Briefings in Bioinformatics, 2025, **26(3)**, bbaf206
https://doi.org/10.1093/bib/bbaf206
Case Study

High-dimensional mediation analysis for longitudinal mediators and survival outcomes

Lili Liu[1], Haixiang Zhang[2], Yinan Zheng[3], Tao Gao[3], Cheng Zheng[4], Kai Zhang[5], Lifang Hou[3], Lei Liu[1,*]

[1]Center for Biostatistics and Data Science, Washington University in St. Louis, 660 S. Euclid Ave, St. Louis, MO 63110, United States
[2]Center for Applied Mathematics, Tianjin University, No. 92 Weijin Road, Nankai District, Tianjin 300072, China
[3]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, 680 N. Lake Shore Drive, Chicago, IL 60611, United States
[4]Department of Biostatistics, University of Nebraska Medical Center, 42nd and Emile Streets, Omaha, NE 68198, United States
[5]Department of Population and Community Health, College of Public Health, The University of North Texas Health Science Center at Fort Worth, 3500 Camp Bowie Blvd, Fort Worth, TX 76107, United States

*Corresponding author. E-mail: lei.liu@wustl.edu

Abstract

Mediation analysis with high-dimensional mediators is crucial for identifying epigenetic pathways linking environmental exposures to health outcomes. However, high-dimensional mediation analysis methods for longitudinal mediators and a survival outcome remain underdeveloped. This study fills that gap by introducing a method that captures mediation effects over time using multivariate, longitudinally measured time-varying mediators. Our approach uses a longitudinal mixed effects model to examine the relationship between the exposure and the mediating process. We connect the mediating process to the survival outcome using a Cox proportional hazards model with time-varying mediators. To handle high-dimensional data, we first employ a mediation-based sure independence screening method for dimension reduction. A Lasso inference procedure is further utilized to identify significant time-varying mediators. We adopt a joint significance test to accurately control the family wise error rate in testing high-dimensional mediation hypotheses. Simulation studies and an analysis of the Coronary Artery Risk Development in Young Adults Study demonstrate the utility and validity of our method.

Keywords: high-dimensional mediation analysis; variable selection; longitudinal data; survival outcome

Introduction

Mediation analysis, as an important topic in causal inference, has been used to explore the mechanisms through which intermediate variables mediate the relationship between an exposure and an outcome. Originally developed in psychology research [1–3], it has since been extended to fields such as epigenomics [4] and microbiome studies [5, 6]. In recent years, substantial research efforts have been devoted to developing methodology for mediation analysis, especially with the rise of high-dimensional data. In these contexts, the sparsity assumption is often applied, implying that only a few mediators have significant effects. Therefore, selecting the relevant mediators is essential for estimating and inferring mediation effects. Zhang *et al.* [7] and Gao *et al.* [8] proposed innovative methods on testing mediation effects in high dimensional epigenetic studies. Additionally, Zhou *et al.* [9] presented an inference procedure for indirect effects in high-dimensional linear mediation models, while Zhang *et al.* [10] and Luo *et al.* [11] proposed a multiple-testing procedure for high-dimensional mediation effects in survival outcomes. However, these methods typically focus on time-invariant mediators. In many studies, mediators are measured repeatedly over time, as in the longitudinal DNA methylation (DNAm) data. When repeatedly measured assessments of mediators are available, a straightforward approach to simplifying mediation analysis is to aggregate the longitudinal mediators into a single one. However, this simplification often fails to capture the full complexity of the mediation process, leading to a weakening of the indirect effect [12]. There is a dearth of suitable high-dimensional mediation models for longitudinal mediators and a survival outcome.

Our motivating example stems from the DNAm research conducted within the Coronary Artery Risk Development in Young Adults (CARDIA) Study, an ongoing longitudinal cohort examining the development and determinants of clinical and subclinical cardiovascular disease (CVD) and their risk factors. CVD remains the leading cause of death and disability in the USA and worldwide, primarily due to the progressive decline in cardiovascular health over the lifespan. In the CARDIA study, epigenome-wide DNAm profiling was conducted repeatedly over time, providing a rich source of longitudinal epigenetic data.

Social determinants of health (SDH) have long been recognized as major contributors to CVD [13, 14]. Both individual (e.g. education, household income, occupation, financial strain) and parental factors (education, occupation) have been associated with CVD risk [15, 16]. Our research aims to investigate the mediating effects of longitudinal DNAm markers on the relationship between the development of CVD and the individual social determinants of

Received: November 14, 2024. **Revised:** March 17, 2025. **Accepted:** April 07, 2025

health (iSDH), which is calculated from individual level factors including personal, parental, and childhood family environments.

In this study, we propose novel statistical techniques within the context of high-dimensional mediation analysis (HIMA), focusing on longitudinal DNAm mediators and a survival outcome. Our approach uses a mixed-effects model to investigate the relationship between the exposure—iSDH and the longitudinal mediators. We connect the mediating process to the survival outcome using a Cox proportional hazards model with time-varying mediators. Our methodology is structured into a three-step approach. First, a screening process reduces the number of mediators in the model from a very large set to a manageable size. Second, we select important time-varying DNAm markers predictive of disease risk through Lasso [17] in a Cox model. Third, we develop a joint significance test for mediation effects, adjusting for multiplicity using Bonferroni's method.

The remainder of this paper is structured as follows: In Section Model and estimation, we introduce our model and propose a three-step testing procedure for mediation analysis. Section Simulation evaluates the performance of our approach through Monte Carlo simulation studies. In Section Application, we demonstrate the proposed method using the CARDIA study as a practical example. Finally, we summarize our methodology and discuss potential future directions in Section Discussion.

Model and estimation

Let X_i denote the exposure (e.g. iSDH) for individual i, $i = 1, \cdots, n$; and let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iq})^\top$ represent the baseline covariates to adjust for. Define $M_{ik}(t)$ as the k-th mediator (e.g. DNAm at the kth CpG site) of individual i measured at time point t, $k = 1, \cdots, p$ and $t = t_0, \cdots, t_T$. $M_{ik}(t)$ changes over time during the follow-up period. The outcome is a survival endpoint, where the observed follow-up time Y_i is the minimum of the censoring time C_i and the failure event time $\tilde{T}_i$. The censoring indicator is given by $\delta_i = I(\tilde{T}_i \leq C_i)$.

For the longitudinal methylation data in the CARDIA study, we use the following mixed-effects model to examine the association of DNAm at each CpG site and the exposure to iSDH:

$$M_{ik}(t) = a_{ik} + \alpha_k X_i + \boldsymbol{\eta}_k^\top \mathbf{Z}_i + e_{ik}(t), \quad k = 1, \ldots, p \tag{2.1}$$

where α_k is the parameter relating the exposure to the kth mediator; $\boldsymbol{\eta}_k$ is the vector of regression coefficients for the covariates. Each DNAm marker has its own random intercept, denoted by a_{ik}, at the DNAm marker level. We assume that $a_{ik} \sim N(0, \sigma_k^2)$ is independent for all i's and k's. The error term $e_{ik}(t) \sim N(0, \sigma_e^2)$ is independent of random effects a_{ik}'s. This model captures the hierarchical structure of the data, which includes two levels: level 1 is the repeated measures for each DNAm marker, and level 2 is the subject level. In the mediators screening step described below, we fit model (2.1) separately for each DNAm marker, leading to the fitting of p two-level models.

