3$  **do**

Solve (3) using  $\mathcal{D}^{(0)}/\mathcal{D}^{(0)[r]}$  to obtain  $\widehat{\mathbf{w}}^{(0)[r]}$ ;

Calculate the C-index (5) on  $\mathcal{D}^{(0)[r]}$ :  $\widehat{C}^{(0)[r]} = C(\widehat{\mathbf{w}}^{(0)[r]}; \mathcal{D}^{(0)[r]})$ ;

Calculate the threshold  $\widehat{C}^{(0)} = 1/3 \sum_{r=1}^3 \widehat{C}^{(0)[r]}$ ;

**for**  $k = 1, \dots, K$  **do**

Solve (3) using  $\mathcal{D}^{(k)} \bigcup \{\mathcal{D}^{(0)}/\mathcal{D}^{(0)[r]}\}$  to obtain  $\widehat{\mathbf{w}}^{(k)[r]}$ ;

Calculate the C-index (5) on  $\mathcal{D}^{(0)[r]}$ :  $\widehat{C}^{(k)[r]} = C(\widehat{\mathbf{w}}^{(k)[r]}; \mathcal{D}^{(0)[r]})$ .

**end for**

**end for**

Initialise  $\widehat{\mathcal{A}}_h = \emptyset$ ;

**for**  $k = 1, \dots, K$ , **do**

Calculate  $\widehat{C}^{(k)} = 1/3 \sum_{r=1}^3 \widehat{C}^{(k)[r]}$ .

**if**  $\widehat{C}^{(k)} > \widehat{C}^{(0)}$ , **then**

$\widehat{\mathcal{A}}_h = \widehat{\mathcal{A}}_h \cup \{k\}$  ;

**end if**

**end for**

---

## 2.3 Theoretical Properties and Confidence Interval Construction

### 2.3.1 Estimation error bound and detection consistency

In this section, we establish the theoretical properties of the proposed methods. We first introduce some notations to be used in this text. For a vector

$\mathbf{v} = (v_1, \dots, v_p)^\top \in \mathbb{R}^p$  and  $q \in [1, \infty)$ , let  $\|\mathbf{v}\|_q = \left(\sum_{j=1}^p |v_j|^q\right)^{\frac{1}{q}}$  be its  $\ell_q$  norm,  $\|\mathbf{v}\|_0 = \#\{j : v_j \neq 0\}$  be its  $\ell_0$  norm and  $\|\mathbf{v}\|_\infty = \max_{1 \leq j \leq p} |v_j|$  be its  $\ell_\infty$  norm. For a matrix  $\mathbf{A}_{p \times q} = [a_{ij}]_{p \times q}$ , let  $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq p} \sum_{j=1}^q |a_{ij}|$ , and  $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$ . Let  $\lambda_{\min}(\mathbf{A})$  and  $\lambda_{\max}(\mathbf{A})$  be its eigenvalue with the smallest value and largest value, respectively. This is the common notation for eigenvalues of a matrix, and  $\lambda_{\min}, \lambda_{\max}$  should not be confused with the penalization parameter used in a penalty function. For a sequence  $\{a_n\}$  and another nonnegative sequence  $\{b_n\}$ , we write  $a_n = \mathcal{O}(b_n)$  or  $a_n \lesssim b_n$  if there exists a constant  $c > 0$  such that  $|a_n| \leq cb_n$  for all  $n \geq 1$ . Also, we use  $a_n = o(b_n)$  or  $a_n \ll b_n$  to represent  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 0$ . We write  $b_n \gg a_n$  if  $a_n \ll b_n$ . All proofs are in Supplementary Materials.

The following Theorem 1 gives the  $\ell_1/\ell_2$ -estimation error bound for the Oracle-Trans estimator, which is based on the known informative sources.

**Theorem 1** ( $\ell_1/\ell_2$ -estimation error bound of Oracle-Trans). *Assume Conditions (C1) (C2) (C3) in Supplementary Materials hold,  $n_0 > Cs^2 \log p$ , and  $h \leq s\sqrt{\log p/n_0}$ , where  $C > 0$  is a constant. We take  $\lambda_{\boldsymbol{\omega}} = C_{\boldsymbol{\omega}}\sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}}$  and  $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}}\sqrt{\frac{\log p}{n_0}}$ , where  $C_{\boldsymbol{\omega}}$  and  $C_{\boldsymbol{\delta}}$  are sufficiently large positive constants, then*

$$\|\hat{\boldsymbol{\beta}}_{\mathcal{A}_h}^{(0)} - \boldsymbol{\beta}^{(0)}\|_2 \lesssim h^{1/2} \left(\frac{\log p}{n_0}\right)^{1/4} + s^{1/2} \left(\frac{\log p}{n_0}\right)^{1/4} \left(\frac{\log p}{n_0 + n_{\mathcal{A}_h}}\right)^{1/4}, \quad (6)$$

$$\|\hat{\boldsymbol{\beta}}_{\mathcal{A}_h}^{(0)} - \boldsymbol{\beta}^{(0)}\|_1 \lesssim s \left(\frac{\log p}{n_{\mathcal{A}_h} + n_0}\right)^{1/2} + \left(\frac{\log p}{n_0 + n_{\mathcal{A}_h}}\right)^{1/4} (sh)^{1/2} + h, \quad (7)$$

with probability at least  $1 - 2p^{-1}$ .

Theorem 1 implies that, when  $\mathcal{A}_h$  is an empty set, the upper bound in (6) is  $\mathcal{O}_P(\sqrt{s \log p / n_0})$ . When  $\mathcal{A}_h$  is non-empty, the upper bound in (6) is sharper than  $\sqrt{s \log p / n_0}$  and the upper bound in (7) is sharper than  $s\sqrt{\log p / n_0}$  if  $n_0 \lesssim n_{\mathcal{A}_h}$  and  $h < s\sqrt{\log p / n_0}$ . Similar to Theorem 4 of Tian and Feng (2023), we can show the following detection consistency for Algorithm 1.

**Theorem 2** (Detection consistency of  $\mathcal{A}_h$ ). *For Algorithm 1, with Condition (C4) in Supplementary Materials satisfied for some  $h$ , for any  $\delta > 0$ , there exist constants  $C'(\delta)$  and  $N = N(\delta) > 0$  such that when  $M_1 = C'(\delta)$  and  $\min_{k \in \{0\} \cup \mathcal{A}_h} n_k > N(\delta)$ , we have  $\mathcal{P}(\hat{\mathcal{A}}_h = \mathcal{A}_h) \geq 1 - \delta$ .*

### 2.3.2 Confidence interval construction

In this section, we construct the asymptotic confidence interval (CI) for each component of  $\boldsymbol{\beta}^{(0)}$  based on the previous transfer learning estimator  $\hat{\boldsymbol{\beta}}_{\mathcal{A}_h}^{(0)}$ . Motivated by Cai et al. (2024), Zhang and Zhang (2014) and Ning and Liu (2017), we consider the desparsified estimator  $\hat{\boldsymbol{\beta}}_{\mathcal{A}_h}^{(0)} + H^{-1}\hat{\boldsymbol{\eta}}$ , where  $H = -\mathbb{E}\left\{\Delta_l^{(0)} I(Y_i^{(0)} > Y_l^{(0)}) S_n''(\boldsymbol{\beta}^{(0)}(\mathbf{X}_i^{(0)} - \mathbf{X}_l^{(0)})) (\mathbf{X}_i^{(0)} - \mathbf{X}_l^{(0)})^\top\right\}$  is the inverse Hessian matrix, and  $\hat{\boldsymbol{\eta}} := \frac{1}{n_0(n_0-1)} \sum_{i \neq l} \Delta_l^{(0)} I(Y_i^{(0)} > Y_l^{(0)}) S_n'(\hat{\boldsymbol{\beta}}_{\mathcal{A}_h}^{(0)}(\mathbf{X}_i^{(0)} - \mathbf{X}_l^{(0)})) (\mathbf{X}_i^{(0)} - \mathbf{X}_l^{(0)})$  is the negative gradient. Unfortunately,  $H$  is unknown in the above formula.

