BMC Medical Research Methodology

RESEARCH

Open Access

High-dimensional mediation analysis for continuous outcome with confounders using overlap weighting method in observational epigenetic study

Weiwei Hu[1], Shiyu Chen[1], Jiaxin Cai[1], Yuhui Yang[1], Hong Yan[1] and Fangyao Chen[1,2*]

Abstract

Background Mediation analysis is a powerful tool to identify factors mediating the causal pathway of exposure to health outcomes. Mediation analysis has been extended to study a large number of potential mediators in high-dimensional data settings. The presence of confounding in observational studies is inevitable. Hence, it's an essential part of high-dimensional mediation analysis (HDMA) to adjust for the potential confounders. Although the propensity score (PS) related method such as propensity score regression adjustment (PSR) and inverse probability weighting (IPW) has been proposed to tackle this problem, the characteristics with extreme propensity score distribution of the PS-based method would result in the biased estimation.

Methods In this article, we integrated the overlapping weighting (OW) technique into HDMA workflow and proposed a concise and powerful high-dimensional mediation analysis procedure consisting of OW confounding adjustment, sure independence screening (SIS), de-biased Lasso penalization, and joint-significance testing underlying the mixture null distribution. We compared the proposed method with the existing method consisting of PS-based confounding adjustment, SIS, minimax concave penalty (MCP) variable selection, and classical joint-significance testing.

Results Simulation studies demonstrate the proposed procedure has the best performance in mediator selection and estimation. The proposed procedure yielded the highest true positive rate, acceptable false discovery proportion level, and lower mean square error. In the empirical study based on the GSE117859 dataset in the Gene Expression Omnibus database using the proposed method, we found that smoking history may lead to the estimated natural killer (NK) cell level reduction through the mediation effect of some methylation markers, mainly including methylation sites cg13917614 in CNP gene and cg16893868 in LILRA2 gene.

Conclusions The proposed method has higher power, sufficient false discovery rate control, and precise mediation effect estimation. Meanwhile, it is feasible to be implemented with the presence of confounders. Hence, our method is worth considering in HDMA studies.

Keywords High-dimensional mediation model, Propensity score, Overlap weighting, Joint significant test, Composite null hypothesis

*Correspondence:
Fangyao Chen
chenfy@xjtu.edu.cn
Full list of author information is available at the end of the article

Background

The analysis of the mediating effect was first proposed by Baron and Kenny (1986) [1] and was broadly applied in many scientific fields, such as psychological, sociological, and biomedical studies [2–4]. Mediation analysis has become a powerful tool to investigate the underlying mechanism of environmental exposures on health outcomes and identify the factors mediating the effect of exposures on outcomes [5]. Currently, analytical methods including the single mediator model [6, 7], multiple-mediators model [8], and high-dimensional mediation model [9] are proposed and available for researchers in many scientific fields.

With the development of advanced data collection techniques, high-dimensional data has become common in biomedical research. For example, in the epigenetic study, the Illumina Infinium HumanMethylation450 BeadChip array platform allows to measure the DNA methylation levels of roughly 480 K probes [10] and generates high dimensional data. Focusing on practical research, smoking affects lung function, and some DNA methylation sites may mediate the effect of smoking on lung function [11, 12]. To identify the significant mediators (CpG sites) between smoking and lung function, we can conduct mediation analysis in the collected high-dimensional data [9, 13, 14]. Obviously, this method can be used to identify the methylation sites mediating the association between environmental factors other than smoking and other health outcomes including some physical signs and diseases.

However, there are also some issues in high dimensional mediation analysis (HDMA), such as the curse of dimensionality, the false positive rate inflation caused by multiplicity and the confounding existing in observational research. To overcome these issues, scholars have proposed a series of statistical methods. Zhang et al. [9]. proposed the HIMA model consisting of variable screening based on sure independence screening (SIS), variable selection techniques based on minimax concave penalty (MCP) estimation and joint significance test. HIMA extends the multiple mediator framework to the high-dimensional setting by incorporating variable screening and variable selection techniques into multiple mediation analysis. The following high-dimensional mediation analysis methods also employ the generic procedure [13–16], which reduces dimensionality from high to moderate or low scale and then conducts multiple mediation test. For example, the HIMA2 procedure proposed by Perera et al. [17], which employs the SIS method based on the indirect effect of every single mediator and conducts debiased Lasso to obtain more accurate estimates, then utilizes the multiple-testing procedure proposed by James et al. [18] to control the false discovery rate. Moreover, to adjust the confounders of observational epigenetic studies, researchers tried to integrate propensity score (PS) into the high-dimensional mediation model by weighting or considering it as a covariate [14, 16], except for the classic regression adjustment.

Although many works have been made to tackle these problems, there are still some issues remaining in the dimensionality reduction and adjustment for confounders. For high dimensional mediation analysis, the previous studies don't take confounders into account, just consider them as covariates [15, 19], such as HDMA, HIMA, and HIMA2 [5, 9, 17]. As is known to all, the multivariable model cannot adequately account for confounding effects in the presence of a large number of confounders [20]. If we only control confounding during the mediation test, but not in the dimension reduction stage, then a biased variable selection result may be obtained [14]. Thus, it is necessary to adjust confounders to improve the performance of variable selection.

To address this issue, researchers have adopted the PS-based method including PS regression adjustment (termed PSR) and classical PS weighting (also called inverse probability weighting, IPW) to adjust confounding during both stages [14]. However, the adjustment for confounders using the IPW based on PS still faces the issue of extreme weights caused by extreme PS distribution [21, 22]. To address the issue of extreme PS distribution, Li et al. [23]. proposed the overlap weighting (OW) method, which emphasizes individuals with the most overlap in their observed characteristics and is beneficial to provide a consistent estimator of the effect of exposure on outcome in the presence of extreme PS tails. OW belongs to the weighting confounding adjustment method based on PS and is gaining more popularity because of excellent statistical properties [24, 25]. However, the above OW method is only applied to traditional epidemic analysis, which needs to be extended to mediation analysis and high-dimensional data setting. Besides, most of the existing methods all hold the independent assumption between potential mediators, which is hard to ensure in high dimensional epigenetic data analysis [5, 9, 13–15, 18].

In this article, we incorporated the OW method into HIMA [9] and HIMA2 [17] models, respectively. In order to develop the accuracy of the screening of potential mediators, we modified the framework of variable screening in the original HIMA2 procedure. Eventually, we proposed the OW-based modified HIMA2 (mHIMA2) procedure for HDMA. We evaluated the performance of the proposed procedure and the existing models through simulation studies. All the above evaluations are based on the simulation study and real data application.

The rest of the article is structured as follows. In the next section, we introduced the notions, assumptions, models, and the procedure of adjustment for confounders in the high-dimensional mediation analysis model. Then, we conducted the Monte Carlo simulation study to evaluate the performance of various methods of confounding adjustment and two different mediation test approaches. Additionally, we applied the proposed method to the dataset GSE117859 in the Gene Expression Omnibus (GEO) databases and identified some DNA methylation markers that mediate the effect of smoking on the estimated natural killer (NK) cell level. Finally, we concluded the advantages and limitations of this study.

