mathematics

MDPI

Article

High-Dimensional Mediation Analysis for Time-to-Event Outcomes with Additive Hazards Model

Meng An and Haixiang Zhang *

Center for Applied Mathematics, Tianjin University, Tianjin 300072, China
* Correspondence: haixiang.zhang@tju.edu.cn

Abstract: Mediation analysis plays an increasingly crucial role in identifying potential causal pathways between exposures and outcomes. However, there is currently a lack of developed mediation approaches for high-dimensional survival data, particularly when considering additive hazard models. The present study introduces two novel approaches for identifying statistically significant mediators in high-dimensional additive hazard models, including the multiple testing-based mediator selection method and knockoff filter procedure. The simulation results demonstrate the outstanding performance of these two proposed methods. Finally, we employ the proposed methodology to analyze the Cancer Genome Atlas (TCGA) cohort in order to identify DNA methylation markers that mediate the association between smoking and survival time among lung cancer patients.

Keywords: high-dimensional mediators; knockoff filter; multiple testing; survival analysis

MSC: 62N02

Citation: An, M.; Zhang, H. High-Dimensional Mediation Analysis for Time-to-Event Outcomes with Additive Hazards Model. *Mathematics* **2023**, *11*, 4891. https://doi.org/10.3390/math11244891

Academic Editor: Manuel Alberto M. Ferreira

Received: 9 November 2023
Revised: 4 December 2023
Accepted: 5 December 2023
Published: 6 December 2023

1. Introduction

The concept of mediation analysis was initially introduced in the field of social psychology [1]. Subsequently, this methodology has been extended to various disciplines, establishing itself as a valuable and widely adopted tool in the social sciences [2,3], epigenomics [4,5], and biomedical sciences [6,7]. In the field of biomedical sciences, mediation analysis plays a pivotal role by quantifying the intermediate effects of mediators on the causal pathway from treatment to outcome. The advancement of data collection technology has led to a significant emphasis on the analysis and processing of high-dimensional mediators. For example, Ref. [8] used a joint significance test for mediation effects in high-dimensional epigenetic studies. Ref. [9] studied sparse principal component-based high-dimensional mediation analysis. Ref. [10] proposed a high-dimensional mediation analysis model with latent variables. Refs. [11,12] studied the mediation effects in high-dimensional and compositional microbiome data. The two review papers by [13,14] provide further insights into high-dimensional mediation analysis.

As the field of mediation analysis continues to advance, its scope has expanded beyond continuous and binary outcomes [15,16] to encompass time-to-event outcomes [17,18], which find significant applications in genetics and biomedical sciences that frequently encounter censored survival data. The availability of high-dimensional mediation analysis methods for time-to-event outcome data remains limited. For example, Ref. [19] introduced a novel approach for high-dimensional mediation analysis in Cox's model, which incorporates sure independent screening and minimax concave penalty techniques to select relevant variables. Ref. [20] proposed a novel approach to accurately control the false discovery rate (FDR) in high-dimensional Cox mediation regression, enabling the identification of potential mediators. Ref. [21] proposed a high-dimensional mediation analysis procedure in the Cox model that utilized propensity scores to adjust for potential confounders. Ref. [22] introduced a novel approach that incorporates the aggregation of multiple knockoffs into the Cox model for analyzing survival outcomes with high-dimensional mediators. The

current literature on high-dimensional mediation analysis of additive hazard models for survival data is limited, with the exception of [23]. This kind of high-dimensional topic also has potential applications in other fields [24–28]. The present study introduces two novel approaches for identifying statistically significant mediators in high-dimensional additive hazard models: a multiple testing-based mediator selection method and a knockoff filter procedure.

The remaining sections of the paper are structured as follows. In Section 2, we present the additive hazard model and its corresponding notation. In Section 3, we propose a three-step approach for mediator selection based on multiple testing. In Section 4, we introduce a knockoff filter procedure designed for high-dimensional mediators. In Section 5, we evaluate the performance of our proposed method through numerical simulations. In Section 6, we apply the proposed method in the context of the TCGA project. The paper concludes with a discussion in Section 7, offering some final remarks.

2. Model and Notations

With the rapid advancement of information technology, certain conventional mediation analysis methods fail to meet the demands of practical analysis, particularly when dealing with high-dimensional survival data. Nevertheless, time-to-event data are prevalent in genomics and bioinformatics research. Consequently, this study presents an introduction to high-dimensional mediation analysis using the additive hazards model for survival data. Let $\tilde{T}_i = \min(T_i, C_i)$ represent the observed failure time, where T_i is the survival time of the ith individual, and C_i denotes the censoring time, $i = 1, ..., n$. Denote $\delta_i = I(T_i \leqslant C_i)$ as the failure indicator, where $I(\cdot)$ is the indicator function. We consider the following high-dimensional survival mediation model with the additive hazards model:

$$M_k = c_k + \alpha_k X + \boldsymbol{\zeta}_k' \mathbf{Z} + e_k, \quad k = 1, \cdots, p, \tag{1}$$

$$\lambda(t|X, \mathbf{M}, \mathbf{Z}) = \lambda_0(t) + \gamma X + \beta_1 M_1 + \cdots + \beta_p M_p + \boldsymbol{\eta}' \mathbf{Z}, \tag{2}$$

where $\lambda_0(t)$ is the baseline hazard function, X is an exposure, $\mathbf{Z} = (Z_1, \cdots, Z_q)'$ is a vector of confounding variables, and $\mathbf{M} = (M_1, \cdots, M_p)'$ is a vector of p-dimensional mediators with $p \gg n$; γ is the "direct effect" of X on the hazard of T, after adjusting for all mediators and covariates. $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)'$ is a vector of parameters relating the exposure to p mediating variables, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is a vector of parameters relating the mediators to T adjusting for the exposure and covariates. $\boldsymbol{\zeta}_k$'s and $\boldsymbol{\eta}$ are the parameters of covariates. In addition, c_k's are the intercept terms; e_k's are error terms. By [29], the "indirect effect" along the path $X \to M_k \to T$ is $\alpha_k \beta_k$ for $k = 1, \cdots, p$. Let $S_0 = \{k : \alpha_k \beta_k \neq 0, k = 1, \cdots, p\}$ be the index set of significant mediators.