To evaluate how DNAm biomarker trajectories influence CVD development, we link $M_{ik}(t)$ with CVD outcomes using a survival model. Specifically, we define a discrete mediator process as follows: Let $t_0 = 0 < t_1 \ldots < t_j < \ldots < t_T$ be an increasing sequence of time points. Define the mediator process $M_{ik}(t)$ as follows: $M_{ik}(t) = M_{ikj}$ for $t_j \leq t < t_{j+1}$ with $j = 0, 1, \ldots, T-1$. Define $\overline{\mathbf{M}}_{ij} = \{M_{ikj'} : k = 1, \ldots, p, j' \leq j\}$ to be the mediator history up to time t_j [18]. The mediators are only defined while the individual is at risk ($Y_i > t$); otherwise they are undefined. We take a survival outcome and apply a high-dimensional Cox model with time-dependent covariates to assess the association between the survival outcome and time-varying mediators:

$$\lambda_i(t|\overline{\mathbf{M}}_{ij}) = \lambda_0(t) \exp\left\{\theta_1 X_i + \boldsymbol{\theta}_2^\top \mathbf{Z}_i + \sum_{k=1}^{p} \beta_k M_{ik}(t)\right\}, \tag{2.2}$$

where $\lambda_i(t)$ is the hazard for the survival endpoint for subject i; $\lambda_0(t)$ is an unspecified baseline hazard function; θ_1 is the direct effect of the exposure on the survival outcome; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the regression parameter vector relating the mediators to the survival outcome adjusting for the effect of the exposure and covariates; $\boldsymbol{\theta}_2$ is the vector of regression coefficients for the covariates.

To draw causal conclusions, assumptions about the absence of confounders must be made. We assume the stable unit treatment value assumption (SUTVA) holds [19]. Let $\mathbf{m} = (m_1(t_0), \ldots, m_p(t_T))$ represent the vector of potential mediator values over time, and let $Y(x, \mathbf{m})$ denote the potential survival time when the exposure is set to x and the mediators to $\mathbf{m}$. $M_k(x, t)$ denotes the observed value for mediator k measured at time point t when the exposure is set to x. Here, we assume that individual mediator process is not causally related to each other. We would like to point out that this assumption does not imply that all mediators are independent given the exposure X and baseline adjusted covariates $\mathbf{Z}$, and it allows for potential unmeasured common causes (whether induced by the exposure or not) between mediators. In addition to the assumptions of positivity and consistency [20], the following assumptions regarding potential confounding [4] will allow us to identify both the joint causal mediation effect and path-specific causal effects in the framework above:

(C1) $X \perp Y(x, \mathbf{m})|\mathbf{Z}$, i.e. no unmeasured confounders between the exposure and the survival outcome;
(C2) For any k, $\{M_k(x, t), t \in [t_0, t_T]\} \perp Y(x, \mathbf{m})|X, \mathbf{Z}$, i.e. no unmeasured confounders between the mediators and the survival outcome;
(C3) For any k, $X \perp \{M_k(x, t), t \in [t_0, t_T]\}|\mathbf{Z}$, i.e. no unmeasured confounders between the exposure and the mediators;
(C4) For any k, $\{M_k(x^*, t), t \in [t_0, t_T]\} \perp Y(x, \mathbf{m})|\mathbf{Z}$, i.e. no exposure-induced confounding between the mediators and the survival outcome, where x^* is the intervention for the exposure X with different value than x.

Figure 1 illustrates the causal mediation pathway of high-dimensional mediators $M_k(t)$, exposure (X), and time-to-event outcome (Y). The path-specific causal effect on the log-hazard difference scale for the mediator $M_k(t)$ ($X \to M_k(t) \to Y$) can be defined as a comparison of log hazard for $Y(x, M_1(x, t), \ldots, M_k(x^*, t), \ldots, M_p(x, t))$ and $Y(x, M_1(x, t), \ldots, M_k(x, t), \ldots, M_p(x, t))$, which can be approximated as $\alpha_k \beta_k (x^* - x)$ [10]. The approximation holds under the rare event assumption [21]. It is important to emphasize that even if the rare event approximation does not hold, testing the null hypothesis $\alpha_k \beta_k = 0$ remains valid for evaluating the existence of a path-specific causal effect through M_k. This approximation and its assumptions have been rigorously discussed in [10].

Our aim is to estimate and test the null hypothesis $\alpha_k \beta_k = 0$ to determine the presence of path-specific causal effect through M_k with $k = 1, \ldots, p$. Denote by $S_0 = \{k : \alpha_k \beta_k \neq 0\}$ the index set of significant mediators. The proposed approach is as follows:

Step 1: Mediators screening. We reduce the dimension of mediators from a very large to a moderate scale (below the sample size). Motivated by the sure independence screening (SIS) [22, 23], we consider a series of models with the single mediator, as shown

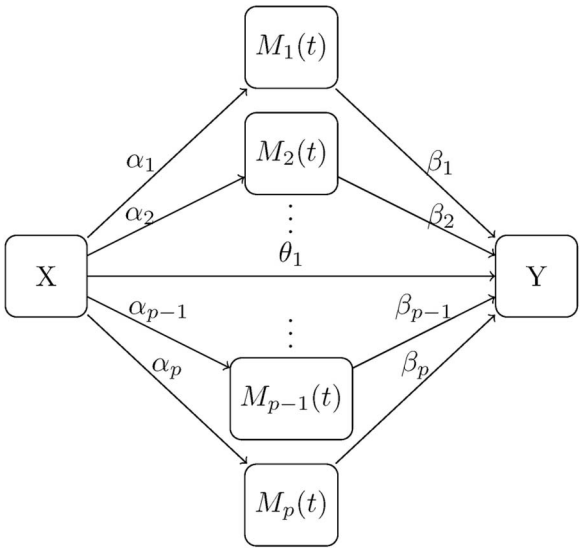

Figure 1. The causal mediation pathway of high-dimensional mediators $M_k(t)$, exposure X, and time-to-event outcome Y.

in equations (2.1) and (2.3):

$$\lambda_i(t|\overline{\mathbf{M}}_{ij}) = \lambda_0(t)\exp\left\{\theta_1 X_i + \boldsymbol{\theta}_2^\top \mathbf{Z}_i + \beta_k M_{ik}(t)\right\}. \quad (2.3)$$

We select a subset $\mathcal{D} = \{k : M_k \text{ is among the top } d = [n/2\log(n)] \text{ largest effect } |\hat{\alpha}_k\hat{\beta}_k|, \text{ for } k = 1, \cdots, p\}$, where $\hat{\alpha}_k$ and $\hat{\beta}_k$ are the estimates based on the above models (2.1) and (2.3), respectively.

Step 2: Variable selection. We consider the following submodel based on the selected set $\mathcal{D}$,

$$\lambda_i(t|\overline{\mathbf{M}}_{ij}) = \lambda_0(t)\exp\left\{\theta_1 X_i + \boldsymbol{\theta}_2^\top \mathbf{Z}_i + \sum_{k\in\mathcal{D}} \beta_k M_{ik}(t)\right\}. \quad (2.4)$$

In Step 2, we apply Lasso to further select important DNAm markers predictive of disease risk. This simplifies the model to a conventional regression with a small number of covariates. The partial log-likelihood of Model (2.4) is defined as

$$\ell(\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{\Theta_i - \log\left[\sum_{i'=1}^{n} I(Y_{i'} \geq Y_i)\exp(\Theta_i)\right]\right\}, \quad (2.5)$$

where $\Theta_i = \theta_1 X_i + \boldsymbol{\theta}_2^\top \mathbf{Z}_i + \sum_{k\in\mathcal{D}} \beta_k M_{ik}(Y_i)$. The penalized Cox partial log-likelihood is defined as

$$\Omega(\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\beta})_\lambda = -\ell(\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}),$$

where $P_\lambda(\boldsymbol{\beta})$ is the Lasso penalty function, and $\lambda \geq 0$ is a tuning parameter. The parameter estimates are obtained by minimizing $\Omega(\theta_1, \boldsymbol{\theta}_2, \boldsymbol{\beta})_\lambda$. The penalized Cox proportional hazards model is especially challenging when covariates vary over follow-up time (i.e. the mediators are time-dependent). The above procedure can be conveniently implemented using R function *pcoxtime* [24], which uses proximal gradient descent algorithm for fitting penalized Cox models.