Even if we can estimate  $H$  by  $\widehat{H}$ , we cannot estimate  $H^{-1}$  directly by  $\widehat{H}^{-1}$ , since the matrix  $\widehat{H}$  may not be invertible when the dimension  $p$  is larger than the sample size  $n_0$ . To address this, we adopt the approach in Cai et al. (2011) to obtain  $\widehat{H}^{-1}$ . Denote the estimator of  $H^{-1}$  as  $\widehat{\Theta}$ . Then we obtain  $\widehat{\Theta}$  via the following convex program:

$$\min_{\Theta \in \mathbb{R}^{p \times p}} \|\Theta\|_\infty, \quad \text{s.t.} \quad \|\Theta \widehat{H} - \mathbf{I}\|_{\max} \leq \gamma_n. \quad (8)$$

Following Cai et al. (2011), we use five-fold cross-validation to select  $\gamma_n$ . Finally, the desparsified estimator is defined as follows:

$$\widetilde{\beta} = \widehat{\beta}_{\mathcal{A}_h}^{(0)} + \widehat{\Theta} \widehat{\eta}. \quad (9)$$

The details about confidence interval construction are shown in Algorithm 2, and the corresponding theories are provided in Theorem 3.

---

**Algorithm 2** Confidence interval construction for the high-dimensional transformation model

---

**Require:** Target dataset  $\mathcal{D}^{(0)} = \left\{ Y_i^{(0)}, \Delta_i^{(0)}, \mathbf{X}_i^{(0)} \right\}_{i=1}^{n_0}$ ; transferring estimator  $\widehat{\beta}_{\mathcal{A}_h}^{(0)}$ ;

**Ensure:** Desparsified Lasso estimator  $\widetilde{\beta}$  and its confidence intervals  $\{\mathcal{I}_j\}_{j=1}^p$  ;

1: Compute the negative gradient  $\widehat{\eta}$  and Hessian matrix  $\widehat{H}$  using the target data:

$$\widehat{\eta} = \frac{1}{n_0(n_0-1)} \sum_{i \neq l} \Delta_l^{(0)} I(Y_i^{(0)} > Y_l^{(0)}) S'_n \left( \widehat{\beta}_{\mathcal{A}_h}^{(0)\top} (\mathbf{X}_i^{(0)} - \mathbf{X}_l^{(0)}) \right) (\mathbf{X}_i^{(0)} - \mathbf{X}_l^{(0)}),$$

and

$$\widehat{H} = -\frac{1}{n_0(n_0-1)} \sum_{i \neq l} \Delta_l^{(0)} I(Y_i^{(0)} > Y_l^{(0)}) S''_n \left( \widehat{\beta}_{\mathcal{A}_h}^{(0)\top} (\mathbf{X}_i^{(0)} - \mathbf{X}_l^{(0)}) \right) (\mathbf{X}_i^{(0)} - \mathbf{X}_l^{(0)}) (\mathbf{X}_i^{(0)} - \mathbf{X}_l^{(0)})^\top,$$

where  $S'_n(x) = S_n(x)(1 - S_n(x))/\sigma_n$  and  $S''_n(x) = S_n(x)(1 - S_n(x))(1 - 2S_n(x))/\sigma_n^2$ .

2: Compute  $\widehat{\Theta}$  by solving the optimization problem (8).

3: Compute the desparsified estimator:  $\widetilde{\beta} = \widehat{\beta}_{\mathcal{A}_h}^{(0)} + \widehat{\Theta} \widehat{\eta}$ .

4: Construct the confidence interval for  $\beta_j^{(0)}$ ,  $j = 1, \dots, p$ :

$$\mathcal{I}_j \leftarrow \left[ \widetilde{\beta}_j - \sqrt{\widehat{\Theta}_j^\top \widehat{G} \widehat{\Theta}_j} q_{\alpha/2} / \sqrt{n_0}, \widetilde{\beta}_j + \sqrt{\widehat{\Theta}_j^\top \widehat{G} \widehat{\Theta}_j} q_{\alpha/2} / \sqrt{n_0} \right],$$

where  $\widehat{G} = n_0^{-1} \sum_{l=1}^{n_0} \{ \nabla \widehat{\tau}_n(\mathbf{V}_l^{(0)}, \widehat{\beta}_{\mathcal{A}_h}^{(0)}) \nabla \widehat{\tau}_n^\top(\mathbf{V}_l^{(0)}, \widehat{\beta}_{\mathcal{A}_h}^{(0)}) \}$ , with  $\nabla \widehat{\tau}_n(v, \beta) = n_0^{-1} \sum_{i=1}^{n_0} \{ \Delta_i^{(0)} I(y \geq Y_i^{(0)}) S'_n(\beta^\top \mathbf{x} - \beta^\top \mathbf{X}_i^{(0)}) (\mathbf{x} - \mathbf{X}_i^{(0)})^\top + \delta I(Y_i^{(0)} \geq y) S'_n(\beta^\top \mathbf{X}_i^{(0)} - \beta^\top \mathbf{x}) (\mathbf{X}_i^{(0)} - \mathbf{x})^\top \}$ , in which  $\mathbf{V}_i^{(0)} = (\Delta_i^{(0)}, Y_i^{(0)}, \mathbf{X}_i^{(0)})$  and  $v = (\delta, y, x)$ . Here  $\widetilde{\beta}_j$  is the  $j$ -th component of  $\widetilde{\beta}$ , and  $q_{\alpha/2}$  is the  $\alpha/2$ -left tail quantile of  $\mathcal{N}(0, 1)$ .

5: Output the confidence intervals  $\{\mathcal{I}_j\}_{j=1}^p$ .

---

**Theorem 3.** Suppose that  $\widehat{\beta}_{\mathcal{A}_h}^{(0)}$  satisfies the estimation error bound in Theorem 1 with  $h \leq s\sqrt{\log p/n_0}$ , based on Conditions (C1) and (C5), as  $n_0 \rightarrow \infty$ , we have

$$\sqrt{n_0}(\widetilde{\beta} - \beta^{(0)}) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Gamma G H^{-1}), \quad (10)$$

where  $G$  is the asymptotic variance of  $\widehat{\eta}(\beta^{(0)})$ ,  $G = \mathbb{E}\{\nabla\tau(\mathbf{V}_i^{(0)}, \beta^{(0)})\nabla\tau^\top(\mathbf{V}_i^{(0)}, \beta^{(0)})\}$

with  $\nabla\tau(v, \beta) = \mathbb{E}\left\{\Delta_i^{(0)}I(y \geq Y_i^{(0)})S'_n(\beta^\top \mathbf{x} - \beta^\top \mathbf{X}_i^{(0)})(x - \mathbf{X}_i^{(0)}) + \delta I(Y_i^{(0)} \geq y)S'_n(\beta^\top \mathbf{X}_i^{(0)} - \beta^\top \mathbf{x})(\mathbf{X}_i^{(0)})\right\}$  in which  $\mathbf{V}_i^{(0)} = (\Delta_i^{(0)}, Y_i^{(0)}, \mathbf{X}_i^{(0)})$  and  $v = (\delta, y, x)$ .

