BMC Bioinformatics

RESEARCH Open Access

Transfer learning for accelerated failure time model with microarray data

Yan-Bo Pei[1], Zheng-Yang Yu[1] and Jun-Shan Shen[1*]

*Correspondence: shenjunshan@cueb.edu.cn

[1] School of Statistics, Capital University of Economics and Business, Beijing, China

Abstract

Background: In microarray prognostic studies, researchers aim to identify genes associated with disease progression. However, due to the rarity of certain diseases and the cost of sample collection, researchers often face the challenge of limited sample size, which may prevent accurate estimation and risk assessment. This challenge necessitates methods that can leverage information from external data (i.e., source cohorts) to improve gene selection and risk assessment based on the current sample (i.e., target cohort).

Method: We propose a transfer learning method for the accelerated failure time (AFT) model to enhance the fit on the target cohort by adaptively borrowing information from the source cohorts. We use a Leave-One-Out cross validation based procedure to evaluate the relative stability of selected genes and overall predictive power.

Conclusion: In simulation studies, the transfer learning method for the AFT model can correctly identify a small number of genes, its estimation error is smaller than the estimation error obtained without using the source cohorts. Furthermore, the proposed method demonstrates satisfactory accuracy and robustness in addressing heterogeneity across the cohorts compared to the method that directly combines the target and the source cohorts in the AFT model. We analyze the GSE88770 and GSE25055 data using the proposed method. The selected genes are relatively stable, and the proposed method can make an overall satisfactory risk prediction.

Keywords: Survival analysis, Auxiliary studies, Gene expression data, Weighted least squares, Transfer learning

Introduction

Modeling the time to disease relapse or death of patients is a crucial aspect of clinical research, especially for chronic diseases like cancer and lymphoma. With the development of high throughput technologies, an important application is to identify genomic markers that are associated with time-to-event outcomes (e.g., [1, 2]). However, when the research data are collected from a single institution or clinical trial, researchers often face the challenge of limited sample size. For example, we consider a relatively rare subtype of breast cancer, named invasive lobular carcinoma (ILC), here. Although ILC accounts for only approximately 5–15% of all breast cancer cases [3], compared to invasive ductal carcinoma (IDC), the most common type of breast cancer, patients with ILC are unlikely to achieve improved outcomes through conventional treatment approaches [3–5]. A study on prognostic analysis of patients with ILC using microarray data has been reported in [6]. This study demonstrated the prognostic value of genomic grade. However, due to the rarity of disease cases and the cost of gene sequencing, the size of samples collected in the study is far from being satisfactory, which poses difficulties in estimating the effects of individual genes and assessing the risk for patients. To address this issue, leveraging information from outside data is a promising solution. Public functional genomics data repositories, such as the Gene Expression Omnibus (GEO) [7] and The Cancer Genome Atlas Program (TCGA) [8], have been increasingly used as auxiliary data sources because of their reliability, easy accessibility and large sample size.

Leveraging information from outside data (i.e., the source cohorts) to enhance the analysis of the target cohort (e.g., the ILC cohort in the motivating example) offers a viable solution to the limited sample size problem in microarray prognostic studies. However, this approach may encounter the challenge of data heterogeneity. In our motivating example, due to worse prognosis and different pathogenesis, there may exist heterogeneity between patients with ILC and those diagnosed with other types of breast cancer. In the field of machine learning, transfer learning [9] provides a robust framework for borrowing information adaptively from related tasks or datasets, even in the presence of certain heterogeneity. Transfer learning has been widely applied in medical research, including medical diagnosis [10], biological imaging analysis [11], and drug sensitivity prediction [12]. Recently, several transfer learning based statistical methods have been studied. A transfer learning method for high-dimensional linear regression has been proposed in [13], which quantifies the heterogeneity using the difference between target and source coefficients. Tian et al. [14] introduces a transfer learning framework into polygenic risk scores (PRS) based on linear regression. Tian and Feng [15] and [16] extend the findings of [13] to the generalized linear model. For time-to-event outcomes, a transfer learning method for the Cox model has been proposed in [17], which allows different levels of information borrowing in the regression coefficients and baseline hazards through tuning parameters. However, in microarray prognostic studies mentioned earlier, transfer learning methods for standard survival analysis techniques are unsuitable for handling high-dimensional gene expression data. Moreover, in such data, only a small subset of genes are typically correlated with disease progression, making it critical to accurately identify these relevant genes. The transfer learning method to analyze the time-to-event outcomes with microarray gene expression data is still worth exploring.

For the aforementioned transfer learning problem in high-dimensional survival analysis, the AFT model is a promising approach. Compared to the Cox proportional hazards model (e.g., [18, 19]) and the additive risk model (e.g., [20, 21]), which both model the hazard function and require simultaneous estimation of regression coefficients and baseline hazards, the AFT model directly regresses the logarithm (or a known monotonic transformation) of failure time on covariates, and thus has a simpler structure and an intuitive linear regression interpretation [22]. Due to these advantages, the AFT model is better suited for transferring knowledge from the source to the target cohorts. In this study, we consider the method proposed by Stute [23], which uses weighted least squares loss function to account for censoring. This method has been widely applied to the penalized estimation of AFT models due to its concise loss function, for example, the Lasso estimator [24] and Bridge estimator [25]. Our concern is to adapt this method to introduce transfer learning into the estimation of the AFT model with microarray data.

In this article, based on Stute's weighted least squares loss function, we propose a transfer learning method for the AFT model with the time-to-event outcomes and gene expression covariates. We measure the heterogeneity between cohorts based on their differences in AFT model coefficients. By incorporating the Lasso penalty [26] and tuning parameters, the method simultaneously performs gene selection and controls the extent of information sharing in the coefficient estimation, all within a unified algorithm. The proposed method addresses the challenge of sharing information across different cohorts under the AFT model with microarray data. Our simulation studies demonstrate that the proposed method exhibits robust stability and accuracy, even in the presence of moderate heterogeneity between target and source cohorts. Furthermore, through a Leave-One-Out (LOO) cross validation evaluation procedure, we show that the method achieves satisfactory predictive performance when applied to GSE88770 and GSE25055 datasets, using other GEO datasets as source cohorts. The remainder is organized as follows: Sect. "Methods" introduces the notation, model, and algorithm. Section "Simulation" presents the results of our simulation studies. Section "Data application" contains our real-data analysis, including the motivating example involving ILC cohorts. We conclude with remarks and discussions in Sect. "Conclusion".

Methods

Notation and model

For the ith subject in a random sample of size n, let T_i be the logarithm of the nonnegative time from an initial event to an event of interest. We treat gene expressions as covariates in this article and denote X_i be the p-dimensional covariates. Consider the following accelerated failure time (AFT) model

$$T_i = \gamma + X_i^\top \theta + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where γ is the intercept, $\theta \in \mathbb{R}^p$ is the regression coefficient vector, and ϵ_is are independent and identically distributed random error terms. Ideally, if T_i is fully observed for all $i = 1, \ldots, n$, then one can consider the following least squares function

$$\sum_{i=1}^{n} \left(T_i - \gamma - X_i^\top \theta \right)^2, \tag{2}$$

the γ and θ can then be estimated by minimizing (2). However, in practice, T_i may be subject to right censoring, and we have $\{(Y_i, \delta_i, X_i); i = 1, \ldots, n\}$, where $Y_i = min\{T_i, C_i\}$, C_i is the logarithm of the censoring time, and $\delta_i = I\{T_i \leq C_i\}$ is the occurrence of the interested event (e.g., disease relapse or death) or censoring indicator. Directly using Y_i may lead to biased estimate. In general, estimation of the AFT model have been extensively studied. Notably, the Buckley-James estimator [27, 28] and the rank-based estimator [29] are widely used. Although effective in cases with a small number of covariates, both methods are computationally intensive in high-dimensional settings, particularly when gene selection is involved. A more computationally feasible alternative is the Stute's weighted least squares approach [23], which uses Kaplan-Meier weights to address right censoring in the least squares criterion of the AFT model. Let $\hat{F}_n$ be the Kaplan-Meier estimator [30] of F, the distribution function of T. Let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the ordered statistic of Y_i, $\delta_{(i)}$ be the corresponding censoring indicator, and $X_{(i)}$ be the corresponding covariates, respectively. Then $\hat{F}_n$ can be written as $\hat{F}_n(t) = \sum_{i=1}^{n} w_i I\{Y_{(i)} \leq t\}$, where the Kaplan-Meier weights w_is are the jumps in the Kaplan-Meier estimator that can be computed as

$$w_1 = \frac{\delta_1}{n} \qquad \text{and} \qquad w_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}, \qquad i \in \{2, \ldots, n\}, \tag{3}$$