Step 3: Joint significance test. We develop a joint significance test for mediation effects. We use Bonferroni's method to adjust for multiplicity. The indirect effect for the kth DNAm marker can be given by $\alpha_k\beta_k$. We thus need to test the significance of the indirect effect as

$$H_{0k} : \alpha_k\beta_k = 0 \quad \text{vs} \quad H_{1k} : \alpha_k\beta_k \neq 0.$$

We can use the joint significant test [7] to test the above hypothesis. Let $\mathcal{S} = \{k : \hat{\beta}_k \neq 0\}$, which is based on the lasso-penalized estimate in Step 2. The raw p-value for testing $H_0 : \beta_k = 0$ is given by $P_{\beta_k} = 2\{1 - \Phi(|\hat{\beta}_k|/\hat{\sigma}_{\beta_k})\}$, where $k \in \mathcal{S}$, $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$, and $\hat{\sigma}_{\beta_k}$ is the estimate of standard error for $\hat{\beta}_k$. Since the lasso penalty can produce biased estimates, we conduct a refitting step to obtain accurate parameter estimates by maximizing the log partial likelihood with the selected variables. This procedure can be conveniently implemented using R function *coxph*, which provides $\hat{\beta}_k$, $\hat{\sigma}_{\beta_k}$ and P_{β_k}. To control the family wise error rate (FWER), we use Bonferroni's method to adjust for multiple comparisons. Thus, the corrected p-value is given by $\tilde{P}_{\beta_k} = \min(P_{\beta_k}|\mathcal{S}|, 1)$, where $k \in \mathcal{S}$ and $|\mathcal{S}|$ is the cardinality, i.e. the number of elements in set $\mathcal{S}$.

Similarly, the raw p-value for testing $H_0 : \alpha_k = 0$ is $P_{\alpha_k} = 2\{1 - \Phi(|\hat{\alpha}_k|/\hat{\sigma}_{\alpha_k})\}$, where $k \in \mathcal{S}$, $\hat{\alpha}_k$ is the estimate of α_k, and $\hat{\sigma}_{\alpha_k}$ is the corresponding estimated standard error. The Bonferroni corrected p-value is $\tilde{P}_{\alpha_k} = \min(P_{\alpha_k}|\mathcal{S}|, 1)$. We will reject the null hypothesis of no mediation effect with M_k only if both α_k and β_k are significant. The Bonferroni corrected p-value for the joint significance test is defined as

$$P_{\max,k} = \max(\tilde{P}_{\alpha_k}, \tilde{P}_{\beta_k}). \quad (2.6)$$

If $P_{\max,k} < 0.05$, we can conclude that there exists significant mediation effect for mediator k.

Simulation

In this section, we conduct simulation studies in four examples to assess our proposed procedure.

Example 1 We generate data from models (2.1) and (2.2). The first nine elements of $\boldsymbol{\beta}$ are $(0.5, 0.4, 0.3, 0, 0, 0, 0.5, 0.4, 0.3)^\top$, and the first nine elements of $\boldsymbol{\alpha}$ are $(0.5, 0.4, 0.3, 0.5, 0.4, 0.3, 0, 0, 0)^\top$. The rest of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are all 0. Thus, the first three mediators M_1, M_2 and M_3 are active (significant) mediators. The exposure X_i is generated from a normal distribution $N(0, 0.5^2)$. The covariate $\mathbf{z}_i = (z_{i1}, z_{i2})^\top$ is generated from a multivariate normal distribution with mean 0, variance 1 and an exchangeable correlation $\rho = 0.3$. The coefficient $\boldsymbol{\eta}_k = (0.4, 0.4)^\top$, $\theta_1 = 0.4$ and $\boldsymbol{\theta}_2 = (0.4, 0.4)^\top$. The time-dependent covariates $M_k(t)$ are generated based on equation $M_{ik}(t) = \alpha_k X_i + \boldsymbol{\eta}_k^\top \mathbf{Z}_i + b_i + a_{ik} + e_{ik}(t), \quad k = 1, \ldots, p$. A subject-level random effect b_i is generated from $N(0, 0.2^2)$; the DNAm marker level random effect a_{ik} is generated from $N(0, 0.2^2)$; $e_{ik}(t)$ is generated from $N(0, 1)$. The simulation of the survival data with time-dependent covariates $M_k(t)$ extends that of [25] from dichotomous time-dependent covariates to continuous time-dependent covariates. We consider two time intervals R1 = [0, 0.2), R2 = [0.2, ∞), where the time-dependent covariates are constant within each interval but can vary between intervals. The mediators are measured at time points 0 and 0.2. The baseline

Table 1. The biases and MSEs of the estimated mediation effects in Example 1

	p = 5000				p = 10 000			
	n = 300		n = 600		n = 300		n = 600	
$\alpha_k\beta_k$	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\alpha_1\beta_1$	0.032	0.006	0.023	0.002	0.040	0.007	0.025	0.003
$\alpha_2\beta_2$	0.041	0.007	0.016	0.001	0.053	0.009	0.017	0.001
$\alpha_3\beta_3$	0.048	0.006	0.005	0.001	0.049	0.006	0.016	0.002
$\alpha_4\beta_4$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 2. The frequency of active mediators being kept after the screening step over 100 repetitions in Example 1

	p = 5000		p = 10 000	
M_k	n = 300	n = 600	n = 300	n = 600
M_1	100	100	100	100
M_2	100	100	99	100
M_3	93	100	90	100

hazard is generated from the Weibull distribution with the scale and shape parameters of 1 and 4, respectively. We consider administrative censoring with $C_i = 1.2$. As a result, the censoring rate is approximately 40%, with a similar number of events observed in the Application. We consider the number of repeated measures for each patient $n_i = 1$ or 2. The simulation is performed under four settings: two sample sizes ($n = 300$ and $n = 600$) and two dimensions of mediators ($p = 5000$ and $p = 10\,000$). All simulation results are obtained via 100 replicates.

We also compare our procedure with the naive joint significance test, where each mediator is tested individually for significance, and p-values are adjusted to account for multiple testing. Specifically, Bonferroni's correction is applied based on the total number of DNAm markers p, controlling the FWER by dividing the nominal significance level (e.g. 0.05) by p. In the above setting, let $S_0 = \{1, 2, 3\}$ denote the index set of significant mediators. We define FWER $= P(\exists k \in S_0^c : P_{\max,k} < 0.05)$, where $P_{\max,k}$ is given in (2.6).

Table 1 presents the estimates and mean square errors (MSE) for the indirect effect $\alpha_k\beta_k, k = 1, \ldots, 4$. Here, we omit the results for $\{\alpha_k\beta_k\}_{k=5}^{p}$, because their performances are similar to that of $\alpha_4\beta_4$. From the results in Table 1, we can see that both biases and MSEs are very small. The estimates are close to the true values of the indirect effects and the MSEs decrease as the sample size n increases.

Table 2 presents the frequency at which active mediators are retained after the screening step, calculated over 100 repetitions for each active mediator. The retention frequency for active mediators is consistently near or at 100%, indicating the effectiveness of our screening method. These results highlight the robust performance of the screening process across various scenarios.