For convenience, we define the counting process and risk process as $N_i(t) = I(\tilde{T}_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(\tilde{T}_i \geq t)$, respectively. Let $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}', \boldsymbol{\eta}')'$ and $\boldsymbol{Q}_i = \left(X_i, \mathbf{M}_i', \mathbf{Z}_i'\right)'$, where $i = 1, \cdots, n$. According to [30], the corresponding pseudo-likelihood score function of the additive hazards model is

$$\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \int_0^{\tau} \{\boldsymbol{Q}_i - \bar{\boldsymbol{Q}}(t)\} \{\mathrm{d}N_i(t) - Y_i(t) \boldsymbol{\theta}' \boldsymbol{Q}_i \mathrm{d}t\}, \tag{3}$$

where $\bar{\boldsymbol{Q}}(t) = \sum_{j=1}^{n} Y_j(t) \boldsymbol{Q}_j / \sum_{j=1}^{n} Y_j(t)$ and τ is the length of study. We can write the score function as

$$\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{h} - \boldsymbol{V}' \boldsymbol{\theta}, \tag{4}$$

where

$$\boldsymbol{h} = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \{\boldsymbol{Q}_i - \bar{\boldsymbol{Q}}(t)\} \mathrm{d}N_i(t),$$

$$\boldsymbol{V} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} Y_i(t)\{\boldsymbol{Q}_i - \bar{\boldsymbol{Q}}(t)\}^{\otimes 2}\,\mathrm{d}t,$$

and $a^{\otimes 2} = aa'$. Based on (4), the loss function of the additive hazards model has the following form:

$$L(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}'\boldsymbol{V}\boldsymbol{\theta} - \boldsymbol{h}'\boldsymbol{\theta}. \tag{5}$$

3. Multiple Testing-Based Mediator Selection

In this section, we are interested in selecting significant mediators in models (1) and (2) with $p \gg n$. Our proposed new approach for achieving this objective involves a three-step multiple testing-based mediator selection method (Figure 1):

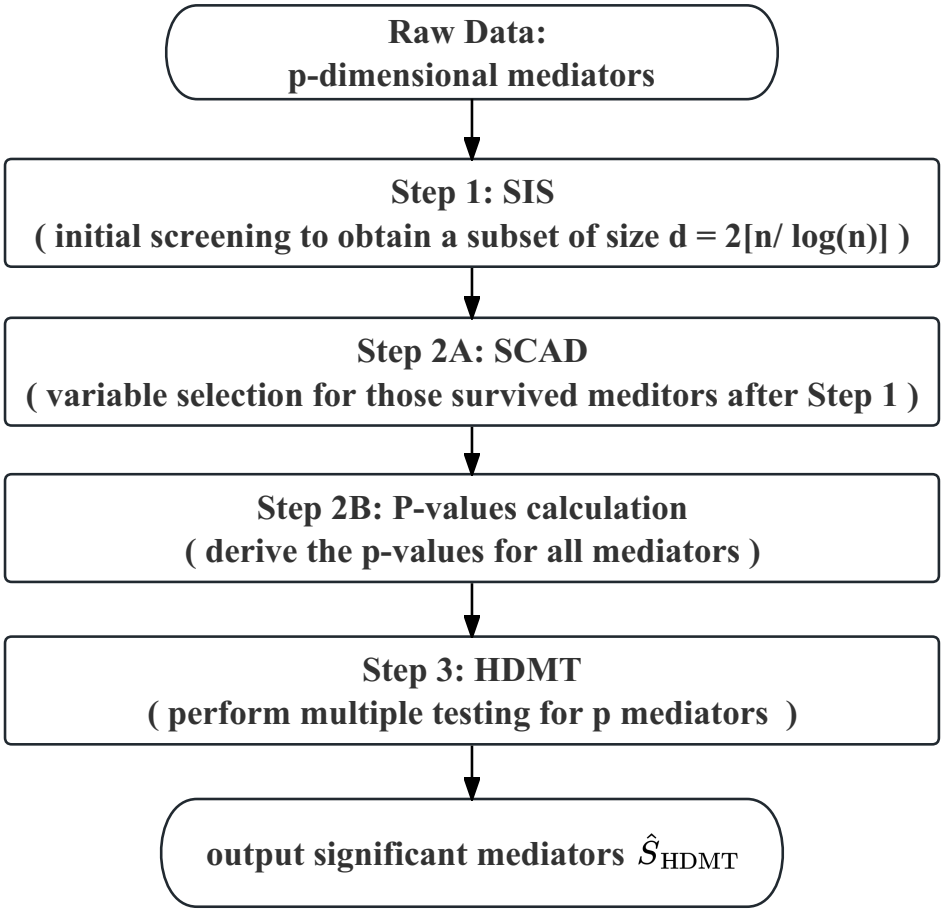

Figure 1. The workflow of multiple testing-based mediator selection procedure.

Step 1. (*Mediator screening*). First, all mediators are standardized with mean zero and variance one. For $k = 1, \cdots, p$, we perform a series of marginal mediation models:

$$M_k = c_k + \alpha_k X + \boldsymbol{\zeta}_k'\mathbf{Z} + e_k, \tag{6}$$

$$\lambda(t|X, M_k, \mathbf{Z}) = \lambda_0(t) + \gamma X + \beta_k M_k + \boldsymbol{\eta}'\mathbf{Z}. \tag{7}$$

Using the idea of sure independence screening [31], we can identify a subset $S_1 = \{k : M_k$ is among the top $d = 2[n/\log(n)]$ mediators with the largest $\min(|\hat{\alpha}_k|, |\hat{\beta}_k|), k = 1, \cdots, p\}$, where $\hat{\alpha}_k$ and $\hat{\beta}_k$ are the estimates based on the marginal models (6) and (7), respectively.