Methods

Model definitions

Our high-dimensional mediation model is shown in Fig. 1. Let X be the exposure variable, where $X = 1$ represents the exposed group and $X = 0$ represents the controlled group. Denote the outcome as Y, here we mainly focus on continuous outcome. Let $M = (M_1, M_2, \cdots, M_P)^T$ be the set of the p-dimensional potential mediators, where $p \gg n$, n is the sample size. Let $C = \left(C_1, C_2, \cdots, C_q\right)^T$ be the q-dimension baseline confounders which influence the relation of exposure-mediator, mediator-outcome, and exposure-outcome. For individual $i, i = 1,2, \cdots, n$, we have the high-dimensional mediation models as follows:

$$M_{ki} = a_k + \alpha_k X + \phi_k^T C_i + e_{ki}, k \in [p], \quad (1)$$

$$Y_i = a + \gamma X_i + \beta^T M_i + \eta^T C_i + \epsilon_i. \quad (2)$$

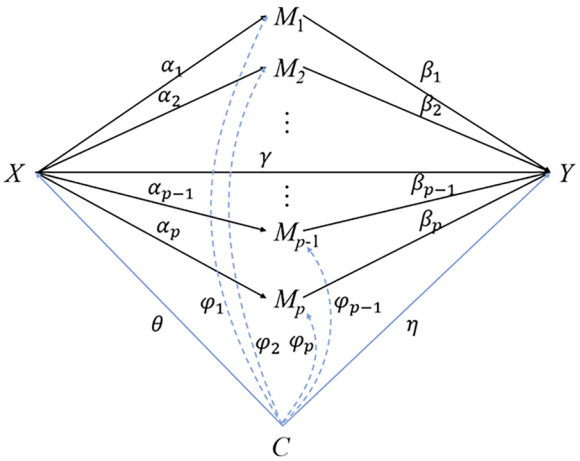

Fig. 1 Causal diagram. High-dimensional mediation model with confounders between exposure, mediator and outcome

where $\alpha = \left(\alpha_1, \alpha_2, \cdots, \alpha_p\right)^T$ is the coefficient vector relating the exposure to the mediators, $\beta = \left(\beta_1, \beta_2, \cdots, \beta_p\right)^T$ represents the effect of the mediators on the outcome, $\alpha_k\beta_k$ corresponds to the mediation effect of M_k according to the definition of coefficients product method, and $[p]$ denotes the set of $\{1,2, \cdots, p\}$. One can consider whether M_k is the statistically significant mediator or not by testing the null hypothesis $H_0 : \alpha_k\beta_k = 0$. ϕ_k and η are the effect of C on M and C on Y, respectively. a_k and a are the intercept term in the Eqs. 1 and 2, respectively. The same as above, e_k and ϵ are each the corresponding error term. We will compare the different variable selection strategies and methods of adjusting confounders.

Assumptions

To ensure the identification of path-specific mediating effects, some assumptions need to be held as below. These assumptions were proposed referring to necessary condition required for high-dimensional mediation analysis suggested in published studies [8, 15, 17, 19, 26, 27]:

A1: There is no causal association between mediators. This means the proposed model contains only parallel mediators.

A2: Sequential ignorability. That consists of four assumptions listed below:

(A2.1) There are no unmeasured confounders between the exposure and the outcome;

(A2.2) There are no unmeasured confounders between the mediators and the outcome;

(A2.3) There are no unmeasured confounders between the exposure and the mediators;

(A2.4) There is no exposure-induced confounding between the mediators and the outcome.

A3: Stable unit treatment value assumption (SUTVA) [28, 29] for both the mediators and the outcome. That is to say, there is no interference between individuals.

A4: Consistency for the mediators and the outcome. That is to say, there are no measurement errors in the mediators.

A5: Positivity assumption [30]. Every individual has some positive probability of being exposed to the factor of interest.

Proposed Procedure

We improved the HIMA procedure proposed by Zhang et al. (2016) [9] and the HIMA2 procedure proposed by Perera et al. (2022) [17] under the condition of adjusting confounders in observational data.

In this study, we developed two processes to conduct the confounding-controlled high-dimensional mediation analysis. The detailed procedure is described in the following text.

Step 1: PS-based methods for adjusting confounders

Since there are always some baseline confounders in observational data, we integrate propensity score (PS) into mediators (and/or outcome) models to reduce the selection bias and acquire as accurate estimates of the mediation effect as possible. Due to the PS approaches allowing the inclusion of a large scale of confounders, PS is widely used in observational research.

PS is defined as the conditional probability that a study individual with baseline covariates $C = (C_1, C_2, \cdots, C_l)$ would be exposed to certain study factors of interest [31]:

$$PS = P(X = 1|C_1, \cdots, C_l).$$

PS can be estimated by classic multivariable statistical methods such as logistic regression [32] or by machine learning methods such as random forest (RF) and generalized boosted model (GBM) [33, 34]. In practice, logistic regression is the most commonly used. The PS of ith individual $\pi_i = P(X_i = 1|C_{1i}, \cdots, C_{li})$ can be expressed as follows:

$$logit(\pi_i = P(X_i = 1)) = \theta_0 + \theta_1 C_{1i} + \cdots + \theta_l C_{li},$$

where $\theta = (\theta_1, \theta_2, \cdots, \theta_l)^T$ represents the effect of the confounders on the exposure. Then we can adopt some PS-based techniques to adjust confounders such as matching [35], stratification [36], regression [31], and weighting [37]. Here, we focus on PS regression (PSR) and PS weighting [14] (PSW, also called IPW short for inverse probability weighting) techniques to adjust potential confounders between exposure, mediators and outcome.

PSR approach incorporates PS as a covariate into the original regression model to adjust for the probability of being exposed to study factors and to reduce confounding [32]. That is similar to taking all confounders as covariates in a classical regression approach which usually uses the linear regression model for continuous outcomes and the logistic regression model for binary outcomes [38]. For the PSR approach, we can estimate the effect through the models below:

$$\begin{aligned} M_{ki} &= a_k + \alpha_k X + \phi_k' \pi_i + e_{ki}, k \in [p], \\ Y_i &= a + \gamma X_i + \beta^T M_i + PS_i + \epsilon_i. \end{aligned} \tag{3}$$

The PSW approach constructs the inverse probability weights by taking the reciprocal of PS. For binary exposure, the weight of the exposed group $X = 1$ is given as $\frac{1}{PS}$, and that of the controlled group $X = 0$ as $\frac{1}{1-PS}$. For ith individual:

$$ipw_i = \frac{1}{P(X_i = 1|C)} = \frac{X_i}{\pi_i} + \frac{(1 - X_i)}{(1 - \pi_i)}.$$

Then, we can estimate the coefficients of X in pathways $X \to M$ and $M \to Y$ by weighted estimation:

$$\begin{aligned} M_{ki} &= a_k + \alpha_{k,ipw} X + \phi_k^T C_i + e_{ki}, k \in [p], \\ Y_i &= a + \gamma_{k,ipw} X_i + \beta^T M_i + \eta^T C_i + \epsilon_i, \end{aligned} \tag{4}$$

where $\alpha_{k,ipw}$ and γ_{ipw} are the weighted estimation according to the *ipw* weight vector. However, the IPW often faces extreme PSs issue which may lead to extreme weights and result in biased estimates and excessive variance [23, 24].

The overlap weighting (OW) approach was proposed to address the issue of extreme PSs [23]. The overlap weight is given as $1 - PS$ for the group $X = 1$ and PS for the group $X = 0$. Note that, individuals with PS of 0.5 make the largest contribution to the effect estimate, and individuals with PS close to 0 and 1 make the smallest contribution. OW is likely to be beneficial in the presence of extreme tail weights [23, 39]. For individual i:

$$ow_i = \begin{cases} 1 - \pi_i, X_i = 1 \\ \pi_i, X_i = 0 \end{cases}.$$

Then, the effect estimation of OW is similar to that of the PSW procedure:

$$\begin{aligned} M_{ki} &= a_k + \alpha_{k,ow} X + \phi_k^T C_i + e_{ki}, k \in [p], \\ Y_i &= a + \gamma_{k,ow} X_i + \beta^T M_i + \eta^T C_i + \epsilon_i, \end{aligned} \tag{5}$$

In the same way, $\alpha_{k,ow}$ and γ_{ow} are the weighted estimation using *ow* weight vector.

Step 2: Confounding-controlled SIS approach for dimensionality reduction

The SIS procedure is a general technique to reduce accurately high dimensions to below sample size [40]. We adopt the SIS method to reduce dimension p from ultra-high dimension to moderate scale $d = \left[\frac{2n}{\log(n)}\right]$ [9, 15].

In this study, we considered two preliminary screening strategies as described in HIMA [9] and HIMA2 [17], based on the effects of M on Y (β_k) and the indirect effect $|\alpha_k \beta_k|$ respectively. Because the indirect effects can be both positive and negative effects, to address the influence of the signs of the estimated indirect effects, the HIMA2 approach uses the absolute values of the indirect effect to obtain the size of the effect estimate regardless of the direction. This approach ensures that mediators with large effect size can be selected.

Due to the lack of screening accuracy in SIS based on indirect effects in the presence of confounders, we conducted the SIS screening based on the effects on the path $M \to Y$ controlling confounding effects using the OW approach.