In Table 3, four criteria are used to assess the model's performance:

(i). Power: the empirical power after the multiple testing;
(ii). FWER: the family wise error rate, with a threshold level of 0.05;

Table 3. The empirical power of three active mediators, FWER, N_{TP}, N_{FP} after the multiple testing in Example 1

		p = 5000		p = 10 000	
		n = 300	n = 600	n = 300	n = 600
Proposed	P_{M_1}	0.96	1.00	0.94	1.00
	P_{M_2}	0.77	0.99	0.71	1.00
	P_{M_3}	0.35	0.93	0.34	0.83
	FWER	0.006	0.003	0.033	0.003
	N_{TP}	2.08	2.92	2.00	2.83
	N_{FP}	0.02	0.01	0.10	0.01
Naive	P_{M_1}	0.75	1.00	0.71	1.00
	P_{M_2}	0.29	0.91	0.19	0.91
	P_{M_3}	0.06	0.55	0.03	0.40
	FWER	0.000	0.000	0.000	0.000
	N_{TP}	1.10	2.46	0.93	2.31
	N_{FP}	0.00	0.00	0.00	0.00

P_{M_1}, P_{M_2}, *and* P_{M_3}: the empirical power of mediators M_1, M_2, and M_3, respectively; FWER: the family-wise error rate; N_{TP}: average number of true positives; N_{FP}: average number of false positives.

(iii). N_{TP}: the average number of true positives (i.e. the average number of relevant predictors being correctly selected);
(iv). N_{FP}: the average number of false positives (i.e. the average number of irrelevant predictors being incorrectly selected).

From Table 3, we observe that the N_{TP} values for our method are close to 3, while the N_{FP} values remain very small. The estimated FWER for mediation effects is below 0.05, demonstrating that our joint significance test procedure effectively controls the FWER under the threshold level. In contrast, the naive procedure is overly conservative with poor control of Type I error. Furthermore, our method shows superior power compared to the naive method.

Example 2. In this example, to simulate "spatially" shared information among CpG sites, we generated a_{ik} from a multivariate normal distribution with an autoregressive (AR) correlation structure, where $cor(a_{ik}, a_{ij}) = \rho^{|k-j|}$ with a moderate correlation coefficient $\rho = 0.5$. The mean and variance of a_{ik} are identical to those specified in Example 1. This structure captures high correlations between adjacent mediators while allowing for low correlations between distant mediators, effectively simulating spatially shared information. The time-dependent covariates $M_{ik}(t)$ are generated based on equation $M_{ik}(t) = \alpha_k X_i + \boldsymbol{\eta}_k^\top \mathbf{Z}_i + b_i + a_{ik} + e_{ik}(t)$. The other parameters, including exposure X_i, the covariate $\mathbf{z}_i$, coefficient $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}_k$, θ_1, $\boldsymbol{\theta}_2$, $e_{ik}(t)$, b_i, the number of repeated measures for each patient n_i, and the time to

Table 4. The biases and MSEs of the estimated mediation effects in Example 2

	p = 5000				p = 10 000			
	n = 300		n = 600		n = 300		n = 600	
$\alpha_k\beta_k$	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\alpha_1\beta_1$	0.043	0.006	0.037	0.003	0.058	0.009	0.036	0.003
$\alpha_2\beta_2$	0.038	0.008	0.024	0.001	0.061	0.009	0.022	0.001
$\alpha_3\beta_3$	0.054	0.006	0.012	0.001	0.058	0.006	0.020	0.002
$\alpha_4\beta_4$	0.000	0.000	0.000	0.000	0.001	2e-4	0.000	0.000

Table 5. The frequency of active mediators being kept after the screening step over 100 repetitions in Example 2

	p = 5000		p = 10 000	
M_k	n = 300	n = 600	n = 300	n = 600
M_1	100	100	100	100
M_2	99	100	98	100
M_3	95	100	91	100

event data are generated in the same way as in Example 1.

Table 4 presents the estimates and MSEs for the indirect effect $\alpha_k\beta_k, k = 1, \ldots, 4$. The results for $\{\alpha_k\beta_k\}_{k=5}^{p}$ are omitted, as their performances are similar to that of $\alpha_4\beta_4$. The findings in Table 4 demonstrate that both the biases and MSEs are small. The estimates closely approximate the true values of the indirect effects, and the MSEs exhibit a decreasing trend as the sample size n increases.

Table 5 reports the retention frequency of active mediators after the screening step, calculated over 100 repetitions for each mediator. The retention frequency for active mediators consistently approaches or reaches 100%, demonstrating the robustness and effectiveness of the proposed screening methods.

Table 6 shows the empirical power of mediators M_1, M_2, M_3, FWER, N_{TP}, and N_{FP} for the proposed and naive methods. The proposed method shows consistently higher power for detecting mediators. Our method maintains FWER < 0.05 across all scenarios, ensuring Type I error control. In contrast, the naive procedure is overly conservative with poor control of Type I error. For true positives, the proposed method identifies more mediators on average. Overall, the proposed method balances low FWER with high power.

In Examples 3 and 4, we examine scenarios where mediators interact under both strong and weak heredity assumptions. The strong heredity assumption requires that an interaction term be included in the model only if both corresponding main effects are significant, whereas the weak heredity assumption relaxes this requirement by allowing interactions when at least one main effect is significant.

Example 3. In this example, we evaluate the performance of our proposed method when mediators exhibit pairwise interactions under the strong heredity assumption. The mediators $M_{ik}(t)$ are generated using the same approach as described in Example 1. We also include the interactions of the first three significant mediators: M_1M_2, M_1M_3, and M_2M_3 in the generation of the time-to-event data. Specifically, we model the hazard function as:

Table 6. The empirical power of three active mediators, FWER, N_{TP}, N_{FP} after the multiple testing in Example 2

		p = 5000		p = 10 000	
		n = 300	n = 600	n = 300	n = 600
Proposed	P_{M_1}	0.96	1.00	0.90	1.00
	P_{M_2}	0.76	1.00	0.64	1.00
	P_{M_3}	0.30	0.88	0.27	0.77
	FWER	0.017	0.008	0.017	0.013
	N_{TP}	1.97	2.88	1.81	2.77
	N_{FP}	0.06	0.03	0.04	0.05
Naive	P_{M_1}	0.65	0.98	0.54	1.00
	P_{M_2}	0.26	0.88	0.17	0.86
	P_{M_3}	0.03	0.37	0.03	0.36
	FWER	0.010	0.003	0.010	0.006
	N_{TP}	0.94	2.23	0.74	2.22
	N_{FP}	0.02	0.01	0.01	0.02

P_{M_1}, P_{M_2}, P_{M_3}: the empirical power of mediators M_1, M_2, and M_3, respectively; FWER: the family-wise error rate; N_{TP}: average number of true positives; N_{FP}: average number of false positives.

$\lambda_i(t|\overline{\mathbf{M}}_{ij}) = \lambda_0(t)\exp\{\theta_1 X_i + \boldsymbol{\theta}_2^\top \mathbf{Z}_i + \sum_{k=1}^{p}\beta_k M_{ik}(t) + \varphi_1 M_{i1}(t)M_{i2}(t) + \varphi_2 M_{i1}(t)M_{i3}(t) + \varphi_3 M_{i2}(t)M_{i3}(t)\}$, where $\varphi_1 = \varphi_2 = \varphi_3 = 0.4$. All other parameters are generated in the same way as in Example 1. In the analysis, we ignore interactions and consider only the main effects during both screening and model fitting. This setup allows us to evaluate the robustness and effectiveness of our method in scenarios where the true data structure includes interactions that are not explicitly accounted for in the analysis.

Table 7 summarizes the estimates and MSE for the indirect effects $\alpha_k\beta_k, k = 1, \ldots, 4$. The results in Table 7 show the robust performance of our method, with consistently low biases and MSEs. Moreover, the MSEs decrease as the sample size n increases, demonstrating improved estimation accuracy with increasing sample size.

