CoxMDS: Multiple Data Splitting for High-dimensional Mediation Analysis with Survival Outcomes in Epigenome-wide Studies

Minhao Yao[1,*], Peixin Tian[2,*], Xihao Li[3,4], Shijia Bian[5], Gao Wang[6,7], Yian Gu[6,7,8,9], Ana Navas-Acien[10], Badri N. Vardarajan[7,8], Daniel W. Belsky[9,11], Gary W. Miller[10], Andrea A. Baccarelli[12], Zhonghua Liu[13,#], for the Alzheimer's Disease Neuroimaging Initiative[†]

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong SAR, China.

[2] First Affiliated Hospital of Kunming Medical University, Yunnan, China.

[3] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

[4] Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

[5] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA.

[6] Center for Statistical Genetics, Gertrude H. Sergievsky Center, Columbia University Medical Center, New York, NY, USA.

[7] Department of Neurology, Columbia University, New York, NY, USA.

[8]Taub Institute on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, USA.

[9] Department of Epidemiology, Joseph P. Mailman School of Public Health, Columbia University, New York, USA.

[10] Department of Environmental Health Sciences, Columbia University, New York, NY, USA.

[11] Butler Columbia Aging Center, Columbia University Mailman School of Public Health, New York, USA.

[12] Office of the Dean, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[13] Department of Biostatistics, Columbia University, New York, NY, USA.

[*] Contributed equally. [#] Correspondence to: Zhonghua Liu (zl2509@cumc.columbia.edu)

Abstract

Causal mediation analysis investigates whether the effect of an exposure on an outcome operates through intermediate variables known as mediators. Although progress has been made in high-dimensional mediation analysis, current methods do not reliably control the false discovery rate (FDR) in finite samples, especially when mediators are moderately to highly correlated or follow non-Gaussian distributions. These challenges frequently arise in DNA methylation studies. We introduce CoxMDS, a multiple data splitting method that uses Cox proportional hazards models to identify putative causal mediators for survival outcomes. CoxMDS ensures finite-sample FDR control even in the presence of correlated or non-Gaussian mediators. Through simulations, CoxMDS is shown to maintain FDR control and achieve higher statistical power compared with existing approaches. In applications to DNA methylation data with survival outcomes, CoxMDS

identified eight CpG sites in The Cancer Genome Atlas (TCGA) that are consistent with the hypothesis that DNA methylation may mediate the effect of smoking on lung cancer survival, and two CpG sites in the Alzheimer's Disease Neuroimaging Initiative (ADNI) that are consistent with the hypothesis that DNA methylation may mediate the effect of smoking on time to Alzheimer's disease conversion.

Keywords: Causal mediation analysis; False discovery rate; Survival outcomes; DNA methylation; Data splitting

1 Introduction

Causal mediation analysis aims to understand the causal pathway from an exposure to an outcome of interest through an intermediate mediator variable [1–4], and has become a valuable tool in genome-wide epigenetic studies, where DNA methylation CpG sites mediate the relationships between modifiable exposures and health outcomes [5–9].

DNA methylation involves the covalent bonding of a methyl group (CH3) to the C5 position of cytosine, which plays a regulatory role in gene expression by either interacting with relevant genes or obstructing the binding of transcription factors to DNA [10]. Since DNA methylation is a reversible biochemical process [11], targeting DNA methylation may offer novel therapeutic opportunities for disease prevention and intervention [12, 13].

On Illumina Infinium DNA methylation microarrays (such as the HumanMethylation450 and EPIC BeadChips), methylation at each CpG site is quantified using paired probe intensities for methylated ($Methy$) and unmethylated ($Unmethy$) signals. The methylation level is usually summarized as either the β-value, $\beta = Methy/(Methy + Unmethy + \alpha)$, which ranges from 0 to 1 with a small offset α, or the M-value, a logit-like transformation of the β-value [14–16]. β-values are often modeled using beta-family distributions [17, 18], whereas M-values, even after transformation, are not guaranteed to follow a Gaussian distribution [19]. Furthermore, methylation levels show strong local correlation among neighboring CpGs, a phenomenon known as co-methylation [20].

As a result, array-based DNA methylation data are typically non-Gaussian and highly correlated, which creates major challenges for statistical modeling and limits the effectiveness of existing methods for high-dimensional mediation analysis. In particular, it remains difficult to reliably identify CpG sites that mediate the effect of an exposure on a survival outcome while controlling the false discovery rate (FDR) [21] in finite samples. Several approaches have extended the Cox proportional hazards model [22] to accommodate high-dimensional mediators. For example, Luo et al. [7] combined sure independence screening (SIS) [23], the minimax concave penalty (MCP) [24], and the Sobel test [25] to select mediators from DNA methylation data, and Zhang et al. [26] proposed a de-biased Lasso estimator for survival outcomes. However, neither method provides finite-sample FDR control. More

recently, Tian et al. [27] introduced the CoxMKF method, which aggregates multiple knockoffs to achieve finite-sample FDR control. Yet, the knockoff filter may perform poorly when mediators are moderately to highly correlated [28], and because knockoff copies are generated under a multivariate Gaussian assumption, power can decline when the mediators deviate from Gaussianity [28]. These limitations highlight the need for new methods that guarantee finite-sample FDR control while retaining power in the presence of correlated and non-Gaussian mediators.

In this paper, we propose CoxMDS, a new method for identifying DNA methylation CpG sites that mediate the effect of an exposure on a survival outcome. Building on the multiple data splitting strategy of Dai et al. [28], CoxMDS guarantees finite-sample FDR control even when mediators are moderately to highly correlated or follow non-Gaussian distributions. The procedure consists of three main steps:

Step 1. **Candidate mediator filtering**: Following prior work [7, 27], we fit a linear regression model for each mediator to test the exposure–mediator association. The resulting p-values are adjusted using the Benjamini–Hochberg (BH) procedure [21] to retain the most significant candidate mediators.

Step 2. **Mirror statistic construction**: The data are randomly split into two groups. In Group I, we fit a Cox proportional hazards model with the minimax concave penalty (MCP) to the candidate mediators from Step 1 and retain those with non-zero coefficients. In Group II, we refit the Cox model including only the selected mediators from Group I. For each mediator, a mirror statistic is then computed by combining the coefficients from both groups. This process is repeated multiple times to stabilize the results.

Step 3. **Mediator selection with finite-sample FDR control**: For each candidate mediator, we calculate its inclusion rate [28], which measures its frequency of selection across data splits. A data-adaptive threshold is applied to the inclusion rates to select mediators at the target FDR level.

Through extensive simulations, we show that CoxMDS guarantees finite-sample FDR control and achieves higher power than CoxMKF when mediators are non-Gaussian or moderately to highly correlated. In real data applications, CoxMDS identified eight CpG sites that may mediate the effect of smoking on lung cancer survival in The Cancer Genome Atlas (TCGA) cohort, and two CpG sites that may mediate the effect of smoking on time to Alzheimer's disease conversion in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort.

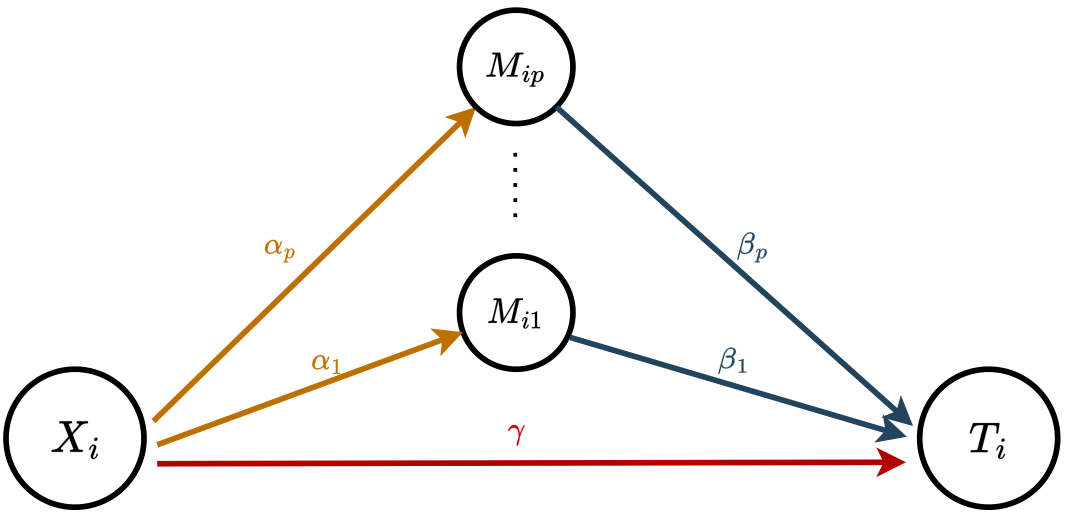

Figure 1: Causal diagram of multiple mediators related with the exposure variable X_i and the survival outcome T_i.