In the proof of Theorem 3, we show that the desparsified estimator  $\widetilde{\beta}$  enjoys the following Bahadur representation:

$$\|\sqrt{n_0}(\widetilde{\beta} - \beta^{(0)}) - \sqrt{n_0}H^{-1}\widehat{\eta}(\beta^{(0)})\|_\infty = \mathcal{O}\left(\sqrt{\log p}\left(\sqrt{\frac{\log p}{n_0}}\|\widehat{\beta}_{\mathcal{A}_h}^{(0)} - \beta^{(0)}\|_1\right)^{1-q} + \sqrt{\log p}\|\widehat{\beta}_{\mathcal{A}_h}^{(0)} - \beta^{(0)}\|_1^2\right)$$

where  $\|\widehat{\beta}_{\mathcal{A}_h}^{(0)} - \beta^{(0)}\|_1 \lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h} + n_0}\right)^{1/4}\sqrt{sh} + h$  shown in Theorem 1.

This suggests that  $\sqrt{n_0}(\widetilde{\beta} - \beta^{(0)})$  can be expressed as a high-dimensional U-statistic up to some negligible terms. With this help, the asymptotic distribution of the estimators can be derived using the central limit theorem for U-statistic  $\widehat{\eta}(\beta^{(0)})$  (Song et al., 2007; Lee, 2019; Lin and Peng, 2013), as shown in Theorem 3.

### 3 Simulation studies

In this section, we conduct comprehensive simulations, evaluate the performance of the proposed approaches, and compare them against multiple alternatives. For simplicity, we refer to the first step of our Oracle-Trans algorithm as the fusion learning step, and the second step as the debiasing step. We compare the following six methods:

- **Target-Only:** Perform the naive estimation using only the target dataset.
- **Naive-Pooled:** Perform the fusion learning step by pooling all data together.
- **Oracle-Pooled:** Assume  $\mathcal{A}_h$  is known and performs the fusion learning step by minimizing (3) using both sources in  $\mathcal{A}_h$  and the target dataset without debiasing.
- **Oracle-Trans:** Perform the debiasing step using the target dataset after Oracle-Pooled.
- **Auto-Pooled:** Run Algorithm 1 to obtain  $\widehat{\mathcal{A}}_h$  and then performs the fusion learning step by minimizing (3) using sources in  $\widehat{\mathcal{A}}_h$  and the target dataset without debiasing.
- **Auto-Trans:** Perform the debiasing step using the target dataset after Auto-Pooled.

### 3.1 Data generation

When generating data, we consider several scenarios with varying discrepancies between the target and source models, different numbers of source datasets, and proportions of informative source datasets. We also consider different dimensions and sample sizes. Specifically, we generate target and source datasets from the accelerated failure time (AFT) models:

$$\log(T_i) = \boldsymbol{\beta}^{(k)\top} \mathbf{X}_i^{(k)} + \varepsilon_i, \quad k = 0, 1, \dots, K, \quad i = 1, \dots, n_k,$$

where  $n_0$  and  $n_k$  are sample sizes of the target dataset and the  $k$ -th source dataset, respectively. For the target dataset,  $\mathbf{X}_i^{(0)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}^{(0)})$ ,  $i = 1, \dots, n_0$  and  $\boldsymbol{\Sigma}^{(0)} = [0.3^{|j-j'|}]$  for  $j, j' = 1, \dots, p$ . For the  $k$ -th source dataset,  $k = 1, \dots, K$ ,  $\mathbf{X}_i^{(k)} \stackrel{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}^{(k)})$ ,  $i = 1, \dots, n_k$ , where  $\boldsymbol{\Sigma}^{(k)} = \boldsymbol{\Sigma}^{(0)} + \mathbf{v} \cdot \mathbf{v}^\top$  and  $\mathbf{v} \sim \mathcal{N}(\mathbf{0}_p, 0.3^2 \cdot \mathbf{I}_p)$ ,  $\mathbf{I}_p$  is a  $p$ -dimensional identity matrix. The random error  $\varepsilon_i \sim \mathcal{N}(0, 0.2)$  is independent of  $\mathbf{X}_i^{(k)}$ . The censoring time is generated from  $\text{Exp}(1/\theta)$ , where  $\theta$  is set to achieve the censoring rate about 40%.

We consider different dimensions of covariates. In Scenarios S1-S5 and S7, we set  $p = 200$  and  $s = \|\boldsymbol{\beta}^{(0)}\|_0 = 12$  with  $\boldsymbol{\beta}^{(0)} = (1 \cdot \mathbf{1}_2^\top, -1 \cdot \mathbf{1}_2^\top, 0.8 \cdot \mathbf{1}_2^\top, -0.8 \cdot \mathbf{1}_2^\top, 0.6 \cdot \mathbf{1}_2^\top, -0.6 \cdot \mathbf{1}_2^\top, \mathbf{0}_{p-s})^\top$ , where  $\mathbf{1}_2$  is a two-dimensional vector of all 1,  $\mathbf{0}_{p-s}$  is a  $p-s$ -dimensional vector of all 0. In Scenario 6, we set  $p = 500$  and  $s = 24$  with  $\boldsymbol{\beta}^{(0)} = (1 \cdot \mathbf{1}_4^\top, -1 \cdot \mathbf{1}_4^\top, 0.8 \cdot \mathbf{1}_4^\top, -0.8 \cdot \mathbf{1}_4^\top, 0.6 \cdot \mathbf{1}_4^\top, -0.6 \cdot \mathbf{1}_4^\top, \mathbf{0}_{p-s})^\top$ . In these scenarios, the coefficients  $\boldsymbol{\beta}^{(k)}$  for the source data are generated by perturbing  $\boldsymbol{\beta}^{(0)}$ . Specifically, let  $\mathcal{J} = \{j : \beta_j^{(0)} \neq 0\}$  and  $\mathcal{J}^c = \{j : \beta_j^{(0)} = 0\}$ . For the  $k$ -th source, we construct three index subsets  $\mathcal{J}_1^{(k)}$ ,  $\mathcal{J}_2^{(k)}$  and  $\mathcal{J}_3^{(k)}$  with sizes  $d_1$ ,  $d_2$  and  $r$ , respectively. The elements in  $\mathcal{J}_1^{(k)}$  and  $\mathcal{J}_3^{(k)}$  are randomly selected from  $\mathcal{J}$  while the elements in  $\mathcal{J}_2^{(k)}$  are randomly selected from  $\mathcal{J}^c$ . The perturbations are performed on the corresponding coefficients as follows:  $\beta_j^{(k)} = \beta_j^{(0)} + \epsilon_j^{(k)}$  for  $j \in \mathcal{J}_1^{(k)}$  and  $\mathcal{J}_2^{(k)}$ ;  $\beta_j^{(k)} = -\beta_j^{(0)}$  for  $j \in \mathcal{J}_3^{(k)}$ , where  $\epsilon_j^{(k)}$  are generated from the uniform distribution  $U[-u, u]$ . To avoid non-identifiability, all coefficients are normalized such that  $\|\boldsymbol{\beta}^{(k)}\|_2 = 1$ , for  $k = 0, 1, \dots, K$ . For each scenario,  $d_1$ ,  $d_2$ ,  $r$  and  $u$  are carefully designed to generate informative and non-informative sources. Here, we introduce the estimated rank correlation (ERC) to assess the similarity between the target and source coefficients:

$$\text{ERC}(\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(k)}) = \frac{\sum_{i \neq j} I(\beta_i^{(0)} > \beta_j^{(0)}) I(\beta_i^{(k)} > \beta_j^{(k)})}{\sum_{i \neq j} I(\beta_i^{(0)} > \beta_j^{(0)})},$$

where a larger ERC indicates a more useful dataset. For helpful datasets  $k \in \mathcal{A}_h$ , we set  $d_1 = 2$  or  $4$ ,  $d_2 = 4$ ,  $u = 0.3$  or  $0.4$ ,  $r = 2$ , to achieve that the  $\text{ERC} > 0.8$ . For unhelpful datasets  $k \in \mathcal{A}_h^c$ , we set  $d_1 = 6$ ,  $d_2 = 6$ ,  $u = 1$ ,  $r = 7$ , and the corresponding  $\text{ERC} < 0.5$ . Specifically, we consider the following seven scenarios:

- S1:  $n_0 = 100$ ,  $n_k = 200$ ,  $p = 200$ ,  $K = 2$ ,  $|\mathcal{A}_h| = 1$ . For the helpful source, we set  $d_1 = 2$ ,  $d_2 = 4$ ,  $r = 2$ ,  $u = 0.3$ . For the unhelpful source, we set  $d_1 = 6$ ,  $d_2 = 6$ ,  $r = 7$ ,  $u = 1$ .

- S2:  $n_0 = 100$ ,  $n_k = 200$ ,  $p = 200$ ,  $K = 2$ ,  $|\mathcal{A}_h| = 1$ . For the helpful source, we set  $d_1 = 4$ ,  $d_2 = 4$ ,  $r = 2$ ,  $u = 0.4$ . For the unhelpful source, we set  $d_1 = 6$ ,  $d_2 = 6$ ,  $r = 7$ ,  $u = 1$ .
- S3:  $n_0 = 100$ ,  $n_k = 100$ ,  $p = 200$ ,  $K = 6$ ,  $|\mathcal{A}_h| = 3$ . Other settings are the same as S1.
- S4:  $n_0 = 100$ ,  $n_k = 60$ ,  $p = 200$ ,  $K = 6$ ,  $|\mathcal{A}_h| = 3$ . Other settings are the same as S1.
- S5:  $n_0 = 100$ ,  $n_k = 100$ ,  $p = 200$ ,  $K = 6$ ,  $|\mathcal{A}_h| = 2$ . Other settings are the same as S1.
- S6:  $n_0 = 200$ ,  $n_k = 200$ ,  $p = 500$ ,  $K = 6$ ,  $|\mathcal{A}_h| = 3$ . For the helpful source, we set  $d_1 = 4$ ,  $d_2 = 10$ ,  $r = 2$ ,  $u = 0.4$ . For the unhelpful source, we set  $d_1 = 12$ ,  $d_2 = 10$ ,  $r = 14$ ,  $u = 1$ .
- S7:  $n_0 = 60$ ,  $n_k = 60$ ,  $p = 200$ ,  $K = 10$ ,  $|\mathcal{A}_h| = 3, 6, 9$ . For the helpful source, we set  $d_1 = 2$ ,  $d_2 = 3$ ,  $r = 2$ ,  $u = 0.3$ . For the unhelpful source, we set  $d_1 = 7$ ,  $d_2 = 7$ ,  $r = 9$ ,  $u = 1$ .

In S1, there are two sources with  $\text{ERC}(\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(k)}) = 0.834$  for  $k \in \mathcal{A}_h$  and 0.418 for  $k \in \mathcal{A}_h^c$ . Compared to S1, more perturbations are added in S2, while S3 considers more sources and retains the same perturbations as in S2. Compared with S3, the sample sizes of the source datasets in S4 are reduced while the proportion of helpful sources in S5 is reduced. In S6, we consider a setting with higher dimensions  $p = 500$ . Besides, we aim to investigate how the various methods perform when the proportion of helpful source datasets is increased through S7. To evaluate the methods, we consider the following measurements: (1) F1-score for assessing variable selection accuracy; (2) RMSE of the estimates for assessing estimation accuracy:  $\text{RMSE} = \left\{ \sum_{j=1}^p (\hat{\beta}_j - \beta_j)^2 \right\}^{1/2}$ ; (3) C-index  $C(\hat{\boldsymbol{\beta}}; \mathcal{D}^*)$  in (5) for evaluating prediction accuracy. It should be noted that the C-index is calculated based on a testing set  $\mathcal{D}^*$  of the target domain data, which is generated independently with sample size  $n_* = 30$ . For each scenario, the results are summarized based on 500 replications.

Additionally, to evaluate the accuracy of the proposed method for constructing confidence intervals, we also conduct a simulation. We use three metrics for evaluation: (1) the bias of the estimator before and after the debiasing process; (2) the coverage probability of the confidence interval, which is the proportion of times the confidence interval covers the true value in 500 repetitions; (3) the length of the confidence interval. We considered two types of variables: signal variables with nonzero coefficients and noise variables with zero coefficients. Due to space limitations, the details of the scenario and simulation results are placed in Supplementary Materials.

### 3.2 Simulation results

Across the entire spectrum of simulation, the proposed Auto-Trans method is observed to have performance either at or near the best in variable selection, estimation accuracy, and prediction accuracy. Specifically, the following conclusions are obtained:

(1). The box-plots of F1-score in Figure 2 show that transfer learning offers advantages in variable selection over the Target-only and Naive-Pooled methods. The four transfer learning related methods produce much larger F1-scores than the other two methods. In almost all settings, Auto-Pooled performs similarly to Oracle-Pooled, while Auto-Trans yields results very close to Oracle-Trans in terms of F1-score. This further reconfirms the accuracy of the detection algorithm for identifying informative sources.

(2). From the results of the C-index presented in Figure 3, we find that both Target-only and Naive-Pooled are inferior to the other four methods related to transfer learning. The Naive-Pooled, in particular, produces the smallest C-index. Among the four transfer learning methods, Oracle-Trans performs the best as expected, since it knows accurately and utilizes the useful sources. Auto-Trans also shows an absolute advantage and is comparable to Oracle-Trans, indicating that the proposed Algorithm 1 can accurately detect the informative sources. Besides, the debiasing step does help to improve the estimates since we observe that Oracle-Trans performs better than Oracle-Pooled and Auto-Trans performs better than Auto-Pooled.

(3). Figure 4 shows the comparison of RMSE among different methods. We can see that Oracle-Trans produces the smallest RMSE, as it accurately utilizes useful data sources, and Auto-Trans has a performance closer to Oracle-Trans in terms of RMSE, which confirms the superiority of the approach of detection for helpful sources. Besides, the transfer learning based methods have a clear advantage over Target-only and Naive-Pooled in estimation accuracy, since Target-Only ignores the helpful sources while Naive-Pooled incorporates many noisy sources. In addition, Oracle-Trans exhibits superior estimation accuracy compared to Oracle-Pooled, empirically confirming the importance of the debiasing step.

(4). By comparing the results across different scenarios, we find that the more similar the source and target data are, the more helpful it is to improve the estimates. Comparing the results in S2 and S3, we can see that the transfer learning methods have much better performances in terms of C-index and F1-score when the number of informative sources increased from one to three. The results from S3 and S4 show that there is a decrease in variable selection, estimation, as well as prediction accuracy as the sample size  $n_k$  drops from 100 to 60 for the transfer learning methods. The performance of transfer learning methods in S5 is slightly worse than in S3 since the proportion of helpful sources is smaller. In the high-dimensional scenario S6, where  $n_0$  is much smaller than  $p$ , the results show that the Fabs algorithm remains effective and the proposed methods also demonstrate superior performances.