Step 2A. (*SCAD-based variable selection*). Using the mediators M_k's that survived from Step 1, we further minimize the following penalized loss function:

$$Q(\boldsymbol{\theta}_{S_1}) = L(\boldsymbol{\theta}_{S_1}) + \sum_{j \in S_1} p_\lambda(|\beta_j|), \tag{8}$$

where $\boldsymbol{\theta}_{S_1} = (\gamma, \boldsymbol{\beta}'_{S_1}, \boldsymbol{\eta}')'$, and $\boldsymbol{\beta}_{S_1}$ denotes the subvector of $\boldsymbol{\beta}$ with index belonging to S_1; $p_\lambda(\cdot)$ is the Smoothly Clipped Absolute Deviation Fan and Li [32], and its derivative function is

$$p'_\lambda(|\beta|) = \lambda I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{a-1} I(|\beta| > \lambda),$$

where $a > 2$ is the adjustable parameter, and $\lambda > 0$ is the regularization. Based on [32], we set $a = 3.7$ in practical applications. Meanwhile, the parameter λ is determined by five-fold cross-validation. The subset S_2 can be obtained as $\{k : \hat{\beta}_k \neq 0, k \in S_1\}$, which is based on the SCAD-penalized estimates $\hat{\beta}_k$'s.

Step 2B. (*p-value calculation*). When the number of mediators p exceeds the sample size n, it becomes challenging to obtain the p-values for β_k's. To address this issue, we propose a method for computing the p-values that is similar to [33]. Let

$$\lambda(t|X, \mathbf{M}_{S_2}, \mathbf{Z}) = \lambda_0(t) + \gamma X + \boldsymbol{\beta}_{S_2} \mathbf{M}_{S_2} + \boldsymbol{\eta}' \mathbf{Z}, \text{ if } j \in S_2; \tag{9}$$

$$\lambda(t|X, \mathbf{M}_{S_2 \cup \{j\}}, \mathbf{Z}) = \lambda_0(t) + \gamma X + \boldsymbol{\beta}_{S_2 \cup \{j\}} \mathbf{M}_{S_2 \cup \{j\}} + \boldsymbol{\eta}' \mathbf{Z}, \text{ if } j \notin S_2. \tag{10}$$

The estimates $\hat{\beta}_j$'s and their standard errors $\hat{\sigma}_{\beta_j}$'s for $j \in S_2$ can be obtained using Equation (9). For $j \notin S_2$, the estimates $\hat{\beta}_j$'s and their standard errors $\hat{\sigma}_{\beta_j}$'s are derived from Equation (10). The p-values for β_j's can be computed accordingly:

$$P_{\beta_j} = 2\Big\{1 - \Phi\Big(|\hat{\beta}_j| / \hat{\sigma}_{\beta_j}\Big)\Big\}, \; j = 1, \cdots, p, \tag{11}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$.

Step 3. (*Multiple testing*). We consider the following multiple testing problem:

$$H_{0k} : \alpha_k \beta_k = 0 \text{ vs. } H_{Ak} : \alpha_k \beta_k \neq 0, \; k = 1, \cdots, p. \tag{12}$$

The above null hypothesis can be equivalently decomposed into three disjoint null hypotheses:

$$H_{00,k} : \alpha_k = 0 \; and \; \beta_k = 0,$$
$$H_{01,k} : \alpha_k = 0 \; and \; \beta_k \neq 0,$$
$$H_{10,k} : \alpha_k \neq 0 \; and \; \beta_k = 0.$$

For $k = 1, \cdots, p$, let

$$P_{max,k} = max(P_{\alpha_k}, P_{\beta_k}), \tag{13}$$

where P_{β_k} is given in (11), $P_{\alpha_k} = 2\{1 - \Phi(|\hat{\alpha}_k| / \hat{\sigma}_{\alpha_k})\}$, $\hat{\alpha}_k$ is the ordinary least square estimator and its standard error is $\hat{\sigma}_{\alpha_k}$. For $t \in [0,1]$, we define $V_{00}(t) = \#\{P_{max,k} \leqslant t | H_{00}\}$, $V_{01}(t) = \#\{P_{max,k} \leqslant t | H_{01}\}$, $V_{10}(t) = \#\{P_{max,k} \leqslant t | H_{10}\}$, $V_{11}(t) = \#\{P_{max,k} \leqslant t | H_{11}\}$, and $R(t) = V_{00}(t) + V_{01}(t) + V_{10}(t) + V_{11}(t)$, where $V_{01}(t), V_{10}(t), V_{00}(t)$ denote the number of false positives in terms of the three different hypotheses. Then we can define the FDR as

$$\text{FDR}(t) = E\left[\frac{V_{00}(t) + V_{01}(t) + V_{10}(t)}{R(t) \vee 1}\right]. \tag{14}$$

Denote π_{00}, π_{01}, and π_{10} as the proportions of the three disjoint null hypotheses $H_{00,k}$,$H_{01,k}$ and $H_{10,k}$, respectively. By [34], the FDR(t) can be expressed as

$$\widehat{\text{FDR}}(t) = \frac{\hat{\pi}_{01} t + \hat{\pi}_{10} t + \hat{\pi}_{00} t^2}{max\{R(t), 1\}/p}.$$

Based on [34], we define a threshold for controlling the FDR:

$$\hat{t}_\delta = \sup\Big\{t : \widehat{\text{FDR}}(t) \leq \delta\Big\}. \tag{15}$$

where δ is the significance level, $\hat{\pi}_{00}$, $\hat{\pi}_{01}$, $\hat{\pi}_{10}$ and $\hat{t}_\delta$ can be obtained from the R package `HDMT` (version 1.0.5). The index set of selected mediators is given as

$$\hat{S}_{\text{HDMT}} = \{k : P_{max,k} \leq \hat{t}_\delta, k = 1, \cdots, p\},$$

where $P_{max,k}$ and $\hat{t}_\delta$ are given in (13) and (15), respectively.