2 Methodology

2.1 High-dimensional mediation model with survival outcomes

Suppose we have an i.i.d. sample of size n. For subject i ($i = 1, \ldots, n$), let D_i denote the true event time, C_i the censoring time, and $T_i = \min(D_i, C_i)$ the observed survival time with event indicator $\Delta_i = I(D_i \leq C_i)$. We aim to investigate the causal mechanism from an exposure X_i to the survival outcome T_i through a p-dimensional mediator vector $\boldsymbol{M}_i = (M_{i1}, \ldots, M_{ip})^\top$, as illustrated in Figure 1.

Specifically, we consider the following two models for the mediators and survival outcome [3, 4, 27]:

$$M_{ij} = c_j + \alpha_j X_i + \boldsymbol{\vartheta}^\top \boldsymbol{Z}_i + \varepsilon_{ij}, \quad j \in [p], \tag{1}$$

$$\lambda_i(t|X_i, \boldsymbol{M}_i, \boldsymbol{Z}_i) = \lambda_0(t) \exp\left\{\gamma X_i + \boldsymbol{\theta}^\top \boldsymbol{Z}_i + \boldsymbol{\beta}^\top \boldsymbol{M}_i\right\}, \tag{2}$$

where $[p]$ denotes the set of $\{1, 2, \cdots, p\}$, $\boldsymbol{Z}_i = (Z_{i1}, \cdots, Z_{iq})^\top$ is a q-dimensional baseline covariate vector (e.g., sex, age), ε_{ij} is the error term, and $\lambda_i(t|X_i, \boldsymbol{M}_i, \boldsymbol{Z}_i)$ and $\lambda_0(t)$ are the Cox proportional hazards model and an unspecified baseline hazard function, respectively. In model (1), let $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)^\top$ denote the regression parameter vector relating the exposure to the mediators, and let $\boldsymbol{c} = (c_1, \cdots, c_p)^\top$ denote the vector of intercepts. In model (2), γ is the direct effect of the exposure on the outcome, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top$ is the effect of the mediators on the survival outcome adjusting for the effect of the exposure. $\boldsymbol{\vartheta}$ and $\boldsymbol{\theta}$ are the regression coefficients relating the covariates to the mediators and to the survival outcome, respectively. Let $H_{1,\alpha} = \{j \in [p] : \alpha_j \neq 0\}$ denote the set of mediators with non-zero exposure-mediator associations, and let $H_{1,\beta} = \{j \in [p] : \beta_j \neq 0\}$ denote the set of mediators with non-zero mediator-outcome associations. Then, $H_1 = H_{1,\alpha} \cap H_{1,\beta}$ denotes the set of true mediators, and $H_0 = \{j \in [p] : j \notin H_1\}$ is the set of non-mediators.

Under the potential outcomes framework [29, 30], we can decompose the effect of the exposure on the survival outcome using the difference in counterfactual log hazards [3, 31]. Let $T(x, \boldsymbol{m})$ denote the survival time when the exposure is set to x and the vector of mediators is set to $\boldsymbol{m}$, $\boldsymbol{M}(x) = (M_1(x), \cdots, M_p(x))^\top$ denote the vector of mediators when the exposure is set to x, $\boldsymbol{Z}$ denote the baseline covariates, and x^* denote the reference level of the exposure. The total effect (TE) of the exposure on the outcome, the natural direct effect (NDE), and the natural indirect effect (NIE) are defined respectively as follows [3, 31]:

$$\begin{aligned}
\text{TE} &= \log \lambda(T(x, \boldsymbol{M}(x)); t|\boldsymbol{Z}) - \log \lambda(T(x^*, \boldsymbol{M}(x^*)); t|\boldsymbol{Z}), \\
\text{NDE} &= \log \lambda(T(x, \boldsymbol{M}(x^*)); t|\boldsymbol{Z}) - \log \lambda(T(x^*, \boldsymbol{M}(x^*)); t|\boldsymbol{Z}), \\
\text{NIE} &= \log \lambda(T(x, \boldsymbol{M}(x)); t|\boldsymbol{Z}) - \log \lambda(T(x, \boldsymbol{M}(x^*)); t|\boldsymbol{Z}).
\end{aligned}$$

According to the definitions, we can decompose the total effect as $\text{TE} = \text{NDE} + \text{NIE}$.

As noted in VanderWeele [3], we require the following four standard assumptions for the identification of NDE and NIE in models (1) and (2):

(A1) $T(x, \boldsymbol{m}) \perp X|\boldsymbol{Z}$: there is no unmeasured confounding between the exposure and the survival outcome.

(A2) $T(x, \boldsymbol{m}) \perp M_j|X, \boldsymbol{Z}$ for $j \in [p]$: there is no unmeasured confounding between mediators and the survival outcome.

(A3) $X \perp M_j|\boldsymbol{Z}$ for $j \in [p]$: there is no unmeasured confounding between the exposure and mediators.

(A4) $T(x, \boldsymbol{m}) \perp M_j(x^*)|\boldsymbol{Z}$ for $j \in [p]$: there is no exposure-induced confounding between the mediators and the survival outcome.

If assumptions (A1) - (A4) hold, then we have the following approximation for the counterfactual log hazard [3, 27, 31]:

$$\log\lambda(T(x, \boldsymbol{M}(x^*)); t|\boldsymbol{Z}) \approx \log \lambda_0(t) + \boldsymbol{\theta}^\top \boldsymbol{Z} + \frac{1}{2}\sigma^2_{W_\beta} + \sum_{j=1}^{p} \beta_j(c_j + \boldsymbol{\vartheta}^\top \boldsymbol{Z}) + \gamma x + \sum_{j=1}^{p} \alpha_j \beta_j x^*, \tag{3}$$

where $\sigma^2_{W_\beta} = \sum_{j=1}^{p} \beta_j^2 \sigma^2_{M_j}$, and $\sigma^2_{M_j}$ is the variance of the jth mediator. Therefore, in models (1) and (2), we have the following expressions of NDE and NIE, respectively:

$$\text{NDE} = \gamma(x - x^*); \quad \text{NIE}_j = \alpha_j \beta_j (x - x^*).$$

2.2 CoxMDS Framework

In high-dimensional mediation analysis, the number of candidate mediators p is often much larger than the sample size n, so the traditional Cox proportional hazards model (2) cannot be directly applied.

Recently, Tian et al. [27] proposed a framework, "CoxMKF," which integrates the Benjamini–Hochberg (BH) procedure [21] with the aggregation of multiple knockoffs [32] to achieve finite-sample FDR control in high-dimensional mediation analysis with survival outcomes. However, the knockoff filter requires exact knowledge of the joint distribution of mediators. When mediators are moderately to highly correlated or follow non-Gaussian distributions, the knockoff filter may suffer from reduced power in identifying true mediators [28]. To overcome this limitation, we propose "CoxMDS," a novel framework that adapts the idea of multiple data splitting [28]. CoxMDS accommodates correlated or non-Gaussian mediators while maintaining finite-sample FDR control in high-dimensional mediation analysis with survival outcomes, as illustrated in Figure 2.