(5). Due to the space limit, we put the results of S7 in Supplementary Materials. It is observed that as the number of helpful source datasets increases, the C-index gradually increases while the RMSE decreases. Similar to the results in scenarios S1-S6, Oracle-Trans outperforms Oracle-Pooled, while Auto-Trans outperforms Auto-Pooled, further illustrating the benefits of using target data for debiasing. Furthermore, as the number of helpful source datasets increases, the gap between Auto-Trans and Oracle-Trans narrows. This indicates that our proposed method is capable of attaining a performance comparable to that of the Oracle estimator when the number of informative source datasets is considerable. Besides, we cal-

culate the Recall for the informative source dataset identification process, and the results show that the methods can accurately identify the informative sources with large Recall values.

(6). The simulation results in Supplementary Materials for constructing confidence intervals show that the bias of the desparsified Lasso estimator is significantly reduced. In a large number of repeated simulations, the confidence intervals constructed by the proposed algorithm cover the true parameter value approximately 95% of the time. Additionally, we plot the histograms of the desparsified Lasso estimator  $\tilde{\beta}$ . The empirical distributions are in high agreement with the standard normal density, providing strong numerical evidence for the asymptotic normality established in Theorem 3.

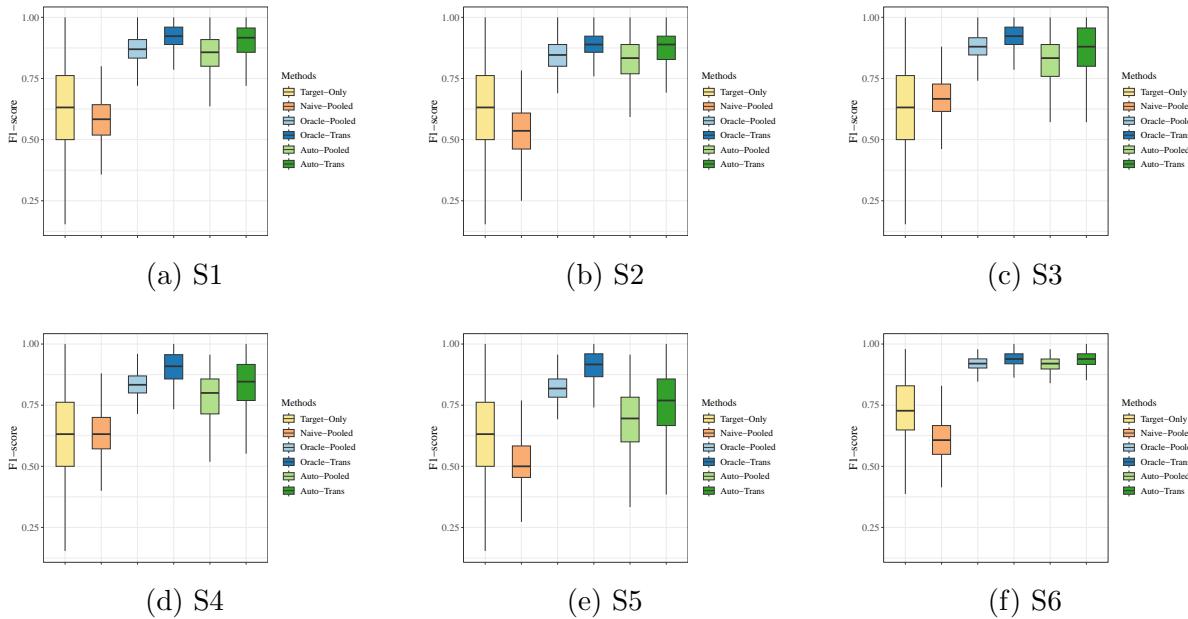


Figure 2: Simulation results for F1-score under Scenarios S1-S6

## 4 Real data analysis

### 4.1 Data preparation

In this section, we focus on the sepsis cohort within the MIMIC-IV database (<https://mimic.mit.edu>), a contemporary electronic health record dataset spanning admissions from 2008 to 2019 (Johnson et al., 2023). Our research targets patients admitted to the ICU for the first time with sepsis. As mentioned in the introduction, MSSA is a highly significant and potentially lethal bacteria, yet current research on MSSA is insufficient and limited by inadequate data. To address this gap, we aim to utilize transfer learning tools for a more comprehensive analysis. Here the MSSA patient data serve as the target dataset, and other sepsis cases as source datasets. Sepsis cases with fewer than 40 samples and those of unspecified etiology are excluded, resulting in a total of nine source datasets, as shown