In simulation, we found that it is hard to select the true mediators based on $|\alpha_k\beta_k|$ in the presence of confounding factors as applied in the original HIMA2 approach. So, we modified the frame of the HIMA2 method and both adopt SIS based on the effects on the path $M \rightarrow Y$ β_k in the preliminary screening to select the subset of potential mediators $M_{SIS} = \{M_k : M_k \text{isamongthetop} d \text{largesteffect} of \beta_k\}$.

Noticing that we need to adopt a two-step weighting method [14] to estimate β_k for the PSW and OW methods.

First, $\gamma_{k,w}$ can be obtained from the following sub-model:

$$Y_i = a + \widehat{\gamma}_{k,w} X_i + \beta_k M_{ki} + \epsilon_{ki}$$

where $\widehat{\gamma}_{k,w}$ is the estimator of $\gamma_{k,ow}$ or $\gamma_{k,ipw}$ for each M_k. In addition, the residual $\widehat{e}_k$ can be derived:

$$\widehat{e}_k = Y - \widehat{\gamma}_{k,w} X.$$

Then β_k can be estimated by regressing $\widehat{e}_k$ on M_k without weighting. Through the above SIS procedure, we can identify the important mediators and achieve the goal of dimensionality reduction.

Step 3: Penalized estimation

According to the HIMA procedure, after the preliminary selection of candidate mediators, further variable selection can be accomplished by the penalized estimation method. Here, we adopt the MCP [41] rather than other penalty functions, since the MCP approach has the oracle property which can select the correct model with probability tending to 1 as $n \rightarrow \infty$ [15, 41, 42].

For the d-dimensional subset M_{SIS}, we employed the MCP-penalized estimation to further select significant mediators set $M_{MCP} = \{M_k : \beta_k \neq 0, M_k \in M_{SIS}\}$, MCP penalty function can be defined as below:

$$P_{\lambda,\delta}(\beta_k) = \lambda\left[|\beta_k| - \frac{|\beta_k|^2}{2\lambda\delta}\right] I\{0 \leq |\beta_k| < \delta\} + \frac{\lambda^2\delta}{2} I\{|\beta| \geq \delta\lambda\}$$

where $\lambda > 0$ is the regularization parameter which can be selected by AIC or BIC, and $\delta > 0$ is the tuning parameter which determines the concavity of MCP. The MCP procedure can be implemented through the R package *ncvreg* [43]. Through MCP penalty estimation, we filtered out the mediators with too weak effects by combining SIS and MCP procedures and then acquired the small number of mediators that needed to be tested. That will help to obtain more accurate effect estimates.

Following the original HIMA2 procedure, the penalized estimation adopts the de-biased Lasso method to get the estimator $\widehat{\beta}_k$ and standard error $\widehat{\sigma}_{\beta_k}$. The sub-model of the de-biased Lasso method can be described below:

$$Y = a + \gamma X + \beta_{SIS}^T M_{SIS} + \eta^T C + \epsilon$$

where β_{SIS} denote the effects of $M_k \in M_{SIS}$ on Y. The corresponding P-values P_{β_k} are given as:

$$P_{\beta_k} = 2\left\{1 - \Phi\left(\left|\widehat{\beta}_k\right| / \widehat{\sigma}_{\beta_k}\right)\right\},$$

where $\Phi(.)$ is the cumulative distribution function of standard normal distribution $N(0,1)$. The de-biased Lasso method can be implemented with the R package *hdi*.

Step 4: PS-based multiple mediation test

After MCP-based penalized estimation, we use the Joint significance test [3, 44] (termed JS-uniform) to test the mediation effect of $M_k \in M_{MCP}$. The Joint significance test considers the M_k as a true mediator when α_k and β_k is significant simultaneously. Here, α_k can be estimated through different confounding adjustment methods as shown in Eqs. 1, 3, 4, and 5. β_k can be obtained using the linear regression with considering all confounders as covariates or only including PS (summary of all confounders) as a covariate.

In other words, that is based on the P-values for testing the path-specific effects $H_0 : \alpha_k = 0$ or $H_0 : \beta_k = 0$. The raw P-value for the joint significance test [3] is defined below:

$P_{raw,k} = \max\left(P_{raw,\alpha_k}, P_{raw,\beta_k}\right)$, #where P_{raw,α_k} and P_{raw,β_k} are the P-values for testing $H_0 : \alpha_k = 0$ and $H_0 : \beta_k = 0$. P_{raw,α_k} and P_{raw,β_k} can be obtained from the mediator model (e.g. Equations 1, 3, 4, and 5) and outcome model (Eq. 2), respectively.

For the multiplicity (Type I error inflation) issue in multiple mediation testing, we adopted the Benjamini–Hochberg (BH) method [45, 46] to acquire the adjusted p-values as below,

$$P_{BH,k} = \min\left(P_{raw,k} \cdot \frac{q}{r_k}, 1\right),$$

where q is the number of potential mediators in the set M_{MCP}, and r_k is the location number of $P_{raw,k}$ when all the P-values $P_{raw,k}$ are sorted ascending.

However, the Joint significance test assumes $P_{raw,k}$ follows a uniform null distribution. Although P_{α_k} and P_{β_k} are each uniformly distributed, their maximum may not. Therefore, the Joint significance test results in a valid but overly conservative test with lower power [13, 17, 47].

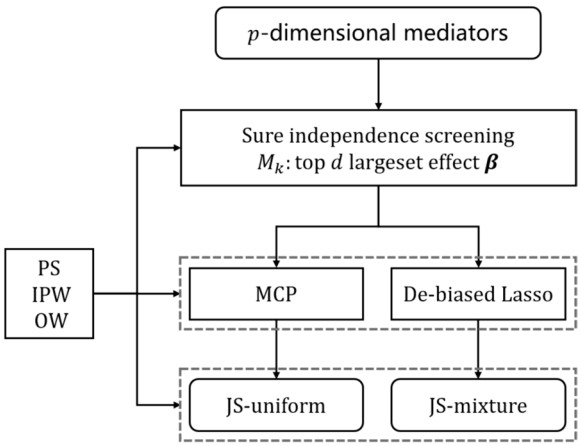

Fig. 2 The overall workflow for high-dimensional mediation analysis under the adjusting for confounders condition

Hence, we adopt the PS-based joint significance with mixture null distribution method [18] (termed JS-mixture) approach to conduct multiple mediation test after de-biased Lasso penalized estimation [17, 48] referring to the classical HIMA2 procedure. The PS-based JS-mixture approach adopts a 3-component mixture distribution as below:

$$\begin{aligned} H_{00,k} &: \alpha_k = 0 \text{ or } \beta_k{=}0, \\ H_{01,k} &: \alpha_k = 0 \text{ or } \beta_k \neq 0, \\ H_{10,k} &: \alpha_k \neq 0 \text{ or } \beta_k = 0. \end{aligned}$$

The estimated pointwise FDR for testing mediation can be computed as:

$$\widehat{FDR}(t) = E\left[\frac{V_{00}(t)+V_{01}(t)+V_{10}(t)}{\max(R(t),1)}\right],$$

where $t \in [0,1]$, $V_{00}(t), V_{01}(t), V_{10}(t)$ denoting the numbers of the three types of false positives and $R(t) = V_{00}(t) + V_{01}(t) + V_{10}(t) + V_{11}(t)$. The $V_{00}(t), V_{01}(t), V_{10}(t)$ and $\widehat{FDR}(t)$ can be obtained using the R package *HDMT*.

We set the significance level of 0.05 for all the tests. The detailed processes of the proposed method are summarized in Fig. 2.

Simulation studies

Simulation design

In this section, we conducted the simulation studies to evaluate the performance of the proposed method. The implementation of the simulation was based on R (version 4.3.0, R Foundation for Statistical Computing, Vienna, Austria) and RStudio (version 2023.9.0.463, RStudio: Integrated Development Environment for R, Boston, MA). The setting of simulation parameters was based on the published studies [9, 14, 16]. The number of replications in simulation study was set to be 500 for each combination of parameter setting referring to the replication times settings in published methodological studies [9, 14–17, 19, 49].