Table 8 presents the retention frequency of active (significant) mediators after the screening step, based on 100 repetitions for each mediator. The results indicate that active mediators are consistently retained with high frequencies, demonstrating the effectiveness of our method in identifying and preserving important variables at the screening step. Moreover, the retention frequencies reach 100 for $n = 600$.

Table 9 shows the empirical power of mediators M_1, M_2, M_3, FWER, N_{TP}, and N_{FP} for the proposed and naive methods. The proposed method achieves higher power in detecting important mediators while effectively controlling the FWER across all settings. It strikes an optimal balance between low FWER and high power. In contrast, the naive procedure is overly conservative.

Table 7. The biases and MSEs of the estimated mediation effects in Example 3

	p = 5000				p = 10 000			
	n = 300		n = 600		n = 300		n = 600	
$\alpha_k\beta_k$	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\alpha_1\beta_1$	0.049	0.008	0.023	0.003	0.066	0.013	0.027	0.003
$\alpha_2\beta_2$	0.030	0.006	0.008	0.001	0.038	0.008	0.012	0.001
$\alpha_3\beta_3$	0.042	0.006	0.007	0.001	0.039	0.006	0.004	0.002
$\alpha_4\beta_4$	0.000	0.000	0.000	0.000	0.000	0.000	7e-4	5e-5

Table 8. The frequency of active mediators being kept after the screening step over 100 repetitions in Example 3

	p = 5000		p = 10 000	
M_k	n = 300	n = 600	n = 300	n = 600
M_1	100.00	100.00	100.00	100.00
M_2	98.00	100.00	98.00	100.00
M_3	81.00	100.00	80.00	100.00

Table 9. The empirical power of three active mediators, FWER, N_{TP}, N_{FP} after the multiple testing in Example 3

		p = 5000		p = 10 000	
		n = 300	n = 600	n = 300	n = 600
Proposed	P_{M_1}	0.95	1.00	0.88	1.00
	P_{M_2}	0.82	0.99	0.75	0.99
	P_{M_3}	0.39	0.94	0.37	0.87
	FWER	0.022	0.008	0.022	0.009
	N_{TP}	2.16	2.93	1.98	2.86
	N_{FP}	0.06	0.03	0.05	0.04
Naive	P_{M_1}	0.38	0.94	0.37	0.94
	P_{M_2}	0.15	0.73	0.13	0.68
	P_{M_3}	0.01	0.25	0.02	0.12
	FWER	0.000	0.000	0.000	0.000
	N_{TP}	0.54	1.92	0.52	1.74
	N_{FP}	0.00	0.00	0.00	0.00

P_{M_1}, P_{M_2}, P_{M_3}: the empirical power of mediators M_1, M_2, and M_3, respectively; FWER: the family-wise error rate; N_{TP}: average number of true positives; N_{FP}: average number of false positives.

Even in the presence of interactions among important mediators, our method remains robust.

Example 4. In this example, we evaluate the performance of our proposed method in more complex scenarios by incorporating interactions under both strong and weak heredity assumptions. The mediators $M_{ik}(t)$ are generated using the same approach as in Example 1. The first three mediators M_1, M_2 and M_3 are active (significant) mediators. In Example 3, the interactions M_1M_2, M_1M_3, and M_2M_3 are constructed under the strong heredity assumption. In this example, we include three additional interactions M_1M_4, M_1M_5, and M_1M_6, which follow the weak heredity assumption. Specifically, the hazard function is modeled as: $\lambda_i(t|\overline{\mathbf{M}}_{ij}) = \lambda_0(t)\exp\{\theta_1 X_i + \boldsymbol{\theta}_2^\top \mathbf{Z}_i + \sum_{k=1}^{p}\beta_k M_{ik}(t) + \varphi_1 M_{i1}(t)M_{i2}(t) + \varphi_2 M_{i1}(t)M_{i3}(t) + \varphi_3 M_{i2}(t)M_{i3}(t) + \varphi_4 M_{i1}(t)M_{i4}(t) + \varphi_5 M_{i1}(t)M_{i5}(t) + \varphi_6 M_{i1}(t)M_{i6}(t)\}$, where $\varphi_1 = \varphi_2 = \varphi_3 = 0.4$, consistent with Example 3, and $\varphi_4 = \varphi_5 = \varphi_6 = 0.1$. All other parameters are generated in the same way as in Example 1. We also ignore interactions and consider only the main effects in the analysis.

Table 10 presents the biases and MSEs for the estimates of $\alpha_k\beta_k, k = 1, \ldots, 4$. Overall, the biases and MSEs remain small. Compared to including only interactions under the strong assumption in Example 3, incorporating interactions under both strong and weak heredity assumptions leads to a slight decrease in performance. This is likely due to the inclusion of additional interactions under the weak heredity condition, which introduces more noise into the model and reduces power when the interactions are omitted from the analysis.

Table 11 presents the retention frequency of active mediators after the screening step, based on 100 repetitions for each mediator. We observe that M_1 is consistently selected in all cases, while M_2 and M_3 are identified with high frequency, particularly when the sample size is increased to $n = 600$.

Table 12 shows the empirical power of mediators M_1, M_2, M_3, FWER, N_{TP}, and N_{FP} for the proposed and naive methods. Compared to the naïve approach, the proposed method achieves higher power across all mediators while maintaining a proper FWER, effectively identifying true mediators while controlling false positives.

Application

We apply our method to the CARDIA Study, which is designed to investigate the factors (behavioral, environmental, and race- and sex-associated) that contribute to the development of CVD. The CARDIA study is a multicenter, longitudinal, population-based cohort involving 5115 Black and White men and women aged 18–30 years old at baseline (1985–1986). Participants were recruited from four urban areas: Birmingham, Alabama; Chicago, Illinois; Minneapolis, Minnesota; and Oakland, California. Within each center, the sample is designed to have approximately equal numbers of participants by sex, race (Black or White), age groups (18–24 years and 25–30 years), and education levels (high school graduate or less, and beyond high school). Eight follow-up examinations have been conducted through 30 years, starting from baseline in 1985–1986 (Year 0), followed by 1987–1988 (Year 2), 1990–1991 (Year 5), 1992–1993 (Year 7), 1995–1996 (Year 10), 2000–2001 (Year 15), 2005–2006 (Year 20), 2010–2011 (Year 25), and 2015–2016 (Year 30). During the follow-up visits, extensive data have been collected, including residential addresses, demographics, psychosocial factors, subclinical and clinical CVD outcomes. DNAm data collection is planned at four time points: Year 15,

Table 10. The biases and MSEs of the estimated mediation effects in Example 4

	p = 5000				p = 10 000			
	n = 300		n = 600		n = 300		n = 600	
$\alpha_k\beta_k$	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\alpha_1\beta_1$	0.066	0.010	0.028	0.003	0.069	0.011	0.033	0.003
$\alpha_2\beta_2$	0.058	0.009	0.023	0.002	0.067	0.010	0.024	0.002
$\alpha_3\beta_3$	0.052	0.006	0.004	0.001	0.053	0.006	0.016	0.002
$\alpha_4\beta_4$	5e-4	2e-5	0.000	0.000	0.000	0.000	0.000	0.000

Table 11. The frequency of active mediators being kept after the screening step over 100 repetitions in Example 4

	p = 5000		p = 10 000	
M_k	n = 300	n = 600	n = 300	n = 600
M_1	100	100	100	100
M_2	97	100	94	100
M_3	81	100	82	96

Table 12. The empirical power of three active mediators, FWER, N_{TP}, N_{FP} after the multiple testing in Example 4

		p = 5000		p = 10 000	
		n = 300	n = 600	n = 300	n = 600
Proposed	P_{M_1}	0.92	1.00	0.91	1.00
	P_{M_2}	0.70	0.97	0.63	0.99
	P_{M_3}	0.29	0.90	0.29	0.81
	FWER	0.032	0.003	0.010	0.003
	N_{TP}	1.91	2.87	1.83	2.80
	N_{FP}	0.09	0.01	0.03	0.01
Naive	P_{M_1}	0.38	0.96	0.29	0.92
	P_{M_2}	0.05	0.61	0.05	0.54
	P_{M_3}	0.01	0.16	0.00	0.08
	FWER	0.00	0.00	0.00	0.00
	N_{TP}	0.44	1.73	0.34	1.54
	N_{FP}	0.00	0.00	0.00	0.00

P_{M_1}, P_{M_2}, P_{M_3}: the empirical power of mediators M_1, M_2, and M_3, respectively; FWER: the family-wise error rate; N_{TP}: average number of true positives; N_{FP}: average number of false positives.