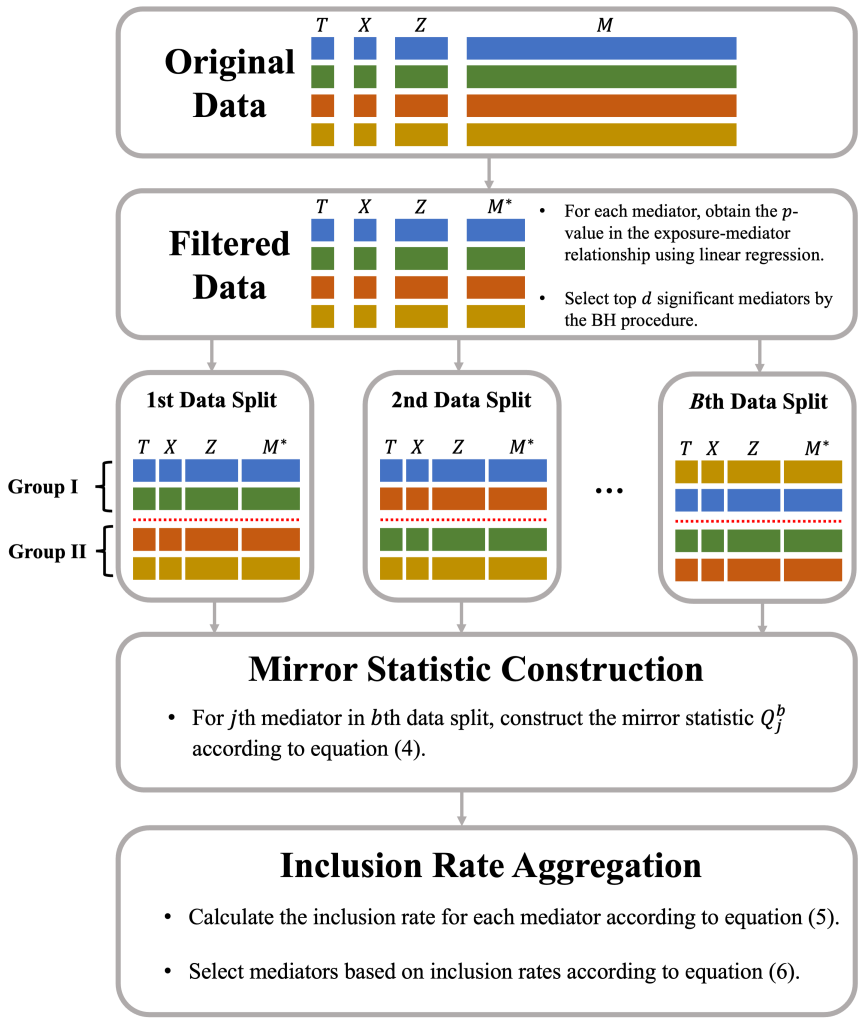

Figure 2: **Framework of CoxMDS.** First, we apply the BH procedure to the exposure–mediator associations and identify d candidate mediators. Second, we randomly split the data into two subsets B times and compute the mirror statistic for each candidate mediator within each split. Third, we aggregate the mirror statistics across the B splits using inclusion rate aggregation to select mediators under the pre-specified FDR level.

First, we select mediators that are strongly associated with the exposure. Specifically, we obtain the p-values in the $X \to M_j$ ($j \in [p]$) relationships by fitting linear regression models, and then apply the Benjamini-Hochberg (BH) procedure [21] to select mediators with significant exposure-mediator

associations by controlling the FDR at level q_1. After this step, the dimension of candidate mediators is reduced from p to d. Without loss of generality, we assume this filtering step preserves the first d mediators and denote the vector of filtered mediators as $\boldsymbol{M}_i^* = (M_{i1}, \cdots, M_{id})^\top$.

Second, we randomly split the samples into two groups for B times. Denote $\boldsymbol{T} = (T_1, \cdots, T_n)^\top$ as the vector of observed survival outcomes, $\boldsymbol{X} = (X_1, \cdots, X_n)^\top$ as the vector of exposures, $\boldsymbol{M}^* = (\boldsymbol{M}_1^*, \cdots, \boldsymbol{M}_n^*)^\top$ as the design matrix of filtered mediators, and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n)^\top$ as the design matrix of baseline covariates. In the bth data split, we randomly split the n samples into two groups, Group I and Group II, denoted as $(\boldsymbol{T}^{1,b}, \boldsymbol{X}^{1,b}, \boldsymbol{M}^{*1,b}, \boldsymbol{Z}^{1,b})$ and $(\boldsymbol{T}^{2,b}, \boldsymbol{X}^{2,b}, \boldsymbol{M}^{*2,b}, \boldsymbol{Z}^{2,b})$. In Group I, we fit a Cox proportional hazards model with the MCP and obtain the coefficients $\widehat{\boldsymbol{\beta}}^{1,b} = (\widehat{\beta}_1^{1,b}, \cdots, \widehat{\beta}_d^{1,b})^\top$, and we denote the set of mediators with non-zero coefficients as $\widehat{S}^b$. In Group II, we fit a Cox proportional hazards model using only mediators in $\widehat{S}^b$ to obtain the corresponding coefficients $\widehat{\beta}_j^{2,b}$; for mediator $j \notin \widehat{S}^b$, we set $\widehat{\beta}_j^{2,b} = 0$. Therefore, we can obtain the vector of coefficients $\widehat{\boldsymbol{\beta}}^{2,b} = (\widehat{\beta}_1^{2,b}, \cdots, \widehat{\beta}_d^{2,b})^\top$ in Group II. We then combine the two vectors of coefficients $\widehat{\boldsymbol{\beta}}^{1,b}$ and $\widehat{\boldsymbol{\beta}}^{2,b}$ by constructing the mirror statistics $\boldsymbol{Q}^b = (Q_1^b, \cdots, Q_d^b)^\top$ as

$$Q_j^b = \text{sign}(\widehat{\beta}_j^{1,b} \times \widehat{\beta}_j^{2,b}) \cdot (|\widehat{\beta}_j^{1,b}| + |\widehat{\beta}_j^{2,b}|), \quad j \in [d]. \tag{4}$$

Third, we select mediators by aggregating the mirror statistics over B data splits. According to Dai et al. [28], if the jth mediator-outcome association $\beta_j = 0$, then the mirror statistic Q_j^b is symmetric about 0; otherwise, Q_j^b is likely to be a positive and relatively large value. Therefore, if a mediator has large and positive mirror statistics across B data splits, then this mediator is more likely to have a non-zero mediator-outcome association. According to Meinshausen et al. [33] and Dai et al. [28], the following two procedures can be used to aggregate the mirror statistics over B data splits and then select the mediators with non-zero mediator-outcome associations.

The first aggregation procedure calculates an importance score, named inclusion rate [28], for each mediator over B data splits, and selects mediators with large inclusion rates as true mediators, thus we refer to this procedure as the "inclusion rate aggregation". Specifically, in the bth data split, if we select mediators whose mirror statistics are no less than a positive threshold $\tau > 0$, then an estimate of the false discovery proportion (FDP) is given by

$$\widehat{\text{FDP}}^b(\tau) = \frac{\#\{j \in [d] : Q_j^b \leq -\tau\}}{\#\{j \in [d] : Q_j^b \geq \tau\} \vee 1},$$

where $a_1 \vee a_2 = \max(a_1, a_2)$ for two real numbers a_1 and a_2. For a target FDR control level $q_2 \in (0, 1)$, we then choose the data-dependent threshold $\tau_{q_2}^b$ as

$$\tau_{q_2}^b = \min\{\tau > 0 : \widehat{\text{FDP}}^b(\tau) \leq q_2\},$$

and the selection set of mediators in the bth data split is $\widehat{H}_1^b = \{j \in [d] : Q_j^b \geq \tau_{q_2}^b\}$. Then we calculate

the inclusion rate for each mediator $j \in [d]$ as

$$V_j = \frac{1}{B} \sum_{b=1}^{B} \frac{I(j \in \widehat{H}_1^b)}{\#\widehat{H}_1^b \vee 1}, \quad j \in [d], \tag{5}$$

where $I(\cdot)$ is the indicator function such that $I(A) = 1$ if event A happens and $I(A) = 0$ if event A does not happen. From equation (5), V_j measures the importance of the jth mediator across B data splits. For example, if the jth mediator is selected in each data split, and each data split selects 10 mediators, then $V_j = 1/10$. Therefore, a large value of V_j indicates that the jth mediator is more likely to have a non-zero mediator-outcome association. We sort the inclusion rates such that $0 \leq V_{(1)} \leq V_{(2)} \leq \cdots \leq V_{(d)}$, and find the largest index $l \in [d]$ such that $V_{(1)} + \cdots + V_{(l)} \leq q_2$. Then, the inclusion rate aggregation procedure selects the mediators as

$$\widehat{H}_1^{\mathrm{IRA}} = \{j \in [d] : V_j > V_{(l)}\}. \tag{6}$$

The second aggregation procedure is referred to as the "quantile aggregation", which is adopted in CoxMKF [27]. Based on the mirror statistic Q_j^b in the bth data split, the quantile aggregation calculates the following intermediate statistic δ_j^b for mediator j:

$$\delta_j^b = \begin{cases} \dfrac{1 + \#\{k \in [d] : Q_k^b \leq -Q_j^b\}}{d}, & Q_j^b > 0; \\ 1, & Q_j^b \leq 0. \end{cases}$$