The model structure is shown in Fig. 1. We consider 8 confounders $C = (C_1, C_2, \cdots, C_8)$ affecting the relationship of X, M, Y, in which continuous confounders $C_1 - C_4$ follow a multivariate normal distribution $N(\mu, \Sigma)$ with a mean vector $\mu = (0,0,0,0)^T$ and a covariance matrix Σ:

$$\Sigma = \begin{bmatrix} 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 \end{bmatrix}.$$

The last four binary confounders $C_5 - C_8$ are independently generated from the Binary distribution $B(n, 0.3)$, where n is the sample size.

Then exposure X can be generated from Binary distribution $B(n, P_c)$, where n is the sample size, $P_c = 1/\left(1 + e^{-(\theta^T C)}\right)$, and $\theta^T = (\theta_1, \theta_2, \cdots, \theta_8) = (0.2,0.2,0.3,0.3,0.2,0.2,0.3,0.3)$.

Mediators M and the outcome variable Y are generated according to Eqs. 1 and 2, respectively. For simplicity, we set all the effects of C on M to be the same. Let $\phi_k = (\phi_{k1}, \cdots, \phi_{k8})^T = (0.2,0.2,0.3,0.3,0.2,0.2,0.3,0.3)^T$ represent the effect of C on M. Let $\eta = (\eta_1, \eta_2, \cdots, \eta_8)^T = (0.2,0.2,0.3,0.3,0.2,0.2,0.3,0.3)^T$ denote the effects of C on Y.

We set the first four potential mediators $M_1 - M_4$ as the true significant mediators in this study. Let $\alpha = \left(\alpha_1, \alpha_2, \cdots, \alpha_p\right)^T = (0.4,0.4,0.5,0.5,0.5,0.5,0,0,\cdots,0)$; $\beta = \left(\beta_1, \beta_2, \cdots, \beta_p\right)^T = (0.4,0.5,0.5,0.6,0,0,0.5,0.5,0,\cdots,0)$. The elements of both α and β are equal to zero except for the first eight elements, and the first four are the significant mediators. The mediation effect size of the true mediators $M_1 - M_4$ is $\alpha\beta_{1-4} = (0.16,0.2,0.25,0.3)$.

Let $\gamma = 0.5$; $a = 0.5$; $a_k \sim U(0,1)$, $\epsilon \sim N(0,1)$. The error term e_k are generated from $N(0,1.2)$ and the correlation between mediators mostly falls between 0.15 and 0.35.

To evaluate the impacts of sample size and potential mediators dimension, we set two sample size levels $n = 300, 500$, and two dimension levels p=1000,10000.

In addition, we take the correlation between mediators into account in the condition of p=1000 dimension. We simulate the strong correlation between mediators by generating the error terms e_k from $N(0, \Sigma_e)$, where $\Sigma_e = \left(\rho^{|k-k'|}\right)_{k,k'}$. It means the correlation between two mediators will decrease as the absolute difference in mediators' subscript $|k - k'|$ increases. We set four correlation levels $\rho = 0, 0.25, 0.5, 0.75$ with dimension p =1000 and sample size $n = 300, 500$. In the simulation

Table 1 TPR and FDP for the four true mediators (M1–M4)

Settings	MT methods[a]	CONF methods[b*]	TPR				Overall TPR	FDP
			M1$\alpha\beta$= 0.16	M2$\alpha\beta$= 0.20	M3$\alpha\beta$= 0.25	M4$\alpha\beta$= 0.30		
$n=300$, $p=1000$	HIMA	RA	0.2340	0.2740	0.4820	0.4880	0.3695	0.0027
		PSR	0.1980	0.2260	0.4380	0.4340	0.3240	0.0031
		IPW	0.2000	0.2060	0.3760	0.4120	0.2985	0.0050
		OW	0.1860	0.2000	0.4140	0.4080	0.3020	0.0033
	mHIMA2	RA	0.3298	0.3510	0.6004	0.5962	0.4693	0.0144
		PSR	0.4754	0.5246	0.7295	0.7520	0.6204	0.0670
		IPW	0.5071	0.5253	0.7333	0.7394	0.6263	0.0909
		OW	0.5396	0.5375	0.7546	0.7789	0.6526	0.0905
$n=300$, $p=10{,}000$	HIMA	RA	0.2660	0.2680	0.4540	0.4940	0.3705	0.0185
		PSR	0.2160	0.2280	0.4040	0.4460	0.3235	0.0046
		IPW	0.2020	0.1820	0.3640	0.4080	0.2890	0.0120
		OW	0.1960	0.2100	0.3840	0.4180	0.3020	0.0082
	mHIMA2	RA	0.3550	0.3203	0.5649	0.5758	0.4540	0.0141
		PSR	0.4714	0.4857	0.6776	0.6918	0.5816	0.0539
		IPW	0.5051	0.5030	0.6707	0.7131	0.5980	0.0800
		OW	0.5466	0.5344	0.7328	0.7672	0.6452	0.0834
$n=500$, $p=1000$	HIMA	RA	0.5440	0.5860	0.8380	0.8740	0.7105	0.0007
		PSR	0.5140	0.5420	0.8160	0.8480	0.6800	0.0022
		IPW	0.4540	0.4480	0.7280	0.7420	0.5930	0.0025
		OW	0.4800	0.4880	0.7680	0.7880	0.6310	0.0024
	mHIMA2	RA	0.6579	0.6984	0.8927	0.9170	0.7915	0.0076
		PSR	0.8060	0.8180	0.9400	0.9540	0.8795	0.0487
		IPW	0.8377	0.8297	0.9198	0.9599	0.8868	0.0679
		OW	0.8620	0.8520	0.9480	0.9760	0.9095	0.0629
$n=500$, $p=10{,}000$	HIMA	RA	0.5440	0.5640	0.8460	0.8360	0.6975	0.0050
		PSR	0.5160	0.5360	0.8220	0.8020	0.6690	0.0000
		IPW	0.4140	0.4480	0.7280	0.7220	0.5780	0.0026
		OW	0.4520	0.4900	0.7920	0.7720	0.6265	0.0016
	mHIMA2	RA	0.6451	0.6326	0.8852	0.8664	0.7573	0.0109
		PSR	0.8116	0.8196	0.9519	0.9479	0.8828	0.0657
		IPW	0.8096	0.8236	0.9419	0.9339	0.8773	0.0774
		OW	0.8417	0.8717	0.9760	0.9559	0.9113	0.0667

[a] MT methods denote two different mediation test approaches, including HIMA Zhang et al. and modified HIMA2 Perera et al. (termed mHIMA2)

[b] CONF methods denote different confounding adjustment methods

[*] RA denotes regression adjustment. PSR denotes propensity score regression adjustment. IPW denotes inverse probability weighting. OW denotes overlapping weighting

setting $\rho = 0, 0.25, 0.5, 0.75$, the corresponding Pearson correlation coefficients between two adjacent mediators are around 0.4, 0.5, 0.7, and 0.8, respectively. We evaluated the performance of the mHIMA2 and PS-based HIMA by conducting 500 replications of simulated data sets for each scenario [9, 14–17, 19, 49].

Simulation results

Simulation results are presented in Tables 1 and 2. Evaluation of the performance of mediator selection of the proposed approach is shown in Table 1 by measuring the true positive rate (TPR) and false discovery proportion (FDP) of selection after the significance test for mediation effects. The mediators have higher TPR as the indirect effect increases (i.e., larger mediation effect, higher detection rate).