Year 20, Year 25, and Year 30. Due to data availability, epigenome-wide DNAm data from blood samples of CARDIA subjects at Year 15 and Year 20 are included for our analysis, encompassing 856 626 CpG sites. Future studies may incorporate data from Years 25 and 30 to validate findings and investigate longitudinal changes in DNAm patterns.

The iSDH index measured at Year 10 is the exposure variable. Developed by Gao *et al.* [26], the iSDH index reflects more detrimental conditions with higher values. It incorporates personal factors (education, household income, occupation, financial strain), parental factors (education, occupation), and adverse childhood family environment, collected through structured questionnaires at the baseline and follow-up visits. The Boosted Regression Tree model is used to determine the contribution of each factor to coronary artery calcification risk. The selected SDHs, along with their contributions (including the direction of influence), are weighted and summed to construct the iSDH index. To enable comparability across study years, the index at each visit is standardized using the visit-specific standard deviation.

We consider the time to the first occurrence of CVD or death as the composite survival outcome, using Year 15 as the time origin. Since mediators are observed at Years 15 and 20, this choice ensures that the exposure at Year 10 precedes the mediators, maintaining the correct temporal sequence for causal inference: exposure → mediator → outcome. To maintain a valid causal relationship, individuals diagnosed with CVD or deceased before Year 15 are excluded. Individuals whose events occurred between Years 15 and 20 are retained, as their DNA methylation measured at Year 15 could still serve as a mediator. Notably, for these individuals, DNA methylation measured at Year 20 is not considered as a mediator since it is observed after the survival outcome. For individuals whose events occur after Year 20, DNAm measured at both Years 15 and 20 is used as the mediator. The model adjusts for priori confounders, including age, sex (coded as female = 0, male = 1), race (coded as white = 1, Black = 0), examining center (i) Birmingham, AL; (ii) Chicago, IL; (iii) Minneapolis, MN; and (iv) Oakland, CA, blood cell type proportion, and batch effect. All covariates are measured at Year 10, and continuous variables are standardized before analysis.

In our study, we address missing data for both the exposure and mediators. For the exposure, we impute missing iSDH factors with the mean of the subject during follow up visits and calculate the iSDH index accordingly. For the mediators, individuals who are not consent to DNAm profiling are excluded. We use the k-nearest neighbors method to impute missing data in the high-dimensional DNAm dataset, ensuring completeness and consistency for downstream analysis. Importantly, a comparison of the characteristics of subjects with and without DNAm data reveals no significant differences, suggesting that the DNAm sample is representative of the broader population. Ultimately, 2032 subjects with complete data—including exposure variables, covariates, DNAm data, and survival outcomes—are included in the analysis, with a total of 3663 measurements. Of these, 146 subjects are diagnosed with CVD or experienced death during the follow-up.

Our research specifically investigates how repeatedly measured DNAm markers mediate the relationship between iSDH and the risk of developing the composite survival outcome of CVD or death. The total effect of the exposure on the survival outcome, denoted as $X \to Y$, is estimated to be $\gamma = 0.322$, with a p-value of 0.0005, indicating a significant association. Table 13 presents the estimates and standard error (SE, in parenthesis) for $\hat{\alpha}_k$ and $\hat{\beta}_k$, the Bonferroni-adjusted p-values (p-value) and the percentage of total effect (per) for the selected DNAm markers that exhibit significant mediation effects based on our method, with a FWER threshold set at 0.05.

Our analysis identifies two significant CpGs as mediators. For cg05575921, located in the aryl hydrocarbon receptor repressor

Table 13. Summary of selected DNAm markers with significant mediation effects

CpGs	Gene	$\hat{\alpha}_k$ (SE)	$\hat{\beta}_k$ (SE)	p-value	per (%)
cg05575921	AHRR	–0.205 (0.023)	–0.164 (0.075)	0.033	10.4
cg06834630	KSR1	–0.062 (0.016)	–0.486 (0.108)	0.015	9.4

(AHRR) gene on chromosome 5, the estimated pathway effect for $X \rightarrow M$ is –0.205, indicating that higher levels of (detrimental) iSDH factors are associated with reduced methylation at cg05575921. The estimated pathway effect for $M \rightarrow Y$ is –0.164, suggesting that lower methylation at cg05575921 is associated with an increased risk of CVD/death. These findings are consistent with previous studies [7]. Similarly, cg06834630, located in the kinase suppressor of ras 1 (KSR1) gene on chromosome 17, exhibits estimated pathway effects of –0.062 for $X \rightarrow M$ and –0.486 for $M \rightarrow Y$. These results mirror the pattern observed for cg05575921. Thus, both selected CpGs contribute to a log-hazard indirect effect that favorable iSDH factors are associated with a decreased risk of CVD/death. Moreover, we examine the relative magnitudes of the total effect mediated through methylation markers, defined as $\alpha_k\beta_k/\gamma$ for each methylation marker. The results, presented in the last column of Table 13, show that 10.4% of total effect between iSDH index and CVD/death risk is mediated via cg05575921, and cg06834630 mediates 9.4% of the total effect.

Interestingly, although smoking during early adulthood is not a component of the iSDH index, it is often associated with detrimental iSDH factors, potentially contributing to some of the observed associations. Previous epigenome-wide association studies have also identified a connection between smoking and methylation changes at the AHRR gene, with cg05575921 being the most significantly affected CpG site [27]. Moreover, AHRR methylation has been explored as an objective marker for smoking behavior and its implications for pulmonary and cardiovascular risk prediction [28]. In addition, the KSR1 gene encodes scaffold proteins that are integral to the Raf/MEK/ERK MAPK signaling pathway. This pathway is crucial for various cellular processes, including those related to cardiovascular function. Notably, KSR1 is linked to isoform A of the Ras association domain-containing protein 1, which has been shown to regulate cardiac hypertrophy [29]. These findings highlight the importance of DNAm markers in mediating the relationship between early life social determinants and cardiovascular outcomes.

Discussion

In this study, we propose a novel model for high-dimensional mediation analysis that accommodates time-varying mediators and a survival outcome. Our approach integrates a longitudinal mixed effects model to assess the relationship between the exposure and mediators over time, and a Cox proportional hazards model to link the mediating process to the survival outcome. This framework addresses the complexities associated with longitudinal and survival data, providing a robust mechanism to explore mediation effects in high-dimensional settings.

There are several potential extensions for our method. First, our method can be extended to mediation analysis with multiomic data, such as lipidomics [30] and proteomics, where mediation analysis offers valuable insights into the comprehensive biological mechanisms. Second, further refinement of the Lasso inference procedure and the multiple-testing correction method can improve the accuracy and power of detecting significant mediators. We can use SCAD [31] and MCP [32] as alternative regularization techniques, which could yield better performance [33–36]. Third, the health outcome might depend on the expected values of markers, rather than the observed values of markers accompanying measurement errors. We will use the joint model to explore this mediation analysis framework, e.g. [37–39].