Based on our notation, the weighted least square loss function is giving as following

$$Q_n(\theta) = \frac{1}{2} \sum_{i=1}^{n} w_i \left(Y_{(i)} - \gamma - X_{(i)}^{\top} \theta \right)^2. \tag{4}$$

We center $Y_{(i)}$ and $X_{(i)}$ with their w_i-weighted means, respectively. That is, let $x_{(i)} = (nw_i)^{1/2}(X_{(i)} - \bar{X}_w)$ and $y_{(i)} = (nw_i)^{1/2}(Y_{(i)} - \bar{Y}_w)$, where $\bar{X}_w = \sum_{i=1}^{n} w_i X_{(i)} / \sum_{i=1}^{n} w_i$ and $\bar{Y}_w = \sum_{i=1}^{n} w_i Y_{(i)} / \sum_{i=1}^{n} w_i$. Using the weighted centered values, the intercept γ is 0. Then we can rewrite the $Q_n(\theta)$ as

$$Q_n(\theta) = \frac{1}{2} \sum_{i=1}^{n} \left(y_{(i)} - x_{(i)}^{\top} \theta \right)^2, \tag{5}$$

which has a concise form. Once the value of the estimator $\hat{\theta}$ is computed, we obtain $\gamma = \bar{Y}_w - \bar{X}_w^{\top} \hat{\theta}$. In relation to the gene selection problem, this article apply a Lasso penalty term to the loss function as follows to obtain a regularized estimate.

$$L_n(\theta) = Q_n(\theta) + \lambda \|\theta\|_1, \tag{6}$$

where $\lambda \geq 0$ is a data-dependent tuning parameter. In an asymptotic sense, λ will generally be of order $\sqrt{\log p/n}$ [31]. In practice, one can set $\lambda = c\sqrt{\log p/n}$, where c is some constant typically ranging between [0, 1] [13]. We will discuss the process of selecting λ in detail in Sect. "Transfer learning algorithm".

Transfer learning algorithm

In this article, we consider the following multi-source transfer learning problem. Suppose we have the target cohort $\{(Y_i^{(0)}, \delta_i^{(0)}, X_i^{(0)}); i = 1, \ldots, n_0\}$ and K source cohorts $\{(Y_i^{(k)}, \delta_i^{(k)}, X_i^{(k)}); i = 1, \ldots, n_k, k = 1, \ldots, K\}$. Assume the outcomes $\{T_i^{(k)}; i = 1, \ldots, n_k, k = 0, 1, \ldots, K\}$ in the target and the source cohorts all follow the AFT models

$$T_i^{(k)} = \gamma^{(k)} + (X_i^{(k)})^{\top} \theta^{(k)} + \epsilon_i^{(k)}, \qquad i = 1, \ldots, n_k, \qquad k = 0, 1, \ldots, K, \tag{7}$$

where $\theta^{(k)} \in \mathbb{R}^p$ is the coefficient vector of the kth model, $\gamma^{(k)}$ is the corresponding intercept and $\epsilon_i^{(k)}$ are random error terms. For $k = 0, 1, \ldots, K$, $\theta^{(k)}$ are possibly different and the same set of covariates are available in every cohort. We denote the target coefficient $\beta = \theta^{(0)}$. Suppose the target model is sparse, which satisfies $\|\beta\|_0 \ll p$. This means that only s of the p covariates are associated with the target outcomes. For each cohort, We can center it respectively based on its Kaplan-Meier weights $w_i^{(k)}$ as aforementioned in (3), and then have the target cohort $\{(y_{(i)}^{(0)}, x_{(i)}^{(0)}); i = 1, \ldots, n_0\}$ and the source cohorts $\{(y_{(i)}^{(k)}, x_{(i)}^{(k)}); i = 1, \ldots, n_k, k = 1, \ldots, K\}$. In transfer learning, our aim is to leverage information from source cohorts to improve the estimation based on the target cohort. One intuitive approach is to combine all centered samples from both the target and the source cohorts and then apply the loss function (6) to obtain an estimator. However, this approach ignores the potential heterogeneity between the target cohort and the source cohorts, which can lead to biased estimates or risk assessments for the target cohort. Even if the heterogeneity is small, the estimation bias by combining all the cohorts can not be neglected as the number of the source cohorts and their sample size increase. To reduce bias while borrowing information from the source cohorts, which may or may not be different from the target cohort, we propose a two-stage transfer learning algorithm for the AFT model to improve the efficiency and accuracy of information borrowing and estimation, the algorithm is motivated by the ideas in [13] and [15], which we call a Trans-AFT algorithm.

In the first stage, we fit an AFT model by pooling all the centered samples $\{(y_{(i)}^{(k)}, x_{(i)}^{(k)}); i = 1, \ldots, n_k, k = 0, 1, \ldots, K\}$, the weighted least square loss function with the Lasso penalty is

$$O^s(\theta) = \frac{1}{2n_s} \sum_{k=0}^{K} \sum_{i=1}^{n_k} \left(y_{(i)}^{(k)} - (x_{(i)}^{(k)})^\top \theta \right)^2 + \lambda_\theta \|\theta\|_1, \tag{8}$$

where $n_s = \sum_{k=0}^{K} n_k$, $\lambda_\theta = c_1 \sqrt{\log p / n_s}$ with some constant c_1. In practice, we can select the value of λ_θ through V-fold cross validation. That is, we first construct a sequence of equally spaced values for c_1 from the interval [0, 1], calculate the corresponding $\lambda_\theta = c_1 \sqrt{\log p / n_s}$ for each value in the sequence, then select the optimal λ_θ that minimizes the cross validation loss based on the loss function (8) through R package glmnet [32]. A rough estimator $\hat{\theta}^s$ can be estimated by minimize the loss function (8). In the second stage, we correct the bias using the target cohort only. In this work, we characterize the bias using the sparsity of the difference between $\theta^{(k)}$ and β. More specifically, assume that the heterogeneity between the target cohort and the source cohorts lies in the shift of their coefficients in AFT models, that is

$$\eta^{(k)} = \beta - \theta^{(k)}. \tag{9}$$

The parameter $\eta^{(k)}$ is used to quantify the difference between the target coefficient and the kth source coefficient. Intuitively, if the kth source cohort is "close enough" to the target cohort, the parameter $\eta^{(k)}$ will degenerate to zero, otherwise there exists at least one none-zero component in $\eta^{(k)}$, in such case, combining all the cohorts roughly would result in a biased estimate. Based on the assumption about $\eta^{(k)}$, we consider estimating its non-zero subset to correct the bias and control the utilization of information from the source cohorts adaptively. With the formulation (9), the loss function to be minimized is

$$O(\eta) = \frac{1}{2n_0}\sum_{i=1}^{n_0}\left(y_{(i)}^{(0)} - (x_{(i)}^{(0)})^{\top}(\hat{\theta}^s + \eta)\right)^2 + \lambda_\eta \|\eta\|_1, \tag{10}$$

where $\lambda_\eta = c_2\sqrt{\log p/n_s}$ with some constant c_2, which can also be obtained using the similar method for selecting λ_θ. Note that we focus on the overall difference in coefficients between the target and the source cohorts rather than the differences in coefficients between individual cohorts. The reason is that the objective of transfer learning is solely to estimate the target coefficient. Theoretically, additional penalty terms and the joint analysis of multiple estimators may not enhance the estimation of the coefficient of interest [13]. Our proposed Trans-AFT algorithm is formally presented in Algorithm 1. The definitions of the symbols in the algorithm follow those in Sect. "Transfer learning algorithm".