As presented in Table 1. Under most settings, the mHIMA2 mediation test approach has a higher TPR than PS-based HIMA while a higher FDP at the same time. Overall, the mHIMA2 is more powerful than the

Table 2 Estimation results of mediation effects, expressed as Mean (MSE)

Settings	MT methods[a]	CONF methods[b*]	M1 $\alpha\beta$= 0.16 (MSE)	M2 $\alpha\beta$= 0.20 (MSE)	M3 $\alpha\beta$= 0.25 (MSE)	M4 $\alpha\beta$= 0.30 (MSE)
$n=300$, $p=1000$	HIMA	RA	0.1547 (0.0043)	0.1909 (0.0067)	0.2414 (0.0065)	0.2907 (0.0084)
		PSR	0.1628 (0.0048)	0.1987 (0.0072)	0.2520 (0.0073)	0.3014 (0.0088)
		IPW	0.1581 (0.0054)	0.1967 (0.0072)	0.2451 (0.0071)	0.2976 (0.0094)
		OW	0.1573 (0.0044)	0.1947 (0.0069)	0.2454 (0.0065)	0.2955 (0.0083)
	mHIMA2	RA	0.1488 (0.0041)	0.1813 (0.0062)	0.2320 (0.0062)	0.2764 (0.0081)
		PSR	0.1487 (0.0042)	0.1850 (0.0064)	0.2337 (0.0063)	0.2837 (0.0083)
		IPW	0.1524 (0.0052)	0.1897 (0.0068)	0.2367 (0.0069)	0.2884 (0.0094)
		OW	0.1522 (0.0043)	0.1879 (0.0064)	0.2369 (0.0064)	0.2871 (0.0083)
$n=500$, $p=1000$	HIMA	RA	0.1596 (0.0024)	0.2016 (0.0038)	0.2458 (0.0040)	0.3040 (0.0046)
		PSR	0.1684 (0.0027)	0.2110 (0.0044)	0.2563 (0.0044)	0.3158 (0.0053)
		IPW	0.1607 (0.0026)	0.2024 (0.0041)	0.2470 (0.0045)	0.3051 (0.0052)
		OW	0.1594 (0.0024)	0.2019 (0.0038)	0.2459 (0.0041)	0.3041 (0.0046)
	mHIMA2	RA	0.1540 (0.0022)	0.1936 (0.0037)	0.2349 (0.0040)	0.2902 (0.0043)
		PSR	0.1548 (0.0023)	0.1966 (0.0037)	0.2392 (0.0040)	0.2977 (0.0045)
		IPW	0.1569 (0.0025)	0.1980 (0.0040)	0.2413 (0.0044)	0.2994 (0.0050)
		OW	0.1556 (0.0023)	0.1976 (0.0038)	0.2402 (0.0040)	0.2986 (0.0045)
$n=300$, $p=10{,}000$	HIMA	RA	0.1217 (0.0045)	0.1514 (0.0061)	0.1873 (0.0088)	0.2278 (0.0119)
		PSR	0.1472 (0.0044)	0.1823 (0.0050)	0.2256 (0.0064)	0.2750 (0.0087)
		IPW	0.1569 (0.0048)	0.1935 (0.0058)	0.2363 (0.0066)	0.2887 (0.0086)
		OW	0.1567 (0.0043)	0.1929 (0.0051)	0.2358 (0.0061)	0.2875 (0.0078)
	mHIMA2	RA	0.1259 (0.0039)	0.1540 (0.0053)	0.1896 (0.0076)	0.2307 (0.0102)
		PSR	0.1350 (0.0039)	0.1698 (0.0049)	0.2089 (0.0064)	0.2565 (0.0083)
		IPW	0.1425 (0.0043)	0.1780 (0.0055)	0.2169 (0.0066)	0.2692 (0.0086)
		OW	0.1425 (0.0039)	0.1780 (0.0049)	0.2170 (0.0062)	0.2670 (0.0080)
$n=500$, $p=10{,}000$	HIMA	RA	0.1502 (0.0022)	0.1911 (0.0037)	0.2364 (0.0040)	0.2836 (0.0055)
		PSR	0.1596 (0.0024)	0.2016 (0.0037)	0.2497 (0.0038)	0.2984 (0.0050)
		IPW	0.1577 (0.0023)	0.2008 (0.0038)	0.2502 (0.0037)	0.2985 (0.0052)
		OW	0.1573 (0.0021)	0.2008 (0.0034)	0.2495 (0.0035)	0.2991 (0.0047)
	mHIMA2	RA	0.1287 (0.0024)	0.1630 (0.0037)	0.2022 (0.0046)	0.2420 (0.0067)
		PSR	0.1402 (0.0021)	0.1805 (0.0032)	0.2243 (0.0035)	0.2719 (0.0048)
		IPW	0.1461 (0.0022)	0.1872 (0.0036)	0.2333 (0.0035)	0.2800 (0.0052)
		OW	0.1456 (0.0020)	0.1871 (0.0031)	0.2324 (0.0033)	0.2808 (0.0047)

[a] MT methods denote two different mediation test approaches, including HIMA and modified HIMA2 (termed mHIMA2).

[b] CONF methods denote different confounding adjustment methods

[*] RA denotes regression adjustment. PSR denotes propensity score regression adjustment. IPW denotes inverse probability weighting. OW denotes overlapping weighting

PS-based HIMA and is less conservative in selecting significant mediators.

As shown in Table 1, for the mHIMA2 mediation test approach, TPR is ranked as OW > IPW > PSR > RA, and FDP is not more than 0.1 and gradually decreases to close to 0.05 as the sample size increases. Among all models, the mHIMA2 mediation test approach with OW adjustment has the highest power and acceptable false positive level. When using the PS-based HIMA mediation test approach, TPR is ranked consistently as RA > PSR > OW > IPW, and all four models also keep FDP at an extremely low level.

Table 2 presents the estimation of mediation effects with the mean and mean square error (MSE). The estimators approach the true values as the mediation effect increases. All models tend to be more accurate as n gets larger and p gets smaller. Overall, the mHIMA2 mediation test approach has a smaller MSE than the PS-based HIMA approach in most cases. RA adjustment has a higher MSE than other adjustment methods especially

Table 3 Baseline characteristics of the HIV-positive patients included in the analysis

Variable		Non-smoker ($N=236$)	Smoker ($N=351$)	Total ($N=587$)	*P*-value
Age		49.72 ± 8.86	49.13 ± 6.69	49.37 ± 7.63	0.391
Race	White	197 (83.5%)	307 (87.5%)	504 (85.9%)	0.008
	Black	21 (8.9%)	36 (10.3%)	57 (9.7%)	
	Others	18 (7.6%)	8 (2.3%)	26 (4.4%)	
ART[a]	Yes	185 (78.4%)	273 (77.8%)	458 (78%)	0.941
	No	51 (21.6%)	78 (22.2%)	129 (22%)	
CD4		0.05 ± 0.05	0.05 ± 0.05	0.05 ± 0.05	0.228
CD8		0.17 ± 0.08	0.18 ± 0.08	0.18 ± 0.08	0.368
NK		0.09 ± 0.06	0.07 ± 0.05	0.07 ± 0.06	< 0.001

[a] *ART* adherence of antiretroviral therapy

when facing the large mediation effect, OW adjustment has the lower MSE among the four adjustment methods.

As shown in Table 2, similarly, the mHIMA2 approach with OW adjustment has the smallest MSE among all models. Moreover, similar results can be seen in the different strong correlation settings in Table S1-S8 in the supplementary file. The mHIMA2 methods have lower MSE (i.e. more precise estimation) and apparently higher TPR. That means the de-biased Lasso technique in mHIMA2 methods performs better when handling the moderate correlation between mediators. However, the FDP of all models slightly increases as the correlation between mediators increases. When correlation among the mediators is strong (for example, $r>0.7$), all models suffer in terms of increased MSE.

Data application

Smoking is an important environmental factor affecting the immune system and blood cell composition [50, 51]. Previous studies have demonstrated smokers had lower natural killer (NK) cell counts and activity [50, 51]. Smoking has also been found to be associated with DNA methylation levels [52]. Meanwhile, DNA methylation levels have also been found to be associated with associated with human NK cell activation [53, 54]. Therefore, DNA methylation may mediate the association between smoking and NK cell level. So we implemented the proposed high-dimensional mediation analysis methods to identify the specific functional CpG sites that may mediate the relationship between smoking and the estimated NK cell level.

Here we apply our method to the GSE117859 dataset obtained from the Gene Expression Omnibus (GEO) database. The aim of the study in which GSE117859 was originally measured is to explore the smoking-associated DNA methylation features linked to AIDS outcomes in the HIV-positive population [55]. The blood samples from the Veteran Aging Cohort Study (VACS) were collected in that study. The HumanMethylation450 BeadChip platform was used to measure the DNA methylation levels.