A small value of δ_j^b indicates that the jth mediator has a strong mediator-outcome association in the bth data split. Following Meinshausen et al. [33] and Tian et al. [27], we calculate the Multiple Cox Statistic (MCS) for each mediator:

$$\bar{\delta}_j = \min\left\{1, \frac{\mathcal{Q}_\eta(\{\delta_j^b : b \in [B]\})}{\eta}\right\}, \quad j \in [d],$$

where η is the pre-specified quantile point with a recommended value $\eta = 0.05$ [27], and $\mathcal{Q}_\eta(\cdot)$ denotes the η-quantile function. The MCS $\bar{\delta}_j$ is another importance measurement that is different from the inclusion rate V_j. Specifically, a small value of $\bar{\delta}_j$ indicates that the jth mediator is likely to have a non-zero mediator-outcome association across B data splits. We then sort the MCSs such that $0 \leq \bar{\delta}_{(1)} \leq \bar{\delta}_{(2)} \leq \cdots \leq \bar{\delta}_{(d)}$, and apply the BH procedure to choose a data-dependent threshold to control the FDR at level q_2:

$$\widehat{k}_{\mathrm{BH}} = \max\left\{k \in [d] : \bar{\delta}_{(k)} \leq q_2 \frac{k}{d}\right\}.$$

Then, the quantile aggregation procedure selects the mediators as

$$\widehat{H}_1^{\mathrm{QA}} = \{j \in [d] : \bar{\delta}_j \leq \bar{\delta}_{(\widehat{k}_{\mathrm{BH}})}\}.$$

We compare inclusion rate aggregation and quantile aggregation with extensive simulations in Section 3.3. Empirically, we find that the inclusion rate aggregation in CoxMDS maintains finite-sample FDR control even for highly correlated mediators, whereas the quantile aggregation in CoxMDS

might have inflated FDR when mediators are highly correlated. Accordingly, we recommend using CoxMDS with inclusion rate aggregation as the default.

3 Simulation Studies

In this section, we present extensive simulation studies to assess the finite-sample performance of the proposed CoxMDS framework. In particular, we investigate the following aspects:

1. the impact of the sample-split proportion and the number of splits;
2. the performance of CoxMDS under correlated mediators;
3. the performance of CoxMDS under non-Gaussian mediators.

Each simulation setting is based on 500 replications. In the second and third settings, we compare the performance of CoxMDS with that of CoxMKF. We set $q_1 = 0.2$ in the screening step and $q_2 = 0.1$ as the target FDR level. Let $\widehat{H}_1$ denote the selection set of mediators. We then evaluate CoxMDS and CoxMKF in terms of the false discovery rate (FDR) and the true positive proportion (TPP):

$$\mathrm{FDR} = \mathbb{E}\Big[\frac{\#\{j \in \widehat{H}_1 \cap H_0\}}{\#\widehat{H}_1}\Big], \quad \mathrm{TPP} = \mathbb{E}\Big[\frac{\#\{j \in \widehat{H}_1 \cap H_1\}}{\#H_1}\Big].$$

We generate the simulated data as follows. The binary exposure X_i is generated from a Bernoulli distribution with success probability 0.6 as Ber(0.6), and the direct effect of the exposure on the outcome is set to $\gamma = 0.5$. The covariates are $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2})^\top$, where Z_{i1} is a binary covariate generated from a Bernoulli distribution with success probability 0.3 as Ber(0.3), and Z_{i2} is a continuous covariate generated from a uniform distribution as $\mathrm{U}(0,1)$. The coefficients of the covariates on the outcome are $\boldsymbol{\theta} = (0.3, -0.2)^\top$. As for the mediators, in the first and second simulation settings, mediators are generated according to model (1), where the coefficients $\boldsymbol{\vartheta} = (0.3, -0.2)^\top$, the error terms follow a multivariate normal distribution with a Toeplitz covariance matrix whose (i,j)-th entry is $\rho^{|i-j|}$, and $\rho \in [0,1)$ is used to control the correlations among mediators. In the third simulation setting, mediators are generated from beta distributions, as described in Section 3.3. The true event time D_i is generated from the exponential model $\lambda_i(t|X_i, \boldsymbol{Z}_i, \boldsymbol{M}_i) = 0.5\exp\left\{\gamma X_i + \boldsymbol{\theta}^\top \boldsymbol{Z}_i + \boldsymbol{\beta}^\top \boldsymbol{M}_i\right\}$. The censoring time C_i is generated from an exponential distribution with parameter c_0, and the average censoring rate is controlled at 30% or 60% by properly choosing c_0. The observed survival time T_i is generated as $T_i = \min(D_i, C_i)$. For $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we fix the first ten elements of $\boldsymbol{\alpha}$ to be $\kappa_\alpha \cdot (1,1,1,1,-1,-1,-1,-1,1,-1)$, the first twelve elements of $\boldsymbol{\beta}$ to be $\kappa_\beta \cdot (1,1,1,1,1,1,1,1,0,0,1,-1)$, where κ_α and κ_β are two scaling factors to prevent the TPP from being too close to 1. The rest elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are set to 0. Therefore, the set of true mediators is $H_1 = \{1, 2, \cdots, 8\}$.

3.1 Choice of the sample-split proportion and the number of data splits

We evaluate the performance of CoxMDS under a range of sample-split proportions and numbers of data splits. The proportion of samples in Group I ranges from 10% to 90%, and the number of data splits is chosen from $\{1, 2, 5, 10, 25, 50\}$. In this simulation setting, we adopt the inclusion rate aggregation and set the target FDR level at $q_2 = 0.1$. The error terms in model (1) are drawn from a multivariate normal distribution with a Toeplitz covariance matrix, where (i, j)-th entry is $\rho^{|i-j|}$ and $\rho = 0.6$. The sample size is set to $n = 500$, and the number of candidate mediators is set to $p = 2,000$.

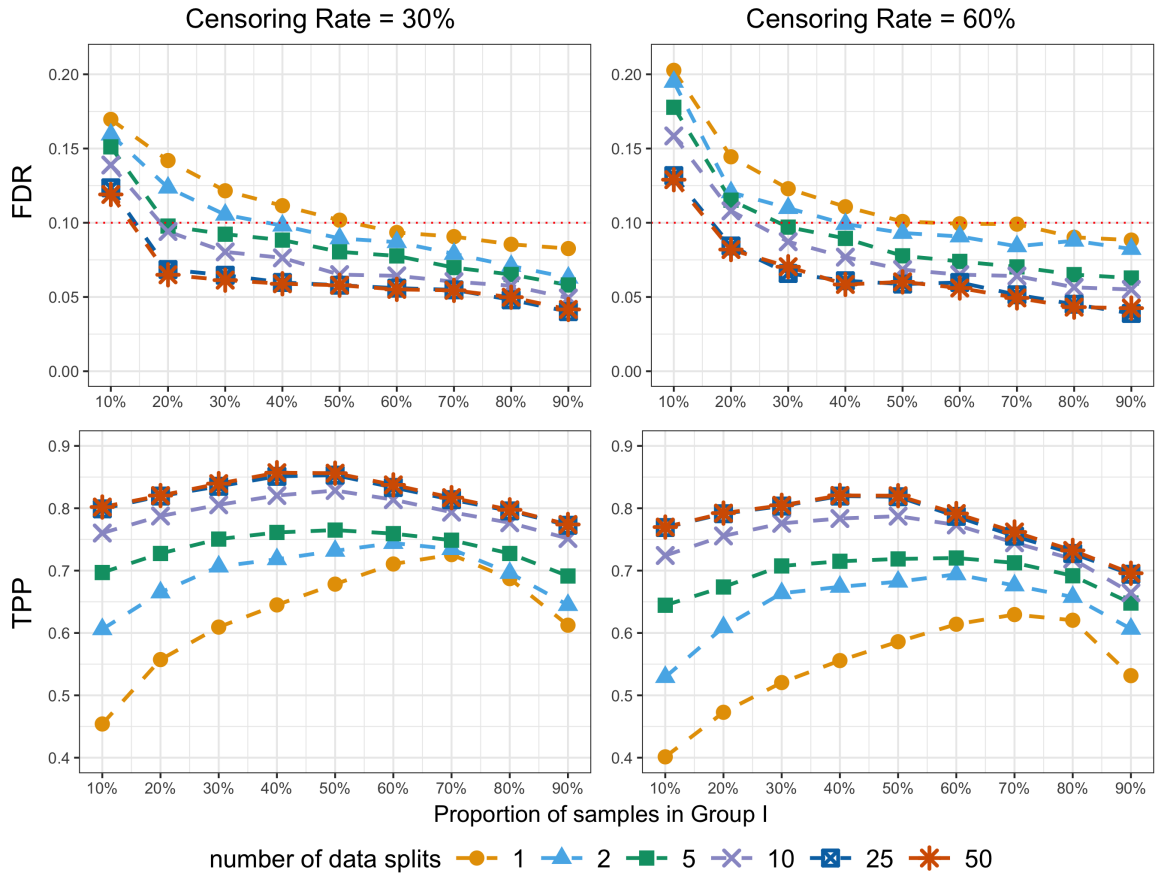

Figure 3: Line plot of the FDR and TPP of CoxMDS against different sample-split proportions and different numbers of data splits. We adopt the inclusion rate aggregation in this simulation. The red dashed line represents the target FDR level $q_2 = 0.1$.