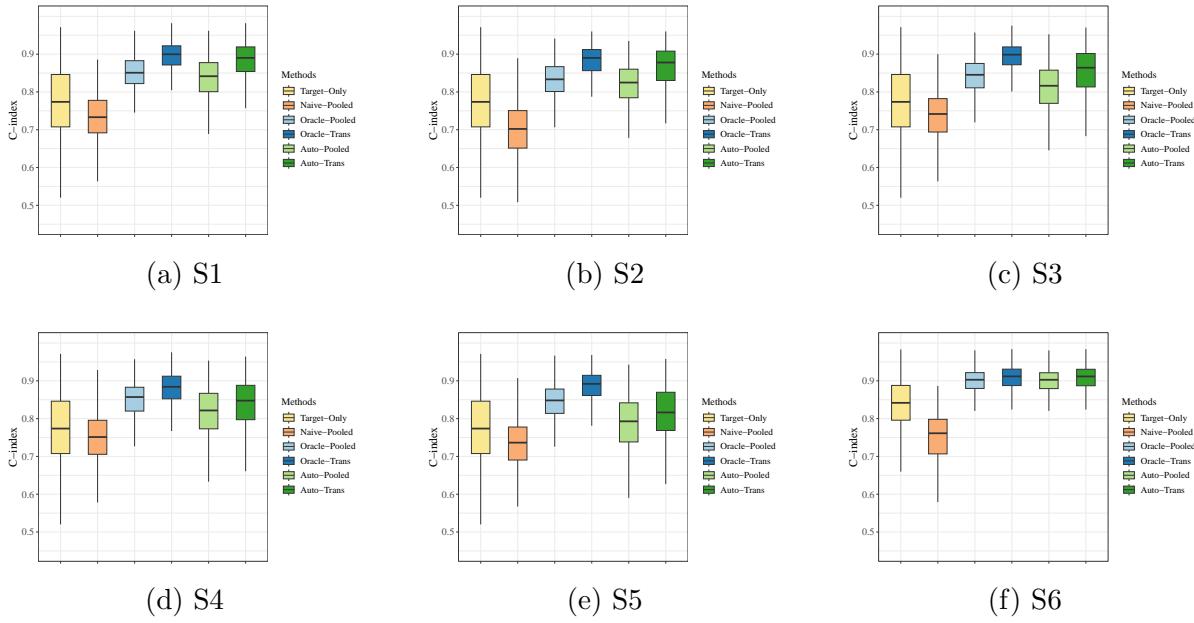


Figure 3: Simulation results for C-index under Scenarios S1-S6

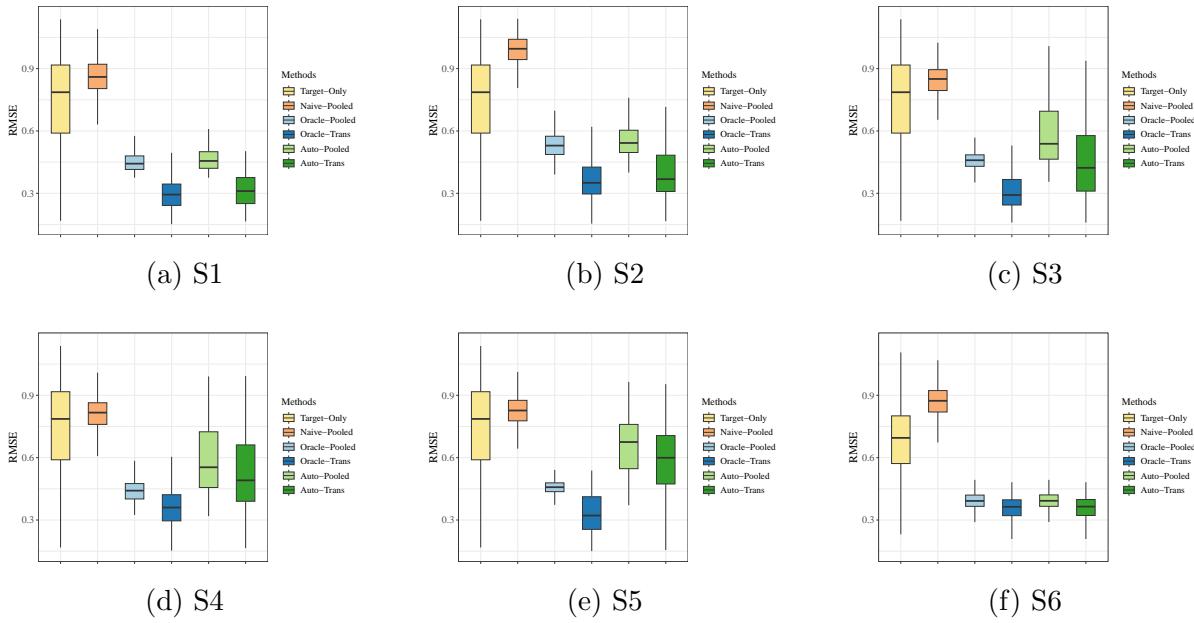


Figure 4: Simulation results for RMSE under Scenarios S1-S6

in Table 1. Table 1 demonstrates the ICD-code, sample size, and corresponding sepsis type for the nine source datasets. It also shows the gains in C-index for each source dataset using our proposed detection algorithm in Algorithm 1, and seven source datasets with positive gains in C-index can improve the learning of the target model, while the other two with negative gains in C-index are not helpful.

The sepsis dataset to be analyzed includes one target dataset and nine source datasets, with a total sample size of 1700. Among these, 347 patients died of sepsis during the study periods leading to a censoring rate of 79.58%. We denote  $T$  as the period from ICU admission to in-hospital death. For those who survive until hospital discharge, the survival time is censored with  $C$  being the gap time between the discharge and the ICU admission dates. For each patient, 279 features are recorded but with a high missing rate, and variables with a missing rate over 25% are discarded in our analysis. For continuous variables with a missing rate less than 10%, mean imputation is employed, and variables with a missing rate 10% – 25% undergo multiple imputations. Missing values in binary categorical variables are filled in using the mode. Ultimately, 102 covariates are included, categorized into: i) Demographic characteristics: age at admission, gender, and ethnicity; ii) Haematological assessments on the first day of ICU admission: haemoglobin, platelet, and white blood cell counts, electrolyte balance, renal and liver function tests, and coagulation profiles; iii) Arterial blood gas evaluations performed on the first day of ICU admission: lactate, pH, and oxygen saturation; iv) Scores measuring organ failure and illness severity: LODS, SOFA, APS III, etc.

## 4.2 Results

Table 1: Description of the target and source datasets

Dataset	ICD-10	Description	Sample size	Gains in C-index	Selection times
Target	A4101	Sepsis due to Methicillin susceptible <i>Staphylococcus aureus</i>	229	/	/
S1	A4102	Sepsis due to Methicillin resistant <i>Staphylococcus aureus</i>	98	-0.046	40
S2	A408	Other streptococcal sepsis	79	0.003	69
S3	A411	Sepsis due to other specified <i>staphylococcus</i>	74	0.001	89
S4	A4151	Sepsis due to <i>Escherichia coli</i> [E. coli]	420	-0.036	22
S5	A4181	Sepsis due to <i>Enterococcus</i>	207	0.042	80
S6	A4152	Sepsis due to <i>Pseudomonas</i>	88	0.006	66
S7	A4189	Other specified sepsis	270	0.041	96
S8	A4150	Gram-negative sepsis, unspecified	64	0.027	98
S9	A4159	Other Gram-negative sepsis	171	0.028	83

We use Target-Only, Naive-Pooled, and Auto-Trans methods for estimation due to the unknown true informative sources. The feature extraction results are shown in Table 2. Our Auto-Trans method identifies 12 features, and Target-Only and Naive-Pooled identify 11 and 17 features, respectively. All three methods select admission age, Sequential Organ Failure Assessment score (SOFA), logistic organ dysfunction system (LODS) score, and three hematological assessment (first ICU day) indexes. Specifically, the estimated coefficients for admission age are negative in all three methods, indicating that older patients have shorter survival times and a higher risk of death. Additionally, SOFA is commonly used as a measure of organ dysfunction and has a high discriminative ability for predicting emergency

and in-hospital mortality (Toker et al., 2021), which is consistent with the negative coefficient estimates. It is worth noting that Auto-Trans identifies the minimum total carbon dioxide level ( $\text{TCO}_2$ ) in blood gases, which is missed by both Target-Only and Naive-Pooled. Literature suggests that this feature is important, as low  $\text{TCO}_2$  levels are associated with higher risks of all-cause mortality (Yang et al., 2023), which is consistent with the negative estimation coefficient.

Table 2: Data analysis: results of variable selection

Variable	Target-Only	Naive-Pooled	Auto-Trans
Admission age	-0.609	-0.481	-0.611
Maximum Partial Thromboplastin Time (PTT) recorded	-0.406	-0.160	-0.122
Logistic Organ Dysfunction System (LODS)	-0.305	-0.480	-0.489
Minimum anion gap recorded	-0.203	-0.160	-0.122
Minimum Partial Thromboplastin Time (PTT) recorded	-0.203	-0.160	-0.244
Sequential Organ Failure Assessment score (SOFA)	-0.102	-0.160	-0.244
Acute Physiology Score III	-0.203	-0.160	\
Minimum blood lactate level recorded	\	-0.320	-0.366
Verbal response score from the Glasgow Coma Scale (GCS)	\	0.320	0.122
Minimum heart rate recorded in vital signs	\	-0.160	-0.244
Mean respiratory rate recorded in vital signs	\	-0.160	-0.122
Simplified Acute Physiology Score II (SAPS II)	\	-0.160	-0.122
Minimum hematocrit level recorded	-0.102	\	\
Motor response score from the Glasgow Coma Scale (GCS)	0.203	\	\
Maximum partial pressure of oxygen ( $\text{PaO}_2$ ) in blood gases recorded	0.305	\	\
Minimum systolic blood pressure recorded in vital signs	0.305	\	\
Maximum hemoglobin level recorded	\	0.160	\
Maximum blood glucose level recorded	\	0.160	\
Minimum absolute eosinophil count recorded	\	0.160	\
Minimum total carbon dioxide level in blood gases recorded	\	\	-0.030
Mean peripheral capillary oxygen saturation ( $\text{SpO}_2$ ) recorded in vital signs	\	0.160	\
Maximum glucose level recorded in vital signs	\	-0.160	\

To gain further insights into the analysis results, we conduct a random splitting-based evaluation. Since most of the source datasets are informative for predicting the target, we compare the performances of the three methods by reducing the number of informative sources. We randomly select 20% of the target data as the testing set, with the remaining data used for training. The model is trained on the training set, and the C-index and Log-rank values are calculated on the testing set to evaluate the performance of the methods. Larger values of the Log-rank statistic indicate more significant differences in the survival curves between the high- and low-risk groups (Harrington and Fleming, 1982), which in turn suggests better predictive accuracy of the model. Here, the high-risk and low-risk differences are delineated by the median of the model’s predicted values  $\hat{\beta}^\top \mathbf{X}$ . This process is repeated 100 times to evaluate the performance of the various methods.