In total 608 samplesand 485,577 probes were included in the dataset. Clinical information such as age, sex, race, smoking history, adherence of antiretroviral therapy (ART), estimated CD4 T cells, estimated CD8 T cells, and estimated NK cells were collected. The estimated CD4/CD8/NK were obtained using a methylation-based cell type deconvolution algorithm proposed by Housman et al. [56]. To some extent, the estimated CD4 and CD8 levels can represent AIDS severity.

Smoking status was collected based on self-report. All included patients were classified into the smoker and the non-smoker groups according to their reported smoking history. After removing the individuals without available clinical information and DNAm sites with missing values, a total of 587 samples and 485,503 probes were included in the analysis.

We adjusted the potential confounders including age, race, adherence of antiretroviral therapy, estimated CD4 T cells, and estimated CD8 T cells. Demographic and clinical variables included in our analysis are presented in Table 3.

The analysis results using the proposed mHIMA2 method are presented in Table 4. Here, we mainly presented the CpGs mediators with a total effect proportion greater than 5%. Due to the limitation of text content, we didn't present the whole summary results of the PS-based HIMA method, but that can be seen in Table S9 in the supplementary file.

As shown in Table 4, we identified two methylation sites cg13917614 in CNP gene and cg16893868 in LILRA2 gene by most of mHIMA2 based methods. The similar result can be seen in Table S9 in the supplementary file. The existing studies have already demonstrated the site cg13917614 is associated with smoking [52, 57]. Although we don't find direct evidence that the CNP gene is associated with immune function based on the

Table 4 Summary of the selected CpGs mediators with a %TE > 5 by the mHIMA2 models

Method[a]	CpG	Chrom	Gene	$\widehat{\alpha}$	$\widehat{\beta}$	%TE[b]	*P*-value
mHIMA2-RA	cg20460771	1	PTAFR	0.0211	-0.0572	5.2398	0.0019
	cg06040872	17	CCL18	0.0157	-0.1092	7.4323	0.0002
	cg16893868	19	LILRA2	-0.0155	0.1015	6.8344	< 0.0001
mHIMA2-PSR	cg13917614	17	CNP	0.0230	-0.1936	19.3678	0.0001
	cg03164561	2	NMUR1	0.0149	-0.1146	7.4161	0.0015
	cg03605454	4	RP11-526A4.1	0.0154	-0.1033	6.9094	0.0028
	cg09529165	17	-	0.0198	-0.2038	17.5660	< 0.0001
	cg01500140	19	LIM2	0.0157	-0.1393	9.5028	0.0013
	cg16893868	19	LILRA2	-0.0155	0.1117	7.5217	0.0012
mHIMA2-IPW	cg13917614	17	CNP	0.0230	-0.0717	7.1289	0.0001
	cg16893868	19	LILRA2	-0.0157	0.0913	6.1962	0.0001
mHIMA2-OW	cg13917614	17	CNP	0.0230	-0.0737	7.3785	0.0001
	cg16893868	19	LILRA2	-0.0155	0.0916	6.1681	0.0001

[a] Method denotes the combination of the mHIMA2 approach and different confounding adjustment methods. Such as mHIMA2-OW denotes the mHIMA2 method with overlapping weighting

[b] %TE total effect proportion

- No related genes were found

existing literature, relevant studies showed a link between CNP and inflammatory responses in which the mechanism remains further study [58, 59].

The encoded protein of the LILRA2 gene can suppress innate immune response [60, 61]. The results reveal that smoking will promote the demethylation of cg16893868, leading to an increase in gene LILRA2 expression and ultimately reducing the estimated NK cell level. It has been found that the remaining CpG sites cg20460771, cg03164561, cg03605454, cg09529165, and cg01500140 are all associated with smoking [11, 52, 62–64]. Further insights into the discovered CpG mediators in genome-wide epigenetic studies will be meaningful.

Discussion

The causal relationship obtained in high-dimensional mediation analysis usually depends on no-confounding assumption. However, confounding is almost inevitable in observational studies owing to the lack of randomization of the baseline covariates in practice. Previous studies show the utilization of PS method such as PS-adjustment and IPW in high-dimensional mediation analysis, but those face the issue of extreme PS distribution.

In this article, we integrated OW approach into the high-dimensional mediation model, which can address extreme PS distribution and better adjust for confounding. Finally, we developed a high-dimensional mediation analysis workflow consisting of OW confounding adjustment, SIS, de-biased Lasso penalization for potential mediator screening, and the high-dimensional mediation test underlying the mixture null distribution of *P*-values.

Simulation results indicate that the mHIMA2 with OW approach presented in this study performs best among all the compared models with the highest TPR, acceptable FDP level, and the smallest MSE in mediating effect estimation. In addition, the mHIMA2 embedded de-biased Lasso method performs better when moderate correlations between mediators exist.

Simulation study also suggestedthe proposed method would perform better when the sample size was increased. This result suggests that when the proposed method is used for the analysis of mediating effects on real data, a sufficient sample size should also be ensured. Such a feature is also consistent with other existing methods [5, 9, 14, 17, 19, 49]. Furthermore, the dimensionality of potential mediators has little effect on the performance of the proposed method.

In most of the previous studies [5, 9, 13, 17], it didn't take confounding adjustment into account in the SIS process. However, we adopted the PS-based method to adjust confounding, thus improving the accuracy of mediators screening. Moreover, it has been assumed that mediators are linearly independent of each other, but such an assumption is often not strictly valid in real data. The violation of the mediators' independence assumption often affects the accuracy of mediators selection and precision of mediating effect estimation. The proposed method can effectively deal with this issue which can tolerate the correlation between the mediators and ensure the robustness of mediators selection, multiple mediation testing, and mediating effect estimation.

Similar to other two-step approaches, the error of the first model may be introduced and cumulated in the second step, because the first-step can not quarantee 100% correctness. To avoid this, we set a relatively loose screening criterion with $d = 2n/log(n)$ to select the top d largest effect mediators [15–17, 49] in the first step to control false negative while avoiding the increase of false positive error according to the application recommendation of SIS approach. Though the errors cannot be totally avoid, this can reduce the error in the preliminary screening of mediators and prevent serious error cumulation in the second step to some extend. As shown in the simulation, the proposed two-step model performed well. Besides, previous published studies also have demonstrated the error cumulation issue in two-step models can be controlled well in the similar way as we did, and well not cause serious bias in the final results [14, 65–70].

Meanwhile, we applied the proposed method to the dataset GSE117859 obtained from the GEO databases and identified several significant DNAm mediators, including the sites cg13917614, cg16893868, cg20460771, cg03164561, cg03605454, cg09529165, and cg01500140. Among them, site cg16893868 in LILRA2 gene has been demonstrated to be associated with smoking and immune function [60, 61]. That indicates that the proposed method can identify reliable mediators in empirical data analysis.

The presence of confounding in observation studies always is a major challenge to obtaining causal relationships. Currently, most genetic studies are based on observational research without randomization of baseline characteristics. Particularly, the high-dimensional mediation analysis always faces some issues, such as the accuracy of the high-dimensional mediation selection and the low power of multiple mediation test [13, 14, 17, 18]. Although the utilization of PSR and IPW offers a solution of confounding adjustment in classical HDMA workflow, it still faces the issue of extreme PS distribution.

The proposed OW-based method can provide a more precise and stable mediating effect estimation. However, the misspecification of the outcome model and PS model can not be avoid in practice. Hence, the doubly robust methods may be desirable to be applied in HDMA workflow in future study. Even if the JS-mixture method was proposed to improve the power of multiple mediation testing, other more powerful test methods still are appealing in large-scale genome-wide epigenetic studies [13, 18]. Conducting further simulation and methodology studies to compare different powerful test methods may provide useful reference for future studies. It should also be noticed that the existence of unmeasured confounding is out of the scope of this paper. Previews published studies have provided serval applicable methods to deal with this issue [49, 71].

Conclusion

Overall, the mHIMA2 with OW adjustment has sufficient power in selecting potential true mediators and obtaining precise estimation for mediation effects. It can be recommended in practical high-dimensional mediation analysis, especially in epigenetic study.