As shown in Figure 3, CoxMDS achieves finite-sample FDR control and maintains high TPP with a balanced sample-split proportion and a large number of data splits. For example, with the number of splits $B = 25$ and a 30% censoring rate, a half-half sample-split proportion yields finite-sample FDR control and the highest TPP (0.853) among all sample-split proportions. Moreover, increasing the number of sample splits B from 25 to 50 does not yield much power gain. Balancing these trade-offs, we find that a half-half sample-split proportion combined with $B = 25$ data splits achieves finite-sample FDR control, maintains high power, and remains computationally efficient, making it our recommended default for practical applications.

3.2 Comparison of CoxMDS and CoxMKF in correlated mediators

In this simulation, we compare the empirical performance of CoxMDS with CoxMKF under different aggregation procedures (inclusion rate aggregation and quantile aggregation) across different values of correlations among mediators, and the results are presented in Figure 4. Specifically, mediators are generated from model (1), where the error terms are drawn from $N(0, \Sigma)$ with $\Sigma = (\Sigma_{ij})_{p\times p} = (\rho^{|i-j|})_{p\times p}$. We vary the correlation $\rho \in \{0, 0.2, 0.4, 0.6, 0.8\}$, and fix the sample size $n = 500$ and number of candidate mediators $p = 2,000$. To avoid the power from approaching 1, the scaling factors $(\kappa_\alpha, \kappa_\beta)$ are set to $(0.50, 0.35)$ in this simulation setting. The default aggregation procedure of CoxMKF is quantile aggregation, but we also implement CoxMKF with inclusion rate aggregation for fair comparison. For quantile aggregation procedure, we adopt the recommended quantile point $\eta = 0.05$ [27]. We set the target FDR level at $q_2 = 0.1$ for both CoxMDS and CoxMKF.

In this simulation setting, CoxMDS with inclusion rate aggregation not only achieves finite-sample FDR control but also demonstrates greater power than CoxMKF under high mediator correlation. For example, with a censoring rate of 30% and $\rho = 0.8$, the TPP of CoxMDS with inclusion rate aggregation reaches 0.810, compared with 0.651 and 0.710 for CoxMKF under inclusion rate aggregation and quantile aggregation, respectively. The quantile aggregation might result in an inflated FDR when mediators are highly correlated, and thus is not recommended for CoxMDS. These results highlight CoxMDS with inclusion rate aggregation as the preferred method in settings with highly correlated mediators, offering both finite-sample FDR control and higher power. Accordingly, we use inclusion rate aggregation as the default aggregation procedure of CoxMDS in subsequent simulation settings and real data analyses.

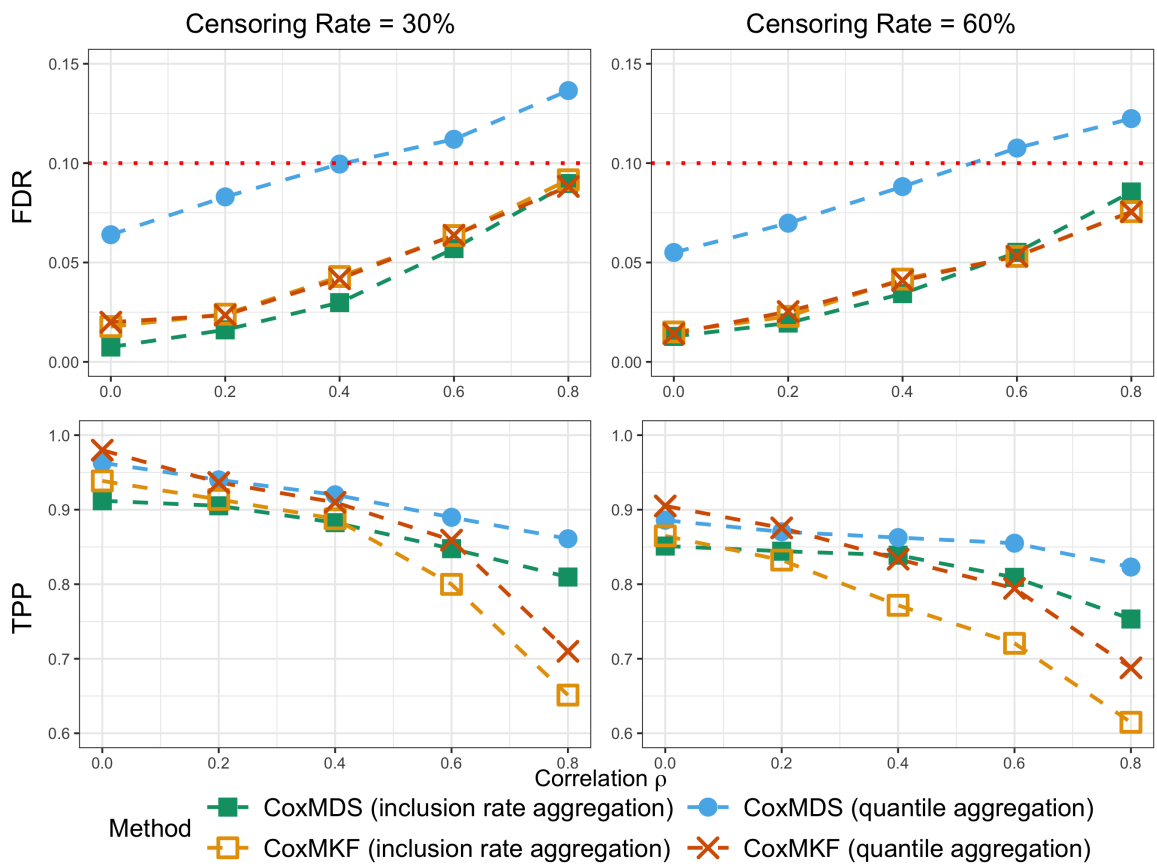

Figure 4: Line plot of the FDR and TPP for CoxMKF and CoxMDS with different values of correlation $\rho \in \{0, 0.2, 0.4, 0.6, 0.8\}$ and different censoring rates (30% or 60%). The numbers of multiple knockoffs and data splits are both $B = 25$. We adopt a half-half sample-split proportion in CoxMDS. The red dashed line represents the target FDR level $q_2 = 0.1$.

3.3 Comparison of CoxMDS and CoxMKF in non-Gaussian mediators

In this simulation setting, we investigate the performance of CoxMDS and CoxMKF when the joint distribution of mediators is non-Gaussian. Specifically, we generate the jth mediator M_{ij} from a beta distribution $\text{Beta}(a_{ij}, b_{ij})$, with $a_{ij} = \phi_j \cdot \mu_{ij}$ and $b_{ij} = \phi_j \cdot (1 - \mu_{ij})$, where $\mu_{ij} = 1/(1 + \exp(-(c_j + \alpha_j X_i + \boldsymbol{\vartheta}^\top \boldsymbol{Z}_i))$ is the mean of M_{ij}, and $\phi_j > 0$ is the precision parameter that controls the variance of M_{ij} and is drawn from an exponential distribution $\text{Exp}(1)$. We set the sample size $n = 500$, the number of mediators $p = 2,000$, the target FDR level $q_2 = 0.1$, and the scaling factors $(\kappa_\alpha, \kappa_\beta) = (0.90, 0.80)$. We present the FDR and TPP of CoxMDS and CoxMKF with inclusion rate aggregation procedure in Table 1.

CoxMDS maintains finite-sample FDR control and achieves higher TPP than CoxMKF in this non-Gaussian setting. For instance, with a 60% censoring rate, the TPP of CoxMDS reaches 0.813, compared with 0.762 for CoxMKF. This performance gap likely arises because CoxMKF relies on knockoffs generated from a multivariate Gaussian distribution, which reduces its power when the candidate mediators are non-Gaussian. In contrast, CoxMDS does not impose distributional assumptions

on the mediators, making it more powerful than CoxMKF in non-Gaussian settings.

Table 1: FDR and TPP for CoxMKF and CoxMDS with different censoring rates (30% or 60%) when mediators are generated from beta distributions. The numbers of multiple knockoffs and data splits are both $B = 25$. The half-half sample-split proportion is adopted in CoxMDS. The target FDR level is set to $q_2 = 0.1$.