4. Knockoff Filter for High-Dimensional Mediators

In this section, we introduce a novel knockoff filter for the additive hazards model with high-dimensional mediators in models (1) and (2). The details of our method are presented as follows:

Step 1. (*Mediator screening*). First, we use the BH-based initial screening method to reduce the dimension of mediators. To be specific, we calculate the p-value of the effect (α_k) along the path $X \to M_k$ as

$$P_{\alpha_k} = 2\{1 - \Phi(|\widehat{\alpha}_k|/\hat{\sigma}_{\alpha_k})\}, \tag{16}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$, $\hat{\alpha}_k$ is the estimate of the ordinary least square (OLS) based on (1), and $\hat{\sigma}_{\alpha_k}$ is the estimated standard error. Given δ_0, the BH-procedure is applied to perform the initial screening such that

$$d = max\left\{k : P_{(k)} \leq \frac{k\delta_0}{p}\right\}, \ k = 1, \cdots, p, \tag{17}$$

where $P_{(k)}$'s are the order statistics of p-values given in (16). Denote $S_1 = \{k$: M_k's are the d mediators with $P_{(1)} \leq \cdots \leq P_{(d)}\}$, and the corresponding mediators are denoted by $\mathbf{M}_{S_1}$.

Step 2. (*Perform knockoffs*). The knockoff filter, introduced by [35,36], presents a novel variable selection procedure. Its core concept involves constructing a set of knockoff variables that are uncorrelated with the response but possess similar structures to the original covariates. We define $\{\tilde{\mathbf{M}}_{S_1}^{[b]}\}_{b=1}^{B}$ as the knockoffs of $\mathbf{M}_{S_1}$, where $\mathbf{M}_{S_1}$ is an n-by-d mediator matrix and $\tilde{\mathbf{M}}_{S_1}^{[b]}$ is the bth knockoff of $\mathbf{M}_{S_1}$. The knockoffs $\{\tilde{\mathbf{M}}_{S_1}^{[b]}\}_{b=1}^{B}$ possess the following two properties, as stated in [36]:

(I) For any subset $S \subseteq S_1$ and $b \in \{1, ..., B\}$,

$$(\mathbf{M}_{S_1}, \tilde{\mathbf{M}}_{S_1}^{[b]})_{\text{swap }(S)} \overset{d}{=} (\mathbf{M}_{S_1}, \tilde{\mathbf{M}}_{S_1}^{[b]}).$$

The expression "$\overset{d}{=}$" denotes having the same distribution, while "swap (S)" represents the operation of swapping elements in set S. The notation $(\mathbf{M}_{S_1}, \tilde{\mathbf{M}}_{S_1}^{[b]})_{\text{swap }(S)}$ is obtained by interchanging the kth columns of $\mathbf{M}_{S_1}$ and $\tilde{\mathbf{M}}_{S_1}^{[b]}$, where $k \in S$. As an example, if $S_1 = \{1,2,3\}$ and $S = \{2,3\}$, then $(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \tilde{\mathbf{M}}_1^{[b]}, \tilde{\mathbf{M}}_2^{[b]}, \tilde{\mathbf{M}}_3^{[b]})_{\text{swap }(S)} = (\mathbf{M}_1, \tilde{\mathbf{M}}_2^{[b]}, \tilde{\mathbf{M}}_3^{[b]}, \tilde{\mathbf{M}}_1^{[b]}, \mathbf{M}_2, \mathbf{M}_3)$.

(II) The random variables $\tilde{\mathbf{M}}_{S_1}^{[b]}$ and $\mathbf{T}$ are mutually independent given $\mathbf{M}_{S_1}$ for any $b \in \{1, ..., B\}$, where $\mathbf{T}$ represents the vector of failure times. For practical application, $\{\tilde{\mathbf{M}}_{S_1}^{[b]}\}_{b=1}^{B}$ can be generated by the R package `knockoff` (version 0.3.6).

Step 3. (*Mediator selection*). Using $\mathbf{M}_{S_1}$ and the bth knockoff $\tilde{\mathbf{M}}_{S_1}^{[b]}$, together with X and $\mathbf{Z}$, we refit an additive hazards model, denoting $\beta_j^{[b]}$ as the coefficient corresponding to the jth mediator M_j, and $\phi_j^{[b]}$ as the coefficient corresponding to the knockoff $\tilde{M}_j^{[b]}$. Then, we consider minimizing the bth loss function with Lasso penalty, i.e.,

$$\boldsymbol{Q}(\boldsymbol{\theta}_b) = L(\boldsymbol{\theta}_b) + \lambda \sum_{j \in S_1} (\, |\beta_j^{[b]}| + |\phi_j^{[b]}|\,), \ b = 1, ..., B, \tag{18}$$

where $\boldsymbol{\theta}_b$ is the vector of regression coefficients in the refitted additive hazards model, and $L(\boldsymbol{\theta}_b)$ is similarly given as that of (5). In accordance with [22], we proceed to construct the matrix $\boldsymbol{\tau} = (\tau_j^{[b]})_{B\times d}$, where each entry is carefully determined by

$$\tau_j^{[b]} = |\hat{\beta}_j^{[b]}| - |\hat{\phi}_j^{[b]}|,\ b = 1, ..., B, j \in S_1, \tag{19}$$

$\hat{\beta}_j^{[b]}$ and $\hat{\phi}_j^{[b]}$ are the Lasso estimates derived from (18). Following [37], the value of B is set to 25 in the practical application.

The property (II) suggests that $|\hat{\phi}_j^{[b]}|$ should be very small. Consequently, when the jth mediator M_j exhibits significance, $\tau_j^{[b]}$ tends to be a relatively large positive value; whereas for non-significant mediators, $\tau_j^{[b]}$ tends to hover around a small range centered at 0. The statistic $\boldsymbol{\pi} = (\pi_j^{[b]})_{B\times d}$ is obtained in a similar manner as described in [22], with its entries being calculated as

$$\pi_j^{[b]} = \begin{cases} \frac{\#\left\{k\in S_1 : \tau_k^{[b]} \le -\tau_j^{[b]}\right\}}{d}, & \tau_j^{[b]} > 0; \\ 1, & \tau_j^{[b]} \le 0. \end{cases} \tag{20}$$

The small value of $\pi_j^{[b]}$ is worth noting as it indicates a strong mediation signal. This is because when the jth mediator is significant, $\tau_j^{[b]}$ tends to be a relatively large positive value. Consequently, among those mediators M_k screened by S_1, it is rare for the corresponding $\tau_k^{[b]}, k \in S_1$, to be smaller than $-\tau_j^{[b]}$.