Algorithm 1 Process of Trans-AFT Algorithm

```
Input:
    Target cohort $\{(Y_i^{(0)}, \delta_i^{(0)}, X_i^{(0)}); i = 1, \ldots, n_0\}$;
    Source cohorts $\{(Y_i^{(k)}, \delta_i^{(k)}, X_i^{(k)}); i = 1, \ldots, n_k, k = 1, \ldots, K\}$;
    A sequence of equally spaced values for $c_1$ and $c_2$ from [0,1], respectively.
Output:
    The estimated coefficient vector $\hat{\beta}$ .
Center every cohort respectively based on its Kaplan-Meier weights defined in equation (3);
For the sequence of $c_1$, calculate the corresponding $\lambda_\theta = c_1\sqrt{\log p/n_s}$;
Based on V-fold cross validation and the loss function (8), select the value of $\lambda_\theta$ that minimizes the cross validation loss.
Transferring step: Estimate preliminar estimator $\hat{\theta}^s$ by minimizing the loss function (8);
For each value in the sequence of $c_2$, calculate the corresponding $\lambda_\eta = c_2\sqrt{\log p/n_0}$;
Based on V-fold cross validation and the loss function (10), select the value of $\lambda_\eta$ that minimizes the cross validation loss.
Debiasing step: Estimate $\hat{\eta}$ by minimizing the loss function (10);
The estimated coefficient is $\hat{\beta} = \hat{\theta}^s + \hat{\eta}$.
```

Evaluation

In practice, since the true effects of the covariates on the outcome are unknown, we need to evaluate aspects such as the predictive performance of the proposed model and other comparable methods. Unfortunately, most of the conventional evaluation techniques are effective only when $p < n$, they are not suitable for gene expression data where $p \gg n$. In this study, we focus on two key aspects: (1) predictive performance, meaning the model and the selected genes should make accurate predictions for external cases; (2) relative stability of selected genes, meaning gene selection should be reproducible on similar data. Motivated by the ideas in [21], which evaluated the performance of additive risk models based on leave-one-out cross validation, we propose an evaluation procedure to assess the relative stability of gene selection and compare the predictive performance of proposed method and other comparable methods. For $i = 1, \ldots, n_0$, compute the proposed transfer learning estimator $\hat{\beta}^{(-i)}$ with the reduced dataset by removing the ith subject, then compute the risk score $(X_i^{(0)})^\top \hat{\beta}^{(-i)}$ corresponding to the ith subject that was not used in the estimation. In the process of obtaining these n_0 estimators, for the jth component of β, denote c_j as the times it is selected among the total n_0 estimations, then compute the corresponding occurrence index $OI_j = c_j/n_0$, which ranges from 0 to 1. Intuitively, if a gene is relatively important, it should be selected in most of the reduced datasets and its occurrence index should approach 1. To evaluate and compare the overall predictive power of different models, we first dichotomize the n_0 predictive risk scores $\{(X_i^{(0)})^\top \hat{\beta}^{(-i)}; i = 1, \ldots, n_0\}$ at their median, thereby creating two hypothetical risk groups. Because the AFT model directly performs regression on survival outcomes, higher risk scores leading to longer survival times, and so lower survival risk. Then compare the K-M survival curve of the two risk groups using a log-rank test. A significant difference between the two groups indicates that the model provides satisfactory predictions for external cases. Additionally, calculate the C-index [33, 34] based on the risk scores to compare model performance, that is,

$$C = \frac{\sum_{i=1}^{n_0} \delta_i^{(0)} (\#\{j : s_i > s_j\} + \#\{j : s_i = s_j\}/2)}{\sum_{i=1}^{n0} \delta_i^{(0)} \#\{j : Y_i^{(0)} > Y_j^{(0)}\}}, \tag{11}$$

where $s_i = (X_i^{(0)})^\top \hat{\beta}^{(-i)}$, $\#\{\cdot\}$ represents the counting of elements in a set. The C-index can be interpreted as the fraction of all pairs of subjects whose predicted risk scores are correctly ordered among all subjects that can actually be ordered. A higher C-index indicates a better prediction performance of the model. C-index equals 1 means perfect prediction accuracy and C-index equals 0.5 is as good as a random predictor.

Note that although Leave-One-Out cross validation may be computationally slower compared to 10-fold or 5-fold cross validation, considering that the target cohort may have a small sample size and a high censoring proportion, the Leave-One-Out cross validation can more effectively utilize the limited information in the data. Moreover, evaluation based on external data is also feasible (e.g., [35, 36]), but in practice, due to patient data privacy concerns and differences among various studies, finding external data that is similar or consistent with the research objective can be quite challenging.

Simulation

We evaluate the empirical performance of the proposed Trans-AFT algorithm and compare it with other methods through a series of simulation studies. Specifically, we evaluate three methods, including AFT models using the target cohort only (Lasso-AFT), the simple combination of all the cohorts (Pooled-AFT), and our proposed Trans-AFT algorithm (Trans-AFT). The R code for implementing all the methods in simulation are available at https://github.com/YuZhengyang-CUEB/Trans-AFT.

Homogeneous designs

We consider $p = 500$, $n_0 = 150$, and $n_1, ..., n_K = 200$. We set $K \in \{8, 12, 16, 20, 24\}$ to observe whether an increase in the source cohorts will improve the estimation performance of transfer learning. The covariates $X_i^{(k)}$ are *i.i.d.* Gaussian with mean zero and identity covariance matrix for all $0 \le k \le K$ and $\epsilon_i^{(k)}$ are *i.i.d.* Gaussian with mean zero and variance one for all $0 \le k \le K$. For the target coefficient β, we set $s = \|\beta\|_0 = 20$, the number of its non-zero elements. We set $\beta_j = 0.3$ for $j \in 1, \ldots, s$, and $\beta_j = 0$ otherwise. For coefficients of the source cohorts, similar to the work in [13], we consider two configurations to simulate different patterns and extents of shift in the source coefficients:

(1) For $1 \le k \le K$, let

$$\theta_j^{(k)} = \beta_j - 0.3I(j \in H_k), \tag{12}$$

where H_k is a random subset of $1, \ldots, p$ with $|H_k| = h \in \{2, 6, 12\}$.

(2) For $1 \le k \le K$, let $H_k = \{1, \ldots, 100\}$ and

$$\theta_j^{(k)} = \beta_j + \xi_j I(j \in H_k), \text{ where } \xi_j \sim_{i.i.d} N(0, h/100), \tag{13}$$

where $h \in \{2, 6, 12\}$ and $N(a, b)$ is Gaussian with mean a and variance b. Configurations 1 and 2 can be seen as scenarios where the source coefficients exist a fixed shift and a random shift, respectively. The value of h characterizes the extent of the shift. For the outcome, we set the proportion of censoring as 20%, 50% and 70%, respectively. We compute the sum of absolute estimation errors (SAE) for each estimator b, $\|b - \beta\|_1$, in Fig. 1, each point is summarized from 200 independent simulations.