	Censoring Rate=30%		Censoring Rate=60%	
	FDR	TPP	FDR	TPP
CoxMDS	0.076	0.871	0.089	0.813
CoxMKF	0.042	0.832	0.056	0.762

4 Real Data Applications

4.1 Identifying CpG Sites Mediating the Effect of Smoking on Lung Cancer Survival

Lung cancer is the most common cancer among men and the third most common among women, with approximately 85% of cases classified as non-small cell lung cancer (NSCLC) [34]. Improving patient survival requires both effective therapies and early diagnosis [10]. Tobacco smoking, the major risk factor for lung cancer, has been shown to induce widespread alterations in DNA methylation [35]. In this section, we apply the proposed CoxMDS method to the TCGA lung cancer dataset [36], which includes patients with lung squamous cell carcinoma and lung adenocarcinoma. Our goal is to identify CpG sites that mediate the effect of smoking on lung cancer survival.

In this dataset, DNA methylation profiles were measured using the Illumina Infinium HumanMethylation450 BeadChip array, and methylation levels were summarized as β-values [36]. The exposure is defined as smoking status (smoker vs. non-smoker) at initial diagnosis. The true survival time is the number of days from diagnosis to death, and the censoring time is the number of days from diagnosis to last follow-up. We include age, sex, cancer stage, and radiotherapy status as covariates. Following Tian et al. [27], we excluded patients with missing survival time, smoking status, or covariates, yielding a final dataset of 754 patients aged 33–90 years with measurements on 365,306 CpG sites. The median survival time is 1,632 days, with a censoring rate of 60%. To explore the distributional properties of the DNA methylation data, we randomly selected four CpG sites and plotted their Pearson correlation heatmap and density curves (Figure 5). These plots reveal that CpG sites are highly correlated and often exhibit non-Gaussian distributions, underscoring the suitability of CoxMDS, which is designed to accommodate both correlated and non-Gaussian mediators.

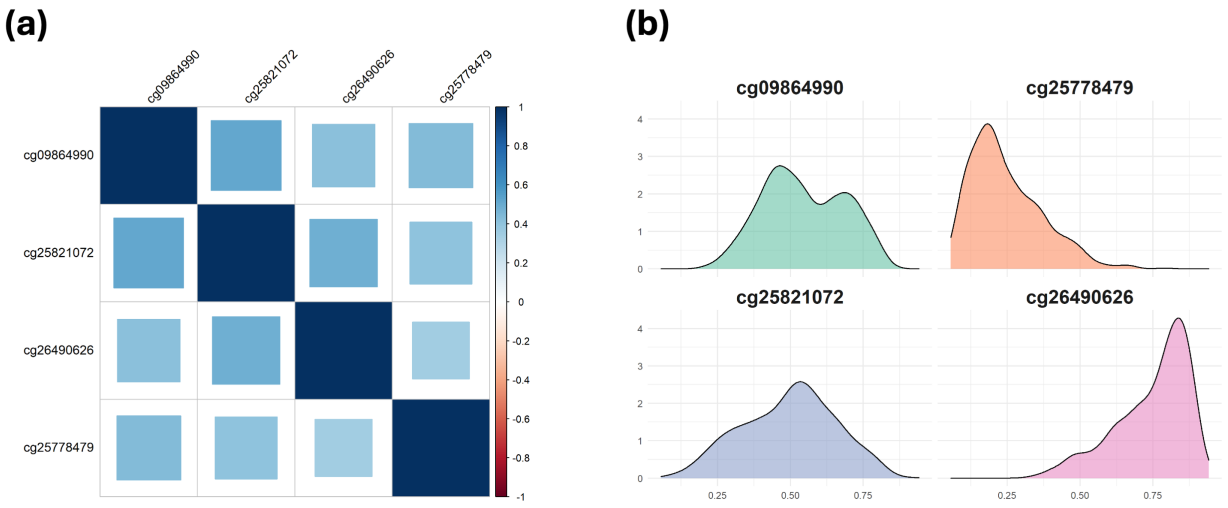

Figure 5: (a) Pearson correlation heatmap and (b) density plots of four DNA methylation CpG sites in the TCGA lung cancer cohort study.

For CoxMDS, we set $q_1 = 5 \times 10^{-4}$ as the threshold in the BH procedure for candidate mediator filtering, $q_2 = 0.1$ as the target FDR level, and $B = 25$ as the number of data splits. CoxMDS identifies eight DNA methylation CpG sites, and the effect estimates and corresponding genes are presented in Table 2. All four CpGs previously reported by CoxMKF [27] are also identified by CoxMDS, while four additional CpGs (cg01023787, cg06320150, cg12823072, and cg15953187) are uniquely identified by CoxMDS. This finding is consistent with our simulation results, where CoxMDS demonstrates greater power in detecting correlated and non-Gaussian mediators.

Among the newly identified CpGs, cg01023787 ($\widehat{\text{NIE}} = 0.031$) lies in the gene SLC6A18 on chromosome 5. SLC6A18 is located within the 5p15.33 chromosomal region that is frequently gained in early-stage NSCLC, suggesting this locus may participate in early genetic changes associated with the disease [37]. The CpG site cg06320150 also exhibits a positive $\widehat{\text{NIE}} = 0.025$ and is associated with the expression of C22orf34, a gene reported to show downregulated expression in lung adenocarcinoma compared with adjacent normal tissue [38]. The CpG site cg12823072 ($\widehat{\text{NIE}} = -0.120$) is located in SCARA5 on chromosome 8. SCARA5 acts as a tumor suppressor in NSCLC, with promoter methylation linked to cancer progression [39]. Finally, cg15953187 is located in MUC5B on chromosome 11, the same gene as cg07690349, which is detected by both CoxMDS and CoxMKF. MUC5B is reported as a favorable prognostic biomarker in NSCLC with EGFR mutations and has diagnostic and prognostic significance in lung adenocarcinoma when combined with TTF-1 expression [40]. These DNA methylation CpG sites may provide valuable insights for researchers in designing interventions to improve treatment strategies for lung cancer patients.

Table 2: The identified DNA methylation CpG sites in the TCGA dataset and their estimated effects obtained using CoxMDS with $q_1 = 5 \times 10^{-4}$, $q_2 = 0.1$, and $B = 25$. Here, CHR denotes the chromosome.

CpG	Gene Name	CHR	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\text{NIE}}$
cg01023787	SLC6A18	5	-0.070	-0.441	0.031
cg06320150	-	22	-0.058	-0.440	0.025
cg07690349	MUC5B	11	-0.075	1.259	-0.094
cg12823072	SCARA5	8	-0.076	1.575	-0.120
cg15953187	MUC5B	11	-0.061	0.546	-0.033
cg21926276	H19	11	-0.058	-2.949	0.172
cg24129177	-	12	-0.056	-1.581	0.089
cg24200525	SBF1	12	-0.024	5.204	-0.126

4.2 Identifying CpG Sites Mediating the Effect of Smoking on Time-to-Conversion to Alzheimer's Disease

In this section, we apply CoxMDS and CoxMKF to identify CpG sites that might mediate the effect of smoking on time-to-conversion to Alzheimer's diseases in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (`https://adni.loni.usc.edu/`). The ADNI study investigates Alzheimer's disease progression through neuroimaging, biomarker measurements, and clinical and cognitive assessments [41]. In this analysis, DNA methylation profiles are measured by the Illumina Infinium HumanMethylationEPIC BeadChip array, and DNA methylation levels are summarized as M-values [42]. The exposure is defined as any history of smoking (smoker/non-smoker) during the subject's lifetime. The true survival time is the number of years from recruitment to Alzheimer's disease diagnosis, and the censoring time is the number of years from recruitment to the last follow-up date. Since the DNA methylation data in the ADNI study are derived from blood samples [42, 43], we estimate the relative proportion of underlying cell types using the R package `ENmix` [44], and include the estimated cell type proportions as covariates. We also adjust for the following covariates: APOE-ε4 allele count (a known genetic risk factor for Alzheimer's disease [45]), age, years of education, gender, and history of hypertension.

After data preprocessing and quality control [43], we include a total of 611 subjects aged 55–91 years, with 865,859 DNA methylation CpG sites. Of these, 404 subjects are from the ADNI2 phase, while the remaining 207 are from the ADNI GO phase. The censoring rate (i.e., the proportion of subjects without diagnosis of Alzheimer's disease during the follow-up period) is 66.1%, and the median of survival outcome is 5.10 years. In Figure 6, we present the descriptive statistics of these subjects included in the analysis.