In this work, we make a simplifying assumption by treating the mediator process as piecewise constant, corresponding to the discrete times at which mediators are observed. While the underlying mediator process may be continuous, it is not directly observable. Our approach is consistent with standard practice in survival analysis, where time-dependent covariates are updated at observed time points. Modeling the continuous underlying processes would introduce significant complexity and is typically not pursued. Furthermore, addressing time-varying effects and adjusting for time-varying confounding covariates, especially in the context of high-dimensional mediation analysis, presents important challenges but is beyond the scope of this framework.

The choice between a point exposure and a time-varying exposure depends on whether the goal is to assess the potential impact of a one-time intervention or a continuous intervention. In our study, we choose a fixed exposure framework (at Year 10) to align with the temporal order necessary for causal inference: exposure → mediator → outcome. This choice simplifies the model while maintaining computational feasibility and interpretability, particularly in a high-dimensional setting with over 850 000 mediators. Importantly, if there is no inverse causation, i.e. early methylation (mediator) does not affect later social determinants of health (exposure) and no lag in the effect, the current method can be extended to incorporate time-varying exposures. However, if inverse causation exists, even low-dimensional models become complex, requiring methods such as the marginal structural model or g-formula [40] to compute causal effects. Extending these methods to a high-dimensional mediator setting is a promising direction for future work but is beyond the scope of this study.

We note that there is no universally accepted standard for the number of variables to retain in the screening step; however, the choice of $d = [n/2\log(n)]$ is consistent with the recommendations of [22]. This choice reflects a balance between false negatives (type II error) and false positives (type I error). In Step 2, we use the lasso penalty for variable selection, as it is currently the only available method for handling high-dimensional Cox models with time-dependent covariates. While lasso is known to be prone to false positives, reducing the number of mediators to $d = [n/2\log(n)]$ helps mitigate this issue, increasing the likelihood of identifying truly significant mediators. Conversely, retaining a larger number of variables, such as $d = [n/\log(n)]$, could reduce statistical power in the joint significance testing step, especially when Bonferroni's method is used for multiplicity correction. A larger d inflates the number of hypotheses tested, resulting in stricter corrections. Our simulation results indicate that the chosen criterion strikes an effective balance, demonstrating strong performance in practice.

In our current study, our model does not explicitly account for the time scale on which DNA methylation changes from baseline and corresponding effects occur. To capture these time-dependent dynamics, we could modify Model (2.2) as follows:

$$\lambda_i(t|\overline{\mathbf{M}}_{ij}) = \lambda_0(t)\exp\left\{\theta_1 X_i + \boldsymbol{\theta}_2^\top \mathbf{Z}_i + \sum_{k=1}^{p}\gamma_k M_{ik}(t_0) + \sum_{k=1}^{p}\beta_k \Delta M_{ik}(t)\right\},$$

where $\Delta M_{ik}(t)$ is the difference between $M_{ik}(t)$ at time t and $M_{ik}(t_0)$ at the time origin t_0 for the kth mediator. If the p-value for the coefficient β_k of $\Delta M_{ik}(t)$ is significant, it indicates that change of the kth mediator from t_0 is significantly associated with the time to event. This model allows us to assess the temporal changes in mediators and their impact on the survival outcome. However, incorporating these temporal changes introduces complexity, particularly in defining and interpreting indirect effects. For example, there are two high-dimensional vectors of variables $M_{ik}(t_0)$ and $\Delta M_{ik}(t)$, $k = 1, \ldots, p$, so variable selection is needed for these two sets of variables. Furthermore, this dual-parameter structure complicates the decomposition of total effects into direct and indirect components, making it more challenging to define and interpret mediation pathways clearly. Despite these challenges, such an extension provides a more nuanced understanding of the mediators' temporal roles in influencing survival outcomes, thus worth further exploration in future research.

Our study has several limitations. First, our mediation methods did not account for interaction effects among CpGs, a limitation also present in our previous studies [7, 10]. In our simulations, we evaluate the performance of our proposed method in the presence of interactions under different heredity conditions. Specifically, Example 3 includes interactions under the strong heredity condition, while Example 4 incorporates interactions under both strong and weak heredity conditions. Our method demonstrates consistent performance, even when these interactions are ignored during analysis. While incorporating interactions between pairs of mediators could enhance statistical power and provide more accurate estimates of indirect effects, doing so would significantly increase dimensionality and computational demands. Given that the primary goal of this study is to perform epigenome-wide mediation analysis with 856 626 potential mediators, incorporating all pairwise interactions is beyond the scope of this paper. Nevertheless, this is a promising direction for future research. Interaction screening methods for high-dimensional data (e.g. [41, 42]), could effectively identify interactions and offer a practical approach to addressing this challenge. Second, our current dataset does not include genomic variant information, and incorporating such data in future analyses could significantly enhance the depth and scope of the study. It is important to recognize that genomic variants could serve as potential confounders, especially if they directly influence the outcome independent of DNA methylation. To mitigate this, genetic variants could be included as additional covariates in the model to account for their confounding effects, thereby ensuring a more robust and accurate interpretation of the results. Finally, while potential DNAm markers are identified as significant mediators in this application, the causal relationships between iSDH, DNAm markers, and the outcomes require further investigation and validation through biological experiments. Incorporating such experimental evidence into future research will be crucial to strengthening the interpretation of these findings.

Key Points

- We develop high-dimensional mediation analysis methods for longitudinal mediators and survival outcomes.
- Our method can accommodate time-varying mediators, avoiding the limitations of simplified aggregate approaches.
- Our approach outperforms existing methods, offering enhanced accuracy and rigorous control of family-wise error rates.
- Our model addresses a critical need in longitudinal studies to identify pathways linking environmental exposures to survival outcomes.

Conflict of interest: None declared.

Funding

This research is partly supported by NIH grants R21 AG063370, R21 AG068955, R01 AG081244, R01 AG069120, UL1 TR002345. Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program were supported by the National Heart, Lung and Blood Institute (NHLBI). Methylomics for "NHLBI TOPMed: Whole Genome Sequence Analysis in Early Cerebral Small Vessel Disease" (phs001612.v3.p3) was performed at Molecular Genomics Core (MGC) at the Keck School of Medicine of the University of Southern California (USC) (HHSN2682016000038I). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN2682018000021). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN2682018000011). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

The CARDIA is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the University of Alabama at Birmingham (HHSN2682018000051 and HHSN2682018000071), Northwestern University (HHSN2682018000031), University of Minnesota (HHSN2682018000061), and Kaiser Foundation Research Institute (HHSN2682018000041). CARDIA was also partially supported by the Intramural Research Program of the National Institute on Aging (NIA) and an intra-agency agreement between NIA and NHLBI (AG0005).

Data availability

Restrictions apply to the availability of these data, which were used under license for CARDIA study. The code for implementing the proposed method is publicly available at https://github.com/liliwustl/high-dimensional-longitudinal-mediation.