Through Fig. 1, as expected, the performance of Lasso-AFT does not change as K increases, and in most cases, it has the largest estimation error. The other two methods that utilize source cohorts' information have estimation errors decreasing as K increases. As h increases, the difference between the coefficients of the target cohort and those of the source cohorts widens, leading to larger estimation errors. As the proportion of censoring increases, the problem also becomes harder and the estimation errors of all three methods increase. Meanwhile, Pooled-AFT method always has larger estimation error than Trans-AFT method, even when h is small. This confirms the importance of the debiasing step in the transfer learning Algorithm 1. This indicates that when there are varying degrees of difference between the coefficients of the target and source cohorts, the proposed transfer learning method is more adept at accommodating these differences and reducing the errors resulting from information borrowing.

Heterogeneous designs

In this section, taking into account the potentially heterogeneity of covariates between the target and the source cohorts, we now consider a heterogeneous setting where $\Sigma^{(k)}$, the covariance matrix of $X_i^{(k)}$, are distinct for $k = 0, 1, \ldots, K$. Specifically, let $X_i^{(k)}, k = 0, 1, \ldots, K$ are *i.i.d.* Gaussian with mean zero and a covariance matrix $\Sigma^{(k)}$, we set $\Sigma^{(0)} = I_p$, and set $\Sigma^{(k)}, k = 1, \ldots, K$ as a Toeplitz covariance matrix whose first row is

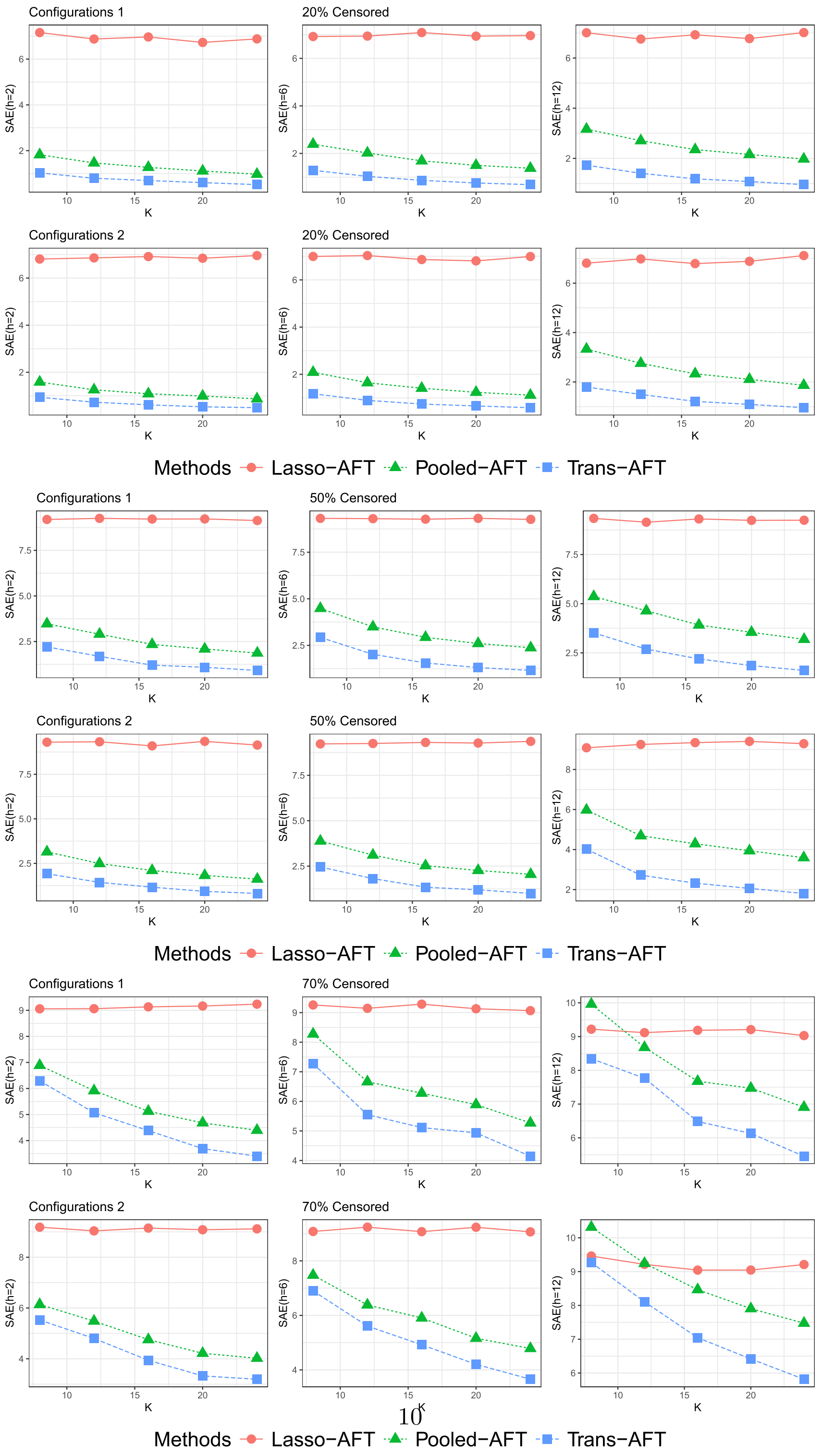

Fig. 1 Sum of absolute estimation errors of the Lasso-AFT, Pooled-AFT, Trans-AFT method with homogeneous designs under two configurations. The proportion of censoring is set to 20%, 50%, 70%. The *y* axis corresponds to $\|b - \beta\|_1$ for some estimator *b*

$$\Sigma_{1,\cdot}^{(k)} = (1, \underbrace{1/(k+1), \ldots, 1/(k+1)}_{2k-1}, 0_{p-2k}). \tag{14}$$

Other settings are set to be the same as in Sect. "Homogeneous designs".

Figure 2 shows that, in heterogeneous designs, the general patterns observed under homogeneous designs still hold. Under heterogeneous designs, the transfer learning method for AFT models continues to exhibit the best estimation performance compared to other methods. As the proportion of censoring increases, the advantage of the proposed transfer learning method becomes more pronounced. Even when there is moderate degree of heterogeneity in the covariates between the target cohort and the source cohorts, the proposed transfer learning method is still capable of making more precise estimates of the target coefficient.

Data application

GSE88770 data

Invasive lobular carcinoma (ILC) is a relatively rare and special subtype of breast carcinoma. ILC displays a poor response to neoadjuvant therapy, a different metastatic pattern compared to invasive breast carcinoma of no special type, as well as unique molecular characteristics [37]. Compared to the invasive ductal carcinoma (IDC), the incidence rate of ILC is increasing steadily [38]. A previous microarray prognostic study on ILC has demonstrated the prognostic value of gene expression and genomic grade [6]. The gene expression data that based on GPL-570 platform ([HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array) is available at GSE88770 from Gene Expression Omnibus (GEO). In this article, our primary goal is to identify genes that are associated with overall survival (OS) and to predict the survival risk for patients with ILC. The GSE88770 data contained OS times for 28 patients, while the remaining 89 patients' outcomes were censored. Due to the small sample size and high proportion of censoring, the gene selection and risk prediction based solely on the GSE88770 cohort yield unsatisfactory results. Therefore, we expect to transfer the information from patients with other types of breast cancer. We have a brief preview of the data analysis workflow in Fig. 3.