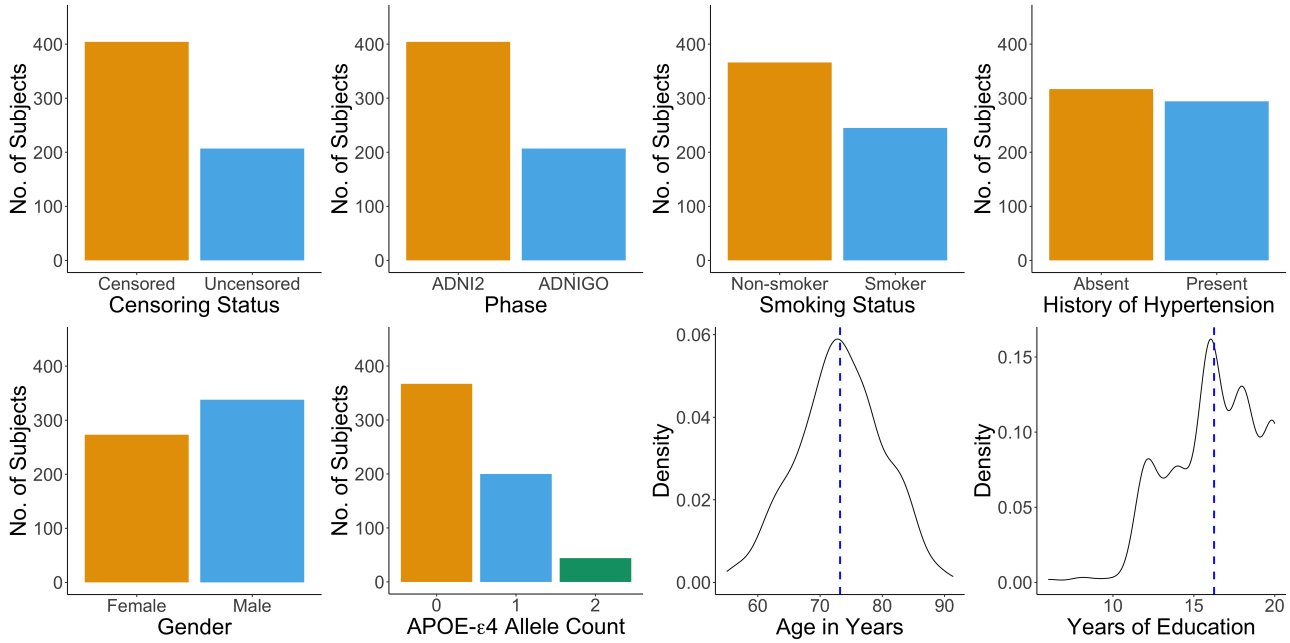

Figure 6: Descriptive statistics of the subjects included in the analysis. Among the 611 subjects, the censoring rate is 66.1%; the proportion of smokers is 40.1%; the proportion of subjects with history of hypertension is 48.1%; the proportion of male subjects is 55.3%; the average APOE-ε4 allele count is 0.471; the average age is 73.2 years; and the average years of education is 16.2 years.

For both CoxMDS and CoxMKF, we set $q_1 = 0.05$ as the threshold for candidate mediator filtering, $q_2 = 0.1$ as the target FDR level, and $B = 25$ as the number of data splits and multiple knockoffs. CoxMDS identifies two DNA methylation CpG sites (cg06644428 and cg01940273) as mediators, while CoxMKF detects none. This finding aligns with our simulations, which show that CoxMDS achieves higher power than CoxMKF when candidate mediators are highly correlated and exhibit non-Gaussian distributions in finite samples.

We report the two CpG sites identified by CoxMDS in Table 3. Among them, cg06644428 ($\widehat{\text{NIE}} = 0.070$) is a known smoking-associated CpG site located in the intergenic region 2q37.1 on chromosome 2 [46]. The CpG site cg01940273 ($\widehat{\text{NIE}} = -0.038$), located in gene ALPPL2 on chromosome 2, is also reported as a smoking-associated DNA methylation CpG site [47]. ALPPL2 encodes an alkaline phosphatase protein, and altered plasma levels of alkaline phosphatase are associated with central nervous system injury and Alzheimer's disease progression [48]. Therefore, these DNA methylation CpG sites may serve as potential targets for future research aimed at understanding disease mechanisms and developing interventions for Alzheimer's disease.

Table 3: The identified DNA methylation CpG sites in the ADNI dataset and their estimated effects using CoxMDS with pre-specified $q_1 = 0.05$, $q_2 = 0.1$, and $B = 25$. CHR stands for the chromosome.

CpG	Gene Name	CHR	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\text{NIE}}$
cg06644428	-	2	-0.391	-0.180	0.070
cg01940273	ALPPL2	2	-0.118	0.319	-0.038

5 Conclusion

In this paper, we propose CoxMDS, a novel mediation analysis method to address the challenge of controlling finite-sample FDR in high-dimensional mediator settings with a survival outcome. By integrating multiple data splitting with high-dimensional mediation analysis, CoxMDS achieves both finite-sample FDR control and high detection power, even when mediators exhibit medium-to-high correlations or follow non-Gaussian distributions. Extensive simulation studies demonstrate that CoxMDS achieves finite-sample FDR control and high detection power even when mediators are highly correlated or non-Gaussian. In real data applications, CoxMDS identifies eight DNA methylation CpG sites that might mediate the effect of smoking on lung cancer survival in the TCGA cohort, and two CpG sites that might mediate the effect of smoking on time-to-conversion to Alzheimer's disease in the ADNI study.

CoxMDS offers two key advantages over competing methods. First, compared with CoxMKF, CoxMDS maintains finite-sample FDR control while achieving higher power in scenarios with correlated and/or non-Gaussian mediators, making it especially suitable for genome-wide DNA methylation data where co-methylation and non-Gaussian distributions are common. Second, CoxMDS is computationally more efficient than CoxMKF, as it avoids generating knockoff copies of mediators and instead uses random data splitting, which is faster and provides practical scalability for high-dimensional mediation analyses. For example, on the TCGA dataset, the average running times of CoxMDS and CoxMKF are 26.25 minutes and 27.77 minutes, respectively, on a server equipped with an Intel Xeon Silver 4116 CPU and 64 GB RAM. In CoxMDS, most of the computation time is spent in the screening step (25.89 minutes), while the data-splitting step requires only 0.36 minutes, making it more efficient than knockoff generation (1.88 minutes). To facilitate its use, we also provide a user-friendly R package, `CoxMDS`.

CoxMDS opens several avenues for future development. First, we will extend the framework to accommodate accelerated failure time models. Second, we plan to generalize the design from a single exposure–outcome pair to settings with multiple, potentially correlated exposures. Third, incorporating time-varying mediators with repeated measurements will allow longitudinal mediation analyses.

Key Points

- DNA methylation data often display non-Gaussian distributions and strong correlations, creating challenges for mediator selection and finite-sample false discovery rate (FDR) control.
- We propose CoxMDS, a multiple-data-splitting approach with Cox proportional hazards models for mediator selection in high-dimensional mediation analysis with survival outcomes.

- CoxMDS guarantees finite-sample FDR control and demonstrates higher power than competing methods with highly correlated or non-Gaussian mediators.
- This framework can be broadly applied to other types of omics data, including RNA-seq, proteomics, and metabolomics.

Code Availability

The CoxMDS method is implemented in an open source R package, which is freely available at https://github.com/MinhaoYaooo/CoxMDS.

Data Availability

The lung cancer dataset is obtained from The Cancer Genome Atlas (TCGA) program and is publicly available through the UCSC Xena Browser at https://xenabrowser.net/datapages/. The Alzheimer's disease dataset is obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and is available upon application at https://adni.loni.usc.edu/.

References

[1] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986.

[2] Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309, 2010.

[3] Tyler J VanderWeele. Causal mediation analysis with survival data. *Epidemiology*, 22(4): 582—585, July 2011. ISSN 1044-3983.

[4] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

[5] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31(2):142–147, 2013.

[6] NT Ventham, Nicholas A Kennedy, AT Adams, Rahul Kalla, Simon Heath, KR O'leary, H Drummond, DC Wilson, et al. Integrative epigenome-wide analysis demonstrates that DNA methylation

may mediate genetic risk in inflammatory bowel disease. *Nature Communications*, 7(1):13507, 2016.

[7] Chengwen Luo, Botao Fa, Yuting Yan, Yang Wang, Yiwang Zhou, Yue Zhang, and Zhangsheng Yu. High-dimensional mediation analysis in survival models. *PLoS Computational Biology*, 16 (4):e1007768, 2020.