According to [38], the statistics $\{\pi_j^{[b]}\}_{b=1}^B$ are aggregated to generate the indicators $\bar{\pi}_j$'s as follows:

$$\bar{\pi}_j = \min\left\{1, \frac{Q_\eta\left(\left\{\pi_j^{[b]} : b \in \{1, ..., B\}\right\}\right)}{\eta}\right\},\ j \in S_1, \tag{21}$$

where η is the pre-specified quantile parameter, and $Q_\eta(\cdot)$ indicates the η-quantile function. Based on the procedure for [37], we use $\eta = 0.3$ in practice. Next, we apply the BH method to determine the thresholds $\hat{t}_\delta$ with FDR control at

$$\hat{t}_\delta = \max\left\{k : \bar{\pi}_{(k)} \le \frac{k\delta}{d}\right\},\ k = 1, ..., d, \tag{22}$$

where $\bar{\pi}_{(k)}, k = 1, ..., d$ is the order statistic of $\bar{\pi}_j, j \in S_1$. The estimated index set of active mediators is given as

$$\hat{\mathcal{S}}_{\mathrm{KF}} = \left\{j \in S_1 : \bar{\pi}_j \le \bar{\pi}_{(\hat{t}_\delta)}\right\}. \tag{23}$$

For the mediator screened by Step 1, we can conclude that this mediator M_j is significant if its statistic $\bar{\pi}_j$ is smaller than the specified value $\bar{\pi}_{(\hat{t}_\delta)}$. The overall workflow of this procedure is illustrated in Figure 2 as a concise summary.

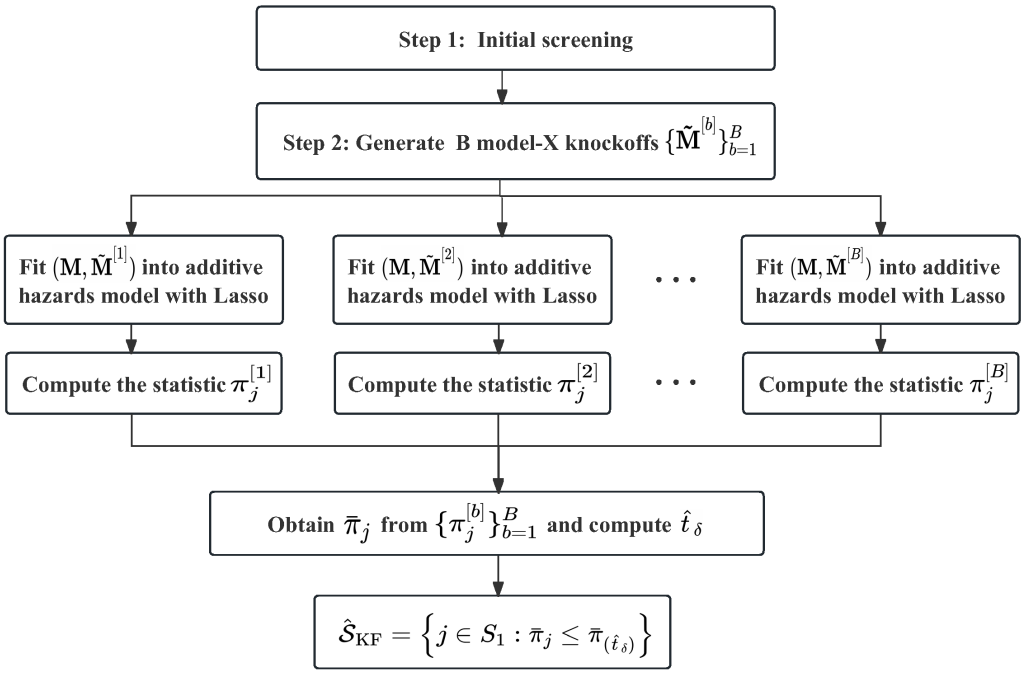

Figure 2. The workflow of knockoff-based mediator selection procedure.

5. Simulation Studies

In this section, we conduct some simulations under two different types of scenarios (binary and continuous exposure) to evaluate the performance of our two proposed approaches, where the ones mentioned in the previous Sections 2 and 3 are denoted as 'HDMT' and 'Knockoff', respectively. We also compare with the one in paper [23], denoted as 'Cui'. Similar to [22], we choose $\delta_0 = 0.2$ for 'Knockoff' in the simulations.

First, we generate failure times $T_1, \cdots, T_n$ from the additive hazards model with $\lambda(t|X, \mathbf{M}, \mathbf{Z}) = \lambda_0(t) + \gamma X + \beta_1 M_1 + \cdots + \beta_p M_p + \boldsymbol{\eta}'\mathbf{Z}$, where $\lambda_0(t) = 1$, $\gamma = 0.3$, and $\boldsymbol{\eta} = (0.5, 0.5)'$; the mediators are generated from linear models $M_k = c_k + \alpha_k X + \boldsymbol{\zeta}_k'\mathbf{Z} + e_k$, where $\boldsymbol{\zeta}_k = (0.3, 0.3)'$, the intercept term c_k is generated by the uniform distribution $U(0, 0.2)$ and the random error term e_k is generated from $N(0, 1)$. We set $\alpha = (0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.2, 0.2, 0, \cdots, 0)'$ and $\beta = (1, 1, 1, 1, 1, 1, 0, 0, 0.2, 0.2, 0, \cdots, 0)'$, where the dimension of mediators is p = 10,000. That is, the first six mediators $\{M_k\}_{k=1}^{6}$ are significant. Moreover, we consider two different scenarios for the exposure X and covariate $\mathbf{Z}$:

Case 1 (*binary exposure*): The exposure X follows from $B(1, 0.6)$, and the covariate $\mathbf{Z} = (Z_1, Z_2)'$, where Z_1 and Z_2 are independently generated from $B(1, 0.3)$ and $U(0, 1)$, respectively.