In the selection of the source cohorts, to avoid excessive heterogeneity, we selected breast cancer patient cohorts with survival information from the same GPL570 platform. For the selected cohorts, we excluded samples from normal breast tissue and those with missing values. Finally, 8 cohorts were selected as the source cohorts, which were GSE58812, GSE48390, GSE42568, GSE31448, GSE21653, GSE20711, GSE20685, and GSE16446, see Table 1 for details. All datasets are publicly available at https://www.ncbi.nlm.nih.gov/geo/. We preprocessed the probe data to match the probes with corresponding genes. In cases where multiple probes correspond to one gene, we took the maximum value among those probes. After preprocessing, the covariates comprised 23,348 genes. Although the proposed transfer learning algorithm does not impose limits on the number of covariates, we followed methods from [21, 25], which screen the covariates to reduce data noise, increase stability, and improve efficiency. Specifically, at first, we applied unsupervised screening (i.e. the outcome is not used in the screening) by removing genes with interquartile ranges smaller than their first quartile, ultimately

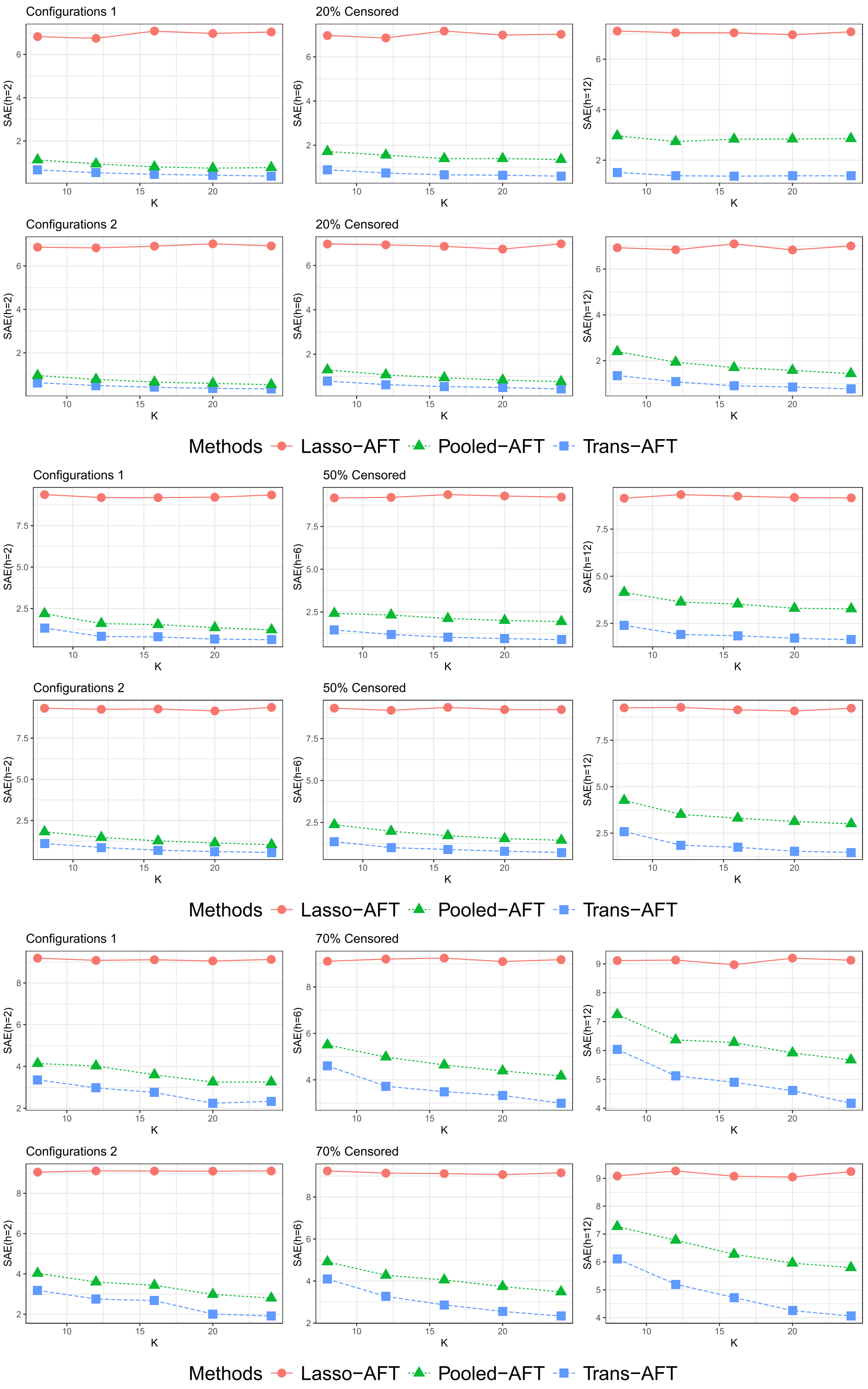

Fig. 2 Sum of absolute estimation errors of the Lasso-AFT, Pooled-AFT, Trans-AFT method with heterogeneous designs under two configurations. The proportion of censoring is set to 20%, 50%, 70%. The *y* axis corresponds to $\|b - \beta\|_1$ for some estimator *b*

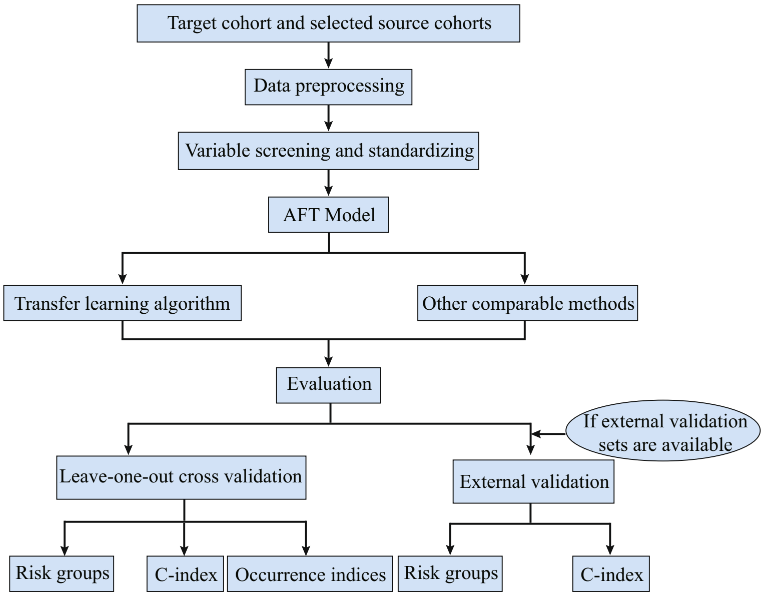
Target cohort and selected source cohorts
Data preprocessing
Variable screening and standardizing
AFT Model
Transfer learning algorithm
Other comparable methods
Evaluation
If external validation sets are available
Leave-one-out cross validation
External validation
Risk groups
C-index
Occurrence indices
Risk groups
C-index

Fig. 3 Flow diagram of date processing, analysis, and evaluation

Table 1 Basic information of cohorts used in Sect. "GSE88770 data"

Type	Cohort	Number of cases	Number of uncensored	Number of censored
Target cohort	GSE88770	117	28	89
Source cohort	GSE58812	107	29	78
Source cohort	GSE48390	81	11	70
Source cohort	GSE42568	104	48	56
Source cohort	GSE31448	246	79	167
Source cohort	GSE21653	248	79	169
Source cohort	GSE20711	86	37	49
Source cohort	GSE20685	327	83	244
Source cohort	GSE16446	107	24	83

retaining 6,608 genes. Next, we conducted supervised screening (i.e. the outcome is used in the screening) by calculating the correlation coefficients between uncensored outcomes and the remaining genes, retaining the 500 genes with the largest absolute correlation coefficients. The purpose of unsupervised screening is to remove some obviously redundant genes, while supervised screening retains an appropriate number of genes for modeling based on their correlation with the outcome. Finally, these 500 genes were standardized to have zero mean and unit variance.