[8] Ruiling Fang, Haitao Yang, Yuzhao Gao, Hongyan Cao, Ellen L Goode, and Yuehua Cui. Gene-based mediation analysis in epigenetic studies. *Briefings in Bioinformatics*, 22(3):bbaa113, 2021.

[9] Luis Alameda, Zhonghua Liu, Pak C Sham, Monica Aas, Giulia Trotta, Victoria Rodriguez, Marta Di Forti, Simona A Stilo, Radhika Kandaswamy, Celso Arango, et al. Exploring the mediation of DNA methylation across the epigenome between childhood adversity and first episode of psychosis—findings from the EU-GEI study. *Molecular Psychiatry*, pages 1–12, 2023.

[10] Lisa D Moore, Thuc Le, and Guoping Fan. DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, 2013.

[11] Hao Wu and Yi Zhang. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell*, 156(1-2):45–68, 1 2014. doi: 10.1016/j.cell.2013.12.019.

[12] Theresa K Kelly, Daniel D De Carvalho, and Peter A Jones. Epigenetic modifications as therapeutic targets. *Nature Biotechnology*, 28(10):1069–1078, 2010.

[13] Yuanyuan Li and Trygve O Tollefsbol. Impact on DNA methylation in cancer prevention and therapy by bioactive dietary components. *Current Medicinal Chemistry*, 17(20):2141–2151, 2010.

[14] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1):587, 2010.

[15] Joanna Zhuang, Martin Widschwendter, and Andrew E Teschendorff. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics*, 13(1):59, 2012.

[16] Georgina Mansell, Tyler J Gorrie-Stone, Yanchun Bao, Meena Kumari, Leonard S Schalkwyk, Jonathan Mill, and Eilis Hannon. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics*, 20(1):366, 2019.

[17] E Andres Houseman, Brock C Christensen, Ru-Fang Yeh, Carmen J Marsit, Margaret R Karagas, Margaret Wrensch, Heather H Nelson, Joseph Wiemels, Shichun Zheng, John K Wiencke, et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for

high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*, 9(1):365, 2008.

[18] Christoph Bock. Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–719, 2012.

[19] Leonie Weinhold, Simone Wahl, Sonali Pechlivanis, Per Hoffmann, and Matthias Schmid. A statistical model for the analysis of beta values in DNA methylation studies. *BMC Bioinformatics*, 17(1):480, 2016.

[20] Tiphaine C Martin, Idil Yet, Pei-Chien Tsai, and Jordana T Bell. coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns. *BMC Bioinformatics*, 16(1):131, 2015.

[21] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[22] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[23] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[24] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[25] Michael E Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13:290–312, 1982.

[26] Haixiang Zhang, Yinan Zheng, Lifang Hou, Cheng Zheng, and Lei Liu. Mediation analysis for survival data with high-dimensional mediators. *Bioinformatics*, 2021.

[27] Peixin Tian, Minhao Yao, Tao Huang, and Zhonghua Liu. CoxMKF: a knockoff filter for high-dimensional mediation analysis with a survival outcome in epigenetic studies. *Bioinformatics*, 38 (23):5229–5235, 2022.

[28] Chenguang Dai, Buyu Lin, Xin Xing, and Jun S Liu. False discovery rate control via data splitting. *Journal of the American Statistical Association*, pages 1–18, 2022.

[29] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

[30] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

[31] Yen-Tsung Huang and Hwai-I Yang. Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology (Cambridge, Mass.)*, 28(3):370, 2017.

[32] Tuan-Binh Nguyen, Jérôme-Alexis Chevalier, Bertrand Thirion, and Sylvain Arlot. Aggregation of multiple knockoffs. In *International Conference on Machine Learning*, pages 7283–7293. PMLR, 2020.

[33] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[34] Amanda Leiter, Rajwanth R Veluswamy, and Juan P Wisnivesky. The global burden of lung cancer: current status and future trends. *Nature Reviews Clinical Oncology*, 20(9):624–639, 2023.

[35] Stephen S Hecht and Dorothy K Hatsukami. Smokeless tobacco and cigarette smoking: chemical mechanisms and cancer prevention. *Nature Reviews Cancer*, 22(3):143–155, 2022.

[36] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[37] Mohammad Al Zeyadi, Ivanka Dimova, Vladislav Ranchich, Blaga Rukova, Desislava Nesheva, Zora Hamude, Sevdalin Georgiev, Danail Petrov, and Draga Toncheva. Whole genome microarray analysis in non-small cell lung cancer. *Biotechnology & Biotechnological Equipment*, 29(1):111–118, 2015.

[38] Juan Li, Ping Li, Wei Zhao, Rui Yang, Shanshan Chen, Yong Bai, Shaozhi Dun, Xiaonan Chen, Yuwen Du, Yuanyuan Wang, et al. Expression of long non-coding RNA DLX6-AS1 in lung adenocarcinoma. *Cancer Cell International*, 15(1):1–8, 2015.

[39] Francisco Exposito, Miriam Redrado, Diego Serrano, Silvia Calabuig-Fariñas, Aida Bao-Caamano, Sandra Gallach, Eloisa Jantus-Lewintre, Angel Diaz-Lagares, Aitor Rodriguez-Casanova, Juan Sandoval, et al. G9a/DNMT1 co-targeting inhibits non-small cell lung cancer growth and reprograms tumor cells to respond to cancer-drugs through SCARA5 and AOX1. *Cell Death & Disease*, 15(11):787, 2024.

[40] Kouki Wakata, Tomoshi Tsuchiya, Koichi Tomoshige, Katsunori Takagi, Naoya Yamasaki, Keitaro Matsumoto, Takuro Miyazaki, Atsushi Nanashima, Jeffrey A Whitsett, Yutaka Maeda,

et al. A favourable prognostic marker for EGFR mutant non-small cell lung cancer: immunohistochemical analysis of MUC5B. *BMJ Open*, 5(7):e008366, 2015.

[41] Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, CR Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease Neuroimaging Initiative (ADNI) clinical characterization. *Neurology*, 74(3): 201–209, 2010.

[42] Aparna Vasanthakumar, Justin W Davis, Kenneth Idler, Jeffrey F Waring, Elizabeth Asque, Bridget Riley-Gillis, Shaun Grosskurth, Gyan Srivastava, Sungeun Kim, Kwangsik Nho, et al. Harnessing peripheral DNA methylation differences in the Alzheimer's Disease Neuroimaging Initiative (ADNI) to reveal novel biomarkers of disease. *Clinical Epigenetics*, 12(1):84, 2020.

[43] Tiago C. Silva, Juan I Young, Lanyu Zhang, Lissette Gomez, Michael A Schmidt, Achintya Varma, X Steven Chen, Eden R Martin, and Lily Wang. Cross-tissue analysis of blood and brain epigenome-wide association studies in Alzheimer's disease. *Nature Communications*, 13(1):4852, 2022.

[44] Zongli Xu, Liang Niu, Leping Li, and Jack A Taylor. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Research*, 44(3):e20–e20, 2016.

[45] DE Schmechel, AM Saunders, WJ Strittmatter, Bo J Crain, CM Hulette, SH Joo, MA Pericak-Vance, D Goldgaber, and AD Roses. Increased amyloid beta-peptide deposition in cerebral cortex as a consequence of apolipoprotein E genotype in late-onset Alzheimer disease. *Proceedings of the National Academy of Sciences*, 90(20):9649–9653, 1993.

[46] Pei-Chien Tsai, Craig A Glastonbury, Melissa N Eliot, Sailalitha Bollepalli, Idil Yet, Juan E Castillo-Fernandez, Elena Carnero-Montoro, Thomas Hardiman, Tiphaine C Martin, Alice Vickers, et al. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. *Clinical Epigenetics*, 10:1–21, 2018.

[47] Yi-Qian Sun, Rebecca C Richmond, Matthew Suderman, Josine L Min, Thomas Battram, Arnar Flatberg, Vidar Beisvag, Therese Haugdahl Nøst, Florence Guida, Lin Jiang, et al. Assessing the role of genome-wide DNA methylation between smoking and risk of lung cancer using repeated measurements: the HUNT study. *International Journal of Epidemiology*, 50(5):1482–1497, 2021.

[48] Pooja Dua, Hye Suk Kang, Seung-Mo Hong, Ming-Sound Tsao, Soyoun Kim, and Dong-ki Lee. Alkaline phosphatase ALPPL-2 is a novel pancreatic carcinoma-associated protein. *Cancer Research*, 73(6):1934–1945, 2013.