References

1. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 1986;**51**:1173–82. https://doi.org/10.1037/0022-3514.51.6.1173
2. MacKinnon D. *Introduction to Statistical Mediation Analysis*. New York: Routledge, 2012. https://doi.org/10.4324/9780203809556.
3. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods* 2008;**40**:879–91. https://doi.org/10.3758/BRM.40.3.879
4. VanderWeele T, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiol Methods* 2014;**2**:95–115. https://doi.org/10.1515/em-2012-0010

5. Zhang H, Chen J, Feng Y. *et al.* Mediation effect selection in high-dimensional and compositional microbiome data. *Stat Med* 2021;**40**:885–96. https://doi.org/10.1002/sim.8808
6. Sohn MB, Li H. Compositional mediation analysis for microbiome studies. *Ann Appl Stat* 2019;**13**:661–81. https://doi.org/10.1214/18-AOAS1210
7. Zhang H, Zheng Y, Zhang Z. *et al.* Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* 2016;**32**:3150–4. https://doi.org/10.1093/bioinformatics/btw351
8. Gao Y, Yang H, Fang R. *et al.* Testing mediation effects in high-dimensional epigenetic studies. *Front Genet* 2019;**10**:1195. https://doi.org/10.3389/fgene.2019.01195
9. Zhou RR, Wang L, Zhao SD. Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika* 2020;**107**:573–89. https://doi.org/10.1093/biomet/asaa016
10. Zhang H, Zheng Y, Hou L. *et al.* Mediation analysis for survival data with high-dimensional mediators. *Bioinformatics* 2021;**37**:3815–21. https://doi.org/10.1093/bioinformatics/btab564
11. Luo C, Fa B, Yan Y. *et al.* High-dimensional mediation analysis in survival models. *PLoS Comput Biol* 2020;**16**:e1007768. https://doi.org/10.1371/journal.pcbi.1007768
12. Vansteelandt S, Linder M, Vandenberghe S. *et al.* Mediation analysis of time-to-event endpoints accounting for repeatedly measured mediators subject to time-varying confounding. *Stat Med* 2019;**38**:4828–40. https://doi.org/10.1002/sim.8336
13. Winkleby MA, Jatulis DE, Frank E. *et al.* Socioeconomic status and health: How education, income, and occupation contribute to risk factors for cardiovascular disease. *Am J Public Health* 1992;**82**:816–20. https://doi.org/10.2105/AJPH.82.6.816
14. Schultz WM, Kelli HM, Lisko JC. *et al.* Socioeconomic status and cardiovascular outcomes: challenges and interventions. *Circulation* 2018;**137**:2166–78. https://doi.org/10.1161/CIRCULATIONAHA.117.029652
15. Lynch JW, Kaplan GA, Cohen RD. *et al.* Do cardiovascular risk factors explain the relation between socioeconomic status, risk of all-cause mortality, cardiovascular mortality, and acute myocardial infarction? *Am J Epidemiol* 1996;**144**:934–42. https://doi.org/10.1093/oxfordjournals.aje.a008863
16. Franks P, Winters PC, Tancredi DJ. *et al.* Do changes in traditional coronary heart disease risk factors over time explain the association between socio-economic status and coronary heart disease? *BMC Cardiovasc Disord* 2011;**11**:1–6. https://doi.org/10.1186/1471-2261-11-28
17. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 1996;**58**:267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
18. Aalen OO, Stensrud MJ, Didelez V. *et al.* Time-dependent mediators in survival analysis: modeling direct and indirect effects with the additive hazards model. *Biom J* 2020;**62**:532–49. https://doi.org/10.1002/bimj.201800263
19. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press, 2015. https://doi.org/10.1017/CBO9781139025751.
20. Huang Y-T, Yang H-I. Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology* 2017;**28**:370.
21. VanderWeele TJ. Causal mediation analysis with survival data. *Epidemiology* 2011;**22**:582–5. https://doi.org/10.1097/EDE.0b013e31821db37e
22. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Series B Stat Methodology* 2008;**70**:849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x
23. Fan J, Yang F, Yichao W. High-dimensional variable selection for Cox's proportional hazards model. In: Berger JO, Cai T, Johnstone I (eds), *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, Beachwood, OH: Institute of Mathematical Statistics, 2010. **6**, 70–87. https://doi.org/10.1214/10-IMSCOLL606.
24. Cygu S, Dushoff J, Benjamin MB. pcoxtime: penalized Cox proportional hazard model for time-dependent covariates. arXiv preprint arXiv: 2102.02297. 2021.
25. Austin PC. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Stat Med* 2012;**31**:3946–58. https://doi.org/10.1002/sim.5452
26. Gao T, Zheng Y, Joyce BT. *et al.* Developing a novel index for individual-level social determinants and cardiovascular diseases in the coronary artery risk development in young adults (CARDIA) study. *Int J Environ Res Public Health* 2025;**22**:422. https://doi.org/10.3390/ijerph22030422
27. Zeilinger S, Kühnel B, Klopp N. *et al.* Tobacco smoking leads to extensive genome-wide changes in dna methylation. *PloS One* 2013;**8**:e63812. https://doi.org/10.1371/journal.pone.0063812
28. Langsted A, Bojesen SE, Stroes ESG. *et al.* AHRR hypomethylation as an epigenetic marker of smoking history predicts risk of myocardial infarction in former smokers. *Atherosclerosis* 2020;**312**:8–15. https://doi.org/10.1016/j.atherosclerosis.2020.08.034
29. Ganesan J, Ramanujam D, Sassi Y. *et al.* Mir-378 controls cardiac hypertrophy by combined repression of mitogen-activated protein kinase pathway factors. *Circulation* 2013;**127**:2097–106. https://doi.org/10.1161/CIRCULATIONAHA.112.000882
30. Getz KR, Jeon MS, Liu L. *et al.* Metabolites and lipid species mediate the associations of adiposity in childhood and early adulthood with mammographic breast density in premenopausal women. *Breast Cancer Research BCR* 2025;**27**:18. https://doi.org/10.1186/s13058-025-01970-6
31. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001;**96**:1348–60. https://doi.org/10.1198/016214501753382273
32. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010;**38**:894–942. https://doi.org/10.1214/09-AOS729
33. Liu L, Lin L. Subgroup analysis for heterogeneous additive partially linear models and its application to car sales data. *Comput Stat Data Anal* 2019;**138**:239–59. https://doi.org/10.1016/j.csda.2019.04.011
34. Liu L, Mae G, Miller JP. *et al.* Capturing heterogeneity in repeated measures data by fusion penalty. *Stat Med* 2021;**40**:1901–16. https://doi.org/10.1002/sim.8878
35. Liu L, He K, Wang D. *et al.* Healthcare center clustering for Cox's proportional hazards model by fusion penalty. *Stat Med* 2023;**42**:3685–98. https://doi.org/10.1002/sim.9825
36. Liu L, He K, Wang D. *et al.* Health care provider clustering using fusion penalty in quasi-likelihood. *Biom J* 2024;**66**:e202300185. https://doi.org/10.1002/bimj.202300185
37. Liu L, Ma JZ, O'Quigley J. Joint analysis of multi-level repeated measures data and survival: an application to the end stage renal disease (ESRD) data. *Stat Med* 2008;**27**:5679–91. https://doi.org/10.1002/sim.3392

38. Liu L, Zheng C, Kang J. Exploring causality mechanism in the joint analysis of longitudinal and survival data. *Stat Med* 2018;**37**:3733–44. https://doi.org/10.1002/sim.7838
39. Zheng C, Liu L. Quantifying direct and indirect effect for longitudinal mediator and survival outcome using joint modeling approach. *Biometrics* 2022;**78**:1233–43. https://doi.org/10.1111/biom.13475
40. VanderWeele TJ, Tchetgen Tchetgen EJ. Mediation analysis with time varying exposures and mediators. *J R Stat Soc Ser B. Stat Methodol* 2017;**79**:917–38. https://doi.org/10.1111/rssb.12194
41. Liu L, Lin L, Liu L. Interaction screening in high-dimensional multi-response regression via projected distance correlation. *Commun Stat-Simul Comput* 2024;1–26. https://doi.org/10.1080/03610918.2024.2393691
42. Hao N, Zhang HH. Interaction screening for ultrahigh-dimensional data. *J Am Stat Assoc* 2014;**109**:1285–301. https://doi.org/10.1080/01621459.2014.881741

Briefings in Bioinformatics, 2025, **26(3)**, bbaf206
https://doi.org/10.1093/bib/bbaf206
Case Study