Case 2 (*continuous exposure*): The exposure X follows from $N(0, 2)$, and the covariate $\mathbf{Z} = (Z_1, Z_2)'$, where both Z_1 and Z_2 are independently generated from $N(0, 2)$.

The censoring time follows a uniform distribution $U(0, c_0)$, where c_0 is adjusted to achieve an average censoring rate of approximately 20% ($c_0 = 1$) and 50% ($c_0 = 0.2$), respectively. The sample size is chosen as n = 300, 500, and 800, respectively. All simulation results are based on 500 repetitions. We compare the two proposed methods with Cui's method, we set the pre-specified FDR level δ as 0.05 in (22). Let S_0 be the set of significant mediators, that is, $S_0 = \{1, 2, \cdots, 6\}$, and let $\hat{S}$ be the estimated index set of significant mediators. To evaluate the performance of mediator selection, we use the following five indicators: the model size (MS) $|\hat{S}|$; the rate that the correct model is selected (CMR) with $I(\hat{S} = S_0)$; the false discovery proportion (FDP) with $FDP = \frac{|\hat{S} \setminus S_0|}{|\hat{S}|}$; the true positive rate (TPR) with $TPR = \frac{|\hat{S} \cap S_0|}{|S_0|}$.

The performance of the three methods for mediator selection is demonstrated with the simulation results presented in Tables 1 and 2. The results with binary exposure are presented in Table 1, indicating that Knockoff exhibits higher TPR and lower FDP, while Cui's method demonstrates a significantly more conservative approach. The results clearly demonstrate that Knockoff's TPR and CMR consistently outperform Cui's method, while the MS is also more aligned with the simulation setup's model size. Although FDP slightly exceeds Cui's method, all values remain below 0.05. However, it should be noted that Cui's approach exhibits excessive conservatism by prioritizing a consistently low FDP at the expense of other indicators. The results with continuous exposure are presented in Table 2, revealing the superiority of the HDMT method over Cui's method across three indicators: MS, CMR, and TPR. The TPR of HDMT consistently outperforms Cui's method, particularly at a censoring rate of 20%, where the TPR remains above 0.875 regardless of the sample size. Additionally, HDMT's CMR and MS indicators demonstrate strong performance across most scenarios. It is worth noting that the performance of the Knockoff method on the FDP indicator is suboptimal in cases of continuous exposure, and similar conclusions have been drawn in [22]. Therefore, we propose the following recommendations for practical applications: if the exposure variable X is continuous, it may be more appropriate to consider using the HDMT method; whereas if X is binary, then the Knockoff method can be considered.

Table 1. The results about mediator selection with binary exposure.

Sample Size		CR = 20%			CR = 50%		
		Cui	HDMT	Knockoff	Cui	HDMT	Knockoff
$n = 300$	MS	3.05	4.42	5.902	1.276	2.018	4.638
	CMR	0.032	0.078	0.524	0	0.008	0.202
	FDP	0.032	0.083	0.039	0.035	0.066	0.034
	TPR	0.481	0.637	0.937	0.193	0.286	0.739
$n = 500$	MS	5.162	6.276	6.222	3.042	4.088	5.866
	CMR	0.362	0.396	0.71	0.044	0.106	0.536
	FDP	0.028	0.092	0.039	0.039	0.075	0.041
	TPR	0.83	0.923	0.989	0.479	0.603	0.93
$n = 800$	MS	6.044	6.59	6.238	4.9	5.852	6.252
	CMR	0.728	0.594	0.786	0.25	0.314	0.736
	FDP	0.027	0.077	0.033	0.019	0.075	0.039
	TPR	0.975	0.996	1	0.796	0.883	0.993

Table 2. The results about mediator selection with continuous exposure.

Sample Size		CR = 20%			CR = 50%		
		Cui	HDMT	Knockoff	Cui	HDMT	Knockoff
$n = 300$	MS	3.47	5.744	7.184	1.012	2.304	5.872
	CMR	0.066	0.454	0.18	0	0.032	0.118
	FDP	0.025	0.063	0.168	0.019	0.042	0.133
	TPR	0.554	0.875	0.978	0.159	0.338	0.833
$n = 500$	MS	5.292	6.376	7.78	1.962	4.666	6.87
	CMR	0.362	0.69	0.098	0.002	0.232	0.152
	FDP	0.025	0.053	0.216	0.013	0.059	0.164
	TPR	0.855	0.994	0.998	0.317	0.702	0.941
$n = 800$	MS	5.974	6.362	7.976	3.482	6.174	7.382
	CMR	0.784	0.73	0.09	0.064	0.522	0.15
	FDP	0.017	0.048	0.232	0.015	0.068	0.182
	TPR	0.976	0.999	0.999	0.567	0.942	0.989

6. Application

The incidence and mortality rates of lung cancer are rapidly increasing, making it the most formidable menace to human health. The findings of numerous studies indicate that individuals who engage in long-term heavy smoking are at a higher risk of developing lung cancer compared to non-smokers. Furthermore, research has demonstrated a correlation between smoking and DNA methylation [39]. The identification of DNA methylation CpG sites that mediate the association between smoking and lung cancer holds significant practical implications. The proposed methods are applied to analyze the TCGA lung cancer cohort, which can be freely accessed at https://xenabrowser.net/datapages/.

In our analysis, we specifically focus on the clinical presentation and epigenetic information of a cohort comprising 754 patients, whose ages span from 33 to 90 years. The survival endpoint is defined as the duration from initial diagnosis to either death or the last follow-up, with a median survival time of 658 days. This also encompasses censoring, indicating that 305 patients experienced mortality during the follow-up period, resulting in a censoring rate of 59%. The Infinium HumanMethylation450 BeadChip array analysis identified a total of 365,306 potential mediators in the form of DNA methylation CpG sites. The exposure is the smoking status (smoker = 1; non-smoker = 0), and the survival time is the outcome variable. The analysis also incorporates four covariates, namely age at initial diagnosis, gender (Male = 1; Female = 0), tumor stage (Stage I = 1; Stage II = 2; Stage III = 3; Stage IV = 4), and radiotherapy (Yes = 1; No = 0). The objective of our study is to ascertain the mediating methylation markers that link smoking and survival in patients with lung cancer.