Abbreviations

FDP	False discovery proportion
GEO	Gene Expression Omnibus
HDMA	High-dimensional mediation analysis
IPW	Inverse probability weighting
JS-uniform	Joint significant test with uniform distribution
JS-mixture	Joint significance test with mixture null distribution
MSE	Mean square error
mHIMA2	Modified HIMA2 model
NK cell	Natural killer cell
OW	Overlapping weighting
PSR	Propensity score regression adjustment
PS	Propensity score
SIS	Sure independence screening
TPR	True positive rate

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02254-x.

Additional file 1: High-dimensional mediation analysis for continuous outcome with confounders using overlap weighting method in observational epigenetic study. Simulation results of different correlation levels ρ =0,0.25,0.5,0.75 with dimension p=1000 and sample size n=300,500 were presented in the Table S1-8. Analysis result using the PS-based HIMA methods was shown in Table S9.

Acknowledgements

We acknowledge GEO database for providing their platforms and contributors for uploading their meaningful datasets.

Authors' contributions

F.C., H.Y. and W.H. led the conception and design of the work. W.H. and S.C. conducted the simulation study. W.H., S.C., and J.C. finished data cleaning and implemented real data application. W.H. completed the original draft. F.C., J.C., and Y.Y. guided analyses and provided advice. F.C. critically reviewed and edited the manuscript.

Funding

This work was supported by the National Social Science Found of China (21CTJ009) and National Nature Science Foundation of China (81703325).

Availability of data and materials

The dataset GSE117859 obtained from GEO database in our real data analysis can be accessed (at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117859) without limitation. Our procedure is implemented using the R software. The corresponding R code can be found at https://github.com/huww1998/CONF_mHIMA2.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University, Xi'an 710061, Shaanxi, China. [2]Department of Radiology, First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, Shaanxi, China.

Received: 15 March 2024 Accepted: 22 May 2024
Published online: 03 June 2024

References

1. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Personal Soc Psychol. 1986;51(6):1173–82. https://doi.org/10.1037//0022-3514.51.6.1173.
2. Huan T, Joehanes R, Schurmann C, Schramm K, Pilling LC, Peters MJ, et al. A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking. Hum Mol Genet. 2016;25(21):4611–23. https://doi.org/10.1093/hmg/ddw288 .
3. MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. Psychol Methods. 2002;7(1):83–104. https://doi.org/10.1037/1082-989x.7.1.83.
4. Biesanz JC, Falk CF, Savalei V. Assessing Mediational models: testing and interval estimation for Indirect effects. Multivar Behav Res. 2010;45(4):661–701. https://doi.org/10.1080/00273171.2010.498292.
5. Gao Y, Yang H, Fang R, Zhang Y, Goode EL, Cui Y. Testing mediation effects in high-dimensional epigenetic studies. Front Genet. 2019;10:1195.https://doi.org/10.3389/fgene.2019.01195.
6. Taylor AB, MacKinnon DP. Four applications of permutation methods to testing a single-mediator model. Behav Res Methods. 2012;44(3):806–44. https://doi.org/10.3758/s13428-011-0181-x.
7. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. Epidemiol (Cambridge Mass). 2009;20(1):18–26. https://doi.org/10.1097/EDE.0b013e31818f69ce .
8. VanderWeele TJ, Vansteelandt S. Mediation analysis with multiple mediators. Epidemiol Methods. 2014;2(1):95–115. https://doi.org/10.1515/em-2012-0010 .
9. Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinf (Oxford England). 2016;32(20):3150–4. https://doi.org/10.1093/bioinformatics/btw351 .
10. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–95. https://doi.org/10.1016/j.ygeno.2011.07.007.
11. Harlid S, Xu Z, Panduri V, Sandler DP, Taylor JA. CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study. Environ Health Perspect. 2014;122(7):673–8. https://doi.org/10.1289/ehp.1307480.
12. Toyooka S, Maruyama R, Toyooka KO, McLerran D, Feng Z, Fukuyama Y, et al. Smoke exposure, histologic type and geography-related differences in the methylation profiles of non-small cell lung cancer. Int J Cancer. 2003;103(2):153–60. https://doi.org/10.1002/ijc.10787 .
13. Liu Z, Shen J, Barfield R, Schwartz J, Baccarelli AA, Lin X. Large-scale hypothesis testing for Causal Mediation effects with Applications in Genome-wide epigenetic studies. J Am Stat Assoc. 2022;117(537):67–81. https://doi.org/10.1080/01621459.2021.1914634.
14. Luo L, Yan Y, Cui Y, Yuan X, Yu Z. Linear high-dimensional mediation models adjusting for confounders using propensity score method. Front Genet. 2022;13:961148. https://doi.org/10.3389/fgene.2022.961148.
15. Luo C, Fa B, Yan Y, Wang Y, Zhou Y, Zhang Y, et al. High-dimensional mediation analysis in survival models. PLoS Comput Biol. 2020;16(4):e1007768. https://doi.org/10.1371/journal.pcbi.1007768.
16. Yu Z, Cui Y, Wei T, Ma Y, Luo C. High-dimensional mediation analysis with confounders in Survival models. Front Genet. 2021;12:688871. https://doi.org/10.3389/fgene.2021.688871.
17. Perera C, Zhang H, Zheng Y, Hou L, Qu A, Zheng C, et al. HIMA2: high-dimensional mediation analysis and its application in epigenome-wide DNA methylation data. BMC Bioinformatics. 2022;23(1):296. https://doi.org/10.1186/s12859-022-04748-1.
18. Dai JY, Stanford JL, LeBlanc M. A multiple-testing procedure for high-dimensional mediation hypotheses. J Am Stat Assoc. 2022;117(537):198–213. https://doi.org/10.1080/01621459.2020.1765785.
19. Zhang H, Zheng Y, Hou L, Zheng C, Liu L. Mediation analysis for survival data with high-dimensional mediators. Bioinf (Oxford England). 2021;37(21):3815–21. https://doi.org/10.1093/bioinformatics/btab564 .
20. Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. Eur Heart J. 2011;32(14):1704–8. https://doi.org/10.1093/eurheartj/ehr031.
21. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Science: Rev J Inst Math Stat. 2010;25(1):1–21. https://doi.org/10.1214/09-STS313 .
22. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. PLoS One. 2011;6(3):e18174. https://doi.org/10.1371/journal.pone.0018174.
23. Li F, Thomas LE, Li F. Addressing Extreme Propensity scores via the Overlap weights. Am J Epidemiol. 2019;188(1):250–7. https://doi.org/10.1093/aje/kwy201 .
24. Thomas LE, Li F, Pencina MJ. Overlap weighting: a propensity score method that mimics attributes of a Randomized Clinical Trial. JAMA. 2020;323(23):2417. https://doi.org/10.1001/jama.2020.7819.
25. Mlcoch T, Hrnciarova T, Tuzil J, Zadak J, Marian M, Dolezal T. Propensity Score Weighting Using Overlap Weights: A New Method Applied to Regorafenib Clinical Data and a Cost-Effectiveness Analysis. Value in Health. 2019;22(12):1370-7. doi: 10.1016/j.jval.2019.06.010.
26. Vanderweele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiol (Cambridge Mass). 2014;25(2):300–6. https://doi.org/10.1097/EDE.0000000000000034 .
27. Perera C, Zhang H, Zheng Y, Hou L, Qu A, Zheng C, et al. HIMA2: high-dimensional mediation analysis and its application in epigenome-wide DNA methylation data. BMC Bioinformatics. 2022;23(1). https://doi.org/10.1186/s12859-022-04748-1.
28. Basu D. Randomization analysis of Experimental Data: the Fisher randomization test. J Am Stat Assoc. 1980;75(371):575–82. https://doi.org/10.1080/01621459.1980.10477512.
29. Rubin DB. Comment. J American Statis Assoc. 1986;81(396):961–2. https://doi.org/10.1080/01621459.1986.10478355.
30. Imbens GW, Rubin DB. Causal inference for statistics, Social, and Biomedical sciences: an introduction. Cambridge: Cambridge University Press; 2015.
31. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55. https://doi.org/10.1093/biomet/70.1.41.
32. Haukoos JS, Lewis RJ. The Propensity score. JAMA. 2015;314(15):1637–8. https://doi.org/10.1001/jama.2015.13480.
33. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010;29(3):337–46. https://doi.org/10.1002/sim.3782.
34. Abdia Y, Kulasekera KB, Datta S, Boakye M, Kong M. Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: a comparative study. Biometrical J Biometrische Z. 2017;59(5):967–85. https://doi.org/10.1002/bimj.201600094 .
35. Rosenbaum PR, Rubin DB. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. Am Stat. 1985;39(1):33–8. https://doi.org/10.1080/00031305.1985.10479383.
36. Rosenbaum PR, Rubin DB. Reducing Bias in Observational studies using subclassification on the Propensity score. J Am Stat Assoc. 1984;79(387):516–24. https://doi.org/10.1080/01621459.1984.10478078.
37. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc. 1995;90(429):106–21. https://doi.org/10.1080/01621459.1995.10476493.

38. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res. 2011;46(3):399–424. https://doi.org/10.1080/00273171.2011.568786.
39. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via Propensity score weighting. J Am Stat Assoc. 2018;113(521):390–400. https://doi.org/10.1080/01621459.2016.1260466.
40. Fan J, Lv J. Sure Independence Screening for Ultrahigh Dimensional Feature Space. J Royal Stat Soc Ser B: Stat Methodol. 2008;70(5):849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x .
41. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. Annals Stat. 2010;38(2):894–942. https://doi.org/10.1214/09-AOS729 .
42. Maity AK, Basu S. Highest posterior model computation and variable selection via simulated annealing. The New England J Statis Data Sci. 2023;1(2):200–7. https://doi.org/10.51387/23-NEJSDS40.
43. Breheny P, Huang J. Coordinate Descent algorithms for Nonconvex Penalized Regression, with applications to Biological feature selection. Annals Appl Stat. 2011;5(1):232–53. https://doi.org/10.1214/10-AOAS388 .
44. Huang Y-T, Pan W-C. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. Biometrics. 2016;72(2):402–13. https://doi.org/10.1111/biom.12421.
45. Benjamini Y, Hochberg Y. Controlling the false Discovery rate: a practical and powerful Approach to multiple testing. J Roy Stat Soc: Ser B (Methodol). 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
46. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800–2. https://doi.org/10.1093/biomet/75.4.800.
47. Huang Y-T. Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics. Annals Appl Stat. 2018;12(3):1535–57. https://doi.org/10.1214/17-AOAS1120.
48. Fang EX, Ning Y, Liu H. Testing and confidence intervals for high Dimensional Proportional hazards Model. J Royal Stat Soc Ser B Stat Methodol. 2017;79(5):1415–37. https://doi.org/10.1111/rssb.12224 .
49. Chen F, Hu W, Cai J, Chen S, Si A, Zhang Y, et al. Instrumental variable-based high-dimensional mediation analysis with unmeasured confounders for survival data in the observational epigenetic study. Front Genet. 2023;14:1092489. https://doi.org/10.3389/fgene.2023.1092489.
50. Qiu F, Liang C-L, Liu H, Zeng Y-Q, Hou S, Huang S, et al. Impacts of cigarette smoking on immune responsiveness: up and down or upside down? Oncotarget. 2017;8(1):268–84. https://doi.org/10.18632/oncotarget.13613.
51. Elisia I, Lam V, Cho B, Hay M, Li MY, Yeung M, et al. The effect of smoking on chronic inflammation, immune function and blood cell composition. Sci Rep. 2020;10:19480. https://doi.org/10.1038/s41598-020-76556-7.
52. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet. 2011;88(4):450–7. https://doi.org/10.1016/j.ajhg.2011.03.003 .
53. Wiencke JK, Butler R, Hsuang G, Eliot M, Kim S, Sepulveda MA, et al. The DNA methylation profile of activated human natural killer cells. Epigenetics. 2016;11(5):363–80. https://doi.org/10.1080/15592294.2016.1163454.
54. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clin Epigenetics. 2015;7:113. https://doi.org/10.1186/s13148-015-0148-3.
55. Zhang X, Hu Y, Aouizerat BE, Peng G, Marconi VC, Corley MJ, et al. Machine learning selected smoking-associated DNA methylation signatures that predict HIV prognosis and mortality. Clin Epigenetics. 2018;10(1):155. https://doi.org/10.1186/s13148-018-0591-z.
56. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13(1):86. https://doi.org/10.1186/1471-2105-13-86.
57. Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet. 2012;21(13):3073–82. https://doi.org/10.1093/hmg/dds135.
58. Bao Q, Zhang B, Zhou L, Yang Q, Mu X, Liu X, et al. CNP Ameliorates Macrophage Inflammatory Response and Atherosclerosis. Circ Res. 0(0). https://doi.org/10.1161/CIRCRESAHA.123.324086.
59. Bae C-R, Hino J, Hosoda H, Arai Y, Son C, Makino H, et al. Overexpression of C-type natriuretic peptide in endothelial cells protects against Insulin Resistance and inflammation during Diet-induced obesity. Sci Rep. 2017;7(1):9807. https://doi.org/10.1038/s41598-017-10240-1.
60. Lu HK, Mitchell A, Endoh Y, Hampartzoumian T, Huynh O, Borges L, et al. LILRA2 selectively modulates LPS-mediated cytokine production and inhibits phagocytosis by monocytes. PLoS ONE. 2012;7(3):e33478. https://doi.org/10.1371/journal.pone.0033478.
61. Lewis Marffy AL, McCarthy AJ. Leukocyte Immunoglobulin-Like receptors (LILRs) on human neutrophils: modulators of infection and immunity. Front Immunol. 2020;11:857. https://doi.org/10.3389/fimmu.2020.00857.
62. Sikdar S, Joehanes R, Joubert BR, Xu C-J, Vives-Usano M, Rezwan FI, et al. Comparison of smoking-related DNA methylation between newborns from prenatal exposure and adults from personal smoking. Epigenomics. 2019;11(13):1487–500. https://doi.org/10.2217/epi-2019-0066.
63. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic signatures of cigarette smoking. Circulation Cardiovasc Genet. 2016;9(5):436–47. https://doi.org/10.1161/CIRCGENETICS.116.001506 .
64. Sun YV, Smith AK, Conneely KN, Chang Q, Li W, Lazarus A, et al. Epigenomic association analysis identifies smoking-related DNA methylation sites in African americans. Hum Genet. 2013;132(9):1027–37. https://doi.org/10.1007/s00439-013-1311-6.
65. Luo C, Wang G, Hu F. Two-Step Gene Feature Selection Algorithm Based on Permutation Test. In: Yao J, Yang Y, Słowiński R, Greco S, Li H, Mitra S, et al., editors. Springer; 2012. p. 249—58. https://doi.org/10.1111/ppe.12382.
66. Liu D, Yeung EH, McLain AC, Xie Y, Buck Louis GM, Sundaram R. A two-step Approach for Analysis of Nonignorable Missing outcomes in Longitudinal Regression: an application to Upstate KIDS Study. Paediatr Perinat Epidemiol. 2017;31(5):468–78. https://doi.org/10.1111/ppe.12382.
67. Newcombe PJ, Connolly S, Seaman S, Richardson S, Sharp SJ. A two-step method for variable selection in the analysis of a case-cohort study. Int J Epidemiol. 2018;47(2):597–604. https://doi.org/10.1093/ije/dyx224.
68. Song J, Shin SJ. A two-step approach for variable selection in linear regression with measurement error. Commun Stat Appl Methods. 2019;26(1):47–55. https://doi.org/10.29220/CSAM.2019.26.1.047 .
69. Liu Y, Qin SJ. A Novel two-step sparse Learning Approach for Variable Selection and Optimal Predictive modeling. IFAC-PapersOnLine. 2022;55(7):57–64. https://doi.org/10.1016/j.ifacol.2022.07.422.
70. Chamlal H, Benzmane A, Ouaderhman T. A Two-Step Feature Selection Procedure to Handle High-Dimensional Data in Regression Problems. 2023 International Conference on Decision Aid Sciences and Applications (DASA). 2023. p. 592–6.
71. Wickramarachchi DS, Lim LHM, Sun B. Mediation analysis with multiple mediators under unmeasured mediator-outcome confounding. Stat Med. 2023;42(4):422–32. https://doi.org/10.1002/sim.9624.

Publisher's Note