We applied the proposed transfer learning method to the processed data, resulting in the selection of 79 genes. Model evaluation and comparison were performed using the Leave-One-Out (LOO) procedure described in Sect. "Evaluation". Since unsupervised screening does not consider survival outcomes, for each reduced dataset, we performed supervised screening on the genes that passed unsupervised screening, selecting a potentially different set of 500 genes each time. The OIs of individual genes that pass

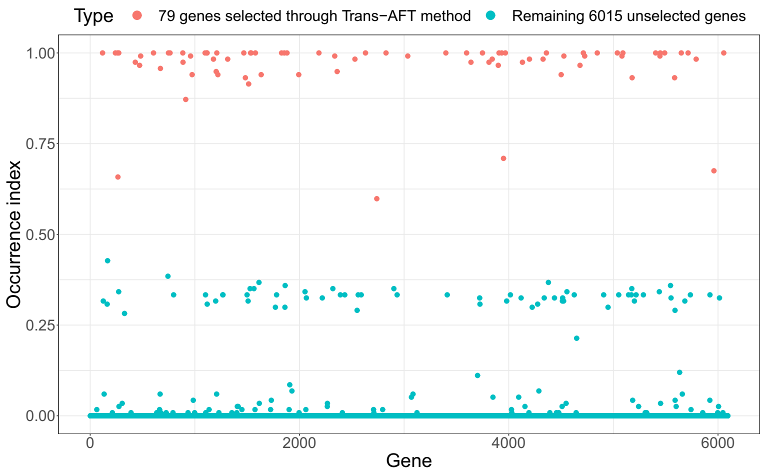

Fig. 4 GSE88770 data: occurrence index of individual genes selected by proposed transfer learning method

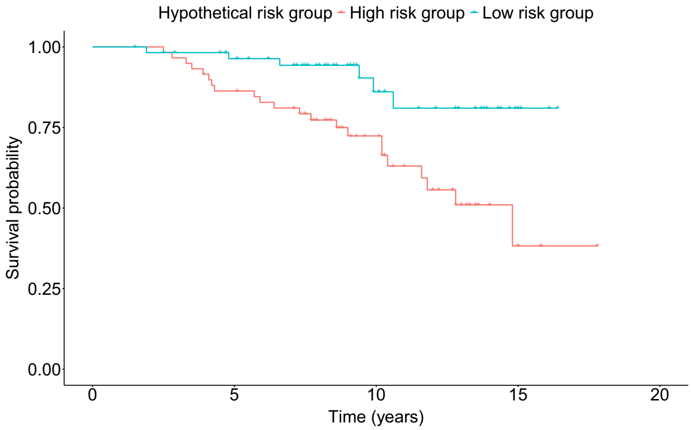

Fig. 5 GSE88770 data: K-M survival curves of two hypothetical risk groups identified by proposed transfer learning method

the unsupervised screening are shown in Fig. 4. In Fig. 4, blue dots represent the 79 genes selected using the proposed method, while red dots represent the rest 6015 genes. As observed, the genes selected via the proposed method exhibit higher OIs compared to the rest of the genes, indicating their greater importance. Moreover, the majority of selected genes have OIs close to 1, suggesting that the proposed method is relatively stable in identifying important genes.

To evaluate and compare the overall predictive power of different models, we generated two risk groups based on the predictive risk scores obtained through the LOO procedure. In Fig. 5, we show the K-M survival curves of the two risk groups generated by the proposed transfer learning method. It is evident that the two survival functions differ significantly, the high-risk group generally has a shorter survival time and a faster decline in survival probability compared to the low-risk group. In Table 2, we present the results of the log-rank test and the C-index for the three methods discussed in Sect. "Simulation". The proposed method shows a significant difference between the two

Table 2 GSE88770 data: Evaluation of the predictive performance of three methods using log-rank test and C-index

Method	Chi-square statistic	*p*-value	C-index
Trans-AFT	8.712	0.0032	0.73
Lasso-AFT	0.069	0.7934	0.50
Pooled-AFT	0.160	0.6888	0.58

Table 3 Basic information of cohorts used in Sect. "GSE25055 data"

Type	Cohort	Number of cases	Number of uncensored	Number of censored
Target cohort	GSE25055	309	65	244
Validation cohort	GSE25065	198	45	153
Source cohort	GSE158309	445	116	329
Source cohort	GSE124647	140	130	10
Source cohort	GSE45255	135	31	104
Source cohort	GSE17705	298	71	227
Source cohort	GSE12093	136	20	116
Source cohort	GSE7390	198	91	107
Source cohort	GSE4922	242	83	159

risk groups and achieves the highest C-index among the compared methods. Therefore, we can conclude that the proposed transfer learning method can satisfactorily predict patients' survival risk based on the selected genes.

GSE25055 data

The GSE25055 data consists of 310 HER2-negative breast cancer patients following neoadjuvant taxane-anthracycline chemotherapy [39, 40], at the same time, it has a validation cohort GSE25065, which includes 198 HER2-negative breast cancer patients who received the same treatment [39, 40]. Our aim is to develop a predictive model for response and survival outcomes following this treatment in patients with HER2-negative invasive breast cancer. After excluding one case with missing values, distant relapse-free survival (DRFS) times of 65 patients were available and the other 244 patients were censored. The gene expression data, based on GPL-96 platform ([HG-U133A] Affymetrix Human Genome U133A Array), is available at GSE25055 from GEO. Similar to Sect. "GSE88770 data", according to the workflow shown in Fig. 3. We selected 7 groups of breast cancer samples based on GPL-96 platform from GEO as source cohorts, which include GSE158309, GSE124647, GSE45255, GSE17705, GSE12093, GSE7390, GSE4922, see Table 3 for details. After data preprocessing, 2439 out of 13,435 genes were selected through unsupervised screening, and then we selected 500 genes through the supervised screening. Gene expressions were then standardized to have zero mean and unit variance.

Using the proposed transfer learning method, 56 genes were identified. Relative stability of selected genes was evaluated using the proposed OI. In Fig. 6, we can see that 56 genes identified by proposed method have higher OIs than the rest 2383 genes, and the

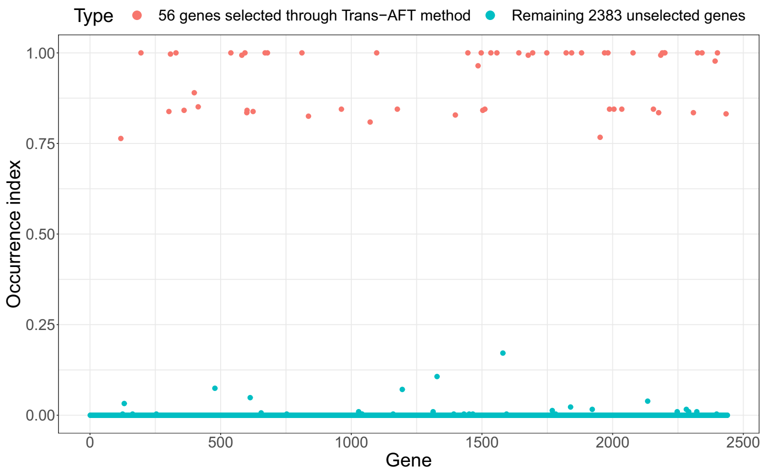

Fig. 6 GSE25055 data: occurrence index of individual genes selected by proposed transfer learning method

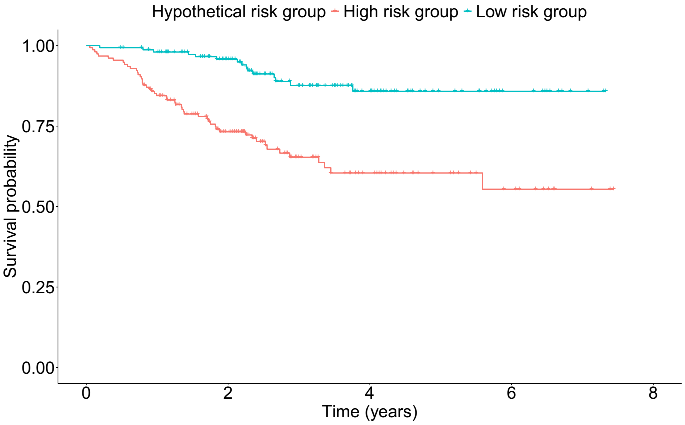

Fig. 7 GSE25055 data: K-M survival curves of two hypothetical risk groups identified by proposed transfer learning methods

majority of selected genes have OIs close to 1, which suggest that the selected genes are relatively stable.