The Knockoff method is adopted due to the binary nature of the exposure. The conclusions drawn from simulations suggest a threshold of $\delta_0 = 3 \times 10^{-4}$ in the initial stage of the screening process. By applying this approach to identify DNA methylation CpG sites, we successfully obtain two CpG sites (cg21926276, cg24200525) with significant mediation effects, as presented in Table 3. From this, we can see that the mediating effect of cg21926276, $\hat{\alpha}_k\hat{\beta}_k$, is greater than 0, suggesting that smoking through this methylation CpG site increases the probability of mortality. The aforementioned conclusion can be further substantiated by previous empirical research. [40] showed that the gene H19 (cg21926276 locate) is associated with lung cancer and tumor growth. For cg24200525, its mediation effect $\hat{\alpha}_k\hat{\beta}_k$ is less than 0, which is consistent with previous studies [19,22]. The analysis led to the identification of two specific methylation sites (smoking $\rightarrow$ cg21926276 $\rightarrow$ survival time, smoking $\rightarrow$ cg24200525 $\rightarrow$ survival time) that serve as significant mediators among the numerous potential methylation sites. The Cui method was also employed for the analysis of this lung cancer dataset. However, this approach proved ineffective in identifying any of the significant mediators. Hence, the proposed method we have presented is of greater practical value, thus making it more suitable for real-world applications.

Table 3. The summary of significant mediating CpGs with Knockoff method.

CpGs	Chromosome	Gene	$\hat{\alpha}_k$(se)	$\hat{\beta}_k$(se)	$P_{max,k}$
cg21926276	Chr11	H19	$-0.0584(0.0106)$	$-0.0007(0.0002)$	0.0016
cg24200525	Chr22	SBF1	$-0.0241(0.0044)$	$0.0019(0.0005)$	0.0005

7. Conclusions

In this paper, we have introduced two innovative approaches for identifying statistically significant mediators in high-dimensional additive hazard models, including a mediator selection method based on multiple testing and the implementation of a knockoff filter procedure. The simulation results demonstrated the exceptional performance of these two proposed methods. Due to the binary nature of the exposure, we applied the Knockoff methodology to analyze the Cancer Genome Atlas (TCGA) cohort to identify two DNA methylation markers (cg21926276, cg24200525) that mediate the association between smoking and survival time among lung cancer patients. Finally, we proposed the following

recommendations for practical applications: if the exposure variable X is continuous, it may be more appropriate to consider using the HDMT method; whereas if X is binary, then the Knockoff method can be considered.

Author Contributions: Methodology, H.Z.; formal analysis, M.A.; writing—original draft preparation, M.A.; writing—review and editing, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The collected data used in this study are available at https://xenabrowser.net/datapages/.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baron, R.M.; Kenny, D.A. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Personal. Soc. Psychol.* **1986**, *51*, 1173. [CrossRef] [PubMed]
2. Valeri, L.; VanderWeele, T.J. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychol. Methods* **2013**, *18*, 137. [CrossRef] [PubMed]
3. VanderWeele, T.J. Mediation analysis: A practitioner's guide. *Annu. Rev. Public Health* **2016**, *37*, 17–32. [CrossRef] [PubMed]
4. Song, Y.; Zhou, X.; Zhang, M.; Zhao, W.; Liu, Y.; Kardia, S.L.R.; Roux, A.V.D.; Needham, B.L.; Smith, J.A.; Mukherjee, B. Bayesian Shrinkage Estimation of High Dimensional Causal Mediation Effects in Omics Studies. *Biometrics* **2020**, *76*, 700–710. [CrossRef] [PubMed]
5. Dowling, C.M.; Hayes, S.L.; Phelan, J.J.; Cathcart, M.C.; Finn, S.P.; Mehigan, B.J.; McCormick, P.H.; Coffey, J.C.; O'Sullivan, J.N.; Kiely, P. Expression of protein kinase C gamma promotes cell migration in colon cancer. *Oncotarget* **2017**, *8*, 72096–72107. [CrossRef] [PubMed]
6. Huang, Y.T. Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics. *Ann. Appl. Stat.* **2018**, *12*, 1535–1557. [CrossRef]
7. Charalambous, A.; Giannakopoulou, M.; Bozas, E.; Paikousis, L. Parallel and serial mediation analysis between pain, anxiety, depression, fatigue and nausea, vomiting and retching within a randomised controlled trial in patients with breast and prostate cancer. *BMJ Open* **2019**, *9*, e026809. [CrossRef]
8. Zhang, H.; Zheng, Y.; Zhang, Z.; Gao, T.; Joyce, B.T.; Yoon, G.; Zhang, W.; Schwartz, J.D.; Just, A.C.; Colicino, E.; et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **2016**, *32*, 3150–3154. [CrossRef]
9. Zhao, Y.; Lindquist, M.A.; Caffo, B.S. Sparse principal component based high-dimensional mediation analysis. *Comput. Stat. Data Anal.* **2020**, *142*, 106835. [CrossRef]
10. Derkach, A.; Pfeiffer, R.M.; Chen, T.H.; Sampson, J.N. High dimensional mediation analysis with latent variables. *Biometrics* **2019**, *75*, 745–756. [CrossRef]
11. Zhang, H.; Chen, J.; Feng, Y.; Wang, C.; Li, H.; Liu, L. Mediation effect selection in high-dimensional and compositional microbiome data. *Stat. Med.* **2021**, *40*, 885–896. . [CrossRef] [PubMed]
12. Zhang, H.; Chen, J.; Li, Z.; Liu, L. Testing for mediation effect with application to human microbiome data. *Stat. Biosci.* **2021**, *13*, 313–328. . [CrossRef] [PubMed]
13. Zeng, P.; Shao, Z.; Zhou, X. Statistical methods for mediation analysis in the era of high-throughput genomics: Current successes and future challenges. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3209–3224. . [CrossRef] [PubMed]
14. Zhang, H.; Hou, L.; Liu, L. A review of high-dimensional mediation analyses in DNA methylation studies. In *Epigenome-Wide Association Studies: Methods and Protocols*; Weihua, G., Ed.; Springer: Berlin/Heidelberg, Germany, 2022; Volume 2432. . [CrossRef]
15. Valeri, L.; Lin, X.; VanderWeele, T.J. Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. *Stat. Med.* **2014**, *33*, 4875–4890. [CrossRef] [PubMed]
16. Gaynor, S.M.; Schwartz, J.; Lin, X. Mediation analysis for common binary outcomes. *Stat. Med.* **2019**, *38*, 512–529. [CrossRef]
17. Tein, J.Y.; MacKinnon, D.P. Estimating mediated effects with survival data. In *New Developments in Psychometrics: Proceedings of the International Meeting of the Psychometric Society (IMPS2001), Osaka, Japan, 15–19 July 2001*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 405–412.
18. Gelfand, L.A.; MacKinnon, D.P.; DeRubeis, R.J.; Baraldi, A.N. Mediation analysis with survival outcomes: Accelerated failure time vs. proportional hazards models. *Front. Psychol.* **2016**, *7*, 423. [CrossRef] [PubMed]
19. Luo, C.; Fa, B.; Yan, Y.; Wang, Y.; Zhou, Y.; Zhang, Y.; Yu, Z. High-dimensional mediation analysis in survival models. *PLoS Comput. Biol.* **2020**, *16*, e1007768. [CrossRef]
20. Zhang, H.; Zheng, Y.; Hou, L.; Zheng, C.; Liu, L. Mediation analysis for survival data with high-dimensional mediators. *Bioinformatics* **2021**, *37*, 3815–3821. [CrossRef]
21. Yu, Z.; Cui, Y.; Wei, T.; Ma, Y.; Luo, C. High-dimensional mediation analysis with confounders in survival models. *Front. Genet.* **2021**, *12*, 688871. [CrossRef]