Among the nine sources, S1 and S4 are not helpful for promoting the target learning due to their negative gains in C-index, and we will keep them in the following evaluation process. We consider the following three scenarios: Scenario I pretends to have only three sources, S1, S4, and S8, where S8 has the largest selection times while S1 and S4 are the two sources with the smallest selection times (Table 1); In Scenario II, we consider six sources including S1, S2, S3, S4, S8, and S9; Scenario III includes all nine sources. Figure 5 shows the results based on 100 replications.

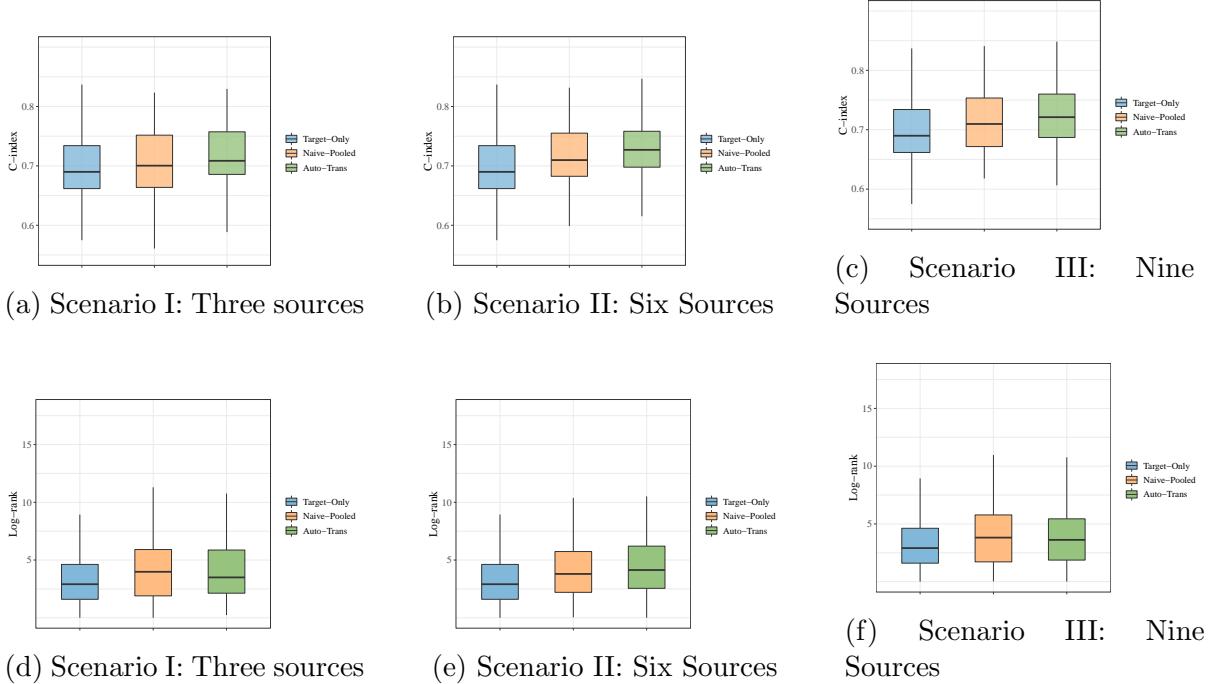


Figure 5: Data analysis: results for random splitting-based evaluation

The results show that Auto-Trans enjoys advantages in prediction accuracy over Naive-Pooled and Target-Only, since it tends to produce the largest C-index and Log-rank statistics. Target-Only has the poorest prediction performances due to the insufficient sample size. As the number of sources increases, the performance of Auto-Trans and Naive-Pooled improve in prediction accuracy as the proportion of helpful sources increases. Specifically, the mean C-index values for Auto-Trans in Scenarios I, II, and III are 0.717, 0.720, and 0.722 respectively, which are larger than the mean values (0.702, 0.714, 0.713) for Naive-Pooled, and 0.697 for Target-Only. For the three scenarios, the mean values of Log-rank statistics are (4.490, 4.572, 4.306) for Auto-Trans, (4.218, 4.156, 3.996) for Naive-Pooled, and 3.371 for Target-Only. Overall, clear advantages of transfer learning are again observed.

## 5 Conclusion

Transfer learning is a powerful tool to enhance model performance on the target dataset by leveraging information from other source datasets with similar but not exactly the same distributions. In this study, we have proposed a transfer learning approach (Auto-Trans) to address the pressing challenge of analyzing high-dimensional time-to-event data in the context of sepsis caused by MSSA. Given the complexity and high-dimensional nature of sepsis data, Auto-Trans, designed for transformation models, offers the flexibility of semiparametric models by capturing the relationship between survival time and predictors without relying on restrictive parametric assumptions. A key innovation of our method is the development of a transferable source detection mechanism based on the C-index, which

can consistently identify informative sources and ensure that valuable information from related datasets is appropriately integrated. Statistical validity is rigorously established. Furthermore, the confidence intervals for each coefficient component are provided with theoretical guarantees. Simulation results have shown that the proposed approach demonstrates competitive performance in enhancing variable selection, estimation, and prediction accuracy. When applied to the sepsis dataset, our method reveals findings that differ from alternative approaches, reaffirming its practical superiority.

Overall, our research contributes to the growing body of literature on transfer learning in survival analysis, offering a robust and scalable solution for high-dimensional time-to-event data with right-censored. Future work may explore the application of Auto-Trans to other data settings, such as truncated data and interval-censored data. Additionally, when several source datasets are available for analysis, it can be important and challenging to recognize the helpful source datasets to avoid negative transfer. A natural question is whether there is a technique that can avoid this effect of misidentification and it would be helpful to develop such methods that can adaptively use the non-informative sources without worrying about the negative transfer. Besides, Li et al. (2024) and He et al. (2024) recently proposed a new framework that jointly incorporates losses from both the target and source domains. In the scenario considered in this paper, applying this new framework yields the objective function:

$$\begin{aligned}\tilde{Q}_n^{\mathcal{A}_h} = & -\sum_{k \in \mathcal{A}_h} \frac{\alpha_k}{n_k(n_k-1)} \sum_{i \neq j} \Delta_i^{(k)} I(Y_j^{(k)} > Y_i^{(k)}) S_n \left( (X_j^{(k)} - X_i^{(k)})^\top (\delta^{(k)} + \beta^{(0)}) \right) \\ & - \frac{\alpha_0}{n_0(n_0-1)} \sum_{i \neq j} \Delta_i^{(0)} I(Y_j^{(0)} > Y_i^{(0)}) S_n \left( X_j^{(0)\top} \beta^{(0)} - X_i^{(0)\top} \beta^{(0)} \right) \\ & + \lambda_0 \|\beta^{(0)}\|_1 + \sum_{k \in \mathcal{A}_h} \lambda_k \|\delta^{(k)}\|_1.\end{aligned}$$

We have compared the numerical performance of our proposed approach against this new framework through simulations in Supplementary material. However, the theoretical understanding in the differences between the two frameworks are complicated, and we leave this topic for future work.