The overall predictive power of different methods was evaluated in the same procedure as described in Sect. "GSE88770 data". The two risk groups identified by the proposed transfer learning method are shown through the K-M survival curves in Fig. 7. There is a significant difference between the two groups, with the high-risk group having an average lower survival time and a faster declining survival probability. Table 4 presents the results of the log-rank test and the C-index, confirming that the transfer learning method can accurately distinguish high-risk and low-risk groups and has the highest C-index among the evaluated methods. At the same time, we also conducted predictions on the external validation cohort GSE25065. The two hypothetical risk groups in Fig. 8 created by the risk scores corresponding to the transfer learning method, also display strong differences. Table 5 shows the results of the log-rank test and the C-index, further confirming that the transfer learning method

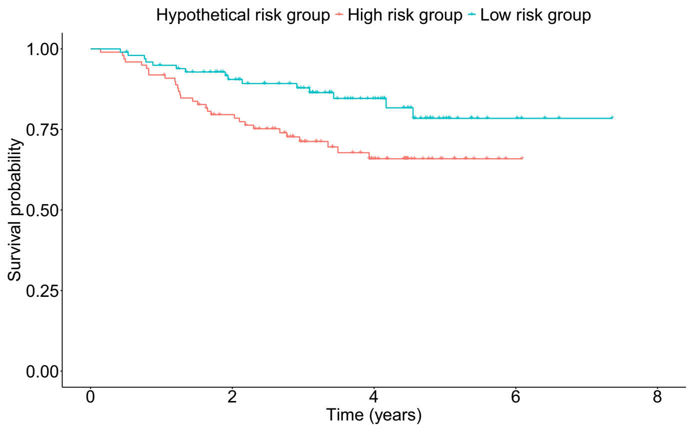

Fig. 8 GSE25065 data: K-M survival curves of two hypothetical risk groups identified by proposed transfer learning methods

Table 4 GSE25055 data: Evaluation of the prediction performance of three methods using log-rank test and C-index

Method	Chi-square statistic	*p*-value	C-index
Trans-AFT	26.376	<1e-04	0.72
Lasso-AFT	0.171	0.6789	0.56
Pooled-AFT	12.308	0.0005	0.64

Table 5 GSE25065 data: Evaluation of the prediction performance of three methods using log-rank test and C-index

Method	Chi-square statistic	*p*-value	C-index
Trans-AFT	6.160	0.0131	0.66
Lasso-AFT	3.080	0.0793	0.57
Pooled-AFT	2.568	0.1090	0.58

has stronger and more significant predictive abilities compared to other methods. Note that although the Pool-AFT method has made effective risk predictions on the GSE25055 cohort, its performance on the external validation set is worse than that of the transfer learning method and it has a lower C-index. We thus conclude that the proposed transfer learning method can provide a more accurate risk assessment based on a small subset of selected genes.

Remark

Analysis of the GSE88770 and GSE25055 data suggests that the AFT model with proposed transfer learning method is capable of identifying a small number of genes and risk assessment. We note that, despite the fact that the target cohorts in these two examples have a high proportion of censoring, their predictions are expected to be valid based on LOO procedure or external data validation.

Conclusion

In microarray prognostic studies, when the sample size of the target cohort is limited, developing a method which can leverage information from the source cohorts to enhance the analysis of the target cohort has important practical implications. In this article, we assume the accelerated failure time (AFT) model for analyzing the time-to-event outcomes and gene expressions. AFT models offer a useful alternative to the Cox and additive hazard models due to their simpler structure and more intuitive interpretation of coefficients. A transfer learning method is proposed for coefficient estimation and gene selection. Our simulation studies demonstrated that the transfer learning method has better performance in terms of estimation error. Gene selection and overall predictive performance were evaluated using the leave-one-out (LOO) procedure.The analysis of GSE88770 and GSE25055 datasets with the proposed method showed that it successfully identifies a small subset of genes with strong predictive power.

The proposed method still faces certain challenges. Firstly, under the current framework, we select the source cohorts for transfer learning based on experience. If the heterogeneity between target and source cohorts is too great, transfer learning may have a negative impact on the target task, which is called negative transfer [9, 41]. It is difficult to assess transferability between target and source cohorts, and define criteria to measure cohort similarity for transferability assessment. [13] introduced an algorithm for transfer learning under high-dimensional linear regression, which aggregates a number of candidate estimators [42] to reduce the impact of unsuitable source cohorts on the estimation. [15] developed an algorithm based on cross validation, which rejects the use of a source cohort when it contributes excessively to the cross-validation error. However, developing a similar method for the AFT model is not trivial, we postpone pursuing this to a separate study.

Secondly, in high-dimensional survival analysis, despite Lasso's impressive performance in practice, it has been shown that the Lasso is in general not variable selection consistent [43]. There are many better penalization methods that have consistent selection, including the Adaptive Lasso, the SCAD and the Bridge. We acknowledge that the optimization of the model's theoretical performance goes beyond the scope of this article and is worthy of future research.

Thirdly, in our framework, we assume the covariates of interest are available for every cohort. In medical institutions or clinical trials, this assumption faces many limitations. Especially in microarray analysis, the probe variables obtained from different platforms are sometimes different. At the same time, we notice that although the proposed method demonstrated good performance in Sect. "Data application" with smaller sample sizes and higher censoring rates, the effectiveness of the proposed algorithm may be limited if the sample size is further reduced or the censoring rate further increases. This could restrict its application in some diseases with high cure rates.Therefore, another interesting direction is extending transfer learning to other specific semiparametric survival analysis models, such as the partial linear regression and Cure Rate Model. Finally, in terms of model evaluation, methods such as hypothesis testing for parameter estimation in transfer learning also remain to be explored.

Acknowledgements

The authors appreciate the editorial team for the careful review and useful comments that significantly improve the initial manuscript.

Author contributions

Y.P. conceived the project, developed the method, wrote the paper, and revised the paper. Z.Y. developed the method, wrote the R code, analyzed the results, wrote the paper, and revised the paper. J.S. developed the method, wrote the paper and revised the paper. All authors have reviewed and approved this manuscript.

Funding

Capital University of Economics and Business special fund for basic scientific research and business expenses of Beijing affiliated universities (QNTD202207).

Availability of data and materials

The data used in this article are all sourced from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and can be accessed using the provided identifiers. The identifiers for all the GEO data used in this article are as follows: GSE88770, GSE58812, GSE48390, GSE42568, GSE31448, GSE21653, GSE20711, GSE20685, GSE16446, GSE25055, GSE25065, GSE158309, GSE124647, GSE45255, GSE17705, GSE12093, GSE7390, GSE4922. The R code for all the methods in simulation are available at https://github.com/YuZhengyang-CUEB/Trans-AFT.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no Conflict of interest.