22. Tian, P.; Yao, M.; Huang, T.; Liu, Z. CoxMKF: A knockoff filter for high-dimensional mediation analysis with a survival outcome in epigenetic studies. *Bioinformatics* **2022**, *38*, 5229–5235. [CrossRef]
23. Cui, Y.; Luo, C.; Luo, L.; Yu, Z. High-dimensional mediation analysis based on additive hazards model for survival data. *Front. Genet.* **2021**, *12*, 771932. [CrossRef] [PubMed]
24. Yang, X.; Wu, L.; Zhang, H. A space-time spectral order sinc-collocation method for the fourth-order nonlocal heat model arising in viscoelasticity. *Appl. Math. Comput.* **2023**, *457*, 128192. [CrossRef]
25. Zhang, H.; Liu, Y.; Yang, X. An efficient ADI difference scheme for the nonlocal evolution problem in three-dimensional space. *J. Appl. Math. Comput.* **2023**, *69*, 651–674. [CrossRef]
26. Tian, Q.; Yang, X.; Zhang, H.; Xu, D. An implicit robust numerical scheme with graded meshes for the modified Burgers model with nonlocal dynamic properties. *Comput. Appl. Math.* **2023**, *42*, 246. [CrossRef]
27. Wang, W.; Zhang, H.; Jiang, X.; Yang, X. A high-order and efficient numerical technique for the nonlocal neutron diffusion equation representing neutron transport in a nuclear reactor. *Ann. Nucl. Energy* **2024**, *195*, 110163. [CrossRef]
28. Zhou, Z.; Zhang, H.; Yang, X. H1-norm error analysis of a robust ADI method on graded mesh for three-dimensional subdiffusion problems. In *Numerical Algorithms*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 1–19.
29. Huang, Y.T.; Yang, H.I. Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology* **2017**, *28*, 370–378. . [CrossRef]
30. Lin, D.Y.; Ying, Z. Semiparametric analysis of the additive risk model. *Biometrika* **1994**, *81*, 61–71. [CrossRef]
31. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser.* **2008**, *70*, 849–911. . [CrossRef]
32. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. . [CrossRef]
33. Hao, N.; Zhang, H.H. Oracle *p*-values and variable screening. *Electron. J. Stat.* **2017**, *11*, 3251–3271. [CrossRef]
34. Dai, J.Y.; Stanford, J.L.; LeBlanc, M. A multiple-testing procedure for high-dimensional mediation hypotheses. *J. Am. Stat. Assoc.* **2022**, *117*, 198–213. [CrossRef] [PubMed]
35. Barber, R.F.; Candès, E.J. Controlling the false discovery rate via knockoffs. *Ann. Stat.* **2014**, *43*, 2055–2085. [CrossRef]
36. Candès, E.J.; Fan, Y.; Janson, L.; Lv, J. Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2016**, *80*, 551–577. [CrossRef]
37. Nguyen, T.B.; Chevalier, J.A.; Thirion, B.; Arlot, S. Aggregation of multiple knockoffs. *Int. Conf. Mach. Learn.* **2020**, *119*, 7283–7293.
38. Meinshausen, N.; Meier, L.; Bühlmann, P. *p*-Values for High-Dimensional Regression. *J. Am. Stat. Assoc.* **2008**, *104*, 1671–1681. [CrossRef]
39. Govindan, R.; Ding, L.; Griffith, M.; Subramanian, J.; Dees, N.D.; Kanchi, K.L.; Maher, C.A.; Fulton, R.S.; Fulton, L.; Wallis, J.W.; et al. Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. *Cell* **2012**, *150*, 1121–1134. [CrossRef]
40. Matouk, I.J.; Halle, D.; Gilon, M.; Hochberg, A. The non-coding RNAs of the H19-IGF2 imprinted loci: A focus on biological roles and therapeutic potential in Lung Cancer. *J. Transl. Med.* **2015**, *13*, 113. [CrossRef]