## Acknowledgements

Lin's work was supported by the MOE Project of the Key Research Institute of Humanities and Social Sciences (22JJD910001).

## References

- Cai, L., Guo, X., Lian, H., and Zhu, L. (2024). Statistical inference for high-dimensional convoluted rank regression. *arXiv preprint arXiv:2405.14652*.
- Cai, T., Liu, W., and Luo, X. (2011). A constrained  $\ell_1$  minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607.

- Chen, S. (2002). Rank estimation of transformation models. *Econometrica* **70**, 1683–1697.
- Chu, J., Lu, W., and Yang, S. (2023). Targeted optimal treatment regime learning using summary statistics. *Biometrika* **110**, 913–931.
- Evans, L., Rhodes, A., Alhazzani, W., Antonelli, M., Coopersmith, C. M., French, C., Machado, F. R., McIntyre, L., Ostermann, M., Prescott, H. C., et al. (2021). Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Critical Care Medicine* **49**, e1063–e1143.
- Faix, J. D. (2013). Biomarkers of sepsis. *Critical Reviews in Clinical Laboratory Sciences* **50**, 23–36.
- Fleischmann-Struzek, C., Mellhammar, L., Rose, N., Cassini, A., Rudd, K., Schlattmann, P., Allegranzi, B., and Reinhart, K. (2020). Incidence and mortality of hospital-and icu-treated sepsis: results from an updated and expanded systematic review and meta-analysis. *Intensive Care Medicine* **46**, 1552–1562.
- Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553–566.
- He, Z., Sun, Y., and Li, R. (2024). Transfusion: Covariate-shift robust transfer learning for high-dimensional regression. In *International Conference on Artificial Intelligence and Statistics*, pages 703–711. PMLR.
- Hu, X. and Zhang, X. (2023). Optimal parameter-transfer learning by semiparametric model averaging. *Journal of Machine Learning Research* **24**, 1–53.
- Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data* **10**, 1.
- Kavanagh, K. T. (2019). Control of mssa and mrsa in the united states: protocols, policies, risk adjustment and excuses. *Antimicrobial Resistance & Infection Control* **8**, 103.
- Khan, S. and Tamer, E. (2007). Partial rank estimation of duration models with general forms of censoring. *Journal of Econometrics* **136**, 251–280.
- Kourtis, A. P. (2019). Vital signs: epidemiology and recent trends in methicillin-resistant and in methicillin-susceptible staphylococcus aureus bloodstream infections—united states. *MMWR. Morbidity and Mortality Weekly Report* **68**, 214–219.
- Lee, A. J. (2019). *U-statistics: Theory and Practice*. Routledge.
- 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: Statistical Methodology* **84**, 149–173.

- Li, S., Zhang, L., Cai, T. T., and Li, H. (2024). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association* **119**, 1274–1285.
- Li, Z., Shen, Y., and Ning, J. (2023). Accommodating time-varying heterogeneity in risk estimation under the cox model: a transfer learning approach. *Journal of the American Statistical Association* **118**, 2276–2287.
- Lin, H. and Peng, H. (2013). Smoothed rank correlation of the linear transformation regression model. *Computational Statistics & Data Analysis* **57**, 615–630.
- Liu, R., Shi, Y., Ji, C., and Jia, M. (2019). A survey of sentiment analysis based on transfer learning. *IEEE Access* **7**, 85401–85412.
- Luhr, R., Cao, Y., Soederquist, B., and Cajander, S. (2019). Trends in sepsis mortality over time in randomised sepsis trials: a systematic literature review and meta-analysis of mortality in the control arm, 2002–2016. *Critical Care* **23**, 1–9.
- Marra, A. R., Bar, K., Bearman, G. M., Wenzel, R. P., and Edmond, M. B. (2006). Systemic inflammatory response syndrome in nosocomial bloodstream infections with pseudomonas aeruginosa and enterococcus species: Comparison of elderly and nonelderly patients. *Journal of the American Geriatrics Society* **54**, 804–808.
- Mo, W., Qi, Z., and Liu, Y. (2021). Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association* **116**, 659–674.
- Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* pages 158–195.
- Paoli, C. J., Reynolds, M. A., Sinha, M., Gitlin, M., and Crouser, E. (2018). Epidemiology and costs of sepsis in the united states—an analysis based on timing of diagnosis and severity level. *Critical Care Medicine* **46**, 1889–1897.
- Parker, S. and Watkins, P. (2001). Experimental models of gram-negative sepsis. *British Journal of Surgery* **88**, 22–30.
- Petegrosso, R., Park, S., Hwang, T. H., and Kuang, R. (2017). Transfer learning across ontologies for phenotype–genome association prediction. *Bioinformatics* **33**, 529–536.
- Pruksachatkun, Y., Phang, J., Liu, H., Htut, P. M., Zhang, X., Pang, R. Y., Vania, C., Kann, K., and Bowman, S. R. (2020). Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *arXiv preprint arXiv:2005.00628* .
- Qiao, S., He, Y., and Zhou, W. (2023). Transfer learning for high-dimensional quantile regression with statistical guarantee. *Transactions on Machine Learning Research* .

- Radhakrishnan, A., Luyten, M. R., Prasad, N., and Uhler, C. (2023). Transfer learning with kernel methods. *Nature Communications* **14**, 5570.
- Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., Colombara, D. V., Ikuta, K. S., Kissoon, N., Finfer, S., et al. (2020). Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet* **395**, 200–211.
- Shi, X., Huang, Y., Huang, J., and Ma, S. (2018). A forward and backward stage-wise algorithm for nonconvex loss functions with adaptive lasso. *Computational Statistics & Data Analysis* **124**, 235–251.
- Song, X., Ma, S., Huang, J., and Zhou, X.-H. (2007). A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics* **8**, 197–211.
- Tan, K. M., Wang, L., and Zhou, W.-X. (2022). High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**, 205–233.
- Tian, Y. and Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association* **118**, 2684–2697.
- Toker, A. K., Kose, S., and Turken, M. (2021). Comparison of sofa score, sirs, qsofa, and qsofa+1 criteria in the diagnosis and prognosis of sepsis. *The Eurasian Journal of Medicine* **53**, 40–47.
- Uehara, M., Kato, M., and Yasui, S. (2020). Off-policy evaluation and learning for external validity under a covariate shift. *Advances in Neural Information Processing Systems* **33**, 49–61.
- Wang, M. and Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing* **312**, 135–153.
- Wu, L. and Yang, S. (2023). Transfer learning of individualized treatment rules from experimental to real-world data. *Journal of Computational and Graphical Statistics* **32**, 1036–1045.
- Yang, C. H., Chen, Y.-A., Bin, P.-J., Ou, S.-M., and Tarn, D.-C. (2023). Associations of the serum total carbon dioxide level with long-term clinical outcomes in sepsis survivors. *Infectious Diseases and Therapy* **12**, 687–701.
- Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76**, 217–242.
- Zhang, W., Li, R., Zeng, T., Sun, Q., Kumar, S., Ye, J., and Ji, S. (2015). Deep model based transfer and multi-task learning for biological image analysis. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1475–1484.