Received: 5 September 2024 Accepted: 17 January 2025
Published online: 17 March 2025

References

1. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):503–11. https://doi.org/10.1038/35000501.
2. Rosenwald A, Wright G, Wiestner A, et al. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. Cancer Cell. 2003;3(2):185–97. https://doi.org/10.1016/s1535-6108(03)00028-x.
3. Cristofanilli M, Angulo AG, Sneige N, et al. Invasive lobular carcinoma classic type: response to primary chemotherapy and survival outcomes. J Clin Oncol. 2005;23(1):185–97. https://doi.org/10.1200/JCO.2005.03.111.
4. Arpino G, Bardou VJ, Clark GM, et al. Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome. Breast Cancer Res. 2004;6(3):149–56. https://doi.org/10.1186/bcr767.
5. Lamovec J, Bracko M. Metastatic pattern of infiltrating lobular carcinoma of the breast: an autopsy study. J Surg Oncol. 1991;48(1):28–33. https://doi.org/10.1002/jso.2930480106.
6. Filho OM, Michiels S, Bertucci F, et al. Genomic grade adds prognostic value in invasive lobular carcinoma. Ann Oncol. 2013;24(2):377–84. https://doi.org/10.1093/annonc/mds280.
7. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets-update. Nucl Acids Res. 2013;41:991–5. https://doi.org/10.1093/nar/gks1193.
8. The Cancer Genome Atlas Research Network, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70. https://doi.org/10.1038/nature11412.
9. Torrey L, Shavlik J, et al. Transfer learning. In: Olivas ES, Guerrero JDM, Sober MM, et al., editors. Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, vol. 42. Hershey: IGI Global; 2010. p. 242–64.
10. Hajiramezanali E, Zamani S. Bayesian Multi-Domain Learning for Cancer Subtype Discovery from Next-Generation Sequencing. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2018. p. 9133–9142.
11. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. 2016;35(5):1285–98. https://doi.org/10.1109/TMI.2016.2528162.
12. Turki T, Wei Z, Wang JT. Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. IEEE Access. 2017;5:7381–93. https://doi.org/10.1109/ACCESS.2017.2696523.
13. Li S, Cai TT, Li HZ. Transfer learning for high-dimensional linear regression: prediction, estimation, and minimax optimality. J R Stat Soc Ser B Stat Methodol. 2022;84(1):149–73. https://doi.org/10.1111/rssb.12479.
14. Tian PX, Chan TH, Wang YF, et al. Multiethnic polygenic risk prediction in diverse populations through transfer learning. Front Genet. 2022;13(906965):1–11. https://doi.org/10.3389/fgene.2022.906965.
15. Tian Y, Feng Y. Transfer learning under high-dimensional generalized linear models. J Am Stat Assoc. 2023;118(544):2684–97. https://doi.org/10.1080/01621459.2022.2071278.
16. Li S, Zhang LJ, Cai TT, Li HZ. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. J Am Stat Assoc. 2024;119(546):1274–85. https://doi.org/10.1080/01621459.2023.2184373.

17. Li ZY, Shen Y, Ning J. Accommodating time-varying heterogeneity in risk estimation under the Cox model: a transfer learning approach. J Am Stat Assoc. 2023;118(544):2276–87. https://doi.org/10.1080/01621459.2023.2210336.
18. Cox DR. Regression models and life-tables. J R Stat Soc Ser B Stat Methodol. 1972;34(2):187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.
19. Gui J, Li HZ. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics. 2005;21(13):3001–8. https://doi.org/10.1093/bioinformatics/bti422.
20. Lin DY, Ying Z. Semiparametric analysis of the additive risk model. Biometrika. 1994;81(1):61–71. https://doi.org/10.1093/biomet/81.1.61.
21. Ma S, Shen Y, Huang J. Additive risk survival model with microarray data. BMC Bioinform. 2007;8(192):1–10. https://doi.org/10.1186/1471-2105-8-192.
22. Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat Med. 1992;11(14–15):1871–9. https://doi.org/10.1002/sim.4780111409.
23. Stute W. Consistent estimation under random censorship when covariables are available. J Multivar Anal. 1993;45(1):89–103. https://doi.org/10.1006/jmva.1993.1028.
24. Huang J, Ma S, Xie HL. Regularized estimation in the accelerated failure time model with high-dimensional covariates. Biometrics. 2006;62(3):813–20. https://doi.org/10.1111/j.1541-0420.2006.00562.x.
25. Huang J, Ma S. Variable selection in the accelerated failure time model via the bridge method. Lifetime Data Anal. 2010;16:176–95. https://doi.org/10.1007/s10985-009-9144-2.
26. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Stat Methodol. 1996;58(1):267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
27. Buckley J, James I. Linear regression with censored data. Biometrika. 1979;66(3):429–36. https://doi.org/10.1093/biomet/66.3.429.
28. Lai TL, Ying Z. Large sample theory of a modified Buckley-James Estimator for regression analysis with censored data. Ann Stat. 1991;19(3):1370–402. https://doi.org/10.1214/aos/1176348253.
29. Ying Z. A large sample study of rank estimation for censored regression data. Ann Stat. 1993;21(1):76–99. https://doi.org/10.1214/aos/1176349016.
30. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–81. https://doi.org/10.1080/01621459.1958.10501452.
31. Van de Geer S. The Lasso. In: Estimation and testing under sparsity: École d'Été de Probabilités de Saint-Flour XLV - 2015. Heidelberg: Springer; 2016. p. 5–25.
32. Friedman J, Tibshirani R, Hastie T. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22. https://doi.org/10.18637/JSS.V033.I01.
33. Raykar VC, Steck H, Krishnapuram B, et al. On ranking in survival analysis: bounds on the concordance index. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2007. p. 1209–1216.
34. Li R, Chang C, Justesen JM, et al. Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. Biostatistics. 2022;23(2):522–40. https://doi.org/10.1093/biostatistics/kxaa038.
35. Tian Z, Tang J, Liao X, et al. An immune-related prognostic signature for predicting breast cancer recurrence. Cancer Med. 2020;9(20):7672–85. https://doi.org/10.1002/cam4.3408.
36. Tian Z, Tang J, Liao X, et al. Identification of a 9-gene prognostic signature for breast cancer. Cancer Med. 2020;9(24):9471–84. https://doi.org/10.1002/cam4.3523.
37. Koufopoulos K, Pateras IS, Gouloumis AR, et al. Diagnostically challenging subtypes of invasive lobular carcinomas: how to avoid potential diagnostic pitfalls. Diagnostics. 2022;12(11):2658. https://doi.org/10.3390/diagnostics12112658.
38. Li CI, Anderson BO, Daling JR, et al. Trends in incidence rates of invasive lobular and ductal breast carcinoma. J Am Med Assoc. 2003;289(11):1421–4. https://doi.org/10.1001/jama.289.11.1421.
39. Hatzis C, Pusztai L, Valero V, et al. A genomic predictor of response and survival following Taxane-anthracycline chemotherapy for invasive breast cancer. J Am Med Assoc. 2011;305(18):1873–81. https://doi.org/10.1001/jama.2011.593.
40. Baldasici O, Balacescu L, Cruceriu D, et al. Circulating small EVs miRNAs as predictors of pathological response to neo-adjuvant therapy in breast cancer patients. Int J Mol Sci. 2022;23(20):12625. https://doi.org/10.3390/ijms232012625.
41. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2009;22(10):1345–59. https://doi.org/10.1109/TKDE.2009.191.
42. Dai D, Rigollet P, Zhang T. Deviation optimal learning using greedy Q-aggregation. Ann Stat. 2012;40(3):1878–905. https://doi.org/10.1214/12-AOS1025.
43. Leng C, Lin Y, Wahba G. A note on the LASSO and related procedures in model selection. Stat Sin. 2006;16(4):1273–84.

Publisher's Note